D212–D220 *Nucleic Acids Research, 2019, Vol. 47, Database issue* *Published online 5 November 2018*
doi: 10.1093/nar/gky1077

RegulonDB v 10.5: tackling challenges to unify classic and high throughput knowledge of gene regulation in *E. coli* K-12

Alberto Santos-Zavaleta[1], Heladia Salgado[1], Socorro Gama-Castro[1], Mishael Sánchez-Pérez[1], Laura Gómez-Romero[1], Daniela Ledezma-Tejeida[1], Jair Santiago García-Sotelo[1], Kevin Alquicira-Hernández[1], Luis José Muñiz-Rascado[1], Pablo Peña-Loredo[1], Cecilia Ishida-Gutiérrez[1], David A. Velázquez-Ramírez[1], Víctor Del Moral-Chávez[1], César Bonavides-Martínez[1], Carlos-Francisco Méndez-Cruz[1], James Galagan[2] and Julio Collado-Vides[1,2,*]

[1]Centro de Ciencias Genómicas, Universidad Nacional Autónoma de México, Cuernavaca, Morelos 62210, México and [2]Department of Biomedical Engineering, Boston University, Boston, MA, USA

Received September 18, 2018; Revised October 16, 2018; Editorial Decision October 18, 2018; Accepted October 19, 2018

ABSTRACT

RegulonDB, first published 20 years ago, is a comprehensive electronic resource about regulation of transcription initiation of *Escherichia coli* K-12 with decades of knowledge from classic molecular biology experiments, and recently also from high-throughput genomic methodologies. We curated the literature to keep RegulonDB up to date, and initiated curation of ChIP and gSELEX experiments. We estimate that current knowledge describes between 10% and 30% of the expected total number of transcription factor- gene regulatory interactions in *E. coli*. RegulonDB provides datasets for interactions for which there is no evidence that they affect expression, as well as expression datasets. We developed a proof of concept pipeline to merge binding and expression evidence to identify regulatory interactions. These datasets can be visualized in the RegulonDB JBrowse. We developed the Microbial Conditions Ontology with a controlled vocabulary for the minimal properties to reproduce an experiment, which contributes to integrate data from high throughput and classic literature. At a higher level of integration, we report Genetic Sensory-Response Units for 200 transcription factors, including their regulation at the metabolic level, and include summaries for 70 of them. Finally, we summarize our research with Natural language processing strategies to enhance our biocuration work.

INTRODUCTION

RegulonDB is a database that offers, in an organized and computable form, the accumulated knowledge obtained through decades of experimentation in many different laboratories around the world, on transcriptional regulation in *Escherichia coli* K-12. It was first published 20 years ago, in 1998 (1), and since then we have periodically published progress reports for our work in database issues of *Nucleic Acids Research*. Our curation is shared with the EcoCyc database (2), which together with RegulonDB provide the major up to date resources of organized knowledge for the best-known bacterial genome model organism.

The major avenues of recent progress are the following. We have made important progress in implementing different components that allowed us to expand RegulonDB to include high throughput (HT)-generated knowledge. This included the design and implementation of the Microbial Conditions Ontology (MCO), which provides a formal framework and defines the set of properties necessary to specify the conditions as well as the genetic material used in a particular study, in order to adequately describe how an experiment was performed in a way that should satisfy its reproducibility. This was inspired by suggestions made by Fred Neidhardt years ago (3). In parallel, we have made progress in the curation of HT-generated literature, particularly binding sites identified from gSELEX and ChIP types of experiments, in conjunction with the corresponding expression profile experiments (4). Here, we report the initial construction of a semi-automatic pipeline that incorporates both binding and expression datasets in order to identify those transcription factor (TF) binding sites (TFBSs) upstream of genes that show change in expression under sim-

[*]To whom correspondence should be addressed. Tel: +52 777 3132063; Fax: +52 777 3175581; Email: collado@ccg.unam.mx

ilar conditions and therefore jointly support evidence for a regulatory interaction (RI).

A different level of integration is offered by GENSOR Units, an acronym for Genetic Sensory-Response Units, in which we gather in a single network the signal, the transduction pathway to the effector that binds the TF, the regulated genes, the corresponding regulated products, and the corresponding metabolic reactions. The description of the construction and analysis of such units can be found elsewhere (5). Recently, we expanded these GENSOR Units to include regulation at the metabolic level, and we have finished both a mechanistic and a physiological written summary for 67 of them. Finally, we summarized our work with strategies using natural language processing (NLP) to facilitate the identification of experimental variables in growth conditions reported in the literature, as well as our work on the partial reconstruction of TF summaries.

MATERIALS AND METHODS

Curation of literature and HT datasets

Original scientific papers are collected every month based on a set of keywords focused on transcription and gene regulation; reading the abstracts we select which to curate, and the information of pertinent papers is curated through EcoCyc capture forms and then transferred to RegulonDB, keeping both databases synchronized with exactly the same data. For the HT data, publications are gathered from PubMed using keywords related to HT methodologies; 'coli' is searched in the title or in the abstract and different synonyms of methods are searched in all fields of publications. Frequently, these publications include additional experimental characterization for a subset of sites based on classic methods. More details about the annotation process can be found at, 'Annotation Process for HT experiments' from the 'Doc & Help' menu.

HT dataset processing

Datasets were obtained from GEO (6). The metadata describing the experimental conditions were retrieved from the corresponding SOFT files (Supplementary file S1). Datasets were chosen because an expression experiment and a TF binding experiment were conducted as part of each dataset. RNA sequencing (RNA- seq) files contain a summary of the expression for each experiment (i.e. RPKM, FPKM, RPM) for each gene. TF binding experiments included ChIP-exo and ChIP-seq datasets that contained intensity values. All coordinates were translated to NC000913.3 using a coordinate dictionary obtained with ECOCYC map-seq-coords tool (https://biocyc.org/ECOLI/map-seq-coords-form?chromosome=COLI-K12) (2). We used StrandedPlot [https://doi.org/10.1101/212654] to visualize intensity values. The processing of files is shown in Supplementary Figure S1.

Acronyms

ChIP, Chromatin immunoprecipitation

ChIP-seq, Chromatin immunoprecipitation (ChIP)-sequencing

ChIP-exo, Chromatin immunoprecipitation (ChIP) combined with lambda exonuclease digestion followed by high-throughput sequencing

gSELEX, Genomic systematic evolution of ligands by exponential enrichment

HT, High throughput

RIs, Regulatory interactions

TFBS, Transcription factor binding site

TF, Transcription factor

GENSOR Unit, Genetic Sensory-Response Unit

Reconstructing regulatory interactions

From the intensity files, macs2 software (7) and BEDOPS (8) commands were run to obtain enrichment regions. These sequences were processed with matrix scan from RSAT tools (9) to obtain putative TFBSs. A HT TFBS and a RegulonDB TFBS were considered the same if they overlapped in >50% of their length. Conditions or phenotypes of RNA-seq experiments were identified to generate a condition contrast. The two conditions in RNA-seq datasets were compared with a two-sample *t*-test and the resulting p-values were adjusted for the false discovery rate (FDR). The $\log_2$ fold change across conditions and the FDR *P*-values were used to construct a volcano plot (10) highlighting the differentially expressed genes (FDR < 5%). To identify RIs, the TFBSs were mapped, when possible, to the regulatory regions of *Escherichia coli* genes using an *ad hoc* python script. A regulatory region per gene per TF was defined either by the distance to the farthest known TFBS (with strong evidence) of such TF or as the interval between −400 and +100 base-pairs with respect to transcription initiation. The processing of files is shown in Supplementary Figure S1.

GENSOR unit assembly

Each GENSOR Unit is centered on one TF. From RegulonDB v10.0, we automatically retrieved its known effectors, active and inactive conformations, its regulated genes, and the effects of the RIs. From Ecocyc (2), we retrieved gene products of the regulated genes; if they were enzymes, we also retrieved the catalyzed reactions, their substrates and products. If encoded proteins were part of a heteromultimeric complex, we included other proteins participating in the complex, even if they were not regulated by the main TF. Additionally, we added Complementary Pathway Reactions. These reactions link pairs of metabolites that are present in the same metabolic pathway. We included this information through a single reaction, i.e. a Complementary Pathway Reaction, that represents all the reactions (one or more) necessary to create a metabolic flux from one molecule to the other. No new metabolites are included, only new connections between those already present in the GENSOR Unit. The detailed assembly algorithm can be found in Ledezma-Tejeida *et al.* (5). Recently, we added all the regulation at the enzymatic level reported in EcoCyc between metabolites and proteins present in the GENSOR Unit. Figures were automatically generated and manually edited in the software CellDesigner v4.4 (11).

RESULTS

Classic curation progress

Here, we report RegulonDB version 10.5 released on September 2018. We have kept the database up to date with the literature of classic experiments using molecular biology methodologies. Supplementary Tables S1 and S2 show the new TFs identified as well as those complexes whose crystal structure has been characterized since the last release in 2016. In addition, our daily curation has increased the number of promoters, transcription units, and overall knowledge of the regulatory network as shown in Figure 1A. Although most of curation entails data from classic experiments, information on HT-supported promoters and recently added RIs show the impact of their contribution (Figure 1B). Figure 1C shows the year of publications contained in RegulonDB. The expansion of knowledge of transcriptional regulation can be appreciated in Figure 1D, as more relevant objects have been added through the years. As mentioned below, we have already initiated the curation of some of the publications reporting binding of TFs using HT-methodologies. This initial inspection is used in the following to estimate the number of RIs in the complete *E. coli* genome.

Estimating the size of the complete regulatory network

The regulatory network comprising the collection of TF–gene interactions of *E. coli* is among the most intensively studied, given the legacy of experiments performed in this bacterial model organism. Previous publications have estimated that RegulonDB has information for around 15–25% of the complete regulatory network (12), but this was before HT technologies showed the existence of many more binding sites for TFs in the genome.

RegulonDB lists evidence for the binding of at least one site for 207 (69%) TFs out of the roughly 300 total TFs we estimated in *E. coli* (13). These 207 TFs affect a total of 1823 genes via 4358 TF–gene interactions; including interactions by CRP that affects 522 genes, down to 10 unique TFs that affect only one gene each. We have known since a while ago that the connectivity follows a power-law distribution (14). Given the currently available HT datasets of binding experiments for around 15 TFs in *E. coli,* including local and global TFs, we estimated the average fold increase in the number of TF–gene interactions in RegulonDB to be close to 7 compared to those from HT experiments (see Table 5 in (4)). Therefore, the estimated total number of interactions for 200 TFs in RegulonDB is 7 times the current one of 4358, giving 30 506 interactions. As mentioned, *E. coli* is estimated to have around 300 TFs, assuming the same average connectivity for these missing TFs, the total number of estimated interactions for the complete set of TFs in the *E. coli* genome adds to ∼45 759 TF–gene interactions. This is the upper-bound estimate, since unknown TFs fall within the larger group of unknown genes which have been observed to be in the low expression range, and thus to be highly specific TFs with a smaller number of regulated target genes than the current average (15,16). Furthermore, global ChIP-seq experiments generate a large number of binding sites at many different positions in the genome, with many lacking a correlation with change in expression of the downstream gene (compare Tables 4 and 5 in (4)). If we take only those clearly involved in transcriptional regulation, i.e. those that result from our manual curation, the fold increase is only 2. The upper-bound estimate indicated that our set of RIs accounts for 9.5% of the total set of interactions in the genome, although many of them will not be easily associated with regulation of transcription initiation. Of these, we estimated 33.3% of interactions will affect the expression of downstream genes.

We conclude that *E. coli* may be subject to around 46 000 TF–gene RIs, of which we know ∼10%; however, of these 46 000 RIs, only around 13 000 will be clearly involved in regulation of transcription initiation, of which we already know a third of them. These numbers and the increased HT-derived knowledge have motivated a good amount of the recent progress in RegulonDB that we report in this publication.

Expanding our curation to HT technologies

We have progressed in expanding our curation to include publications based on HT technologies (4,17,18). This has been quite a challenge, since we do not want to dilute the solid knowledge accumulated through the years from molecular biology experiments with the new information from HT methodologies. We published years ago our own protocol of classification of types of HT evidence (19). We have now recently curated around 51 HT publications, accounting for 1048 new RIs of 9 TFs, in addition to 107 previously known RIs. These publications generated 16 609 interactions of 36 TFs and sigma factors that have some missing information and therefore are included only as datasets containing what we call plain ‘interactions’ (as opposed to ‘regulatory interactions’), for which no evidence is yet available supporting their regulatory role (4). Figure 2 shows that both HT-curated RIs and HT datasets are available in RegulonDB as a result of our curation work.

Preliminary processing and inference of regulatory interactions from HT binding and expression datasets

In addition to curating the literature, we have initiated the direct processing of datasets available in GEO. Here, we report the processing of an initial set of selected HT experiments that have both binding and expression evidence under comparable growth conditions, such that we can combine them and identify those cases where binding correlates with a change in expression of the downstream gene. The adequate integration of these two data sets can predict novel RIs, provided there is a careful normalization of the data and proper identification of the experimental conditions. Properly processed datasets can be displayed in tracks of the RegulonDB JBrowse.

As a proof of concept, we included five GEO series that contain 32 samples, each one corresponding to a dataset in RegulonDB, including a variety of methods (ChIP-exo, ChIP-chip, ChIP-seq; RNA-seq and microarrays) (20–23). This collection includes experiments for Fur, Cra, OxyR, SoxR, GadE, GadW, GadX and OmpR. All of them can be visualized using JBrowse, which can be accessed from the

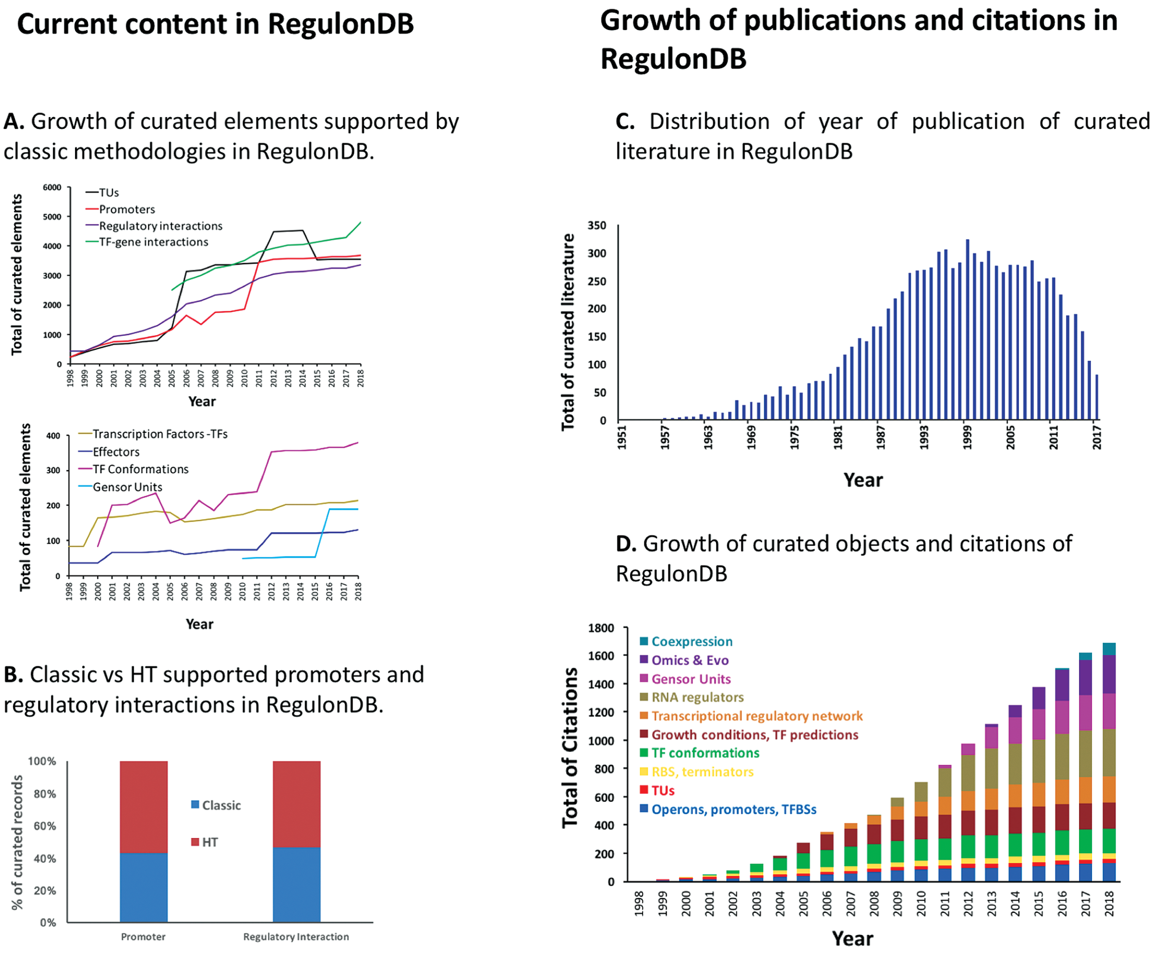


Figure 1. Growth through the years of types of elements and of citations in RegulonDB. (**A**) Curated elements over time supported by classic experimental methods. (**B**) Breakdown of classic versus HT curated records for promoters and RIs. (**C**) Publications curated by RegulonDB over time. (**D**) Curated objects reported in the *NAR* special issue publications and citations over time in RegulonDB.

main menu (go to 'Integrated Views and Tools', and select 'Tracks in JBrowse'). For each dataset, all processed data files and SOFT files were downloaded from GEO, and all available experimental information was curated using the MCO standard (see below), adding rich knowledge about the experimental conditions. The above process identified 4780 RIs of which 161 were already in RegulonDB.

When looking in detail to the HT data for Cra, we noticed that 79 genes have been reported by classic experiments to be under Cra regulation. HT experiments in three different conditions (i.e. presence of fructose, glucose and acetate) revealed 338 new regulated genes with limited overlap between the three conditions, as can be seen in Supplementary Figure S2. The majority of the genes show a change in expression unique for each condition, in spite of being subject to common regulation by Cra. Using KEGG pathway enrichment analysis, we identified TCA and glyoxylate present in fructose and acetate, but not in glucose (P-value < 0.05), whilst glycolysis is enriched in acetate and glucose, but not in fructose (P-value < 0.05). The enzymes of TCA and glyoxylate cycles are encoded by 21 genes, the Cra regulon from classical experiments includes 8 of these genes; when adding the genes obtained by HT processed data, five new genes are added. The missing genes in the pathways are genes that are regulated by CRP (*fumEDABC* and *mdh*). This illustrates the complexity of the connection between metabolism and regulation.

A framework to curate growth conditions based on an ontology (MCO)

In order to systematize the annotation of experimental growth conditions (GCs) using a controlled vocabulary, we recently developed the Microbial Conditions Ontology (MCO), which contains terms to describe the set of minimal properties necessary to reproduce an experiment (24). We have started using this framework of properties to annotate in RegulonDB the GCs that affect gene expression or the binding of a TF to the regulatory region of a gene. The framework contains the following properties: Genetic background, Medium, Medium supplements, Aeration, Temperature, pH, Pressure, Optical density (OD), Growth phase and Growth rate (24). In order to curate HT literature, it is convenient to subdivide the Genetic background item into four elements, i.e. Organism, Strain, Substrain and Genotype. We included additional items such as Vessel type or container, and Agitation conditions. We also consider the

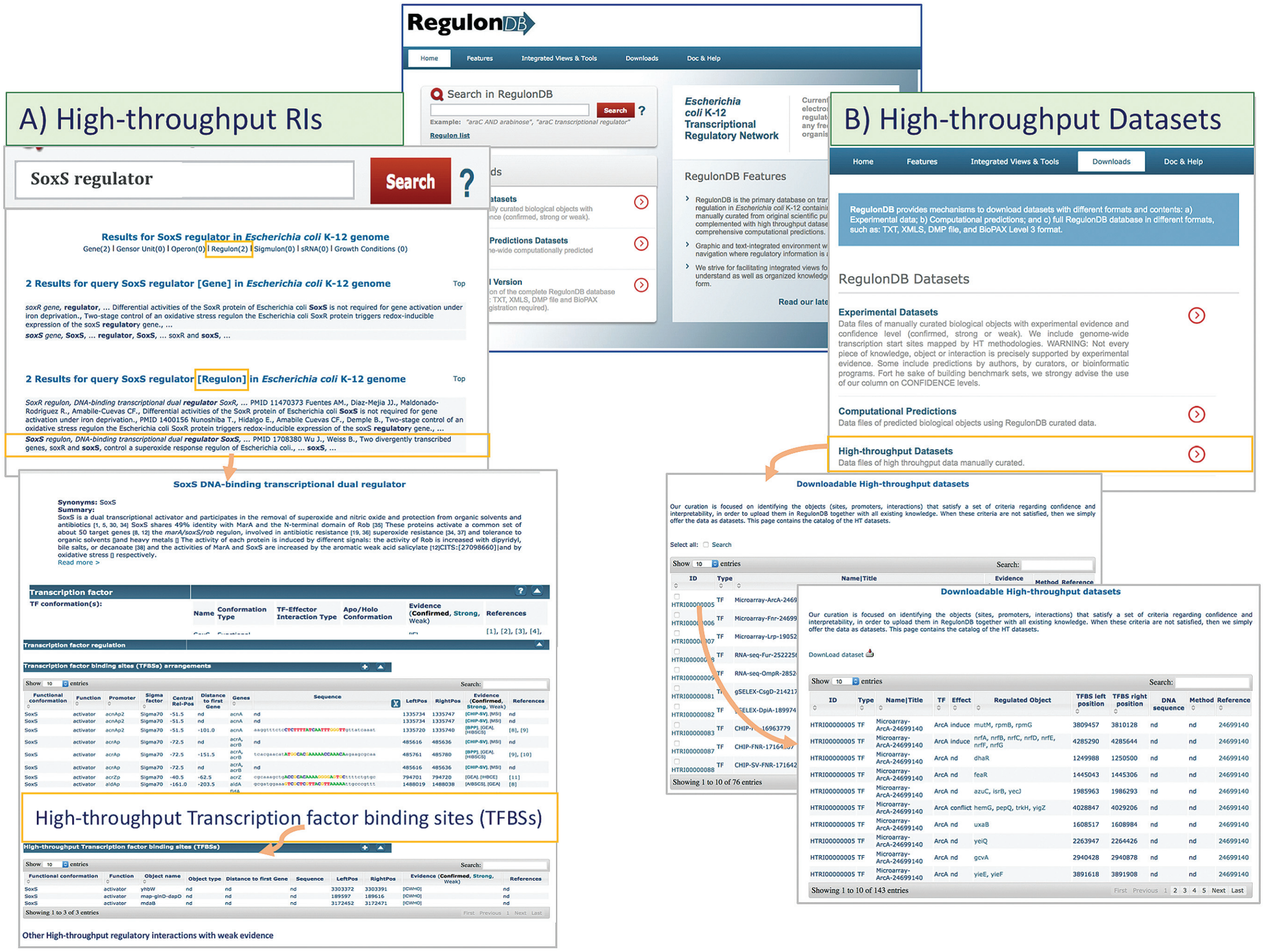

RegulonDB
Home
Features
Integrated Views & Tools
Downloads
Doc & Help
Search in RegulonDB
Search
Escherichia coli K-12 Transcriptional Regulatory Network
RegulonDB Features
A) High-throughput RIs
SoxS regulator
Search
?
Results for SoxS regulator in Escherichia coli K-12 genome
Gene(2) | Gensor Unit(0) | Operon(0) | Regulon(2) | Sigmulon(0) | sRNA(0) | Growth Conditions (0)
2 Results for query SoxS regulator [Gene] in Escherichia coli K-12 genome
2 Results for query SoxS regulator [Regulon] in Escherichia coli K-12 genome
SoxS DNA-binding transcriptional dual regulator
Transcription factor
Transcription factor regulation
Transcription factor binding sites (TFBSs) arrangements
High-throughput Transcription factor binding sites (TFBSs)
High-throughput Transcription factor binding sites (TFBSs)
Other High-throughput regulatory interactions with weak evidence
B) High-throughput Datasets
RegulonDB provides mechanisms to download datasets with different formats and contents: a) Experimental data; b) Computational predictions; and c) full RegulonDB database in different formats, such as: TXT, XMLS, DMP file, and BioPAX Level 3 format.
RegulonDB Datasets
Experimental Datasets
Computational Predictions
Data files of predicted biological objects using RegulonDB curated data.
High-throughput Datasets
Data files of high throughput data manually curated.
Downloadable High-throughput datasets
Showing 1 to 10 of 76 entries
Downloadable High-throughput datasets
Showing 1 to 10 of 143 entries

Figure 2. Display of HT curated regulatory interactions and datasets in RegulonDB. (**A**) The HT RIs are part of the Regulon web page results, in a section called 'High-throughput transcription factor binding sites'. (**B**) The HT datasets are available in the Downloads menu, and the user can filter them by any field, using the text box.

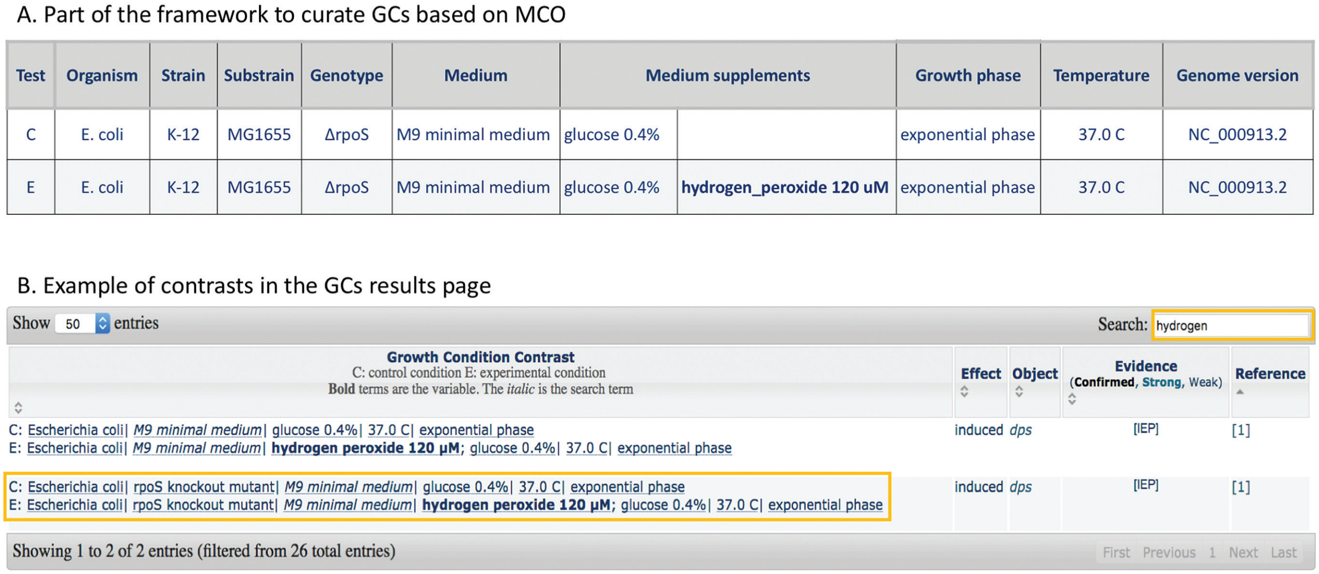
Genome version and the experimental technique in dataset descriptions. Of course, we are limited to what authors report in the literature, and thus we can fill all these items of information defined in the proposed framework in rare occasions. We call 'GC phrases' to the list of these properties for a given experiment, which we have initiated to link to genes or RIs within RegulonDB. See Figure 3 as an example.

As shown in Figure 3B, when two GC phrases are compared, a contrast is generated and the experimental variable (in bold) is identified. One of the phrases is the control experiment (C), while the other one is the test experiment (E). The variable is the experimental property that is being analyzed. Note that a GC phrase can be part of more than one contrast, as a control or as a test, and contrasts can have more than one variable. In Table 1, we show the number of instances of different variable types, where the Medium Supplement variables are grouped in high-level terms related to their biological role, to biological processes or to some stress types; all of them represented in the MCO.

We have annotated GCs from 40 classic papers and from nine papers with HT methodologies. We have also annotated the GCs derived from 49 experiments from the GEO datasets. Table 2 shows the number of variables, contrasts,

Table 1. Number of curated instances of growth condition variables types

Variable type	Variable instances
Genotype	
Knockout of genes	26
Insertion of a plasmid with gene(s)	28
Medium	
Microbiological culture medium	9
Medium supplements	
Antimicrobial agents	5
Carbon sources	23
Electron acceptors (respiration)	10
Iron depletion/repletion	7
Nitrogen sources	4
Nucleotide availability	4
Oxidative stress	10
Protein cofactors	11
Quorum sensing	4
Growth phase	
Growth phase	3



A. Part of the framework to curate GCs based on MCO

Test	Organism	Strain	Substrain	Genotype	Medium	Medium supplements		Growth phase	Temperature	Genome version
C	E. coli	K-12	MG1655	ΔrpoS	M9 minimal medium	glucose 0.4%		exponential phase	37.0 C	NC_000913.2
E	E. coli	K-12	MG1655	ΔrpoS	M9 minimal medium	glucose 0.4%	**hydrogen_peroxide 120 uM**	exponential phase	37.0 C	NC_000913.2

B. Example of contrasts in the GCs results page

Show 50 entries Search: hydrogen

Growth Condition Contrast C: control condition E: experimental condition **Bold** terms are the variable. The *italic* is the search term	**Effect**	**Object**	**Evidence** (**Confirmed**, **Strong**, Weak)	**Reference**
C: Escherichia coli| *M9 minimal medium*| glucose 0.4%| 37.0 C| exponential phase E: Escherichia coli| *M9 minimal medium*| **hydrogen peroxide 120 μM**; glucose 0.4%| 37.0 C| exponential phase	induced	*dps*	[IEP]	[1]
C: Escherichia coli| rpoS knockout mutant| *M9 minimal medium*| glucose 0.4%| 37.0 C| exponential phase E: Escherichia coli| rpoS knockout mutant| *M9 minimal medium*| **hydrogen peroxide 120 μM**; glucose 0.4%| 37.0 C| exponential phase	induced	*dps*	[IEP]	[1]

Showing 1 to 2 of 2 entries (filtered from 26 total entries)

Figure 3. Annotation framework and display in RegulonDB of the GCs contrasts. (**A**) Part of the framework to curate GCs based on the MCO. (**B**) Display of example of contrasts in the GCs page result. The variable hydrogen peroxide concentration 120 μM i.e., that induces the *dps* gene expression, is shown in bold. The evidence and reference of the induction of the gene under that GCs is shown in each contrast.

Table 2. Number of curated instances of growth condition elements

GC elements	Number of GC instances
GC phrases	378
GC contrasts	269
GC controls	164
GC experimental tests	256
GC variables	162

and phrases produced in our work. 73 genes and 4059 RIs have GCs affecting their expression and DNA binding linked to them, respectively. Some genes and RIs are affected by more than one GC.

This highly detailed description of experimental conditions will grow as our curation continues, providing the basis for biologically significant comparisons across all knowledge of gene expression within RegulonDB.

Searching growth conditions in RegulonDB

Recently we published a protocol explaining in detail several aspects of the navigation within RegulonDB (25), but it was before our incorporation of GCs. All the MCO terms are now included like search terms, so users can search any specific term. The results page provides the number of GC contrasts where the term is present. Each result is a link to the contrasts and their affected genetic elements (genes, transcription units, RIs, etc.) with their corresponding evidence and references. Also, via the 'Integrated Views and Tools', there is an option for the MCO browser to navigate through the terms in the ontology. The tool has a search box that allows filtering and selecting the desired term, to later move inside the browser to the node where the term is defined. When the user selects the desired node, the properties or attributes that define it are shown; additionally, a tab was added for those terms that have conditions phrases with contrasts. Each phrase is a link to the Growth Condition results page.

GENSOR units

GENSOR Units (see definition on RegulonDB glossary) have been updated using the latest EcoCyc and RegulonDB releases. They have been assembled through our previously reported semiautomatic pipeline (5). We have added 16 new GENSOR Units, centered on the following TFs: BCCP, BtsR, CecR, HigA, HigB-HigA, HprR, MraZ, NimR, PdeL, RclR, SrlR, SutR, TtdR, UvrY, YhaJ and YjjQ. All GENSOR Units now depict reversible reactions to better describe the information flow that each TF controls. The vocabulary has been homogenized, and a unique name is used for each metabolite across all GENSOR Units. We have reorganized their classification (See: 'Integrated Views & Tools/GENSOR Unit Groups') using the COG classification of genes (26) with 22 functional categories. Additionally, signal annotation has been extended for two-component systems, indicating the physiologically relevant stimulus that elicits the phosphorylation process. For instance, structural changes in the membrane's lipopolysaccharides signal RcsB. The Cra GENSOR has been updated given the recent knowledge of its allosteric effector (27).

We extended GENSOR Units to include regulation at the metabolic level with metabolites that affect the activity of enzymes. Interactions between metabolites and enzymes already present in the GENSOR Unit were retrieved from EcoCyc. A total of 310 interactions were added to 86 GENSOR Units. The biological significance of adding these will be reported elsewhere (Ledezma-Tejeida *et al.*, manuscript in preparation). In RegulonDB 7.0, superreactions were added when metabolites of the same metabolic pathway could be connected by a series of reactions but their enzymes were not directly regulated by the TF, and thus are absent from the GENSOR Unit. Previously superreactions were limited to metabolites that were at most three reactions apart in their metabolic pathway, and they were depicted together as one superreaction. In this new version, the limit for intermediate reactions is the size of the metabolic pathway, such that we recover all possible links between metabolites in the same pathway. We have now termed these reactions 'Complementary Pathway Reactions', because they complement GENSOR Units with genes not directly regulated by the TF. More than 145 complementary pathway reactions were identified in all GENSOR Units, 50% of which involve only 1 intermediate reaction. The longest complementary pathway reaction includes 13 intermediate reactions, suggesting that TFs tend to coregulate reactions close to each other in metabolic pathways (see Supplementary Figure S2).

Short summaries of 70 GENSOR Units have been split in a Molecular Biology description, centered on the mechanistic details of the GENSOR Unit, and a physiology summary, centered on the functional effect of the regulated genes. The molecular description begins by pointing the signal, which is frequently also the effector metabolite that binds to the TF regulating its active/inactive conformation. It explains the consequence of the TF conformation on the expression of the target genes. It includes the biological processes in which the regulated genes participate, according to the Gene Ontology (The Gene Ontology Consortium, 2016). One or more biological process was used in order to include the functions of all the genes in the GENSOR Unit. Whenever there is auto-regulation this was included. The physiological summary describes the effect of the signal either inducing or repressing the expression of the regulated genes and the biological processes they are involved in, from a functional perspective. Both summaries include a less detailed version**.**

RegulonDB and BioNLP research

In 2014, we initiated research in the use of Natural language processing (NLP) strategies to facilitate or accelerate our biocuration work, with an approach that helps us to extract the experimental contrasting variable in growth conditions from the literature (28). More recently, we focused in extracting biological processes of regulated genes and information on the structural domains of TFs, two features that are part of our TF extensive summaries. We used the manual summaries to train an automatic summarizer that collects sentences concerning TF properties from article collections (29). Since the automatic summary is a simple concatenation of extracted sentences, clear differences in wording stand out when the results are compared to those from the manual summary (See Supplementary Table S3). We invite users to visit our 'NLP research' section within the 'Integrated Views and Tools' main menu. We offer datasets of curated sentences as a resource for the BioNLP community for tasks of automatic classification, passage detection, and relation extraction.

DISCUSSION

We recently reported curation of close to 50 publications based on HT methodologies including gSELEX and various ChIP (exo, chip, seq) experiments (4). Our work focused essentially in extracting those bound sites with enough information (sequence in the genome, as well as evidence of the effect on the downstream gene) to be identified as regulatory interactions. These are available with the rest of browsable elements in RegulonDB v10.5, and all other interactions reported in those papers are available as datasets (Figure 2).

HT methodologies are very powerful in generating huge amounts of data, although this occurs at the expense of being, in a sense, 'isolated collections of information'. That is, having all binding sites for a TF is a partial story for the understanding of its physiological role. As a consequence, HT methods rely precisely on confident comparisons across different experiments, since while binding evidence may come from a ChIP-seq experiment, the evidence of effects on gene regulation may come from a different experiment like RNA-seq or a microarray. We have initiated quite a different strategy that starts not from the literature but from the dataset sample files in GEO. In this case, we can merge evidence to support RIs as shown in our initial processing of a small number of experiments. This semiautomatic processing pipeline opens the door to a new type of biocuration that would expand knowledge in RegulonDB to be sensitive to the large number of datasets from HT experiments, as well as those to be produced in the coming years. The major benefit of this approach would be a collection of experiments processed by exactly the same pipeline of bioinformatics programs, with the same versions and thresholds. This would generate an overall better foundation to make significant comparisons across this repertoire of binding and expression experiments.

In order to combine findings from these different experiments, our curation has to consider the precise conditions that were used, to make sure merged experiments are comparable. This includes not only the growth conditions, but also technical properties, such as the experimental technique and the genome version. The framework to do this is our recently published microbial condition ontology, or MCO, which has been built in a bottom up approach based on the initial detailed curation of close to 600 experiments with these required properties (24). We have started using the framework and controlled vocabulary of the MCO in our curation, and we are aware that curating the large number of RIs available in RegulonDB at the level of detail required by the MCO is an enormous task that will not be completed in the short term. Nonetheless, it is essential to fully exploit the accumulated knowledge of the *E. coli* network. For instance, discrepancies between what is curated in RegulonDB and information on ChIP-seq binding sites for a given TF may be due to specific differences in the growth conditions employed.

Understanding the role of the new TF–gene interactions that may be revealed by HT methods will be quite a challenge given the many variables that participate in the final expression of genes. TFs work frequently in combination and deciphering their composite rules is not straightforward; 58% of regulated genes have binding sites for more than one TF, and RNA-seq expression is the final consequence of as many interactions. TFs can work differently with different promoters, eliciting the expression of different transcription units. For instance, AraC can both activate and repress the *araB* promoter by itself and in the presence of arabinose. There is also evidence that TFs can work differently depending on the growth conditions, as illustrated by the Cra example here analyzed. Furthermore, we know from the construction of GENSOR Units that the mapping between metabolism and regulation is not simple, starting with the fact that the complete set of reactions in a pathway is not always jointly coregulated. Since gene regulation and the organization of the genome in transcription units is well characterized in *E. coli*, using all this knowledge as a model to compare with the rich binding and expression experiments will no doubt be a major challenge for research in the future.

Finally, we considered it valuable to offer within RegulonDB a summary of our efforts to incorporate semiautomatic processes in our curation in order to be able to expand, with a reduced number of curators, our curation more efficiently, particularly given the challenge of processing large amounts of HT-generated datasets. The organized curation of literature can offer datasets to be used for training new methods with machine learning approaches that could be applicable to curate similar literature on other microbial organisms.

We are excited by the possible research questions that can be powered by RegulonDB v10.5 and amazed by how far we have come since RegulonDB was first published in *NAR* 20 years ago (1). Since 1998 we have continuously enriched knowledge representation pertinent to gene regulation of *E. coli* K- 12, as graphically summarized in Figure 1D. Experimental and computational advances maintain the excitement in the deciphering of the *E. coli* transcriptional regulatory network.

DATA AVAILABILITY

URL: http://regulondb.ccg.unam.mx/.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

We acknowledge contributions by Gabriel Martínez-Posadas, Diego Palomares, Alfredo Hernández, Delfino García-Alonso, Shirley Alquicira-Hernández, Gerardo Salgado-Osorio for their technical help; we acknowledge Víctor Tierrafría and Citlali Mejía-Almonte for their help in adding GCs into the MCO vocabulary. We acknowledge useful observations of the manuscript made by Leonardo Collado-Torres.
We want to thank years of support by UNAM, CONACyT and National Institutes of Health funding for most of these 20 years.

FUNDING

Universidad Nacional Autónoma de México; National Institute of General Medical Sciences of the National Institutes of Health [R01GM110597, U24GM077678]. Funding for open access charge: National Institutes of Health.

Conflict of interest statement. None declared.

REFERENCES

1. Huerta,A.M., Salgado,H., Thieffry,D. and Collado-Vides,J. (1998) RegulonDB: a database on transcriptional regulation in Escherichia coli. *Nucleic Acids Res.*, **26**, 55–59.
2. Keseler,I.M., Mackie,A., Santos-Zavaleta,A., Billington,R., Bonavides-Martinez,C., Caspi,R., Fulcher,C., Gama-Castro,S., Kothari,A., Krummenacker,M. *et al.* (2017) The EcoCyc database: reflecting new knowledge about Escherichia coli K-12. *Nucleic Acids Res.*, **45**, D543–D550.
3. Neidhardt,F.C., Ingraham,J.L. and Schaechter,M. (1990) *Physiology of the bacterial cell: a molecular approach.*. Sinauer Associates, Sunderland, p. 507.
4. Santos-Zavaleta,A., Sanchez-Perez,M., Salgado,H., Velazquez-Ramirez,D.A., Gama-Castro,S., Tierrafria,V.H., Busby,S.J.W., Aquino,P., Fang,X., Palsson,B.O. *et al.* (2018) A unified resource for transcriptional regulation in Escherichia coli K-12 incorporating high-throughput-generated binding data into RegulonDB version 10.0. *BMC Biol.*, **16**, 91.
5. Ledezma-Tejeida,D., Ishida,C. and Collado-Vides,J. (2017) Genome-wide mapping of transcriptional regulation and metabolism describes information-processing units in Escherichia coli. *Frontiers in microbiology*, **8**, 1466.
6. Barrett,T., Wilhite,S.E., Ledoux,P., Evangelista,C., Kim,I.F., Tomashevsky,M., Marshall,K.A., Phillippy,K.H., Sherman,P.M., Holko,M. *et al.* (2013) NCBI GEO: archive for functional genomics data sets–update. *Nucleic Acids Res.*, **41**, D991–995.
7. Zhang,Y., Liu,T., Meyer,C.A., Eeckhoute,J., Johnson,D.S., Bernstein,B.E., Nusbaum,C., Myers,R.M., Brown,M., Li,W. *et al.* (2008) Model-based analysis of ChIP-Seq (MACS). *Genome Biol.*, **9**, R137.
8. Neph,S., Kuehn,M.S., Reynolds,A.P., Haugen,E., Thurman,R.E., Johnson,A.K., Rynes,E., Maurano,M.T., Vierstra,J., Thomas,S. *et al.* (2012) BEDOPS: high-performance genomic feature operations. *Bioinformatics (Oxford, England)*, **28**, 1919–1920.
9. Nguyen,N.T.T., Contreras-Moreira,B., Castro-Mondragon,J.A., Santana-Garcia,W., Ossio,R., Robles-Espinoza,C.D., Bahin,M., Collombet,S., Vincens,P., Thieffry,D. *et al.* (2018) RSAT 2018: regulatory sequence analysis tools 20th anniversary limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res.*, **46**, W209–W214.
10. Cui,X. and Churchill,G.A. (2003) Statistical tests for differential expression in cDNA microarray experiments. *Genome Biol.*, **4**, 210.
11. Funahashi,A., Matsuoka,Y., Jouraku,A., Morohashi,M., Kikuchi,N. and Kitano,H. (2008) CellDesigner 3.5: a versatile modeling tool for biochemical networks. *Proc. IEEE*, **96**, 1254–1265.
12. Thieffry,D., Huerta,A.M., Perez-Rueda,E. and Collado-Vides,J. (1998) From specific gene regulation to genomic networks: a global analysis of transcriptional regulation in Escherichia coli. *BioEssays*, **20**, 433–440.
13. Perez-Rueda,E. and Collado-Vides,J. (2000) The repertoire of DNA-binding transcriptional regulators in Escherichia coli K-12. *Nucleic Acids Res.*, **28**, 1838–1847.
14. Freyre-Gonzalez,J.A., Alonso-Pavon,J.A., Trevino-Quintanilla,L.G. and Collado-Vides,J. (2008) Functional architecture of Escherichia coli: new insights provided by a natural decomposition approach. *Genome biology*, **9**, R154.
15. Lozada-Chavez,I., Angarica,V.E., Collado-Vides,J. and Contreras-Moreira,B. (2008) The role of DNA-binding specificity in the evolution of bacterial regulatory networks. *J. Mol. Biol.*, **379**, 627–643.
16. Seshasayee,A.S., Fraser,G.M., Babu,M.M. and Luscombe,N.M. (2009) Principles of transcriptional regulation and evolution of the metabolic system in E. coli. *Genome Res.*, **19**, 79–91.
17. Gama-Castro,S., Salgado,H., Santos-Zavaleta,A., Ledezma-Tejeida,D., Muniz-Rascado,L., Garcia-Sotelo,J.S., Alquicira-Hernandez,K., Martinez-Flores,I., Pannier,L., Castro-Mondragon,J.A. *et al.* (2016) RegulonDB version 9.0: high-level integration of gene regulation, coexpression, motif clustering and beyond. *Nucleic Acids Res.*, **44**, D133–D143.
18. Salgado,H., Peralta-Gil,M., Gama-Castro,S., Santos-Zavaleta,A., Muniz-Rascado,L., Garcia-Sotelo,J.S., Weiss,V., Solano-Lira,H.,

Martinez-Flores,I., Medina-Rivera,A. *et al.* (2013) RegulonDB v8.0: omics data sets, evolutionary conservation, regulatory phrases, cross-validated gold standards and more. *Nucleic Acids Res.*, **41**, D203–D213.

19. Weiss,V., Medina-Rivera,A., Huerta,A.M., Santos-Zavaleta,A., Salgado,H., Morett,E. and Collado-Vides,J. (2013) Evidence classification of high-throughput protocols and confidence integration in RegulonDB. *Database*, **2013**, bas059.
20. 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.
21. 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.
22. Kim,D., Seo,S.W., Gao,Y., Nam,H., Guzman,G.I., Cho,B.K., Palsson,B.O., Mendez-Cruz,C.F., Gama-Castro,S., Mejia-Almonte,C. *et al.* (2018) Systems assessment of transcriptional regulation on central carbon metabolism by Cra and CRP First steps in automatic summarization of transcription factor properties for RegulonDB: classification of sentences about structural domains and regulated processes. *Nucleic Acids Res.*, **46**, 2901–2917.
23. Seo,S.W., Kim,D., Szubin,R. and Palsson,B.O. (2015) Genome-wide reconstruction of OxyR and SoxRS transcriptional regulatory networks under oxidative stress in Escherichia coli K-12 MG1655. *Cell Rep.*, **12**, 1289–1299.
24. Tierrafria,V.H., Mejia-Almonte,C., Camacho-Zaragoza,J.M., Salgado,H., Alquicira,K., Gama-Castro,S., Ishida,C. and Collado-Vides,J. (2018) MCO: towards an ontology and unified vocabulary for a framework-based annotation of microbial growth conditions. *Bioinformatics (Oxford, England)*, 1–9.
25. Salgado,H., Martinez-Flores,I., Bustamante,V.H., Alquicira-Hernandez,K., Garcia-Sotelo,J.S., Garcia-Alonso,D. and Collado-Vides,J. (224018) Using RegulonDB, the Escherichia coli K-12 gene regulatory transcriptional network database. *Curr. Protoc. Bioinformatics*, **61**, 1.32.31–31.32.30.
26. Galperin,M.Y., Makarova,K.S., Wolf,Y.I. and Koonin,E.V. (2015) Expanded microbial genome coverage and improved protein family annotation in the COG database. *Nucleic Acids Res.*, **43**, D261–D269.
27. Bley Folly,B., Ortega,A.D., Hubmann,G., Bonsing-Vedelaar,S., Wijma,H.J., van der Meulen,P., Milias-Argeitis,A. and Heinemann,M. (2018) Assessment of the interaction between the flux-signaling metabolite fructose-1,6-bisphosphate and the bacterial transcription factors CggR and Cra. *Mol. Microbiol.*, **109**, 278–290.
28. Gama-Castro,S., Rinaldi,F., Lopez-Fuentes,A., Balderas-Martinez,Y.I., Clematide,S., Ellendorff,T.R., Santos-Zavaleta,A., Marques-Madeira,H. and Collado-Vides,J. (2014) Assisted curation of regulatory interactions and growth conditions of OxyR in E. coli K-12. *Database*, **2014**, bau049.
29. Mendez-Cruz,C.F., Gama-Castro,S., Mejia-Almonte,C., Castillo-Villalba,M.P., Muniz-Rascado,L.J. and Collado-Vides,J. (2017) First steps in automatic summarization of transcription factor properties for RegulonDB: classification of sentences about structural domains and regulated processes. *Database*, **2017**, bax070.