

RESEARCH

Open Access



Flexible comparative genomics of prokaryotic transcriptional regulatory networks

Sefa Kiliç¹, Miquel Sánchez-Osuna², Antonio Collado-Padilla², Jordi Barbé² and Ivan Erill^{1*} 

From 15th International Symposium on Bioinformatics Research and Applications (ISBRA '19)
Barcelona, Spain. 3-6 June 2019

Abstract

Background: Comparative genomics methods enable the reconstruction of bacterial regulatory networks using available experimental data. In spite of their potential for accelerating research into the composition and evolution of bacterial regulons, few comparative genomics suites have been developed for the automated analysis of these regulatory systems. Available solutions typically rely on precomputed databases for operon and ortholog predictions, limiting the scope of analyses to processed complete genomes, and several key issues such as the transfer of experimental information or the integration of regulatory information in a probabilistic setting remain largely unaddressed.

Results: Here we introduce CGB, a flexible platform for comparative genomics of prokaryotic regulons. CGB has few external dependencies and enables fully customized analyses of newly available genome data. The platform automates the merging of experimental information and uses a gene-centered, Bayesian framework to generate and integrate easily interpretable results. We demonstrate its flexibility and power by analyzing the evolution of type III secretion system regulation in pathogenic Proteobacteria and by characterizing the SOS regulon of a new bacterial phylum, the Balneolaeota.

Conclusions: Our results demonstrate the applicability of the CGB pipeline in multiple settings. CGB's ability to automatically integrate experimental information from multiple sources and use complete and draft genomic data, coupled with its non-reliance on precomputed databases and its easily interpretable display of gene-centered posterior probabilities of regulation provide users with an unprecedented level of flexibility in launching comparative genomics analyses of prokaryotic transcriptional regulatory networks. The analyses of type III secretion and SOS response regulatory networks illustrate instances of convergent and divergent evolution of these regulatory systems, showcasing the power of formal ancestral state reconstruction at inferring the evolutionary history of regulatory networks.

Keywords: Comparative genomics, Regulon, Operon, Promoter, Transcription, Bacteria, SOS response, Type III secretion system

* Correspondence: erill@umbc.edu

¹University of Maryland Baltimore County, Baltimore, MD 21250, USA

Full list of author information is available at the end of the article



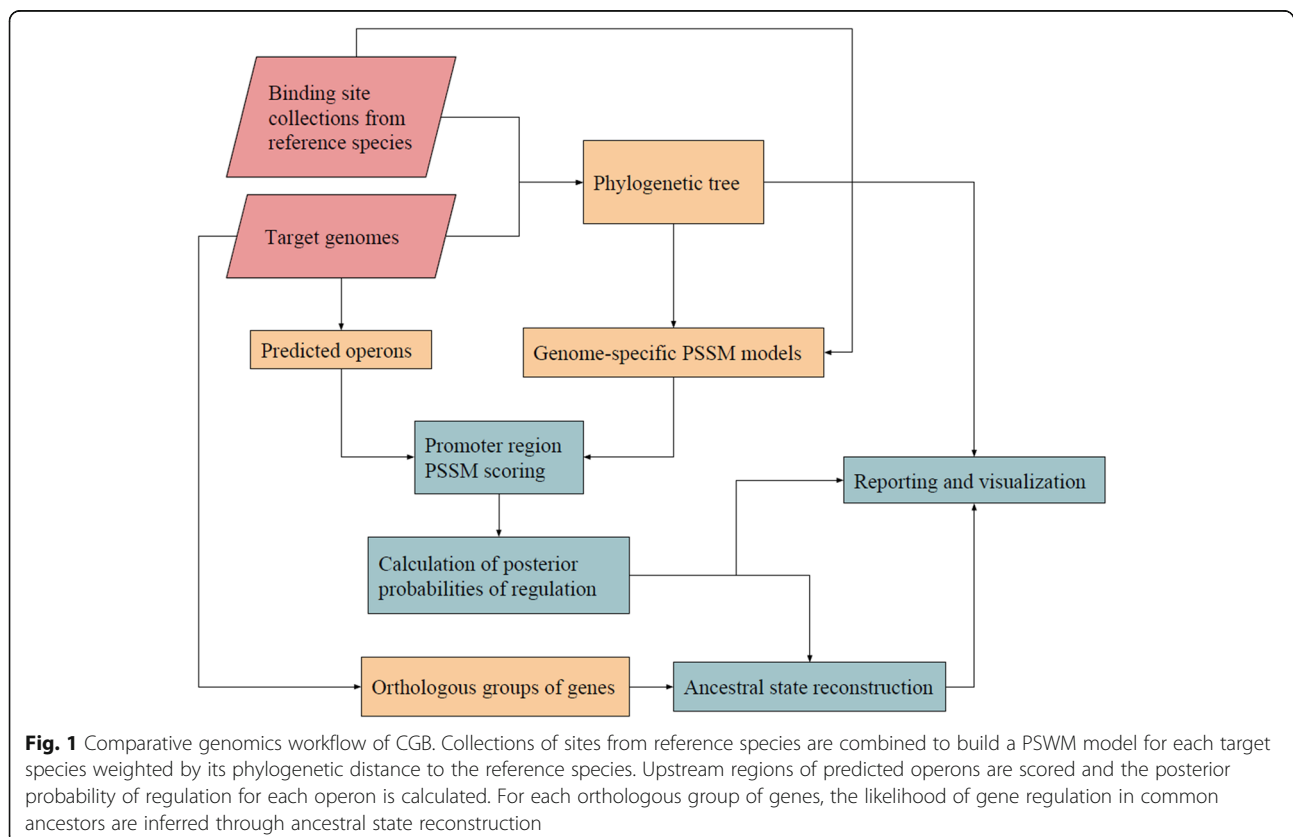
© The Author(s). 2020 **Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>. The Creative Commons Public Domain Dedication waiver (<http://creativecommons.org/publicdomain/zero/1.0/>) applies to the data made available in this article, unless otherwise stated in a credit line to the data.

Background

Transcriptional regulation is the dominant mechanism for regulation of gene expression in bacteria [1, 2]. Transcription factors (TFs) bind promoter regions in sequence-specific manner and can either hinder or promote transcription of target operons containing genes expressed from a shared promoter [2, 3]. Given prior knowledge on the binding specificity of a transcription factor, genomic sequence data can be leveraged to identify putative target sites and reconstruct the transcriptional network, or regulon, under control of a given transcription factor [4, 5]. In theory, this provides the means to elucidate the transcriptional regulatory networks they encode, yielding insights into the molecular mechanisms used by bacteria to orchestrate and coordinate diverse physiological processes. In practice, however, the short and degenerate nature of TF-binding patterns, or motifs, leads to high false positive rates in genome-wide searches, limiting their applicability [6].

Comparative genomics methods for bacterial regulon reconstruction exploit the notion that only functional TF-binding sites should be preserved across substantial evolutionary spans. Hence, the identification of a conserved TF-binding site in the promoter region of two or more orthologous operons should intuitively bolster our confidence in its prediction as a functional element

[7–9]. In spite of its applicability and potential, few integrated frameworks have been developed to automate comparative genomics analyses of bacterial regulatory networks using available motif information [10, 11]. Furthermore, several formal and practical aspects of the comparative genomics pipeline remain largely unaddressed. For instance, currently available solutions rely on precompiled databases to predict orthologous operons [12], precluding their use on the vast amount of genomic sequence data representing newly discovered bacterial clades [13–16]. Similarly, formal methods to define what constitutes a functional TF-binding site prediction, and the integration of such predictions across multiple genomes to define what constitutes a conserved binding site, have not been implemented in available tools [17, 18]. Other issues concern the automated integration of multiple sources of experimental information and the generation of interpretable probabilistic results for gene regulation in the light of operon reorganization [19]. Here we present CGB, an integrated pipeline for comparative genomics of bacterial regulatory networks with minimal external dependencies that provides a flexible environment for comparative genomics analyses while introducing a formal probabilistic framework for the integration and interpretation of analysis results. We showcase essential features of CGB through the analysis



of HrpB-mediated type III secretion system regulation and in the discovery and validation of a novel TF-binding motif in the *Balneolaeota* SOS response.

Results and discussion

A flexible platform for comparative reconstruction of bacterial regulons

CGB implements a complete computational workflow for the comparative reconstruction of bacterial regulons using available knowledge of TF-binding specificity (Fig. 1). Execution starts with the read-in of a JSON-formatted input file. This file contains the NCBI protein accession number and list of aligned binding sites for at least one transcription factor instance, accession numbers for chromids [20] or contigs mapping to one or more target species, and several configuration parameters. Reference TF-instances are used to detect orthologs in each target genome and a phylogenetic tree of TF instances is generated. The tree is used to combine available TF-binding site information into a position-specific weight matrix (PSWM) for each target species. Operons are predicted in each target species and promoter regions are scanned to identify putative TF-binding sites and estimate their posterior probability of regulation. Groups of orthologous genes are predicted across target species and their aggregate regulation probability is estimated using ancestral state reconstruction methods. CGB outputs multiple CSV files reporting identified sites, ortholog groups, derived PSWMs and posterior probabilities of regulation, as well as plots depicting hierarchical heatmap and tree-based ancestral probabilities of regulation. The following sections describe the novel strategies used to implement the different components of this computational workflow in order to generate an efficient and highly customizable comparative genomics platform.

Gene-centered, species-specific regulon reconstruction

Previous approaches to regulon reconstruction have focused on the operon as the fundamental unit of regulation [10, 11]. This poses problems for both analysis and reporting due to the frequent reorganization of operons. It is well known, for instance, that after an operon split, genes in the original operon may be regulated by the same transcription factor through independent promoters [2, 3, 21]. CGB uses instead a gene-centered framework, wherein operons become logical units of regulation, but the comparative analysis and reporting of regulons is based on the gene as the fundamental unit of regulation. This enables a rapid assessment of the regulatory state of each gene, while providing the user with detailed information on the operon setup in each organism.

Experimental information on TF-binding specificity is often available in different reference bacterial species, yet the problem of how to transfer and combine this information to target species in a comparative genomics analysis remains largely unaddressed. TF-binding motif information can be transferred across species, but the efficacy of this process decays with evolutionary distance [19]. CGB takes in prior knowledge in the form of a list of TF instances (NCBI protein accessions) in different bacterial strains, together with reported (or inferred) TF-binding sites for each of these TF instances. The collections of TF-binding sites for each TF instance must be aligned, so that the resulting motifs have the same dimensions (i.e. compatible PSWMs). This alignment can be performed manually by the user, or using dedicated tools [22]. CGB automates the transfer of TF-binding motif information from multiple sources across target species. CGB estimates a phylogeny of the reference and target TF orthologs, and uses the inferred distances between reference and target species to generate a weighted mixture PSWM in each target species, following the weighting approach used in CLUSTALW [23]. This provides a principled and reproducible approach for the dissemination of TF-binding motif information, forgoing the need to manually adjust inferred collections of TF-binding sites in each target organism [10].

Promoter scoring and probabilistic framework

The frequency information in a PSWM can be transformed into a position-specific scoring matrix (PSSM) and used to identify TF-binding sites in genomic sequences. The use of a PSSM score cut-off for predicting putative TF-binding sites in promoter regions has long been the de facto standard in regulon analysis [8–10, 24, 25]. However, this approach is not well-suited for the comparative genomics framework, because thresholds may often need to be tuned in different bacterial genomes owing to their particular distribution of oligomers [6]. To circumvent this problem, here we adopt a Bayesian probabilistic framework originally developed for regulon analysis in metagenomic sequences [26]. This framework estimates posterior probabilities of regulation that are easily interpretable and directly comparable across species.

For each position i of a promoter region, we first combine the PSSM scores obtained in the forward (f) and reverse (r) strands using the function [26]:

$$PSSM(s_i) = \log_2 \left(2^{PSSM(s_i^f)} + 2^{PSSM(s_i^r)} \right) \quad (1)$$

To estimate the posterior probability of regulation of a promoter, we define two distributions of PSSM scores within a promoter region. In a promoter not regulated

by the TF, we expect a background distribution of scores (B). We approximate this distribution using a normal distribution parametrized by the statistics of PSSM scores in promoters genome-wide:

$$B \sim N(\mu_G, \sigma_G^2) \quad (2)$$

In a promoter regulated by the TF, however, we expect that the distribution of PSSM scores (R) be a mixture of both the background distribution (B) and the distribution of scores in functional sites. The latter can be approximated with a normal distribution parametrized by the statistics of the TF-binding motif (M):

$$R \sim \alpha N(\mu_M, \sigma_M^2) + (1-\alpha)N(\mu_G, \sigma_G^2) \quad (3)$$

The mixing parameter α is a prior that corresponds to the probability of a functional site being present in an average-length regulated promoter and can be easily estimated from experimental data. Bacterial transcription factors regulate most of their target genes by binding to a given number of sites in the promoter region, and the average length of the promoter region in a given organism can be readily approximated as the average intergenic distance between the first genes in opposing directions. For a transcription factor known to typically bind one site per regulated promoter, and an estimated average promoter length of 250 bp, we obtain $\alpha = 1/250 = 0.004$. This results in a mixture distribution for the regulated promoter (R) drawing predominantly (99.6% of the time) from the background distribution of scores (B).

For any given promoter, we can define the posterior probability of regulation $P(R|D)$ given the observed scores (D):

$$\begin{aligned} P(R|D) &= \frac{P(D|R)P(R)}{P(D)} \\ &= \frac{P(D|R)P(R)}{P(D|R)P(R) + P(D|B)P(B)} \end{aligned} \quad (4)$$

Assuming independence among the scores at each promoter position, the likelihood functions can be estimated for a given score s_i using the density function of the R and B distributions defined above:

$$P(D|R) = \prod_{s_i \in D} L(s_i | \alpha N(\mu_M, \sigma_M^2) + (1-\alpha)N(\mu_G, \sigma_G^2)) \quad (5)$$

$$P(D|B) = \prod_{s_i \in D} L(s_i | N(\mu_G, \sigma_G^2)) \quad (6)$$

The priors $P(R)$ and $P(B)$ can be inferred from the reference collections. $P(R)$ can be approximated as the

number of known regulated promoters in a reference genome, divided by the total number of operons, and $P(B)$ is trivially $1-P(R)$. Alternatively, the number of regulated promoters can be estimated from the information content of the species-specific TF-binding motif [6, 27].

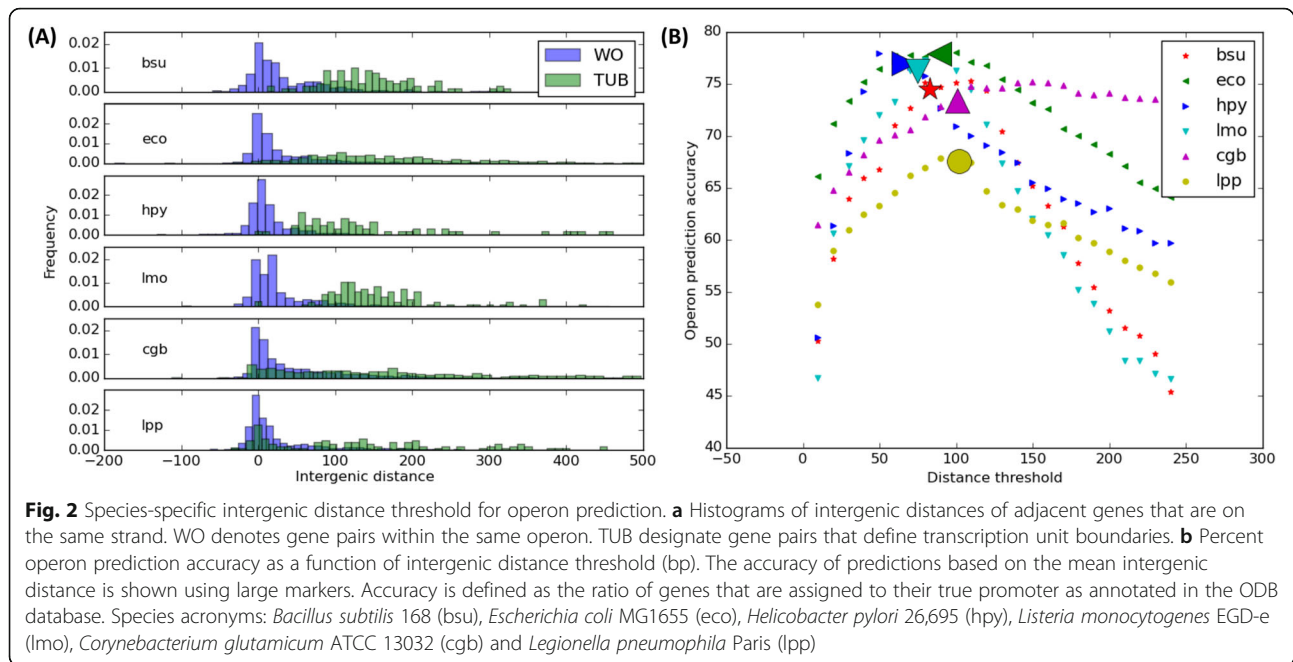
Operon prediction

Operon prediction remains a challenging problem in bacterial genomics [28]. Available comparative genomics platforms rely on curated operon databases to improve accuracy, but this limits their applicability to a preselected set of complete bacterial genomes [10, 11]. To enable analyses including newly sequenced, complete or incomplete bacterial genomes, CGB implements a two-tiered operon prediction sequence. Intergenic distance is an effective and widely-used predictor of operons. Genes pairs in a same direction (adjacent in the same orientation with no intervening genes in the opposite strand) are considered to belong to an operon if their intergenic distance is below a pre-established threshold [29–31]. Because different genomes can have different coding densities, CGB defines this threshold in an adaptive manner as the average intergenic distance in all directions within a given genome. We benchmarked this approach using experimental operon data from the ODB database for six bacterial species (Fig. 2), revealing that dynamically adapting the threshold to each bacterial genome significantly enhances the prediction accuracy.

When reconstructing regulons through comparative genomics, errors in operon prediction can yield two different scenarios: an operon may be split and regulation for some of its genes hence not properly detected, or spurious genes may be incorporated into an operon and their regulation inaccurately predicted. To circumvent this problem CGB predicts operons using a conservative threshold to minimize undesired operon splits (Fig. 2). It then scans the upstream region of all genes to detect any genes within a predicted operon harboring a high-scoring TF-binding site. This is defined as a site scoring above a PSSM score threshold that satisfies the equality between the negative logarithm of the false positive rate (FPR) and the information content of the TF-binding motif [32]. In such cases, the operon is split on the gene with the identified high-scoring site. This allows recovering relevant regulation information for genes that may have inaccurately included within an operon in the initial prediction.

Ortholog detection and ancestral state reconstruction

Ortholog detection remains a challenging, computationally-intensive problem in bioinformatics [33]. Available comparative genomics platforms make use of precompiled ortholog sets [10, 11], but this restricts the range of bacterial species that can be analyzed. To provide the user with



greater flexibility, CGB implements automated detection of orthologous groups in the species under analysis. Orthologous genes are detected as reciprocal best BLAST hits for each pair of species using tBLASTX with default parameters and a 10^{-10} cutoff e-value. Pairwise reciprocal BLAST hits are used to generate a graph where vertices correspond to gene products and edges denote best reciprocal BLAST hit relationships, and orthologous gene groups are detected as cliques in the graph [34]. Crucially, regulon reconstruction through comparative genomics does not require that all orthologous groups across target species be identified. CGB limits ortholog detection to those genes present in operons with a posterior probability of regulation higher than a user-specified cut-off in any of the target species. This dramatically reduces the complexity of the ortholog detection step, enabling it to be performed in real time.

Integrating the regulon information inferred from each target genome is a critical step in comparative genomics in order to generate insights on the overall makeup of the regulon and its evolutionary history. A common rule of thumb in many comparative genomic analyses has been to assume that the detection of putative TF-binding sites in the promoter region of orthologous operons from two or more sufficiently divergent genomes represents strong evidence of regulation [7–9]. More recently, comparative regulon reconstruction has been formally recast as an ancestral state reconstruction problem, wherein one seeks to infer the likelihood of regulation for a given operon on a phylogenetic tree [18]. CGB implements this approach through bootstrapping ancestral state reconstruction for any given gene on a phylogenetic tree of TF instances. For each bootstrap

replicate, CGB assigns discrete regulated (s_r) or non-regulated (s_o) states to each target species by sampling randomly according to the inferred posterior probability of regulation. If the species does not encode an ortholog for the given gene, the absent (s_a) state is assigned. CGB then uses BayesTraits to infer the discrete regulation states on ancestral nodes for each bootstrap replicate [35]. These inferred ancestral states are averaged over all bootstrap replicates to obtain ancestral posterior probabilities of regulation.

Comparative analysis of the LexA regulatory network in gram-positive bacteria

The SOS response is a transcriptional regulatory network that responds to DNA damage and activates the expression of genes to address DNA lesions and their effects. The SOS response was first described in *Escherichia coli*, where it was shown to regulate over 40 genes involved in three primary functions: DNA repair, inhibition of cell division and translesion synthesis [36, 37]. DNA damage is sensed by the recombination protein RecA, which can promote the autocatalytic cleavage and inactivation of the transcription factor LexA, leading to de-repression of its target genes [36, 38]. Later research has shown that the SOS response is widespread in bacteria but, in contrast with other regulatory networks, multiple LexA-binding motifs have been reported in different bacterial phyla [39]. Both the binding motif and the regulatory network for LexA have been amply documented the Actinobacteria and the Firmicutes [40–44], providing an ideal test case for assessing the performance of CGB. We performed a comparative analysis of

the LexA regulon across seven bacterial species: five for which the SOS response has been reported (*Corynebacterium glutamicum* ATCC 13032, *Bacillus subtilis* 168, *Staphylococcus aureus* NCTC 8325, *Listeria monocytogenes* EGD-e and *Mycobacterium tuberculosis* H37Rv) and two related species where the SOS response remains uncharacterized (*Leifsonia xyli* CTCB07, *Acidothermus cellulolyticus* 11B).

In agreement with previous reports [9], our analysis reveals that the core SOS regulon in Gram-positive bacteria encompasses the LexA and RecA proteins, as well as error-prone polymerases, a radical SAM protein and a

cell-division inhibitor (Fig. 3). The plot also illustrates how CGB distributes the available experimental information on TF-binding motifs across all target species, generating phylogenetically weighted mixture motifs that smooth out motifs with low experimental support. We assessed the accuracy of CGB at determining the regulatory state of target genes using the predicted posterior probability of regulation in different reference species. Our results show that, among genes predicted to be regulated in at least one species, the posterior probability generates sharp distinction between regulated and non-regulated genes, yielding accuracy across a broad range

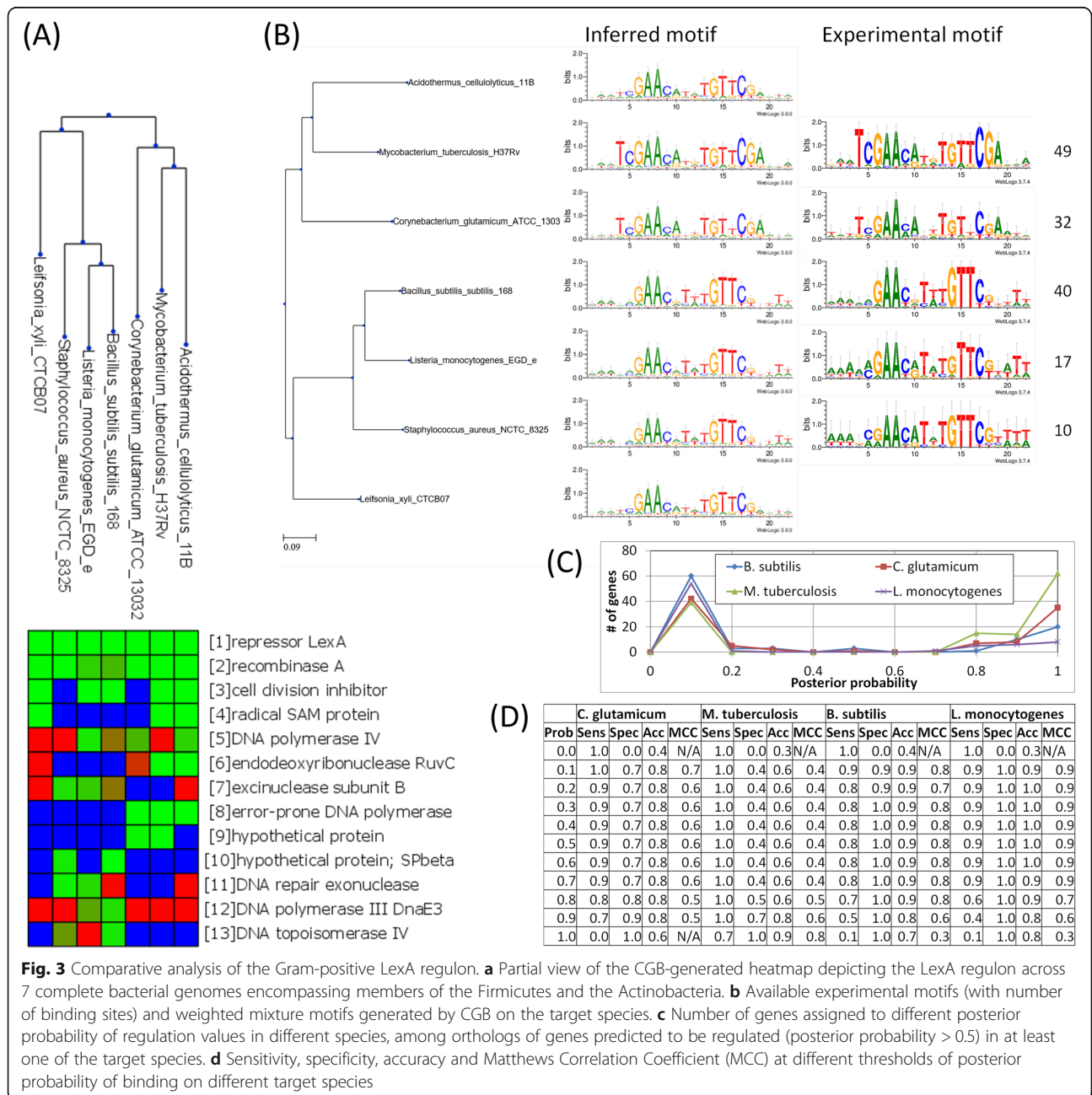


Fig. 3 Comparative analysis of the Gram-positive LexA regulon. **a** Partial view of the CGB-generated heatmap depicting the LexA regulon across 7 complete bacterial genomes encompassing members of the Firmicutes and the Actinobacteria. **b** Available experimental motifs (with number of binding sites) and weighted mixture motifs generated by CGB on the target species. **c** Number of genes assigned to different posterior probability of regulation values in different species, among orthologs of genes predicted to be regulated (posterior probability > 0.5) in at least one of the target species. **d** Sensitivity, specificity, accuracy and Matthews Correlation Coefficient (MCC) at different thresholds of posterior probability of binding on different target species

of thresholds. Analysis of the prediction accuracy for individual genes reveals that there is a consistent positive correlation ($0.33 \pm SD 0.08$; Spearman correlation coefficient) between true positives in individual species and the number of species in which the gene was predicted to be regulated, supporting the fundamental assumption that conservation of regulatory elements is indicative of functionality.

Analysis of type III secretion system regulation by HrpB/HrpX

Regulation of the bacterial type III secretion system (T3SS) has been described as mediated by the orthologous

transcription factors HrpX (*Xanthomonas oryzae*) and HrpB (*Ralstonia solanacearum*) [45, 46]. In both systems, this TF has been shown to regulate the expression of genes governing the assembly of the T3SS apparatus and several of the effectors that these pathogenic bacteria translocate into host cells. Here we used experimental TF-binding motif from *Xanthomonas* and *Ralstonia* species available in the CollecTF database (Fig. 4a) to analyze the evolution of this regulatory network in several groups of pathogenic bacteria harboring HrpB/X orthologs. This is accomplished by first propagating the experimental motif among all target species (Fig. 4b), and then performing the comparative analysis.

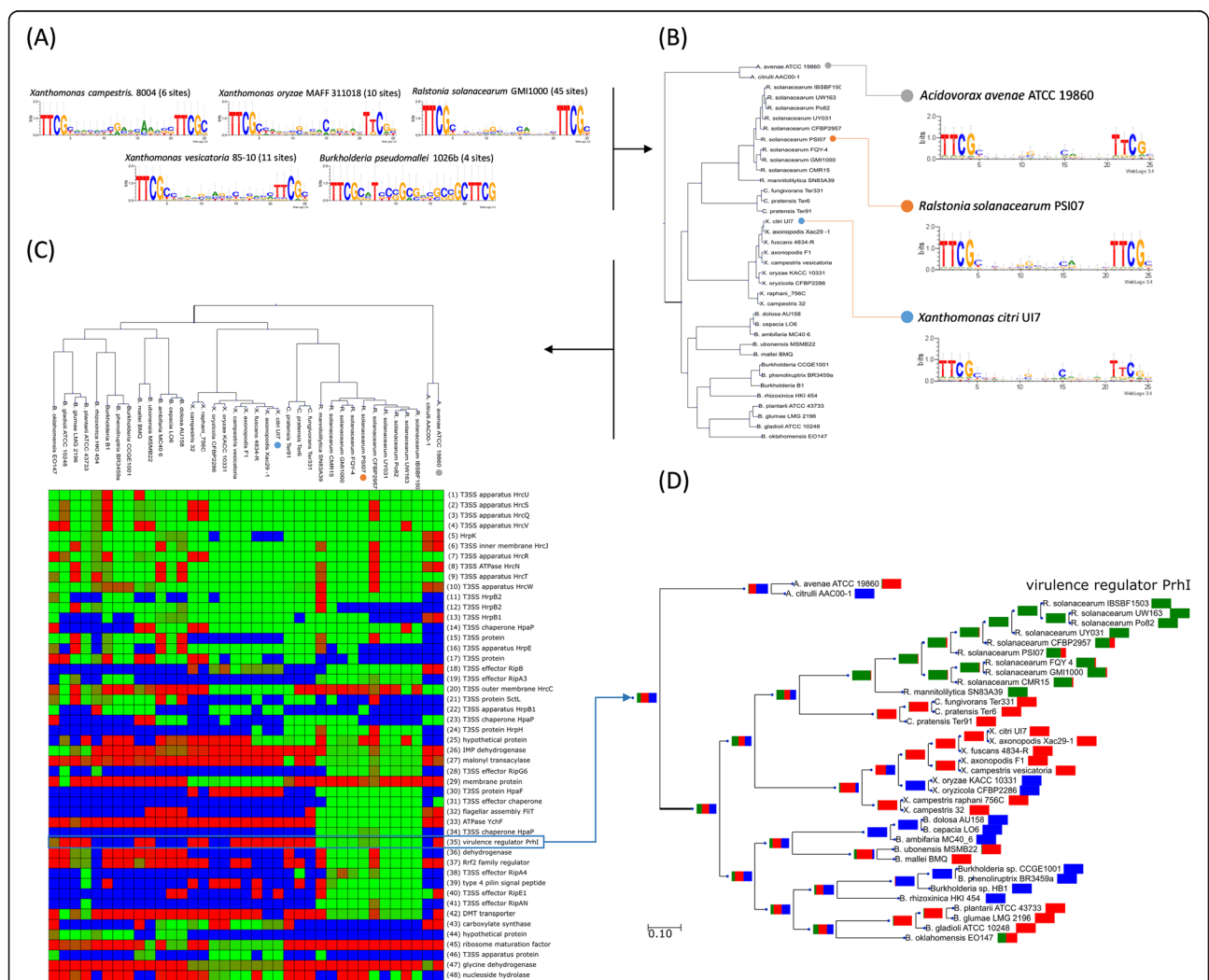


Fig. 4 Comparative analysis of the HrpB/X regulon. **a** Experimental HrpB and HrpX motifs used in the analysis, with number of available binding sites indicated. **b** Weighted mixture motifs generated by CGB on different target species following the HrpB/X-based phylogenetic tree. **c** Partial view of the CGB-generated heatmap depicting the HrpB/X regulon across 37 complete bacterial genomes encompassing members of the *Comamonadaceae*, *Burkholderiaceae*, *Oxalobacteraceae* and *Xanthomonadaceae* families. Each row represents an orthologous group, sorted by average posterior probability of regulation. Cells are colored from green (regulation) to red (no regulation); blue denotes ortholog absence. Colored circles next to species names indicate propagated motif (panel B) used in the genome search for the species. **d** Bootstrapped ancestral state reconstruction of the posterior probability of regulation for virulence regulator PhrI

The results (Fig. 4c) reveal that the only elements of the HrpB/X regulon conserved across all the species analyzed are the members of the two core operons that define the structural elements of the type III secretion system. Beyond this core HrpB/X regulon, the analysis (Fig. 4c) shows the divergent uptake of different effectors and chaperones in *Ralstonia*, *Burkholderia* and *Xanthomonas* species, indicating that the core T3SS effectome under HrpB/X regulation in each of these genera diverged soon after each obtained its T3SS module, providing specialized functions in each group [47]. This divergent uptake can be traced to specific lineages through ancestral state reconstruction (Fig. 4d), as in the case of the *Ralstonia* genus for the virulence regulator PrhI [48]. Our results showcase the ability of CGB to automatically disseminate experimental TF-binding motif information among target species in a principled manner (Fig. 4a&b), to use all available sequence files for any given genome (multiple chromosomes and plasmids), to highlight the core elements of a transcriptional regulatory network spanning multiple bacterial orders (Fig. 4c), and to provide a formal inference of the ancestral state of regulation for any given gene (Fig. 4d).

Reconstruction of the SOS response network in the Balneolaeota phylum

Metagenomic, single-cell and systematic large-scale genomic sequencing studies have uncovered the existence of several new major bacterial phyla [13–16]. Genomic data corresponding to these phyla is often only available as whole-genome shotgun assemblies and is hence not amenable to comparative genomics studies using available platforms that rely on precompiled complete genome datasets. The Balneolaeota phylum comprises several genera of halotolerant bacteria, but there is scant experimental information on their physiology [49]. Here we coupled motif discovery with comparative genomics using CGB to reconstruct the LexA regulon of the Balneolaeota. Motif discovery (Fig. 5a) was performed on the promoter regions of Balneolaeota LexA homologs with MEME, and the resulting motif was used as input for CGB analysis without phylogenetic weighting. After performing the comparative analysis of the putative LexA regulon across all species with available genome sequence in the Balneolaeota phylum (Fig. 5b), we validated *in vitro* predicted LexA-binding sites in the promoter region of all genes with orthologs in at least six of the seven species analyzed and presenting an average inter-species posterior probability of regulation above 0.5 (Fig. 5c). The resulting EMSAs on *Balneola vulgaris* and *Rhodohalobacter halophilus* promoters (Fig. 5c) reveal that all the predicted LexA-regulated promoters are bound by LexA. The high precision illustrates the usefulness of leveraging the comparative genomics approach

implemented in CGB to boost the accuracy of *in silico* prediction of TF-regulated genes on individual genomes.

Our results (Fig. 5b) reveal that the Balneolaeota LexA protein binds a novel direct repeat motif with consensus sequence TTACACATATTTTWTACATA (Fig. 5a; Additional File 1). In spite of substantial operon rearrangement, the Balneolaeota LexA regulon encodes a SOS response network encompassing the LexA repressor and the inducer of the system (RecA), as well as several translesion synthesis polymerases (polymerases IV and V), a helicase, a radical SAM protein and a predicted SOS peptidase [50] (Fig. 5c). These results represent the first description of the LexA regulon in the Balneolaeota phylum, reinforcing the notion that translesion synthesis is the primary role of the SOS response and that this system has convergently evolved to regulate error-prone polymerases following drastic changes to the LexA-binding motif [51]. They also illustrate the complex evolutionary history of the LexA protein, which appears to have independently evolved in several instances the ability to recognize a direct repeat motif structure, as opposed to the canonical palindromic LexA-binding motif [51–53]. This example illustrates the ability of CGB to leverage draft genome sequence data to infer regulons in novel phyla, and to operate in tandem with comparative motif discovery methods, using a single instance of the inferred TF-binding motif uniformly distributed across target species.

Conclusions

Comparative genomics is a powerful method to infer the composition and evolutionary history of prokaryotic transcriptional regulatory networks. The pipeline reported here, CGB, automates the comparative genomics analysis under a formal Bayesian probabilistic framework, enabling the use of ancestral state reconstruction to infer the evolutionary make up of TRNs. The system also enables automated integration of multiple sources of experimental information, forsaking the need to manually port TF-binding motif information. By means of a dynamic operon prediction algorithm and real-time ortholog detection, CGB enables users to analyze any record of draft and complete genomic data available in the NCBI GenBank and RefSeq databases, providing an unparalleled degree of flexibility for this type of comparative analyses. The lack of reliance on external databases, such as operon databases, makes CGB resilient to unexpected downtimes on third-party systems, with the exception of NCBI services. The results reported for two transcriptional regulatory different systems showcase the flexibility of CGB and provide evidence of convergent and divergent evolution in regulatory networks.

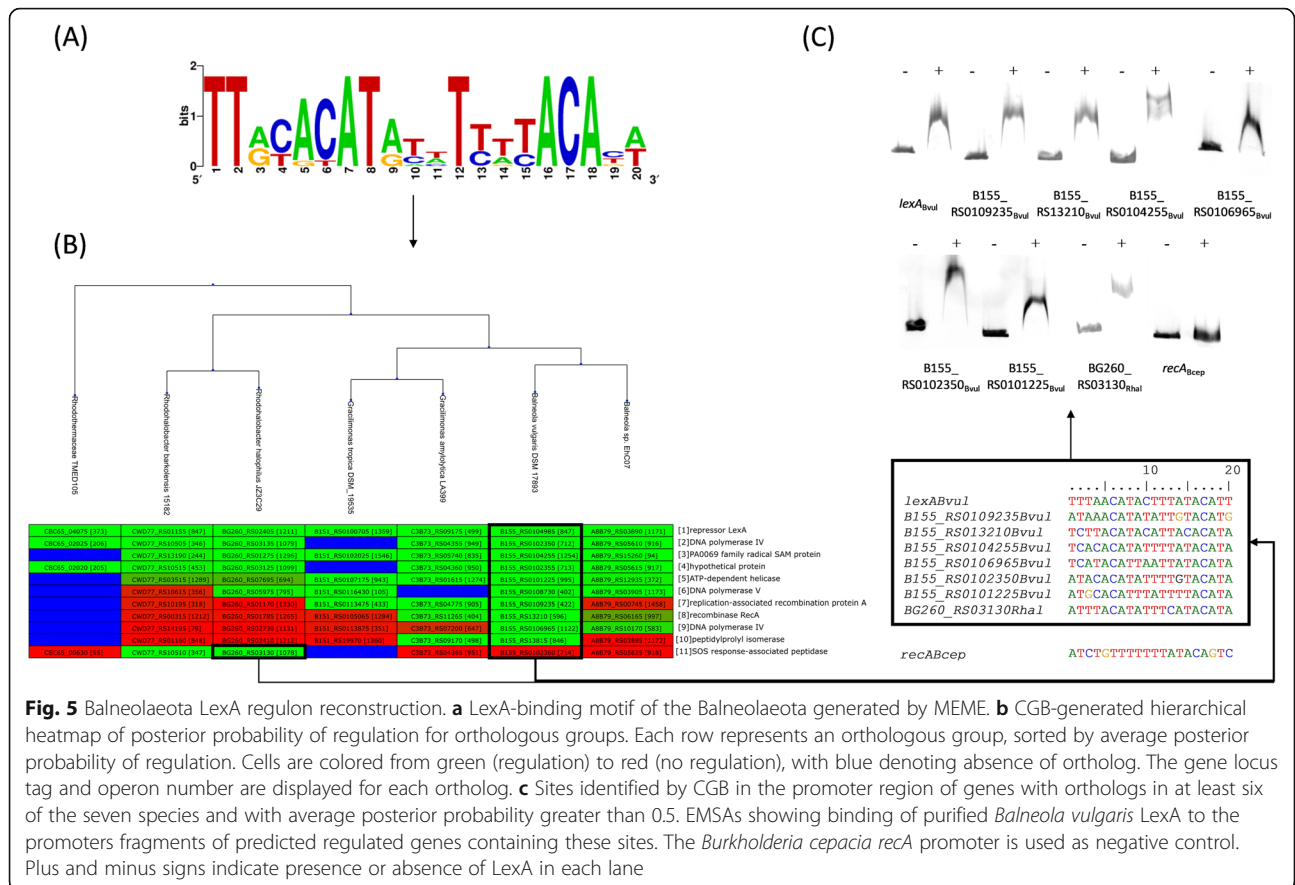


Fig. 5 Balneolaeta LexA regulon reconstruction. **a** LexA-binding motif of the Balneolaeta generated by MEME. **b** CGB-generated hierarchical heatmap of posterior probability of regulation for orthologous groups. Each row represents an orthologous group, sorted by average posterior probability of regulation. Cells are colored for green (regulation) to red (no regulation), with blue denoting absence of ortholog. The gene locus tag and operon number are displayed for each ortholog. **c** Sites identified by CGB in the promoter region of genes with orthologs in at least six of the seven species and with average posterior probability greater than 0.5. EMSAs showing binding of purified *Balneola vulgaris* LexA to the promoters fragments of predicted regulated genes containing these sites. The *Burkholderia cepacia* *recA* promoter is used as negative control. Plus and minus signs indicate presence or absence of LexA in each lane

Materials and methods

CGB platform

CGB is a Python library for comparative genomics of transcriptional regulation in prokaryotes. It is written entirely in Python 2.7 using the object-oriented programming paradigm and deployed as a conda virtual environment (Continuum Analytics). CGB is freely available under a GPL license on GitHub [54]. CGB external requirements are Clustal Omega [55], NCBI BLAST+ [56] and BayesTraits [35].

Motif data and motif discovery

Collections of experimentally-validated TF-binding sites for HrpB and HrpX were downloaded from the CollecTF database [57]. HrpB orthologs in Proteobacteria and LexA orthologs in the Balneolaeta phylum were detected as best reciprocal BLAST hits using, respectively, the *Ralstonia solanacearum* HrpB protein [WP_011004170.1] and the *Verrucomicrobium spinosum* DSM 4136 LexA protein [WP_009959117] as a queries and a cut-off e-value of 10^{-30} . The upstream regions (-250, +50 bp from predicted translational start site) of genes coding for identified LexA orthologs were downloaded from the NCBI GenBank database and input to MEME for motif discovery using the any number of repetitions

(ANR) site distribution and motif width limits of 10–22 bp. CGB configuration files for the analyses here reported are provided as supplementary material (Additional File 2).

Protein purification and electro-mobility shift assays

The *Balneola vulgaris* DSM 17893 *lexA* gene [B155_RS0104985] was synthesized by ATG:biosynthetics GmbH, Germany, subcloned into the pUA1108 vector [58] and overexpressed in *E. coli* BL21-CodonPlus (DE3)-RIL (Stratagene) cells. The resulting LexA His-tagged protein was purified following a previously described protocol [52]. Electro-mobility shift assays (EMSA) were performed using 100 bp-long DNA probes (Additional File 3). Probes were generated using two complementary synthetic oligonucleotides centered on predicted LexA-binding sites and performing PCR with M13 forward and reverse digoxigenin-labeled oligos, as described previously [59]. EMSAs were carried out on a mixture containing 20 ng of each digoxigenin-marked DNA probe and 40 nM of purified LexA protein [60]. Samples were loaded onto 6% non-denaturing Tris-glycine polyacrylamide gels and digoxigenin-labeled DNA-protein complexes were detected using the manufacturer’s protocol (Roche NimbleGen).

Supplementary information

Supplementary information accompanies this paper at <https://doi.org/10.1186/s12864-020-06838-x>.

Additional file 1.

Additional file 2.

Additional file 3.

Abbreviations

ANR: Any Number of Repetitions; BLAST: Basic Local Alignment Search Tool; CGB: Comparative Genomics of Bacteria; CSV: Comma Separated Value; DNA: Deoxyribonucleic Acid; EMSA: Electro-Mobility Shift Assay; FPR: False Positive Rate; GPL: GNU General Public License; JSON: JavaScript Object Notation; MEME: Multiple EM for Motif Elicitation; NCBI: National Center for Biotechnology Information; ODB: Operon DataBase; PCR: Polymerase Chain Reaction; PSSM: Position-Specific Scoring Matrix; PSWM: Position-Specific Weight Matrix; SAM: S-adenosyl-L-methionine; SOS: Save Our Souls; T3S5: Type III Secretion System; TF: Transcription factor; TRN: Transcriptional regulatory network

Acknowledgements

The authors wish to thank Patrick O'Neill for insightful discussions on the CGB mathematical framework, Stéphane Genin for assistance in the interpretation of results of the HrpB/X analysis and Mark Pagel for guidance in the application of BayesTraits for ancestral state reconstruction in the CGB setting.

About this supplement

This article has been published as part of BMC Genomics Volume 21 Supplement 5, 2020: Selected articles from the 15th International Symposium on Bioinformatics Research and Applications (ISBRA-19): genomics. The full contents of the supplement are available online at <https://bmcbgenomics.biomedcentral.com/articles/supplements/volume-21-supplement-5>.

Authors' contributions

IE and SK designed and coded the CGB platform. MSO performed all in vitro techniques. ACP wrote scripts and performed motif discovery analyses. IE run comparative genomics analyses. IE and JB coordinated the research. IE, SK and MSO drafted the manuscript. All authors revised and approved the manuscript and interpreted results.

Funding

This work was funded by the US National Science Foundation award MCB-1158056 (UMBC), and by grant BIO2016-77011-R from the Spanish Ministerio de Economía y Competitividad (UAB). MS-O was supported by a predoctoral fellowship from the Ministerio de Educación, Cultura y Deporte de España. Publication costs are funded by the UMBC Department of Biological Sciences.

Availability of data and materials

Nucleotide sequences analyzed in the current study are available in the NCBI GenBank repository, <https://www.ncbi.nlm.nih.gov/nuccore>. TF-binding motifs are available in the CollecTF database, <http://www.collectf.org/>.

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing interests

The authors declare that they have no competing interests.

Author details

¹University of Maryland Baltimore County, Baltimore, MD 21250, USA.

²Universitat Autònoma de Barcelona, 08193 Bellaterra, Spain.

Received: 7 May 2020 Accepted: 16 June 2020

Published: 16 December 2020

References

- Ptashne M. Regulation of transcription: from lambda to eukaryotes. *Trends Biochem Sci.* 2005;30:275–9.
- Ishihama A. Prokaryotic genome regulation: multifactor promoters, multitarget regulators and hierarchic networks. *FEMS Microbiol Rev.* 2010;34:628–45.
- Orphanides G, Reinberg D. A unified theory of gene expression. *Cell.* 2002;108:439–51.
- Gelfand MS. Prediction of function in DNA sequence analysis. *J Comput Biol.* 1995;2:87–115.
- Osada R, Zaslavsky E, Singh M. Comparative analysis of methods for representing and searching for transcription factor binding sites. *Bioinformatics.* 2004;20:3516–25.
- Erill I, O'Neill MC. A reexamination of information theory-based methods for DNA-binding site identification. *BMC Bioinformatics.* 2009;10:57.
- Gelfand MS, Novichkov PS, Novichkova ES, Mironov AA. Comparative analysis of regulatory patterns in bacterial genomes. *Brief Bioinformatics.* 2000;1:357–71.
- Tan K, Moreno-Hagelsieb G, Collado-Vides J, Stormo GD. A comparative genomics approach to prediction of new members of regulons. *Genome Res.* 2001;11:566–84.
- Cornish JP, Matthews F, Thomas JR, Erill I. Inference of self-regulated transcriptional networks by comparative genomics. *Evol Bioinformatics Online.* 2012;8:449–61.
- Novichkov PS, Rodionov DA, Stavrovskaya ED, Novichkova ES, Kazakov AE, Gelfand MS, et al. RegPredict: an integrated system for regulon inference in prokaryotes by comparative genomics approach. *Nucleic Acids Res.* 2010;38(Web Server issue):W299–307.
- Liu B, Zhou C, Li G, Zhang H, Zeng E, Liu Q, et al. Bacterial regulon modeling and prediction based on systematic cis regulatory motif analyses. *Sci Rep.* 2016;6:23030.
- Mao F, Dam P, Chou J, Olman V, Xu Y. DOOR: a database for prokaryotic operons. *Nucleic Acids Res.* 2009;37(Database issue):D459–63.
- Rinke C, Schwientek P, Sczyrba A, Ivanova NN, Anderson IJ, Cheng J-F, et al. Insights into the phylogeny and coding potential of microbial dark matter. *Nature.* 2013;499:431–7.
- Di Rienzi SC, Sharon I, Wrighton KC, Koren O, Hug LA, Thomas BC, et al. The human gut and groundwater harbor non-photosynthetic bacteria belonging to a new candidate phylum sibling to cyanobacteria. *Elife.* 2013;2:e01102.
- Herlemann DPR, Geissinger O, Ikeda-Ohtsubo W, Kunin V, Sun H, Lapidus A, et al. Genomic Analysis of “*Elusimicrobium minutum*,” the First Cultivated Representative of the Phylum “*Elusimicrobia*” (Formerly Termite Group 1). *Appl Environ Microbiol.* 2009;75:2841–9.
- Wu D, Hugenholtz P, Mavromatis K, Pukall R, Dalin E, Ivanova NN, et al. A phylogeny-driven genomic encyclopaedia of Bacteria and Archaea. *Nature.* 2009;462:1056–60.
- Oberto J. FITBAR: a web tool for the robust prediction of prokaryotic regulons. *BMC Bioinformatics.* 2010;11:554.
- Bykova NA, Favorov AV, Mironov AA. Hidden Markov models for evolution and comparative genomics analysis. *PLoS One.* 2013;8:e65012.
- Kılıç S, Erill I. Assessment of transfer methods for comparative genomics of regulatory networks in bacteria. *BMC Bioinformatics.* 2016;17 Suppl 8:277.
- Harrison PW, Lower RPJ, Kim NKD, Young JPW. Introducing the bacterial “chromid”: not a chromosome, not a plasmid. *Trends Microbiol.* 2010;18:141–8.
- Erill I, Campoy S, Mazon G, Barbe J. Dispersal and regulation of an adaptive mutagenesis cassette in the bacteria domain. *Nucleic Acids Res.* 2006;34:66–77.
- Lee C, Huang C-H. LASAGNA: a novel algorithm for transcription factor binding site alignment. *BMC Bioinformatics.* 2013;14:108.
- Thompson JD, Higgins DG, Gibson TJ. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res.* 1994;22:4673–80.
- Munch R, Hiller K, Grote A, Scheer M, Klein J, Schobert M, et al. Virtual Footprint and PRODORIC: an integrative framework for regulon prediction in prokaryotes. *Bioinformatics.* 2005;21:4187–9.

25. Yellaboina S, Seshadri J, Kumar MS, Ranjan A. PredictRegulon: a web server for the prediction of the regulatory protein binding sites and operons in prokaryote genomes. *Nucleic Acids Res.* 2004;32 suppl_2:W318–20.
26. Hobbs ET, Pereira T, O'Neill PK, Erill I. A Bayesian inference method for the analysis of transcriptional regulatory networks in metagenomic data. *Algorithms Mol Biol.* 2016;11. <https://doi.org/10.1186/s13015-016-0082-8>.
27. Schneider TD, Stormo GD, Gold L, Ehrenfeucht A. Information content of binding sites on nucleotide sequences. *J Mol Biol.* 1986;188:415–31.
28. Taboada B, Estrada K, Ciria R, Merino E. Operon-mapper: a web server for precise operon identification in bacterial and archaeal genomes. *Bioinformatics.* 2018;34:4118–20.
29. Westover BP, Buhler JD, Sonnenburg JL, Gordon JI. Operon prediction without a training set. *Bioinformatics.* 2005;21:880–8.
30. Price MN, Huang KH, Alm EJ, Arkin AP. A novel method for accurate operon predictions in all sequenced prokaryotes. *Nucleic Acids Res.* 2005;33:880–92.
31. Chuang L-Y, Chang H-W, Tsai J-H, Yang C-H. Features for computational operon prediction in prokaryotes. *Brief Funct Genomics.* 2012;11:291–9.
32. Hertz GZ, Stormo GD. Identifying DNA and protein patterns with statistically significant alignments of multiple sequences. *Bioinformatics.* 1999;15:563–77.
33. Nichio BTL, Marchaukoski JN, Raittz RT. New tools in Orthology analysis: a brief review of promising perspectives. *Front Genet.* 2017;8:165.
34. O'Neill PK, Or M, Erill I. scnRCA: A Novel Method to Detect Consistent Patterns of Translational Selection in Mutationally-Biased Genomes. *PLoS ONE.* 2013;8:e76177.
35. Pagel M, Meade A, Barker D. Bayesian estimation of ancestral character states on phylogenies. *Syst Biol.* 2004;53:673–84.
36. Walker GC. The SOS response of *Escherichia coli*. In: Neidhart FC, Ingram JL, Low KB, Magasanik B, Schaechter M, Umberger HE, editors. *Escherichia coli and Salmonella typhimurium: cellular and molecular biology*. Washington, D.C.: American Society for Microbiology; 1987. <http://www.utoronto.ca/medicalgenetics/student%20site/undergraduate/MGY/WalkerSOS.pdf>.
37. Fernandez De Henestrosa AR, Ogi T, Aoyagi S, Chafin D, Hayes JJ, Ohmori H, et al. Identification of additional genes belonging to the LexA regulon in *Escherichia coli*. *Mol Microbiol.* 2000;35:1560–72.
38. Harmon FG, Rehrauer WM, Kowalczykowski SC. Interaction of *Escherichia coli* RecA protein with LexA repressor. II. Inhibition of DNA strand exchange by the uncleavable LexA S119A repressor argues that recombination and SOS induction are competitive processes. *J Biol Chem.* 1996;271:23874–83.
39. Erill I, Campoy S, Barbe J. Aeons of distress: an evolutionary perspective on the bacterial SOS response. *FEMS Microbiol Rev.* 2007;31:637–56.
40. Cirz RT, Jones MB, Gingles NA, Minogue TD, Jarrahi B, Peterson SN, et al. Complete and SOS-mediated response of *Staphylococcus aureus* to the antibiotic ciprofloxacin. *J Bacteriol.* 2007;189:531–9.
41. Durbach SI, Andersen SJ, Mizrahi V. SOS induction in mycobacteria: analysis of the DNA-binding activity of a LexA-like repressor and its role in DNA damage induction of the recA gene from *Mycobacterium smegmatis*. *Mol Microbiol.* 1997;26:643–53.
42. Au N, Kuester-Schoeck E, Mandava V, Bothwell LE, Canny SP, Chachu K, et al. Genetic composition of the *Bacillus subtilis* SOS system. *J Bacteriol.* 2005;187:7655–66.
43. van der Veen S, van Schalkwijk S, Molenaar D, de Vos WM, Abee T, Wells-Bennik MHJ. The SOS response of *Listeria monocytogenes* is involved in stress resistance and mutagenesis. *Microbiology.* 2010;156:374–84.
44. Jochmann N, Kurze A-K, Czaja LF, Brinkrolf K, Brune I, Huser AT, et al. Genetic makeup of the *Corynebacterium glutamicum* LexA regulon deduced from comparative transcriptomics and in vitro DNA band shift assays. *Microbiology.* 2009;155:1459–77.
45. Valls M, Genin S, Boucher C. Integrated regulation of the type III secretion system and other virulence determinants in *Ralstonia solanacearum*. *PLoS Pathog.* 2006;2:e82.
46. Xue X, Zou L, Ma W, Liu Z, Chen G. Identification of 17 HrpX-regulated proteins including two novel type III effectors, XOC_3956 and XOC_1550, in *Xanthomonas oryzae* pv. *oryzicola*. *PLoS ONE.* 2014;9:e93205.
47. Saier MH. Evolution of bacterial type III protein secretion systems. *Trends Microbiol.* 2004;12:113–5.
48. Brito B, Aldon D, Barberis P, Boucher C, Genin S. A signal transfer system through three compartments transduces the plant cell contact-dependent signal controlling *Ralstonia solanacearum* hrp genes. *Mol Plant-Microbe Interact.* 2002;15:109–19.
49. Hahnke RL, Meier-Kolthoff JP, García-López M, Mukherjee S, Huntemann M, Ivanova NN, et al. Genome-based taxonomic classification of Bacteroidetes. *Front Microbiol.* 2016;7:2003.
50. Aravind L, Anand S, Iyer LM. Novel autoproteolytic and DNA-damage sensing components in the bacterial SOS response and oxidized methylcytosine-induced eukaryotic DNA demethylation systems. *Biol Direct.* 2013;8:20.
51. Erill I, Campoy S, Kiliç S, Barbé J. The *Verrucomicrobia* LexA-binding motif: insights into the evolutionary dynamics of the SOS response. *Front Mol Biosci.* 2016;3. <https://doi.org/10.3389/fmolb.2016.00033>.
52. Sánchez-Osuna M, Barbé J, Erill I. Comparative genomics of the DNA damage-inducible network in the Patescibacteria. *Environ Microbiol.* 2017;19:3465–74.
53. Mazon G, Campoy S, Erill I, Barbe J. Identification of the *Acidobacterium capsulatum* LexA box reveals a lateral acquisition of the Alphaproteobacteria *lexA* gene. *Microbiology.* 2006;152 Pt 4:1109–18.
54. Kiliç S, Erill I. CGB: Comparative genomics of transcriptional regulation in Bacteria, <https://github.com/erilllab/cgb>, last accessed 2019/02/07. English. <https://github.com/erilllab/cgb>. Accessed 7 Feb 2019.
55. Sievers F, Wilm A, Dineen D, Gibson TJ, Karplus K, Li W, et al. Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal omega. *Mol Syst Biol.* 2011;7:539.
56. NCBI Resource Coordinators. Database resources of the National Center for biotechnology information. *Nucleic Acids Res.* 2017;45:D12–7.
57. Kiliç S, White ER, Sagitova DM, Cornish JP, Erill I. CollecTF: a database of experimentally validated transcription factor-binding sites in Bacteria. *Nucleic Acids Res.* 2014;42(Database issue):D156–60.
58. Mayola A, Irazoki O, Martínez IA, Petrov D, Menolascina F, Stocker R, et al. RecA protein plays a role in the chemotactic response and chemoreceptor clustering of *Salmonella enterica*. *PLoS One.* 2014;9:e105578.
59. Campoy S, Fontes M, Padmanabhan S, Cortes P, Llagostera M, Barbe J. LexA-independent DNA damage-mediated induction of gene expression in *Myxococcus xanthus*. *Mol Microbiol.* 2003;49:769–81.
60. Sanchez-Alberola N, Campoy S, Barbe J, Erill I. Analysis of the SOS response of *Vibrio* and other bacteria with multiple chromosomes. *BMC Genomics.* 2012;13:58.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Ready to submit your research? Choose BMC and benefit from:

- fast, convenient online submission
- thorough peer review by experienced researchers in your field
- rapid publication on acceptance
- support for research data, including large and complex data types
- gold Open Access which fosters wider collaboration and increased citations
- maximum visibility for your research: over 100M website views per year

At BMC, research is always in progress.

Learn more [biomedcentral.com/submissions](https://www.biomedcentral.com/submissions)

