UNIVERSIDAD NACIONAL AUTÓNOMA DE MÉXICO

DOCTORADO EN CIENCIAS BIOMÉDICAS
CENTRO DE CIENCIAS GENÓMICAS

EFECTO DE LA VECINDAD GÉNOMICA EN LA COEXPRESIÓN DE GENES CORREGULADOS EN *Escherichia coli* K-12

TESIS

QUE PARA OPTAR POR EL GRADO DE:

DOCTORA EN CIENCIAS

PRESENTA:

LUCIA PANNIER

TUTOR PRINCIPAL

DR. JULIO COLLADO VIDES
CENTRO DE CIENCIAS GENÓMICAS

MIEMBROS DEL COMITÉ TUTOR

DR. ENRIQUE MERINO
INSTITUTO DE BIOTECNOLOGÍA

DRA. KATHLEEN MARCHAL
DEPARTMENT OF PLANT BIOTECHNOLOGY AND BIOINFORMATICS, GHENT, BÉLGICA

CUERNAVACA, MORELOS. Junio, 2017

UNAM – Dirección General de Bibliotecas Tesis Digitales Restricciones de uso

DEDICATORIA

Deze thesis is opgedragen aan mijn ouders,

voor hun liefde en steun door dik en dun

AGRADECIMIENTOS

Gracias a Julio Collado he crecido mucho como persona y como científica, disfruté y aprendí de su devoción a la ciencia, sus ganas de vivir y aprender, su positivismo, su pasión e infinita fuente de energía, sus ideas geniales y su inteligencia impresionante. Aprecio mucho sus enseñanzas y amistad.

A Enrique Merino que con su ayuda en mi proyecto me enseñó, con paciencia, cosas importantes en la investigación científica y que con su apoyo incondicional, consejos y ánimos, tuvo un rol muy importante en mi doctorado.

A Kathleen Marchal, aprendí mucho de su ojo crítico y su actitud de trabajo pragmatica. Ella mejoró considerablemente mi trabajo y tuvo un rol muy importante en el éxito de mi doctorado.

~

Aan Kathleen Marchal, ik heb veel geleerd van haar kritisch oog en haar pragmatische werkhouding. Ze heeft mijn werk substantieel verbeterd en had een doorslaggevende rol in het slagen van mijn doctoraat.

A Susana Brom, por ayudarme en los períodos dificiles, estoy muy agradecida por su paciencia y apoyo.

A todas las personas en el laboratorio del Programa de Genómica Computacional, que han brindado ayuda y sonrisas todos los días, creando un ambiente amigable y positivo. Un mejor entorno de trabajo no me hubiera podido pedir!

A CONACYT, al Centro de Ciencias Genómicas de la UNAM y al Dr. David Romero por los apoyos otorgados.

A México, un país mágico con gente bella, estoy muy feliz de haber llegado.

ÍNDICE

CAPÍTULO 1. INTRODUCCIÓN

En células bacterianas, el control de la actividad de los genes depende en gran parte de proteínas llamadas Factores de la Transcripción (FTs) que pueden activar o reprimir la expresión de sus genes blancos (Lewin 2008). Se dice que genes blancos de un cierto FT son *corregulados* por ese FT, por ejemplo, araA, araB, adaD, araJ, etc son corregulados por AraC. Genes corregulados tienden a tener perfiles de expresión similares a través de multiples condiciones experimentales, i.e. se tienden a *coexpresar* (Lemmens et al. 2009).

El *contexto genómico* de genes, siendo donde los genes están localizados el uno relativo al otro en el cromosoma, es importante tanto para la corregulación que para la coexpresión. Varias observaciones demuestran que existe una correlación entre 1) corregulación y cercanía genómica, y 2) coexpresión y cercanía genómica de genes. Primero, se ha observado que varios genes que son corregulados están localizados en cercanía, tanto en procariontes (Janga et al. 2009) como en eucariontes (Schneider & Grosschedl 2007). Segundo, se ha encontrado que genes ubicados en cercanía son más altamente coexpresados que genes lejanos en *E. coli* (Zampieri et al. 2008; Korbel et al. 2004).

También, varias observaciones demuestran que genes corregulados y situados en vecindad muchas veces son *muy* altamente coexpresados, como cuando hay cotranscripcón bidireccional en promotores divergentes (Beck & Warren 1988; Korbel et al. 2004; Rhee et al. 1999) o cuando los genes blancos de FTs con pocos genes blancos están agrupados en el genoma (Janga et al. 2009; Michoel et al. 2009; Zhang et al. 2012).

Estas observaciones sugieren que la cercanía genómica podría tener un efecto adicional, o "sinérgico" sobre el efecto de la corregulación en la coexpresión, es decir, que el efecto combinado sobre la coexpresión es más grande que la suma de los efectos independientes de la corregulación y la cercanía genómica.

Sin embargo, el efecto adicional de la distancia genómica entre genes corregulados en la coexpresión aún no se ha estudiado sistemáticamente.

En este trabajo, por lo tanto, evaluamos cómo la distancia genómica de los genes corregulados en *E. coli* influye en su coexpresión. Elegimos *E. coli* como organismo modelo dada la disponibilidad de datos abundantes de expresión y regulación transcripcional. Consideramos pares de genes como corregulados cuando son controlados por mínimo un FT común y con el mismo efecto (activador, represor o dual) tal como reportado en la base de datos RegulonDB (Gama-Castro et al. 2015). Se excluyen del estudio los genes dentro de un mismo operón para no confundir los análisis. Se estimó el nivel de coexpresión entre pares de genes corregulados por la similitud de los perfiles de expresión diferencial a través de todas los contrastes, tal como se encuentran en la base de datos de microarreglos COLOMBOS (Moretto et al. 2016).

Para medir la coexpresión propusimos una nueva métrica llamada la Spearman Correlation Rank (SCR), descrita en el primer artículo (sección 2.1). Usamos la SCR para dos fines; 1) en el presente análisis, *i.e.* para evaluar el impacto de la distancia genómica en la coexpresión de los genes corregulados, descrito en el primer artículo (sección 2.1), y 2) en una herramienta de análisis de coexpresión, implementada en la version 9.0 de RegulonDB y descrita en el segundo artículo (sección 2.2).

En general, hemos observado que los genes corregulados muestran gradualmente mayores grados de coexpresión si están más cercanos en el genoma. El efecto de la cercanía es obvio sobre todo en genes corregulados que también tienen FTs no comunes. Pudimos excluir la posibilidad de que este efecto se debiera a la cotranscripción divergente, transcripción *readthrough* o que fuera causado por la cercanía del gen codificante del FT. Tambien vimos que la tendencia de genes corregulados altamente coexpresados de estar localizados cerca es conservada a través de otras especies gammaproteobacterianas.

Nuestra hipótesis para explicar nuestras observaciones dice que la cercanía de genes corregulados aumenta la coexpresión de genes corregulados porque a cortas distancias hay una accessibilidad similar de proteínas FT en sus respectivos promotores.

En este trabajo demostramos que la distancia entre los genes y la corregulación no trabajan en forma aislada, sino que conjuntamente y en forma sinérgica influyen para controlar la coexpresión de genes.

CAPÍTULO 2. RESULTADOS

2.1 Primer artículo: Effect of genomic distance on coexpression of corregulated genes in *E.coli*

En este artículo se presentan los resultados del análisis genómico sobre la influencia de la distancia genómica entre los genes corregulados en su coexpresión.

PLOS ONE

RESEARCH ARTICLE

Effect of genomic distance on coexpression of coregulated genes in *E. coli*

Lucia Pannier[1], Enrique Merino[2], Kathleen Marchal[3,4,5,6]*, Julio Collado-Vides[1]*

1 Programa de Genómica Computacional, Centro de Ciencias Genómicas, Universidad Nacional Autónoma de México, Cuernavaca, Morelos, México, **2** Departamento de Microbiología Molecular, Instituto de Biotecnología, Universidad Nacional Autónoma de México, Cuernavaca, Morelos, México, **3** Department of Microbial and Molecular Systems, KU Leuven, Centre of Microbial and Plant Genetics, Leuven, Belgium, **4** Department of Plant Biotechnology and Bioinformatics, Ghent University, Technologiepark, Ghent, Belgium, **5** Department of Information Technology, Ghent University, IMinds, Ghent, Belgium, **6** Department of Genetics, University of Pretoria, Hatfield Campus, Pretoria, South Africa

* Kathleen.Marchal@intec.ugent.be (KM); collado@ccg.unam.mx (JCV)

Citation: Pannier L, Merino E, Marchal K, Collado-Vides J (2017) Effect of genomic distance on coexpression of coregulated genes in *E. coli*. PLoS ONE 12(4): e0174887. https://doi.org/10.1371/journal.pone.0174887

Editor: Akira Ishihama, Hosei University, JAPAN

Received: December 13, 2016

Accepted: March 16, 2017

Published: April 18, 2017

Data Availability Statement: Expression data are available at the COLOMBOS database ("download" tab at www.colombos.net). Transcriptional regulation data are available at the RegulonDB database (http://regulondb.ccg.unam.mx/menu/download/datasets/files/BindingSiteSet.txt). We confirm that future interested researchers will be able to gain access to data from COLOMBOS and RegulonDB databases in the same manner as the authors without any special privileges needed.

Funding: Lucia Pannier is a doctoral student from Programa de Doctorado en Ciencias Biomédicas (PDCB) in Centro de Ciencias Genomicas (CCG) of

Abstract

In prokaryotes, genomic distance is a feature that in addition to coregulation affects coexpression. Several observations, such as genomic clustering of highly coexpressed small regulons, support the idea that coexpression behavior of coregulated genes is affected by the distance between the coregulated genes. However, the specific contribution of distance in addition to coregulation in determining the degree of coexpression has not yet been studied systematically. In this work, we exploit the rich information in RegulonDB to study how the genomic distance between coregulated genes affects their degree of coexpression, measured by pairwise similarity of expression profiles obtained under a large number of conditions. We observed that, in general, coregulated genes display higher degrees of coexpression as they are more closely located on the genome. This contribution of genomic distance in determining the degree of coexpression was relatively small compared to the degree of coexpression that was determined by the tightness of the coregulation (degree of overlap of regulatory programs) but was shown to be evolutionary constrained. In addition, the distance effect was sufficient to guarantee coexpression of coregulated genes that are located at very short distances, irrespective of their tightness of coregulation. This is partly but definitely not always because the close distance is also the cause of the coregulation. In cases where it is not, we hypothesize that the effect of the distance on coexpression could be caused by the fact that coregulated genes closely located to each other are also relatively more equidistantly located from their common TF and therefore subject to more similar levels of TF molecules. The absolute genomic distance of the coregulated genes to their common TF-coding gene tends to be less important in determining the degree of coexpression. Our results pinpoint the importance of taking into account the combined effect of distance and coregulation when studying prokaryotic coexpression and transcriptional regulation.

Univerdidad Nacional Autónoma de México (UNAM) and received PhD fellowship (420430) from Consejo Nacional de Ciencia y Tecnología México (CONACyT) and was partially supported by the National Institutes of Health under grant number R01GM110597 and FOINS CONACyt Fronteras de la Ciencia under project number 15. Ghent University Multidisciplinary Research Partnership "Bioinformatics: from nucleotides to networks"; Fonds Wetenschappelijk Onderzoek-Vlaanderen (FWO) [G.0329.09, 3G042813, G.0A53.15N]; Agentschap voor Innovatie door Wetenschap en Technologie (IWT) [NEMOA]. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing interests: The authors have declared that no competing interests exist.

Abbreviations: AUC, Area Under the Curve; MI, Mutual Information; MIR, Mutual Information Rank; PCC, Pearson Correlation Coefficient; PCR, Pearson Correlation Rank; SCC, Spearman Correlation Coefficient; SCR, Spearman Correlation Rank; TF, Transcription Factor; TFBS, Transcription Factor Binding Site; TN, True Negative; TP, True Positive.

Introduction

Transcriptional coregulation in general implies coexpression: genes that are regulated by the same Transcription Factors (TFs) are more likely to be coexpressed. RegulonDB defines the transcriptional programs of genes in *E. coli* K-12 based on curated information. A distinction is often made between simple and complex transcriptional regulatory programs depending on whether a gene's regulatory program consists of at most one or more TFs. Genes are defined to be **coregulated** if their respective regulatory program overlaps, i.e. if they are coregulated by at least one TF with the same role (activator, repressor or dual). The complexity of their individual regulatory programs in combination with the extent to which their program overlaps defines **the tightness of the coregulation**. Genes with a completely identical regulatory program are expected to be more tightly coregulated under all conditions than in case of an incomplete overlap. In the latter case different gene-specific TFs can be involved in tuning the expression at the individual gene level (less tight coregulation). Also if more TFs are shared by the coregulated genes, their coregulation can be expected to be tighter.

Evidence exists that besides coregulation also the genomic distance between two genes contributes to their coexpression. Closely located genes are more coexpressed than faraway located genes in *E. coli* [1,2], yeast [3,4], *Arabidopsis* [5], zebrafish [6] and humans [5]. Several mechanisms supporting coexpression behavior of closeby located genes have been reported in prokaryotes, including operonic organization, bidirectional cotranscription at divergent promoters [2,7,8] and genomic clustering of highly coexpressed small regulons, i.e. of TFs such as GntR and GadW that only regulate a few operons [9–11]. These observations suggest that coregulation and genomic vicinity both can contribute to the degree to which two genes tend to be coexpressed. However, assessing the contribution of the genomic distance added to coregulation in determining coexpression is complicated as in many cases the close distance between genes is also at the basis of their mechanism of coregulation (genes located in the same operon, read-through transcription of contiguous operons [12], and bidirectional cotranscription at divergent promoters [2,7]). In this study we exploited the large body of information in RegulonDB together with publicly available expression data to systematically assess whether the genomic distance affects the degree of coexpression, independently of the coregulation mechanism.

We tested to what extent the distance between coregulated genes is associated with their degree of coexpression. Our results confirm that genomic vicinity of coregulated genes is an important factor that contributes to higher levels of coexpression, also for genes that are not tightly coregulated. This observation was further supported by the finding that there was an evolutionary constraint in maintaining the distance between coregulated genes that are highly coexpressed.

Results

Assessing the degree of coexpression between coregulated genes

In bacteria, genomic distance between genes is a feature that, in addition to coregulation, affects coexpression. In this study, we aimed at assessing whether and how genomic distance between coregulated genes associates with their degree of coexpression. The degree of coexpression between genes was assessed by calculating the pairwise similarity between their gene expression profiles obtained from a large scale expression compendium assessing expression under 4077 condition contrasts (Materials and methods) [13].

To identify the measure that best reflects the degree of pairwise coexpression between any pair of coregulated genes we tested six similarity measures based on respectively correlation

and mutual information (see Materials and methods and Supplementary file S1 File part 1). The measure referred to as Spearman Correlation Rank (SCR) performed best in separating the coexpression behavior of genes that were expected (genes within the same operon) to be highly coexpressed from those that were not (genes not known to be coregulated). In addition, we could show that this rank-based measure better normalized for the unequal number of samples present in the compendium that represent the conditions under which the different TFs are active, as explained in detail in the Supplementary File S1 File part 2.

In the remainder of the analysis the degree of coexpression between two genes is thus defined as the pairwise similarity between the expression profiles of these genes as measured by SCR. High SCR values between two genes correspond to a low degree of coexpression whereas low SCR values correspond to a high degree of coexpression.

To gain a first insight into the overall degree to which coregulated genes are coexpressed, we calculated their average degree of coexpression using SCR (Supplementary File S1 File part 3). In the context of this study, coregulated genes were defined as any set of two genes that have at least one common TF in their respective regulatory programs with the same regulatory effect on each of the considered genes (activation, repression or both). Whether two genes were coregulated was derived from curated information on TF-gene regulatory interactions in RegulonDB [14] (Materials and methods). We deliberately excluded pairs of coregulated genes originating from the same operon as for operonic transcription, coregulation and distance are confounded (i.e. the closeby location is the cause of the coregulation) and including these operonic coregulated genes would blur assessing the effect of the genomic distance between coregulated genes on their degree of coexpression.

We observed that on average, the degree of coexpression between genes known to be coregulated was rather low, as was also previously reported [15]. In particular, genes coregulated by a common *global* TF (here defined as a TF with more than 130 target genes), but not by any other additional common more specific TF, showed a relatively low degree of coexpression. Those coregulated genes that only have a global TF in the common part of their regulatory program were excluded from further analysis as they are known to be only loosely coregulated (Materials and methods) and including them results in underestimating the average degree of coexpression between coregulated genes. In Supplementary Table S1 Table we provided a full list of 91 TFs that together control 11339 pairs of coregulated genes considered in this study, as well as per TF the mean pairwise genomic distances and the mean degree of pairwise coexpression between the target genes coregulated by that TF.

Distance between coregulated genes inversely correlates with the mean degree of coexpression

We hypothesized that the distance between coregulated genes has an influence on their coexpression degree. To test this hypothesis, we examined the relationship between the pairwise genomic distance between coregulated genes and their degree of coexpression. The pairwise linear distance between genes along the circular chromosome, hereafter referred to as *distance*, was determined by the number of base pairs separating the start positions of two genes.

In Fig 1 the mean degree of coexpression is shown as a function of the distance between genes, i.e., the median SCR (y-axis) of a pair of genes for which the distance between the two genes is smaller than a given value (x-axis). The mean coexpression degrees between genes that were not known as coregulated was shown as a negative control (Fig 1, red curve).

Overall, we observed a clear influence of the distance on the degree of coexpression: coregulated genes tend to be pairwisely more coexpressed when they are closely located than when they are more distantly located (see Fig 1, slope of dark-blue curve). Also in the negative

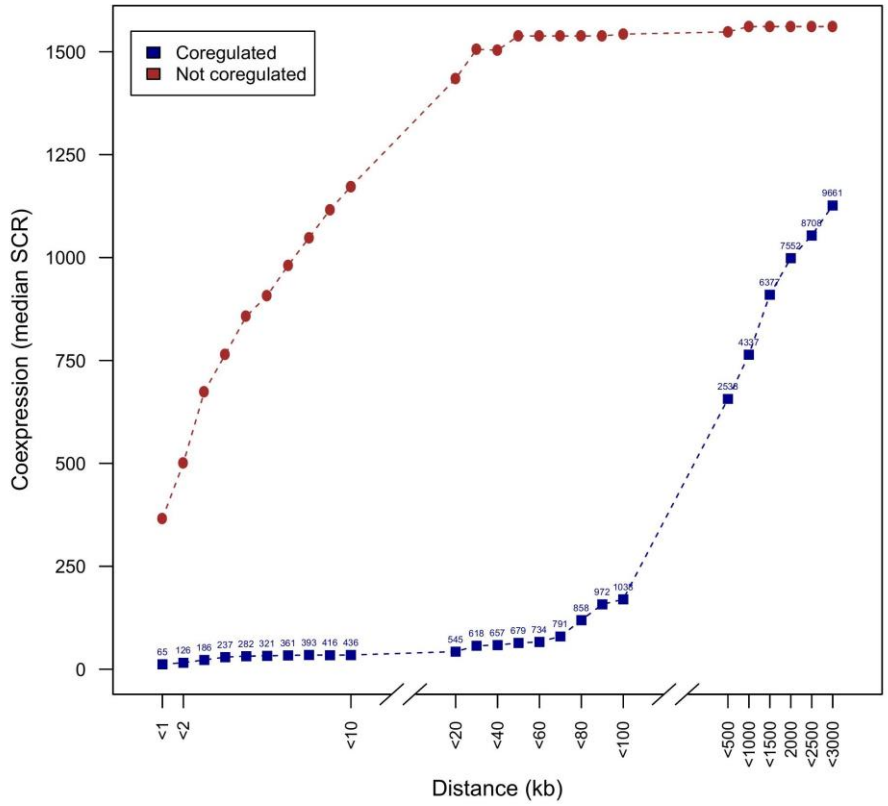

Fig 1. The distance between coregulated genes negatively influences their coexpression degree. The plot shows the mean coexpression degree as a function of the maximum distance between two genes. The distance (x-axis) is measured by the number of kb (kilo base pairs, equal to 1000 base pairs) between the structural gene start positions of two genes. Coexpression degree (y-axis) is measured by the median SCR (a low median SCR implies high degree of coexpression) of genes with a distance lower or equal to the distance indicated on the x-axis. The effect of distance on coexpression is shown for all coregulated genes (dark-blue curve). Coexpression degree of coregulated genes can be compared to the negative control containing all genes not known to be coregulated (red curve). Note that breaks in the x-axis between distances <10 and <20 kb and between distances <100 and <500 correspond to scale differences. The numbers above each data point of the dark-blue curve represent the number of pairs of coregulated genes for which the median SCR was calculated.

https://doi.org/10.1371/journal.pone.0174887.g001

control (genes not known to be coregulated) at small distances (see slope of red curve, distances <20 kb) a relative high degree of coexpression was observed. Because genomic clustering and coexpression tend to be associated [16], genomically colocalized genes might tend to be more coexpressed, irrespective of whether they are coregulated by the same TF. According to Sobetzko et al. [16], colocalization of genes tends to trigger some degree of coexpression because at close distances, levels of DNA supercoiling tend to be similar, hereby leading to

similar gene expression patterns. So genes that are clustered on the genome might therefore be coexpressed as a mere result of their closeby positioning rather than because of coregulation. To test whether this was indeed the case, we have identified genes that belong to distinct regulons (i.e. genes regulated by distinct TFs) that are genomically colocalized with each other (Materials and methods). We have compared the degree of coexpression of pairs of colocalized genes that were also coregulated versus the degree of coexpression of gene pairs that were colocalized but not coregulated. We still observed a significant difference in degree of coexpression between both gene classes (Kruskal-Wallis p-value $<< 0.001$), indicating that genomic colocalization alone most likely cannot be responsible for the high degrees of coexpression observed for some gene pairs in the (negative) reference set.

It thus is more likely that the relatively high degree of coexpression in the negative control at small genomic distances is the result of the incompleteness of the information in RegulonDB rather than being the consequence of the small distance: because of missing information in RegulonDB, we cannot exclude that a minor fraction of these so-called non-coregulated gene pairs are in fact coregulated. Further analysis (data not shown) indeed showed that the observed relatively high average coexpression degree of non-coregulated genes at small distances visible in Fig 1 could be attributed to a small fraction of the non-coregulated genes showing high degrees of coexpression but that the majority of the non-coregulated genes were not highly coexpressed. An additional overlay of the set of genes reported to be non-coregulated and having high degrees of coexpression with sets of genes that were predicted to be in vitro coregulated based on SELEX results [17] confirmed that indeed several of the so-called non-coregulated genes with high degree of coexpression might actually be coregulated (listed in Supplementary Table S2 Table).

The effect of the distance on coexpression decreases as the tightness of the coregulation increases

To assess whether the effect of the distance in determining the degree of coexpression was dependent on the coregulation tightness, we first subdivided coregulated genes in two groups depending on whether **their regulatory programs overlapped completely versus partially**: if two coregulated genes have a completely overlapping regulatory program they are assumed to be more tightly coregulated than when their regulatory programs are only partially overlapping. A partial overlap means that at least one of the coregulated genes has TFs in its regulatory program that are not shared by the other gene or when the same TF has different effects on each gene. Indeed, as shown in Fig 2 the degree of coexpression between coregulated genes with complete overlap of regulatory programs is higher than that of coregulated genes with only a partial overlap of regulatory programs and that this is true over all distances considered (blue versus orange curve). Regarding the effect of distance on the degree of coexpression, this effect exists for both genes that have completely overlapping versus those that have only a partially overlapping regulatory program. However the distance effect is most pronounced for genes that have a partially overlapping program but lasts at larger distances for genes with a completely overlapping regulatory program (respectively around $<$20 kb versus around $<$100 kb).

In addition, we made a distinction between genes **that are coregulated by one versus those that are coregulated by more than one TF**, as we assume that coregulation by multiple common TFs can also contribute to a larger coregulation tightness with a possible effect on the degree of coexpression [18]. The effect of the coregulation tightness determined by the number of common TFs in the overlapping part of the regulatory programs is confounded with the degree to which the regulatory program overlaps (e.g. it is hard to compare the degree of

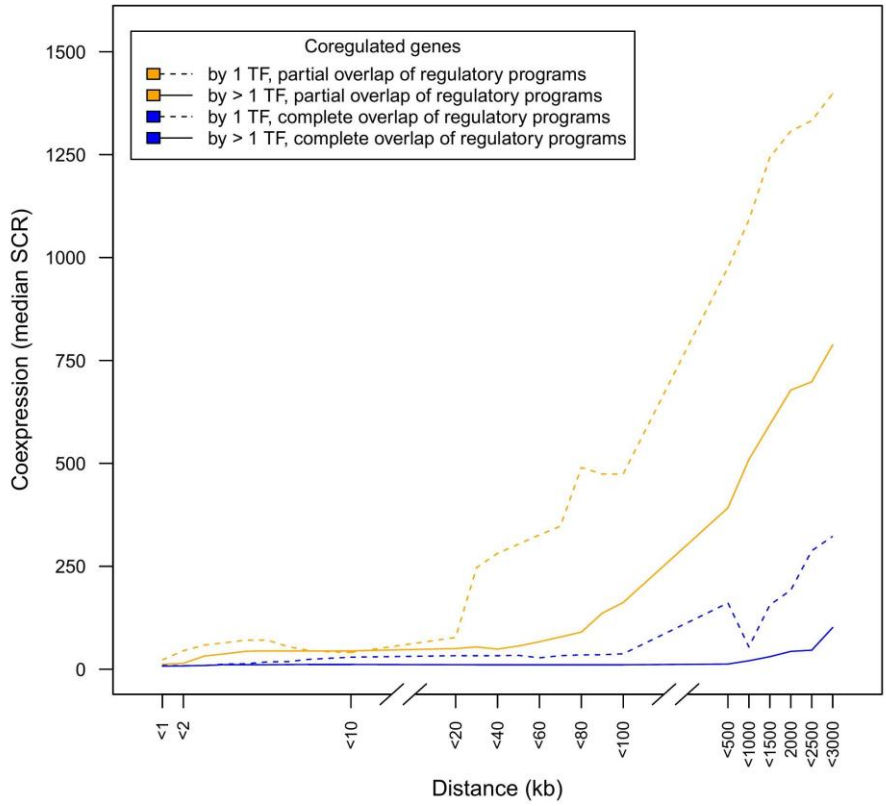

Fig 2. The distance between coregulated genes has a larger influence on their coexpression degree if genes are less tightly coregulated. The coexpression behavior of coregulated genes was disentangled, depending on whether the regulatory programs displayed complete versus partial overlap (blue versus orange) and depending on the number of common TFs present in the overlapping part of their regulatory program (dotted line for 1 TF versus full line for >1 TF).

https://doi.org/10.1371/journal.pone.0174887.g002

tightness of coexpression between a partial overlapping program with three shared TFs and a completely overlapping program with one TF). Therefore we conditioned the effects of the number of TFs in the shared part of their regulatory programs on whether the regulatory programs of these coregulated genes were completely versus partially overlapping. Both in case of a complete or a partial overlap of regulatory programs, we observed a higher degree of coexpression for those genes that have more than one common TF than for those that have only one common TF in the overlapping part of their regulatory programs. Also both in case of complete and partial overlap of regulatory programs, the degree of coexpression remains higher at larger distances for genes with more than one common TF versus for those with just one common TF (Fig 2, full orange curve versus dotted orange curve for partial overlap of

regulatory programs and full blue curve versus dotted blue curve for complete overlap of regulatory programs).

In general, distance thus affects the degree of coexpression, irrespective of the tightness of coregulation. For the most tightly coregulated genes the effect of distance is less visible as the genes tend to be highly coexpressed anyway and thus the contribution of small distances in increasing the degree of coexpression is the least pronounced for the most tightly coregulated genes. This indicates that the effect of distance is relatively small compared to effect of the tightness of the coregulation in determining the degree to which coregulated genes are coexpressed.

Non-operonic adjacent genes that are coregulated show a high degree of coexpression independently of their coregulation tightness or their genomic orientation

Focusing on coregulated genes that are located in each other's close neighborhood (<1 kb to <20 kb), it seems that their degree of coexpression is almost independent of the tightness of their coregulation: at such small distances, the mean degree of coexpression is not significantly different for coregulated genes with a completely overlapping or a partially overlapping regulatory program, and not significantly different for coregulated genes that share one or that share more TFs in the overlapping part of their regulatory program (Kruskal-Wallis $p << 0.001$, see also Fig 2, for respectively orange versus blue, and full versus dotted lines).

We argued that for genes that are involved in the same biological processes but are the *least tightly coregulated* i.e. by 1 TF and not the same regulatory program, their nearby location might be a way to guarantee the high degree of coexpression that would be needed to make them available together. To assess whether this was true in our data, we assessed whether indeed the least tightly coregulated genes that are located nearby were associated more frequently to the same biological processes than the least tightly coregulated genes located at larger genomic distances (>10 kb); to associate genes to biological processes Gene Ontology (GO) annotations were used (Materials and methods). This seemed indeed to be the case (Kruskal-Wallis $p = 0.007$).

From Fig 2 also appears that at an *extremely small* distance between coregulated genes (<2 kb), a high degree of coexpression of the coregulated genes is almost guaranteed irrespective of the tightness of coregulation (except for the very least tightly coregulated genes, see orange dotted curve). However, at such small distances we cannot exclude that the observed high degree of coexpression is caused by the occurrence of shared promoter elements (in divergently oriented adjacent promoters), or, *read-through* transcription [12] or not yet annotated operons (in codirectionally oriented promoters).

As these alternative causes of the observed high degree of coexpression can only exist for cases of divergently and codirectionally oriented gene pairs, we tested to which extent the high degree of coexpression observed between coregulated genes located at small distances from each other also held for convergently oriented genes.

Hereto we analyzed how the coexpression of genes that are members of coregulated adjacent operons, referred to as *coregulated proximally located genes*, depends on their relative orientation. Proximally located genes with divergent orientation are overrepresented in our dataset compared to those with other orientations (368 out of 490 proximally located pairs of genes or 75%) supporting the idea that divergent orientation indeed has evolved as a prevalent mechanism of assuring coexpression between adjacent coregulated genes as was also described by Korbel et al. [2]. Our results reveal that indeed proximally located coregulated genes are highly coexpressed when divergently oriented (median SCR 47). Also codirectionally oriented

proximally located coregulated genes are highly coexpressed as expected (median SCR 44). Having a divergent or codirectional promoter orientation can thus definitely account for part of the observed high degree of coexpression between proximally located coregulated genes. However, interestingly, also proximally located coregulated genes with convergent orientation showed equally high coexpression as those with the divergent and codirectional orientation (median SCR 34): coexpression was not significantly different between the divergent, codirectional or convergent orientation as indicated by the Kruskal-Wallis test ($p = 0.84$).

This observation indicates that at proximal distances, not only with distance confounded mechanisms of coregulation such as bidirectional cotranscription, readthrough transcription or unannotated operons, but also mere close distance can account for the observed high degrees of coexpression, independently from coregulation tightness. Note that the latter conclusion relies heavily on the evidence of 30 pairs only of proximally located genes in convergent orientation. One might argue therefore that we cannot rule out that the high degree of coexpression observed for coregulated genes at small distances is not the mere consequence of confounded mechanisms such as readthrough transcription.

To specifically assess the effect of readthrough transcription we evaluated whether the degree of coexpression of coregulated genes that are not proximally located but still located at small distances, was significantly lower than that of proximally located genes (with 'small distance' being defined as an intergenic distance of maximally 12 kb, equal to the maximum distance that is observed between proximally located genes). The mean degree of coexpression of coregulated genes that are not proximally located (145 pairs of genes) is not significantly different from that of proximally located genes (490 pairs of genes) (Kruskall-Wallis $p = 0.41$), indicating that besides known mechanisms, such as read-through transcription, also the mere effect of the small distance plays a role in determining levels of coexpression.

So, given that the relative orientation does not bias the coexpression degree of proximally located genes we conclude that the relative orientation causes no bias for the observed effect of the distance on the degree of coexpression of coregulated genes.

Coregulated genes are more coexpressed when they are located equidistantly relative to their common TF coding gene

To find a potential mechanism by which close distance of coregulated genes that is not mediated by read-through transcription or bidirectional cotranscription can explain higher degrees of coexpression, the following reasoning was made: assuming that the availability of TF molecules is limited by diffusion and assuming that coregulated genes that are exposed to similar quantities of TF proteins will be more coexpressed than coregulated genes that are not, we reasoned that coregulated genes that are more equidistantly located from their common TF coding gene are exposed to a more similar quantity of the TF encoded gene product and as a consequence will tend to be more coexpressed than coregulated genes that are not located equidistantly from their common TF gene.

To test this assumption, we compared the degree of coexpression hereby distinguishing between 1) coregulated genes located equidistantly with respect to their common TF gene and 2) coregulated genes not located equidistantly to their common TF gene. Equidistant means that the two distances, i.e. between the common TF gene and the two target genes, are within 90% of one another (Materials and methods). We restricted the analysis to genes that are coregulated by at most one TF in order to unequivocally define equidistancy to one and the same common TF and to exclude possible interferences of distances to other common TFs.

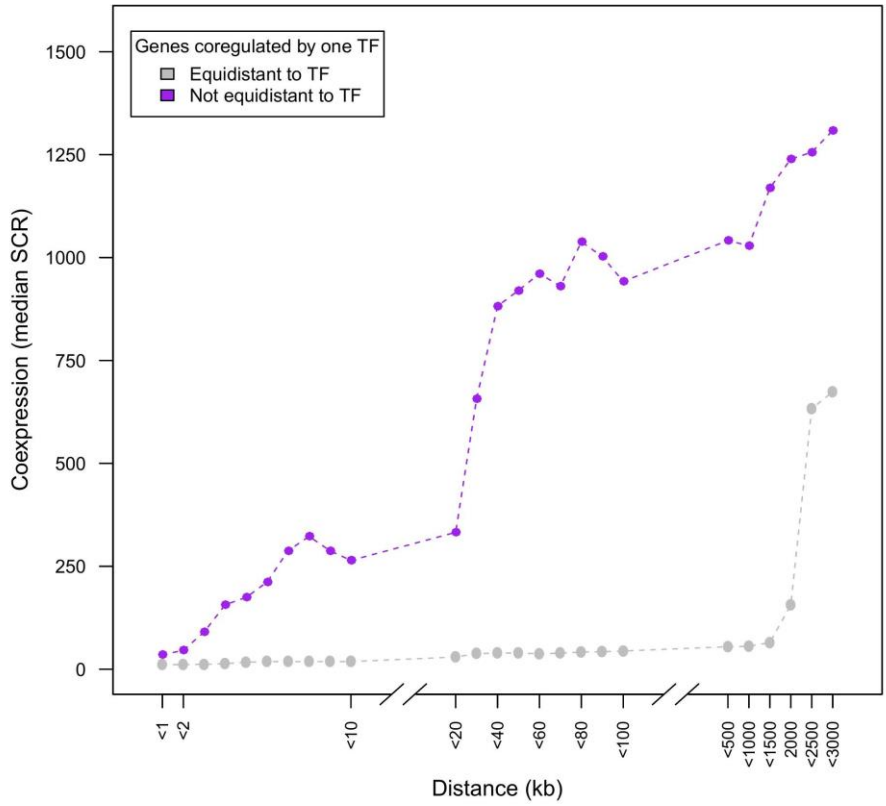

Fig 3. Effect of relative distance between TF and target genes on the degree of coexpression of the target genes. The coexpression behavior of genes that are coregulated by one TF is disentangled, depending on whether genes are equidistantly located (grey) or not equidistantly located (purple) relative to their common TF-coding gene. Y-axis displays the degree of coexpression (SCR), X-axis displays the maximum genomic distance between the coregulated genes.

https://doi.org/10.1371/journal.pone.0174887.g003

Fig 3 shows that coregulated genes that are equidistantly located from their common TF(s) (grey curve) generally are more coexpressed than genes that are not equidistantly located (purple curve).

The degree of coexpression between coregulated genes does not depend on the nearby location of their common TF coding gene

The previous paragraph supported the hypothesis that coregulated genes located equidistantly from their common TF are subject to similar local quantities of TF proteins and therefore show a higher degree of coexpression. One could also hypothesize that the closeness of the TF

coding gene itself could result in higher absolute local TF quantities in the target neighborhood, and as such increases the degree of coexpression of coregulated genes.

Fig 3 however shows that coexpression remains remarkably high for coregulated genes that are located equidistantly from their common TF, even when the genes themselves are located relatively distant from each other and thus by definition also relatively further from their common TF. This implies that for tightly coregulated genes sharing one common TF, coexpression is not only independent of the distance between those genes but, as a consequence, also independent of the distance of those genes relative to their common TF coding gene.

We further statistically tested this independence of the degree of coexpression on the distance between the common TF coding gene and the coregulated target genes. Hereto coregulated genes were classified in two groups referred to as *near to TF* or *far from TF*, depending on whether the distance between the common TF and the coregulated targets was smaller than or larger than 30 kb, respectively (Materials and methods). We included in these groups only those coregulated genes that were (1) located equidistantly from their common TF in order to study the mere effect of the distance between the TF and the coregulated genes on the degree of coexpression and to exclude the effect of unequal distances between the common TF and coregulated targets (see previous paragraph) and (2) coregulated by at most one TF to exclude effects caused by multiple common TFs between the coregulated genes or the effect of additional TFs that were not shared by the analyzed coregulated genes. Interestingly no statistically significant difference in degree of coexpression was observed between the two groups of coregulated genes referred to as respectively *near to TF* or *far from TF*, i.e. the null hypothesis of the Kruskal-Wallis test was rejected and the SCR of *near to TF* and *far from TF* are samples that come from the same population ($p = 0.50$).

In conclusion, our results demonstrate that in contrast to equidistancy from a common TF, a closer distance of a common TF to coregulated genes does not result in a higher degree of coexpression.

Close distance between highly coexpressed coregulated genes is evolutionarily constrained

Here we assumed that if the distance between coregulated genes plays a key role in affecting the degree of coexpression between those coregulated genes, this distance should be evolutionarily constrained. To test this assumption, we performed a comparative study in the subclass of gamma-proteobacteria [19–21] to assess whether the distance between coregulated genes is evolutionarily more conserved when the coregulated genes display a high degree of coexpression than when they do not. We started the analysis using all pairs of coregulated genes in *E. coli* and determined the orthologous of those genes in other gamma-proteobacteria. We defined as metric of *distance conservation* the proportion of the number of ortholog gene pairs in the different species for which genes have a distance equal to or smaller than the distance between the two corresponding coregulated genes in *E. coli* on the total number of considered orthologous pairs (Materials and methods).

In Fig 4 we plotted the average distance conservation of highly coexpressed ($SCR < 100$) and not (highly) coexpressed ($SCR > 1000$) coregulated genes in *E. coli* as a function of the distance between the genes. It can be observed that coregulated genes located at small distances (< 10 intervening genes) have a stronger distance conservation when they are highly coexpressed (30–40%, black curve) than when they are not coexpressed (25–30%, blue curve). This observation indicates that for highly coexpressed genes located in each other's neighborhood on the genome there is an evolutionary constraint on conserving their small distance. Because evolutionary conservation of close distance of genes has been associated with horizontal gene

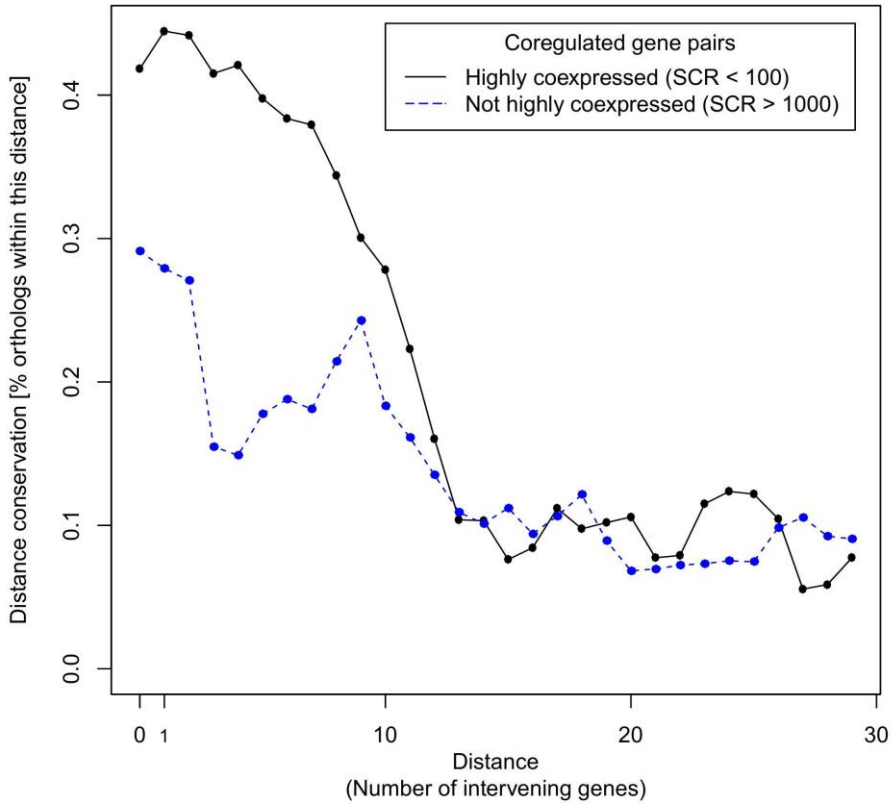

Fig 4. Evolutionary conservation of distance between coregulated genes. The x-axis represents the pairwise genomic distance between coregulated genes in E. coli, measured in intervening genes. The y-axis represents the degree to which for coregulated genes in E. coli the genomic distance is evolutionarily conserved in other gamma-proteobacteria which is expressed as the fraction of orthologous gene pairs for which the distance is equal or smaller (y-axis) than the distance between the corresponding genes in E. coli (x-axis) over the total number of analyzed orthologous genes. Orthologous genes are pairs of genes in other species that are orthologous to a pair considered in E. coli, i.e. a pair of coregulated genes in E. coli is expected to have an orthologous counterpart in other gamma-proteobacterial species if both genes in the E. coli pair have an orthologous counterpart in the considered gamma-proteobacterial species. Results are shown for respectively pairs of genes that are highly coexpressed (SCR < 100) (black curve) versus pairs of genes that are not coexpressed (SCR > 1000) (blue curve).

https://doi.org/10.1371/journal.pone.0174887.g004

co-transfer [22], we further hypothesized that highly coexpressed genes that are nearby located and that show strong distance conservation are likely to show evidence of horizontal co-transfer. We indeed found evidence of horizontal gene co-transfer for several of these cases (see Supplementary File S1 File part 5) which is thus an additional indication that for highly coexpressed nearby located genes there exists an evolutionary constraint for maintaining their small distance.

These results provide further evidence that for nearby coregulated genes for which coexpression is crucial, the vicinity or small distance is a driving force for guaranteeing high coexpression.

Discussion

In prokaryotes, genomic distance is a feature that in addition to coregulation affects coexpression. In this work, we evaluated how the genomic distance of genes known to be coregulated in *E. coli* contributes to their coexpression behavior. Hereto, we combined information on regulation in *E. coli* K-12 reported in RegulonDB, one of the largest curated and continually updated transcriptional databases, with publicly available expression data. Based on the information available for *E. coli* K-12 we observed that in general coregulated genes display higher degrees of coexpression as they are more closely located on the genome.

For genes that display very tight coregulation (e.g. genes with the exact same regulatory programs), this additional effect of genomic vicinity on coexpression is less pronounced compared to the distance effect observed for genes that are less tightly coregulated. This indicates that the contribution of genomic distance in determining the degree of coexpression is relatively small compared to the degree of coexpression that was determined by the tightness of the coregulation. As a consequence especially for non-tightly coregulated genes, distance seems to have a critical role in guaranteeing coexpression: only when located at small distances, the effect of the common TFs in increasing coexpression is large enough to compensate for the effect of the non-common TFs in potentially lowering coexpression. We found indications that non-tightly coregulated genes are located nearby to guarantee high coexpression in order to coordinate their common involvement in a particular biological process.

We showed that at very small distances, coexpression is high irrespective of the tightness of coregulation. This is because the small distance is at least partially the cause of coregulation, as is the case for *read-through* transcription or potentially unannotated operons (in codirectionally oriented operons) or for bidirectional cotranscription through common regulatory elements (in divergently oriented operons). However genes located in convergently oriented operons are also found to be highly coexpressed. In the latter case, the small distance cannot be causal to the coregulation and thus supports the idea of a *distance effect* as an additional factor independent of coregulation triggering high coexpression of closeby located coregulated genes.

We hypothesized that part of the distance effect can be explained by the fact that coregulated genes that are more closely located to each other are subject to more similar levels of TF molecules and are therefore more highly coexpressed. We could support this hypothesis by showing that coregulated genes that were located at similar distances relative to their common TF tend to be more coexpressed than genes that were not located equidistantly relative to their common TF. At very small distances, coregulated genes were found to be highly coexpressed, irrespective of whether or not they are located equidistantly relative to their TF. This may be explained by the fact that both coregulated genes are so close to each other that their distance to the common TF can only slightly differ.

Unlike the distance between target genes, the distance of the targets to the common TF coding gene does not seem to play a major role in determining coexpression of coregulated genes. This shows that, even when limited TF diffusion [23] may reduce TF availability at distances far away from the TF coding gene, coexpression can still be guaranteed because both targets are subject to a minimal, but comparable quantity of TF proteins. This hypothesis assumes that both target genes have the same response to their common TF, i.e. an equal concentration of TF proteins is needed to trigger gene expression (in the case of an activator TF) or to inhibit

gene expression (in the case of a repressor TF). Even though the assumption of equal responses to a TF in different target genes seems a major simplification of reality, for example because of different affinities or different numbers of binding sites for the common TF, in general, the effect of distance on coexpression is still visible.

Alternatively, one could imagine that when promoter regions reside at small distances, they are likely to be subject to the same degree of DNA supercoiling, bending or looping and thus more equally accessible to common TFs than more distantly located promoter regions [16,24–29]. In addition, colocalized promoter regions are more likely to be subject to the same degree of RNA polymerase molecules and the same degree of DNA phosphorylation which may add to the tightness of coregulation of nearby genes and thus to their coexpression. The observation that nearby coregulated genes tend to conserve their close distance more if they are highly coexpressed further adds to the importance of the vicinity in driving coexpression.

It is important to remark that our definition of *distance* being the linear distance along the chromosome is a strong simplification of the dynamic three-dimensional (3D) genome structure. As we currently do not have sufficient data available on dynamic 3D distances between genes, it is difficult to know the effect of the 3D distances. However, given that TF diffusion not only happens through 3D space but also by one-dimensional movement of TFs along the DNA segment such as "sliding" and "hopping" [30], it is not surprising that we find that also simply the linear genomic distance is a critical factor for coexpression of coregulated genes.

In conclusion, we systematically demonstrated that as much as genes are controlled by common TFs, their genomic distance functions is an additional and independent factor determining their coexpression. Our assumption that TF accessibility seems to be an important cause for enhancing coexpression at small distances, opens the door to more studies on local levels of TF molecules and their role in driving coexpression. In future studies on transcriptional regulation, distance is a critical factor to be taken into account in driving coexpression.

Materials and methods

Expression data

To retrieve *E. coli* expression data, we used the publicly available large-scale expression compendium COLOMBOS v3.0 compiling 4077 condition *contrasts* for 4321 genes [13]. 'Condition contrasts' do not represent single experimental conditions, but represent the difference between a test and reference condition (the differential expression values between the respective test and reference conditions in a particular contrast is expressed as a logratio). This concept 'condition contrast' is used in COLOMBOS to render expression values comparable across platforms and experiments. A full list of growth conditions from which the contrasts were derived as well as a more detailed explanation for condition *contrasts* is available at www.colombos.net.

Operon definitions

Operons were taken from direct literature curation at RegulonDB and bioinformatics predictions from ProOpDB [31]. Operon architectures were taken from ProOpDB because the accuracy of predictions of this database is one of the highest reported thus far (94.6%). In addition, the operon prediction of this database did not include coexpression as information source, whereby we avoided any circularity problem.

Coexpression measure

To quantify the degree of coexpression between any two genes, all pairwise similarities between gene expression profiles across all experimental conditions of the expression compendium

were calculated. We tested six different similarity measures: Pearson Correlation Coefficient (PCC), Spearman Correlation Coefficient (SCC), Mutual Information (MI), Pearson Correlation Rank (PCR), Spearman Correlation Rank (SCR) and Mutual Information Rank (MIR) and selected SCR as the similarity measure for our study as explained in Supplementary File S1 File part 1.

Note that our assessment of coexpression only took into account positive correlation and no anticorrelation. Although theoretically an inverse correlation could be expected, for example, between a repressor TF and its target genes, based on this work and our previous experience [10] it appears that negative correlation coefficients are not at all common. We therefore deliberately omitted assessing negative correlations as they would contribute relatively more spurious associations than true correlations.

The PCR, SCR and MIR, mentioned above are rank-based derivatives of respectively the PCC, SCC and MI and quantify how similar the expression profiles of two genes are relative to how similar these genes' expression profiles are to the expression profiles of all other genes (i.e. the similarity of expression profiles measured by PCC, SCC and MI respectively). The calculation of these rank-based derivatives of the PCC, SCC and MI is based on the work of Obayashi and coworkers [32,33]. In their work they propose the 'mutual rank' which is the ranked derivative of the PCC (here referred to as PCR). Below we provide details on the derivation of the SCR from the SCC according to the procedure described by Obayashi et al. [33]. The derivation of the PCR from the PCC and the MIR from the MI is calculated analogously. The derivation of the SCR from the SCC is as follows: calculating the SCC results in a symmetrical matrix in which each value contains the Spearman correlation between the gene expression profiles of any two genes A and B. (Supplementary File S1 File, part 2). This SCC matrix is converted into an asymmetric ranked matrix. To this end we assign a rank to each value in the row direction of the correlation matrix i.e. we rank all correlation values of gene A where the lowest rank i.e. 1 is assigned to the highest SCC value of gene A in the row and further ranks are assigned in descending order of the row SCC values of gene A. Each ranked value thus expresses how correlated gene A is with gene B relative to its correlation with all other genes (see Supplementary file S1 File part 2). This results in an asymmetric matrix in which the rank assigned to the correlation of gene A to B is not necessarily the same as the rank assigned to the correlation of gene B to A.

For each gene pair A-B an *SCR value* is subsequently derived by calculating the geometric mean of the two ranked values of A-B and B-A. We used the geometric mean rather than the arithmetic mean as this performed better as a measure of coexpression; this has been proved by Obayashi et al [33] and also showed the best results on our benchmark (data not shown). The added value of these rank-based derivatives of correlation in assessing the degree of coexpression between genes was described in the work of Obayashi and colleagues [33] and the advantage of using these measures particularly in our setup is explained in the Supplementary File S1 File part 2.

Modes of coregulation

Sets of coregulated genes were derived from regulatory interactions derived in RegulonDB v9.0 [14], a database containing information on the transcriptional regulation of *E. coli* strain K-12. Depending on the information that is available to support TF-gene regulatory interactions, RegulonDB distinguishes between interactions with "strong" or "weak" evidence. To ensure that the results derived in the main text were not influenced by whether or not we included interactions with weak evidence, we tested the impact of using different sets of interactions on our results, more specifically we tested a set including all interactions (i.e. those

supported by weak plus those by strong evidence i.e. a total of 3430 interactions), a set excluding interactions supported by one type of weak evidence only (2961 interactions), and a set containing interactions based on strong evidence only (in comparison to the previous setting here also interactions that are supported by two types of weak evidence are excluded, i.e. 2213 interactions). Results of these tests are presented in (Supplementary Material part 4) and show that in general the results and general conclusions hold irrespective of the type of dataset that was used as input. In the main text the results are shown for a dataset that containing all interactions except those supported by at most one source of weak evidence as this dataset offers a trade-off between containing the most reliable interactions, but still being sufficiently large to make statistical inferences.

Starting from the defined 2961 interactions, we derived 76891 coregulated genes used for our analysis; these were selected by taking all combinations of two genes that are not in the same operon (known and predicted operons as described above) and share at least one common TF with the same regulatory effect (activation, repression or dual). In total, 56235 out of 76891 coregulated gene pairs were coregulated only by a global TF and were left out: TFs with at least 130 target genes were considered global TFs, i.e. CRP (380 target genes), FNR (150), IHF (131), ArcA (133), Fis (268), and H-NS (140). The filtered dataset contained 11399 pairs of genes that are coregulated by at least one of 91 non-global TFs. A full list of the 91 TFs along with the number of pairs of genes they coregulate and per TF the mean of all pairwise distances and mean degrees of coexpression between the genes coregulated by that TF is given in the Supplementary Table S1 Table. Genes with a complete overlap of regulatory programs are defined as pairs of coregulated genes for which all TFs known to be involved in the regulation of either gene and with the same role (activator, repressor, or dual) are shared between both genes. Genes with a partially overlapping regulatory program are pairs of coregulated genes that do not share all of the TFs known to be involved in their regulation.

Distance measures

The distance between two genes is equal to the shortest distance (in base pairs) between the two structural gene start positions, i.e. by taking the shortest distance along the circular chromosome. Hereby, the shortest distance between two genes by definition is always smaller than half the chromosome length (4600 base pairs or 4,6 kilo base pairs). Note that for the assessment of distance conservation a different measure of distance was used (see below).

Measure of average degree of coexpression

In the plots of Figs 1–3 we took the median SCR as a measure of average coexpression degree because the median is less susceptible to *outliers* (here pairs of genes with extreme low degrees of coexpression (which means a high value of SCR) than the mean.

Identification of equidistancy to TF coding gene

To analyze the effect of equidistancy and the effect of distance to the common TF on coexpression of coregulated genes, we only considered genes that are coregulated by one common TF (1238 pairs of genes) to exclude additional and/or confounded effects due to coregulation by multiple TFs. Coregulated genes were considered to be located equidistantly from their common *TF* (i.e., TF coding gene), if the proportion of the smallest and the largest of the two corresponding distances for each of the two genes to the TF exceeded 0.9. Pairs of genes for which this was not the case were considered to be not equidistantly located relative to their common *TF*. In total 122 pairs of genes were located equidistantly and 1116 pairs of genes were located not equidistantly to their common TF coding gene.

When both genes in a pair have a distance to the common *TF* coding gene that was < = 30 kb the gene pair was considered to be located near to their common TF coding gene or *near to TF*. Alternatively if both genes in the pair had a distance to the common TF coding gene that was >30 kb the pair was considered to be located far from the TF coding gene or *far from TF*. A cut-off of 30 kb was taken by plotting the median SCR as a function of the distance of both target genes to the TF coding gene; at a distance of 30 kb, the slope of the median SCR changes, i.e., the rate at which the degree of coexpression decreases with the distance becomes lower (data not shown), 30 kb thus determines the range below which the effect of the distance on coregulation is most visible.

Measure to assess functional similarity

To assess whether pairs of genes belonged to the same functionality class according to Gene Ontology (GO) we used the BioConductor package GOSemSim [34] that allowed calculating the degree to which similar GO terms were associated to the considered pairs of genes. Gene Ontology annotations were downloaded from the gene ontology website (http://geneontology.org/page/go-annotation-file-format-20) and GO similarity between genes was calculated by taking semantic similarity between GO terms within the "Biological Process" ontology that were associated to the genes.

Identification of colocalized regulons

To identify different sets of coregulated genes that were genomicallycolocalized, we selected a set of coregulated genes between which the distance genes was <10 kb. Coregulated gene pairs within this set were used to calculate the degree of coexpression of colocalized coregulated genes. To assess the degree of coexpression of colocalized non-coregulated genes we used the combinations of genes from the set that were colocalized but did not share the same TF. The Kruskal-Wallis test was used to assess differences in mean degree of coexpression between the two sets.

Evolutionary conservation of distance

All genes and distances (as measured by the number of intervening genes) of 267 species of gamma-proteobacteria were collected with their respective orthologs for each gene in *E. coli* from GeConT [35]. In GeConT, two genes were considered to be orthologs by using Bidirectional Best Hit [36]. For each pair of coregulated genes with distance D in *E. coli*, we extracted N orthologous pairs of genes (with distance d) in N of 286 gamma-proteobacterial species, i.e., species in which orthologs existed for both genes of the *E. coli* pair. Conservation of the distance or *distance conservation* was defined as the proportion of orthologous pairs with distance d< = D relative to the total number of orthologous pairs, with orthologous pairs being the pairs of genes in a given species which are orthologous to two coregulated genes in *E. coli*. To select orthologous pairs we only considered species for which both genes in a coregulated pair of genes in *E. coli* contained an orthologous counterpart. For the evaluation of this metric, we considered the *distance* between any two genes as the *number of intervening genes* to normalize for the fact that the length of intergenic regions between orthologous genes can differ in different organisms. For the selection of highly and not highly coexpressed coregulated genes we took pairs of coregulated genes with SCR < 100 (5347 pairs of genes) and with SCR > 1000 (54936 pairs of genes), respectively.

Supporting information

S1 File. This file contains the following sections:

- Selection of a similarity measure to quantify coexpression
- Rank-based similarity measures compensate for conditional dependency
- The degree of coexpression of genes that are coregulated in E. coli is generally low
- Evidence of horizontal gene co-transfer in genes with strong distance conservation
(DOC)

S1 Table. List of TFs that control the pairs of coregulated genes used in this study. This table shows all TFs considered in our analysis with at least one pair of coregulated genes (see Materials and methods for the definition of coregulated genes). For each TF we showed the number of pairs of genes coregulated by that TF, the mean distance between every two genes in a pair (in base pairs), and the mean coexpression (as measured by SCR).
(DOC)

S2 Table. List of so-called non-coregulated genes that are potentially coregulated as derived from SELEX. This table shows an excerpt of pairs of genes that belong to the negative control of genes not known to be coregulated but highly coexpressed and located nearby and that were predicted to be coregulated according to SELEX. Gene 1 and gene 2 correspond to a pair of genes selected from the set of so-called non-coregulated genes with selection criteria 1) small distance (< 10 kb) 2) high degree of coexpression (SCR<10) and 3) at least one common TF in their respective set of TFs as predicted by SELEX. The coexpression degree between gene 1 and gene 2 is given in SCR. One or more predicted common TF(s) was (were) given. The numbers between brackets refer to the Nth best hit (which means Nth highest % similarity of that TF for that gene) that TF was for respectively gene 1 and gene 2.
(DOCX)

S3 Table. List of all coregulated genes with their genomic distances and coexpression degrees. This table contains the following fields: TF (coregulating TF), gene 1 (first gene in the pair), gene 2 (second gene in the pair), distance (genomic distance between gene 1 and gene 2), TF-gene1_distance (genomic distance between the TF and the first gene), TF-gene2_distance (genomic distance between the TF and the second gene), role (a = activator, r = repressor, d = dual), SCR (Spearman Correlation Rank as a measure of coexpression degree).
(TXT)

Acknowledgments

Lucia Pannier is a doctoral student from Programa de Doctorado en Ciencias Biomédicas (PDCB) in Centro de Ciencias Genomicas (CCG) of Univerdidad Nacional Autónoma de México (UNAM) and received PhD fellowship (420430) from Consejo Nacional de Ciencia y Tecnología México (CONACyT) and was partially supported by the National Institutes of Health under grant number R01GM110597 and FOINS CONACyt Fronteras de la Ciencia under project number 15. We would like to thank Cesar Bonavides-Martínez for technical support. Ghent University Multidisciplinary Research Partnership 'Bioinformatics: from nucleotides to networks'; Fonds Wetenschappelijk Onderzoek-Vlaanderen (FWO) [G.0329.09, 3G042813, G.0A53.15N]; Agentschap voor Innovatie door Wetenschap en Technologie (IWT) [NEMOA]. We thank the anonymous reviewers for their careful reading of our manuscript and their many insightful comments and suggestions.

Author Contributions

Conceptualization: LP EM KM JCV.

Data curation: LP EM.

Formal analysis: LP EM.

Funding acquisition: LP KM JCV.

Investigation: LP KM EM JCV.

Methodology: LP KM.

Project administration: LP KM.

Resources: JCV.

Software: LP.

Supervision: KM JCV.

Validation: LP KM EM.

Visualization: LP EM.

Writing – original draft: LP.

Writing – review & editing: LP EM KM.

References

1. Zampieri M, Soranzo N, Bianchini D, Altafini C. Origin of Co-Expression Patterns in E.coli and S.cerevisiae Emerging from Reverse Engineering Algorithms. Isalan M, editor. PLoS One. San Francisco, USA: Public Library of Science; 2008; 3: e2981. https://doi.org/10.1371/journal.pone.0002981 PMID: 18714358
2. Korbel JO, Jensen LJ, von Mering C, Bork P. Analysis of genomic context: prediction of functional associations from conserved bidirectionally transcribed gene pairs. Nat Biotech. Nature Publishing Group; 2004; 22: 911–917. Available: http://dx.doi.org/10.1038/nbt988
3. Kruglyak S, Tang H. Regulation of adjacent yeast genes. Trends Genet. 2000; 16: 109–111. http://dx.doi.org/10.1016/S0168-9525(99)01941-1 PMID: 10689350
4. Homouz D, Kudlicki AS. The 3D Organization of the Yeast Genome Correlates with Co-Expression and Reflects Functional Relations between Genes. Khodursky AB, editor. PLoS One. San Francisco, USA: Public Library of Science; 2013; 8: e54699. https://doi.org/10.1371/journal.pone.0054699 PMID: 23382942
5. Williams EJB, Bowles DJ. Coexpression of Neighboring Genes in the Genome of Arabidopsis thaliana. Genome Res. Cold Spring Harbor Laboratory Press; 2004; 14: 1060–1067.
6. Ng YK, Wu W, Zhang L. Positive correlation between gene coexpression and positional clustering in the zebrafish genome. BMC Genomics. BioMed Central; 2009; 10: 42.
7. Beck CF, Warren RA. Divergent promoters, a common form of gene organization. Microbiol Rev. 1988; 52: 318–326. Available: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC373147/ PMID: 3054465
8. Rhee KY, Opel M, Ito E, Hung S, Arfin SM, Hatfield GW. Transcriptional coupling between the divergent promoters of a prototypic LysR-type regulatory system, the ilvYC operon of Escherichia coli. Proc Natl Acad Sci U S A. The National Academy of Sciences; 1999; 96: 14294–14299. Available: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC24430/
9. Janga SC, Salgado H, Martínez-Antonio A. Transcriptional regulation shapes the organization of genes on bacterial chromosomes. Nucleic Acids Res. Oxford University Press; 2009; 37: 3680–3688.
10. Michoel T, De Smet R, Joshi A, Van de Peer Y, Marchal K. Comparative analysis of module-based versus direct methods for reverse-engineering transcriptional regulatory networks. BMC Syst Biol. BioMed Central; 2009; 3: 49.

11. Zhang H, Yin Y, Olman V, Xu Y. Genomic Arrangement of Regulons in Bacterial Genomes. Badger JH, editor. PLoS One. San Francisco, USA: Public Library of Science; 2012; 7: e29496. https://doi.org/10.1371/journal.pone.0029496 PMID: 22235300
12. Lee F, Yanofsky C. Transcription termination at the trp operon attenuators of Escherichia coli and Salmonella typhimurium: RNA secondary structure and regulation of termination. Proc Natl Acad Sci U S A. 1977; 74: 4365–4369. Available: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC431942/ PMID: 337297
13. Moretto M, Sonego P, Dierckxsens N, Brilli M, Bianco L, Ledezma-Tejeida D, et al. COLOMBOS v3.0: leveraging gene expression compendia for cross-species analyses. Nucleic Acids Res. Oxford University Press; 2016; 44: D620–D623.
14. Gama-Castro S, Salgado H, Santos-Zavaleta A, Ledezma-Tejeida D, Muñiz-Rascado L, García-Sotelo JS, et al. RegulonDB version 9.0: high-level integration of gene regulation, coexpression, motif clustering and beyond. Nucleic Acids Res. 2015;
15. Lemmens K, De Bie T, Dhollander T, De Keersmaecker SC, Thijs IM, Schoofs G, et al. DISTILLER: a data integration framework to reveal condition dependency of complex regulons in Escherichia coli. Genome Biol. BioMed Central; 2009; 10: R27–R27.
16. Sobetzko P. Transcription-coupled DNA supercoiling dictates the chromosomal arrangement of bacterial genes. Nucleic Acids Res. Oxford University Press; 2016; 44: 1514–1524.
17. Riley TR, Slattery M, Abe N, Rastogi C, Mann R, Bussemaker H. SELEX-seq, a method for characterizing the complete repertoire of binding site preferences for transcription factor complexes. Methods Mol Biol. 2014; 1196: 255–278. https://doi.org/10.1007/978-1-4939-1242-1_16 PMID: 25151169
18. Yu H, Luscombe NM, Qian J, Gerstein M. Genomic analysis of gene expression relationships in transcriptional regulatory networks. Trends Genet. 2003; 19: 422–427. http://dx.doi.org/10.1016/S0168-9525(03)00175-6 PMID: 12902159
19. Pérez AG, Angarica VE, Vasconcelos ATR, Collado-Vides J. Tractor_DB (version 2.0): a database of regulatory interactions in gamma-proteobacterial genomes. Nucleic Acids Res. Oxford University Press; 2007; 35: D132–D136.
20. González Pérez AD, González González E, Espinosa Angarica V, Vasconcelos ATR, Collado-Vides J. Impact of Transcription Units rearrangement on the evolution of the regulatory network of gamma-proteobacteria. BMC Genomics. BioMed Central; 2008; 9: 128.
21. Lozada-Chávez I, Janga SC, Collado-Vides J. Bacterial regulatory networks are extremely flexible in evolution. Nucleic Acids Res. Oxford University Press; 2006; 34: 3434–3445.
22. Dilthey A, Lercher MJ. Horizontally transferred genes cluster spatially and metabolically. Biol Direct. London: BioMed Central; 2015; 10: 72.
23. Kuhlman TE, Cox EC. Gene location and DNA density determine transcription factor distributions in Escherichia coli. Mol Syst Biol. Nature Publishing Group; 2012; 8: 610.
24. Browning DF, Grainger DC, Busby SJW. Effects of nucleoid-associated proteins on bacterial chromosome structure and gene expression. Curr Opin Microbiol. 2010; 13: 773–780. http://dx.doi.org/10.1016/j.mib.2010.09.013 https://doi.org/10.1016/j.mib.2010.09.013 PMID: 20951079
25. Dillon SC, Dorman CJ. Bacterial nucleoid-associated proteins, nucleoid structure and gene expression. Nat Rev Micro. Nature Publishing Group; 2010; 8: 185–195. Available: http://dx.doi.org/10.1038/nrmicro2261
26. Dorman CJ. Co-operative roles for DNA supercoiling and nucleoid-associated proteins in the regulation of bacterial transcription. Biochem Soc Trans. 2013; 41: 542–7. Available: http://www.ncbi.nlm.nih.gov/pubmed/23514151 https://doi.org/10.1042/BST20120222 PMID: 23514151
27. Marr C, Geertz M, Hütt M-T, Muskhelishvili G. Dissecting the logical types of network control in gene expression profiles. BMC Syst Biol. BioMed Central; 2008; 2: 18.
28. Toth A, Tischler ME, Pal M, Koller A, Johnson PC. A multipurpose instrument for quantitative intravital microscopy. J Appl Physiol. 1992; 73: 296–306. Available: http://jap.physiology.org/content/73/1/296.abstract PMID: 1506384
29. Peter BJ, Arsuaga J, Breier AM, Khodursky AB, Brown PO, Cozzarelli NR. Genomic transcriptional response to loss of chromosomal supercoiling in Escherichia coli. Genome Biol. London: BioMed Central; 2004; 5: R87–R87.
30. Halford SE, Marko JF. How do site-specific DNA-binding proteins find their targets? Nucleic Acids Res. Oxford, UK: Oxford University Press; 2004; 32: 3040–3052.
31. Taboada B, Ciria R, Martinez-Guerrero CE, Merino E. ProOpDB: Prokaryotic Operon DataBase. Nucleic Acids Res. Oxford University Press; 2012; 40: D627–D631.

32. Mutwil M, Klie S, Tohge T, Giorgi FM, Wilkins O, Campbell MM, et al. PlaNet: Combined Sequence and Expression Comparisons across Plant Networks Derived from Seven Species. Plant Cell. American Society of Plant Biologists; 2011; 23: 895–910.

33. Obayashi T, Kinoshita K. Rank of Correlation Coefficient as a Comparable Measure for Biological Significance of Gene Coexpression. DNA Res An Int J Rapid Publ Reports Genes Genomes. Oxford University Press; 2009; 16: 249–260.

34. Yu G, Li F, Qin Y, Bo X, Wu Y, Wang S. GOSemSim: an R package for measuring semantic similarity among GO terms and gene products. Bioinforma. 2010; 26: 976–978.

35. Ciria R, Abreu-Goodger C, Morett E, Merino E. GeConT: Gene context analysis. Bioinformatics. 2004; 20: 2307–2308. https://doi.org/10.1093/bioinformatics/bth216 PMID: 15073003

36. Smith TF, Waterman MS. Identification of Common Molecular Subsequences. J Mol Biol Vol 147, No 1 (25 March 1981), pp 195–197 Key citeulike668527. 1981; 147: 195–197.

2.2 Segundo artículo: RegulonDB version 9.0: High-level integration of gene regulation, coexpression, motif clustering and beyond.

Mi contribución en este artículo fue (1) implementar una herramienta en la nueva versión de la base de datos RegulonDB llamada *Coexpression Page* y (2) una página que resume la coexpresión en operones y regulones llamada *Coexpression Overview.* La implementación de la *Coexpression Page* permite evaluar la coexpresión para uno o más genes de interés; A) la coexpresión entre ellos y B) los genes mejor coexpresados con cada uno de ellos. En la *Coexpression Overview* integramos una página que contiene el resumen de la coexpresión para dos grupos de interés biológico siendo operones y regulones.

El uso y la aplicación de la herramienta *Coexpression Page* se explica en el articulo (Gama-Castro et al. 2015) y la *Coexpression Overview* se puede consultar en http://regulondb.ccg.unam.mx/ Home page > Integrated views and tools > Coexpression Browser.

Published online 2 November 2015 *Nucleic Acids Research, 2016, Vol. 44, Database issue* ***D133–D143***
doi: 10.1093/nar/gkv1156

RegulonDB version 9.0: high-level integration of gene regulation, coexpression, motif clustering and beyond

Socorro Gama-Castro[1,†], Heladia Salgado[1,†], Alberto Santos-Zavaleta[1], Daniela Ledezma-Tejeida[1], Luis Muñiz-Rascado[1], Jair Santiago García-Sotelo[1], Kevin Alquicira-Hernández[1], Irma Martínez-Flores[1], Lucia Pannier[1], Jaime Abraham Castro-Mondragón[2], Alejandra Medina-Rivera[3], Hilda Solano-Lira[1], César Bonavides-Martínez[1], Ernesto Pérez-Rueda[4], Shirley Alquicira-Hernández[1], Liliana Porrón-Sotelo[1], Alejandra López-Fuentes[1], Anastasia Hernández-Koutoucheva[1], Víctor Del Moral-Chávez[1], Fabio Rinaldi[5] and Julio Collado-Vides[1,*]

[1]Programa de Genómica Computacional, Centro de Ciencias Genómicas, Universidad Nacional Autónoma de México, A.P. 565-A, Cuernavaca, Morelos 62100, Mexico, [2]UMR_S 1090 TAGC, INSERM, Marseille, 13000 France, [3]Laboratorio Internacional de Investigación sobre el Genoma Humano, Universidad Nacional Autónoma de México, Campus Juriquilla, Boulevard Juriquilla 3001, Juriquilla 76230, Santiago de Querétaro, QRO, Mexico, [4]Departamento de Microbiologia Molecular, IBT, Universidad Nacional Autónoma de México, Cuernavaca, Morelos 62100, Mexico and [5]Institute of Computational Linguistics, University of Zurich, Binzmühlestrasse 14, CH-8050 Zurich, Switzerland

Received September 15, 2015; Revised October 17, 2015; Accepted October 19, 2015

ABSTRACT

RegulonDB (http://regulondb.ccg.unam.mx) is one of the most useful and important resources on bacterial gene regulation,as it integrates the scattered scientific knowledge of the best-characterized organism, *Escherichia coli* K-12, in a database that organizes large amounts of data. Its electronic format enables researchers to compare their results with the legacy of previous knowledge and supports bioinformatics tools and model building. Here, we summarize our progress with RegulonDB since our last *Nucleic Acids Research* publication describing RegulonDB, in 2013. In addition to maintaining curation up-to-date, we report a collection of 232 interactions with small RNAs affecting 192 genes, and the complete repertoire of 189 Elementary Genetic Sensory-Response units (GENSOR units), integrating the signal, regulatory interactions, and metabolic pathways they govern. These additions represent major progress to a higher level of understanding of regulated processes. We have updated the computationally predicted transcription factors, which total 304 (184 with experimental evidence and 120 from computational predictions); we updated our position-weight matrices and have included tools for clustering them in evolutionary families. We describe our semiautomatic strategy to accelerate curation, including datasets from high-throughput experiments, a novel coexpression distance to search for 'neighborhood' genes to known operons and regulons, and computational developments.

INTRODUCTION

RegulonDB is a relational database that offers, in an organized and computable form, updated knowledge on transcriptional regulation in *Escherichia coli* K-12 (1). RegulonDB, first published in 1998, captures the results of a continuous effort to this day (2). Our curation work also feeds the EcoCyc database (3), which together with RegulonDB are the major sources of organized information for the best-known bacterial genome model organism. For years we have expanded the number of objects and their properties in our database, always enriching the modeling of the molecular components governing transcription initiation, as we strive to keep up-to-date with new methodologies. We have also enriched the modeling of gene regulation, proposing new concepts, such as regulatory phrases (4) and, more recently, GENSOR units (genetic sensory-response units) that link signals, the associated regulatory interactions and the regulated response as metabolic and cellular capabilities (5). Briefly, RegulonDB facilitates access to organized information on the mechanisms of tran-

*To whom correspondence should be addressed. Tel: +52-777-3132063; Fax: +52-777-317-5581; Email: collado@ccg.unam.mx
†These authors contributed equally to the paper as first authors.

scription initiation; more precisely, RegulonDB organizes the available information on the shadows and fingerprints of these mechanisms in the genome.

Here we present progress since the last *Nucleic Acids Research* (NAR) paper, published in 2013 (1). We have kept our curation up-to-date, including regulation by small RNAs (sRNAs); we report a rather complete repertoire of elementary GENSOR units, each one integrating the network of regulatory interactions and metabolic pathways affected by one transcription factor (TF). This additional information represents major progress for our focus on integrative approaches to facilitate not only information but also summarized knowledge given the high granularity of most biological processes. We have updated the set of computationally predicted TFs, as well as the high-quality position-weight matrices (PWMs) for each TF with sufficient binding sites, and we have included a novel browser based on a clustering of such matrices that reflects their grouping into TF evolutionary families.

We are well aware that a critical barrier in genomics is how to accelerate access to and processing of the large amounts of information and knowledge that are continuously generated. Curation is a bottleneck for facilitating the capability to digest the tsunami of genomic knowledge. This motivated us to initiate the implementation of assisted curation by means of natural language processing (NLP) strategies, thanks to a collaboration with Dr Fabio Rinaldi, an expert in the field. Our initial results capturing growth conditions are promising, although there is a long way to go. Curation of high-throughput (HT) datasets (i.e., chromatin immunoprecipitation [ChIP] variants, microarrays, gSELEX and transcriptional start site [TSS] mapping) is a delicate issue, since we do not want to dilute the high-quality classical experiments with the massive but more fragmented knowledge that these methodologies produce. Our current solution, as discussed in detail below, is at the crossroads of two paradigms: the classic one of a relatively well-organized genome with promoters and binding sites involved in regulation of transcription initiation, and one inundated by scattered promoters and binding sites, many of which we do not yet know if they are involved in transcriptional regulation.

Another avenue linking RegulonDB data at this time with HT-generated profiles of expression is the capability we have implemented for evaluating the similarity of coexpression of any two genes, based on the COLOMBOS microarray library of experiments. We offer those similarity values for all operons and regulons. Finally, additional computational developments are summarized.

RESULTS

The RegulonDB version 9.0 release contains all the data described below, the sRNAs, elementary GENSOR units and HT datasets. Literature curation is typically up-to-date within 2 months on average for each release.

AN UPDATED COLLECTION OF INFORMATION ON REGULATORY sRNAs

Classic regulation of transcription initiation governed by TFs affecting promoter activity has been the major focus of RegulonDB. However, as years have passed we have expanded our curation to include regulation by small metabolites and proteins targeting RNA polymerase directly, as well as regulation by sRNAs. The regulatory potential of sRNAs is magnified when we take into account the fact that some sRNAs regulate the expression of genes themselves involved in regulation of many genes, such as sigma factors (like sigma32), global TFs (like H-NS) and other local TFs (like OmpR) which indirectly affect the expression of numerous genes.

We present an updated, integrative view of the known *E. coli* sRNAs. In our manual curation, we considered only data supported by experimental evidence, with the large majority supported by strong evidence (i.e., based on RT-PCR), except for 10 sRNAs supported by microarray experimental data. A total of 120 sRNAs with 231 total interactions are included in this collection, which all together regulate 192 genes. This collection includes detailed and high-quality information about the known regulatory interactions of sRNAs, such as the binding motifs in the targets.

COMPREHENSIVE SEMIAUTOMATIC CURATED ELEMENTARY GENSOR UNITS

A GENSOR unit, a short term for 'genetic sensory-response unit,' initially defined by Gama Castro *et al.* in 2011 (5), is a novel concept that from our perspective places regulatory mechanisms in their natural biological context, as part of a flux of information that starts with a change (appearance of a signal) that elicits a regulated response.

Since the 2013 article, we have updated 45 already-curated GENSOR units and added 144 new GENSOR units, to a total of 189. We defined the boundaries of the GENSOR unit concept and its constituents: currently all GENSOR units are elementary, since they are limited to a single TF, starting with the signal, all reactions from the signal to the effector binding the TF, the effect of the TF active conformation on the regulated genes, the regulated transcription units (TUs), their mRNAs, products and the reactions of these products. If any enzyme is part of a multimeric complex, all the monomers of the complex are added (even if they are not directly regulated by the TF). These 144 new GENSOR units have been curated by a semiautomatic method that starts with a pipeline of programs that extract all information pertinent to a GENSOR unit from the RegulonDB and EcoCyc databases; such data are subsequently manually revised and curated and used to generate the visual map available in RegulonDB. The full methodology, motivation and relevance of this integrative new concept will be published elsewhere (Ledezma *et al.*, manuscript in preparation).

GENSOR unit components and their interactions place a TF and its regulatory mechanisms in a larger context, providing evidence in many cases for the TF's role in decision-making processes and information flux from the signal to the elicited genetically encoded response.

For the process of GENSOR unit construction, we considered the need to reflect relationships between metabolites where two or more are in the same metabolic pathway only a few reactions apart, specifically in coregulated pathways, since some reactions are not necessarily regulated by the TF

that defines a particular GENSOR unit. In RegulonDB version 7.0, we included 'super-reactions' to include these reaction gaps. In this most recent version of RegulonDB, we have limited the number of reaction gaps to a maximum of three. Reactions have to be successive and present among EcoCyc's metabolic pathways. Three is the average number of total reactions in EcoCyc pathways; since the median number of reactions is 2, this limit allows more than 50% of the pathways to be completed by reaction gaps in their respective GENSOR unit.

RegulonDB 9.0 hosts a GENSOR unit for each local TF (6) for which there is experimental evidence in the database. A total of 103 TFs have a known effector in RegulonDB, including 25 two-component systems. When available, the four components of the GENSOR unit are highlighted: the signal, the signal processing, the genetic switch and the response. By default, effectors are considered signals, unless a reaction that produces the effector is present in the GENSOR unit, in which case the substrate of that reaction is deemed the signal and the reaction itself becomes part of the signal processing. Directionality of reactions is considered in the identification of the four components, as well as for the addition of reaction gaps.

A total of 78 GENSOR units have their four components highlighted; 119 include the genetic switch and the response, and 2 contain only the genetic switch. We believe this gradient of knowledge is a reflection of both the information that we have yet to discover and the cooperation among TFs to orchestrate complete biological processes. GENSOR units, apart from revealing the precise role of the activity of a TF in cellular metabolism, are the building blocks of larger GENSOR units that will describe decisions encoded in the genome in response to changes in the environment.

GENSOR units for which there is sufficient information about their four components have a short written summary describing the higher-level flow of information portrayed. For example, the BetI GENSOR unit (Figure 1) shows that external choline is transported inside the cell, where it binds to BetI and allows the activation of genes involved in the conversion of imported choline to glycine betaine. This information comes from the interactions between the elements in the GENSOR unit, rather than from the elements alone, thus describing the unit with the higher granularity of description, which is more appropriate to understanding physiological and biochemical processes.

The RegulonDB portal has a Web page with the complete list of GENSOR units grouped either by the transduction mechanism or by the signal that initiates a flux of regulated processes (available in the menu under 'Integrated Views & Tools/RegulonDB Overviews/GENSOR Unit Groups'). Images comply with the Cell Designer (7) graphical notation (8). Cell Designer 4.4 XML format files are available for download for all GENSOR units and their components. Users interested in importing GENSOR units into SBGN-compatible tools can download pure SBML level 2 version 4 XML format files.

We redesigned the Web page for GENSOR units, and this page now contains three sections: the graphical map of the elementary GENSOR unit, its general properties, including the written summary and a section for the properties of each reaction.

UPDATED TF FAMILIES, POSITION-WEIGHT MATRICES AND THEIR GROUPING IN CLUSTERS

A core component of the transcription machinery is the TF-binding site (TFBS) interaction. In this section, we describe, first our incorporation of an updated set of computationally predicted TFs. Second, we have updated the construction of PWMs for TFs with sufficient known TFBSs. Third, we offer the clustering of TFs based on the similarities of their matrices.

Updated set of predicted TFs

We updated the set of computationally predicted TFs, based on recent work by Perez-Rueda *et al*. (9). A total of 184 TFs experimentally characterized and for which information was deposited in RegulonDB (1) were used as seeds in BLASTP searches against the complete proteome of *E. coli*. *E*-values of ≤1e-6 and a coverage of 70% were required for a TF to be considered a putative TF. In addition, TFs specifically associated with *E. coli* K-12 and deposited in the DBD, HAMAP (10), Superfamily DB (11) and PFAM (12) databases were retrieved. Superfamily and family assignations were based on Superfamily annotations (11), PFAM (12) and the Conserved Domain Database (CDD) (11–14). Forty-two groups of paralogs defined by BLASTP (14) comparisons for which the *E*-values were ≤1e-6 and for which coverage was at least 50% of any of the proteins in the alignment were identified in the total set of TFs.

In total, the repertoire of TFs comprises 304 proteins. Of these, 184 are experimentally described in RegulonDB and 120 are predictions. These proteins can be classified in 78 different evolutionary families based on PFAM, CDD and Superfamily annotations. The most abundant family (LysR) entails 46 proteins (15% of the total number of TFs), although for almost 50% of these there is not any experimental evidence. Two additional large families were also identified, AraC/XylS (26 proteins) and GntR (20 proteins). An important improvement of the previous predictions is the elimination of false positives, such as for transposases and integrases.

Of the 184 experimentally described TFs, those for which we have identified binding sites are the subject for PWM construction and clustering, as described in the following section.

Updated set of PWMs

A minimum of four annotated TFBSs is required for building a motif in the form of a PWM. There were enough sites to build a motif for 93 TFs, 7 more than in the previous version; the full set of sites include 3195 TF → gene regulatory interactions. Using different sequence lengths (variation of ±4 bp around the annotated binding site length), programs (MEME and consensus) and background models (orders 0 and 1), we evaluated the different motifs available for each TF and selected the one with the best quality (15). At the time of the RegulonDB version 8.0 release, we had evaluated the quality of PWMs by taking into account (i) the information content conservation across the PWM; (ii) the false-positive rate for recovering 70% of the annotated

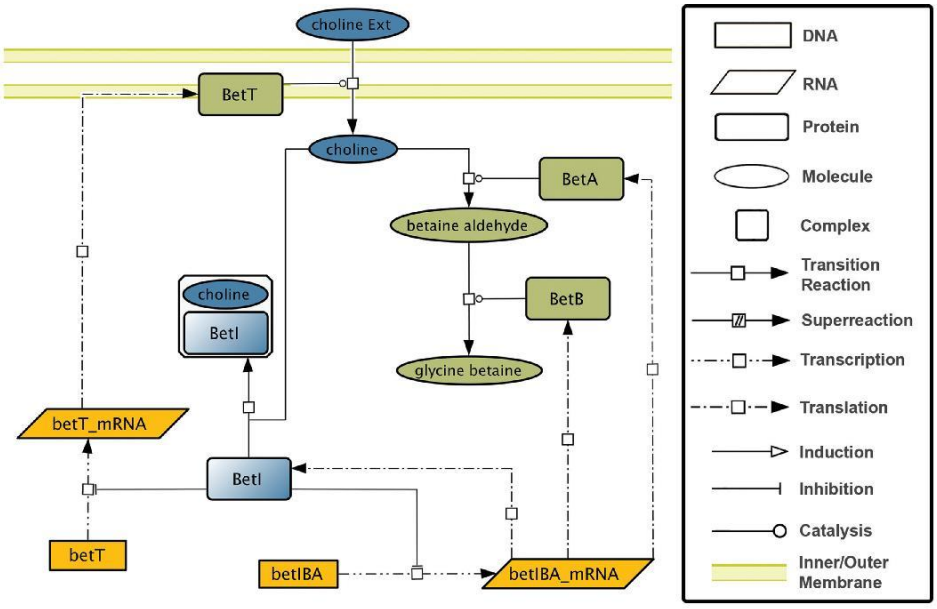

Figure 1. The BetI GENSOR unit. The signal and signal processing, in this case transport of choline through the membrane, are shown in blue. The genetic switch, i.e., repression of *betT* and *betIBA* transcription units, is shown in yellow. The response is shown in green: production of BetT, a choline transporter, and BetA and BetB, enzymes responsible for the utilization of choline.

sites; (iii) the difference between the observed distribution of scores in the upstream regions on *E. coli* K-12 versus the theoretical distribution; and (iv) the level of overfitting of the PWM to the original sequences used to build it (1).

We were able to obtain a high-quality matrix for 60% of the TFs, 10% more than in the previous version. This version includes motifs for seven TFs that fulfilled only the requirement for four binding sites, plus nine TFs for which the quality increased from low to good. A flat file with the PWMs in consensus format was added in the downloads page of the website. Additionally, in the 'Integrated Views & Tools' section, a browser allows navigation through each TF and the distributions that support the quality of the PWMs.

Using the evaluated set of motifs, we found 16 207 predicted binding sites in the upstream region of the *E. coli* K-12 MG1655 *uid57779* genes, where upstream regions are defined as −400 bp upstream to 50 bp downstream from the start codon (16,17). This set of predicted binding sites corresponds to 12 574 TF → gene regulatory interactions; this represents a recovery of 52% of the 1592 annotated regulatory interactions in the database for the 93 TFs for which we have a PWM, which represents a 9% improvement from the previous RegulonDB version. If only TFs with a good-quality PWM are taken into account, the total number of predicted TF → gene interactions is 8714, recovering 672 (57%) of annotated interactions for this TF subset. The TFBS predictions can be obtained from the 'Dataset' menu in the 'Computational Prediction' section.

Clustering of PWMs

TFs belong to evolutionarily related families, where members of the same family tend to share a significant similarity of protein domains that bind to DNA, which in bacteria are most frequently helix-turn-helix motifs.

The 93 PWMs available in RegulonDB, built as mentioned before, were analyzed with the program *matrix-clustering* (16), a tool that groups similar PWMs. Given the high similarity of motifs of proteins of the same family, this program can be used to identify TF-binding motifs (TFBMs) that belong to phylogenetically related TFs or DNA-binding proteins that recognize similar DNA sequences. The clustering can be displayed as a collection of hierarchical trees (forest), where each tree represents a cluster with its global alignment of PWMs. Additionally, a heat map representation with an all-versus-all PWMs comparison is also now possible.

We found 47 clusters formed by PWMs corresponding exclusively to TFs of the same family (e.g. AraC, LacI, NarL, GntR and NagC). The alignments of these PWMs show the conserved and non-conserved positions between them. These groups of PWMs are summarized as Familial Bind-

ing Profiles (FBP) (14), a general PWM that represents a collection of similar motifs highlighting the similar positions of the clustered PWMs, which allow us to potentially have one FBP for each TF family.

The PWMs were grouped as follows: (i) all the motifs were compared to each other using two metrics to measure their similarity (the Pearson correlation coefficient and a normalized version of the Pearson correlation relative to the width of the match between two aligned PWMs) (18,19). (ii) The motifs were grouped with hierarchical clustering, using the standard UPGMA method (http://arxiv.org/abs/1105.0121). (iii) The hierarchical tree was cut using as the threshold a combination of different metrics values; the tree is cut in a collection of trees (a forest). (iv) Each tree is used as a guide to create a progressive alignment of the PWMs. (v) The clusters are represented both as trees and as heat maps (see http://www.rsat.eu/).

The browser that enables the user to see the collection of PWMs in a hierarchical tree is available via the 'Integrated Views & Tools' menu, in the 'Browse RegulonDB' section in the 'Clustering of RegulonDB PWMs' option. Also in the same section, there is a link to display a circular browser (Figure 2), developed with the D3.js JavaScript library (http://d3js.org/), that integrates the information from families, the TFs and their PWMs.

IMPLEMENTING A SEMIASSISTED CURATION STRATEGY

Given the large amount of biological data generated day by day as a result of research in various laboratories, manual curation represents a bottleneck to facilitate access to knowledge in an organized way. We therefore have initiated the implementation of NLP methods in collaboration with the OntoGene group, to enhance the efficiency of curation to keep up with the flood of knowledge and publications, as reported recently (20).

We developed an *ad hoc* interface called ODIN (The OntoGene Document INspector) to curate the literature supporting the knowledge in RegulonDB. The input for ODIN entails full papers, and the output is an interface with several tools to facilitate their curation. We have initiated the process of assisted, or semiautomatic, curation in a very cautious manner, focusing on missing pieces of knowledge, such as growth conditions (GCs) under which specific regulatory interactions (RIs) have been identified.

To do so, filters were created that display in ODIN only those sentences in a paper that contain the data we need to curate; therefore, we do not have to read the full article, but only the phrases that should contain the RIs and GCs. The data we have curated in the traditional way, as is the case with RIs, serve as a control to benchmark this new method. In the case of OxyR, we identified all 20 RIs (100%) that had been previously curated, and we identified the GCs for 16 of them (20).

After reporting the work of OxyR, we used the same strategy to identify the GCs of SoxR and SoxS RIs; we identified 27 of the 28 (96%) RIs of SoxS and obtained the GCs for 13 of them. This lower number may be due to non-specified growth conditions reported in the papers, such as when performing *in vitro* experiments for overexpression of TFs. We also identified 3 out of 3 RIs of SoxR and the GC for 2 of them (see Table S1 in supplementary material). Therefore, now we have in RegulonDB the GCs for 31 RIs, including the OxyR GC-RI pairs, and we will continue to work with other TFs to identify their GC-RI pairs. We will use these results for a cyclic improvement of our assisted curation strategies.

This semiautomatic process enables us to increase the efficiency of curation; however, currently there is a total of 3195 RIs for 199 TFs. This shows the long way we have to go in order to curate all GCs for the RIs. These numbers make clear the necessity for implementation of an assisted curation strategy. We are also motivated to implement NLP filters, not only for new properties but also for a more precise and comprehensive curation, as in the case for methods associated with evidence codes.

Identification and annotation of methods

All data added to the database contain the evidence that demonstrates the existence of each object or interaction of regulation. We use a set of evidence codes (see http://regulondb.ccg.unam.mx/evidenceclasification) for this purpose. These evidence codes are derived from more than one method that is reported in the literature, as in the case of the evidence to identify TSSs of promoters, 'Transcription initiation mapping,' that could be related to the methods of primer extension, S1 mapping or 5'-RACE, among others. This level of description is common to major resources, such as the GO (Gene Ontology) and EcoCyc databases.

A virtue within RegulonDB that we implemented since 2008 is, on the one hand, a simplified classification of 'weak' and 'strong' levels of confidence for all evidence sources. Strong confidence essentially requires physical evidence for the existence of the object or interaction (21). What is most interesting is the specific algebra of the combinations of evidence that can be considered independent from each other and therefore can be added to increase the overall degree of confidence for an object or interaction, including cross-validation of strong confidence supporting the 'confirmed' level of confidence (22).

Motivated by this previous work, we started a project with the group OntoGene to extract the methods for all objects and interactions contained in RegulonDB. We initiated the project by identifying the experimental methods of primer extension and northern blotting. These methods are the most easy to identify by text-mining methods, because very specific words are used to describe them, as opposed to other methods. Northern blotting is commonly used to identify TUs, whereas primer extension is used to identify TSSs of promoters. The objects that were identified with these methods are listed in RegulonDB with strong evidence codes. We took all the papers that are linked to only one promoter or one TU in RegulonDB, where the promoter or TU had the evidence code for 'Transcription initiation mapping' and 'Length of transcript experimentally determined,' respectively. Subsequently, using ODIN filters, we did a search for the words 'primer extension' and 'northern blot' in each set of papers. To increase the confidence in these text-mining strategies, we included the requirement that in the phrase(s) that identifies the method, the name

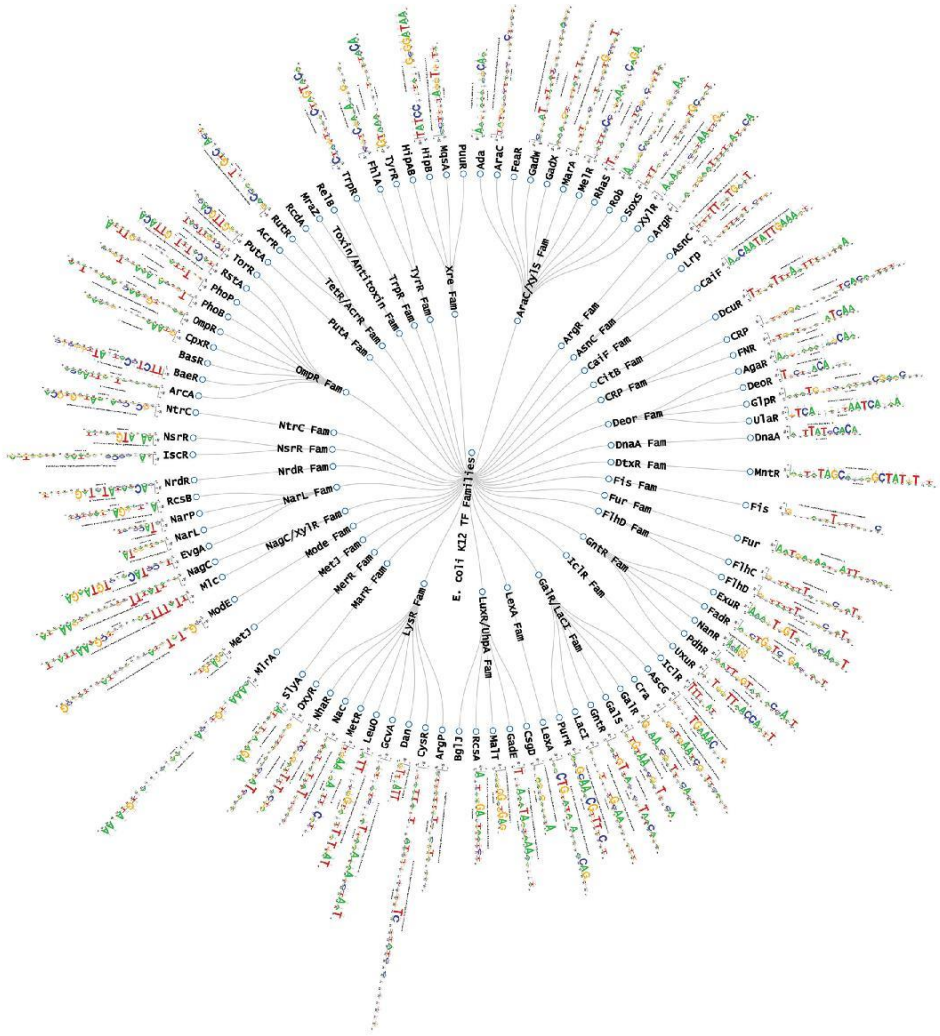

Figure 2. Circular Browser. Transcription factors (TF) classified in their evolutionary families, based on PFAM, CDD and Superfamily annotations.

of the TU/promoter has to be present too. Of course, we read all these phrases and checked if they were correct. We identified the method of primer extension for 227 promoters and that of northern blotting for 110 TUs. We plan to expand these strategies to extract additional methods and other objects and knowledge from the literature.

CURATION OF HIGH-THROUGHPUT DATASETS

HT experiments generate a large number of scattered fragments of knowledge. By default, we add such information, with peaks already processed by the authors of the curated papers, as datasets separated from the database, but available, for instance, for display as tracks as part of the different resources in RegulonDB. Our manual curation efforts are focused on extracting the subset of objects (binding sites and promoters) that have additional evidence supporting them, as well as to combine different experiments that congruently support an object. For instance, ChIP-based experiments identify sites that occur within coding regions (23), which may have nothing to do with transcriptional regulation or at least there is not yet evidence of such involvement. Similar concerns may be raised with TSSs identification.

In some cases, a subset of the results is subject to further analysis, such as EMSA, footprinting, northern blotting and/or matrix analyses for site identification. Currently, we add a regulatory interaction in the database only for sites with strong evidence for TFBSs-validated sites (22) and where additional knowledge assigns the function of the TF on the regulated gene, such as ChIP-exo complemented with RNA-seq analysis (24,25). This illustrates the rationale of our approach to combine the data of different HT experiments to integrate these new data with existing knowledge in the database.

There are more than 38 transcriptional regulators (TFs) whose sites have been identified by ChIP methodologies, and this number may increase to 200 TFs for which genomic SELEX screening has been done (26), with data available and published for 17 TFs. A summary of the currently curated datasets is shown in Supplementary Tables S2 and S3. Extraction and curation of this type of data is particularly difficult, because the generated information shows a great variety in terms of formats of the results, and only the central peaks around the binding site for a TF are shown (e.g. 200–300 nt).

There are several HT-dedicated repositories and resources of microbial experimental results, such as TBDB (27), CollectTF (28) and RegTransBase (29), which hold curated motif data from HTs sources with links to GEO from NCBI (30) and to ArrayExpress from EMBL-EBI (31); databases with expression profiles, such as COLOMBOS (32), which offers a variety of tools for analysis, M3D (33) and GenExpDB (http://genexpdb.ou.edu/main/). For a broader context of these resources and many more related to gene regulation, users can visit our link to additional resources (http://regulondb.ccg.unam.mx/menu/about_regulondb/additional_resources/index.jsp).

Our work does not duplicate those efforts, since our main goal is to detect evidence that can be added to either existing objects in RegulonDB and/or that can be combined to support knowledge of higher granularity.

HT data generated by gSELEX and ChIP-exo

We curated in RegulonDB ChIP-chip data for the PurR regulon by using validation data (22). For gSELEX, we generated two tables, one with raw data (data gSELEX peaks) and the other for cross-data comparisons, i.e., gSELEX and microarrays data for H-NS and LeuO (34); gSELEX and consensus sequences for the transcription factor CRP (35). Only a few cases were uploaded to the database from ChIP-exo plus RNA-seq analysis (24,25). The results generated for both methodologies are summarized in Supplementary Tables S2 and S3.

HT dataset for TSSs under three conditions

The dataset for 14 868 TSSs from the Storz lab has been curated. It includes 5495 TSSs corresponding to potential antisense RNAs (asRNAs). These data were generated from RNA-seq and prediction algorithms under three different biological conditions: the MG1655 wild-type strain grown to exponential phase or stationary phase in LB medium as well as the wild-type strain grown to exponential phase in M63 minimal glucose medium (36) (see also http://regulondb.ccg.unam.mx/menu/download/high_throughput_datasets/index.jsp).

COEXPRESSION DISTANCE AROUND THE REGULATORY NETWORK

One of the extensive uses of HT technologies is for the development of global expression profiles. As mentioned before, dedicated databases with information on *E. coli* include COLOMBOS (32), M3D (33) and GenExpDB (http://genexpdb.ou.edu/main/).

For years, RegulonDB has offered links that allow users to upload gene sets to search for their expression profiles in COLOMBOS (www.colombos.net). In addition to these links, we have implemented tools for a full comparison of expression of groups of genes across all conditions.

The 'Coexpression' page can be reached directly from the search option. A single query gene or a group of genes are added either manually, based on the set of interest to the user, or are automatically uploaded as a collection of genes defining operons or regulons (from their corresponding pages). The result will be a list of the top 20 genes (the default quantity) that have the highest similarity in coexpression, from the set of all experiments present in COLOMBOS across all conditions. There is a single best list for each one of the genes in the input list, which can be browsed on the 'Coexpression' page. These lists include relevant information for the input genes, i.e. the gene product name, the operon to which the gene belongs, the regulators for which the gene has binding sites, and ontological classes of processes in which the gene participates. In the next release of RegulonDB, in an additional section, we will show coexpression by providing color charts to facilitate visualization.

In addition, in the most recent RegulonDB release, we offer a coexpression overview for two groups of input genes: operons and regulons. For dual regulators, regulons are also separated into what we call 'strict regulons,' that is to say,

groups of target genes subject to the same effect (activator, repressor or dual effect) by a TF. For each group, i.e. each operon, regulon or strict regulon, we display a browser containing the following sections: the name of the group, the genes contained in the group, a 'coexpression matrix,' the 'coexpression distribution' of the group, and the 'top best coexpressed' genes. The 'coexpression matrix' section enables the user to see the coexpression values of genes with other genes within the group, and the 'coexpression distribution' section shows a plot of the probability density distribution of the coexpression values of the genes within the group, contrasted with a background. For example, the coexpression distribution of the strict regulon 'CRP,+' shows the coexpression probability density distribution of all genes activated by CRP with each other, in contrast with the coexpression probability density distribution of all remaining genes in the genome. The 'top best coexpressed' section offers a list of additional genes that show the highest coexpression with the genes in the group.

Additional genes that are most highly coexpressed with a group of genes are identified by calculating the top best-scoring medians of the set of coexpression values of any additional gene with each gene of the group. This is certainly an interesting question for any set of query genes, but it is computationally intensive, since for every pair of input genes, we need to identify the intersection of output coexpressed genes. We have therefore precalculated the group values for operons and regulons.

To quantify coexpression for all combinations of gene pairs, we implemented a rank-based approach, using data available in COLOMBOS version 2.0, which contains expression profiles of 2470 different, contrasting conditions. The method and results will be described in detail in a paper to be submitted by Pannier *et al.* Gene coexpression is typically quantified by pairwise correlation analysis across large expression compendia. However, these analyses are difficult to interpret because of the highly variable distributions of such correlation coefficients. We use a rank-based approach that normalizes the differences between the range of correlation coefficients between genes, which allows comparisons of coexpression strengths among genes despite the large variability of expression values.

COMPUTATIONAL ADDITIONS

In order to facilitate searching for information, we implemented a free word-searching tool based in Elastic Search (https://www.elastic.co). This tool enables identification of synonyms for any object, in order by relevance and highlights the searched elements.

New features of our website

Search results. A new view was added to the display of search results by regulon, at the request of our users. When the user selects 'regulon search' without giving a term, all the regulons are displayed in a table with the regulon name, the total regulated genes, the total regulated operons, the total binding sites and the total regulatory interactions. The user can sort using any column of this table.

Gene page. We created a new section named 'Elements in the selected gene context region unrelated to any TU in RegulonDB.' In this section, users can find biological objects in the vicinity of a gene that are not part of its TU, such as the many TSSs near the *micF* gene, to mention one example. In addition, the same gene page in the section called 'Operon arrangement' has links to the operon page. Each promoter is linked with the corresponding TU that it transcribes.

Sigmulon page. We have included the sigma signal transduction map with a link showing the details of the reactions contained in the map.

Datasets. In the submenu related to the datasets, included in downloads, we have integrated new information related to the TSSs experimentally determined in the laboratory of Dr Morett. The TSSs are included in the file named 'High-throughput transcription initiation mapping. Illumina directional RNA-seq experiments where total RNA received different treatments to enrich for 5′-monophosphate or 5′-triphosphate ends.' These objects are included in the new section 'Elements in the selected gene context region unrelated to any TU in RegulonDB,' previously described.

Impact of RegulonDB. Figure 3 shows the accumulated citations for each RegulonDB paper by year and the concomitant expansion of new objects and properties related to the regulation of gene expression that we curate. RegulonDB plays a central role in the development and testing of novel approaches of gene regulation in bioinformatics, comparative genomics, and systems biology, and it is the model to inspire similar approaches and studies for any other organism, including pathogenic bacteria (37–39). Evidence of its usefulness is apparent from the more than 1200 citations in published articles, in addition to the many citations for the EcoCyc database, which incorporates our curation work. Within the 'Features' menu, we have added this type of information, showing the impact of RegulonDB, such as the number and type of journals for publications that have cited our RegulonDB-related publications.

Releases. The release that corresponds to this paper is version 9.0. Major changes to the overall navigation and structure of the main pages have been made, offering more structured access to the data, based on the two dominant types of users: biologists, usually conducting individual search queries, and those interested in data collections.

CONCLUSIONS

RegulonDB is a complex evolving system. As mentioned in the 'Introduction' section, through the years we have gradually expanded both the content and level of detail of the biological data as well as the diverse types of experimental and bioinformatics sources of knowledge that nurture our understanding of gene regulation in *E.coli* K-12. A simplified diagram of our work is shown in Figure 4, with a triangle representing integration of data, information and knowledge; we do not mean absolute definitions of what is data or information, but simply provide relative concepts of the hierarchical nature of a highly granular knowledge. In the

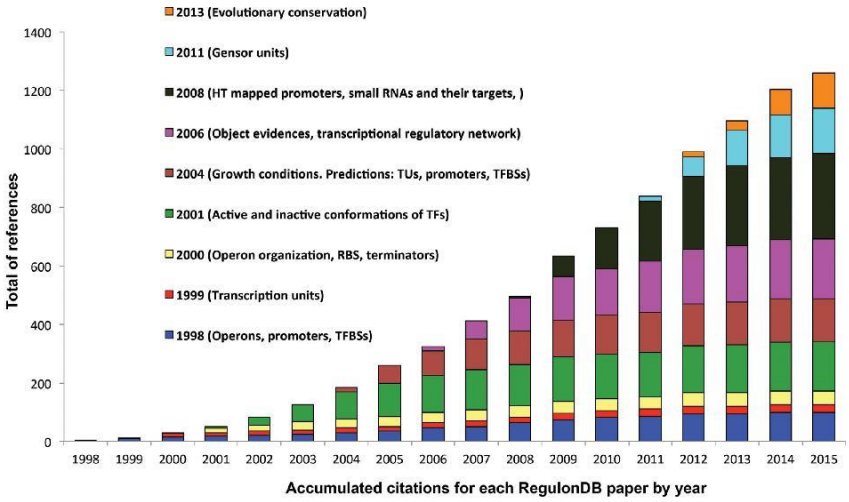

Figure 3. Impact of RegulonDB. Accumulated citations for each RegulonDB paper by year and the concomitant expansion of domains of the biology that we curate.

following discussion, we try to locate the major advances reported in this paper in this context.

For instance, the tsunami of HT-generated data is an effort that appears at the bottom, where decisions of what to represent where (datasets versus integrated sites in the database) reflect the tension of a classic paradigm of a reasonably well-organized genome, as opposed to one inundated by promoters and binding sites of unknown function.

Since the challenge of encoding this flux of knowledge is occurring faster than our human abilities and resources to keep our work up to date, we need to develop novel strategies. The assisted curation by means of NLP methodologies illustrates our efforts to implement strategies and test and improve them to accelerate our curation work. A direct benefit in years to come will be to have curated all specific contrasting conditions for each regulatory switch. Expansions to our work include the coexpression comparative metrics, enhanced collection of sRNAs regulation and clustering of PWMs and their grouping into TF evolutionary families. A landmark for this publication is the clear progress of the comprehensive collection of GENSOR units in an effort to enrich the top of the pyramid via overviews agglutinating large amounts of information that should make sense as a unit.

Briefly, we are guided by electronically editing in a structured way from the low granularity of details of mechanisms to the higher granularity regarding description and abstraction that offer broader perspectives to our understanding of the machinery and processes of *E. coli*'s way of life.

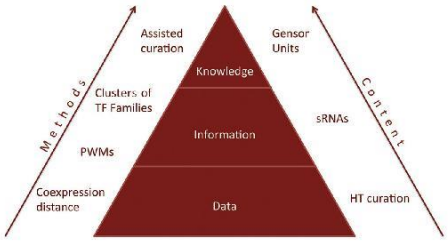

Figure 4. Schema of types of methods and content in RegulonDB.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

We acknowledge Alfredo Hernández for technical support, Diego Z. Palomares Lagunas for his contribution to the development of the GENSOR units webpage, Gerardo Salgado for administrative help, Mishael Sánchez for the processing of HT datasets, Ingrid Keseler for periodically sending us selected literature for curation, José M. Camacho Zaragoza, María Cecilia Ishida Gutierrez and Sara B. Martínez Luna for their contributions for curation of GENSOR Units, and finally Gustavo Engstrom, Víctor Bustamante, Yalbi Balderas, David A. Velázquez-Ramírez and Citlalli Mejía for their participation in the prototype evaluation of coexpression.

FUNDING

Universidad Nacional Autónoma de México; National Institute of General Medical Sciences of the National Institutes of Health [R01GM110597]; Department of Health and Human Services; National Institutes of Health; National Institute of General Medical Sciences, under Cooperative Agreement [2U24GM077678-24A1]. Funding for open access charge: National Institutes of Health [R01GM110597].
Conflict of interest statement. None declared.

REFERENCES

1. Salgado,H., Peralta-Gil,M., Gama-Castro,S., Santos-Zavaleta,A., Muniz-Rascado,L., Garcia-Sotelo,J.S., Weiss,V., Solano-Lira,H., Martinez-Flores,I., Medina-Rivera,A. *et al.* (2013) RegulonDB v8.0: omics data sets, evolutionary conservation, regulatory phrases, cross-validated gold standards and more. *Nucleic Acids Res.*, **41**, D203–D213.
2. Huerta,A.M., Salgado,H., Thieffry,D. and Collado-Vides,J. (1998) RegulonDB: a database on transcriptional regulation in *Escherichia coli*. *Nucleic Acids Res.*, **26**, 55–59.
3. Keseler,I.M., Mackie,A., Peralta-Gil,M., Santos-Zavaleta,A., Gama-Castro,S., Bonavides-Martinez,C., Fulcher,C., Huerta,A.M., Kothari,A., Krummenacker,M. *et al.* (2013) EcoCyc: fusing model organism databases with systems biology. *Nucleic Acids Res.*, **41**, D605–D612.
4. Collado-Vides,J. (1992) Grammatical model of the regulation of gene expression. *Proc Natl. Acad. Sci. U.S.A.*, **89**, 9405–9409.
5. Gama-Castro,S., Salgado,H., Peralta-Gil,M., Santos-Zavaleta,A., Muniz-Rascado,L., Solano-Lira,H., Jimenez-Jacinto,V., Weiss,V., Garcia-Sotelo,J.S., Lopez-Fuentes,A. *et al.* (2011) RegulonDB version 7.0: transcriptional regulation of *Escherichia coli* K-12 integrated within genetic sensory response units (Gensor Units). *Nucleic Acids Res.*, **39**, D98–D105.
6. Martinez-Antonio,A. and Collado-Vides,J. (2003) Identifying global regulators in transcriptional regulatory networks in bacteria. *Curr. Opin. Microbiol.*, **6**, 482–489.
7. Funahashi,A., Matsuoka,Y., Jouraku,A., Morohashi,M., Kikuchi,N. and Kitano,H. (2008) CellDesigner 3.5: a versatile modeling tool for biochemical networks. *Proc. IEEE*, **96**, 1254–1265.
8. Kitano,H., Funahashi,A., Matsuoka,Y. and Oda,K. (2005) Using process diagrams for the graphical representation of biological networks. *Nat. Biotechnol.*, **23**, 961–966.
9. Perez-Rueda,E., Tenorio-Salgado,S., Huerta-Saquero,A., Balderas-Martinez,Y.I. and Moreno-Hagelsieb,G. (2015) The functional landscape bound to the transcription factors of *Escherichia coli* K-12. *Comput. Biol. Chem.*, **58**, 93–103.
10. Pedruzzi,I., Rivoire,C., Auchincloss,A.H., Coudert,E., Keller,G., de Castro,E., Baratin,D., Cuche,B.A., Bougueleret,L., Poux,S. *et al.* (2015) HAMAP in 2015: updates to the protein family classification and annotation system. *Nucleic Acids Res.*, **43**, D1064–D1070.
11. Oates,M.E., Stahlhacke,J., Vavoulis,D.V., Smithers,B., Rackham,O.J., Sardar,A.J., Zaucha,J., Thurlby,N., Fang,H. and Gough,J. (2015) The SUPERFAMILY 1.75 database in 2014: a doubling of data. *Nucleic Acids Res.*, **43**, D227–D233.
12. Finn,R.D., Bateman,A., Clements,J., Coggill,P., Eberhardt,R.Y., Eddy,S.R., Heger,A., Hetherington,K., Holm,L., Mistry,J. *et al.* (2014) Pfam: the protein families database. *Nucleic Acids Res.*, **42**, D222–D230.
13. Marchler-Bauer,A., Anderson,J.B., Derbyshire,M.K., DeWeese-Scott,C., Gonzales,N.R., Gwadz,M., Hao,L., He,S., Hurwitz,D.I., Jackson,J.D. *et al.* (2007) CDD: a conserved domain database for interactive domain family analysis. *Nucleic Acids Res.*, **35**, D237–D240.
14. Altschul,S.F., Madden,T.L., Schaffer,A.A., Zhang,J., Zhang,Z., Miller,W. and Lipman,D.J. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.*, **25**, 3389–3402.
15. Medina-Rivera,A., Abreu-Goodger,C., Thomas-Chollier,M., Salgado,H., Collado-Vides,J. and van Helden,J. (2011) Theoretical and empirical quality assessment of transcription factor-binding motifs. *Nucleic Acids Res.*, **39**, 808–824.
16. Medina-Rivera,A., Defrance,M., Sand,O., Herrmann,C., Castro-Mondragon,J.A., Delerce,J., Jaeger,S., Blanchet,C., Vincens,P., Caron,C. *et al.* (2015) RSAT 2015: Regulatory Sequence Analysis Tools. *Nucleic Acids Res.*, **43**, W50–W56.
17. van Helden,J. (2003) Regulatory sequence analysis tools. *Nucleic Acids Res.*, **31**, 3593–3596.
18. Thomas-Chollier,M., Defrance,M., Medina-Rivera,A., Sand,O., Herrmann,C., Thieffry,D. and van Helden,J. (2011) RSAT 2011: regulatory sequence analysis tools. *Nucleic Acids Res.*, **39**, W86–W91.
19. Tanaka,E., Bailey,T., Grant,C.E., Noble,W.S. and Keich,U. (2011) Improved similarity scores for comparing motifs. *Bioinformatics*, **27**, 1603–1609.
20. Gama-Castro,S., Rinaldi,F., Lopez-Fuentes,A., Balderas-Martinez,Y.I., Clematide,S., Ellendorff,T.R., Santos-Zavaleta,A., Marques-Madeira,H. and Collado-Vides,J. (2014) Assisted curation of regulatory interactions and growth conditions of OxyR in *E. coli* K-12. *Database (Oxford)*, **2014**.
21. Gama-Castro,S., Jimenez-Jacinto,V., Peralta-Gil,M., Santos-Zavaleta,A., Penaloza-Spinola,M.I., Contreras-Moreira,B., Segura-Salazar,J., Muniz-Rascado,L., Martinez-Flores,I., Salgado,H. *et al.* (2008) RegulonDB (version 6.0): gene regulation model of *Escherichia coli* K-12 beyond transcription, active (experimental) annotated promoters and Textpresso navigation. *Nucleic Acids Res.*, **36**, D120–D124.
22. Weiss,V., Medina-Rivera,A., Huerta,A.M., Santos-Zavaleta,A., Salgado,H., Morett,E. and Collado-Vides,J. (2013) Evidence classification of high-throughput protocols and confidence integration in RegulonDB. *Database (Oxford)*, **2013**, bas059.
23. Grainger,D.C., Hurd,D., Harrison,M., Holdstock,J. and Busby,S.J. (2005) Studies of the distribution of Escherichia coli cAMP-receptor protein and RNA polymerase along the *E. coli* chromosome. *Proc. Natl. Acad. Sci. U.S.A.*, **102**, 17693–17698.
24. Seo,S.W., Kim,D., O'Brien,E.J., Szubin,R. and Palsson,B.O. (2015) Decoding genome-wide GadEWX-transcriptional regulatory networks reveals multifaceted cellular responses to acid stress in *Escherichia coli*. *Nat. Commun.*, **6**, 7970.
25. Seo,S.W., Kim,D., Szubin,R. and Palsson,B.O. (2015) Genome-wide Reconstruction of OxyR and SoxRS Transcriptional Regulatory Networks under Oxidative Stress in *Escherichia coli* K-12 MG1655. *Cell Rep.*, **12**, 1289–1299.
26. Ishihama,A. (2012) Prokaryotic genome regulation: a revolutionary paradigm. *Proc. Jpn. Acad. Ser. B Phys. Biol. Sci.*, **88**, 485–508.
27. Galagan,J.E., Sisk,P., Stolte,C., Weiner,B., Koehrsen,M., Wymore,F., Reddy,T.B., Zucker,J.D., Engels,R., Gellesch,M. *et al.* (2010) TB database 2010: overview and update. *Tuberculosis (Edinb)*, **90**, 225–235.
28. Kilic,S., White,E.R., Sagitova,D.M., Cornish,J.P. and Erill,I. (2014) CollecTF: a database of experimentally validated transcription factor-binding sites in Bacteria. *Nucleic Acids Res.*, **42**, D156–D160.
29. Kazakov,A.E., Cipriano,M.J., Novichkov,P.S., Minovitsky,S., Vinogradov,D.V., Arkin,A., Mironov,A.A., Gelfand,M.S. and Dubchak,I. (2007) RegTransBase–a database of regulatory sequences and interactions in a wide range of prokaryotic genomes. *Nucleic Acids Res.*, **35**, D407–D412.
30. Barrett,T., Wilhite,S.E., Ledoux,P., Evangelista,C., Kim,I.F., Tomashevsky,M., Marshall,K.A., Phillippy,K.H., Sherman,P.M., Holko,M. *et al.* (2013) NCBI GEO: archive for functional genomics data sets–update. *Nucleic Acids Res.*, **41**, D991–D995.
31. Kolesnikov,N., Hastings,E., Keays,M., Melnichuk,O., Tang,Y.A., Williams,E., Dylag,M., Kurbatova,N., Brandizi,M., Burdett,T. *et al.* (2015) ArrayExpress update–simplifying data submissions. *Nucleic Acids Res.*, **43**, D1113–D1116.
32. Meysman,P., Sonego,P., Bianco,L., Fu,Q., Ledezma-Tejeida,D., Gama-Castro,S., Liebens,V., Michiels,J., Laukens,K., Marchal,K. *et al.* (2014) COLOMBOS v2.0: an ever expanding collection of bacterial expression compendia. *Nucleic Acids Res.*, **42**, D649–D653.
33. Faith,J.J., Driscoll,M.E., Fusaro,V.A., Cosgrove,E.J., Hayete,B., Juhn,F.S., Schneider,S.J. and Gardner,T.S. (2008) Many Microbe Microarrays Database: uniformly normalized Affymetrix compendia with structured experimental metadata. *Nucleic Acids Res.*, **36**, D866–D870.

34. Shimada,T., Bridier,A., Briandet,R. and Ishihama,A. (2011) Novel roles of LeuO in transcription regulation of *E. coli* genome: antagonistic interplay with the universal silencer H-NS. *Mol. Microbiol.*, **82**, 378–397.
35. Shimada,T., Fujita,N., Yamamoto,K. and Ishihama,A. (2011) Novel roles of cAMP receptor protein (CRP) in regulation of transport and metabolism of carbon sources. *PLoS One*, **6**, e20081.
36. Thomason,M.K., Bischler,T., Eisenbart,S.K., Forstner,K.U., Zhang,A., Herbig,A., Nieselt,K., Sharma,C.M. and Storz,G. (2015) Global transcriptional start site mapping using differential RNA sequencing reveals novel antisense RNAs in *Escherichia coli*. *J. Bacteriol.*, **197**, 18–28.
37. Shen-Orr,S.S., Milo,R., Mangan,S. and Alon,U. (2002) Network motifs in the transcriptional regulation network of *Escherichia coli*. *Nat. Genet.*, **31**, 64–68.
38. Barabasi,A.L. and Oltvai,Z.N. (2004) Network biology: understanding the cell's functional organization. *Nat. Rev. Genet.*, **5**, 101–113.
39. Isalan,M., Lemerle,C., Michalodimitrakis,K., Horn,C., Beltrao,P., Raineri,E., Garriga-Canut,M. and Serrano,L. (2008) Evolvability and hierarchy in rewired bacterial gene networks. *Nature*, **452**, 840–845.

CAPÍTULO 3: PERSPECTIVAS

3.1 Perspectivas del análisis genómico

Perspectiva general:

En estudios futuros de la corregulación transciptional será importante considerar que **la cercanía genómica y la corregulación tienen un efecto sinérgico sobre los niveles de coexpresión**.

Perspectivas especificas:

- ✓ La combinación de cercanía genómica, corregulación (predicha) y coexpresión alta ofrece una opción de validar predicciones para ligas en las redes de regulación transcripcional.
- ✓ En cuanto se vaya completando la información sobre la red de regulación transcripcional en *E. coli* K-12, al rehacer este estudio se tendrá un conocimiento robusto sobre el efecto de cercanía genómica en el control coordinado de expresión génica.
- ✓ Es factible que en el futuro se obtenga información de distancias 3D dinámicas entre los genes, por lo cual este análisis se podrá realizar de forma más completa ya que será conforme al contexto genómico real de los genes.
- ✓ Al estudiar el efecto de la distancia en la coexpresión de genes corregulados, habrá que tener en cuenta que la distancia tiene un papel *particularmente* importante en genes con una corregulación no estricta (i.e. cuando el sobrelape de sus programas de regulación es pequeño relativo a la parte de sus programas de regulación que no tienen en común).
- ✓ En este trabajo encontramos indicaciones que la cercanía de genes corregulados implica coexpresión alta por que a distancias pequeñas, los genes blancos de un FT común estarían expuestos a cantidades de proteínas de ese FT *similares*. Para poder validar estos resultados, será necesario realizar un estudio experimental para determinar la localización dinámica (en el tiempo y en el espacio) del FT, por ejemplo, visualizando la difusión de las proteínas FT en el tiempo agregándoles una etiqueta fluorescente.
- ✓ El mismo análisis se podrá realizar en otros genomas microbianos con redes de regulación conocida.
- ✓ Hay varias indicaciones de que la corregulación, la coexpresión y la vecindad de genes, aumentan la probabilidad de que esos genes compartan una o múltiples funciones metabólicas (Michalak 2008). Entonces, la integración combinada de vecindad y coexpresión podría aumentar el poder predictivo de herramientas que permitan construir redes regulatorias, metabólicas u otras redes funcionales.
- ✓ Este estudio constituye un paso en el camino de entender los patrones de cómo la distancia genómica entre los genes participa en controlar la coexpresión génica a través de la corregulación transcripcional. Con este tipo de análisis, poco a poco podremos anticipar cómo los sistemas biológicos y su

actividad genética se comportarán, tanto en gammaproteobacterias, en otras familias de procariontes, o incluso en eucariontes.

3.2 Perspectivas de la metodología

Hemos propuesto una nueva métrica de coexpresión llamada *Spearman Correlation Rank* (SCR) que fue inspirada en la métrica propuesta por Obayashi y collegas (Obayashi *et al.* 2011) (ver el primer articulo, sección 3.1, para una descripción). SCR tiene dos ventajas sobre las medidas de coexpresión que hoy en día se utilizan comúnmente (la más común siendo *Pearson Correlation Coefficient*): que es intuitiva, y que es comparable entre genes a nivel genómico. Es decir, la SCR de un par de genes de interés es intuitiva y fácil de interpretar: un SCR de 1 para un par de genes significa que su nivel de coexpresión es mas alto que el nivel de coexpresión de ellos con todos los demás genes, un SCR de 2 el segundo mas alto, etc. Ademas, la SCR es fácil de comparar entre genes por que es una medida directa de la significancia de su coexpresión de este par de genes *relativamente* a los SCR de este par con los demás pares de genes.

Por estas razones, consideramos la SCR como métrica comprensiva de coexpresión apta para análisis de coexpresión, sobre todo los de escala genómica.

REFERENCIAS

Beck, C.F. & Warren, R.A., 1988. Divergent promoters, a common form of gene organization. *Microbiological Reviews*, 52(3), pp.318–326. Available at: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC373147/.

Gama-Castro, S. et al., 2015. RegulonDB version 9.0: high-level integration of gene regulation, coexpression, motif clustering and beyond. *Nucleic Acids Research* . Available at: http://nar.oxfordjournals.org/content/early/2015/11/01/nar.gkv1156.abstract.

Janga, S.C., Salgado, H. & Martínez-Antonio, A., 2009. Transcriptional regulation shapes the organization of genes on bacterial chromosomes. *Nucleic Acids Research*, 37(11), pp.3680–3688. Available at: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2699516/.

Korbel, J.O. et al., 2004. Analysis of genomic context: prediction of functional associations from conserved bidirectionally transcribed gene pairs. *Nat Biotech*, 22(7), pp.911–917. Available at: http://dx.doi.org/10.1038/nbt988.

Lemmens, K. et al., 2009. DISTILLER: a data integration framework to reveal condition dependency of complex regulons in Escherichia coli. *Genome Biology*, 10(3), pp.R27–R27. Available at: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2690998/.

Michalak, P., 2008. Coexpression, coregulation, and cofunctionality of neighboring genes in eukaryotic genomes. *Genomics*, 91(3), pp.243–8. Available at: http://www.sciencedirect.com/science/article/pii/S0888754307002807.

Michoel, T. et al., 2009. Comparative analysis of module-based versus direct methods for reverse-engineering transcriptional regulatory networks. *BMC Systems Biology*, 3, p.49. Available at: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2684101/.

Moretto, M. et al., 2016. COLOMBOS v3.0: leveraging gene expression compendia for cross-species analyses. *Nucleic Acids Research*, 44(Database issue), pp.D620–D623. Available at: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC4702885/.

Rhee, K.Y. et al., 1999. Transcriptional coupling between the divergent promoters of a prototypic LysR-type regulatory system, the ilvYC operon of Escherichia coli. *Proceedings of the National Academy of Sciences of the United States of America*, 96(25), pp.14294–14299. Available at: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC24430/.

Schneider, R. & Grosschedl, R., 2007. Dynamics and interplay of nuclear architecture, genome organization, and gene expression. *Genes and Development*, 21(23), pp.3027–3043.

Zampieri, M. et al., 2008. Origin of Co-Expression Patterns in E.coli and S.cerevisiae Emerging from Reverse Engineering Algorithms M. Isalan, ed. *PLoS ONE*, 3(8),

p.e2981. Available at: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2500178/.

Zhang, H. et al., 2012. Genomic Arrangement of Regulons in Bacterial Genomes J. H. Badger, ed. *PLoS ONE*, 7(1), p.e29496. Available at: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3250446/.

APÉNDICE

Material supplementario del primer articulo

S1 File. *This file contains the following sections:*

1. *Selection of a similarity measure to quantify coexpression*
2. *Rank-based similarity measures compensate for conditional dependency*
3. *The degree of coexpression of genes that are coregulated in E. coli is generally low*
4. *Evidence classification of interactions in RegulonDB*
5. *Evidence of horizontal gene co-transfer in genes with strong distance conservation*

1. Selection of a similarity measure to quantify coexpression

To select the similarity measure that best captured the degree of coexpression between two coregulated genes in our setting, we compared six similarity measures, i.e. three similarity measures commonly used to quantify coexpression, the Pearson Correlation Coefficient (PCC), Spearman Correlation Coefficient (SCC) and Mutual Information (MI), and their rank-based derivatives, which we defined as respectively the Pearson Correlation Rank (PCR), Spearman Correlation Rank (SCR) and Mutual Information Rank (MIR) (the calculation of these measures was explained in Materials and Methods).

As a benchmark, we used genes located within the same operon (using all combinations of genes within the same operon according to the operon set of RegulonDB [1]), as these are expected to be highly coexpressed. To exclude the effect of regulatory elements within operons, we only considered in the benchmark pairs of operonic genes that are contiguous and that are not separated from each other by an internal promoter or terminator.

This resulted in a positive set of 602 gene pairs (N = 602) which were expected to be highly coexpressed. These were referred to as constituting the True Positive (TP) set. As a negative control, 10000 random gene pairs were sampled and referred to as True Negative (TN) set. For both the positive and negative set, we calculated the PCC, SCC, MI, PCR, SCR and MIR across all conditions in COLOMBOS.

As an illustration, Fig A shows the frequency distributions, i.e. the number of gene pairs, of respectively the TP and TN sets that were coexpressed within a given range of the SCC and within a given range of SCR (its rank-based derivative). SCR values of contiguous operonic genes are localized at the utter left of the SCR distribution which is the most significant region, while the SCR distribution of the TN set (genes in random pairs) is uniform.

In contrast, for the SCC the majority of TP pairs have a degree of coexpression that ranges from approximately 0.1 to approximately 0.7. In contrast to what is observed for the SCR, for the SCC the majority of TP seem to cover a large range of values: 86% of TP gene pairs have an SCC within the interval [0.25, 0.75], which covers 50% of the full positive range of SCC (positive range is the range that looks at correlation and not at anticorrelation i.e. [0-1]), whereas 86% of TP gene pairs have an SCR within the interval [1, 30], which covers only 7% of the full range of SCR.

This means that according to the SCR the majority of TP gene pairs are highly coexpressed, whereas when assessing the coexpression with the SCC it is intuitively more difficult to interpret whether the true positives have a relatively high or low coexpression degree.

This was formally confirmed by assessing the performance of each measure for its ability to classify TP (within operonic genes) and TN (random gene pairs) based on their degree of coexpression. This ability was quantified by calculating the Area Under the Curve (AUC) from the ROC curve (Table A). The AUC equals the probability that a classifier will rank a randomly chosen positive instance higher than a randomly chosen negative one. In other words, the higher the AUC is, the better the measure is able to separate TP pairs from TN pairs based on their differences in coexpression behaviour.

In Table A the Area Under the Curve (AUC) is given for each of the six tested similarity measures as a quantification of how well each measure distinguishes between the TP and TN gene pairs. The highest AUC (0.998) corresponded to SCR using the corresponding distributions of the TP and TN pairs. This implies that TP and TN can be best separated using their coexpression behaviour measured by SCR.

Overall, because SCR a) performs best in distinguishing TP from TN (Table A), b) provides a measure of coexpression behaviour that is more comparable and interpretable between gene pairs (Fig A) SCR was used as coexpression measure in all our analyses.

	PCC	SCC	MI	**PCR**	**SCR**	**MIR**
AUC	0.9738	0.9888	0.9609	**0.9816**	**0.9933**	**0.9726**

Table A. Area Under the Curve (AUC) as a performance measure for the similarity measures PCC, SCC, MI, PCR, SCR, and MIR. The AUC calculated from the ROC curve quantifies the ability of a measure to separate TP from TN, in this case a measure of coexpression to separate contiguous pairs of operonic genes from random gene pairs. The first three columns of Table A represent PCC, SCC and MI, and the next three columns represent their corresponding PCR, SCR and MIR values.

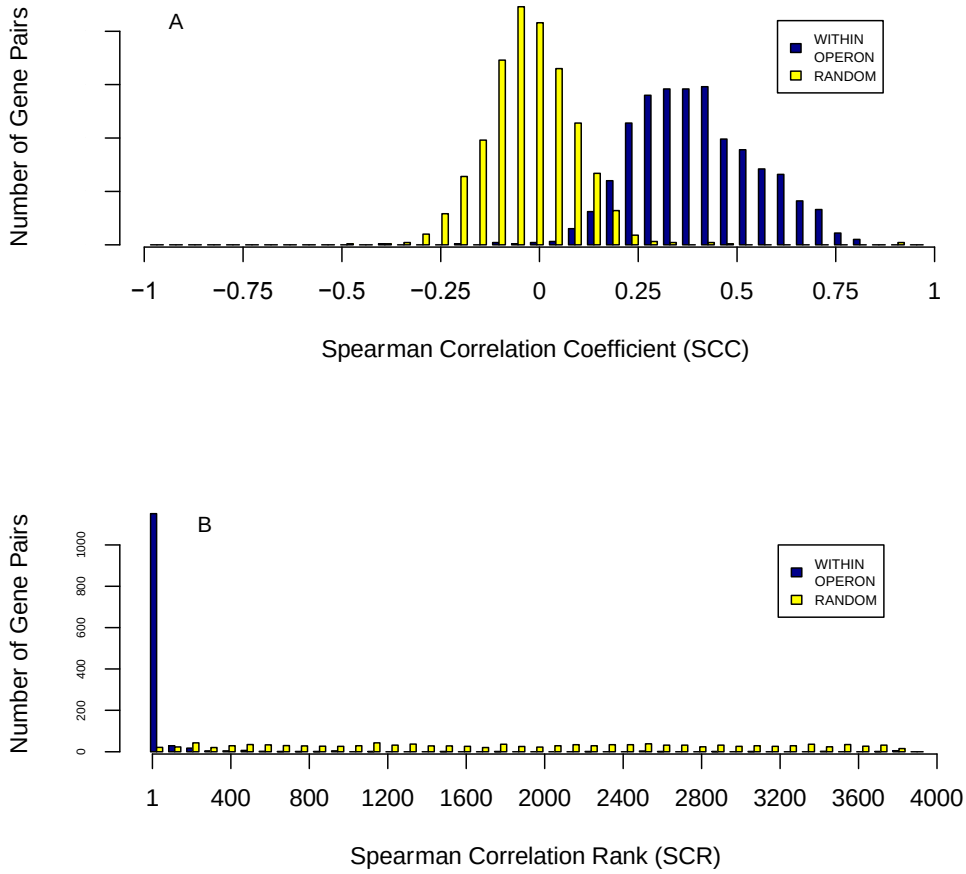

Fig A. Comparison of the SCC and SCR in assessing the degree of coexpression. When using the SCR as a measure of the degree of coexpression between pairs of genes belonging to the same operon, the assessed degree of coexpression is consistently high. SCC (A) and SRC (B) distributions based on expression data from the COLOMBOS compendium for a set of random gene pairs (TN) (yellow) and contiguous within operon gene pairs (TP) (blue). Each bar plot shows two histograms representing the coexpression distribution for TP gene pairs (N = 1,238) (explained in the text) and TN random gene pairs (N = 1,000).

2. Rank-based similarity measures compensate for conditional dependency

The rank derivatives of the standard used PCC, SCC and MI inherently normalize for the variability in ranges of PCC, SCC and MI values that can be observed between genes in a given dataset and hereby facilitate comparing degrees of coexpression between gene pairs.

Consequently part of the reason why the SCR, as a rank-based derivative of the more classically used Spearman Correlation Coefficient performed so well in our study is it's improved ability to compensate for the conditional dependency of transcriptional regulation than the standard used coexpression measures, such as PCC, SCC or MI. In our study, coexpression between genes was measured across all experiments of the expression compendium, irrespective of the conditions under which the genes were effectively coexpressed and thus assumed to be coregulated. When using standard correlation measures such as PCC, SCC or MI, genes that are coregulated under a low number of conditions only because of sample biases in the compendium, will by definition exhibit a low degree of measured coexpression [38]. As a result with standard coexpression measures, such as PCC, SCC or MI it is difficult to distinguish between a low degree of coexpression and/or coregulation or a high degree of coexpression and/or coregulation that was observed in a small subset of the conditions only. Both situations give rise to low measured degrees of coexpression. For the rank-based derivatives of the PCC, SCC or MI on the contrary this is less of an issue, as they express the expression similarity of one gene versus the other gene in a gene pair (i.e., A versus B) relative to the expression similarity of both A versus all other genes and B versus all other genes as mentioned. Thus, even when two genes are highly coexpressed in a small subset of the conditions only, their SCR value might still be equally high as that of genes that are coexpressed under a large set of conditions. Therefore, rank-based derivatives of PCC, SCC or MI are expected to be more robust against biases in the number of samples of specific conditions in the compendium.

Our results show that this is indeed the case (Fig B): in our positive control, pairs of genes that are supposed to be coexpressed consistently receive consistently high coexpression values (low SCR) when using a coexpression measure based on the SCR whereas the range of their Spearman correlation values (SCC) is much wider. For instance for the two operonic genes essD and rrrD, the SCC is 0.28 (low correlation which means a low degree of coexpression degree) whereas the SCR of 3.87 (i.e. a low SCR which means high coexpression degree).

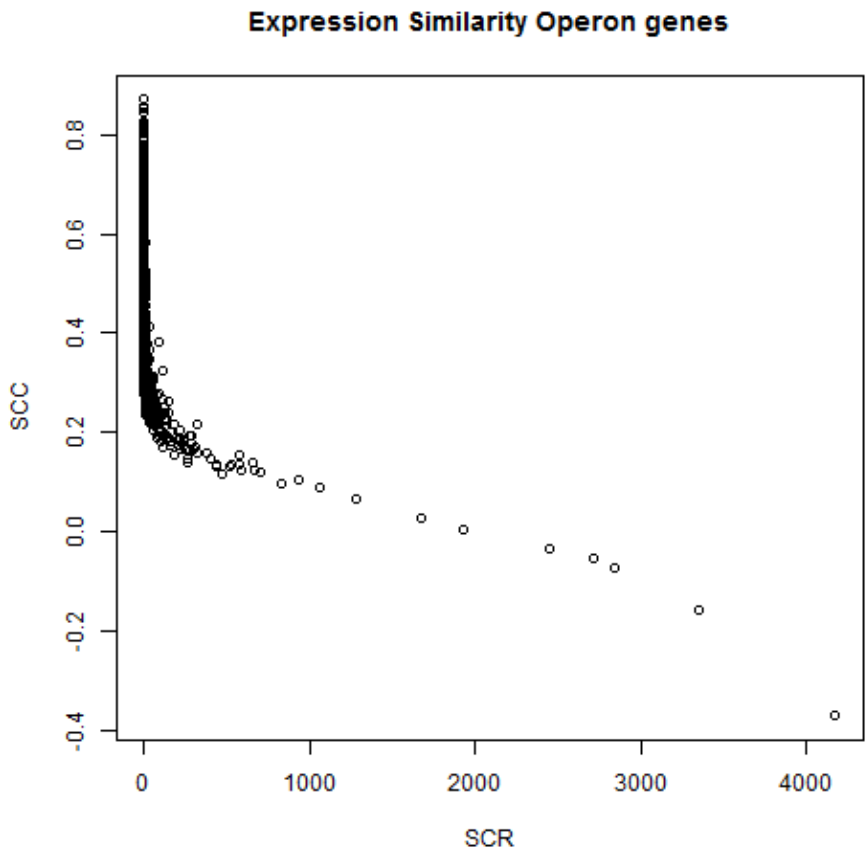

Figure B This figure displays for the positive control i.e. operonic genes that are expected to be well coexpressed, their pairwise Spearman Correlation Coefficient or SCC (Y-axis) as a function of their pairwise Spearman Correlation Rank or SCR (rank-based derivative of SCC) (X-axis). It shows that for most pairs of within operonic genes, their coexpression degree as measured by the SCR is generally higher (low SCR meaning high coexpression degree) than their coexpression degree assessed by the SCC.

3. The degree of coexpression of genes that are coregulated in *E. coli* is generally low

In general it is assumed that genes that are coregulated by the same Transcription Factors (TFs) tend to be highly coexpressed [2]. To have an intuition of the absolute degree of coexpression of coregulated genes in *E. coli* we compared their coexpression with that of genes that are located in the same transcription unit (operonic genes) and that thus should display the maximal levels of coexpression.

To this end we evaluated the degree of coexpression genes located (1) in the same operon, versus the degree of coexpression of genes that are (2) coregulated but not within the same operon (definitions of operons and coregulated genes are described in Materials and Methods).

The degree of coexpression of operonic genes and coregulated genes as measured by SCR was shown in Fig C by boxplotting the SCR values for respectively operonic (left panel) and coregulated genes (right panel).
Operonic genes were mostly highly coexpressed (low SCR), while the majority of coregulated non-operonic genes displayed much lower degrees of coexpression (high SCR).

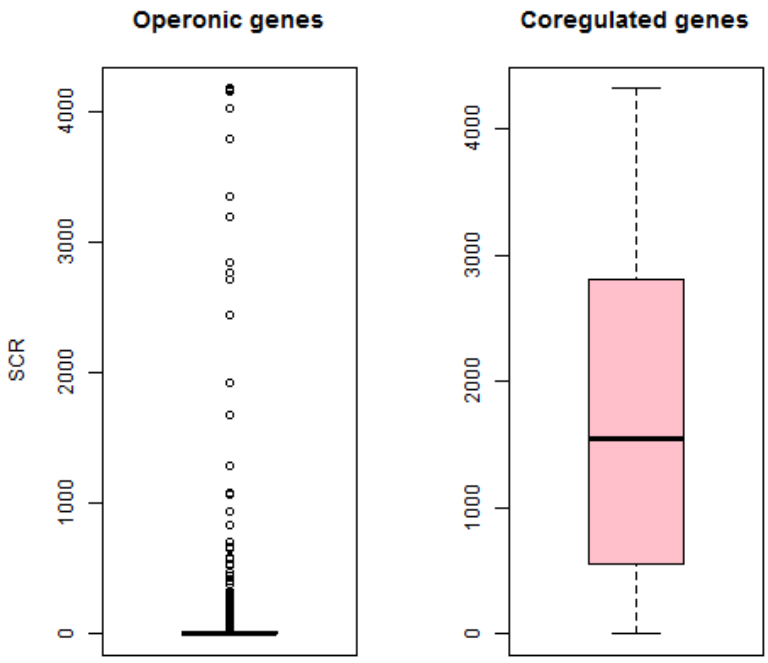

Fig C. Coexpression of genes within operons and of coregulated genes in E. coli. Coexpression degrees of operonic genes and coregulated genes are shown by boxplots of SCR values of gene pairs extracted from respectively operons (left panel) and of gene pairs coregulated by at least one TF (right panel).

4. Evidence classification of interactions in RegulonDB

RegulonDB distinguishes between TF-gene interactions supported by strong versus weak evidence. Interactions are classified as 'based on strong evidence' if they are supported by at least one source of strong evidence and 'based on weak evidence' if they are supported by weak evidence only. According to RegulonDB, "**Weak evidence** is a single evidence with more ambiguous conclusions, where alternative explanations, indirect effects, or potential false positives are prevalent, as well as computational predictions; for instance gel mobility shift assays with cell extracts or gene expression analysis and **Strong evidence** is a single evidence with direct physical interaction or solid genetic evidence with a low probability for alternative explanations; for instance, footprinting with purified protein or site mutation."

To ensure that the main conclusions of our analyses were not influenced by whether or not we included interactions with weak evidence, we tested the impact of using different sets of interactions on our result, more specifically we tested:

- a set including **all interactions** supported by both weak or strong evidence (i.e. 3430 interactions corresponding to 98795 coregulated gene pairs)

- a **partially reduced set of interactions** excluding interactions supported by at most one type of weak evidence only (i.e. 2961 interactions corresponding to 78772 coregulated gene pairs or 86% of the number of coregulated gene pairs of the full set of coregulated gene pairs)

- a set of **strongly evidenced interactions only** containing interactions based on strong evidence only - in comparison to the previous setting here also interactions that are

supported by two types of weak evidence are excluded (i.e. 2213 interactions corresponding to 30894 coregulated gene pairs or 31% of the full set of coregulated gene pairs).

We redid the analysis represented in the main text with each of the datasets mentioned above (all interactions, partially reduced dataset and the dataset containing strongly evidenced interactions only). Fig D represents the effect of the distance on the degree of coexpression as obtained for each of the datasets. Overall tendencies were similar, irrespective of the dataset that was used. Except for the case where genes are 'coregulated by 1 TF with complete overlap of regulatory programs' the tendency observed for the effect of the distance on the degree of coexpression was different between the results obtained for the different datasets and especially non-monotonic for the most stringent dataset (only interactions supported by strong evidence, blue dotted curve, right lower panel). Because of the non-monotonic behavior in case of the most stringent condition, we believe that in this setting the dataset becomes too small to observe a consistent behavior (this dataset contained 1046 pairs of genes instead of 1461 pairs of genes in case of the partially reduced dataset).

Results thus show that in general conclusions are not affected by including weak interactions. As the partially reduced dataset offers the best trade-off between using high confidence interactions and still offering sufficient data to observe tendencies, all the results in the main text were obtained with this dataset.

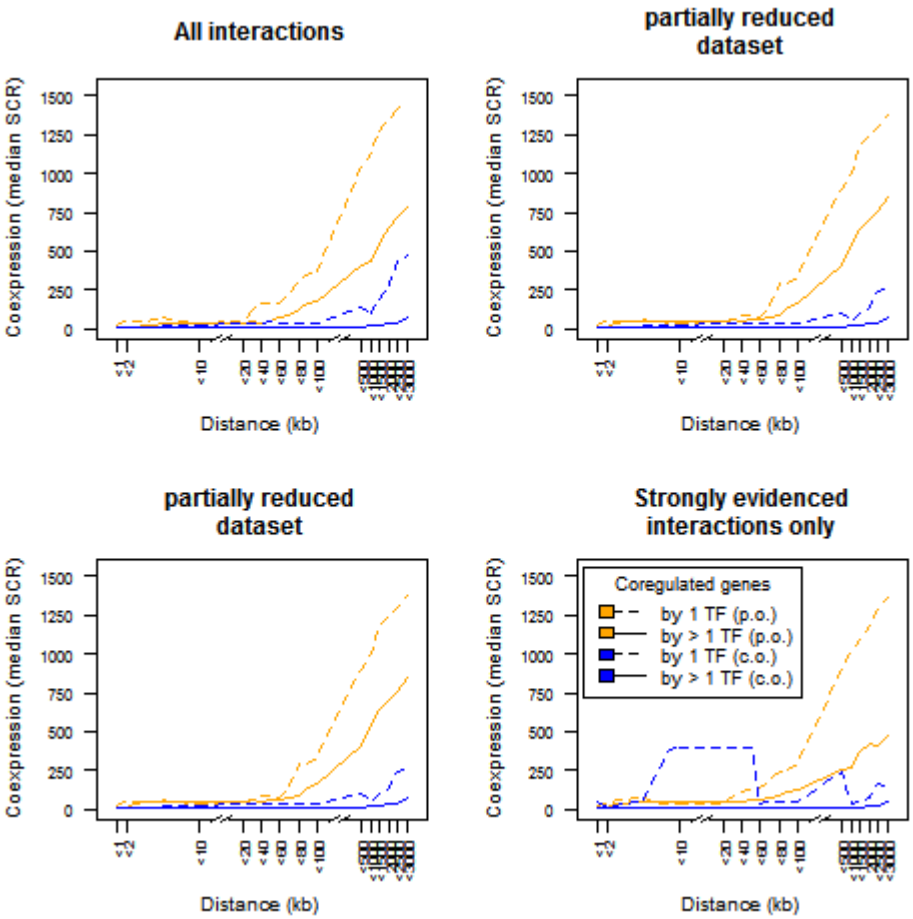

Fig D. ***Effect of coregulation tightness and of the distance between coregulated genes on the coexpression degree for different types of datasets****. Left upper panel: all interactions, right upper and left lower panel: partially reduced dataset, right lower panel: strongly supported interactions only. The coexpression behavior of coregulated genes was disentangled, depending on whether the regulatory programs displayed complete overlap (c.o.) versus partial overlap (p.o.) (blue versus orange) and*

depending on the number of common TFs present in the overlapping part of their regulatory program (dotted line for 1 TF versus full line for >1 TF).

5. Evidence of horizontal gene co-transfer in genes with strong distance conservation

As explained in the results we found indications that for highly coexpressed genes located in each other's neighborhood on the genome there is an evolutionary constraint on conserving their small distance. Because evolutionary conservation of close distance of genes has been associated with horizontal gene co-transfer we evaluated whether highly coexpressed genes that are nearby located and that have strong distance conservation show evidence of horizontal gene co-transfer.

Hereto we selected cases of coregulated genes that were located at small distances (< 5 intervening genes), that were highly coexpressed (SCR < 100), which had an orthologous counterpart in most other species (> 90% of gamma-proteobacteria) and for which the intergenic distance was highly conserved (distance conservation > 0.4). This resulted in 75 pairs of nearby located coregulated genes. We then assessed whether these pairs of genes were co-acquired by horizontal gene co-transfer.

We found that these 75 pairs of genes belonged to 12 different pairs of operons of which 9 show evidence of horizontal operon co-transfer: *araBAD-araC*, whose genes are co-transferred as found repeatedly within γ-Proteobacteria; *csgBAC-csgDEFG* and *narGHJI-narK* [3]; *rhaSR-rhaBAD*, which shows ancestral co-transfer; and *cusRS-cusCFBA*, a more recent co-transfer within the *E. coli - Shigella* lineage [4]. Of the remaining operons we did not find any evidence in the literature of HGT, nor did we find support for their HGT in computational predictions [5].

This indicates that for highly coexpressed nearby located genes there exists an evolutionary constraint for maintaining their small distance.

References

1. Adhya S, Gottesman M. Control of transcription termination. Annu Rev Biochem. 1978;47: 967–96. doi:10.1146/annurev.bi.47.070178.004535

2. Yu H, Luscombe NM, Qian J, Gerstein M. Genomic analysis of gene expression relationships in transcriptional regulatory networks. Trends Genet. 2003;19: 422–427. doi:http://dx.doi.org/10.1016/S0168-9525(03)00175-6

3. Price MN, Dehal PS, Arkin AP. Horizontal gene transfer and the evolution of transcriptional regulation in Escherichia coli. Genome Biol. BioMed Central; 2008;9: R4–R4. doi:10.1186/gb-2008-9-1-r4

4. Skippington E, Ragan MA. Within-species lateral genetic transfer and the evolution of transcriptional regulation in Escherichia coli and Shigella. BMC Genomics. BioMed Central; 2011;12: 532. doi:10.1186/1471-2164-12-532

5. Garcia-Vallve S, Guzman E, Montero MA, Romeu A. HGT-DB: a database of putative horizontally transferred genes in prokaryotic complete genomes. Nucleic Acids Res. Oxford, UK: Oxford University Press; 2003;31: 187–189. Available: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC165451/