bioRxiv preprint doi: https://doi.org/10.1101/410621; this version posted September 7, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Title

Tracing the phylogenetic history of the Crl regulon through the *Bacteria* and *Archaea* genomes.

Keywords

Crl regulon; Stress response; Transcription factors; Comparative genomics; Bacteria; Archaea

Authors

Santos-Zavaleta A[1], Pérez-Rueda E[2], Sánchez-Pérez M[1], Velázquez-Ramírez D A[1], and Collado-Vides J[1].

Affiliations

[1] Programa de Genómica Computacional, Centro de Ciencias Genómicas, Universidad Nacional Autónoma de México. Cuernavaca, Morelos 62210, México.

[2] Instituto de Investigaciones en Matemáticas Aplicadas y en Sistemas. Universidad Nacional Autónoma de México, Unidad Académica Yucatán. Mérida, Yucatán 97302, México.

Correspondence should be addressed to:

asantos@ccg.unam.mx and ernesto.perez@iimas.unam.mx

bioRxiv preprint doi: https://doi.org/10.1101/410621; this version posted September 7, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Abstract

Crl, identified for curli production, is a small transcription factor that stimulates the association of the σ^S factor (RpoS) with the RNA polymerase core through direct and specific interactions, increasing the transcription rate of genes during the transition from exponential to stationary phase at low temperatures, and it uses indole as an effector molecule. The lack of a comprehensive collection of information on the Crl regulon makes it difficult to identify a dominant function of Crl and to generate any hypotheses concerning its taxonomical distribution in archaeal and bacterial organisms. In this work, based on a systematic literature review, we identified the first comprehensive dataset of 86 genes under the control of Crl in the bacterium *Escherichia coli* K-12; those genes correspond to 40% of the σ^S regulon in this bacterium. Based on an analysis of orthologs in 18 archaeal and 69 bacterial taxonomical divisions and using *E. coli* K-12 as a framework, we suggest three main events that resulted in this regulon's actual form: (i) in a first step, *rpoS*, a gene widely distributed in bacteria and archaea cellular domains, was recruited to regulate genes involved in ancient metabolic processes, such as those associated with glycolysis and the tricarboxylic acid cycle; (ii) in a second step, the regulon recruited those genes involved in metabolic processes, which are mainly taxonomically constrained to *Proteobacteria*, with some secondary losses, such as those genes involved in responses to stress or starvation and cell adhesion, among others; and (iii) in a posterior step, Crl was recruited as a consequence of its emergence in *Enterobacteriaceae*. Therefore, we suggest that the regulon Crl is highly flexible for phenotypic adaptation, probably as consequence of the diverse

bioRxiv preprint doi: https://doi.org/10.1101/410621; this version posted September 7, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

growth environments associated with all organisms in which members of this regulatory network are present.

Introduction

Gene expression in bacteria is coordinated through the interplay of sigma (σ) factors on the core RNA polymerase (RNAP) [1] and DNA-binding transcription factors (TFs), providing bacteria with the ability to express multiple genes under different metabolic stimuli or growth conditions. In the bacterium *Escherichia coli* K-12, seven sigma factors have been experimentally identified, together with around 300 different TFs responsible for recognizing and activating almost all of their genes. Among these, RpoD, or σ^{70}, regulates around 40% of the total gene repertoire, whereas alternative sigma factors such as RpoS (σ^{S}), the master regulator of the stationary-phase response [2], regulate between 5 and 10% of the total genes in *E. coli* K-12 [3,4].

Sigma factors and TFs regulate a large diversity of genes, hierarchically organized in regulons [5]. Previous comparative genomics studies have suggested that regulons exhibit considerable plasticity across the evolution of bacterial species [6]. In this regard, comparison of the gene composition of the PhoPQ regulon in *E. coli* and *Salmonella enterica* serovar Typhimurium revealed a very small overlap in both species, suggesting a low similarity in composition between the target genes that are specifically PhoP regulated in *S.* Typhimurium strains and in *E. coli* K-12 [7]. Incidentally, this plasticity in bacterial regulons is evidence of lineage-specific modifications [8].

bioRxiv preprint doi: https://doi.org/10.1101/410621; this version posted September 7, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

In this regard, we conducted an exhaustive analysis concerning the conservation of the Crl regulon in Bacteria and Archaea cellular domains, using as a reference the currently known system in *E. coli* K-12. Contrary to the most common regulatory mechanisms that involve the direct binding to operators or activators, Crl is an RNAP holoenzyme assembly factor that was originally identified in curli production. It is expressed at low temperatures (30°C) [9] during the transition phase between the exponential and stationary phases, under low osmolarity, as well as in stationary phase [10]. In *E. coli*, Crl has a global regulatory effect in stationary phase, through σ^S, as it reorganizes the transcriptional machinery [11], stimulating the association of σ^S with the RNAP core, tilting the competition between σ^S and σ^{70} during the stationary phase in response to different stress conditions [12, 13] [9, 14]; its production is concomitant with the accumulation of σ^S [9].

Assembling the different pieces of the Crl regulon and its regulatory network into one global picture is one of our objectives in this work. The evaluation of this regulon in *Bacteria* and *Archaea* will provide clues about how the regulation of genes by Crl has been recruited in all the organisms, i.e., if the regulated genes were recruited similar to Crl or if they followed different pathways. To this end, 86 genes under the control of Crl in *E. coli* K-12 were compiled from exhaustive literature searches. To our knowledge, this is the first attempt to describe the genes regulated by Crl in *E. coli* K-12; in addition, Crl homologs were searched among bacterial and archaeal genomes and identified in low copy numbers, constrained to *Enterobacteriaceae* species. Finally, members of the regulon were identified as widely distributed beyond enterobacteria, suggesting that Crl was recruited in a secondary evolutionary event to regulate a specific subset of genes,

bioRxiv preprint doi: https://doi.org/10.1101/410621; this version posted September 7, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

most likely genes involved in a functional response in enterobacteria to contend against starvation.

Methods

Identification of Crl-regulated genes

We performed an exhaustive search of the literature related to Crl in *E. coli* K-12 in PubMed [15] under the following search strategy: “coli in the title (to exclude spurious articles) and Crl in all fields,” and “regulation” and “*rpoS*” with different combinations. We obtained 21 manuscripts with relevant information. In addition, genes under the control of Crl were obtained from microarray data analysis with *crl* mutants and with data processed by our authors (Table 1). Finally, we searched for gene/operon notes in RegulonDB and EcoCyc [3, 16] for Crl interactions and σ^S promoters; for assembling the network of regulation of Crl; and to identify associations between the TF and regulatory role for each member of the Crl regulon. We must remember that RegulonDB is the main database on transcriptional regulation in *E. coli* K-12 of manually curated data from scientific publications.

The regulatory network generated was displayed using the Cytoscape program, version 3.3.0 [17], with information obtained in the identified papers as well as information contained in RegulonDB [3]. Genes under Crl control were classified based on Gene Ontology (GO) annotations (http://www.geneontology.org/) using the Gene Association Format (GAF 2.0) as well as the Multifun classification scheme [18]. An enrichment analysis was carried out to find overrepresented annotations, using the PANTHER Classification system program, version 12.0.

bioRxiv preprint doi: https://doi.org/10.1101/410621; this version posted September 7, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Based on our results, we selected biological processes and *E. coli* as parameters [19, 20]. In addition, we used KEGG to categorize the functions of GOs (http://www.genome.jp/kegg-bin/show_organism?menu_type=pathway_maps&org=eco) [21].

Identification of Crl homologs

The Crl protein sequence of *E. coli* K-12 (ID: P24251) was used as the seed to scan all the bacterial and archaeal genomes via a BLASTp search [22] (E-value ≤ 10^{-3} and coverage ≥ 60%). All proteins were compared and aligned using the Muscle algorithm [23] with default parameters, and results were manually edited with the program Jalview. Finally, a phylogeny was inferred by the maximum likelihood method with 1,000 replicates by using the program MEGA [24] and the Tamura-Nei model.

Identification of orthologous genes

Orthologous genes have been classically defined as encoding proteins in different species that evolved from a common ancestor via speciation [25] and have retained the same function. In this work, orthologs were identified by searching for bidirectional best hits (BDBHs) in other organisms [26].

Clustering of orthologous genes

In order to evaluate the taxonomical distribution of the genes belonging to the Crl regulon, their orthologs were traced along 18 archaeal and 69 bacterial

bioRxiv preprint doi: https://doi.org/10.1101/410621; this version posted September 7, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

taxonomical divisions. To this end, the relative abundance of the orthologs was calculated as the fraction of genomes in the group that had one ortholog, divided by the total number of genomes per phylum, i.e., the ratio (total number of orthologs in a phylum) / (total number of organisms in phylum). Thus, the value goes from 0 (absence of orthologs) to 1, or 100%, when all organisms in the division contain an ortholog. The corresponding matrix was analyzed with a hierarchical complete linkage-clustering algorithm with correlation uncentered as the similarity measure. We used the program MeV to perform the analyses (http://www.tm4.org/mev).

Results

A total of 86 genes were identified as members of the Crl regulon

Available information regarding the Crl regulon was integrated through an exhaustive review of the literature. In this regard, diverse experimental evidences were considered significant for determining the association between the regulated genes and Crl protein regulator, such as gene expression analysis (transcriptional fusions), mapping of signal intensities (RNA-seq or microarray analysis), and inferences made from a mutant phenotype (mutation of a TF with a visible cell phenotype), among other analyses. In total, 52 genes of the 86 were identified in this work as new members of the σ^S sigmulon based on microarray data and *crl rpoS* double mutants [9-14, 27-29] (Supplementary material Table S1). From the 86 genes identified as members of this regulon (see Table 1 and Figure 1), 34 have a σ^S-type promoter that were experimentally determined [3] and 8 genes have 13 σ^S-type promoters that were predicted by computational approaches.

bioRxiv preprint doi: https://doi.org/10.1101/410621; this version posted September 7, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

These 86 genes are organized in 77 transcription units (TUs), where 52% are TUs with only one gene.

Previously, genes under the control of Crl were classified in four main categories depending on their role(s) in the cell: DNA metabolism, central metabolism, response to environmental modifications, and miscellaneous [12]. Based on Gene Ontology (GO) annotations, multifunctional classification, and KEGG pathway maps to categorize functions, Crl-regulated genes appear to be involved in metabolic processes such as energy metabolism, amino acid, carbohydrate, and lipid metabolism, and biosynthetic processes such as glycan biosynthesis and biosynthesis of other secondary metabolites, among other metabolic processes. This function correlates with results of the enrichment analysis using PANTHER, which showed that catabolic processes, metabolic processes, and cellular responses to xenobiotic stimuli were overrepresented among the functions associated with genes under the control of Crl (See Figure 2).

In general, genes under Crl control are involved in regulating many aspects of cellular metabolism through Crl's interaction with a subset of genes of the σ^S regulon [9] in addition to quorum sensing playing a major role in cell-to-cell communication during stationary phase and in different processes such as biofilm formation or virulence, and also transporters [12] and genes involved in the uptake and utilization of β-glucosides [29].

Composition of the Crl regulon

bioRxiv preprint doi: https://doi.org/10.1101/410621; this version posted September 7, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

In order to determine whether additional TFs also regulate the genes under the control of Crl, RegulonDB was used to evaluate how genes associated with Crl are also regulated by alternative TFs or sigma factors. A total of 24 genes were identified as exclusively controlled by Crl, whereas 62 are regulated by TFs beyond Crl (Supplementary material Table S2). In this regard, 55 different TFs are involved in the regulation of genes associated with Crl, including Crp, IHF, H-NS, Fis, FNR, ArcA, GadX, GadW, GadE, and CsgD (Table 1), suggesting that all genes regulated by Crl are also involved in multiple functions beyond the stationary phase. It is interesting that six of seven global regulators identified in the regulatory network of *E. coli* are also associated with the set regulated by Crl. In addition, 19 genes of the total of Crl-regulated genes are regulated by one TF, 11 by two TFs, and 14 by three different TFs. Finally, 73 (85%) genes are regulated positively, whereas 12 (15%) genes are regulated negatively (Table 1). The predominance of positive regulation suggests that genes associated with this regulon are in high demand [30], and the activities of their proteins are enhanced to contend with varied environmental stimuli. Thirty-four of the 86 genes have a σ^S-type promoter that was experimentally determined (RegulonDB). Finally, the promoters of 52 genes identified as members of Crl and of the σ^S sigmulon, based on transcriptional fusions and microarray analysis data, remain to be experimentally determined [3]. These findings suggest that Crl was probably recruited to regulate genes under previous regulation within the σ^S sigmulon.

Phylogenetic analysis of Crl

bioRxiv preprint doi: https://doi.org/10.1101/410621; this version posted September 7, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

In order to evaluate the phylogenetic history of Crl across the bacterial and archaeal cellular domains, its homologs were identified as described above in the Methods section, and a phylogenetic tree with maximum likelihood was generated (Figure 3). From this analysis, we found that Crl and its homologs are distributed almost exclusively among *Gammaproteobacteria* but do not share homology with proteins from other taxonomical divisions, as has been previously noted for *E. coli*, *Vibrio* spp., *Citrobacter* spp., *Salmonella* spp., and *Enterobacter aerogenes* [29]. Additional information suggests that Crl is less widespread and less conserved at the sequence level than σ^S [31]. In this regard, four conserved residues (Y22, F53, W56, and W82) are important for Crl activity and for Crl-σ^S interaction but not for Crl stability in *S.* Typhimurium [31]. On one hand it is probable that Crl homologs exist in some σ^S-containing bacteria; however, some species might use alternative strategies to favor σ^S interaction with the core of the RNAP [31]. Therefore, our phylogenetic analysis suggests that Crl is a protein conserved and constrained to *Gammaproteobacteria*, such as in *Vibrio* spp., *Klebsiella* spp., *Enterobacter* spp, and *Escherichia coli*. In addition, this TF was found in low copy numbers, i.e., one member of *crl* per genome. This information, together with the distribution of σ^S, suggests that the regulator was recruited as an element to regulate a subset of σ^S-regulated genes in *Gammaproteobacteria*. This result opens the question of whether genes regulated by Crl are also constrained to this taxonomical division.

Taxonomical distribution of Crl-regulated genes

Based on the identification of orthologs of 86 Crl-regulated genes, we evaluated their taxonomical distribution across archaea and bacteria sequence genomes, as

bioRxiv preprint doi: https://doi.org/10.1101/410621; this version posted September 7, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

described in Methods (See Figure 4). Based on a taxonomical profile, we determined that the evolution of the Crl regulon seems to have involved diverse losses and gains of regulatory interactions. It is possible that large portions of the regulatory network associated with Crl evolved through extensive genetic changes during the evolution of the species studied. Indeed, we suggest three main events modeled the evolution of this regulon: (i) the recruitment of a large number of genes widely distributed among *Bacteria* and *Archea*, such as those genes involved in ancient metabolic processes such as glycolysis (*fbaB*, *pykF*, *pfkA,* and *sucA*) and those involved in the tricarboxylic acid cycle (*gltA* and *sucD*) [32]; (ii) the recruitment of genes with a distribution pattern mainly constrained to *Proteobacteria*, with some secondary losses in other organisms, such as those genes involved in response to stress and starvation (*cstA* and *hdcA*) or cell adhesion (*csgA* and *csgB*), among others; and (iii) the recruitment of Crl as a consequence of its emergence in *Enterobacteriales*. It is interesting that Crl-regulated genes are also part of the σ^S sigmulon, where there are no essential genes [33-35]. All these elements suggest that the Crl regulon is highly flexible for phenotypic adaptation, probably as a consequence of the diverse growth environments associated with the organisms in which members of this regulatory network are present.

Conclusions

Crl stimulates the association of σ^S with the RNAP core in *E. coli* K-12 through direct and specific interactions, increasing the transcription rate of a subset of genes of the σ^S sigmulon. This TF has been described during the transition to

bioRxiv preprint doi: https://doi.org/10.1101/410621; this version posted September 7, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

stationary phase at low temperatures, and a recent review on the structural characterization of the Crl σ^S has been done [36]. In our work, based on an exhaustive literature search, we found 86 genes under the control of Crl in *E. coli*. These protein-coding genes were retrieved mainly from microarray and mutation analyses, among other experimentally supported evidence. These genes are associated with multiple functions, including xenobiotic processes, biofilm formation, metabolic, catabolic, and biosynthetic processes, responses to different stress conditions, and protein assembly, amino acid transport, and transcriptional processes, among others. The diverse functions regulated by Crl suggest that these genes play a fundamental role in multiple functions to respond to environmental changes, mainly those associated with stationary-phase growth at low temperatures [9]. In addition, we conducted an exhaustive analysis concerning the conservation of the regulon Crl among the Bacteria and Archaea genomes, using as a reference the knowledge gathered for *E. coli* K-12. From this analysis, Crl was identified in low copy numbers and constrained to the *Enterobacteriales* order, whereas the homologs of all regulated genes were found to be widely distributed beyond enterobacteria, suggesting that Crl was recruited in a secondary event to regulate a specific subset of genes for which the stimulation of Crl and σ^S is necessary.

References

1. Browning, D.F. and S.J. Busby, *The regulation of bacterial transcription initiation.* Nat Rev Microbiol, 2004. **2**(1): p. 57-65.
2. Landini, P., et al., *sigmaS, a major player in the response to environmental stresses in Escherichia coli: role, regulation and mechanisms of promoter recognition.* Environ Microbiol Rep, 2014. **6**(1): p. 1-13.

bioRxiv preprint doi: https://doi.org/10.1101/410621; this version posted September 7, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

3. Gama-Castro, S., et al., *RegulonDB version 9.0: high-level integration of gene regulation, coexpression, motif clustering and beyond.* Nucleic Acids Res, 2016. **44**(D1): p. D133-43.
4. Weber, H., et al., *Genome-wide analysis of the general stress response network in Escherichia coli: sigmaS-dependent genes, promoters, and sigma factor selectivity.* J Bacteriol, 2005. **187**(5): p. 1591-603.
5. Dufour, Y.S., P.J. Kiley, and T.J. Donohue, *Reconstruction of the core and extended regulons of global transcription factors.* PLoS Genet, 2010. **6**(7): p. e1001027.
6. Lozada-Chavez, I., S.C. Janga, and J. Collado-Vides, *Bacterial regulatory networks are extremely flexible in evolution.* Nucleic Acids Res, 2006. **34**(12): p. 3434-45.
7. Monsieurs, P., et al., *Comparison of the PhoPQ regulon in Escherichia coli and Salmonella typhimurium.* J Mol Evol, 2005. **60**(4): p. 462-74.
8. Liu, R. and H. Ochman, *Origins of flagellar gene operons and secondary flagellar systems.* J Bacteriol, 2007. **189**(19): p. 7098-104.
9. Bougdour, A., C. Lelong, and J. Geiselmann, *Crl, a low temperature-induced protein in Escherichia coli that binds directly to the stationary phase sigma subunit of RNA polymerase.* J Biol Chem, 2004. **279**(19): p. 19540-50.
10. Arnqvist, A., et al., *The Crl protein activates cryptic genes for curli formation and fibronectin binding in Escherichia coli HB101.* Mol Microbiol, 1992. **6**(17): p. 2443-52.
11. Typas, A., et al., *Stationary phase reorganisation of the Escherichia coli transcription machinery by Crl protein, a fine-tuner of sigmas activity and levels.* Embo j, 2007. **26**(6): p. 1569-78.
12. Lelong, C., et al., *The Crl-RpoS regulon of Escherichia coli.* Mol Cell Proteomics, 2007. **6**(4): p. 648-59.
13. Lelong, C., et al., *Mutual regulation of Crl and Fur in Escherichia coli W3110.* Mol Cell Proteomics, 2007. **6**(4): p. 660-8.
14. Dudin, O., S. Lacour, and J. Geiselmann, *Expression dynamics of RpoS/Crl-dependent genes in Escherichia coli.* Res Microbiol, 2013. **164**(8): p. 838-47.
15. Geer, L.Y., et al., *The NCBI BioSystems database.* Nucleic Acids Res, 2010. **38**(Database issue): p. D492-6.
16. Keseler, I.M., et al., *The EcoCyc database: reflecting new knowledge about Escherichia coli K-12.* Nucleic Acids Res, 2017. **45**(D1): p. D543-D550.
17. Shannon, P., et al., *Cytoscape: a software environment for integrated models of biomolecular interaction networks.* Genome Res, 2003. **13**(11): p. 2498-504.
18. Serres, M.H. and M. Riley, *MultiFun, a multifunctional classification scheme for Escherichia coli K-12 gene products.* Microb Comp Genomics, 2000. **5**(4): p. 205-22.
19. Ashburner, M., et al., *Gene ontology: tool for the unification of biology. The Gene Ontology Consortium.* Nat Genet, 2000. **25**(1): p. 25-9.
20. Thomas, P.D., et al., *PANTHER: a library of protein families and subfamilies indexed by function.* Genome Res, 2003. **13**(9): p. 2129-41.
21. Kanehisa, M., et al., *KEGG: new perspectives on genomes, pathways, diseases and drugs.* Nucleic Acids Res, 2017. **45**(D1): p. D353-D361.
22. Singh, H. and G.P. Raghava, *BLAST-based structural annotation of protein residues using Protein Data Bank.* Biol Direct, 2016. **11**(1): p. 4.

bioRxiv preprint doi: https://doi.org/10.1101/410621; this version posted September 7, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

23. Edgar, R.C., *MUSCLE: multiple sequence alignment with high accuracy and high throughput.* Nucleic Acids Res, 2004. **32**(5): p. 1792-7.
24. Tamura, K., et al., *MEGA6: Molecular Evolutionary Genetics Analysis version 6.0.* Mol Biol Evol, 2013. **30**(12): p. 2725-9.
25. Fitch, W.M., *Distinguishing homologous from analogous proteins.* Syst Zool, 1970. **19**(2): p. 99-113.
26. Janga, S.C. and G. Moreno-Hagelsieb, *Conservation of adjacency as evidence of paralogous operons.* Nucleic Acids Res, 2004. **32**(18): p. 5392-7.
27. Olsen, A., et al., *The RpoS sigma factor relieves H-NS-mediated transcriptional repression of csgA, the subunit gene of fibronectin-binding curli in Escherichia coli.* Mol Microbiol, 1993. **7**(4): p. 523-36.
28. Pratt, L.A. and T.J. Silhavy, *The response regulator SprE controls the stability of RpoS.* Proc Natl Acad Sci U S A, 1996. **93**(6): p. 2488-92.
29. Schnetz, K., *Silencing of the Escherichia coli bgl operon by RpoS requires Crl.* Microbiology, 2002. **148**(Pt 8): p. 2573-8.
30. Savageau, M.A., *Demand theory of gene regulation. I. Quantitative development of the theory.* Genetics, 1998. **149**(4): p. 1665-76.
31. Monteil, V., et al., *Identification of conserved amino acid residues of the Salmonella sigmaS chaperone Crl involved in Crl-sigmaS interactions.* J Bacteriol, 2010. **192**(4): p. 1075-87.
32. Dandekar, T., et al., *Pathway alignment: application to the comparative analysis of glycolytic enzymes.* Biochem J, 1999. **343 Pt 1**: p. 115-24.
33. Chen, G., C.L. Patten, and H.E. Schellhorn, *Positive selection for loss of RpoS function in Escherichia coli.* Mutat Res, 2004. **554**(1-2): p. 193-203.
34. Dong, T. and H.E. Schellhorn, *Control of RpoS in global gene expression of Escherichia coli in minimal media.* Mol Genet Genomics, 2009. **281**(1): p. 19-33.
35. Zambrano, M.M., et al., *Microbial competition: Escherichia coli mutants that take over stationary phase cultures.* Science, 1993. **259**(5102): p. 1757-60.
36. Cavaliere, P. and F. Norel, *Recent advances in the characterization of Crl, the unconventional activator of the stress sigma factor sigmaS/RpoS.* Biomol Concepts, 2016. **7**(3): p. 197-204.
37. Pratt, L.A. and T.J. Silhavy, *Crl stimulates RpoS activity during stationary phase.* Mol Microbiol, 1998. **29**(5): p. 1225-36.
38. Hao, Y., et al., *Protection against deleterious nitrogen compounds: role of sigmaS-dependent small RNAs encoded adjacent to sdiA.* Nucleic Acids Res, 2016. **44**(14): p. 6935-48.

Footnotes

bioRxiv preprint doi: https://doi.org/10.1101/410621; this version posted September 7, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

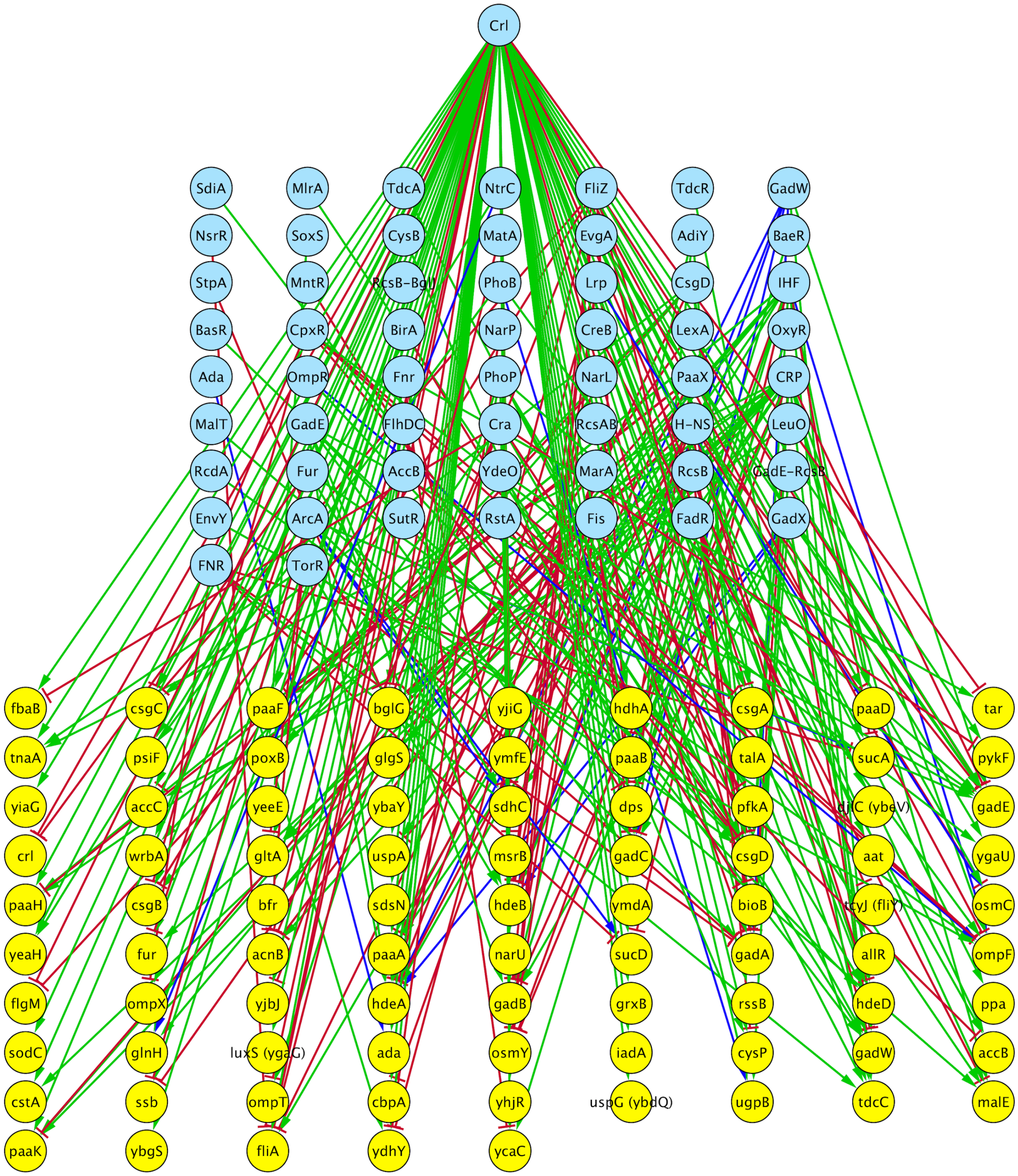

Figure 1. Crl regulatory network in *E. coli* K-12. Crl is shown in the upper part in the light blue circle. TF are shown in the middle of the network in light blue circles, and genes under Crl control are shown in yellow circles at the bottom. The effects of

bioRxiv preprint doi: https://doi.org/10.1101/410621; this version posted September 7, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

both Crl and TFs are shown as green solid lines for activation and red solid lines for repression.

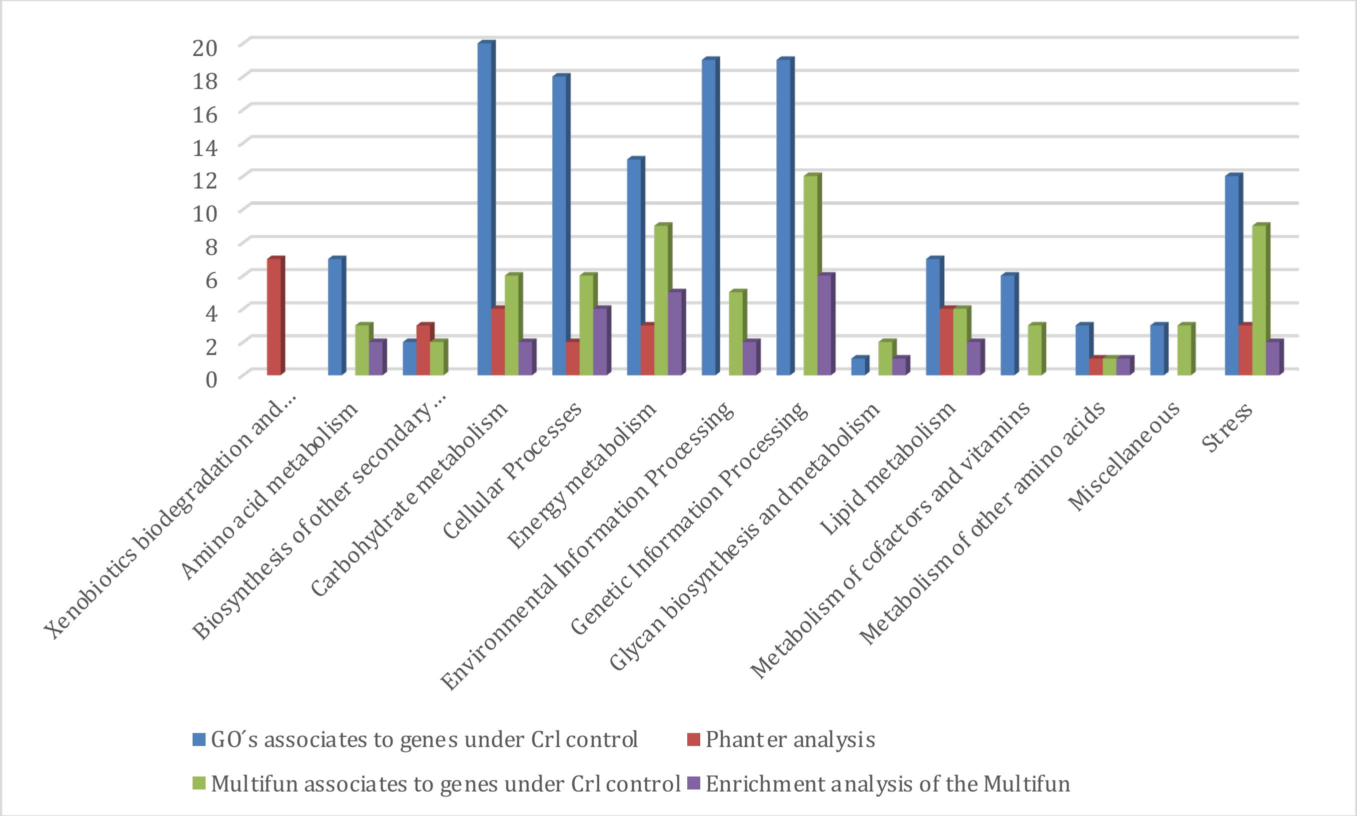

Figure 2. Gos and Multifun-associated genes under Crl control and enrichment analysis with the PANTHER classification system and Multifun. Categories of KEGG used to classify GOs and Multifun terms are shown on the X-axis, and the number of GOs associated with each category are shown on the Y-axis.

bioRxiv preprint doi: https://doi.org/10.1101/410621; this version posted September 7, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

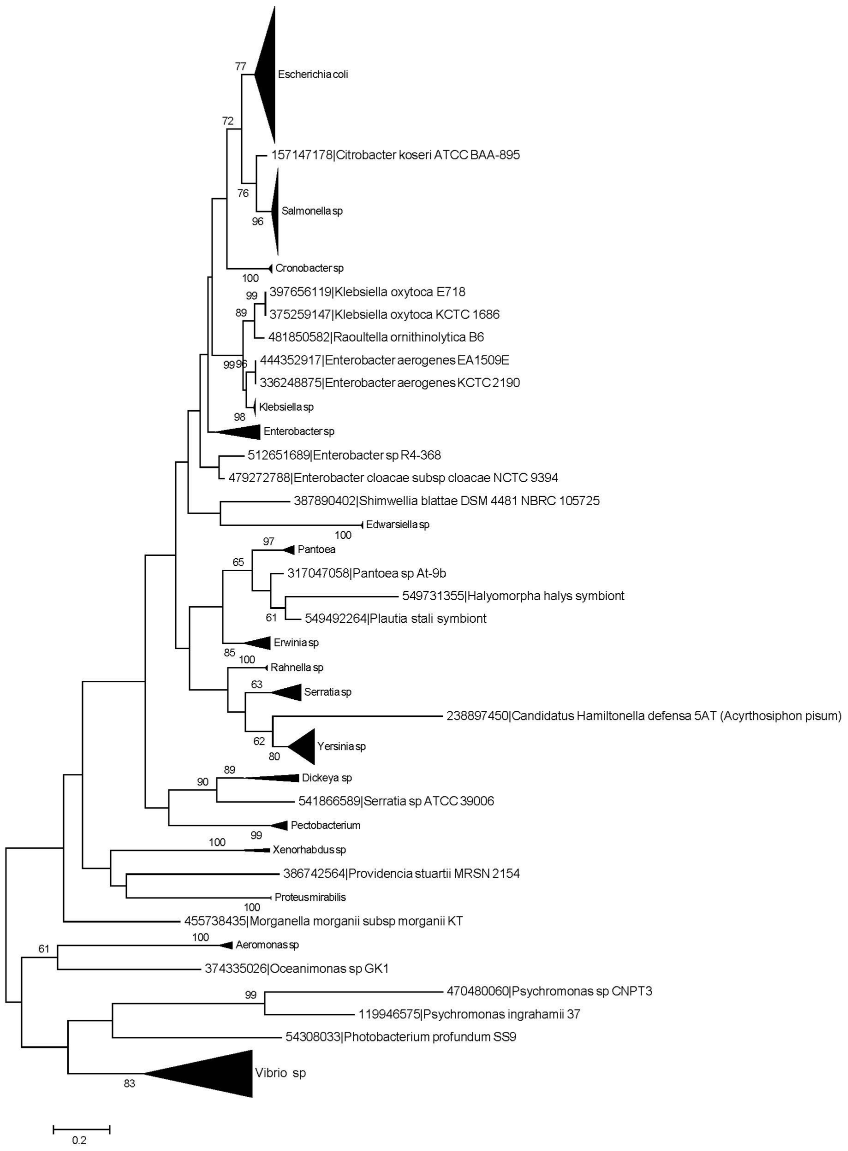
77
Escherichia coli
72
157147178|Citrobacter koseri ATCC BAA-895
76
Salmonella sp
96
Cronobacter sp
100
99
397656119|Klebsiella oxytoca E718
89
375259147|Klebsiella oxytoca KCTC 1686
481850582|Raoultella ornithinolytica B6
9996
444352917|Enterobacter aerogenes EA1509E
336248875|Enterobacter aerogenes KCTC 2190
Klebsiella sp
98
Enterobacter sp
512651689|Enterobacter sp R4-368
479272788|Enterobacter cloacae subsp cloacae NCTC 9394
387890402|Shimwellia blattae DSM 4481 NBRC 105725
Edwarsiella sp
100
97
Pantoea
65
317047058|Pantoea sp At-9b
549731355|Halyomorpha halys symbiont
61
549492264|Plautia stali symbiont
Erwinia sp
85
100
Rahnella sp
63
Serratia sp
238897450|Candidatus Hamiltonella defensa 5AT (Acyrthosiphon pisum)
62
Yersinia sp
80
89
90
Dickeya sp
541866589|Serratia sp ATCC 39006
Pectobacterium
99
100
Xenorhabdus sp
386742564|Providencia stuartii MRSN 2154
Proteus mirabilis
100
455738435|Morganella morganii subsp morganii KT
100
61
Aeromonas sp
374335026|Oceanimonas sp GK1
99
470480060|Psychromonas sp CNPT3
119946575|Psychromonas ingrahamii 37
54308033|Photobacterium profundum SS9
Vibrio sp
83
0.2

Figure 3. Phylogenetic tree based on Crl of *E. coli* and homologs in other organisms generated via maximum likelihood analysis, with 1,000 replicates. Species with bootstrap values higher than 60% are displayed. The black triangles to the right of the branches indicate multiple species for those genera.

bioRxiv preprint doi: https://doi.org/10.1101/410621; this version posted September 7, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

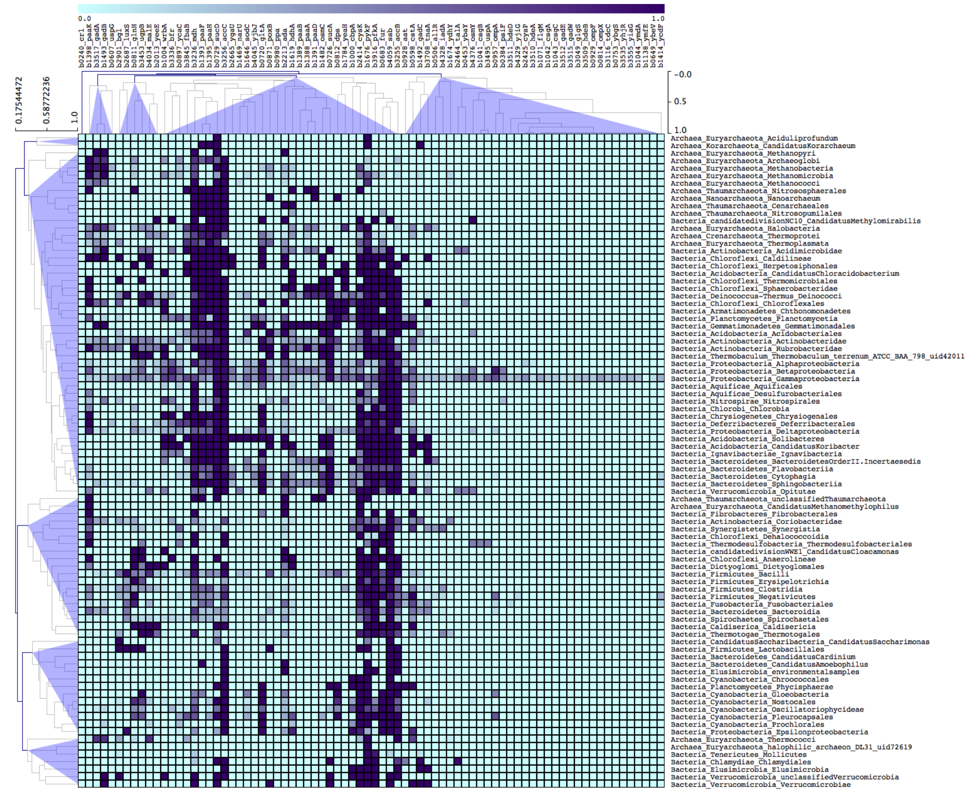

Figure 4. Clustering of orthologues from the perspective of *E. coli* K-12. A single linkage-clustering algorithm with no leaf order optimization was applied with Pearson distance as the similarity measure. The display clustering results were obtained using the MeV program (http://www.tm4.org/mev/). The conserved groups across the different taxonomic groups are indicated. Each column denotes Crl-regulated genes, whereas rows denote taxonomic groups. The bar at the top of the figure indicates the relative abundance of orthologs per group, represented as a percentage, where a value of 1 corresponds to 100% presence and 0% indicates a division without any ortholog of the Crl regulon in the taxonomic group.

Table 1. Genes regulated by Crl, TUs to which they belong (in red are possible candidates regulated by Crl, since they are controlled by Crl and σ^S, but they did

bioRxiv preprint doi: https://doi.org/10.1101/410621; this version posted September 7, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

not have a change of expression in the data we evaluated), TFs regulating the TU, the effect of Crl, evidences, references, and associated GO terms. Experimental evidence types supporting regulation by Crl: APPH = assay of protein purified to homogeneity; GEA = gene expression analysis, transcriptional fusions (*lacZ*), MSI = mapping of signal intensities, such as RNA-seq or microarray analysis; IMP = inferred from mutant phenotype (such as a mutation of a TF that has a visible cell phenotype and it is inferred that the regulator might be regulating the genes responsible for the phenotype). Growth conditions were 30ºC, as the stationary phase was induced for all experiments. All experiments were done with *E. coli* K-12 or derivative strains. All this information can be found in RegulonDB.

Gene	**TU(s)**	**TFs**	**Effect of Crl**	**Evidence**	**Reference(s)**	**GO Terms**
aat	*aat*		+	GEA and IMP	[12]	protein catabolic process, ubiquitin-dependent protein catabolic process via the N-end rule pathway
accB	*accBC*	AccB (-), FadR(+)	+	GEA and IMP	[12]	lipid metabolic process, fatty acid metabolic process, fatty acid biosynthetic process
accC	*accBC*	AccB (-), FadR(+)	+	GEA and IMP	[12]	lipid metabolic process, fatty acid metabolic process, fatty acid biosynthetic process, metabolic process, negative regulation of fatty acid biosynthetic process, malonyl-CoA biosynthetic process
acnB	*acnB*	CRP(+) ArcA(-), Cra(-), Fis (-)	-	IMP	[14]	regulation of translation, propionate catabolic process, 2-methylcitrate cycle, glyoxylate cycle, tricarboxylic acid cycle metabolic process
ada	*ada-alkB*	Ada(+/-)	+	GEA	[14]	DNA dealkylation involved in DNA repair, regulation of transcription, cellular response to DNA damage stimulus, metabolic process, methylation, DNA demethylation
allR	*allR*		+	MSI	[11]	regulation of transcription, cellular response to DNA damage stimulus, negative regulation of transcription
bfr	*bfd-bfr*		+	MSI, IMP	[4]^{MSI}, [12]^{IMP}	iron ion transport, cellular iron ion homeostasis, intracellular sequestering of iron ion, oxidation-reduction process
bglG	*bglG* *bglGFB*	CRP (+), Fis (-), H-NS (-), LeuO	-	MSI, IMP	[29]	regulation of transcription, positive regulation of

bioRxiv preprint doi: https://doi.org/10.1101/410621; this version posted September 7, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

		(+), RcsB-BglJ (+), StpA (-)				transcription
bioB	*bioBFCD*	BirA (-)	-	IMP	[14]	biotin biosynthetic process
cbpA	*cbpAM*	Fis (-)	+	GEA	[14]	protein folding
crl	*crl*	Fur (-)	-	MSI, IMP	[13]	regulation of transcription, DNA-templated, cellular protein complex assembly, positive regulation of transcription
csgA	*csgBAC*	CpxR (-), CsgD (+), FliZ (-)	+	APPH, MSI, IMP, GEA, IMP	[9]$^{APPH, MSI, IMP}$, [10]$^{GEA, IMP}$	cell adhesion, single-species biofilm formation, amyloid fibril formation
csgB	*csgBAC*	CpxR(-), CsgD(+), FliZ(-)	+	APPH, MSI, IMP, GEA, IMP	[9]$^{APPH, MSI, IMP}$, [10]$^{GEA, IMP}$	cell adhesion, single-species biofilm formation, amyloid fibril formation
csgC	*csgBAC*	CpxR (-), CsgD (+), FliZ (-)	+	MSI	[11]	
csgD	*csgDEFG*	BasR (+), Cra (+), CRP (+), CsgD (+), IHF (+), MlrA (+), OmpR (+), RcdA (+), CpxR(-), FliZ (-), RcsAB (-), RstA (-)	+	IMP	[14]	regulation of single-species biofilm formation
cstA	*cstA*	CRP (+)	+	GEA, IMP	[12]	cellular response to starvation
cysP	*cysPUWAM*	CysB (+), H-NS (-)	+	MSI	[11]	sulfur compound metabolic process, transport, sulfate transport, sulfate transmembrane transport
djlC (ybeV)	*ybeU-djlC*		+	MSI	[11]	positive regulation of ATPase activity
dps	*dps*	Fis(-), H-NS(-) ,IHF(+), MntR(-), OxyR(+)	+	GEA, IMP	[12]	cellular iron ion homeostasis, response to stress, chromosome condensation, response to starvation, oxidation-reduction process
fbaB	*fbaB*	Cra(-)	+	GEA, IMP	[12]	glycolytic process, transcription
flgM	*flgMN*, *flgAMN*	CsgD(-)	-	GEA, IMP	[14]	regulation of transcription, bacterial-type flagellum organization, negative regulation of proteolysis, negative regulation of transcription
fliA	*fliAZ-tcyJ*	H-NS(+), MatA(-), SutR(-), NsrR(-), CsgD(-), FlhDC(+)	-	IMP	[14]	transcription initiation from bacterial-type RNAP promoter, sporulation resulting in formation of a cellular spore
fur	*fur* *fldA-uof-fur* *uof-fur*	CRP(+), Fur(-)	+	MSI, IMP	[13]	regulation of transcription, negative regulation of transcription
gadA	*gadAX*	AdiY(+), ArcA(+), CRP(-), FNR(-), Fis(-), GadE-RcsB(+), GadW(+-), GadX(+), H-NS(-), RcsB(-), TorR(-)	+	MSI	[11]	glutamate metabolic process, carboxylic acid metabolic process, intracellular pH elevation
gadB	*gadBC*	AdiY(+) ,CRP(-), Fis(-), FliZ(-), GadE(+), GadW(+-), GadX(+), RcsB(+)	+	MSI, GEA	[4]^{MSI}, [14]^{GEA}	glutamate metabolic process, carboxylic acid metabolic process, intracellular pH elevation

bioRxiv preprint doi: https://doi.org/10.1101/410621; this version posted September 7, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

gadC	*gadBC*	AdiY(+), CRP(-), Fis(-), FliZ(-), GadE(+), GadW(+-), GadX(+), RcsB(+)	+	MSI	[11]	amino acid transmembrane transport, transport, amino acid transport, intracellular pH elevation
gadE	*gadE-mdtEF* *gadE*	ArcA(+), CRP(-), EvgA(+), FliZ(-), GadE(+), GadW(+), GadX(+), H-NS(-), PhoP(+), YdeO(+)	+	MSI	[11]	regulation of transcription
gadW	*gadW*	GadW (+), GadX (-), H-NS(-), PhoP (+), SdiA (+), YdeO (+)	+	MSI	[11]	regulation of transcription, cellular response to DNA damage stimulus
glgS	*glgS*	CRP(+)	+	GEA	[14]	glycogen biosynthetic process, positive regulation of cellular carbohydrate metabolic process, negative regulation of single-species biofilm formation on inanimate substrate, negative regulation of bacterial-type flagellum-dependent cell motility
glnH	*glnHPQ*	IHF(+), NtrC(+/-)	+	GEA, IMP	[12]	transport, amino acid transport
gltA	*gltA*	ArcA(-), CRP(+), IHF(+)	+	GEA, IMP	[12]	tricarboxylic acid cycle, metabolic process, cellular carbohydrate metabolic process
grxB	*grxB*		+	GEA, IMP	[12]	cell redox homeostasis, oxidation-reduction process
hdeA	*hdeAB-yhiD*	FliZ(-), GadE(+), GadW(+/-), GadX(+/-), H-NS(-), Lrp(-), MarA(-), PhoP(+), RcsB(+), TorR(+)	+	MSI, GEA	[4]^{MSI}, [14]^{GEA}	cellular response to stress, cellular response to acidic pH
hdeB	*hdeAB-yhiD*	FliZ(-), GadE(+), GadW(+/-), GadX(+/-), H-NS(-), Lrp(-), MarA(-), PhoP(+), RcsB(+), TorR(+)	+	MSI	[11]	response to pH change, cellular response to stress
hdeD	*hdeD*	GadE(+), GadX(+), H-NS(-), PhoP(+) ,RcsB(+)	+	MSI	[11]	response to pH change
hdhA	*hdhA*		+	GEA, IMP	[12]	lipid metabolic process, metabolic process, steroid metabolic process, lipid catabolic process, bile acid, catabolic process, protein homotetramerization, oxidation-reduction process
iadA	*yjiHG-iadA*		+	MSI	[11]	proteolysis
luxS *(ygaG)*	*luxS*		+	MSI, GEA, IMP	[11]^{MSI}, [12]$^{GEA, IMP}$	cell-cell signaling involved in quorum sensing, L-methionine biosynthetic process from S-adenosylmethionine, quorum sensing

bioRxiv preprint doi: https://doi.org/10.1101/410621; this version posted September 7, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

malE	*malEFG*	CRP(+), CreB(-), Fis(+), MalT(+)	+	GEA, IMP	[12]	cellular response to DNA damage stimulus, carbohydrate transport, maltose transport, detection of maltose stimulus, maltodextrin transport, cell chemotaxis
msrB	*msrB*		-	IMP	[14]	protein repair, response to oxidative stress
narU	*narU*		+	MSI	[11]	nitrate transport, nitrite transport, nitrate assimilation
ompF	*ompF*	CRP(+), CpxR(-), EnvY(+), Fur(+), IHF(+/-), OmpR(+/-), PhoB(+), RstA(-)	-	GEA, IMP	[37]	transport, ion transport, ion transport, drug transmembrane transport, bacteriocin transport
ompT	*ompT* *envY-ompT*	PhoP(+)	-	IMP	[14]	proteolysis
ompX	*ompX*	FNR(-)	+	GEA, IMP	[12]	
osmC	*osmC*	H-NS(-), Lrp(+/-)	+	GEA, IMP	[12]	hyperosmotic response, response to oxidative stress, response to hydroperoxide, oxidation-reduction process
osmY	*osmY*	CRP(-), Fis(-), FliZ(-), IHF(-), Lrp(-)	+	GEA, IMP	[12]	response to osmotic stress
paaA	*paaABCDEFGHIJK*	CRP(+), IHF(+), PaaX(-)	+	MSI	[11]	phenylacetate catabolic process, oxidation-reduction process
paaB	*paaABCDEFGHIJK*	CRP (+), IHF (+), PaaX (-)	+	MSI	[11]	phenylacetate catabolic process
paaD	*paaABCDEFGHIJK*	CRP (+), IHF (+), PaaX (-)	+	MSI	[11]	phenylacetate catabolic process
paaF	*paaABCDEFGHIJK*	CRP (+), IHF (+), PaaX (-)	+	MSI	[11]	lipid metabolic process, fatty acid metabolic process, phenylacetate catabolic process
paaH	*paaABCDEFGHIJK*	CRP(+), IHF(+), PaaX(-)	+	MSI	[11]	fatty acid metabolic process, phenylacetate catabolic process, oxidation-reduction process
paaK	*paaABCDEFGHIJK*	CRP(+), IHF(+), PaaX(-)	+	MSI	[11]	metabolic process, phenylacetate catabolic process
pfkA	*pfkA*	Cra(-)	+	GEA, IMP	[12]	fructose 6-phosphate metabolic process, glycolytic process
poxB	*poxB,* *poxB-ltaE-ybjT*	Cra(+), MarA(+), SoxS (+)	+	GEA, IMP	[12]	pyruvate metabolic process, oxidation-reduction process
ppa	*ppa*		+	GEA, IMP	[12]	phosphate-containing compound metabolic process
psiF	*phoA-psiF*	PhoB(+)	+	MSI	[11]	
pykF	*pykF*	Cra(-)	+	GEA, IMP	[12]	glycolytic process, metabolic process, response to heat, phosphorylation
rssB	*rssB*		+	IMP	[11]	protein destabilization, positive regulation of proteolysis, regulation of nucleic acid-templated transcription (phosphorelay signal transduction system)
sdhC	*sdhCDAB-sucABCD*	CRP(+), Fur(+), ArcA(+/-), Fnr(-)	-	IMP	[14]	aerobic respiration. cytochrome complex assembly, tricarboxylic acid cycle, oxidation-reduction process
sdsN	*sdsN*		+	GEA	[38]	small RNA
sodC	*sodC*		+	GEA	[14]	superoxide metabolic process, removal of superoxide radicals, oxidation-reduction process
ssb	*ssb*	ArcA(-), LexA(-)	+	GEA, IMP	[12]	recombinational repair, DNA replication, cellular response to DNA damage stimulus, SOS

bioRxiv preprint doi: https://doi.org/10.1101/410621; this version posted September 7, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

						response
sucA	*sucAB* *sucABCD*	ArcA(+/-), FNR(-), IHF(-)	+	GEA, IMP	[12]	glycolytic process, tricarboxylic acid cycle, metabolic process, oxidation-reduction process
sucD	*sucAB* *sucABCD*	ArcA(+/-),FNR(-),IHF(-)	+	GEA, IMP	[12]	tricarboxylic acid cycle, metabolic process, protein autophosphorylation
talA	*talA-tktB*	CreB(+)	+	GEA, IMP	[12]	carbohydrate metabolic process, pentose-phosphate shunt
tar	*tar-tap-cheRBYZ*	Fnr(+)	-	IMP	[14]	chemotaxis, signal transduction
tcyJ *(fliY)*	*tcyJ* *fliAZ-tcyJ*	H-NS(+), MatA(-), SutR(-), NsrR(-), CsgD(-), FlhDC(+)	-	IMP	[14]	L-cystine transport
tdcC	*tdcABCDEFG,* *tdcBCDEFG*	CRP(+), FNR(+), IHF(+), TdcA(+), TdcR (+)	+	MSI	[11]	L-serine transport, threonine transport, proton transport, serine transport
tnaA	*tnaCAB*	CRP(+), TorR (+)	+	GEA, IMP	[12]	cellular amino acid metabolic process, aromatic amino acid family metabolic process
ugpB	*ugpBAECQ*	CRP(+), PhoB(+/-)	+	GEA, IMP	[12]	glycerophosphodiester transport, transport, glycerol-3-phosphate transport
uspA	*uspA*	FadR(-), IHF(+)	+	GEA, IMP	[12]	response to stress
uspG *(ybdQ)*	*uspG*		+	GEA, IMP	[12]	response to stress, protein adenylylation, protein autophosphorylation, nucleotide phosphorylation, regulation of cell motility
wrbA	*wrbA-yccJ*	CsgD(+)	+	MSI, GEA, IMP	[14]$^{\text{MSI}}$, [12]$_{\text{GEA, IMP}}$	response to oxidative stress, negative regulation of transcription
ybaY	*ybaY*		+	MSI	[11]	
ybgS	*ybgS*		+	MSI	[11]	
ycaC	*ycaC*	BaeR(+), Fnr(-)	+	MSI, GEA, IMP	[11]$^{\text{MSI}}$, [12]$_{\text{GEA, IMP}}$	metabolic process
ydhY	*ydhYVWXUT*	FNR (+), NarL (-), NarP (-)	+	MSI	[11]	oxidation-reduction process
yeaH	*yeaGH*	NtrC (+)	+	MSI	[11]	
yeeE	*yeeED*		+	MSI	[11]	
ygaU	*ygaU*	CpxR (+)	+	GEA, IMP	[12]	
yhjR	*yhjR*		+	MSI	[11]	bacterial cellulose biosynthetic process
yiaG	*yiaG*		+	MSI	[11]	regulation of transcription
yjbJ	*yjbJ*	FliZ (-)	+	MSI	[11]	
yjiG	*yjiHG-iadA*		+	MSI	[11]	
ymdA	*ymdA*		+	MSI	[11]	
ymfE	*ymfED*		+	MSI	[11]	