doi:10.1016/j.jmb.2012.03.009 *J. Mol. Biol.* (2012) **xx**, xxx–xxx

Contents lists available at www.sciencedirect.com

Journal of Molecular Biology

journal homepage: http://ees.elsevier.com.jmb

Genomic Organization of Evolutionarily Correlated Genes in Bacteria: Limits and Strategies

Ivan Junier[1,2*], Joan Hérisson[1] and François Képès[1*]

[1]*Epigenomics Project/Institute of Systems and Synthetic Biology, Genopole, CNRS, University of Evry, 91030 Evry, France*
[2]*Institut des Systèmes Complexes Paris Île-de-France, 57-59 rue Lhomond, 75005 Paris, France*

Received 19 December 2011;
received in revised form
12 March 2012;
accepted 13 March 2012

Edited by M. Yaniv

The need for efficient molecular interplay in time and space within a cell imposes strong constraints that could be partially relaxed if relative gene positions along chromosomes were appropriate. Comparative genomics studies have demonstrated the short-scale conservation of gene proximity along bacterial chromosomes. Additionally, the long-range periodic positioning of evolutionarily correlated genes within *Escherichia coli* has recently been highlighted. To gain further insight into these different genetic organizations, we examined the compromise between chromosomal proximity and periodicity for all available eubacterial genomes by evaluating groups of evolutionarily correlated genes from a benchmark data set.

In enterobacteria, strict chromosomal proximity is found to be limited to groups under 20 genes, whereas periodicity is significant in all groups over 50. The *E. coli* K12 genome bears 511 periodic genes (12% of the genome), whose orthologs are found to be periodic in all eubacterial phyla. These periodic genes predominantly function in macromolecular synthesis and spatial organization of cellular components. They are enriched in essential and housekeeping genes and tend to often be constitutively expressed.

On this basis, it is argued that chromosomal proximity and periodicity are ubiquitous complementary genomic strategies that favor the build-up of local concentrations of co-functional molecules. In particular, the periodic layout may facilitate chromosome folding to spatially organize the construction of major cell components. The transition at 20 genes is reminiscent of the size of the longest operons and of topological microdomains. The range for which DNA neighborhood optimizes biochemical interactions might therefore be defined by DNA topology.

Keywords:
genome functional organization;
spatial organization of chromosome;
synteny properties;
periodical organization of genes;
gene expression at the cellular scale

Introduction

Chromosome conformation, genome organization and gene expression may be expected to be interdependent. Evidence for nonrandom genome layout, defined as relative gene positioning, stems from two main approaches. Firstly, the analysis of contiguous genome segments across species has highlighted the conservation of gene order (synteny) along short chromosomal stretches.[1–3] Secondly, the study of long-range regularities along the chromosomes of one given species has

**Corresponding authors.* I. Junier is to be contacted at Center for Genomic Regulation, Barcelona, Spain. E-mail addresses: i.junier@gmail.com; francois.kepes@epigenomique.genopole.fr.

Present address: I. Junier, Center for Genomic Regulation, Barcelona, Spain.

Abbreviations used: TF, transcription factor; TU, transcription unit; cTU, correlated transcription unit; PTU, periodic transcription unit; GO, Gene Ontology; HD, hypergeometric distribution; COG, Cluster of Orthologous Gene.

emphasized periodic positioning of microbial genes that are either co-regulated,[4,5] co-expressed,[6] evolutionarily correlated[7] or highly codon-biased.[8]

Képès and Vaillant have postulated the existence of a positive feedback loop connecting periodic genome layout to the cellular conformation of chromosomes on to transcriptional control.[4] Indeed, through the use of a thermodynamic model of chromosome folding where transcription factors (TFs) cross-link distant binding sites, a periodic relative gene positioning has been shown to be crucial for achieving chromosome conformations that favor the formation of "transcription factories."[9] These transcription factories are discrete spatial foci formed, in eukaryotes, by an elevated concentration of TFs and their target genes.[10–12] In particular, they gather genes that can be located far from each other along DNA.[13] As demonstrated in the *lac* operon, the increased local concentration of TFs and their cognate binding sites can then lead to both enhanced[14–16] and robust[17] transcriptional control. Accordingly, the spatial localization of these genes has been shown to be tightly related to their biological function (see, e.g., Ref. 18).

More generally, colocalization of molecules effectively enhances their concentration at specific cellular locations. By enhancing physical interactions and chemical reactions, this locally elevated concentration can trigger functional interdependencies that would not occur otherwise.[19] In this respect, bacteria offer an interesting case because functional constraints and opportunities arise from the coupling of transcription and translation, reactions that have been shown to be spatially organized.[20–25] Furthermore, comparative genomics efficiently delineate fundamental bacterial processes because it is now supported by a thousand sequenced genomes.

In this context, our goal in this article is to shed new light on the interplay between genome layout and genome function and to test the generality of the uncovered phenomena by comparing genomes. In previous studies, we have used transcriptional co-regulation as a proxy for co-function.[5] We found that, in bacteria, genes regulated by the same TF should gather into transcription factories.[4] Here, we widen the scope of co-function by analyzing a benchmark data set consisting of 2254 protein-coding genes contributing to 22,500 gene pairs.[7] These pairs had been identified by comparing 105 bacterial genomes on the basis of two types of evolutionary correlations:[7] a tendency to be located close together in many genomes, independently of their respective order, and phylogenetic co-occurrence,[26] meaning that one gene tends to be present in a particular genome only if the other gene is also present. Genes in such pairs, hereafter referred to as "correlated genes," are known to often share function or transcriptional regulation or have gene products that physically associate.[1,26] In their analysis, Wright *et al.* have shown that correlated genes tend to be separated by integral multiples of 117 kb along the *Escherichia coli* genome.[7] This periodic trend was highlighted by a Fourier-based spectral analysis of the pair–distance histogram. Here, we refine this analysis by applying a new algorithm to the *E. coli* genome that, unlike Fourier analysis, detects complex periodical patterns in sparse, noisy and incomplete data sets.[27] Next, we identify the set of genes that are effectively contributing to this periodical pattern, and we characterize their striking properties. We then extend our results across all sequenced eubacterial genomes. Finally, we present our findings in terms of a unified framework based on the interplay between genome layout and spatial colocalization of co-functional genes.

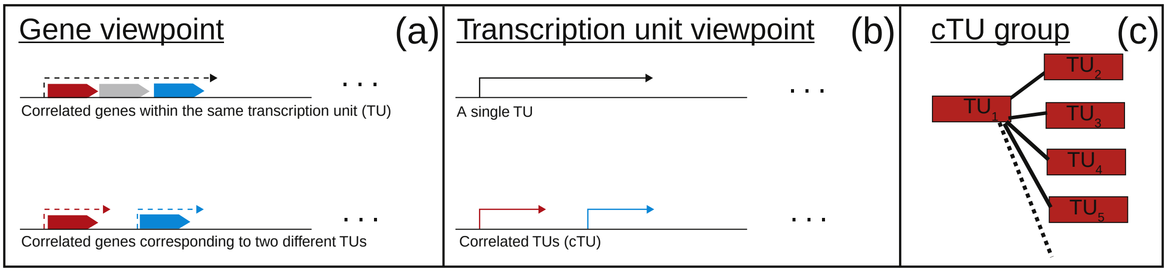
Gene viewpoint
(a)
. . .
Correlated genes within the same transcription unit (TU)
. . .
Correlated genes corresponding to two different TUs
Transcription unit viewpoint
(b)
. . .
A single TU
. . .
Correlated TUs (cTU)
cTU group
(c)
TU1
TU2
TU3
TU4
TU5

Fig. 1. Groups of cTUs. One TU corresponds to a single transcript and may include several genes or cistrons. (a) to (c) then show how groups of cTUs are built from the data of Wright *et al.*[7] In this data set, pairs of correlated genes were identified on the basis of a tendency to be located close together in many genomes and of phylogenetic co-occurrence. (a) Pairs of evolutionary correlated genes, schematically depicted as blue and red genes, were determined on the basis of a tendency to co-occur in 105 bacterial genomes and a tendency to be proximal along the chromosomes.[7] Evolutionary correlated genes may belong to a single TU (broken arrows). (b) Alternatively, they may belong to different TUs, which are then said to be correlated. (c) A cTU group is defined with respect to a given TU. More precisely, it is defined with respect to a specific gene belonging to this TU: it gathers all the TUs that contain at least one gene correlated to this gene.

Please cite this article as: Junier, I. *et al.*, Genomic Organization of Evolutionarily Correlated Genes in Bacteria: Limits and Strategies, *J. Mol. Biol.* (2012), doi:10.1016/j.jmb.2012.03.009

Results

From genes to transcription units

Contiguous protein-coding loci within operons, or "cistrons," are expected to affect genome layout statistics. To observe genome layout unhindered by the operonic organization of bacterial chromosomes, the *E. coli* benchmark data set of 2254 genes obtained by Wright *et al.*[7] is hence first reduced to its 1303 transcription units (TUs), each expressing one mRNA. Evolutionarily correlated transcription units (cTUs) are then defined as two TUs that share at least one pair of correlated genes in the benchmark data set (Fig. 1). A cTU group is next defined as comprising one given TU plus all TUs containing at least one gene that pairs with a specific gene of the given TU (Fig. 1). In other words, a cTU group contains all the TUs correlated to a given TU through a specific gene. Thus, each cTU group represents a *bona fide* list of the evolutionarily correlated partners of one TU. Overall, the benchmark data set contains 1903 cTU groups containing between 2 and 129 TUs.

Chromosomal proximity *versus* periodic spacing

To assess positional regularities within each of these 1903 cTU groups, they were subjected to the solenoidal coordinate method. This method has been shown to be particularly efficient in highlighting the presence of complex periodic patterns in small data sets that contain wrong information (false positives), missing data (false negatives) and noise (positional errors).[27] At a given period, it provides a score (Materials and Methods) that is all the higher that the likelihood for the data set to present a periodic pattern with this period is large (*periodicity score*). High scores at the period equal to full chromosome length represent particular cases that reflect proximity of TUs in some chromosomal regions (*proximity score*). Because this method scans many different periods, it provides for each cTU group a "spectrum" of scores as a function of the period.

To assess the global trend of the whole benchmark data set, the spectra of all cTU groups were normalized for the number of groups with equal sizes and then averaged (Fig. 2). This average spectrum reaches its maximum at a period equal

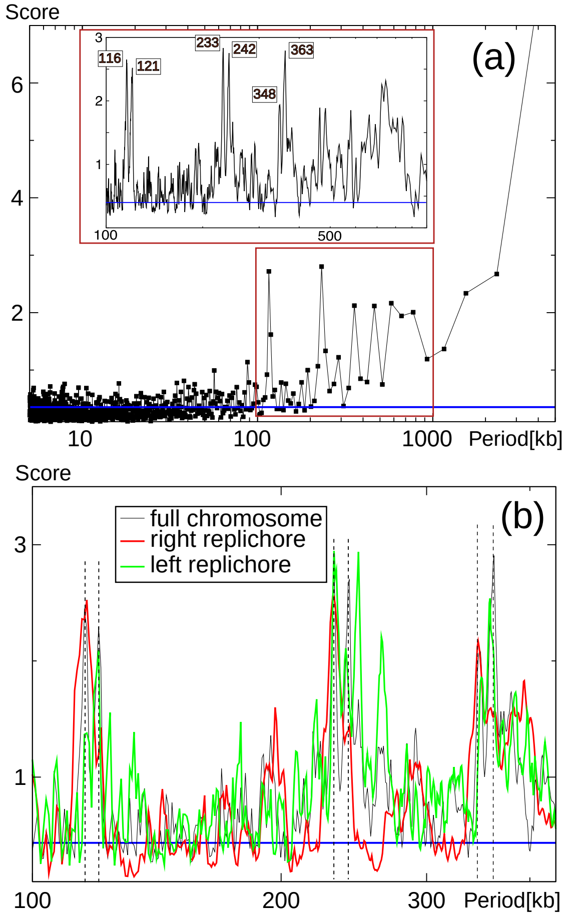

Fig. 2. Periodicity spectra averaged over all groups of cTUs. (a) Full *E. coli* data set. The benchmark data set reduced to its cTU groups (Fig. 1) was globally analyzed for periodicity with the solenoidal coordinate method.[9] The main curve with black squares reports the periodicity scores (Materials and Methods) averaged over all cTU groups, at discrete periods (see Supplementary Fig. 2 for the spectrum of single groups). These periods are submultiples of the chromosome length and range from 5 kb to full chromosome length. All group sizes are equally represented because the contribution from each group to the average is normalized by the number of groups having the same size. That is, at a given period, a score in these panels is given by $\sum_g \frac{S_g}{N_g}$, where S_g is the score of a peculiar cTU group g and N_g is equal to the number of groups having the same size as g, and the sum goes over the 1903 cTU groups. A subset of these periods (inset), from 100 kb to 1 Mb, is scanned with a higher resolution after subtracting the basal line drift due to chromosomal proximity. The labels indicate peak periods in kilobase pairs. In particular, if periodicity was not present, scores would typically be on the same order as those observed below 100 kb. For instance, scores equal to 3 correspond to *z*-scores ~ 26, whereas scores equal to 2.5 correspond to *z*-scores ~ 21 (Supplementary Material). The large islet of high scores around 720 kb can be explained as the combination of harmonics equal to the triple period of 233–242 kb and to the double period of 348–363 kb. (b) Full *E. coli* data set split between right and left replichores. A 100- to 400-kb window is plotted for the full chromosome (black) and separately for the right (red) or left (green) replichores. Each main peak identified in the inset of (a) is indicated here by a double broken vertical line, corresponding to the slightly differing periods of the right and left replichores.

Please cite this article as: Junier, I. *et al.*, Genomic Organization of Evolutionarily Correlated Genes in Bacteria: Limits and Strategies, *J. Mol. Biol.* (2012), doi:10.1016/j.jmb.2012.03.009

to full chromosome length, reflecting the chromosomal proximity of TUs. Chromosomal proximity is expected since the correlated genes have been selected so that their orthologs tend to be close to each other across bacterial species.[7] Besides this trivial result, six major peaks are observed at 116, 233 and 348 kb and 121, 242 and 363 kb (inset, Fig. 2a). These are consistent with the 117-kb period detected previously,[7] the two last periods respectively being the double and triple harmonics of the first ones, 116 and 121 kb.

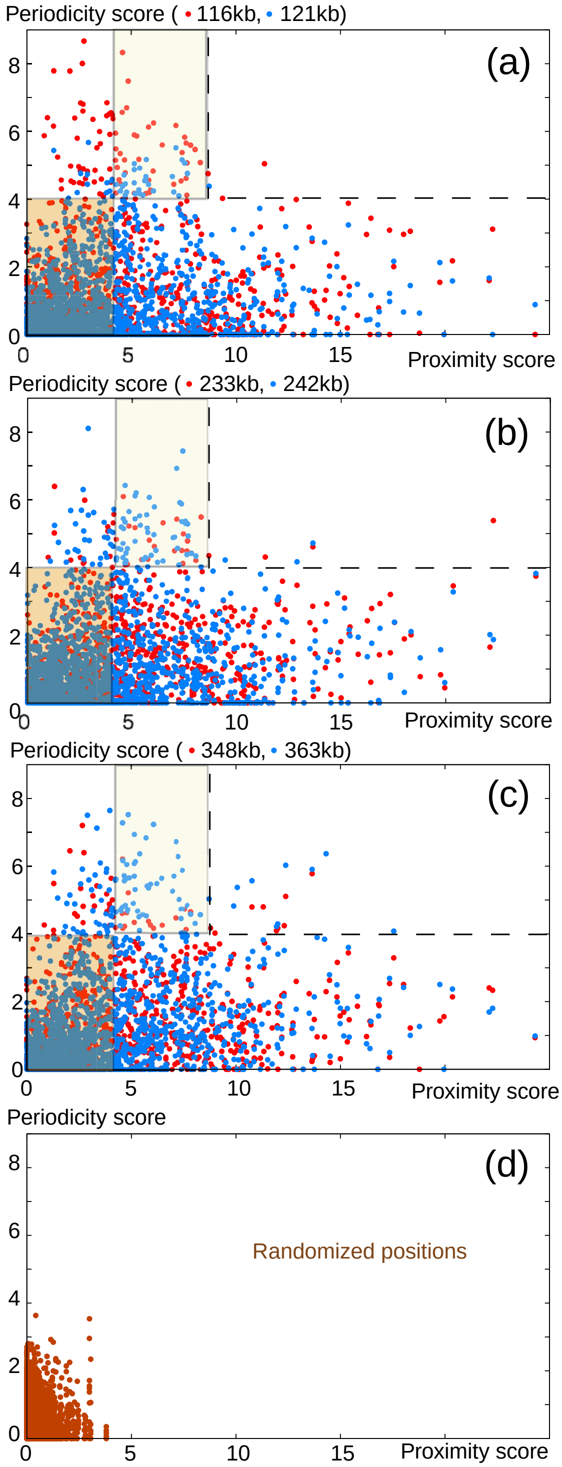

Figure 2b shows the spectra when the positions are analyzed at a higher resolution, separately for each of the two replichores in the circular chromosome—replichores are half chromosomes bordered by replication origin and terminus. For the right replichore, three major peaks are observed at 116, 233 and 348 kb. For the left replichore, two peaks are observed at 121 and 363 kb. In addition, period 242 kb is at the center of a symmetric islet composed of three close peaks. Thus, the presence of pairs of close periods in the global trend (Fig. 2a) can be attributed to a slight difference of period between the right and left replichores. Furthermore, all harmonic periods appear to be significant, and none of them can be considered as a methodological artifact of another one (Supplementary Fig. 1).

To investigate how individual cTU groups contribute to this overall trend, the periodicity scores at the six major peaks were plotted against the proximity scores for all groups (Fig. 3). It appears that the groups can be divided into three main categories. The first category contains 533 groups, which show significant proximity scores (≥4); spectra of these groups have their highest peak mostly at full chromosome length (Supplementary Fig. 2a). The second category contains 167 groups, which show significant periodicity scores (≥4); spectra of these groups have their highest peak mostly at one of the six major periods (Supplementary Fig. 2b). They will later be referred to as the 167 "periodic" cTU groups. The last category contains 1306 groups (brown background) whose scores are no more significant than a randomized set (Fig. 3d).

As highlighted by the green background in Fig. 3a–c, 103 groups show both significant periodicity

Fig. 3. Chromosomal proximity *versus* periodic positioning for each group of cTUs. Each point represents a cTU group. The x-axis represents the proximity score, which is computed at a period equal to full chromosome length. The y-axis represents the periodicity scores at 116–121 kb (a), 233–242 kb (b) and 348–363 kb (c) corresponding to the peaks on Fig. 2b. (d) shows that, by randomizing TU positions along the whole genome according to a uniform law, scores do not exceed 4. In (a) to (c), scores below these values of 4 are accordingly denoted as nonsignificant by a light-brown background. A light-green background highlights areas with simultaneously significant periodic and proximal scores larger than 4. The broken upper-right rectangle highlights the quasi-absence of cTUs with simultaneously very high periodicity and proximity scores. The upper-left rectangle delineates areas of cTUs with significant periodicity and nonsignificant proximity scores; vice versa for the lower-right rectangle.

and significant proximity, that is, they simultaneously belong to the above first two categories. However, a zone of exclusion appears in the upper-right corner of the plots, demonstrating that cTU groups cannot simultaneously be strongly proximal and periodical.

Proximity/periodicity balance as a function of group size

To investigate how these features relate to cTU group sizes, the periodicity and proximity scores for each cTU group were plotted against its size (Fig. 4).

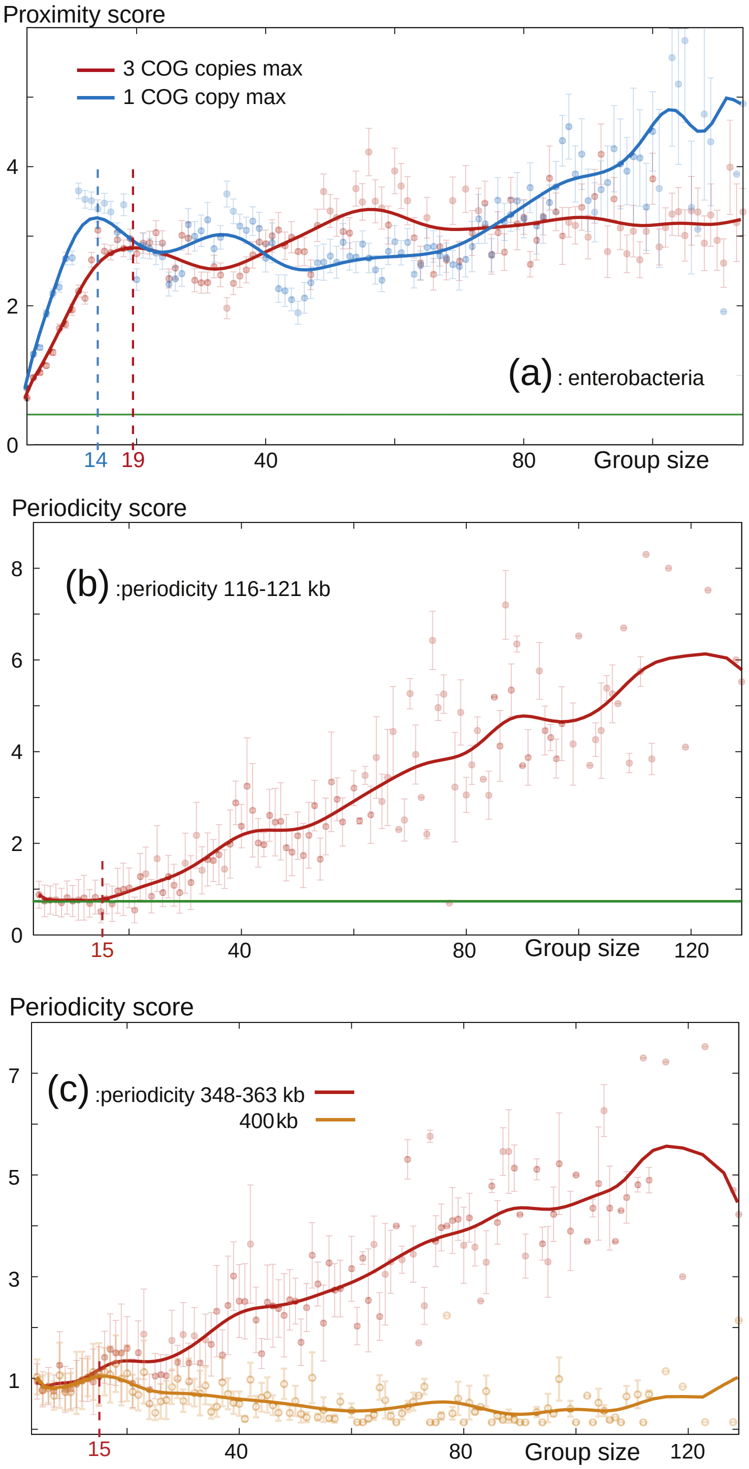

Fig. 4. Balance between periodicity and proximity, as a function of cTU group sizes. (a) Chromosomal proximity score for 62 enterobacterial genomes. Species are equally represented by normalizing the contribution of each strain by the number of studied strains in the species. To reduce false positives of ortholog prediction,[28] the analysis is restricted to COGs having at most three copies in the same target genome (red) or more stringently to COGs having only one copy (blue). The proximity score increases up to groups of 14–19 TUs (vertical broken bar) and then reaches a quasi-plateau. The green horizontal line corresponds to the average value for a randomized data set. *E. coli* results are qualitatively similar, with a transition around 17 TUs, but show a larger variance, which requires a proper statistical analysis (Supplementary Fig. 3a). (b) Periodic positioning score for the *E. coli* genome. For each cTU group, a pair of scores is measured at periods 116 and 121 kb, and the highest score of the pair is plotted. Periodicity departs from the randomized level depicted by a green horizontal line, when group size is above 15 TUs (vertical broken line), and then steadily increases. (c) Periodic positioning score for the *E. coli* genome. The score is measured as in (b), but at periods 348 and 363 kb (red curve). As a control, the score is also measured for a period of 400 kb (orange curve), which in Fig. 2 does not yield significant periodicity. The two curves diverge above a group size of 15 TUs (vertical broken line), and then the red curve steadily increases, while the control orange curve remains at a low level. All panels: smooth curves are Bézier interpolations of real data; the error bars indicate sample standard deviations (the absence of error bars means that only one group is concerned by the size reported along the *x*-axis) (see Supplementary Fig. 3 for additional statistical analysis).

Please cite this article as: Junier, I. *et al.*, Genomic Organization of Evolutionarily Correlated Genes in Bacteria: Limits and Strategies, *J. Mol. Biol.* (2012), doi:10.1016/j.jmb.2012.03.009

In the case of chromosomal proximity, scores increase with size until about 15 TUs and then are roughly constant (Supplementary Fig. 3a). These data display a large variance, though. We hence extended the analysis to all 62 enterobacteria whose genomes have been sequenced and annotated. The same trend is observed, whichever the stringency used in determining orthology (Fig. 4b), but with a much lower variance due the larger data set on which the analysis feeds. By contrast, scores related to periodic spacing stay constant and close to the random level until about 15 TUs and then increase steadily to become extremely significant for large sizes (Fig. 4c and d). As a control of non-periodic behavior, a period of 400 kb was chosen because it is close to 348–363 kb but does not display any periodicity, as evidenced in Fig. 2. The score for this period of 400 kb remains roughly flat and low for all sizes (Fig. 4d, orange). All these trends are maintained when the number of correlated genes from the original data set, rather than the number of cTUs, is considered. In this case, the breakpoint is found at ~20 genes, consistent with 15 TUs and ~1.3 genes per TU (Supplementary Fig. 4).

These results highlight a breakpoint for cTU groups around a size of 15 TUs. Below this limit, successive TUs can be accommodated along the chromosome. For larger group sizes, periodicity appears as a complementary strategy.

Mixed patterns in the large-scale organization of cTUs

To better understand how periodic and proximal patterns coincide, cTUs groups were scrutinized individually. Two extreme illustrations follow. The first case is the cTU group with the strongest periodicity score of 8.7 at period 116 kb (Fig. 5a). This most periodic group comprises genes functioning in macromolecular synthesis and in cell organization (Supplementary Table 1). On top of the periodic pattern, many of the 87 TUs (236 genes) in this cTU group locate within short dense clusters along the chromosome. The second case concerns the cTU group with the highest proximity score of 24.3 (Fig. 5b). This most proximal group comprises many "unknown" genes, thereby precluding any conclusion on their functional enrichment (Supplementary Table 1). It contains 37 TUs (88 genes), which exactly cover the "Ori macrodomain," that is, the macro-structured domain that symmetrically surrounds the origin of DNA replication.[29] More generally, it appears that this group and some others containing up to ~40 TUs can show strong proximity along wide chromosomal regions without displaying any periodic trend. These wide regions typically comprise a succession of shorter dense clusters. By contrast, periodicity is statistically significant for all groups over ~40 TUs (Supplementary Fig. 5). Altogether, it appears that chromosomal proximity and periodic spacing coexist in the 15- to 40-TU range.

Identification of periodic transcription units

To better understand the role of periodicity in the organization of the *E. coli* genome, it is important to identify the TUs that contribute most to the periodic trend, or "periodic transcription units" (PTUs). To this end, a period-dependent positional score[27] is allocated to each TU. This individual score reflects the tendency for the TU to be well positioned with respect to its cTU group for a given period—it is therefore of a different nature from the periodicity scores computed for each cTU group (as reported in Fig. 3). The procedure is applied to the 167 periodic cTU groups (Fig. 3), which contain altogether 511 TUs

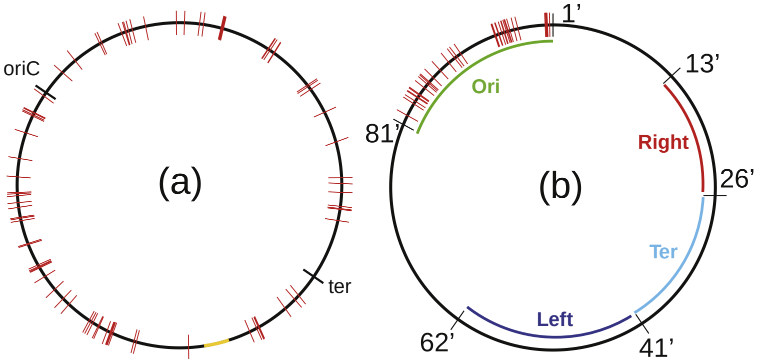

Fig. 5. Interplay between chromosomal proximity and long-range periodic spacing. The black circle represents the circular chromosome of *E. coli*. (a) Mixed patterns of the most periodic cTU group. Red bars indicate the positions of the 87 TUs contained in the cTU group showing the strongest periodicity score at 116 kb. The yellow scale bar represents 116 kb. The DNA replication origin (*oriC*) and terminus (*ter*) are indicated with black bars. Note on top of the periodic pattern the concomitant organization of several TUs in dense clusters. (b) Series of dense clusters of the most proximal cTU group. The macrodomains of the origin (Ori) and terminus (Ter) of replication are, respectively, indicated by colored arcs (for a review, see Ref. 29). The red bars indicate the positions of the 37 TUs contained in the cTU group showing the strongest proximity score. These 37 TUs are disposed in a series of short stretches covering altogether the Ori macrodomain. Just as in (a), some of the stretches are dense clusters. The lists of genes in both cTU groups are provided in Supplementary Table 1.

(1015 genes or cistrons). Two significance thresholds are routinely compared, at the usual *p*-values of 0.05 or 0.01, corresponding to individual scores either over 1.3 or more stringently over 2 (in parentheses in the following). Note that removing the TUs scoring over 1.3 suffices to lose periodic trends in the benchmark data set (Supplementary Fig. 5).

PTUs are found to be equally distributed among all periods (Supplementary Fig. 7), thus confirming the singular role of harmonics for the periodic organization of the cTUs. They sum up to 246 (126) PTUs, which correspond to 511 (278) "periodic correlated genes" (Supplementary Table 2). These constitute about 20 (10)% of the cTUs or 22 (12)% of all correlated genes in the initial benchmark data set or equivalently 12 (6)% of all *E. coli* genes.

Connectivity properties of PTUs

Given that cTU group size matters for the proximity/periodicity balance, the question arises whether PTUs have a higher number of functional partners than average. To solve this issue, a "cTU network" is defined, with nodes representing TUs and edges linking two TUs if they are correlated (Fig. 6a). The network core is defined as the set of TUs with highest degrees (number of connections to functional partners). This is illustrated by a fake example on Fig. 6b, with the gray core in the center of the graph. From this network view, it appears that the proportion of PTUs among the TUs in the data set increases as a function of the degree (Fig. 6c). Above degree ~60 (vertical broken line), PTUs represent more than half of all TUs. Above degree ~100, TUs systematically belong to cTU groups showing strong

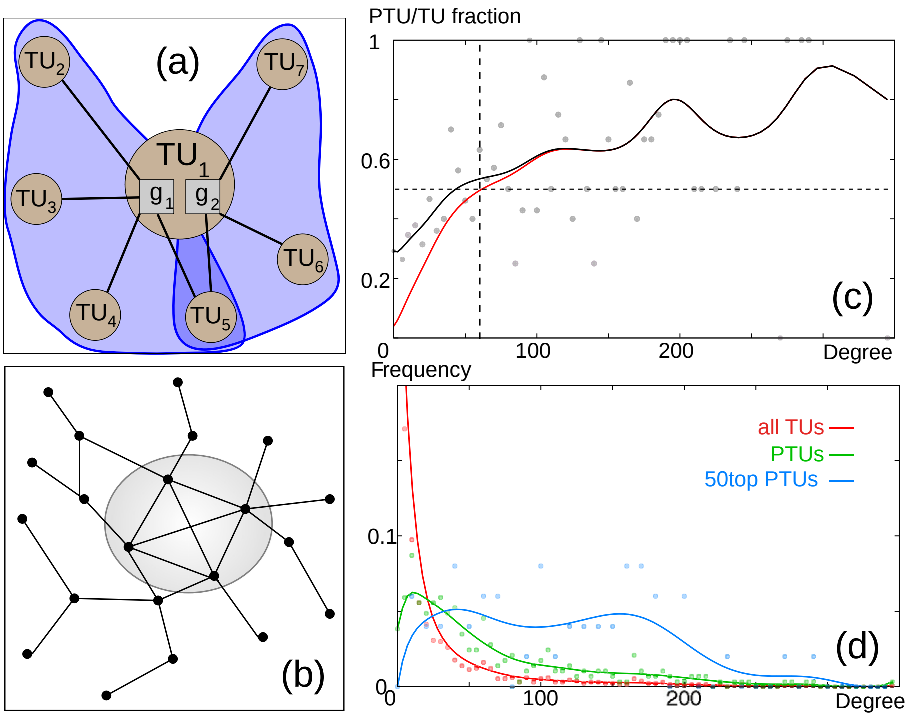

Fig. 6. PTUs have numerous partners. (a) Schematic view of the rules that are used to build the *cTU network*. Given a TU TU$_1$ with two genes g_1 and g_2, the set of adjacent TUs stems from the union of all cTU groups (blue shapes) associated to the genes in TU$_1$ (g_1 and g_2). Note that the cTU groups can overlap, as shown for TU$_5$. (b) Schematic representation of the *core* of a fake network. The core (gray area) contains the nodes with highest degree (number of connected partners). (c) Fraction of TUs that are periodic, as a function of their degree in the cTU network. Black points and curve: the PTU/TU fraction is computed with respect to the TUs that belong to the cTU groups with periodicity scores larger than 4. Red curve: the PTU/TU fraction is computed with respect to all TUs that belong to a cTU group containing more than 15 TUs (periodicity onset according to Fig. 4). The fusion of the red and black curves shows that all TUs with a degree larger than 100 belong to cTU groups showing strong periodicity. In addition, TUs with the highest degrees are systematically periodic. The two outlying points on the *x*-axis correspond to TUs whose positional score is under threshold 1.3 (used for PTU selection) but still above 1. Finally, above degree 60 (vertical broken line), PTUs represent more than half (horizontal broken line) of all TUs belonging to periodic cTU groups. (d) Degree distribution of TUs. Red, all TUs; green, only PTUs; and blue, the 50 PTUs with highest individual positional scores (Table 1). All 50 top PTUs have a degree larger than 13. (c and d) Smooth curves are Bézier interpolations of real data.

Please cite this article as: Junier, I. *et al.*, Genomic Organization of Evolutionarily Correlated Genes in Bacteria: Limits and Strategies, *J. Mol. Biol.* (2012), doi:10.1016/j.jmb.2012.03.009

Table 1. The 50 *E. coli* PTUs with highest positional scores, ranked by decreasing significance

PTU	Score	PTU	Score
rsmE-gshB	9.4	ftsK	8.5
yqgEF	9.4	lolCDE	8.5
yggSTU-rdgB-yggW	9.4	yoaA	8.5
rpsJ-rplCDWB-rpsS-rplV-rpsC-rplP-rpmC-rpsQ	9.4	pabB-nudL	8.5
rplKAJL-rpoBC	9.1	yicC	8.5
purEK	8.8	ispG	8.4
gmk	8.7	murA	8.4
yebA	8.7	map-glnD-dapD	8.3
znuCB	8.7	pyrH	8.2
bamB-der	8.7	efp	8.2
tilS	8.7	yjeFE-amiB-mutL-miaA-hfq-hflXKC	8.2
ybeZYX-lnt	8.7	yjeS	8.2
mfd	8.7	clpS	8.2
secE-nusG	8.6	fliFGHIJK	8.2
ruvAB	8.6	metK	8.1
cysS	8.6	rpoZ-spoT-trmH-recG	8.1
tig	8.6	dgt	8.1
yfgO	8.6	frr	8.0
bamA-hlpA-lpxD-fabZ-lpxAB-rnhB-dnaE	8.5	rlmN	8.0
yhbE-obgE	8.5	hisS	8.0
rplU-rpmA	8.5	rseP	8.0
rpsB-tsf	8.5	dut-slmA	7.8
trxB	8.5	rluD-yfiH	7.8
yeaZY	8.5	yihA	7.8
yicR-rpmBG-mutM	8.5	xseA	7.8

Individual positional scores were calculated for all six significant periods determined in Fig. 2. Top scores are reported in the right column.

periodicity, as evidenced by the merging of the black and red curves (Fig. 6c). Finally, the TUs with highest degrees are all periodic. Reciprocally, the 50 TUs with the largest positional scores (Table 1) are systematically connected to more than 13 TUs (Fig. 6d). Thus, the core of the cTU network is exclusively composed of PTUs.

Functional properties of PTUs

To evaluate the functional coverage of periodic genes, the full set of 511 periodic correlated genes described in the previous paragraph (scores over 1.3) was searched for shared ontology terms as provided by the Gene Ontology (GO) project.[30] The GO analysis reveals functional enrichment of a selected set of genes by allocating *P*-values, hereafter denoted as P_{GO}, associated to the frequency of ontological terms with respect to the original set of genes (see Materials and Methods for further details about hypothesis testing procedures used in this article). Statistical validity is assessed by the hypergeometric distribution (HD) test with respect to the initial set of correlated genes and includes a Bonferroni correction due to the repetitive nature of the analysis.[31] Two general functional categories display strong biases: macromolecular synthesis and structural organization of the cell (Table 2).

Genes coding for macromolecular synthesis are those involved in protein, RNA and DNA synthesis.

Table 2. GO analysis of the 511 periodic genes

GO term	*P*	Frequency in periodic genes	Frequency in correlated genes
GO:0006412—translation	2.58e-11	63/494 (12.8%)	113/2117 (5.3%)
GO:0019538—protein metabolic process	9.47e-08	101/494 (20.4%)	249/2117 (11.8%)
GO:0043170—macromolecule metabolic process	2.65e-07	220/494 (44.5%)	692/2117 (32.7%)
GO:0044260—cellular macromolecule metabolic process	4.62e-07	205/494 (41.5%)	636/2117 (30.0%)
GO:0044267—cellular protein metabolic process	7.32e-07	83/494 (16.8%)	197/2117 (9.3%)
GO:0071806—protein transmembrane transport	1.09e-05	17/494 (3.4%)	20/2117 (0.9%)
GO:0009987—cellular process	4.54e-05	392/494 (79.4%)	1476/2117 (69.7%)
GO:0044249—cellular biosynthetic process	4.79e-05	225/494 (45.5%)	745/2117 (35.2%)
GO:0034645—cellular macromolecule biosynthetic process	4.98e-05	141/494 (28.5%)	418/2117 (19.7%)
GO:0009059—macromolecule biosynthetic process	7.15e-05	141/494 (28.5%)	420/2117 (19.8%)
GO:0009058—biosynthetic process	1.12e-04	226/494 (45.7%)	755/2117 (35.7%)
GO:0033036—macromolecule localization	1.71e-04	45/494 (9.1%)	95/2117 (4.5%)
GO:0071841—cellular component organization or biogenesis at cellular level	2.89e-04	69/494 (14.0%)	172/2117 (8.1%)
GO:0071840—cellular component organization or biogenesis	1.34e-03	72/494 (14.6%)	188/2117 (8.9%)
GO:0043064—flagellum organization	3.93e-03	16/494 (3.2%)	23/2117 (1.1%)
GO:0015986—ATP synthesis coupled proton transport	5.24e-03	10/494 (2.0%)	11/2117 (0.5%)
GO:0015985—energy coupled proton transport, down electrochemical gradient	5.24e-03	10/494 (2.0%)	11/2117 (0.5%)
GO:0016043—cellular component organization	5.59e-03	55/494 (11.1%)	137/2117 (6.5%)
GO:0034613—cellular protein localization	5.84e-03	14/494 (2.8%)	19/2117 (0.9%)
GO:0070727—cellular macromolecule localization	5.84e-03	14/494 (2.8%)	19/2117 (0.9%)
GO:0008104—protein localization	7.23e-03	36/494 (7.3%)	78/2117 (3.7%)

The table reports terminology enrichment of biological processes with respect to the initial set of correlated genes. The first column indicates the GO term. The second column gives the statistical significance of the enrichment (only $P \le 10^{-2}$ are reported). The third and fourth columns, respectively, list the frequency of the term in the set of periodic genes and in the set of correlated genes. Four hundred ninety-four of the 511 periodic genes and 2217 of the 2254 correlated genes had at least one GO classification.

More precisely, the set of periodic genes contains 63 of the 113 correlated genes that are categorized as "translation process" ($P_{GO}\sim 3\times 10^{-11}$); 101 of 249, as "cellular protein metabolic process" ($P_{GO}\sim 10^{-7}$); and 220 of 692, as "macromolecule metabolic process" ($P_{GO}\sim 3\times 10^{-11}$). Concerning RNA transcription, the PTU set comprises the five genes encoding Sigma factors required during exponential growth (*rpoA-D*, *rpoZ*), and the elongation factor *greA*. Concerning DNA replication, the PTU set contains three of the six subunits forming the DNA polymerase III pre-initiation complex (*holA*, *B*,*D*), its core α-subunit (*dnaE*) and three of the six subunits forming the replicative primosome (*dnaB*, *G*, *priB*). Notice that all these genes belong to different operons, except two cases (*rpoB-C*, *rpoD-dnaG*). In many cases, by relaxing the significance threshold, genes encoding additional subunits of these protein complexes appear to also be periodically positioned (Supplementary Table 2). For instance, the second transcriptional elongation factor (*nusA*) and a fourth DNA polymerase III pre-initiation subunit (*holC*) located in the same operon as a ninth aminoacyl-tRNA synthetase (*valS*) have a positional score of 1.2; a fourth and a fifth primosome subunits (dnaT, *priA*) have a score of 1.

Concerning the structural organization of the cell, the PTU set contains 45 of the 95 correlated genes that are categorized as "macromolecule localization" ($P_{GO}\sim 2\times 10^{-4}$), 55 of the 137 correlated genes that participate in "cellular component organization" and 14 of the 19 correlated genes that contribute to "cellular protein organization" ($P_{GO}\sim 5\times 10^{-3}$ in both cases). Most of these genes do not belong to identical operons.

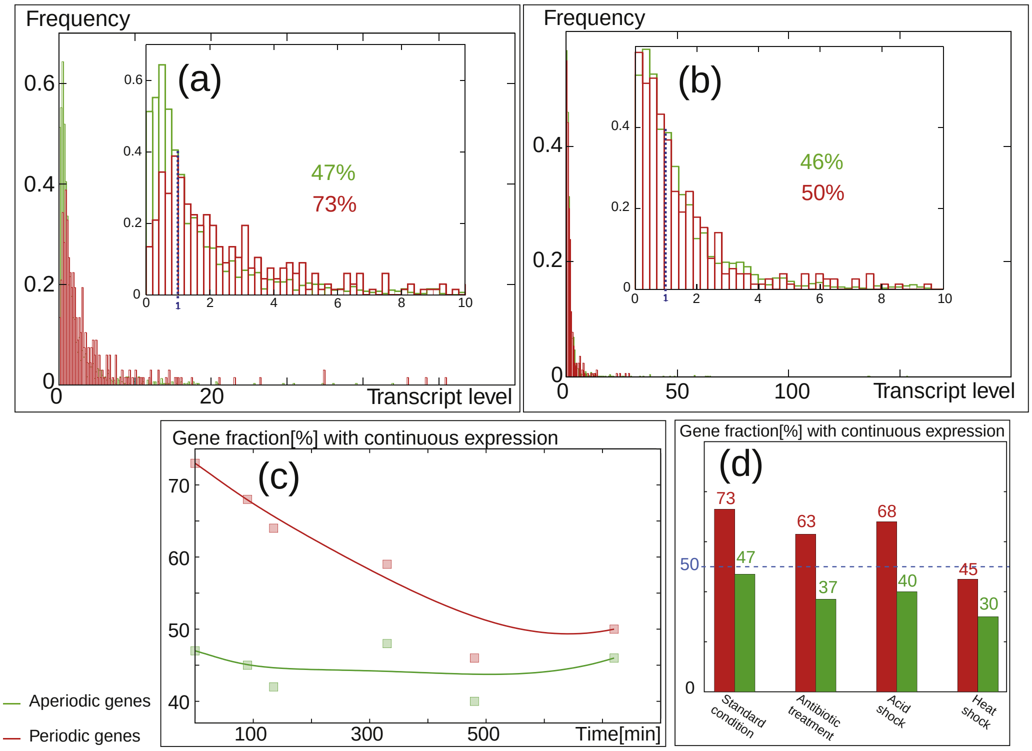

Fig. 7. Expression of periodic genes that are neither essential for cell viability nor involved in housekeeping functions. (a and b) Histograms of the absolute levels of *E. coli* transcripts. *E. coli* cells were grown in minimal glucose medium.[35] Time zero was when the culture started to grow exponentially. Early stationary phase started approximately at 90 min. Late stationary phase was sampled at 720 min. The absolute transcript numbers per cell were measured and stored.[36] Because essential and housekeeping genes tend to be highly expressed, they are removed from the initial set of correlated genes. From these nonessential and non-housekeeping genes, two sets are derived, containing 344 periodic (red) and the remaining 1570 aperiodic (green) genes (Materials and Methods). The insets show a magnification of the histograms for transcript numbers between 0 and 10. Percentages indicate the fraction of genes that are expressed at medium to high levels, which are defined as those with more than one transcript per cell on average (blue vertical broken line). (a) Exponential growth phase. Of the periodic genes, 73% are expressed at medium to high levels, compared to 47% for the aperiodic ones. (b) Late stationary phase. In both sets, about 50% of the genes have less than one transcript per cell (low expression). (c) Temporal evolution of transcript levels as a function of growth phases. The fraction of expressed genes at medium to high levels is plotted *versus* growth time. (d) Fraction of periodic and aperiodic genes with medium to high levels of expression under three different stress conditions. Continuous expression of periodic genes is strongly affected by a heat shock while it is little affected by an acid shock or an antibiotic treatment.

Table 3. Statistics on the transcriptional regulation of periodic genes

a

	TF	Number of correlated genes (n_{reg})	Number of periodic genes (n_{per})	$p(N \geq n_{per})$	$p(N \leq n_{per})$
Enrichment analysis of ssc genes (n_{tot} = 2254)					
	ALL TFs	966	159	1	4.6E-10
	CRP	300	35	1	1.6E-07
	Lrp*	46	0	1	6.4E-06
	NarP	46	0	1	6.4E-06
	Fis*	110	8	1	1.1E-05
	ModE	39	0	1	4.0E-05
	NarL	87	6	1	6.3E-05
	IHF*	159	19	1	2.8E-04
	ArgR	27	0	1	9.2E-04
	FruR	58	5	1	4.3E-03
	SoxS	18	0	1	9.6E-03
	FhlA	17	0	1	0.012
	PhoP	29	2	1	0.025
	AraC	14	0	1	0.027
	FadR	12	0	1	0.045
	Rob	12	0	1	0.045
	MarA	19	1	1	0.049
	PaaX	11	0	1	0.059
	H-NS*	67	10	0.96	0.078
	PdhR	16	1	0.98	0.092
	ArgP	9	0	1	0.098
	NrdR	9	0	1	0.098
	RbsR	6	6	1.3E-04	1
	Zur	4	4	2.6E-03	1
Enrichment analysis of essential ssc genes (n_{tot} = 250)					
	ALL TFs	85	34	1	1.1E-03
	Fis*	11	1	1	2.3E-03
	LexA	16	3	1	3.4E-03
	DnaA	7	0	1	3.9E-03
	NsrR	6	0	1	8.8E-03
	ArgP	5	0	1	0.02
	H-NS*	5	0	1	0.02
	CRP	13	3	1	0.021
	SdiA	4	0	1	0.043
	ArcA	18	16	1.4E-03	1
	FNR	24	18	0.023	1
	NagC	4	4	0.083	1
Enrichment analysis of non-essential ssc genes (n_{tot} = 2004)					
	ALL TFs	881	125	1	8.4E-07
	NarP	46	0	1	5.7E-05
	Lrp*	46	0	1	5.7E-05
	CRP	287	32	1	9.7E-05
	ModE	39	0	1	2.6E-04
	NarL	85	5	1	4.4E-04
	Fis*	99	7	1	6.1E-04
	IHF*	153	15	1	1.1E-03
	ArgR	25	0	1	5.1E-03
	FNR	177	21	1	6.0E-03
	FruR	54	4	1	0.015
	SoxS	17	0	1	0.028
	FhlA	17	0	1	0.028
	AraC	14	0	1	0.053
	NagC	21	1	0.99	0.071
	PhoP	28	2	0.98	0.078
	Rob	12	0	1	0.08
	PaaX	11	0	1	0.099
	RbsR	6	6	4.4E-05	1
	Zur	4	4	1.3E-03	1

Please cite this article as: Junier, I. *et al.*, Genomic Organization of Evolutionarily Correlated Genes in Bacteria: Limits and Strategies, *J. Mol. Biol.* (2012), doi:10.1016/j.jmb.2012.03.009

Table 3. (*continued*)

	TF	Number of correlated genes (n_{reg})	Number of periodic genes (n_{per})	$p(N \geq n_{per})$	$p(N \leq n_{per})$
Enrichment analysis of non-essential ssc genes (n_{tot} = 2004)					
	CsgD	13	7	4.9E-03	1
	BirA	5	4	5.4E-03	1
	DcuR	5	4	5.4E-03	1
	LexA	24	10	8.1E-03	1
	FlhDC	68	20	0.022	0.99
	LeuO	8	4	0.047	0.99
	OxyR	15	6	0.047	0.99
	DnaA	6	3	0.086	0.99
	AsnC	3	2	0.094	0.99

b

	Sigma factor	Number of correlated genes (n_{reg})	Number of periodic genes (n_{per})	$p(N \geq n_{per})$	$p(N \leq n_{per})$
Enrichment analysis of ssc genes (n_{tot} = 2254)					
	Sigma70	1354	305	6.0	0.46
	Sigma54	112	7	1	1.7E-06
	Sigma38	256	32	1	9.6E-06
	Sigma32	213	67	0.0012	1
	Sigma28	116	29	0.3	077
	Sigma24	221	49	0.6	0.46
	Sigma19	4	0	1	0.36
	None	715	174	0.11	0.91
Enrichment analysis of essential ssc genes (n_{tot} = 250)					
	Sigma70	162	68	1	1.40E-07
	Sigma54	2	1	0.79	0.71
	Sigma38	26	3	1	2.80E-06
	Sigma32	22	12	0.57	0.61
	Sigma28	11	10	2.9359e-06	1
	Sigma24	26	14	0.59	0.57
	Sigma19	0	0	-	-
	None	88	52	0.15	0.91
Enrichment analysis of non-essential ssc genes (n_{tot} = 2004)					
	Sigma70	1192	237	0.1	0.92
	Sigma54	110	6	1	2.97E-05
	Sigma38	230	29	1	4.70E.03
	Sigma32	191	55	3.20e-03	1
	Sigma28	105	19	0.63	0.47
	Sigma24	195	35	0.67	0.4
	Sigma19	4	0	1	0.43
	None	611	122	0.23	0.81

(a) Genes regulated by at least one TF. In each table, the second column displays the number (n_{reg}) of TF-regulated genes that are contained in the set indicated in orange. Three sets are studied: the set of correlated genes, the set of essential correlated genes and the complementary set of nonessential correlated genes. The third column indicates the number (n_{per}) of TF-regulated genes that are periodic. Given n_{tot} the total number of genes (first column), the fourth column (respectively, the fifth column) shows the probabilities to obtain more (respectively, less) than n_{per} periodic genes in a process where n_{reg} genes are randomly drawn without replacement among the initial set of n_{tot} genes (Materials and Methods). Asterisks indicate TFs known to be nucleoid-associated proteins. Only TF showing p-values lower than 0.1 are reported (see Supplementary Table 5 for full list of TFs). (b) Same study but analyzing the regulation of Sigma factors. Note that the p-values, in order to be well defined, correspond to probabilities for observing *at least* more (or less) events in the random process. That is, $p(N \geq n_{per}) + p(N \leq n_{per})$ is *a priori* not equal to 1, in contrast to $p(N > n_{per}) + p(N \leq n_{per})$ or $p(N \geq n_{per}) + p(N < n_{per})$. In all tables, the red colors indicate p-values that are lower than 0.05.

Note that these cellular functions remain statistically significant when considering an annotation of TUs instead of an annotation of genes, that is, by eliminating the repetition of GO terms within the same TU. A careful analysis of the translation process actually reveals that periodic genes gather in the largest (periodic) operons (Supplementary Material).

Essentiality properties of periodic genes

To assess whether these 511 periodic correlated genes either are essential for cell viability or perform housekeeping functions, they were compared to relevant and widely recognized data sets. Gil *et al.* published a comprehensive set of genes involved in the maintenance of the basal cellular functions, the "housekeeping core" of all bacterial cells.[32] Of these 206 housekeeping genes, 180 belong to the initial data set of correlated genes used here; 98 are periodic (HD: $P_{\mathrm{HD}} \sim 10^{-22}$). Most housekeeping genes are actually essential, that is, their inactivation is lethal for bacteria. Among the 303 essential genes of *E. coli* reported in the Keio collection,[33] 250 belong to the initial data set of correlated genes; 135 are periodic ($P_{\mathrm{HD}} \sim 10^{-29}$). Altogether, 324 correlated genes are either essential or classified as housekeeping; 167 are periodic ($P_{\mathrm{HD}} \sim 10^{-35}$). In the same spirit, a set of 258 bacterial genes was identified on the basis of a strong cross-species persistence and a tendency for chromosomal proximity.[34] Of them, 248 belong to the initial data set of correlated genes; 130 are periodic ($P_{\mathrm{HD}} \sim 10^{-22}$). In sum, essential and housekeeping genes are highly enriched in genes belonging to the periodic set. Reciprocally, about one-third of the periodic genes either are essential for cell viability or perform housekeeping functions.

Expression properties of periodic genes

Periodic genes are expressed at medium to high levels

Among essential genes, many are highly expressed, for instance, ribosomal protein-encoding genes. Thus, the question arises whether the other periodic genes, those that are neither essential nor involved in housekeeping, are also over-expressed with respect to the rest of the genome. For this purpose, the 324 essential or housekeeping genes were removed from the initial data set. The transcript levels of two gene sets were then compared. The first set is composed of the 344 remaining periodic genes. The second set comprises the 1570 remaining correlated genes that are said to be "aperiodic." Transcript levels are those reported by Allan *et al.* in *E. coli*, grown on standard minimal media under various regimens.[35]

Results show that, during the exponential growth phase under standard conditions, 73% of the periodic genes are expressed at medium to high levels, meaning that they have on average more than one transcript per cell (Fig. 7a). By contrast, the typical transcript number for aperiodic genes is lower than unity, and 53% of them have less than one transcript per cell. This difference does not depend on the carbon source used for growth (data not shown). Instead, it appears to depend on growth phase. Indeed, in deep stationary phase, the transcript level distributions for periodic and aperiodic genes are similar, except that a few aperiodic genes specific to the stationary phase are highly expressed (Fig. 7b). Figure 7c reveals the progressive shift of periodic genes from a regime dominated by medium to high levels of expression under exponential phase to a regime dominated by low expression in prolonged stationary phase. By contrast, the fraction of aperiodic genes that are expressed at medium to high levels stays constant around 45%. Finally, medium to high levels of expression are also observed for the majority of periodic genes after an acid shock or a treatment with the antibiotic ciprofloxacin, but not after a heat shock (Fig. 7d).

Periodic genes are poorly regulated by TFs

Transcriptional regulation properties of periodic genes provide insight into expression patterns and clarify the absence of certain essential genes. First, Table 3a shows that the vast majority of periodic genes are not regulated by TFs (excluding Sigma factors). In addition, genes coding for TFs are also under-represented. Second, Table 3b shows that periodic genes are enriched in genes that are regulated by the heat shock Sigma factor σ 32. In contrast, periodic genes are poorly regulated by σ 38 and σ 54, two Sigma factors that correspond to starvation conditions. Finally, the set of essential periodic genes contains a majority of genes that are not regulated by the primary Sigma factor σ 70.

Altogether, these data show that periodic genes, both essential and nonessential ones, are often constitutively expressed, that is, their expression is generally not regulated by TFs. They also rationalize the absence of genes having a biological function that is over-represented in the set of periodic genes, such as the replication-associated genes *dksA* (regulated by CRP), *seqA* (regulated by HU) or *dnaA* (self-regulating TF).

Periodic organization of E. coli PTUs in distant eubacteria

To probe the generality of the periodicity of PTUs, the study is now extended to other enterobacteria and next to all eubacteria for which a predictive map of operons is available.[37] In each bacterium, the positions of the TUs that contain at least one gene orthologous to one of the 511 periodic correlated genes in *E. coli* are considered. These positions are then analyzed as a single set. Specifically, 47 genomes from 14 species, among the 62 studied enterobacteria from 23 species, show a significant periodicity, that is, with an overall *P*-value lower than 0.05 (score above 1.3; see Materials and Methods). In particular, all 23 strains of *E. coli* show a periodicity around 110–120 kb, except one strain whose *P*-value just misses the significance threshold. However, neither period values nor the ratios of period to genome length are conserved across species (Supplementary Table 3).

Periodicity of PTU orthologs was also observed in remote eubacterial families, covering fairly uniformly all phyla (Supplementary Table 4). Among all 501 non-enterobacterial species, 68 contain at least one strain exhibiting significant periodicity with a score above 1.3: $P_{\mathrm{HD}}\sim 10^{-13}$. By relaxing the threshold score to 1, the number of species increases to 123 species: $P_{\mathrm{HD}}\sim 10^{-20}$. In sum, periodicity appears to be a ubiquitous eubacterial strategy for organizing genomes.

Discussion

Enterobacterial genomes appear to exhibit two major strategies to dispose their evolutionarily correlated genes: genes belonging to small groups under 20 genes tend to cluster along chromosomes, whereas genes in large groups tend to be periodically positioned over long chromosomal distances. In the intermediate size range—about 20–50 genes—mixed strategies are observed. In particular, extensive proximity takes the form of a succession of dense clusters that are sometimes periodically disposed, in which case proximity and periodicity coexist. Based on these observations, we hypothesize that enterobacterial genomes cannot accommodate more than 20 successive genes in one neighborhood. Operons constitute a different means to achieve chromosomal proximity, as they express several proteins from one TU. In this study, operons were reduced to their first cistron to separately analyze non-operonic chromosomal proximity. It is therefore noteworthy that the observed limit of 20 genes corresponds approximately to 20 kb, to the short-range pattern in spatial series of transcriptional activity[38] and to the size range of the longest operons.[37] These comparable operonic and non-operonic sizes suggest that a common organizational principle of bacterial genomes limits neighborhood effectiveness beyond 20 kb. As a working model, this organizational principle could be provided by topological microdomains, that is, superhelicity-independent domains measured *in vivo* to be in the 10- to 20-kb range.[39] The exact nature of this limitation, however, remains mysterious. Several scenarios are possible. For example, the supercoiling constraints generated by transcription[40] could become critical for ∼20-kb-long DNA segments. Another possibility could be an inability to coordinate the expression of many TUs because of insurmountable constraints of DNA folding. These constraints would then be relaxed by splitting the set of contiguous TUs into several smaller sets, which would be located far from each other along the genome.

By attributing periodicity scores to individual genes, we found that close to 12% of *E. coli* genes were periodic. These 511 genes defined in *E. coli* appear to also have periodic orthologous counterparts in representatives of all eubacterial phyla. Thus, periodicity appears to be a ubiquitous strategy for positioning evolutionarily correlated genes. In some phyla, however, periodicity could be asserted only for some, but not all, available sequenced genomes. Notice, then, that the set of correlated genes was originally derived from *E. coli*.[7] Hence, the results for phylogenetically remote organisms may be underestimated by the presence of false negatives.

Periodic genes turn out to predominantly function in spatial organization of cellular components and in macromolecular synthesis. A possible interpretation, then, would be that periodicity preferentially affects highly expressed genes such as those functioning in transcription and translation. However, other highly expressed genes such as those involved in glycolysis are not periodic, while periodic genes include some that are not exceptionally expressed, such as those involved in replication or in spatial cell organization. Along the same line, the sets of "essential," "highly persistent" or "housekeeping" genes[3] overlap with the set of periodic genes only partially, even though significantly.

Instead, we submit that the partly periodic layout of eubacterial genomes favors the overall organization of the cell and more particularly the construction and proper localization of major cell components. The connection between periodic layout and cell organization could be provided by chromosomal conformation. Indeed, we have previously demonstrated the crucial importance of genome layout for specific chromosome folding and conformation.[9] This demonstrated relation suggests that the periodic layout observed here would facilitate chromosome folding such that corresponding genes can be expressed with a good temporal and spatial coordination. More generally, we posit that chromosomal proximity and periodic gene positioning are ubiquitous complementary strategies that contribute to local concentration effects in bacteria. Indeed, gathering molecular partners in space and time can dramatically enhance reaction rate kinetics.[19] Improved kinetics may provide the selective pressure to evolutionarily fix "efficient" genome layouts. This has been demonstrated for protein–protein interactions[19] and for transcriptional regulation in the case of chromosomal proximity.[15,17,41] In an extension of the latter case, named the solenoidal framework, Képès and Vaillant had hypothesized that periodic genome layouts would facilitate the formation of long and regular DNA loops that, by analogy, would also favor the spatial grouping of genes.[4] While in this case the driving force for spatial clustering was TF bivalency and DNA binding site multivalency,[15,17] it is difficult to invoke such causes in the present case due to the paucity of TF-regulated genes in the co-functional periodic set. This suggests that coordination of the expression of these genes, which is necessary given the similarity of their biological functions, is based on a different mechanism. In the light of the elevated local concentrations and the capacity of proteins to interact specifically,[19] we propose that protein–protein interactions, allosteric effects and the coupling between transcription and translation could play a major role.

Regarding the resulting overall organization of the chromosome, any proposed model should be consistent with the different organizations that have been observed *in vivo* (see, e.g., Refs. 29, 42 and 43). In this regard, we note that the relatively large periods discussed both in this article and in Ref. 7 may be related to the high-order organization of DNA that has recently been observed in *E. coli*.[43] In any case, the relation between these high-order

Table 4. GO analysis of the groups showing both strong periodicity and strong chromosomal proximity (upper-right corner in Fig. 3c)

GO term	*P*	Frequency in group genes	Frequency in correlated genes
a. Flagellum/chemotaxis enrichment (two groups)			
GO:0040011—locomotion	8.90e-37	33/63 (52.4%)	58/2117 (2.7%)
GO:0001539—ciliary or flagellar motility	5.96e-31	25/63 (39.7%)	34/2117 (1.6%)
GO:0048870—cell motility	2.05e-30	25/63 (39.7%)	35/2117 (1.7%)
GO:0051674—localization of cell	2.05e-30	25/63 (39.7%)	35/2117 (1.7%)
GO:0006928—cellular component movement	4.11e-28	25/63 (39.7%)	40/2117 (1.9%)
GO:0043064—flagellum organization	3.15e-22	18/63 (28.6%)	23/2117 (1.1%)
GO:0006935—chemotaxis	4.31e-21	18/63 (28.6%)	25/2117 (1.2%)
GO:0042330—taxis	4.31e-21	18/63 (28.6%)	25/2117 (1.2%)
GO:0030030—cell projection organization	4.03e-20	18/63 (28.6%)	27/2117 (1.3%)
GO:0009605—response to external stimulus	5.57e-13	19/63 (30.2%)	63/2117 (3.0%)
GO:0071842—cellular component organization at cellular level	1.30e-08	18/63 (28.6%)	92/2117 (4.3%)
GO:0051179—localization	1.05e-07	39/63 (61.9%)	525/2117 (24.8%)
GO:0009296—flagellum assembly	1.90e-07	8/63 (12.7%)	13/2117 (0.6%)
GO:0030031—cell projection assembly	4.34e-07	8/63 (12.7%)	14/2117 (0.7%)
GO:0016043—cellular component organization	1.72e-06	19/63 (30.2%)	137/2117 (6.5%)
GO:0071841—cellular component organization or biogenesis at cellular level	1.41e-05	20/63 (31.7%)	172/2117 (8.1%)
GO:0042221—response to chemical stimulus	4.14e-05	18/63 (28.6%)	148/2117 (7.0%)
GO:0071840—cellular component organization or biogenesis	6.61e-05	20/63 (31.7%)	188/2117 (8.9%)
GO:0051649—establishment of localization in cell	4.94e-03	8/63 (12.7%)	39/2117 (1.8%)
b. Cell cycle/division/shape/wall enrichment (three groups)			
GO:0008360—regulation of cell shape	1.01e-08	14/85 (16.5%)	36/2117 (1.7%)
GO:0007049—cell cycle	1.84e-07	15/85 (17.6%)	51/2117 (2.4%)
GO:0051301—cell division	4.54e-07	15/85 (17.6%)	54/2117 (2.6%)
GO:0006023—aminoglycan biosynthetic process	7.90e-07	13/85 (15.3%)	40/2117 (1.9%)
GO:0006024—glycosaminoglycan biosynthetic process	7.90e-07	13/85 (15.3%)	40/2117 (1.9%)
GO:0009273—peptidoglycan-based cell wall biogenesis	7.90e-07	13/85 (15.3%)	40/2117 (1.9%)
GO:0009252—peptidoglycan biosynthetic process	7.90e-07	13/85 (15.3%)	40/2117 (1.9%)
GO:0044038—cell wall macromolecule biosynthetic process	1.12e-06	13/85 (15.3%)	41/2117 (1.9%)
GO:0042546—cell wall biogenesis	1.12e-06	13/85 (15.3%)	41/2117 (1.9%)
GO:0070589—cellular component macromolecule biosynthetic process	1.12e-06	13/85 (15.3%)	41/2117 (1.9%)
GO:0006022—aminoglycan metabolic process	1.18e-06	14/85 (16.5%)	49/2117 (2.3%)
GO:0030203—glycosaminoglycan metabolic process	1.18e-06	14/85 (16.5%)	49/2117 (2.3%)
GO:0000270—peptidoglycan metabolic process	1.18e-06	14/85 (16.5%)	49/2117 (2.3%)
GO:0010382—cellular cell wall macromolecule metabolic process	1.57e-06	13/85 (15.3%)	42/2117 (2.0%)
GO:0070882—cellular cell wall organization or biogenesis	2.81e-06	14/85 (16.5%)	52/2117 (2.5%)
GO:0071554—cell wall organization or biogenesis	3.69e-06	14/85 (16.5%)	53/2117 (2.5%)
GO:0044036—cell wall macromolecule metabolic process	4.07e-06	13/85 (15.3%)	45/2117 (2.1%)
GO:0065008—regulation of biological quality	5.33e-05	16/85 (18.8%)	85/2117 (4.0%)
GO:0071555—cell wall organization	1.13e-04	10/85 (11.8%)	32/2117 (1.5%)
GO:0007047—cellular cell wall organization	1.13e-04	10/85 (11.8%)	32/2117 (1.5%)
GO:0015920—lipopolysaccharide transport	3.06e-04	5/85 (5.9%)	6/2117 (0.3%)
GO:0071843—cellular component biogenesis at cellular level	8.71e-04	15/85 (17.6%)	91/2117 (4.3%)
GO:0045229—external encapsulating structure organization	1.86e-03	10/85 (11.8%)	42/2117 (2.0%)
GO:0071841—cellular component organization or biogenesis at cellular level	3.98e-03	20/85 (23.5%)	172/2117 (8.1%)
GO:0044085—cellular component biogenesis	4.07e-03	18/85 (21.2%)	143/2117 (6.8%)
GO:0071840—cellular component organization or biogenesis	4.27e-03	21/85 (24.7%)	188/2117 (8.9%)

The tables report terminology enrichment of biological processes with respect to the initial set of correlated genes for the five groups showing the strongest periodicity trends. These five groups have been divided into two categories. The first category (a) shows an enrichment in flagellum/chemotaxis-coding genes. The second category (b) shows an enrichment in regulation of cell shape, cell cycle, cell division and cell wall biogenesis. In each table, the first column indicates the GO term. The second column gives the statistical significance of the enrichment (only $P \leq 10^{-2}$ are reported).The third and fourth columns, respectively, give the frequency of the term in the groups and in the set of correlated genes.

Please cite this article as: Junier, I. *et al.*, Genomic Organization of Evolutionarily Correlated Genes in Bacteria: Limits and Strategies, *J. Mol. Biol.* (2012), doi:10.1016/j.jmb.2012.03.009

structures and those at a smaller scale, such as the 10-kb topological domains,[39] the plectonemic loops induced by negative supercoiling or the ≤20 kb neighborhood effects described here, remains to be understood. In a multiscale view of chromosomal organization, one possibility is that such large structures would encompass an integral number of smaller ones.

Let us finally mention that this chromosome-centric interpretation for periodic genes is further supported by four observations. Firstly, periodic genes organize spatial processes in the cell. In this regard, Montero Llopis *et al.* have recently proposed a chromosome-centric scheme of the cellular organization of mRNA processes in bacteria.[24] In this view, the chromosome layout is used as an organizing template by exploiting the fact that proteins tend to be translated close to their coding genes. This is consistent with the observed periodicity of genes involved in translation[22,44] and can easily be extended to other cellular processes with a strong spatial component, such as replication. In the same spirit, Saberi and Emberly have shown that the cellular location of chromosomes is crucial for the proper spatial patterning of proteins in bacteria.[45] Along these lines, the high connectivity of periodic genes and their strong functional enrichment in cell spatial organization point to the exceptional organizing capacity of these genes. Secondly, the majority of periodic genes are involved in processes that encompass the whole cell. They encode macrocomplex subunits (e.g., replicase or transcriptase) and enzymes that synthesize large macrocomplexes (e.g., 8 of the 13 enzymes that participate to peptidoglycan synthesis in the cell wall). Moreover, the few cTU groups exhibiting both strong periodicity and proximity are enriched in genes coding for flagellum biosynthesis, chemotaxis, cell cycle, division, wall biogenesis and shape regulation (Table 4). Thirdly, periodic genes appear to be at the root of cell functioning. One-third of periodic genes are either essential to cell viability or involved in universal maintenance functions. Of the remaining two-thirds, most are expressed at medium to high levels during rapid growth. Fourthly, periodic and proximal genes appear to favor chromosome conformations that enhance their own transcriptional control in dedicated factories.[9] The ubiquity of these strategies in the eubacterial world despite constant gene shuffling on evolutionary timescales further suggests that they confer a selective advantage.

Summary

It is important to investigate genome organization to understand the evolutionary forces that have shaped chromosomes and to aid us with the design of synthetic ones. Comparison of bacterial genomes has highlighted the conservation of gene order over short chromosomal regions. More recently, gene periodic positioning along the whole chromosome has been observed. The rationale balancing these different organization scales remains to be elucidated.

In this article, we investigate the compromise between chromosomal proximity and periodicity within groups of evolutionarily correlated genes in bacteria. We observe that clustering along DNA is limited to groups containing less than the typical number (20) of genes contained in topological microdomains of chromosomes. Beyond this limit, groups increasingly adopt a periodic type of organization.

We find that periodic genes constitute 12% of the *E. coli* genome. These genes predominantly function in macromolecular synthesis and spatial organization of cellular components. They are enriched in essential and housekeeping genes and tend to be constitutively expressed. Their homologs are also periodically spaced in diverse bacteria.

Altogether, our findings suggest that periodic genome layout optimizes the construction and the spatial organization of cellular components. DNA topology might define the range for which DNA neighborhood optimizes biochemical interactions.

Materials and Methods

Solenoidal coordinate method

The periodic trend of a cTU group is assessed using the solenoidal coordinate method. In this method, the score at a given period reflects the likelihood for the data set to present a periodic pattern with this period. A high score stems from (i) the remarkable alignment properties of periodic positions when they are represented in a solenoidal coordinate system with the right period (Supplementary Fig. 8) and (ii) the use of an information-theoretic measure *à la* Shannon that rewards both exceptionally dense and void regions of the solenoid (see Ref. 27 for details). Specifically, scores at a given period are the co-logarithms of probabilities that the voids and the clusters on the face view of the corresponding solenoid (Supplementary Fig. 8) could have been obtained with randomized data—for example, a score equal to 8 (as in Supplementary Fig. 2) corresponds to a P-value of 10^{-8}. Figure 2 reports scores that have been averaged over all single groups.

The period equal to full chromosome length plays a singular role in the analysis. Indeed, for this period, the "solenoid" is composed of only one loop. Thus, the analysis does not bear on periodicity tendency but rather on the tendency for TUs to cluster along the chromosome. Accordingly, scores at this peculiar period are referred to as proximity scores.

Period-dependent positional scores

At a given period and for a given cTU group, a positional score is allocated to each TU. It is related to the density of TUs, on the face view of the solenoid, in the vicinity of the TU of interest (Supplementary Fig. 8); the higher the density, the higher the score. The score reported here is defined as the co-logarithm of the probability that the density in the vicinity of the TU could be obtained with randomized data.

Periodicity detection among all phyla

Positions of the genes orthologous to the 511 periodic genes of *E. coli* were determined in the 760 eubacterial strains for which a genome sequence and an operonic map were available.[37] For each strain, the corresponding TUs were analyzed as a single set. To this end, periodic positioning was assessed by scanning periods from 5 kb to full chromosome length. Two pitfalls must be considered to properly evaluate the significance of any observed periodicity. First, each tested period represents one additional possibility for observing an exceptional score. Second, a high proximity score means that several TUs are next to each other; if one TU is in periodic alignment with other TUs, its neighbors will also tend to be, thus boosting the periodicity score. Hence, significance must be corrected for both the multiplicity of tested periods and the possible presence of chromosomal proximity. To correct for chromosomal proximity at each given period, the probability of obtaining a higher significativity score for randomly generated positions is computed while imposing a proximity score as good as that in the original TU set. To correct for multiple tested periods, this probability is then subjected to a Bonferroni procedure, considering that i periods ($P_1,\ldots,P_{i-1},P_i$) have been tested at P_i—the Bonferroni procedure is a stringent type of correction that drastically reduces the rate of false positives at the expense of increasing false negatives.[46] Calling S_i the significativity score at P_i and n_p the number of tested periods in the spectrum, the significativity score of the TU set reads $S_{\text{set}} = \underset{i=1,\ldots,n_p}{max}\{S_i\}$.

Hypothesis testing procedures

The P-value (e.g., P_{GO} or P_{HD}) associated to a property P and that concerns a subset of n periodic genes corresponds to the probability of obtaining more than n genes with property P when drawing 511 genes among the 2254 initial ones according to a random process without replacement. This is computed using the HD, which is the probability distribution associated to the random process. The p-value is given by $\sum_{k=n}^{511} \frac{\binom{Np}{k}\binom{2254-Np}{511-k}}{\binom{2254}{k}}$, where $\binom{\bullet}{\bullet}$ indicates binomial coefficients. P_{GO} further includes a Bonferronni correction due to the repetitive nature of the enrichment test.[31]

In the same spirit, the p-values associated to the fact that, among all 501 non-enterobacterial species, n_{sp} contains at least one strain exhibiting significant periodicity (defined by a probability p for being observed) read $\sum_{k=n_{sp}}^{501}\binom{501}{k}p^k(1-p)^{501-k}$ (binomial distribution).

Bacterial genome annotation

Genes orthologous to the correlated genes of *E. coli* were identified using the Cluster of Orthologous Gene (COG) classification.[47] Operons were predicted using the MicrobesOnline Web site for comparative genomics.[37]

Gene expression

The gene expression data come from Allen *et al.*[35] They contain transcript numbers obtained from Affymetrix GeneChip arrays in different minimal media, at different times of the growth phase and under different stress conditions. They are available in the ASAP database.[36] More precisely, they concern *E. coli* cells that were exponentially growing at 37 °C in four Mops-buffered minimal media that differed by their carbon source: glucose (five arrays), acetate, glycerol or proline (two arrays in each case). Other samples came from late exponential growth phase (90 min after the culture reached the exponential phase; three arrays) and stationary phases (at 135, 330, 480 and 720 min; two arrays in each case) in glucose minimal medium. Additionally, three stress conditions were imposed to bacteria growing exponentially on glucose minimal medium: (i) acid shock at pH 2 with HCl, (ii) antibiotic treatment with ciprofloxacin and (iii) heat shock at 50 °C (two arrays in each case). Samples were collected 10 min later for (i) and (iii) and 30 min later for (ii). Here, we report values averaged over the set of arrays specific to a single condition. Further details about the experimental protocol can be found online†, as well as data‡.

Acknowledgements

We thank L. Jannière for discussions, B. Jester for a careful reading of the manuscript and the referees for useful suggestions. I.J. also thanks R. Guigó for useful suggestions and Région Île-de-France and L'institut des Systèmes Complexes Paris Île-de-France for financial and logistic supports. This work was supported by the Sixth European Research Framework (project number 034952, GENNETEC project), PRES UniverSud Paris, Centre National de la Recherche Scientifique and Genopole.

Supplementary Data

Supplementary data to this article can be found online at doi:10.1016/j.jmb.2012.03.009

† https://asap.ahabs.wisc.edu/glasner/Protocols/DataDefinitionDefinitions.txt

‡ http://asap.ahabs.wisc.edu/asap/experiment_data.php

References

1. Huynen, M. & Bork, P. (1998). Measuring genome evolution. *Proc. Natl Acad. Sci. USA*, **95**, 5849–5856.
2. Tamames, J. (2001). Evolution of gene order conservation in prokaryotes. *Genome Biol.* **2**; RESEARCH0020.
3. Rocha, E. P. C. (2008). The organization of the bacterial genome. *Annu. Rev. Genet.* **42**, 211–233.
4. Képès, F. & Vaillant, C. (2003). Transcription-based solenoidal model of chromosomes. *ComPlexUs*, **1**, 171–180.
5. Képès, F. (2004). Periodic transcriptional organization of the *E. coli* genome. *J. Mol. Biol.* **340**, 957–964.
6. Mercier, G., Berthault, N., Touleimat, N., Képès, F., Fourel, G., Gilson, E. & Dutreix, M. (2005). A haploid-specific transcriptional response to irradiation in *Saccharomyces cerevisiae*. *Nucleic Acids Res.* **33**, 6635–6643.
7. Wright, M., Kharchenko, P., Church, G. & Segrè, D. (2007). Chromosomal periodicity of evolutionarily conserved gene pairs. *Proc. Natl Acad. Sci. USA*, **104**, 10559.
8. Mathelier, A. & Carbone, A. (2010). Chromosomal periodicity and positional networks of genes in *Escherichia coli*. *Mol. Syst. Biol.* **6**, 366.
9. Junier, I., Martin, O. & Képès, F. (2010). Spatial and topological organization of DNA chains induced by gene co-localization. *PLoS Comput. Biol.* **6**, e1000678.
10. Schoenfelder, S., Sexton, T., Chakalova, L., Cope, N. F., Horton, A., Andrews, S. *et al.* (2010). Preferential associations between co-regulated genes reveal a transcriptional interactome in erythroid cells. *Nat. Genet.* **42**, 53–61.
11. Xu, M. & Cook, P. R. (2008). Similar active genes cluster in specialized transcription factories. *J. Cell Biol.* **181**, 615–623.
12. Cook, P. (2002). Predicting three-dimensional genome structure from transcriptional activity. *Nat. Genet.* **32**, 347–352.
13. Osborne, C. S., Chakalova, L., Brown, K. E., Carter, D., Horton, A., Debrand, E. *et al.* (2004). Active genes dynamically colocalize to shared sites of ongoing transcription. *Nat. Genet.* **36**, 1065–1071.
14. Müller, J., Oehler, S. & Müller-Hill, B. (1996). Repression of lac promoter as a function of distance, phase and quality of an auxiliary lac operator. *J. Mol. Biol.* **257**, 21–29.
15. Dröge, P. & Müller-Hill, B. (2001). High local protein concentrations at promoters: strategies in prokaryotic and eukaryotic cells. *BioEssays*, **23**, 179–183.
16. Kuhlman, T., Zhang, Z., Saier, M. H. & Hwa, T. (2007). Combinatorial transcriptional control of the lactose operon of *Escherichia coli*. *Proc. Natl Acad. Sci. USA*, **104**, 6043–6048.
17. Vilar, J. M. G. & Leibler, S. (2003). DNA looping and physical constraints on transcription regulation. *J. Mol. Biol.* **331**, 981–989.
18. Fraser, P. & Bickmore, W. (2007). Nuclear organization of the genome and the potential for gene regulation. *Nature*, **447**, 413–417.
19. Kuriyan, J. & Eisenberg, D. (2007). The origin of protein interactions and allostery in colocalization. *Nature*, **450**, 983–990.
20. Azam, T. A., Hiraga, S. & Ishihama, A. (2000). Two types of localization of the DNA-binding proteins within the *Escherichia coli* nucleoid. *Genes Cells*, **5**, 613.
21. Lewis, P. J., Thaker, S. D. & Errington, J. (2000). Compartmentalization of transcription and translation in *Bacillus subtilis*. *EMBO J.* **19**, 710–718.
22. Cabrera, J. & Jin, D. (2006). Active transcription of rRNA operons is a driving force for the distribution of RNA polymerase in bacteria: effect of extrachromosomal copies of rrnB on the *in vivo* localization of RNA polymerase. *J. Bacteriol.* **188**, 4007–4014.
23. Berger, M., Farcas, A., Geertz, M., Zhelyazkova, P., Brix, K., Travers, A. & Muskhelishvili, G. (2009). Coordination of genomic structure and transcription by the main bacterial nucleoid-associated protein HU. *EMBO Rep.* **11**, 59–64.
24. Montero Llopis, P., Jackson, A. F., Sliusarenko, O., Surovtsev, I., Heinritz, J., Emonet, T. & Jacobs-Wagner, C. (2010). Spatial organization of the flow of genetic information in bacteria. *Nature*, **466**, 77–81.
25. Taniguchi, Y., Choi, P. J., Li, G. W., Chen, H., Babu, M., Hearn, J. *et al.* (2010). Quantifying E. coli proteome and transcriptome with single-molecule sensitivity in single cells. *Science*, **329**, 533–538.
26. Pellegrini, M., Marcotte, E. M., Thompson, M. J., Eisenberg, D. & Yeates, T. O. (1999). Assigning protein functions by comparative genome analysis: protein phylogenetic profiles. *Proc. Natl Acad. Sci. USA*, **96**, 4285–4288.
27. Junier, I., Hérisson, J. & Képès, F. (2010). Periodic pattern detection in sparse boolean sequences. *Algorithms Mol. Biol.* **5**, 31.
28. Koonin, E. V. (2005). Orthologs, paralogs, and evolutionary genomics. *Annu. Rev. Genet.* **39**, 309–338.
29. Espeli, O. & Boccard, F. (2006). Organization of the *Escherichia coli* chromosome into macrodomains and its possible functional implications. *J. Struct. Biol.* **156**, 304–310.
30. Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., Cherry, J. M. *et al.* (2000). Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat. Genet.* **25**, 25–29.
31. Boyle, E. I., Weng, S., Gollub, J., Jin, H., Botstein, D., Cherry, J. M. & Sherlock, G. (2004). GO:TermFinder—open source software for accessing Gene Ontology information and finding significantly enriched Gene Ontology terms associated with a list of genes. *Bioinformatics*, **20**, 3710–3715.
32. Gil, R., Silva, F. J., Peretó, J. & Moya, A. (2004). Determination of the core of a minimal bacterial gene set. *Microbiol. Mol. Biol. Rev.* **68**, 518–537.
33. Baba, T., Ara, T., Hasegawa, M., Takai, Y., Okumura, Y., Baba, M. *et al.* (2006). Construction of *Escherichia coli* K-12 in-frame, single-gene knockout mutants: the Keio collection. *Mol. Syst. Biol.* **2**, 2006.0008.
34. Fang, G., Rocha, E. P. C. & Danchin, A. (2008). Persistence drives gene clustering in bacterial genomes. *BMC Genomics*, **9**, 4.
35. Allen, T. E., Herrgard, M. J., Liu, M., Qiu, Y., Glasner, J. D., Blattner, F. R. & Palsson, B.Ø. (2003). Genome-scale analysis of the uses of the *Escherichia coli* genome: model-driven analysis of heterogeneous data sets. *J. Bacteriol.* **185**, 6392–6399.

Please cite this article as: Junier, I. *et al.*, Genomic Organization of Evolutionarily Correlated Genes in Bacteria: Limits and Strategies, *J. Mol. Biol.* (2012), doi:10.1016/j.jmb.2012.03.009

36. Glasner, J. D., Liss, P., Plunkett, G., Darling, A., Prasad, T., Rusch, M. *et al.* (2003). ASAP, a systematic annotation package for community analysis of genomes. *Nucleic Acids Res.* **31**, 147–151.
37. Alm, E. J., Huang, K. H., Price, M. N., Koche, R. P., Keller, K., Dubchak, I. L. & Arkin, A. P. (2005). The MicrobesOnline web site for comparative genomics. *Genome Res.* **15**, 1015–1022.
38. Jeong, K. S., Ahn, J. & Khodursky, A. B. (2004). Spatial patterns of transcriptional activity in the chromosome of *Escherichia coli*. *Genome Biol.* **5**, R86.
39. Postow, L., Hardy, C. D., Arsuaga, J. & Cozzarelli, N. R. (2004). Topological domain structure of the *Escherichia coli* chromosome. *Genes Dev.* **18**, 1766–1779.
40. Deng, S., Stein, R. A. & Higgins, N. P. (2005). Organization of supercoil domains and their reorganization by transcription. *Mol. Microbiol.* **57**, 1511–1521.
41. Kolesov, G., Wunderlich, Z., Laikova, O. N., Gelfand, M. S. & Mirny, L. A. (2007). How gene order is influenced by the biophysics of transcription regulation. *Proc. Natl Acad. Sci. USA*, **104**, 13948.
42. Wang, X., Liu, X., Possoz, C. & Sherratt, D. J. (2006). The two Escherichia coli chromosome arms locate to separate cell halves. *Genes Dev.* **20**, 1727–1731.
43. Wiggins, P. A., Cheveralls, K. C., Martin, J. S., Lintner, R. & Kondev, J. (2010). Strong intranucleoid interactions organize the *Escherichia coli* chromosome into a nucleoid filament. *Proc. Natl Acad. Sci. USA*, **107**, 4991–4995.
44. Cabrera, J. E. & Jin, D. J. (2003). The distribution of RNA polymerase in *Escherichia coli* is dynamic and sensitive to environmental cues. *Mol. Microbiol.* **50**, 1493–1505.
45. Saberi, S. & Emberly, E. (2010). Chromosome driven spatial patterning of proteins in bacteria. *PLoS Comput. Biol.* **6**, e1000986.
46. Shaffer, J. (1995). Multiple hypothesis testing. *Annu. Rev. Psychol.* **46**, 561–584.
47. Tatusov, R., Natale, D., Garkavtsev, I., Tatusova, T., Shankavaram, U., Rao, B. S. *et al.* (2001). The COG database: new developments in phylogenetic classification of proteins from complete genomes. *Nucleic Acids Res.* **29**, 22.