MICROBIAL GENOMICS

RESEARCH ARTICLE

Tierrafría *et al.*, *Microbial Genomics* 2022;8:000833

DOI 10.1099/mgen.0.000833

MICROBIOLOGY SOCIETY

OPEN DATA OPEN ACCESS

RegulonDB 11.0: Comprehensive high-throughput datasets on transcriptional regulation in *Escherichia coli* K-12

Víctor H. Tierrafría[1,2]†, Claire Rioualen[1]†, Heladia Salgado[1], Paloma Lara[1], Socorro Gama-Castro[1], Patrick Lally[2], Laura Gómez-Romero[3], Pablo Peña-Loredo[1], Andrés G. López-Almazo[1], Gabriel Alarcón-Carranza[1], Felipe Betancourt-Figueroa[1], Shirley Alquicira-Hernández[1], J. Enrique Polanco-Morelos[1], Jair García-Sotelo[4], Estefani Gaytan-Nuñez[1], Carlos-Francisco Méndez-Cruz[1], Luis J. Muñiz[1], César Bonavides-Martínez[1], Gabriel Moreno-Hagelsieb[5], James E. Galagan[2], Joseph T. Wade[6,7] and Julio Collado-Vides[1,2,8,*]

Abstract

Genomics has set the basis for a variety of methodologies that produce high-throughput datasets identifying the different players that define gene regulation, particularly regulation of transcription initiation and operon organization. These datasets are available in public repositories, such as the Gene Expression Omnibus, or ArrayExpress. However, accessing and navigating such a wealth of data is not straightforward. No resource currently exists that offers all available high and low-throughput data on transcriptional regulation in *Escherichia coli* K-12 to easily use both as whole datasets, or as individual interactions and regulatory elements. RegulonDB (https://regulondb.ccg.unam.mx) began gathering high-throughput dataset collections in 2009, starting with transcription start sites, then adding ChIP-seq and gSELEX in 2012, with up to 99 different experimental high-throughput datasets available in 2019. In this paper we present a radical upgrade to more than 2000 high-throughput datasets, processed to facilitate their comparison, introducing up-to-date collections of transcription termination sites, transcription units, as well as transcription factor binding interactions derived from ChIP-seq, ChIP-exo, gSELEX and DAP-seq experiments, besides expression profiles derived from RNA-seq experiments. For ChIP-seq experiments we offer both the data as presented by the authors, as well as data uniformly processed in-house, enhancing their comparability, as well as the traceability of the methods and reproducibility of the results. Furthermore, we have expanded the tools available for browsing and visualization across and within datasets. We include comparisons against previously existing knowledge in RegulonDB from classic experiments, a nucleotide-resolution genome viewer, and an interface that enables users to browse datasets by querying their metadata. A particular effort was made to automatically extract detailed experimental growth conditions by implementing an assisted curation strategy applying Natural language processing and machine learning. We provide summaries with the total number of interactions found in each experiment, as well as tools to identify common results among different experiments. This is a long-awaited resource to make use of such wealth of knowledge and advance our understanding of the biology of the model bacterium *E. coli* K-12.

Received 18 December 2021; Accepted 24 April 2022; Published 18 May 2022

Author affiliations: [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]Instituto Nacional de Medicina Genómica, INMEGEN, Periférico Sur 4809, Arenal Tepepan, Tlalpan 14610, CDMX, Mexico; [4]Laboratorio Internacional de Investigación sobre el Genoma Humano, Universidad Nacional Autónoma de México, Querétaro 76230, Querétaro, Mexico; [5]Department of Biology, Wilfrid Laurier University, 75 University Ave W, Waterloo, ON N2L 3C5, Canada; [6]Wadsworth Center, New York State Department of Health, Albany, NY, USA; [7]Department of Biomedical Sciences, University at Albany, SUNY, Albany, NY, USA; [8]Centre for Genomic Regulation (CRG), The Barcelona Institute of Science and Technology, Dr. Aiguader 88, Barcelona 08003, Universitat Pompeu Fabra(UPF), Barcelona, Spain.

***Correspondence:** Julio Collado-Vides, colladojulio@gmail.com

Keywords: ChIP-seq; ChIP-exo; RNA-seq; gSELEX; DAP-seq; Transcriptional Regulatory Network; High-Throughput Nucleotide Sequencing; *Escherichia coli* K-12.

Abbreviations: CRF, conditional random field; GC, growth condition; HT, high-throughput; LT, low-throughput; MCO, microbial conditions ontology; NLP, natural language processing; PWM, position weight matrix; TF, transcription factor; TFBS, transcription factor binding site; TFRS, transciption factor regulatory site; TSS, transcription start site; TTS, transcription termination site; TU, transcription unit.

†These authors contributed equally to this work

Data statement: All supporting data, code and protocols have been provided within the article or through supplementary data files. Nine supplementary tables are available with the online version of this article.

DATA SUMMARY

All the data are available on the RegulonDB portal (https://regulondb.ccg.unam.mx/). We also provide all the code and documentation associated with these new collections:

RegulonDB software project (https://github.com/regulondbunam/). Database, web services, and web interface.

RegulonDB-HT documentation (https://github.com/PGC-CCG/RegulonDB-HT). Programs used to generate uniform collections of HT objects, mapping them to low-throughput (LT) data, and a manual describing the associated processes and formats.

RegulonDB-HT dataset docker (https://doi.org/10.5281/zenodo.6376425). From Zenodo, the users can find a link to this docker container with the dataset collections in MongoDB, the web services in GraphQL, and the web interface in React.

ChIP-seq pipeline (https://github.com/PGC-CCG/SnakeChunks). A library based on the snakemake workflow management system, which was used to design a generalizable workflow to perform reproducible ChIP-seq analyses [1].

EcoliGenes library (https://github.com/PGC-CCG/EcoliGenes). This R-based library was developed to efficiently deal with frequent and all too-often fastidious tasks related to the programmatic manipulation and comparison of genes and TFs. This library was used in multiple scripts and pipelines mentioned in this article to identify the wide variety of names and IDs used to report genes and TFs in databases and literature, the existence of multiple synonyms, spellings, and outdated bnumbers, and to convert them all into the most up-to-date symbols and bnumbers. It also includes a variety of functions that allow to efficiently get additional information on genes (coordinates, length, product, etc.) or specific genome coordinates (type of region, closest gene) directly into R data.frames, and to convert genomic coordinates from *E. coli* K-12 genome version NC_000913.2 to NC_000913.3.

The authors confirm all supporting data, code, and protocols have been provided within the article or through supplementary data files.

INTRODUCTION

Genomics has enabled a variety of technologies for the genome-wide identification of different elements defining transcription initiation, gene regulation, and transcription unit organization in any organism, provided its genome has been sequenced. In bacteria, these elements include TFs, TF binding sites (TFBS) that show specific binding of TFs, out of which we distinguish TF regulatory sites (TFRS; defined as TFBSs that are involved in transcription regulation) [2]. Moreover, genes can be transcribed either individually, or in polycistronic units, defining transcription units (TUs), which are delimited by transcription start sites (TSSs) and transcription termination sites (TTSs). As reported recently, with the development of technologies and the extension of our knowledge of transcriptional regulation, several classic definitions had to be extended. For instance, both promoters and terminators can have multiple TSSs and TTSs, respectively [2]. These updated definitions have been timely incorporated in RegulonDB [3] and in EcoCyc [4], another major resource containing information on transcriptional regulation of *E. coli* K-12.

Genome-scale technologies allow for the identification of several types of elements, such as TFBSs, gene expression profiles, and genomic elements including TUs, promoters and terminators. Approaches for TFBS identification include *in vivo* chromatin immunoprecipitation sequencing (ChIP-seq) [5, 6], its higher-resolution variant ChIP-exo [7], in addition to *in vitro* approaches, such as biotin-DNA affinity purification sequencing (DAP-seq) [8] and genomic systematic evolution of ligands by exponential enrichment (gSELEX) [9]. Note that given the binding evidence, it is not certain that proteins considered as TFs in these HT binding experiments are *bona fide* TFs, since many of them lack evidence of change in gene expression. Gene expression profiles are obtained using RNA-seq. Higher-resolution variants of RNA-seq, protecting the 5′-end of transcripts, allow for TSS identification at single-nucleotide resolution [10–12], and more recently, for the determination of full-length transcripts, along with their TSSs and TTSs [13, 14].

Publications reporting these experiments frequently describe a subset of regulatory objects, either spread along the main text [15] or compiled in tables [16–19]. Authors also provide processed datasets as supplementary material [20, 21], whereas the raw data are deposited in public repositories, such as NCBI's Gene Expression Omnibus (https://www.ncbi.nlm.nih.gov/gds), the Sequence Read Archive (https://www.ncbi.nlm.nih.gov/sra), and the EMBL-EBI's ArrayExpress (https://www.ebi.ac.uk/arrayexpress/). Extracting and processing such datasets can be challenging. Gathering these types of data in a single resource, such as RegulonDB, saves a lot of work and accelerates research facilitating data comparison with the accumulated existing knowledge based on classic molecular biology experiments, as well as comparisons with future novel knowledge.

E. coli K-12 is the prokaryote with the largest number of regulatory systems studied by classic experimental methods of molecular biology. Our laboratory at UNAM has gathered this rich, classic low-throughput (LT) knowledge for more than two decades, feeding both RegulonDB and EcoCyc [3, 4]. With the publication of large collections resulting from HT sequencing methods, we were concerned by the potential dilution of the LT classic corpus, historically considered as the gold standard, with larger collections identified by novel approaches that involve a large number of processing steps in the final identification of regulatory objects. We therefore considered offering users HT results as separate collections, the way we were offering a few genome-wide

Impact Statement

RegulonDB has been the main resource for knowledge about transcriptional regulation and organization in *E. coli* K-12, and has been accessed intensively since its first publication in 1998 [52]. For instance, in the last 4 years, RegulonDB was accessed an average of ~16300 times per year, and citations to RegulonDB articles quickly count in the hundreds. This curated database started more than 20 years ago, before the advent of high-throughput (HT) experimentation, gathering data obtained by traditional methods, with some HT data added later on. Here we present a major undertaking in ensuring high coverage of the latest HT experimental data in RegulonDB, by incorporating more than 500 HT datasets for transcription factor (TF) DNA-binding, in addition to 1864 RNA-seq datasets generated under different growth conditions and/or genetic backgrounds. Another novelty in this BioResource is the curation effort to associate each dataset with its corresponding detailed metadata that is key for its utilization. The value of having the derived genomic features, or objects, from different kinds of experiments, available in a single repository, will add to the already acknowledged value of RegulonDB to the scientific community.

datasets of TSSs generated by our collaborators back in 2009 [22]. We thus gathered datasets of TFBSs obtained by ChIP-seq and gSELEX in versions 8.0 [10] and 9.0 of RegulonDB [23]. Detailed manual curation has been devoted to extract TFRSs from those publications, for which additional evidence showing a change in expression of a nearby target gene [24] supports a regulatory interaction. Those have been uploaded into EcoCyc and RegulonDB with a clear HT evidence type along with those identified by classic LT methods. In addition to COLOMBOS with expression data [25], the Transcription Profile of *Escherichia coli* (TEC) database [26], released in 2016, offers gSELEX data in *E. coli* and the PROkaryotic Chromatin ImmunoPrecipitation database (proChIPdb, [27]), recently released, offers ChIP-seq and ChIP-exo datasets. However, to our knowledge, there is no comprehensive resource facilitating access in a single place to the diverse wealth of data of different types of objects relevant to the regulation of gene expression in *E. coli* K-12.

In this article we present a radical upgrade of RegulonDB, offering up-to-date collections of TFBSs identified from ChIP-seq, ChIP-exo, gSELEX, and biotin-modified DAP-seq approaches, as well as TSSs, TTSs, TUs and a large collection of RNA-seq expression profiles. For most of them we offer the data published by the authors, extracted either from publications or from dedicated databases. We also processed some collections from available raw data using uniform pipelines reducing their methodological differences or batch effects.

Knowing the biological conditions and genetic background supporting a binding site, an expression profile, the mapping of transcription initiation, or a transcription unit, is crucial to compare them and locate them in the wider context of additional knowledge. We used the Microbial Conditions Ontology (MCO) [28] as our theoretical framework to organize this knowledge, and, as explained below, we also implemented an assisted curation strategy applying Natural language processing (NLP) and machine learning (jointly named: *NLP method*) to automatically extract this knowledge. This assisted curation strategy consists in curating the automatically extracted growth conditions instead of curating conditions from the sources of this knowledge, saving human effort. Additionally, we added search capabilities, besides reorganizing displays in a way that should considerably improve the browsing and visualization of the different datasets and collections.

METHODS

RegulonDB-HT data model and definitions

In this work, we offer facilitated access to HT *collections*. Each collection comprises the curated *datasets* resulting in a specific type of object (Fig. 1); and a *metadata* table containing the complete list of *datasets* and their curated properties. The specific collection of TF binding objects has several *subcollections* based on the type of technology. We conceive a *dataset* as a set of data from a given experiment and its growth conditions as detailed in the MCO (culture medium, medium supplements, aeration, temperature, pH, agitation, growth phase, optical density, genetic background). *Metadata* tables also include additional information such as the genome version, features associated with the publications (author list, year of publication, PMID), as well as reported database identifiers, and any additional pertinent information. Datasets contain data files provided in the original publications (referred to as 'author files'), data files with results from our in-house processing pipelines (referred to as 'uniformized files'), or both types of files.

A new repository was designed to store the different types of datasets. The classes representing the organization of information within RegulonDB-HT, and the types of datasets processed, include TFbindingPeak, TFbindingSite, TranscriptionUnit, TranscriptionStartSite, TranscriptionTerminationSite and GeneExpression. Each of these are accompanied by their metadata and growth conditions, and at least one author data file or uniformized data file (Fig. 1). Growth conditions in the GeneExpression collection were obtained using the NLP method explained below.

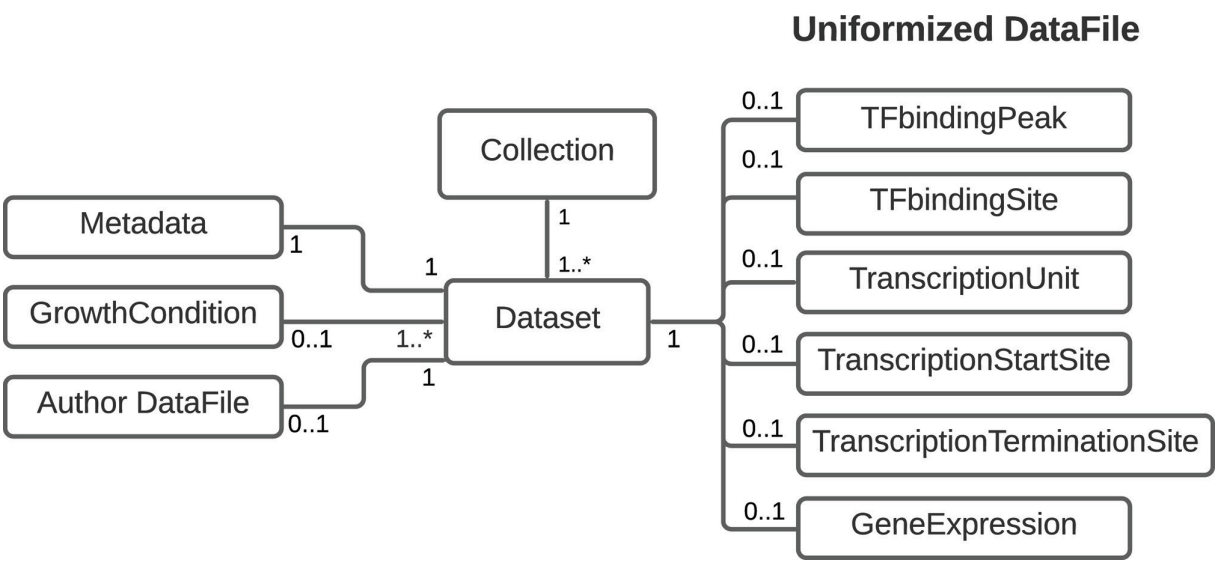

Fig. 1. Data model for HT dataset collections represented as a Unified Modelling Language (UML) class diagram. The links represent bidirectional associations between two classes, and the numbers 1, 0.*, 1.* represent the multiplicity value. For example, the class *Dataset* can have 0 or 1 *Author DataFile*. The components of datasets are the *Metadata*, defined as properties in the *Dataset* class, the *Growth Conditions*, curated manually or using the NLP method, and related data files, either gathered from authors or processed for uniformity.

The data repository was implemented in MongoDB v4.4.5 (https://www.mongodb.com/), a document-oriented database manager that provides the flexibility to deal with the variety of information of each type of dataset and collection. The package for processing the authors' and uniformized data files, and to extract, transform, and load data, was developed under python 3.9. The ChIP-seq workflows were implemented in snakemake 6.10.0 [29]. Access to data was implemented through web services that use Node v16.13.0 (https://nodejs.org/es/), the query language GraphQL v15.5.0 (https://graphql.org/), and Apollo Server Express v2.21.0. A component-based web interface was developed using React v17.0.2 (https://es.reactjs.org/). The tracks display uses igv.js, an embeddable JavaScript implementation of the Integrative Genomics Viewer (IGV) [30]. The software and applications related to the database are available at GitHub (https://github.com/regulondbunam/).

Gathering and processing of the HT data collections

To implement this new framework, we carefully coordinated the different steps involved: manual curation and annotation of literature, data uniformization, computational mapping and display of the HT collections (Fig. 2).

Data gathering

Original scientific papers about transcriptional regulation in *E. coli* K-12 are monthly searched in PubMed (https://pubmed.ncbi.nlm.nih.gov/). Then, articles are selected and curated as described previously [3]. For this work, databases associated with the publications were also explored, these include: Gene Expression Omnibus (GEO https://www.ncbi.nlm.nih.gov/gds) and the Sequence Read Archive (SRA https://www.ncbi.nlm.nih.gov/sra) from the NCBI, ArrayExpress (https://www.ebi.ac.uk/arrayexpress/) from EMBL-EBI, Digital Expression Explorer 2 (DEE2 http://dee2.io/), proChIPdb (https://prochipdb.org/), and TEC (https://shigen.nig.ac.jp/ecoli/tec/top/).

Curation and annotation

The information provided within the original publications was carefully collected and organized into custom metadata tables (one per collection, or one per subcollection in the case of TF-binding), with metadata and growth conditions for each dataset. The datasets constructed from authors sources were annotated and organized into the RegulonDB-HT repository.

Normalization and uniformization

To facilitate processing, display and analysis of these datasets, several strategies were used to uniformize and/or normalize certain datasets.

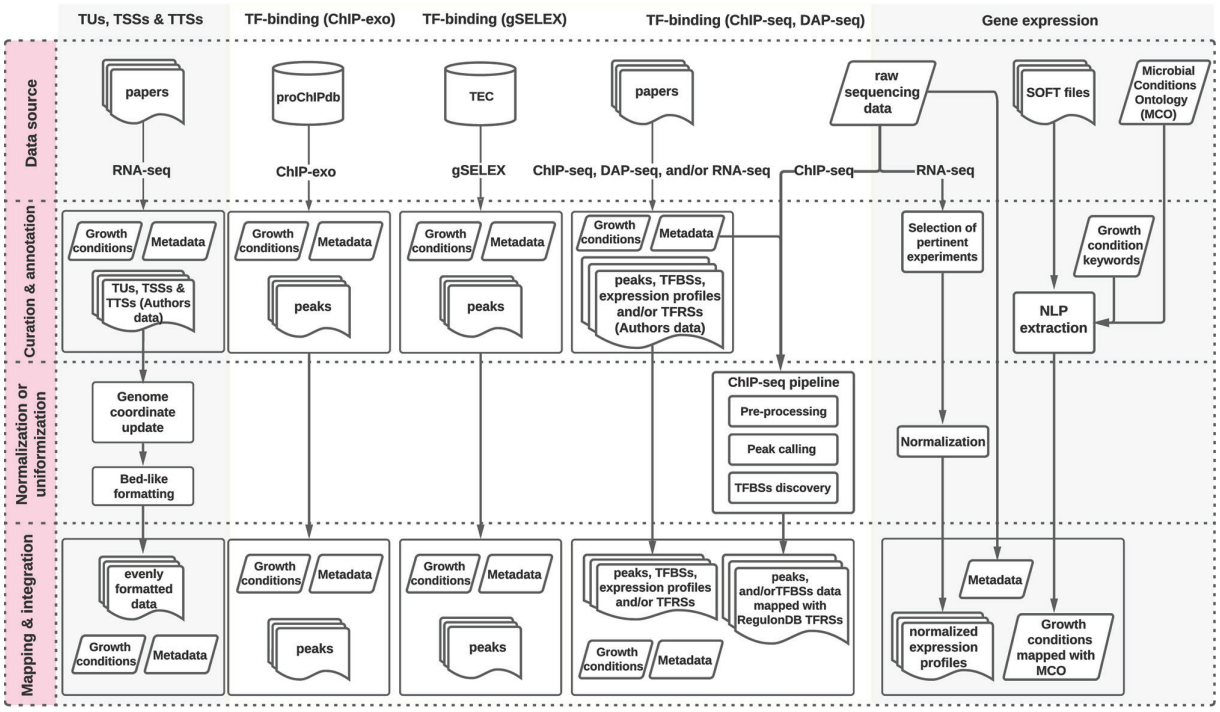

Fig. 2. Overview of the RegulonDB HT framework. This diagram summarizes the three types of dataset collections built in RegulonDB HT: i) genomic features (TUs, TSSs, and TTSs), ii) TF binding and iii) gene expression, displayed as grayscale background columns; and the steps implemented to generate them: i) data gathering, ii) curation, iii) normalization and iv) integration, displayed as horizontal lanes. Further details are described in Methods sections regarding datasets.

Mapping and integration

The resulting uniform HT objects were mapped to reference datasets from LT experiments as curated in RegulonDB. As already mentioned, growth conditions were mapped to the MCO terms and annotated, when available, according to the annotation framework reported in [28].

TF binding datasets

Data gathering

We are including binding data from four HT technologies: ChIP-seq, ChIP-exo, gSELEX and DAP-seq. The ChIP-seq datasets encompass two types of data contained in two different tables: data as reported by the authors, and data generated from our in-house processing of the raw data reported by the same authors. The TFBSs and/or peaks reported by authors were obtained mostly from supplementary material and the associated information described in the main text of their publications. ChIP-seq raw samples and metadata were downloaded systematically from the SRA. The ChIP-exo subcollection was retrieved from the recently published proChIPdb [27]. This subcollection includes datasets tagged in proChIPdb as 'curated', as well as TF binding information for OxyR, SoxR, SoxS, and UvrY, from [31, 32].

The gSELEX datasets were extracted from the TEC database [26]. Each TF was searched in the Tab 'Gene/TF search' with a selected cut-off (indicated in the metadata). The data were obtained by copy and paste since it was not possible to download it otherwise. The datasets contain the TF name, peak center coordinates, target gene, peak location relative to the target, and binding intensity (%) relative to the highest peak intensity in the experiment. We built 63 datasets for 41 TFs using a defined threshold either indicated in the corresponding references, or, in their absence, inferred by us to include all targets indicated in the publications. Ninety-four gSELEX datasets (corresponding to 74 different TFs) were not analysed by the authors, they were only listed in one publication [26]; for these we took the forty targets with the top binding intensities, and the lowest binding intensity was registered in the metadata as the cut-off for each dataset. To allow comparisons with data derived from other methodologies, we offer complete datasets from gSELEX, i.e. with no cut-off, for the nucleoid-associated proteins H-NS, Fis, IHF and HU, as well as Dps and Dan which have also been proposed as nucleoid-associated proteins [26, 33–36]. Overall, a total of 164 TFBS datasets (corresponding to 121 different TFs) derived from gSELEX were generated.

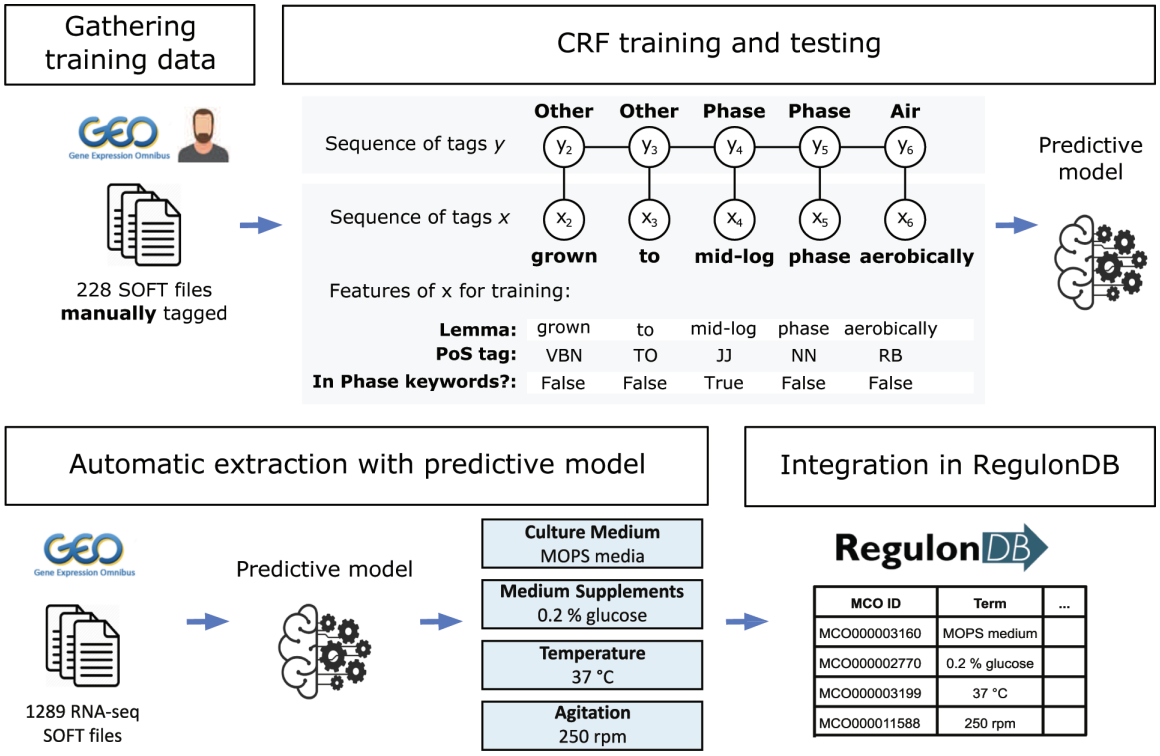

Fig. 3. Steps for growth conditions extraction using our NLP method.

Finally, we obtained the collection of experiments and metadata for 215 TFs in *E. coli* using biotin DAP-seq from the supplementary material available in [37].

Curation and annotation

To build the dataset component with data as reported by authors (Fig. 2, Curation and Annotation lane), we retrieved the following features when available: TF name, peak and TFBS features, such as start- and end genomic coordinates, genomic sequence, statistical values from peak calling or motif prediction, experimental or computational evidence, and the closest gene, considered as the target gene. The associated metadata, including growth conditions, were also extracted from the publications and databases mentioned above. Finally, when ChIP-seq experiments were linked to gene expression in the same publication, we flagged target genes which showed changes in expression and a significant *p-value* for differential expression, annotating the resulting TF function as either activator or repressor. These TFRSs support regulatory interactions which are in the process of being uploaded into EcoCyc and RegulonDB.

Uniformization

We gathered a total of 185 raw data files from 28 ChIP-seq datasets associated with 11 TFs. We processed them in a uniform and reproducible way using the SnakeChunks library of workflows for HT analysis [1, 29]. This framework ensures the consistency of analyses, keeps track of the tools and versions used, while also allowing parameter customization. Adapter and quality trimming were performed using cutadapt with a quality and length threshold of 20 [38]. Read alignment was performed using Bowtie 2 [39] in local alignment mode against the *E. coli* K-12 MG1655 genome (version NC_000913.3). Overall sample quality was checked using FastQC [40] (http://www.bioinformatics.babraham.ac.uk/projects/fastqc/) and multiQC [41]. Peak calling was performed using the latest version of Macs 3 [42], with a *q-val* threshold of 1.10–3 and the following options: *--nomodel --shift 0 --extsize 200*. Then, TFBSs were identified from the peak sequences via pattern-matching using RSAT matrix-scan [43] and the reference TF motifs built from RegulonDB 10.5 [3], and motif-specific thresholds defined by RSAT matrix-quality [44]. Two exceptions were made with GlaR and Nac, where a putative binding motif was obtained through *de novo* motif search using RSAT peak-motifs with a significance threshold of 0 [45], in order to detect binding sites. A new motif was generated for each individual dataset, using TFBS sequences and the RSAT tool convert-matrix [46]. For the other types of binding datasets, we retrieved the data as reported by the authors, in particular: start- and end positions, intensity, and the closest gene to each peak.

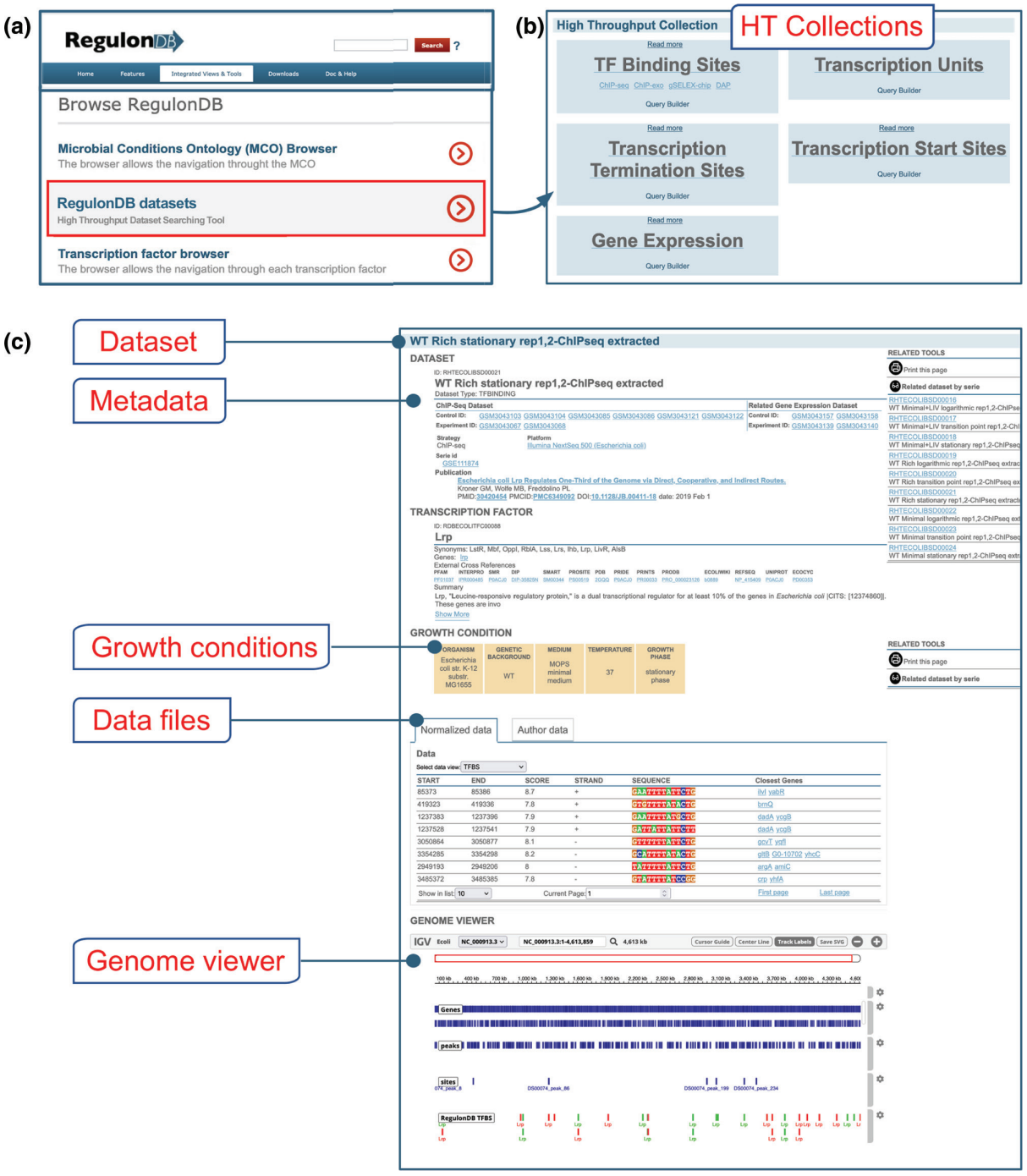

Fig. 4. RegulonDB-HT search tool. This tool gives access to all types of HT datasets retrieved so far, but an example of access to a TF binding HT dataset is shown. (a) RegulonDB portal. (b) RegulonDB HT collections. (c) Content of a TF binding dataset, from the ChIP-seq subcollection.

Mapping and integration

With the aim of comparing the TF binding data derived from HT technologies with the knowledge derived from LT studies, we performed the mapping of TFBS datasets to the RegulonDB subset of TFRSs with classical evidence. We mapped our in-house processed ChIP-seq datasets at the level of peaks and sites: a peak is considered a match when a known binding site falls within its coordinates, and a site matches when its centre position is at most 30 bp away from a known site (in average, motifs are 20 bp long, and a 10-bp distance may be close enough for protein interaction). Mapping datasets from authors proved to be more difficult since not all of them were generated using the same version of the genome, and the precise location of peaks or motifs is

Table 1. Number and content of RegulonDB HT datasets

Object	Strategy	No. of datasets	No. of objects		Additional information
			Curated from papers	Identified from raw data	
EXPRESSION PROFILES					
Gene expression	RNA-seq	1864[a]	ND	4618[b]	
TF BINDING					
TF Binding	ChIP-seq	29[c]	6585 peaks	13167 peaks 5108 sites	Table S2
	ChIP-exo	94	23170 peaks	ND	Table S3
	gSELEX	164	35022 peaks	ND	Table S4
	DAP-seq	215	19540 peaks	ND	Table S5
TUs, TSSs and TTSs					
TUs	RNA-seq	5	12347[d]	ND	Table S6
TSSs	RNA-seq	16	68049[d]	ND	Table S7
TTSs	RNA-seq	5	5326[d]	ND	Table S8

a, The total of SRRs retrieved, which include 575 only in DEE2, 914 (820 GSMs) only in GEO, and 375 (337 GSMs) in both DEE2 and GEO
b, Average number of genes per dataset.
c, Including 27 processed by authors and 28 processed in house.
d, The number of these objects may be higher from the original publications as they were calculated per *dataset*, after our uniformization process. ND. Object identification not determined by the RegulonDB Team.

not always available in publications. Thus, the datasets processed by authors were mapped at the level of the TF-gene interactions. For each TF binding dataset, target genes were compared against the known regulatory interactions from RegulonDB, taking into account the evidence they are associated with (Table S1). Positive mapping results display the type of evidence (classical strong or weak, or computational prediction) of the corresponding interaction in RegulonDB.

TU, TSS, and TTS datasets

Data gathering

Datasets of TUs, TSSs and TTSs came from different sources, though their growth conditions were not always consistently documented. TSS datasets generated by the group of Enrique Morett [10, 22], as well as those from the laboratory of Gisela Storz [47], were already available in RegulonDB [23]. Four collections are from Cho, B. K., *et al.* [48], with additional collections obtained from publications that implemented the identification of TUs using different approaches, which concomitantly identified TSSs and TTSs as TU boundaries [11, 13, 14]. A dataset not-yet-published of more than 5000 TSSs was kindly provided by Joseph T. Wade.

Curation and annotation

Given that transcriptional regulation involves a machinery that deals with different growth conditions, we gathered the precise growing conditions under which these different elements were identified, directly requesting authors for the information when it was not detailed in the publications. Key growth conditions obtained through personal communication include culture medium, either minimal or rich, and growth phase, either exponential or stationary phase.

Uniformization

When necessary, we updated object coordinates to the current genome version NC_000913.3. While the original datasets came in a variety of formats, we extracted the most relevant features for each type of collection, and generated uniform bed files for each dataset to allow their visualization in our genome browser. Objects that shared the same start, stop and strand information were considered duplicates and merged as single objects. Finally, when objects provided in a single file by the author were associated with distinct growth conditions, they were separated in distinct datasets (see *dataset* definition in the Methods section).

Mapping and integration

The uniform TSS datasets were mapped against RegulonDB promoters, and were considered a match when they fell within a 5-bp distance of a known TSS. TUs and TTSs will also be mapped in the near future. Those three uniformized collections were

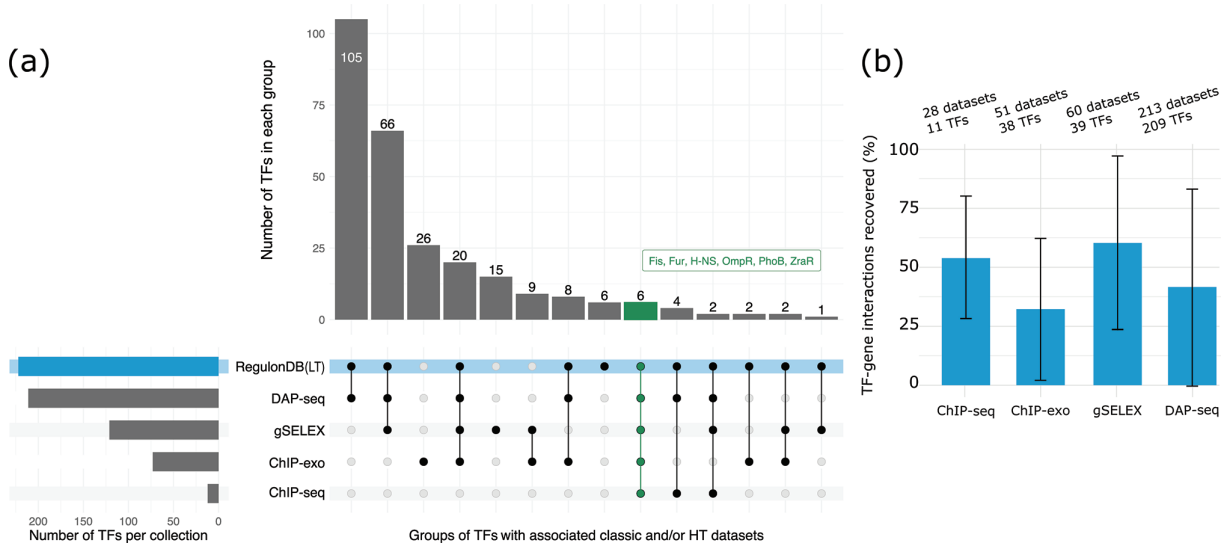

Fig. 5. TFs with binding identified by ChIP-exo, ChIP-seq, DAP-seq and/or gSELEX. (a) Comparison of TFs studied with LT approaches available in RegulonDB, with TFs examined with HT technologies. In RegulonDB, 222 TFs have been confirmed by classical LT evidence with at least one regulatory interaction (displayed as a horizontal blue bar). Each vertical bar represents a group of TFs associated with LT and/or HT experiments, as displayed by the black dots in the bottom rows. (b) Average percentage of TF-gene interactions with classical evidence in RegulonDB, identified in data processed by authors.

integrated into our genome viewer. The original author datasets were not mapped nor integrated into the genome viewer, since they come in a variety of formats and genome versions.

Gene expression datasets

Data gathering

We collected RNA-seq experiments from two different sources, GEO and DEE2. A total of 1429 experiments were retrieved from GEO using 'RNA-seq' and *E. coli*'s taxon id (txid562) as a query. We also obtained 1255 experiments from DEE2 that were not found in our initial GEO query.

Curation and annotation

We filtered these datasets based on the type of experiment and sequencing format used, retaining only RNA-seq experiments, and removing those performed with SOLiD sequencing, as our pipeline is tailored towards Illumina. We also filtered out the datasets that were associated with strains other than K-12. Of the 2684 total samples, we uploaded into RegulonDB the 1864 that could be processed by our pipeline (see Normalization subsection below). This collection is up-to-date as of the end of October 2021. The metadata were also retrieved from the corresponding database. We used the NLP method to extract growth conditions from the metadata files provided by the authors to complement the datasets obtained from GEO. For experiments only found in the SRA (retrieved from DEE2), we used NCBI's Entrez tool, along with custom software, to gather the metadata. In particular, when the metadata were missing or scarce, we used the python package Beautiful Soup four to perform web-scraping.

To gather training data for our NLP method, we selected GEO SOFT files containing metadata of studies performed with different technologies such as RNA-seq, ChIP-seq, and ChIP-exo, available in previous versions of RegulonDB. In total, the SOFT files of 228 GEO samples from 27 GEO series were gathered (Fig. 3). We automatized SOFT files download using the R package GEOquery. We manually curated and tagged the following features describing growth conditions: organism, genetic background, culture medium, medium supplements, growth phase, OD, temperature, pH, aeration, agitation, and genome version.

Manually tagged contexts from 228 SOFT files were used to train and test a linear chain Conditional Random Field (CRF): 70% for training and cross-validation, and 30% for testing. In addition, we manually obtained lists of keywords related to some types of growth conditions. A CRF is a probabilistic framework for tagging and segmenting sequence data based on the conditional probability $P(y|x)$ of a sequence of tags $y = y_1 \ldots y_n$ given a sequence of observations $x = x_1 \ldots x_n$ [49]. In this case, x is the sequence of words of contexts from the SOFT files, and y is the sequence either of tagged growth conditions ('Air', 'Phase', etc.), or the label 'Other' in other cases. The CRF probabilities are based on feature functions which may consider any feature of x_i (e.g. the part-of-speech tag, the lemma, if it contains the symbol '°', if it appears in a list of keywords) and the transition $y_{i-1} \rightarrow y_i$ (e.g. 'Phase' before 'Air'). For the final output, the consecutive words with the same label were collapsed

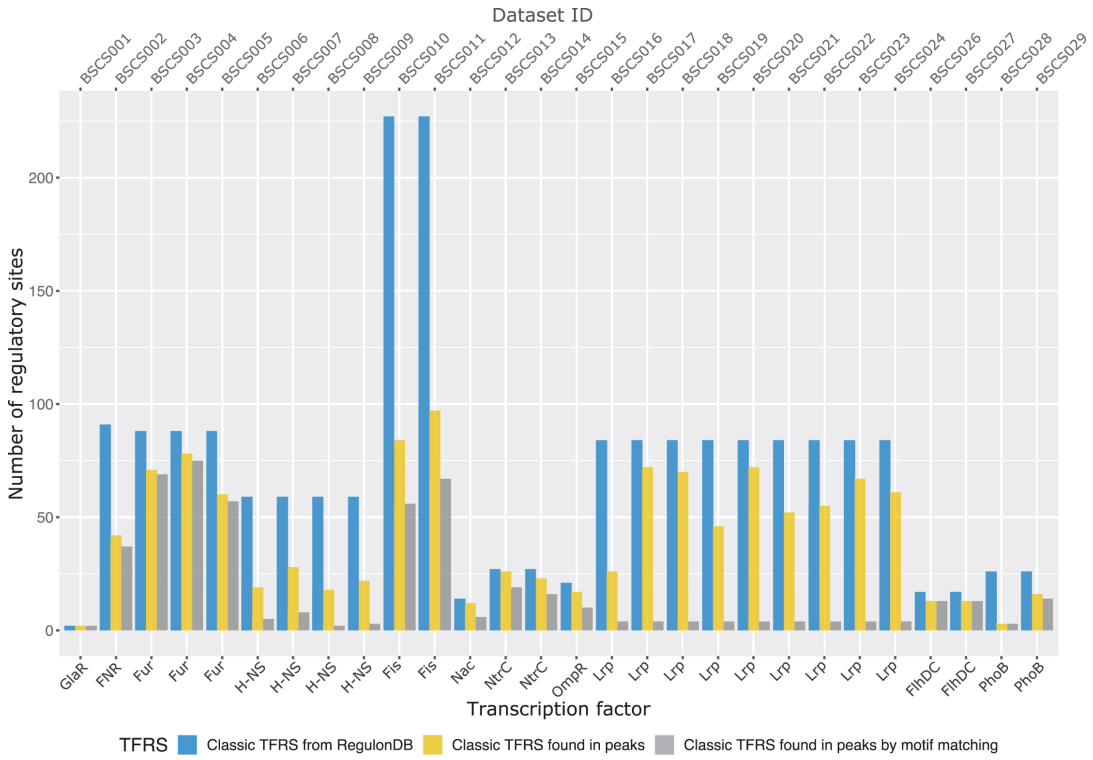

Fig. 6. Number of TFRSs from classic RegulonDB (blue bars), those found in in-house processed ChIP-seq peaks (yellow bars), and those identified in peaks through pattern-matching, using RegulonDB TF motifs (grey bars).

into a fragment of text, while the probabilities were summarized as the mean. This approach has been successfully applied previously for information extraction and it does not require a lot of training data [50].

Normalization

We downloaded the fastq files from the SRA for all datasets to be homogeneously processed by our sequence analysis pipeline. We aligned all samples to the *E. coli* reference genome NC_000913.3 using HISAT2. Our alignments are always run as unpaired; and when the metadata allow determination of the library preparation kit used, we provide the appropriate strandedness parameters, which indicate whether reads are to be expected on the same, or opposite strand of the mRNA transcript. We performed DEseq-normalization to facilitate comparisons across different datasets. Shortly, we created a 'pseudo-reference' sample, where we obtained the geometric mean of each gene's expression, measured in counts, FPKM/RPKM (depending if the experiment is paired-end or single-end, respectively), and TPM. Each gene in a given sample was divided by its pseudo-reference value, and a scaling factor for each sample was obtained by taking the median of these values. The final DEseq-normalized values were obtained by dividing each sample's expression by the sample scaling factor. In total, 1864 samples were processed without errors by our pipeline.

Mapping and integration

We took two approaches for mapping the automatically extracted growth conditions to MCO identifiers comparing the extracted term with the MCO term: (i) exact term matching and (ii) string similarity. String similarity was implemented using the python library fuzzywuzzy v0.18.0 (https://pypi.org/project/fuzzywuzzy/) taking into account string length differences calculated as Levenshtein distances, i.e. the minimum number of edits of one character (insertions, deletions or substitutions) required to change one word into the other. String similarity allowed us to match, for example, the extracted term 'W2 minimal medium' with the MCO term 'W2 minimal media' (ID: MCO000003317).

RESULTS

General overview of HT datasets and objects

As mentioned above, we report several collections of HT datasets that hold distinct types of objects (genomic features, TF binding sites, gene expression profiles) from distinct types of HT experiments (RNA-seq, ChIP-seq, gSELEX, DAP-seq, ChIP-exo). Some

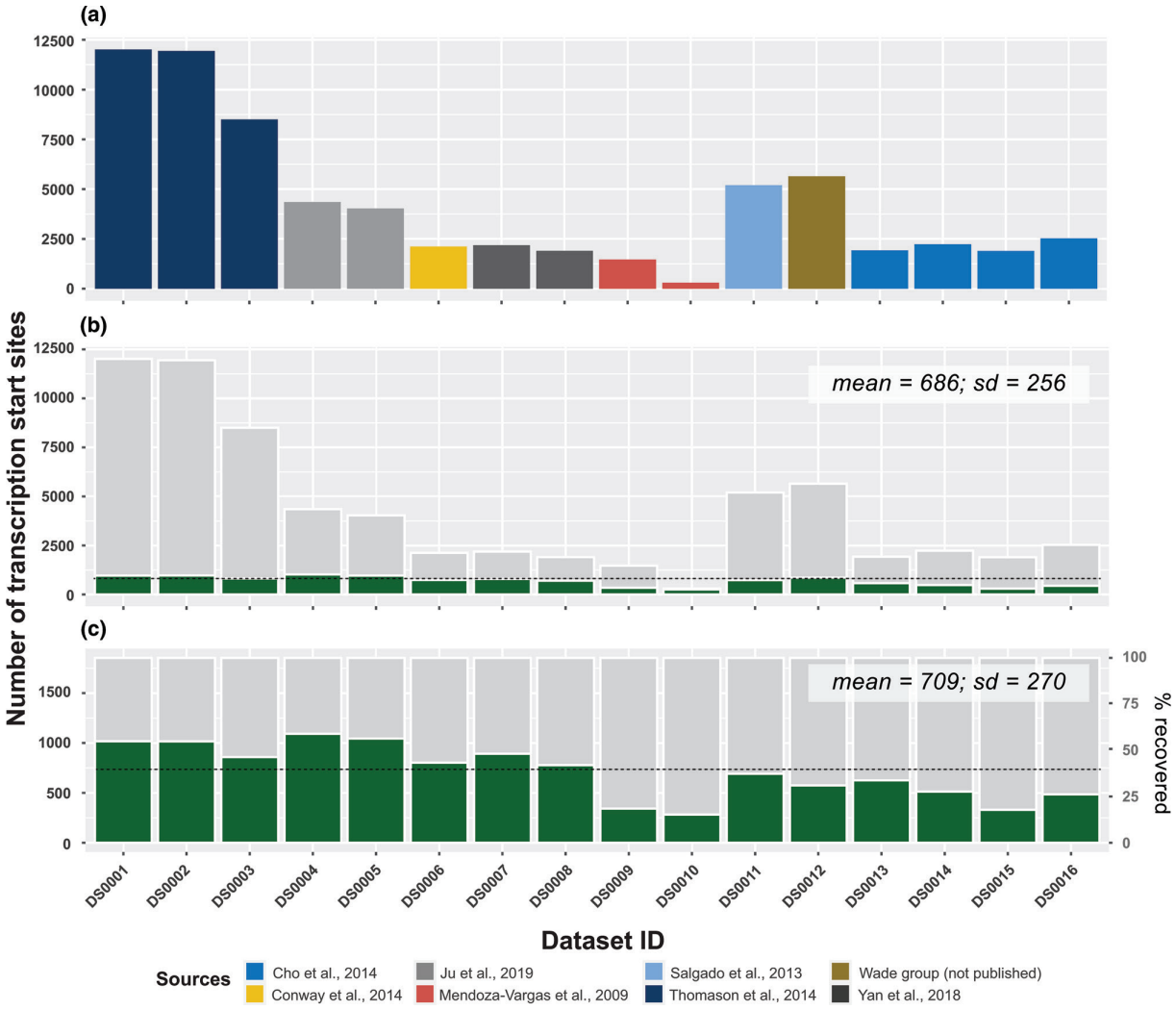

Fig. 7. High-throughput TSS datasets collected and mapped to RegulonDB classic TSSs. (a) Number of TSSs per HT dataset. (b) Number of HT TSSs that match with at least one classic TSS. (c) Number of classic TSSs that match with at least one HT TSS, for each HT dataset.

collections contain two dataset tables: data as reported by the authors, and data uniformized and/or normalized in-house. Data as reported by authors were obtained from publications curated by us, or from the authors' databases, such as TEC and proChIPdb, generated by the Ishihama and Palsson groups respectively (Fig. 2, lane 1). Data processed by other authors frequently vary in reference genome used and/or format, so we processed the author datasets to map TUs, TSSs, and TTSs with the latest reference genome and to display them in the same format (Fig. 2, lane 3). Finally, we integrated (i) data files, (ii) metadata, and (iii) growth conditions to build the RegulonDB HT datasets (Fig. 2, lane 4 and Fig. 4).

We generated three classes of RegulonDB HT datasets, roughly grouped by type of objects (described in more detail in the following sections). For example, *gene expression* datasets comprise the largest collection of datasets and objects, as expected, but are associated with only one object type and strategy, i.e. RNA-seq. In contrast, *TF binding* datasets were produced using several strategies, i.e., ChIP-seq, ChIP-exo, gSELEX, and DAP-seq. Lastly, TU, TSS, and TTS datasets include different objects identified using variations of one strategy, i.e. RNA-seq (Table 1).

Browsing the data

All the curated and annotated information, as well as the standardized data, can be found in the RegulonDB portal (https://regulondb.ccg.unam.mx/). From the menu 'Integrated Views and Tools', in the 'Browse RegulonDB' section, the option 'RegulonDB-HT datasets' is available (Fig. 4a).

Table 2. F1-score in testing for types of growth condition

Growth condition	Precision	Recall	F1-score	Support*
Optical density	1.00	1.00	1.00	21
pH	1.00	1.00	1.00	10
Technique	1.00	1.00	1.00	33
Culture medium	1.00	0.80	0.89	56
Temperature	0.86	0.80	0.83	15
Agitation	1.00	0.29	0.44	7
Growth phase	0.94	0.76	0.84	21
Aeration	0.63	0.59	0.61	88
Genetic background	0.89	0.86	0.88	78
Medium supplements	0.88	0.84	0.86	136
Genome version	1	0.5	0.667	6

**Support* stands for the number of growth conditions available in testing data for evaluation.

An initial page allows the user to select from all types of RegulonDB collections (Fig. 4b). The search builder, which is the subsequently displayed page, allows users to choose search filters associated with the RegulonDB collections' metadata. Any dataset that meets the search criteria will be displayed in a list ordered according to the number of terms found in it. The user will be able to select the desired RegulonDB dataset by clicking its link in the results list. The content of the selected dataset looks as shown in Fig. 4c), and is composed of three main components: (i) metadata, (ii) growth conditions, and iii) related data files. In the Data Files section, users can navigate through two tabs, one to access data as reported by authors, and the other one to access the standardized data produced by the RegulonDB Team.

When uniformized data are available, it is possible to visualize them in the IGV Tool, where the genes, peaks, TFBSs found in peaks, and TFRSs of the TF already stored in RegulonDB (RegulonDB TFRSs) are displayed as tracks. In the RegulonDB TFRSs track, the colour of sites is associated with the function of the TF in line with the current EcoCyc and RegulonDB TFRSs colour code, i.e. green for activators and red for repressors (Fig. 4c).

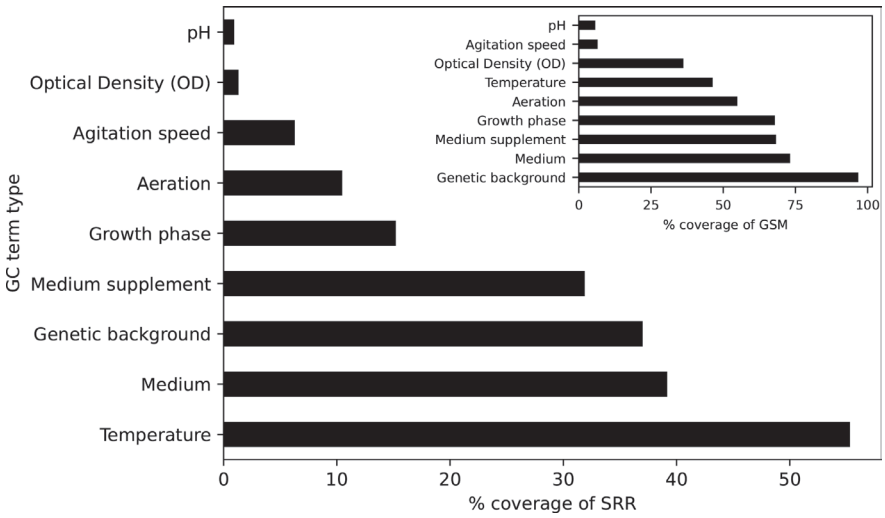

Fig. 8. Foreground bar plot: fraction of SRRs for each type of growth condition. GC term types retrieved for RNA-seq datasets from GEO (1289 SRRs, 1157 GSMs, 95 GSEs), 3224 extracted GC terms: 2680 were mapped and 544 non-mapped with MCO entities. Background bar plot: fraction of GSMs for each type of growth condition in the training data (228 GSMs from 27 GSEs).

HT data content details

In every uniformized RegulonDB HT dataset (Table 1 and Fig. 2, bottom lane), we provide the precise genomic coordinates of objects together with additionally processed information, such as the closest downstream gene(s) in the case of TFBSs and TSSs, and the gene content, in the case of TUs. Another column indicates the list of objects that match previously known objects identified by LT methods as indicated in the evidence type in RegulonDB. This pre-processed column should be highly valuable for users performing comparative analyses. In a future version we will pre-process the comparisons across the multiple HT collections.

In datasets with information provided by the authors, the confidence may vary. For instance, of the TUs identified by Yan, B. *et al.* using SMRT-Cappable-seq, some have a well-identified terminator either by sequence structure or because a significant fraction of transcripts that start at a given TSS terminate at a well defined TTS position. These TUs have a higher confidence level than the other TUs, defined by the end of one or very few long transcripts [13]. In the case of TFBSs, users can identify sites matching previously known sites stored in RegulonDB LT and/or additional evidence supporting change in expression of downstream genes.

TF binding datasets

The ChIP-seq subcollection is conformed by 29 datasets corresponding to 12 different TFs, of which 28 were processed using our dedicated pipeline (one dataset does not come with raw data), and 27 are associated with author files (two datasets are not associated with a publication). Overall, besides those exceptions, 26 datasets associated with ten TFs are provided with two tables: one with data processed by authors and built from the publications, and one with data uniformly processed in-house from raw data (Table S2).

The ChIP-exo subcollection consists of 94 datasets built with data processed by authors, which include 87 datasets corresponding to 73 different TFs assayed independently, and seven datasets derived from assays of a mixture of various TFs (Table S3).

The gSELEX subcollection consists of 164 datasets built with data processed by authors and extracted from the TEC database, corresponding to the binding of 121 different TFs assayed *in vitro* in presence or absence of effector molecules (Table S4). However, as mentioned in methods, this is a heterogeneous collection given the limitations in their extraction: 63 datasets for 41 TFs had thresholds defined by the authors, for 94 datasets of 74 TFs we arbitrarily included the top 40 sites, and for seven datasets we included all interactions with no threshold (see Methods section).

Finally, the DAP-seq subcollection comprises 215 datasets of data processed by authors and built from the supplementary material of a single publication [37], which corresponds to the binding of 211 different TFs assayed *in vitro*. Some datasets correspond to the same TF because their different subunits were assayed independently (Table S5). Some TFs have been studied by more than one of these methodologies. For example, H-NS, Fur, Fis, OmpR, ZraR and PhoB are represented in all four subcollections. Moreover, some TFs without classical evidence of regulatory interactions have been studied exclusively by one of these four HT strategies, this is the case for 26 and 15 TFs from ChIP-exo and gSELEX, respectively. Six TFs with at least one regulatory interaction with classical evidence have no data in any HT binding dataset. Fig. 5a shows the total number of TFs present in the different subcollections and their comparison with classic data from RegulonDB.

We estimated the proportion of TF-gene classic interactions present in RegulonDB that were recovered in the datasets that we constructed from author data. This percentage for every dataset is shown in Tables S2–S5 (available in the online version of this article), Fig. 5b displays the average of such percentages for all datasets within each methodology. However, these numbers have to be taken with a grain of salt, first because the TFs shared by the different methodologies are quite variable, as shown in Fig. 5a, second the recovery is quite variable for different datasets provoking a large standard deviation. Furthermore, this was done only for 63 datasets from 41 TFs of the gSELEX collection since only those have a cut-off defined by the authors. The recovery of known sites is an index frequently reported in HT publications. Note that in spite of the fact that classic evidence is mostly *in vitro* binding, there is not a clear cut tendency of HT *in vitro* methods to recover more classic interactions than the *in vivo* methods.

As mentioned already, for 28 ChIP-seq experiments we also used a uniform bioinformatics pipeline to identify TF binding sites from raw data. In such cases we provide in the same dataset two tables, one with the data as extracted from authors, and one with the results of our in-house pipeline. We generated position weight matrices (PWMs) based on the in-house obtained sites for each dataset in addition to those existing in RegulonDB, and provide the distribution of sites in relation to the start of genes or promoters. Fig. 6 shows the number of classic TFRSs in RegulonDB that are found in the peak sequences as well as those found in peaks by motif matching. The results are quite variable depending on the TF studied. In particular, the Lrp and H-NS datasets show a low rate of recovery, which can be explained by the poor specificity of their PWMs in RegulonDB.

TSS, TU and TTS datasets

We gathered a collection of 16 TSS datasets from seven articles and one unpublished dataset (see Methods section), for a total of 68049 objects (Fig. 7). The TU and TTS collections each comprise five datasets from three articles, for a total of 12347 and 5326 objects respectively (see Tables S6–S8). The original data processed by the authors as well as our uniform datasets were compared with the RegulonDB classic collection. HT TSSs were mapped to classic TSSs when located within five bases on the

same strand. It is interesting to note that even though the total number of TSSs varies from 12000 to slightly less than 300 in the different datasets (Fig. 7a), the number of HT TSSs that match with LT TSSs is much less variable (Fig. 7b), just like the numbers of classic TSSs that match with HT datasets (Fig. 7c). It should be noted that those matches, although similar in number, are not symmetrical, as a result of the window-based mapping.

Gene expression datasets

To ensure high-quality comparisons of expression data, we assessed RNA-seq samples based on sequence read alignment metrics. We tagged as 'PASS' those samples with more than five million raw reads, more than 90% of their reads aligned to the *E. coli* reference genome, and more than 90% of genes with non-zero coverage. Out of 1864 total experiments, 648 were tagged as PASS. This collection offers processed expression values at the gene level. The expression values (counts, RPKM/FPKM, and TPM) from all 1864 experiments were normalized using the DEseq method described above, allowing users to make comparisons among any desired combination of experiments, whether or not they are tagged as 'PASS'.

The growth conditions for the GEO collection were extracted by our NLP method as mentioned in the Methods section. The trained predictive model (CRF) was used to automatically extract the growth conditions from the SOFT files associated with RNA-seq data (Fig. 3). The F1-score (the harmonic mean of precision and recall) of our predictive model was 0.81 in a five-fold cross-validation, and 0.83 in testing. Precision, also known as positive predictive value, was the proportion of true positive growth conditions among all conditions classified as positive by the model. Recall, also known as sensitivity, was the proportion of known positive growth conditions classified as positive by the model. Most growth conditions attained F1-scores above 0.80 (Table 2).

Following our assisted curation strategy, the most accurately predicted NLP-extracted growth conditions terms (probability >0.7) were manually reviewed. Only the correctly predicted terms were uploaded to the searching tool for RNA-seq datasets. These correct terms of growth conditions were mapped to MCO IDs before uploading to RegulonDB.

Our NLP method was applied to 1289 SOFT RNA-seq files, associated with 95 GSEs, 1289 SRRs (SRA accession IDs) for a total of 1157 GSMs or samples. We mapped to MCO IDs ~83% of terms (15% by exact matching, and ~68% by string similarity). The unmapped terms were also included in the RNA-seq searching tool of RegulonDB.

In summary, our NLP method provided 3224 terms supporting queries for 84 GSEs, 1131 SRRs for a total of 1001 GSMs. The percentage of SRRs (coverage) with any type of growth condition was different for each type (foreground bar plot in Fig. 8). For instance, temperature, medium and genetic background are reported in more than 35% of the 1001 (100%) SRRs. In spite of our good F-scores, we know from the training set that a large fraction of data is simply missing (background bar plot in Fig. 8). A lack of data for pH, agitation speed and optical density in the training set is shown, as in the NLP-extracted data. This is a pity since it limits the comparability and usability of the data, a well-known problem in database efforts in genomics [51].

On the other hand, we gathered metadata for 575 SRRs that were not found in GEO and had no available SOFT RNA-seq files. Using NCBI's Entrez tool we were able to retrieve at least one attribute for 520 SRRs. Genetic background and medium supplements were often recovered (520 and 506 SRRs, respectively). Culture medium and growth phase were recovered for only 91 and 80 SRRs, respectively (Table S9). Thus, we have metadata that allow datasets to be searched for 928 out of 1157 RNA-seq datasets from GEO, and for 520 out of 575, SRA experiments that could not be found in GEO.

All expression data is linked to a specific SRR ID. One GEO sample (GSM ID) could include more than one SRR ID and some SRRs are not found in GEO. We processed 1289 SRRs (1157 GEO samples) by the NLP method described earlier and the remaining 575 by the NCBI's Entrez tool strategy. We were able to retrieve at least one metadata attribute for 1131 (out of 1289) and 520 SRRs (out of 575), respectively. This implies that we do not have any metadata associated with 416 SRRs. All these experiments can only be searched based on their SRR ID in RegulonDB HT.

DISCUSSION

As mentioned before, gathering all publicly available HT data from *E. coli* K-12 in a single place would be of great benefit to advance research. In this work we present RegulonDB version 11.0, a major upgrade that offers the largest variety of publicly available HT data relevant to transcriptional regulation of *E. coli* K-12. We did not however update our ChIP-chip nor microarray datasets, and we did not include any Hi-C data.

Most HT data are deposited in repositories like GEO and ArrayExpress. Although GEO requires users to complete major fields to upload genomic datasets in a uniform way, there is a lack of guidelines, or final supervision, to guarantee standardized annotations. The lack of essential information allowing the reproducibility of experiments in the literature about transcriptional regulation became evident when we curated 600 papers in high detail to build the MCO, and found none that described the growth rate, and less than 100 provided the pH, among other properties [28] This represents a major known bottleneck for proper identification and use of HT datasets in downstream analyses [51], requiring manual curation of metadata prior to choosing a final collection to work with. Our application of a method combining Natural language processing and machine learning for the automatic

extraction of growth conditions from GEO files may greatly facilitate re-analysis of these datasets. We are working on improving the predictive model for growth condition extraction.

Another recurrent issue with HT datasets is that there is no standard way of processing the raw data, and a wide variety of tools and approaches can be used, depending on the original publications. Curation has been historically limited to reflect, as precisely as possible, what authors publish and report. A major novelty in RegulonDB 11.0 is the addition of in-house processed collections. The normalized RNA-seq collection standardizes analyses across individual datasets, in principle setting the basis for future tools that would allow users to select their 'control' and 'experimental' RNA-seq datasets and obtain the relative expression of novel comparisons. The uniformized ChIP-seq subcollection was generated using our publicly-available pipeline (see Methods and Data summary sections for details). This ensures its reproducibility, which is a frequent concern when analysing published datasets from numerous sources [1, 29]. Finally, we also offer uniformized TSS, TTS and TU collections. These data were all updated to the current annotated version of the *E. coli* K-12 genome. As updates occur, traceability will be supported by the corresponding versions in GitHub, keeping all details of the tools, parameters and thresholds at hand. The diversity of information and formats provided by authors makes it difficult to compare in a comprehensive way the results of our in-house processing with those provided originally, so we leave users with the liberty of choosing which dataset to use. In a future version we will add comparisons between them.

In the current version we are offering comparisons of some HT datasets with classic LT data from RegulonDB, considered as a gold standard. This way, users can easily evaluate how each HT dataset reproduces known data from classical experiments, which is the first question to arise when applying HT strategies. In the future we will compare as many HT datasets as possible with their corresponding classic corpus, and we plan also to provide comparisons across HT datasets. This information should be highly valuable for users to compare results from different sources and technologies.

Finally, we designed a new integrated web interface, including a genome viewer and increased search capabilities. Previous RegulonDB searching capabilities were limited to TF and object type. We now allow searching for many other fields like, author, PMID, TF, growth conditions, and many more. These metadata are valuable for search and re-analysis of more than two thousand HT datasets gathered in this version.

Besides the technical aspects of the management of HT datasets we described above, we have been revisiting fundamental biological concepts. An important conceptual distinction that HT methods require for their precise description is the one between the ability to bind to specific DNA operator sites, and the capacity to alter the activity of a given promoter. Current HT publications frequently combine a binding experiment like ChIP-seq for instance, with a global expression experiment (i.e. RNA-seq) performed in the same experimental conditions. In this way it is possible to identify those sites that bind, defining TFBSs, and those that bind and modify the expression of a downstream gene, defining TFRSs. The distinction between TFBSs and TFRSs was proposed in the recent update of concepts of gene regulation [2], motivated in fact by the type of data generated with novel post-genomic technologies. By the same token, there are many potential TFs that have been assayed for instance with DAP-seq and gSELEX but have no evidence yet of any concomitant change of expression for a target gene, and therefore, as mentioned before, they do not satisfy the requirements to be fully identified yet as TFs. Lastly, we formally distinguished promoters from TSSs and terminators from TTSs, terms that are frequently used interchangeably in publications.

The version 11.0 of RegulonDB, presented here, represents an important quantitative and qualitative upgrade, offering novel features that make our repository the most comprehensive resource to utilize the wealth of HT data available, together with knowledge accumulated through decades of research with classic molecular biology approaches. We expect this unique resource will help advance research in *E. coli* K-12.

Funding information
We acknowledge funding from Universidad Nacional Autónoma de México (UNAM), as well as funding by NIGMS-NIH grant number 5RO1GM131643, and by UNAM-PAPIIT IA203420. GM-H is funded by The Natural Sciences and Engineering Council of Canada (NSERC). CR is a doctoral student from the Programa de Doctorado en Ciencias Biomédicas, UNAM, and has received fellowship 929687 from CONACyT. PL acknowledges a postdoctoral fellowship from DGAPA-UNAM. EGN thanks DGAPA-UNAM for the scholarships 181821, 369220.

Acknowledgements
We acknowledge Peter L. Freddolino, Laurence Ettwiller, Bo Yan, Xiangwu Ju and Akira Ishihama for fruitful discussion on proper interpretation of their datasets. We acknowledge the observations by anonymous referees, and also thank IT support by Víctor Del Moral.

Author contributions
V.H.T.: Data curation, conceptualization, methodology, validation, writing, review and editing. C.R.: Conceptualization, software, formal analysis, validation, visualization, writing, review and editing. H.S.: Conceptualization, validation, visualization, software, writing, review and editing. P.L.: Data curation, conceptualization, formal analysis, writing, review and editing. L.G.R.: Software, writing. P.P.L.: Software. A.G.L.A.: Software. G.A.C.: Software. F.B.F.: Software. S.A.H.: Resources. J.E.P.M.: Software. J.G.S.: Software. S.G.C.: Data curation, formal analysis. E.G.N.: Formal analysis, software, writing, review. C.F.M.C.: Formal analysis, methodology, supervision, writing, review. C.B.M.: Software, visualization. L.J.M.: Software. G.M.H.: Formal analysis, writing, review and editing. J.E.G.: Conceptualization, formal analysis, funding acquisition. J.T.W.: Conceptualization,

funding acquisition, data contributor, writing, review and editing, J.C.V.: Conceptualization, supervision, funding acquisition, writing, review and editing.

Conflicts of interest
The authors declare that there are no conflicts of interest.

References

1. Rioualen C, Charbonnier-Khamvongsa L, Collado-Vides J, van Helden J. Integrating bacterial ChIP-seq and RNA-seq data with snakechunks. *Curr Protoc Bioinformatics* 2019;66:e72.
2. Mejía-Almonte C, Busby SJW, Wade JT, van Helden J, Arkin AP, *et al*. Redefining fundamental concepts of transcription initiation in bacteria. *Nat Rev Genet* 2020;21:699–714.
3. Santos-Zavaleta A, Salgado H, Gama-Castro S, Sánchez-Pérez M, Gómez-Romero L, *et al*. RegulonDB v 10.5: tackling challenges to unify classic and high throughput knowledge of gene regulation in *E. coli* K-12. *Nucleic Acids Res* 2019;47:D212–D220.
4. Keseler IM, Gama-Castro S, Mackie A, Billington R, Bonavides-Martínez C, *et al*. The EcoCyc Database in 2021. *Front Microbiol* 2021;12:711077.
5. Johnson DS, Mortazavi A, Myers RM, Wold B. Genome-wide mapping of *in vivo* protein-DNA interactions. *Science* 2007;316:1497–1502.
6. Robertson G, Hirst M, Bainbridge M, Bilenky M, Zhao Y, *et al*. Genome-wide profiles of STAT1 DNA association using chromatin immunoprecipitation and massively parallel sequencing. *Nat Methods* 2007;4:651–657.
7. Seo SW, Kim D, Latif H, O'Brien EJ, Szubin R, *et al*. Deciphering Fur transcriptional regulatory network highlights its complex role beyond iron metabolism in *Escherichia coli*. *Nat Commun* 2014;5:4910.
8. O'Malley RC, Huang S-SC, Song L, Lewsey MG, Bartlett A, *et al*. Cistrome and epicistrome features shape the regulatory DNA landscape. *Cell* 2016;165:1280–1292.
9. Shimada T, Ogasawara H, Ishihama A. Genomic SELEX screening of regulatory targets of *Escherichia coli* transcription factors. *Methods Mol Biol* 2018;1837:49–69.
10. Salgado H, Peralta-Gil M, Gama-Castro S, Santos-Zavaleta A, Muñiz-Rascado L, *et al*. RegulonDB v8.0: omics data sets, evolutionary conservation, regulatory phrases, cross-validated gold standards and more. *Nucleic Acids Res* 2013;41:D203-13.
11. Conway T, Creecy JP, Maddox SM, Grissom JE, Conkle TL, *et al*. Unprecedented high-resolution view of bacterial operon architecture revealed by RNA sequencing. *mBio* 2014;5:e01442-14.
12. Ettwiller L, Buswell J, Yigit E, Schildkraut I. A novel enrichment strategy reveals unprecedented number of novel transcription start sites at single base resolution in a model prokaryote and the gut microbiome. *BMC Genomics* 2016;17:199.
13. Yan B, Boitano M, Clark TA, Ettwiller L. SMRT-Cappable-seq reveals complex operon variants in bacteria. *Nat Commun* 2018;9:3676.
14. Ju X, Li D, Liu S. Full-length RNA profiling reveals pervasive bidirectional transcription terminators in bacteria. *Nat Microbiol* 2019;4:1907–1918.
15. Kurata T, Katayama A, Hiramatsu M, Kiguchi Y, Takeuchi M, *et al*. Identification of the set of genes, including nonannotated morA, under the direct control of ModE in *Escherichia coli*. *J Bacteriol* 2013;195:4496–4505.
16. Shimada T, Kori A, Ishihama A. Involvement of the ribose operon repressor RbsR in regulation of purine nucleotide synthesis in *Escherichia coli* . *FEMS Microbiol Lett* 2013;344:159–165.
17. Shimada T, Katayama Y, Kawakita S, Ogasawara H, Nakano M, *et al*. A novel regulator RcdA of the csgD gene encoding the master regulator of biofilm formation in *Escherichia coli*. *Microbiologyopen* 2012;1:381–394.
18. Aquino P, Honda B, Jaini S, Lyubetskaya A, Hosur K, *et al*. Coordinated regulation of acid resistance in *Escherichia coli*. *BMC Syst Biol* 2017;11:1.
19. Fitzgerald DM, Bonocora RP, Wade JT, Søgaard-Andersen L. Comprehensive mapping of the *Escherichia coli* flagellar regulatory network. *PLoS Genet* 2014;10:e1004649.
20. Gao Y, Lim HG, Verkler H, Szubin R, Quach D, *et al*. Unraveling the functions of uncharacterized transcription factors in *Escherichia coli* using ChIP-exo . *Nucleic Acids Res* 2021;49:9696–9710.
21. Seo SW, Kim D, O'Brien EJ, Szubin R, Palsson BO. Decoding genome-wide GadEWX-transcriptional regulatory networks reveals multifaceted cellular responses to acid stress in *Escherichia coli*. *Nat Commun* 2015;6:7970.
22. Mendoza-Vargas A, Olvera L, Olvera M, Grande R, Vega-Alvarado L, *et al*. Genome-wide identification of transcription start sites, promoters and transcription factor binding sites in *E. coli*. *PLoS One* 2009;4:10.
23. Gama-Castro S, Salgado H, Santos-Zavaleta A, Ledezma-Tejeida D, Muñiz-Rascado L, *et al*. RegulonDB version 9.0: high-level integration of gene regulation, coexpression, motif clustering and beyond. *Nucleic Acids Res* 2016;44:D133-43.
24. Santos-Zavaleta A, Sánchez-Pérez M, Salgado H, Velázquez-Ramírez DA, Gama-Castro S, *et al*. A unified resource for transcriptional regulation in *Escherichia coli* K-12 incorporating high-throughput-generated binding data into RegulonDB version 10.0. *BMC Biol* 2018;16:91.
25. Moretto M, Sonego P, Dierckxsens N, Brilli M, Bianco L, *et al*. COLOMBOS v3.0: leveraging gene expression compendia for cross-species analyses. *Nucleic Acids Res* 2016;44:D620-3.
26. Ishihama A, Shimada T, Yamazaki Y. Transcription profile of *Escherichia coli*: genomic SELEX search for regulatory targets of transcription factors. *Nucleic Acids Res* 2016;44:2058–2074.
27. Decker KT, Gao Y, Rychel K, Al Bulushi T, Chauhan SM, *et al*. proChIPdb: a chromatin immunoprecipitation database for prokaryotic organisms. *Nucleic Acids Res* 2022;50:D1077–D1084.
28. Tierrafría VH, Mejía-Almonte C, Camacho-Zaragoza JM, Salgado H, Alquicira K, *et al*. MCO: towards an ontology and unified vocabulary for a framework-based annotation of microbial growth conditions. *Bioinformatics* 2019;35:856–864.
29. Mölder F, Jablonski KP, Letcher B, Hall MB, Tomkins-Tinch CH, *et al*. Sustainable data analysis with Snakemake. *F1000Res* 2021;10:33.
30. Robinson JT, Thorvaldsdóttir H, Turner D, Mesirov JP. IGV.js: an embeddable JavaScript implementation of the Integrative Genomics Viewer (IGV). *Bioinformatics* 2020.
31. Seo SW, Kim D, Szubin R, Palsson BO. Genome-wide reconstruction of OxyR and SoxRS transcriptional regulatory networks under oxidative stress in *Escherichia coli* K-12 MG1655. *Cell Rep* 2015;12:1289–1299.
32. Zere TR, Vakulskas CA, Leng Y, Pannuri A, Potts AH, *et al*. Genomic targets and features of Bara-uvry (-sira). *Signal Transduction Systems PLoS One* 2015;10:12.
33. Ueguchi C, Mizuno T. The *Escherichia coli* nucleoid protein H-NS functions directly as a transcriptional repressor. *EMBO J* 1993;12:1039–1046.
34. Antipov SS, Tutukina MN, Preobrazhenskaya EV, Kondrashov FA, Patrushev MV, *et al*. The nucleoid protein Dps binds genomic DNA of *Escherichia coli* in a non-random manner. *PLoS One* 2017;12:e0182800.
35. Prieto AI, Kahramanoglou C, Ali RM, Fraser GM, Seshasayee ASN, *et al*. Genomic analysis of DNA binding and gene regulation by homologous nucleoid-associated proteins IHF and HU in *Escherichia coli* K12. *Nucleic Acids Res* 2012;40:3524–3537.
36. Lim CJ, Lee SY, Teramoto J, Ishihama A, Yan J. The nucleoid-associated protein Dan organizes chromosomal DNA through rigid nucleoprotein filament formation in *E. coli* during anoxia. *Nucleic Acids Res* 2013;41:746–753.

37. Baumgart LA, Lee JE, Salamov A, Dilworth DJ, Na H, *et al*. Persistence and plasticity in bacterial gene regulation. *Nat Methods* 2021;18:1499–1505.
38. Martin M. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet j* 2011;17:10.
39. Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. *Nat Methods* 2012;9:357–359.
40. Wingett SW, Andrews S. FastQ Screen: A tool for multi-genome mapping and quality control. *F1000Res* 2018;7:1338.
41. Ewels P, Magnusson M, Lundin S, Käller M. MultiQC: summarize analysis results for multiple tools and samples in a single report. *Bioinformatics* 2016;32:3047–3048.
42. Feng J, Liu T, Zhang Y. Using MACS to identify peaks from ChIP-Seq data. *Curr Protoc Bioinformatics* 2011;Chapter 2:Unit.
43. Turatsinze J-V, Thomas-Chollier M, Defrance M, van Helden J. Using RSAT to scan genome sequences for transcription factor binding sites and cis-regulatory modules. *Nat Protoc* 2008;3:1578–1588.
44. Medina-Rivera A, Abreu-Goodger C, Thomas-Chollier M, Salgado H, Collado-Vides J, *et al*. Theoretical and empirical quality assessment of transcription factor-binding motifs. *Nucleic Acids Res* 2011;39:808–824.
45. Thomas-Chollier M, Darbo E, Herrmann C, Defrance M, Thieffry D, *et al*. A complete workflow for the analysis of full-size ChIP-seq (and similar) data sets using peak-motifs. *Nat Protoc* 2012;7:1551–1568.
46. Nguyen NTT, Contreras-Moreira B, Castro-Mondragon JA, Santana-Garcia W, Ossio R, *et al*. RSAT 2018: regulatory sequence analysis tools 20th anniversary. *Nucleic Acids Res* 2018;46:W209–W214.
47. Thomason MK, Bischler T, Eisenbart SK, Förstner KU, Zhang A, *et al*. Global transcriptional start site mapping using differential RNA sequencing reveals novel antisense RNAs in *Escherichia coli*. *J Bacteriol* 2015;197:18–28.
48. Cho BK, Kim D, Knight EM, Zengler K, Palsson BO. Genome-scale reconstruction of the sigma factor network in *Escherichia coli*: topology and functional states. *BMC Biol* 2014;12:4.
49. Lafferty JD, McCallum A, Pereira FCN. Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data. Proceedings of the Eighteenth International Conference on Machine Learning. 2001; Morgan Kaufmann Publishers Inc.: 282–9.
50. Peng F, McCallum A. Information extraction from research papers using conditional random fields. *Information Processing & Management* 2006;42:963–979.
51. Bernstein MN, Doan A, Dewey CN. MetaSRA: normalized human sample-specific metadata for the Sequence Read Archive. *Bioinformatics* 2017;33:2914–2923.
52. Huerta AM, Salgado H, Thieffry D, Collado-Vides J. RegulonDB: a database on transcriptional regulation in *Escherichia coli*. *Nucleic Acids Res* 1998;26:55–59.