Molecular Systems Biology 6; Article number 366; doi:10.1038/msb.2010.21
Citation: ***Molecular Systems Biology*** **6:366**

www.molecularsystemsbiology.com

molecular systems biology

Chromosomal periodicity and positional networks of genes in *Escherichia coli*

Anthony Mathelier[1,2] and Alessandra Carbone[1,2,*]

[1] UPMC Univ Paris 06, FRE3214, Génomique Analytique, 15 rue de l'École de Médecine, Paris, France and [2] CNRS, FRE3214, Génomique des Microorganismes, Paris, France
* Corresponding author. Génomique Analytique, Université Pierre et Marie Curie-Paris 6, FRE3214 CNRS-UPMC, 15 rue de l'École de Médecine, Paris 75006, France. Tel.: +33 01 44 27 73 45; Fax: +33 01 44 27 73 36; E-mail: alessandra.carbone@lip6.fr

Received 17.4.09; accepted 18.3.10

The structure of dynamic folds in microbial chromosomes is largely unknown. Here, we find that genes with a highly biased codon composition and characterizing a *functional core* in *Escherichia coli K12* show to be periodically distributed along the arcs, suggesting an encoded three-dimensional genomic organization helping functional activities among which are translation and, possibly, transcription. This extends to functional classes of genes that are shown to systematically organize into two independent positional gene networks, one driven by metabolic genes and the other by genes involved in cellular processing and signaling. We conclude that functional reasons justify periodic gene organization. This finding generates new questions on evolutionary pressures imposed on the chromosome. Our methodological approach is based on single genome analysis. Given either core genes or genes organized in functional classes, we analyze the detailed distribution of distances between pairs of genes through a parameterized model based on signal processing and find that these groups of genes tend to be separated by a regular integral distance. The methodology can be applied to any set of genes and can be taken as a footprint for large-scale bacterial and archaeal analysis.
Molecular Systems Biology **6**: 366; published online 11 May 2010; doi:10.1038/msb.2010.21
Subject Categories: functional genomics; computational methods
Keywords: chromosome structure; COGs classes; core genes; *Escherichia coli K12*; essential genes

Introduction

Multiple experiments have shown that several hundreds of genes are essential for the life of a microbial organism, in the sense that the organism would not survive without them. This estimation is dependent on the biological complexity and environmental specificity of the organism. On the other hand, several important genes may not be directly involved in growth, but rather in conditions of starvation or stress, and their loss may lead to such a lower degree of fitness that their deletion will never be fixed in natural populations. We address the question of whether these genes, forming a functional *core* of genes for the organism, are organized in regularly spaced groups within the *Escherichia coli K12* genome, possibly depending on transcription regulation patterns or on common functional activities of genes in the groups. Both these possibilities explaining the distribution of genes as a product of structural periodicity are attractive. In fact, the localization of certain core genes along structural chromosomal 'faces' would have the advantage of creating spatial chromosomal subregions in which core genes could be accessed by limited diffusion of RNA polymerase or RNA polymerase fixed in factories (Sinden and Pettijohn, 1981; Cook, 2002; Thanbichler and Shapiro, 2006). The solenoid model (Képès and Valliant, 2003; Képès, 2004) and the rosettes model (Cook, 2002) of chromosomes have been proposed as possible functional and spatial organizations of chromosomes. The idea behind these models is to bring close in space different genes through an encoded three-dimensional genomic organization. The solenoid model organizes loops of DNA along a solenoidal three-dimensional arrangement and the rosettes model organizes DNA loops radially in a flower-like three-dimensional structure.

Evidence for spatial organization of genes along chromosomes has been already noticed for *E. coli K12*. It has been shown that groups of genes regulated by the same transcription factors reveal chromosomal periodicity (Képès, 2004), that spatial series of transcriptional activity exist (Jeong *et al*,

2004), and that evolutionarily conserved gene pairs also reveal chromosomal periodicity (Wright *et al*, 2007). These analyses start on pairs of genes (co-evolving pairs or gene-regulator pairs), which satisfy co-localization based on (already known) interactions and check whether pairs relative positioning along the chromosome follows a periodic pattern. Contrary to what has been done earlier, we analyze sets of genes that have undergone evolutionary or functional constraints but that are not pairwise organized, and ask whether a large-scale signal of periodic gene organization exists. The ultimate aim is to be able to extract numerical properties from the signal and use those to derive biological insights on gene organization, if any.

Here, we consider core genes to be those genes that have undergone an important evolutionary pressure and that are especially encoded with a very biased codon composition. On the basis of a computational method allowing for the identification of highly biased genes (Carbone, 2006), we define a pool of core genes, some of which are conserved across many species, some depend on the environmental living conditions of the organism, some are involved in the stress response, and others have no yet identified function. These predicted core genes cover roughly 10% of all genes in *E. coli K12*, they are expected to be either highly expressed or rapidly expressed when needed (Grantham *et al*, 1980; Sharp and Li, 1987) and they tend to be overrepresented in the class of genes deemed to be essential for *E. coli* (Gerdes *et al*, 2003; Carbone, 2006). An important property of core genes pointed out in Carbone (2006) is that they cover all the spectrum of microbial functions. This means that for any functional class of genes, some representative of the class belongs to the functional core. Consequently, we reasoned, the three-dimensional chromosomal arrangement of these genes may be important to fulfill basic functional responses.

Our goal is to test the existence of periodicity for core genes and other sets. On the back of our mind, there are two main biological hypothesis that we try to approach. First, that a structured localization of core genes can help them to reach fast expression. Second, that the organization of core genes is justified by the joint functional activity of closely localized genes. Both these hypothesis support the idea that significant similarities in gene activities extend beyond the length of an operon and that local patterns of co-expression are dependent on DNA supercoiling. Our analysis highlights that the pool of genes and rRNAs involved in transcription and translation (after Clusters of Orthologous Groups (COGs) classification; Tatusov *et al*, 2003) is organized by a periodic distribution of 117 kb period as found earlier in Wright *et al* (2007). But the striking result is that a stronger signal of periodicity appears at 33 kb when considering core genes, the set of genes involved in metabolism and, in particular, the larger set of genes covering all functional classes classified in COGs (Tatusov *et al*, 2003). Core genes are (up to a few unclassified genes) the highly biased subset of this larger set. We found that gene periodic arrangement systematically organizes functional classes into two independent positional gene networks. The two networks are out of phase among each other but they preserve a 33-kb period as a common parameter.

Beside core genes and functional classes in COGs, we tested several other data sets of genes with different functional importance, some of them based on *in silico* analysis and some on experimental evidence. All sets confirmed the existence of a large-scale periodic gene organization at 33 kb.

Results

We identified a set of core genes for *E. coli K12* by applying a computational method introduced and validated on gene knockout experimental data in Carbone (2006). This set is characterized by genes having a highly biased codon composition and contains ubiquitous genes, non-orthologous genes, environment-specific genes, genes involved in the stress response, and genes with no identified function but highly likely to be essential for the cell. All genes in the set have been selected to have a high Self-Consistent Codon Index (SCCI), where SCCI is highly correlated in *E. coli K12* to the known Codon Adaptation Index (Carbone *et al*, 2003). An asymmetric distribution of core genes preferring the leading strands is observed and the uneven distribution is even stronger for the top 100 core genes. Also, core genes are slightly more numerous around the ORI than the TER (see Supplementary Table I).

We studied the distribution of core genes on two types of chromosomal models (Figure 1; Supplementary Figure 1). The models are general and they do not depend on genes to belong to the functional core. For this, we describe them for a generic set of genes X, where X will become a set of core genes in our first usage of the models. The first model, called *circular model*, does not make any hypothesis on the chromosomal structure and it considers minimal distances between all pairs of genes in X along the circular chromosome. On this basis, we look for a statistically significant period within the distribution of all minimal distances. Notice that, given n genes in X, the number of pairs is about n^2. The second type of models, called *arc-based models*, assume the origin (ORI) and the terminus (TER) of the *E. coli K12* chromosome to have a structurally privileged function in chromosomal architecture and, in consequence, that genes in X belonging to the left and to the right arcs of the chromosome organize independently. We defined three different arc-based models, which are dependent on gene location in the chromosome and consider X to be either the set of genes located on the lagging strands, or on the leading strands (see Supplementary Figure 1) or anywhere along the chromosome, respectively. In this latter case, we speak about 'full arc-based model.' For each arc-based model, we considered all distances between pairs of genes in X, which are located on the right arc and all distances between pairs of

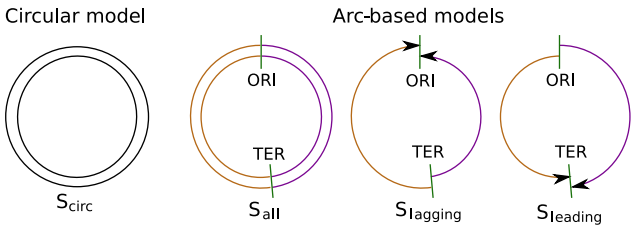

Figure 1 Different models of the circular chromosome. Leading and lagging strands are described, with the arrows coding for the 5′–3′ direction. Chromosomal arcs have distinguished color (purple, brown). S_{circ}, S_{all}, $S_{lagging}$, $S_{leading}$ are the sets of distances between genes associated to the models.

genes that are located on the left arc, and looked for a statistically significant period within the union of these two sets of distances. Intuitively, the arc-based models relative to lagging and leading strands have been introduced to analyze the positional origin of the periodic signal relative to chromosomal strands. They help the understanding of the full arc-based model.

We analyzed the detailed distribution of distances between pairs of core genes for the circular model and for all arc-based models and found that core genes tend to be separated by a regular integral distance of 33 kb. This period displays the strongest signal for the circular model and for the arc-based models defined over the whole chromosome and over the lagging strands. The analysis of core genes for the full arc-based model is illustrated in Figure 2B (see Supplementary Figure 3 for the analysis on the circular model). As expected, the shape of the histogram describing distribution of distances in arc-based models looks roughly like a 'triangle,' where most gene pairs are combined with short distances and a fewer are distantly spaced (Supplementary Figure 2). The number of distances is expected to remain essentially the same across distances (Supplementary Figure 3) for the circular model. To evaluate whether there is a spatial regularity in the organization of core genes along the chromosome, we computed, based on a signal processing parameterized model and a Fast Fourier Transform (FFT) analysis, the spectral response of the distance distributions. Ideally, the best information content we would hope for corresponds to a single period with a sharply detected amplitude of the signal. This would correspond to genes, which are perfectly positioned in a periodic phase along the chromosome. In general, this is not the case, and we should expect several periods, which are associated to amplitudes of variable intensity. Significance of the periods is established by comparing the strength of the signal with a random model based on the generation of random genomes that satisfy properties, which are sufficiently close to the ones of the *E. coli K12* genome. Our random genomes resemble the *E. coli K12* genome: the distribution of intergenic regions is the same as for the *E. coli K12* chromosome, and the number of genes and the distribution of gene lengths are the same for corresponding leading and lagging strands. The statistical significance computed on such random structures is more demanding than the one based on random structures generated by allowing chromosomal length to grow (due to length variability of intergenic regions and genes).

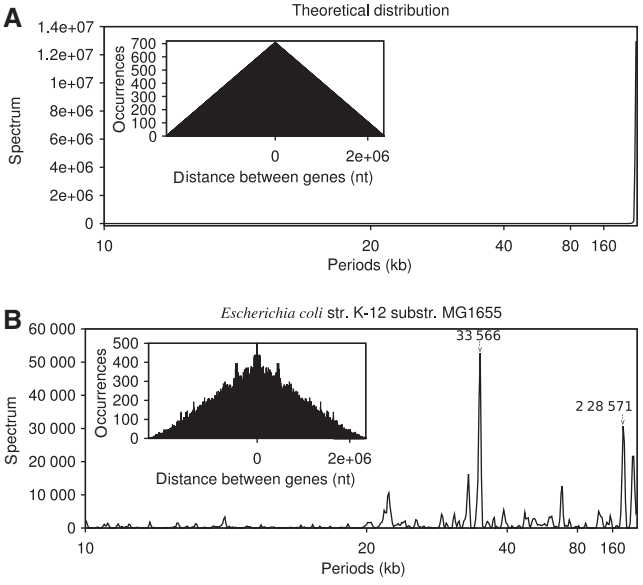

Figure 2 Histograms and periodograms are computed on the distance distribution of *E. coli K12* core genes (with the full arc-based model; **B**) and on the theoretical distribution (**A**). The original distance distribution has been symmetrized giving origin to an 'isoscele triangle' shape. Detrend is applied to the histogram series in the FFT analysis. *E. coli K12* histogram is truncated at the central column; the maximum *y* value is at 706. Distances near zero are overrepresented, and the corresponding peak has been taken from the periodogram (this peak is on the plot for the ideal distribution). Notice that no periodic signal can be detected from the ideal distribution.

The periodic distance of 33 566 nt between core genes of *E. coli K12* appears with a very pronounced spectral value, which turns out to be a few s.d. away from the mean of the peaks distribution (*P*-value $<10^{-4}$, computed on 10 000 randomizations; the exceptional difference between this spectral peak and the mean value of the peak distribution is measured by a *Z*-score=7.81) for the full arc-based model (Figure 2B; Supplementary Table III). There is another peak that we detect at 228 571 nt (with a lower Z=4.39). A 33-kb periodicity is also detected (but with a weaker signal) by the arc-based model that considers core genes located on lagging strands (P=0.0123, Z=7.13). Core genes on leading strands give rise to much weaker signals (Supplementary Table III). The circular model shows a period of 33 566 nt ($P<10^{-4}$, Z=8.19; Supplementary Table III). Core genes are therefore not randomly spaced along the genome but prefer specific genomic intervals of $k \times 33$ kb.

Because of the way we defined pairs of distances (by taking distances between all pairs of genes in an arc or by taking all minimal distances between genes in the circular chromosome), there are no many different distributions of locations for the genes in *E. coli K12* that could generate the periodic distributions that we found. Any sufficiently large subgroup of genes that is periodically spaced with a period of 33 kb and which is out of phase with the majority of periodically spaced core genes, could negatively interfere with the detection of the signal due to the quadratic effect of the method that looks at the distribution of roughly n^2 possible gene pairs for n given core genes. This means that one should expect genes to be strongly co-localized in a series of positions that are spread across the entire chromosome (see the periodically spaced peaks of the SCCI curve in Figure 3A). We found that core genes tend to be localized in specific regularly spaced islands along each arc of the chromosome. These islands form positional gene networks, which are governed by the period of 33 kb (Figures 4C, D and 5). The existence of sharp peaks collecting the majority of genes has been observed also in Wright *et al* (2007). Compared with Wright *et al* (2007), we show the existence of a periodic gene organization for a much smaller period (less than a third of the 117 kb detected in Wright *et al* (2007)). Notice that 452 out of the 563 core genes belong to Wright's data set (Wright *et al*, 2007) and that nevertheless, no 117 kb period appears as significant for core genes. On the other hand, when the full arc-based model is applied to the Wright's data set, the resulting maximal period

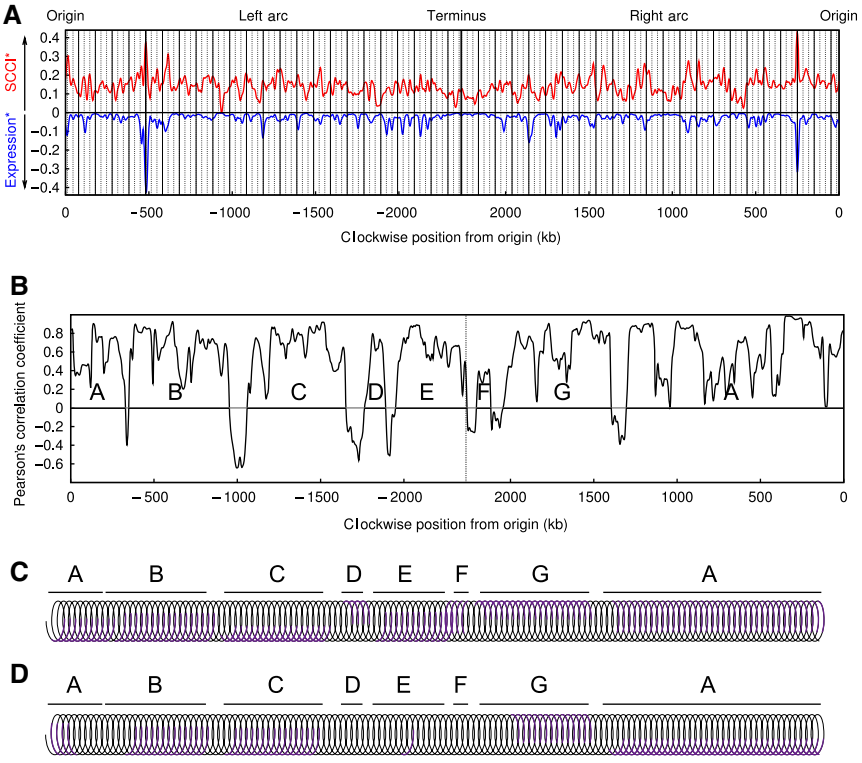

Figure 3 Periods and correlation between SCCI values and expression levels. (**A**) Plots of expression levels for wild type in log-phase growth (bottom, blue curve, values of 3982 genes) and SCCI values (top, red curve, values of 4295 genes) of *E. coli K12* genes, along the chromosome, with a periodically spaced grid in solid lines of period 100 698 nt=3 × 33 566 nt and a grid in dashed points of period 33 566 nt. Many of the highest peaks in the transcription profile appear to fall near the 33-kb grid lines defined by core genes. Smoothed values (denoted SCCI* and Expression*) are used to plot the curves. Distance above the horizontal axis indicates increasing SCCI* and below the horizontal axis indicates increasing expression. (**B**) Local correlation is computed for the two curves in (A); chromosomal sectors are highlighted by thick lines and named with capital letters; sector A overlaps right and left arcs. (**C**) Chromosomal spiral of period 33 566 nt; strips are highlighted in violet along sectors. Most peaks of the SCCI* curve are located within these strips (Table I) and their narrow localization supports the visual effect of peaks matching the 33-kb grid in (A). (**D**) Sectors and strips collecting the strongest peaks of the SCCI* curve, that is peaks with a spectral value $\mu + \sigma$, where μ and σ are the mean and the s.d. of the SCCI* distribution.

is an integer multiple of 33 566 nt (Supplementary Table X). Functional analysis of these groups of genes puts some light on these data and we shall come back to this later.

To explore the functional basis of the distribution, we examined the relationship between SCCI values of *E. coli K12* genes and transcription data for three different growth conditions. These transcriptional data have been analyzed earlier in Wright *et al* (2007). In all three cases, we found that the smoothed SCCI curve, called SCCI*, fits well the high levels of transcription obtained experimentally. This is expected because SCCI values are known to correlate well with CAI values (Carbone *et al*, 2003) and high CAI values are known to correlate well with high expression (Sharp and Li, 1987). We found that the SCCI* curve mirrors the log-phase transcript data along the chromosome (Allen *et al*, 2003) (see Figure 3A; Pearson correlation coefficient R=0.59, s.d. σ_R=0.0017 and $P<2.2e^{-16}$), and that this correlation decreases for transcriptomic data under stress conditions, as heat shock (R=0.47 and σ_R=0.0034) and acid shock (R=0.54 and σ_R=0.0016) (Supplementary Figure 15) (Allen *et al*, 2003). Similar correlations were found in Wright *et al* (2007).

To study the SCCI* curve, we looked at the local maxima of the curve. These peaks do not necessarily correspond to positions for genes with a high SCCI value but rather to positions for genes that display higher SCCI values than their neighbors. To analyze the distribution of peaks, we considered the local correlation of the SCCI* curve with the Expression* curve (i.e. the smoothed curve of transcriptional values), constructed a Pearson local correlation curve between them (Figure 3B) and studied the contiguous regions along the chromosome where the SCCI* values are positively correlated with Expression* values, called *sectors*. One expects a high Pearson correlation coefficient between the two curves to correlate with the existence of a small number of sectors. Hence, the highly biased genome of *E. coli K12* is expected to be organized in few large sectors.

Only seven sectors varying from 87 to 1634 kb cover 89% of the *E. coli K12* chromosome (Figure 3B; Table I; Supplementary Table VII), against an average of 15 sectors that has been estimated on 1000 randomly generated genomes (see Supplementary Figure 17, top). For each sector, in the presence of a uniform distribution of peaks, one would expect that about a third of peaks would fall in windows of 11 189 nt (i.e. a third of the period 33 566 nt), which are periodically spaced of 22 377 nt along sectors. The ensemble of these intervals within a sector is called *chromosomal strip* (see violet strips in Figure 3C and D, where the chromosome is enrolled around a spiral to simplify the representation). A strong bias in peaks periodic

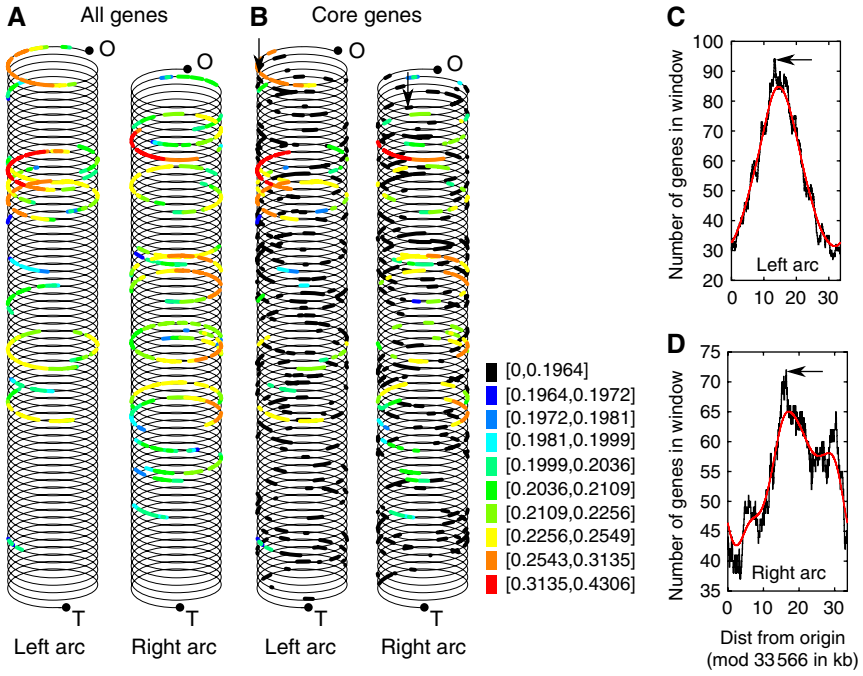

Figure 4 *E. coli K12* genes arranged on left and right arcs of the chromosomal spiral (period 33 566 nt). (**A**) Genes are plotted in colors corresponding to intervals of the smoothed SCCI* curve. Only genes whose value is 1.1 s.d. above the mean (i.e. $\geqslant 0.1964$) of the SCCI* distribution are plotted (they correspond to the highest peaks in Figure 3A). For each gene, we expand the diameter of the dot corresponding to its position from the origin (O). (**B**) All core genes are plotted, with colors corresponding to smoothed values. (**C**, **D**) Distribution of all core genes on the period 33 566 nt (starting from the origin O) and identified on successive windows of 5500 nt along chromosomal arcs; smoothed curves (red) are computed with an s.d. of 2000 nt.

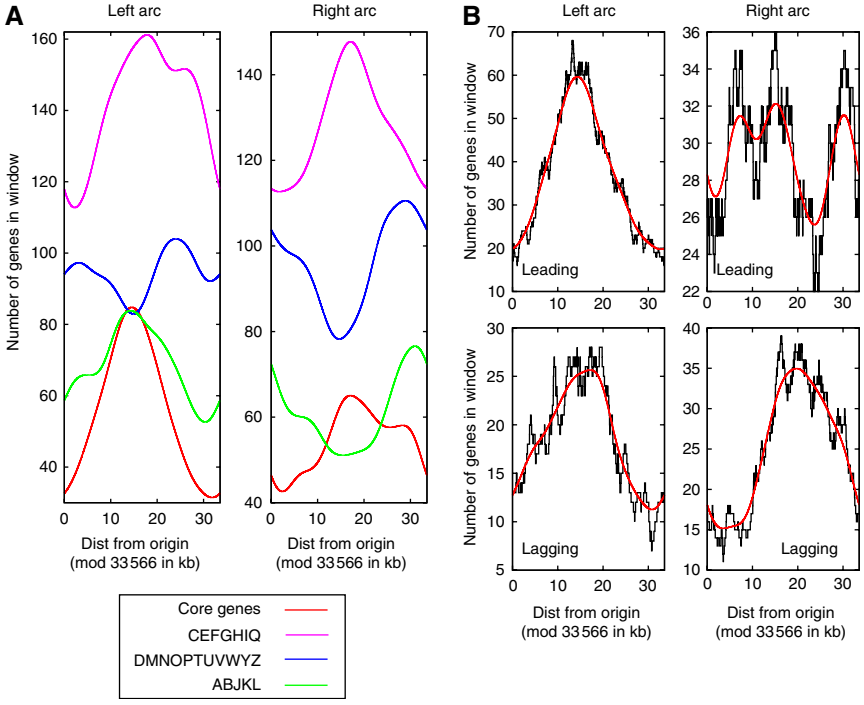

Figure 5 Functional distribution of genes on the chromosomal spiral. (**A**) Distribution of genes belonging to three main functional groups is computed by counting the number of genes located on the chromosomal spiral of period 33 566 nt (starting from the origin O) and on successive windows of 5500 nt along chromosomal arcs. The three functional groups are: information storage and processing (based on COGs classes A, B, J, K, L), cellular processing and signaling (D, M, N, O, P, T, U, V, W, Y, Z), and metabolism (C, E, F, G, H, I, Q). The distribution of core genes (red) is plot for comparison. Curves are smoothed with a Gaussian smoothing window with an s.d. of 2000 nt. (**B**) Curves (smoothed and non-smoothed) tracing chromosomal distribution of core genes along leading (top) and lagging strands (bottom) for left and right arcs. Two sharp periodic co-localizations of genes are observed on lagging strands, both on the right and on the left arc. The same clear signal appears on leading strands on the left arc.

Table I Positions of SCCI* peaks, sector, and strips

	SCCI* peaks								SCCI* peaks $>\mu+\sigma$							
	Left arc							Right arc	Left arc							Right arc
	A	B	C	D	E	F	G	A	A	B	C	D	E	F	G	A
Sector size (kb)	330	605	593	127	429	87	650	1304	330	605	593	127	429	87	650	1304
#Peaks in strip	8	12	9	3	8	3	9	18	0	7	4	0	1	0	4	9
#Total peaks	9	19	17	3	16	3	18	42	2	8	5	0	1	0	6	13

Peaks in the smoothed SCCI curve (Figure 3A, red) are organized in sector along the two chromosomal arcs (Figure 3B). A strip (Figure 3C and D) is defined to cover the largest number of peaks within a sector; we report the number of peaks (left) and of peaks with large spectral value (right) that are captured by strips.

organization appears within *E. coli K12* sectors: all seven sectors present strips where at least 50% of the SCCI* peaks fall into Table I. A *P*-value $<10^{-3}$ is associated to this event (Supplementary Figure 17, bottom). If one considers SCCI* peaks having highest spectral values, that is $>\mu+\sigma$, where μ,σ are mean and s.d. of the SCCI* distribution, then there exist strips that cover at least 60% of the peaks for each sector (Figure 3D; Table I, right). This biased periodic distribution of peaks at 33 kb does not appear for the 117-kb period determined in Wright *et al* (2007) where only two sectors (the shorter ones) capture at least 50% of the peaks (Supplementary Table VIII). The analysis of Expression* peaks provides similar conclusions (Supplementary Table IX).

By considering the Expression* curve for the acid shock stress condition, the same sectors as for log-phase are detected (Supplementary Figure 16, top and middle; Figure 3B) with the exception of the large sector A, located around the ORI, that breaks in three subsectors. Under heat shock stress, only three (sectors B, C, and D) of the seven chromosomal sectors of log-phase are stably preserved, and they are located on a contiguous chromosomal region on the left arc. Sector A breaks into several very small sectors and the region around the TER reorganizes into new sectors. The similar sector organization induced in log-phase and under acid shock stress suggests the existence of a three-dimensional stable chromosomal structure, where positive local correlations remain detected even though gene expression levels might change. The radically different sectors organization of the right arc under heat shock stress suggests that some chromosomal regions may be more affected than others by environmental changes (see 'Discussion' section). Overall, sectors appear robust to experimental noise as shown by several unchanged sector boundaries (B, C, and D) that have been identified under different experimental conditions.

To study the properties of the periodic organization, we represent (without loss of generality) the *E. coli K12* chromosome as being enrolled around a spiral of period 33 566 nt as illustrated in Figure 4A and B. By comparing periodic gene positions along the spiral, we can show a strong tendency of core genes to concentrate on a specific location of the 33-kb interval in both the right and left arcs (Figure 4C and D; Supplementary Figure 13). On the left arc, the signal is stronger and a single peak is uniquely determined. On the right arc, we observe a bimodal distribution of genes and a milder signal due to the much less regular distribution of core genes on the leading strand (Figure 5B, top). To explain the bimodality and to check whether the periodic signal is intrinsically linked to the functional core or it has a more specific functional origin, we analyzed the distribution of functional groups of genes along the chromosome.

We considered those genes in *E. coli K12* that have been organized in COGs functional classes (Tatusov *et al*, 2003) and applied our approach to verify whether the distribution of distances for genes in COGs classes provides a significant period, possibly coinciding with the one obtained for core genes. The 3533 genes (542 of which are core) in COGs classes display the first peak at 369 231 ($=11\times33\,566$ nt, P=0.3514, Z=9.25) and the second one at 33 566 nt (Z=3.82) for the full arc-based model; the peak at 33 566 nt appears to be the only statistically significant one for the circular model (P=0.007, Z=10.52). When the analysis is done on the three main COGs groups separately (see Supplementary Table XI), the signal at 33 566 nt is detected for genes involved in metabolism (1365 genes, 253 of those are core) with a main peak for the circular model (P=0.0041, Z=5.63). The circular model recovers the 33-kb period for the information storage and processing genes (681 genes, 140 of those are core) and for the cellular processing and signaling genes (965 genes, 116 of those are core) but not with a main peak (see Supplementary Table X).

Genes sharing the same function are preferably localized on specific facets (i.e. periodic intervals of $\approx$10 kb kb) of the spiral at 33 kb in Figure 5A. There are two such facets along each chromosome arc, one organizing genes involved in cellular processing and signaling (blue curve in Figure 5A) and the other organizing genes involved in metabolism (violet). Genes involved in information storage and processing (green) seem to undergo the same evolutionary positional pressure as the cellular processing and signaling gene positional network in the right arc (curves with Pearson correlation coefficient R=0.89 and $P<2.2e^{-16}$), and the metabolism gene positional network in the left arc (R=0.58 and $P<2.2e^{-16}$). Core genes (red) share the same chromosomal facet of genes involved in metabolism (R=0.65 on the left arc and R=0.89 on the right arc with $P<2.2e^{-16}$), and in information storage and processing on the left arc (R=0.97 and $P<2.2e^{-16}$) (see Supplementary Table XII). The bimodality present in their distribution on the right arc is explainable by the overlapping of the two independent networks with which the red curve shares the maximal peaks.

Our analysis applied to all *E. coli K12* genes (Figure 6C; Supplementary Table X) provides periods that are multiple of 33 kb for arc-based models and for the circular model with c=3. All genes tend to localize in two specific periodic subintervals of 33 kb. The bimodality is much more

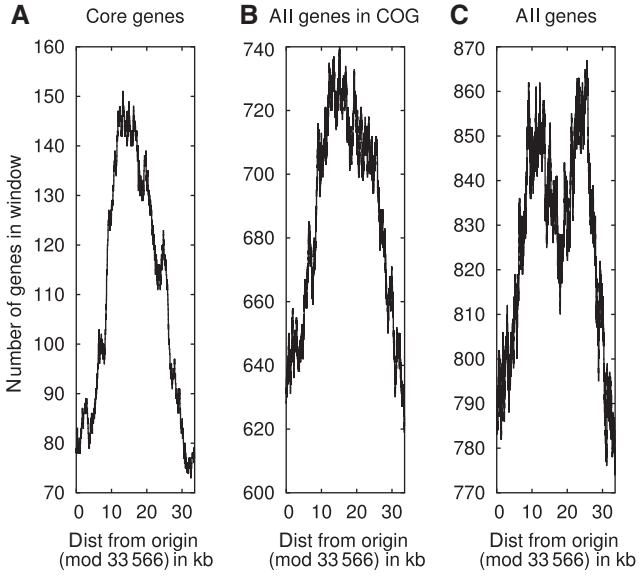

Figure 6 Chromosomal gene mapping: curves (non-smoothed) tracing chromosomal distribution of core genes (**A**), all genes in COGs classes (**B**), and all *E. coli K12* genes (**C**) along the entire chromosome with a period of 33 566 nt. The periodic co-localization of most core genes around a peak between 10 and 20 kb from the origin is clearly defined. Two peaks corresponding to the two identified networks are announced for COGs-classified genes, and clearly observed for all genes.

pronounced for all genes though (Figure 6C). The weaker signal identified by looking at all genes, justifies our strategy to analyze core genes first, where the strength is much stronger.

To test further our results, we applied our method to two sets of essential genes identified experimentally in Gerdes *et al* (2003) (Gerdes' data set, 602 genes) and in Baba *et al* (2006) (Baba's data set, 300 genes) (see Supplementary Table XI). The two data sets share 201 genes (this showing the definition of experimental essentiality to be ambiguous). The signal at (integer multiples of) 33 kb is detected in both data sets through several statistically significant peaks. Baba's data set displays a main periodic signal at 117 kb for both the full arc-based model ($P<10^{-4}$, Z=5.47) and the circular model ($P<10^{-4}$, Z=6.66). To search for the genes in Baba's data set that contribute most to the 117-kb period, we considered the smoothed gene distribution curve defined on the 117-kb interval (see Supplementary Figure 18). Then, we focused on the subinterval defined by the two minima preceding and following the maximum of the curve. We determined 172 Baba's genes lying within the subinterval and we characterized the COGs functional classes whose number of genes within the subinterval is overrepresented (i.e. for a COGs class, we asked that >50% of its genes occurring in Baba's data set should lie in the subinterval). All three main COGs functional groups are represented (classes ACHIJLMO satisfy the above condition), this result showing a non-obvious functional interpretation of the 117-kb period.

By filtering out from the COGs group of genes involved in information storage and processing (ABJKL) those genes that are not biased (i.e. genes with an SCCI value smaller than the mean value of the SCCI distribution of all *E. coli K12* genes), we obtain a main peak at 117 kb for the full arc-based model and

Table II Classes of periodically distributed genes with main period at 33 kb (or a multiple of it)

Data sets	Full arc based		Circular	
	P	*Z*	*P*	*Z*
All genes (4295)	0.2092	10.03	0.9459	3.68
Core genes (563)	$<10^{-4}$	7.81	$<10^{-4}$	8.19
All genes in COGs (3533)	0.3514	9.25	0.007	10.52
Wright's genes (2247)	0.0230	8.21	Not principal	
Core genes in COGs (542)	0.0002	7.9	$<10^{-4}$	8.78
Core genes in Wright's set (452)	0.0001	5.94	$<10^{-4}$	7.63

P-values and *Z*-scores are reported for each main peak obtained on full arc-based and circular models. Wright's set display a 33-kb signal as third best period for the circular model. The data set of all genes displays its fifth best peak at 33 kb, and the third and forth best periods at integer divisor of 33 kb for the circular model. Full information is found in Supplementary Table X.

the 117 kb signal as a second best period for the circular model (Supplementary Table X). Notice that 72% of ABJKL genes in Baba's data set are involved in translation, ribosomal structure, and biogenesis (COGs class J) and most of them (85%) are core genes, therefore highly biased genes. Also, a large portion (36%) of Baba's genes (against 25% of Gerdes' genes) are classified in ABJKL, and this corresponds to the 16% of the genes in ABJKL. No other COGs functional group highlights a 117-kb period. This suggests that Baba's set selects those genes, especially involved in translation (only the 13% of ABJKL genes in Baba's data set lie in the COGs class K involved in transcription), which are periodically spaced at 117 kb. We need to add here that when COGs analysis is restricted to Wright's data set (see Supplementary Table XI), then the 117-kb period characterized in Wright *et al* (2007) is identified as the second most important peak of COGs classes involved in information storage and processing (ABJKL) for the full arc-based model and the third best peak for the circular model. No other COGs functional group highlights the 117-kb period on Wright's data set.

No signal enhancement was found when the analysis of core genes considers their localization into operons (Supplementary Tables IV and V).

Discussion

Bacterial transcription and translation are known to be intimately connected, and we exploit genomic translational signals in *E. coli K12* to detect a statistically significant periodic distribution of core genes, that is highly biased genes, along its chromosome. We show that the periodic signal in *E. coli K12* genome does not come exclusively from core genes but that it is definitely enhanced by them. Core genes carry a strong periodic signal at 33 kb that we also find in functionally organized groups and, at a minor extent, in the whole set of *E. coli K12* genes (see Table II for a summary). We show a positional distribution of genes belonging to functional classes (these genes do not necessarily belong to the functional core of the organism). They organize in two different positional networks, which occupy distinguished chromosomal 'facets' over the same period (Figure 5A; Supplementary Figure 8). It might appear striking that the set of genes involved in *E. coli*

K12 metabolism is periodically organized in the same way as the set of core genes (compare violet and red curves in Figure 5A), but in fact, core genes have been used already in Carbone and Madden (2005) to identify key metabolic networks for a number of bacteria, including *E. coli K12*, and to highlight that genes involved in metabolism are subject to a strong evolutionary pressure on their codon bias. Our current finding adds a new insight on our understanding of the evolutionary pressure undergoing metabolic pathways, reporting that these genes need also to preserve a regular positional distribution along the chromosome. The link between essential metabolic pathways and chromosomal superhelicity had been observed experimentally. Mutations in metabolic genes was observed to affect DNA topology (Hardy and Cozzarelli, 2005), and experimental analysis of the citric acid cycle, known to be essential for *E. coli K12* vitality, showed DNA relaxation to be coupled to crucial metabolic steps of the cycle (Blot *et al*, 2006).

The two main chromosomal networks highlighted for the three main COGs functional groups (Figure 5A) would favor specific chromosomal facets to transcription (Képès and Valliant, 2003; Képès, 2004) and induce in this manner a regulation of gene activity by DNA superhelicity. By running FFT analysis on the three COGs groups separately in the full arc-based model (data are shown in Supplementary Table X), we obtained quite large main periods (685 714 nt for metabolism, 145 455 nt for cellular processing and signaling, and 114 286 nt for information storage and processing) and all of them are multiples of 28 571 nt. The plot of the gene distribution for the three functional groups of COGs classes for 28 kb (Supplementary Figure 19) shows the existence of 'complementary' facets favoring metabolism on one side, cellular processing and signaling and information storage and processing on the other side. This complementary distribution is sharply present on both arcs and highlights once more the existence of two independent gene networks separating metabolic genes from other functional classes (see the complementarity of the violet and the blue curves in Supplementary Figure 19 as already observed in Figure 5A). The positional gene organization confirmed by the 28-kb period suggests that there might be an assortment of small supercoiled loops, with variable sizes, that preserve positional networks. This goes along the lines already pointed out in Postow *et al* (2004). Notice that the 33 and 28 kb periods are compatible with the critical loop size range described in Postow *et al* (2004). The fluidity of loop sizes proposed there is not supported by the functional role of gene positional networks though, as only certain periods define positional gene networks and consequently, might be of interest for cell functioning.

Our finding can be considered as the *in silico* counterpart to the experimental observations that uncovered the function of DNA superhelicity in cellular activity. Alterations of global DNA superhelicity have been shown to be linked to modulation of genomic transcription (Azam and Ishihama, 1999; Azam *et al*, 1999; Jeong *et al*, 2004; Peter *et al*, 2004; Willenbrock and Ussery, 2004; Travers and Muskhelishvili, 2005a) and be associated to both growth transition and stress response to environmental challenge (Balke and Gralla, 1987; Dorman, 1996; Tse-Dinh *et al*, 1997; Cabrera and Ding, 2003; Cheung *et al*, 2003; Travers and Muskhelishvili, 2005b). Only the 8% of specific genes are found to respond to supercoiling in *E. coli K12* though (Peter *et al*, 2004), and the function of superhelicity in organizing transcription remains obscure to experimental approaches. Global transcription of the bacterial genome during cellular growth has been shown to be coordinated with a homeostatic regulation of supercoiling (Blot *et al*, 2006).

Besides the experimental evidence of chromosomal superhelical formation, a chromosome structuring into four macrodomains was observed with fluorescent microscopy in Espéli and Boccard (2006), and the impact on genome plasticity due to structural constraints was experimentally addressed in Esnault *et al* (2007). Even though we analyze the global periodic signal all along the chromosome, and we use the chromosomal spiral for representing the 33-kb periodic gene arrangement (Figure 3C and Figure 4A and B), we looked whether there were periodic patterns at 33 kb, which were localized on specific regions along the chromosome. We found that periodic patterns at 33 kb are present along most parts of the *E. coli K12* chromosome, and that the large majority of maximal peaks in the SCCI* curve (Figure 3A) follows the 33-kb pattern. Strips identified for different sectors might be out of phase across sectors (see Figure 3C), and this suggests that the chromosomal superhelical structure might likely not be a regular spiral. Sectors accommodate the idea of spiral breaks along the chromosomal structure, suggesting a structure that is flexible to local arrangements. The analysis of three Expression* curves defined out of transcriptomic data recorded for three different living conditions (Figure 3A; Supplementary Figure 15), points out the existence of a stable chromosomal structure where variations in local correlation appear to be much more robust than one might imagine when thinking of drastic changes in living conditions. Chromosomal stability is reflected by positive local correlations that remain detected even though gene expression levels might change for different living conditions. Some chromosomal regions are more affected than others by environmental changes and our method allows for a detection of these regions (see Supplementary Figure 16). An observation that confirms further our finding is illustrated in Supplementary Figure 14, where we show that three of our sector boundaries are localized on the same chromosomal regions of three (out of the six) macrodomain boundaries in Esnault *et al* (2007). One of these macrodomains (the Right domain of Esnault *et al* (2007)) is formed by the two sectors C and D, which we have detected to be resistant to changements of *E. coli* living conditions (see Supplementary Figure 16).

Sectors might also be a signal of an extra level of organization for DNA loops (Postow *et al*, 2004) similarly to the domain structure in Esnault *et al* (2007). Overall, mechanisms and effects of chromosomal structuring are far from being understood and any highlights on specific periodic distribution of genes along the chromosome might be of interest for further experimental design uncovering missing biological information.

A chromosomal architecture organized in a spiral, or in a 'rosette-like' arrangement of domains (Cook, 2002), with several 33-kb long loops, would provide a way to far apart genes to be localized close by in the chromosome, as a few

turns of the spiral can bring closer together distant parts of the genome. An organization into a spiral was proposed in Wright *et al* (2007), based on the much larger period of 117 kb. We tested whether a rosette-like architecture could be suggested by our gene networks but no meaningful arrangement of 33 kb long loops was found.

Our single genome analysis to detect a periodic gene distribution in *E. coli K12* can be compared with the genomic comparison approach of Wright *et al* (2007). The latter uses proximity conservation of pairs of homologous genes across a set of genomes and defines a distribution of distances for these pairs in *E. coli K12* genes that turns out to follow a periodic organization at 117 kb. In contrast, our approach considers the set of 563 *E. coli K12* predicted core genes (452 of these genes belong to the set of 2255 conserved genes considered in Wright *et al* (2007)) and all possible distances between pairs of these genes that are located in the right arc and in the left arc of the chromosome. Our period is much smaller and helps to find two positional gene networks that could not be detected in Wright *et al* (2007) (Supplementary Figure 9). When the two methods are compared with the same transcriptional data, our method can explain peaks in the transcriptional curve at a finer scale (through strips and sectors). Finally, our method can be applied easily at large scale to any bacterial organism and it does not demand a thorough genomic comparison of the bacterium of interest with several others as in Wright *et al* (2007). Namely, genes shared by several species are few, and larger the number of compared species fewer is the number of common genes (Mushegian and Koonin, 1996; Brown *et al*, 2001; Harris *et al*, 2003; Koonin, 2003; Charlebois and Doolittle, 2004). In addition, genes might not be physically close in a genome but be close in another and the integration of a large number of genomes to ensure meaningful comparisons is crucial when the method in Wright *et al* (2007) is applied at large scale. This means that any database of precomputed statistically significant pairs of genes could be used only through an appropriate pair filtering, and that a genome addition should reconsider all genomes previously analyzed for the updating of new gene pairs. This approach, once applied at large scale, requires an important computational cost.

Observe that the periodic signal at 33 kb has been found for a large set of genes that does not *a priori* satisfy any particular evolutionary constraint on coupling pairs of genes (in contrast to Wright *et al* (2007)) and that the combinatorial method we used amplifies noise quadratically. In light of these very loose conditions satisfied by the sets of genes under study and by the method, the strength of the periodic signal that we detect is even more reinforced. In particular, the results support the idea that strong evolutionary functional constraints are imposed to gene organization along the chromosome.

In conclusion, an overall chromosomal organization exists and it appears to help the expression of genes that belong to a functional core or that are involved in metabolism, and especially within the lagging strands (Figure 5B, bottom). Genes located in the lagging strands have a known transcriptional disadvantage due to the bacterial-replication mechanism that might indeed involve a stronger evolutionary pressure in lagging strands (Omont and Képès, 2004; Mirkin and Mirkin, 2005). A very clear periodic signal is identified on the left arc of the chromosome compared with a weaker signal appearing on the right arc (see Figure 5B; Supplementary Figure 12). This latter is known to undergo more important gene conversions (French, 1992). Inversions of strands might result, at times, in extremely deleterious rearrangements (Louarn *et al*, 1985; Segall *et al*, 1988), and a disruption by inversion of the periodic gene distribution affects the whole chromosome and not just the inverted part, since the period, which is a global parameter, might be affected as well as the positional gene networks. New models for explaining transcription versus translation, the interplay of transcription-replication, and gene inversion might be found profitable to account for the effect of the periodic chromosomal arrangement of genes.

To know which three-dimensional model (structure) of the chromosome, if any, is closer to the truth, one needs to conceive appropriate experiments. Through the theoretical analysis, we can find signals that can lead to a precise conception of experiments and to hypothesis to be tested. Exchanging the position of metabolic genes or genes involved in translation in the chromosome (by keeping them in the same chromosomal region but shifted by an interval Képès and Valliant, 2003; Jeong *et al*, 2004; Képès, 2004). These results will hopefully lead to conceiving new models or bring new functional insights on old ones.

Materials and methods

Genomes and annotation

E. coli K12 MG1655 genome flat file was retrieved from GenBank. It contains 4295 annotated genes. This genome is referred as *E. coli K12*. *E. coli K12* COGs (Tatusov *et al*, 2003) classification is found at http://www.ncbi.nlm.nih.gov/sutils/coxik.cgi?gi=115. The COGs class 'Inorganic ion transport and metabolism,' named *P*, has been found to follow the same pattern of chromosomal gene distribution observed for all COGs classes grouped as 'cellular processing and signaling' instead of the one followed by COGs classes grouped as 'metabolism.' Several characteristics of the *E. coli K12* genome and lifestyle are collected in Supplementary Table II.

Calculation of highly biased genes

The method used to automatically detect highly biased genes in a genome is explained in Carbone *et al* (2003). It is based on a generalized notion of Codon Adaptation Index (Sharp and Li, 1987), called SCCI, which ranks genes in a genome depending on how much they are affected by the dominant compositional bias governing the genome. For fast growing organisms like *E. coli K12*, genes with high SCCI rank correspond to genes carrying a strong signal of translational bias in their codon usage, which have to be expressed rapidly in specific moment of the life of the organism but not necessarily always.

Calculation of core genes

To detect the set of core genes, that is genes affected by a strong bias in codon composition, we use the computational approach described in Carbone (2006). This method is not based on comparative genomics but on codon bias analysis of the genome. Many of the highly biased genes detected by the method are experimentally proven to be essential, and those which are found to be significant by the method but not experimentally are usually stress response genes, which could never be detected experimentally due to the extremely good living conditions under which experiments are run. The numerical criterium for prediction says that a core gene g has $\mathrm{SCCI}(g) > \mu + \sigma$, where μ and σ are mean and s.d. of the distribution of SCCI values over all the genes in the genome (for the *E. coli K12* genome, μ=0.31 and σ=0.1028).

This definition ensures that genes in the tail of the distribution largely deviate from the average behavior of the genome, and that a significant number of genes belong to the tail. The set of predicted core genes is called *functional genomic core*. For *E. coli K12*, there are 563 predicted core genes ($SCCI(g) > 0.41$), 486 of which are functionally classified in COGs, 28 of which have unknown function, and 22 have an hypothetical function; 223 core genes lie on the lagging strands and 340 on the leading strands.

Essential genes

Two data sets of essential genes have been experimentally identified by Gerdes *et al* (2003) and Baba *et al* (2006). Gerdes' essential genes were retrieved from the Supplementary Table S1 of Gerdes *et al* (2003), at http://www.integratedgenomics.com/online_material/gerdes/ table_s1.xls. Among the 620 essential genes in the original set, we consider the 602 that are annotated in the *E. coli K12* GenBank file. Baba's essential genes were retrieved from the Supplementary Table VI of Baba *et al* (2006), at http://www.nature.com/msb/journal/v2/n1/extref/msb4100050-s8.xls. Among the 303 essential genes defined in the original set, we consider the 300 that are annotated in the *E. coli K12* GenBank file.

Genes on the chromosome arcs, ORI, and TER

Given a set of genes X, we separate them into different classes depending on their location in the chromosome. Leading and lagging chromosomal strands, read from ORI to TER and from TER to ORI, give rise to six distinguished subsets of genes in X depending on their location (Supplementary Figure 1). Two sets, $leading_1$ and $leading_2$, are constituted by all genes in X lying on the two leading strands, respectively, located on the right and left arcs. Similarly, $lagging_1$ and $lagging_2$ are constituted by all genes in X lying on the two lagging strands. Two sets, *left* and *right*, are constituted by all genes in X located on the left and on the right arc, respectively. Accordingly to the analysis that is performed, the set X is the set of all core genes, experimentally identified essential genes (Gerdes *et al*, 2003; Baba *et al*, 2006), evolutionarily conserved gene pairs (Wright *et al*, 2007), genes classified in specific COGs classes eCOG, or all *E. coli K12* genes. We speak about core genes 'around the ORI (TER)' and mean those core genes located on the half chromosome around the origin (termination site). To determine the half chromosome around the ORI (TER), we consider the left and right arcs separately, cut them into two equal parts, and consider the two quarters neighboring the ORI (TER).

Smoothing of curves

Curves are smoothed at times. This is done, all along the paper by using a Gaussian smoothing window of s.d. σ. For Figure 3A and Supplementary Figure 15, σ=6 kb. For Figures 4C and D, and 5 and Supplementary Figures 12, 18, and 19, σ=2 kb. Values of a smoothed curve are indicated with * (when necessary, to avoid confusion).

Pearson correlation coefficient of pairs of curves

The Pearson correlation coefficient and corresponding P-values between pairs of curves in Figure 5A are computed using the R function cor.test in the R-package (R Development Core Team, 2008) (http://www.r-project.org/).

Plots of genes and the periodic (smoothed) spiral

We generated Figure 3A by plotting SCCI* and Expression* values for all genes along the arcs, where a gene location is determined by the gene middle point position on the arc. Distributions of genes in Figures 4C and D, 5, and 6 and Supplementary Figures 8–12, 18, and 19 have been realized by sliding consecutive windows of 5500 nt along the chromosome arcs and by counting for each window the number of overlapping genes (one nucleotide is sufficient for a gene to overlap a window). The chromosome arc is represented on a spiral and each distribution of genes is projected on the interval $[0, \ldots, Y]$ (where Y=33 565 when the period is 33 566 for instance). For each integer $i \in [0, \ldots, Y]$, a distribution records the sum of the number of genes overlapping all the windows that are centered at $I + k\ Y + 1$, for $k \geqslant 0$, and covering the arc. (To center a window we use its middle point.) (The ORI corresponds to i=0 and k=0.) Figure 4A and B plot the full length of a gene that is all nucleotide positions of a selected gene are colored in the spirals. Similarly, all nucleotide positions in a periodic interval are plotted in Figure 3C and D.

Calculation of chromosomal periodic distributions

Let X be a set of genes as above. The distance between the two genes in a set X is computed as the distance between starting points of those genes. For the circular model, we consider minimal distances between all pairs of genes in X along the chromosome, and call S_{circ} the set of all these distances. For arc-based models, we compute distances between all pairs of genes in X lying in the appropriate strand or arc; we call the associated sets of distances S_{left}, S_{right}, $S_{leading1}$, $S_{leading2}$, $S_{lagging1}$, $S_{lagging2}$. The set of distances between the pairs of genes in X located in the same chromosomal arc is $S_{all}=S_{left} \cup S_{right}$. (Notice that distances between genes located in different arcs are not considered.) The set of distances between pairs of genes in X located in the leading chromosomal strands is $S_{leading}=S_{leading1} \cup S_{leading2}$. Similarly, we define $S_{lagging}=S_{lagging1} \cup S_{lagging2}$ (Figure 1).

Chromosomal periodicity is computed as follows. We construct a distribution diagram on a set of distances (where histogram bins record distances measured on steps of 5000 nucleotides) and symmetrize it by centering the first histogram bin (see below on the effects of symmetrization). Then, we compute the periodogram of the histogram using the FFT of the corresponding series (in practice, we used the function *spec.pgram* of the *R* language). To help the FFT detection of the signal, if necessary, the series is padded with zeros until its length is a highly composite number, data are detrended (i.e. a linear trend is removed from the series; compare with Supplementary Figures 4 and 6 for an analysis without detrend) and the mean is removed. A proportion of 0.5 of the data has being tapered. Chromosomal periodicity has been computed for several sets X of genes (Supplementary Tables III–VI and X) on the four models of Figure 1. Sets of distances S_{circ}, S_{all}, $S_{leading}$, and $S_{lagging}$ were defined accordingly to the set X used in the application.

For each periodogram, we identify a period, possibly several periods, whose spectral values are sufficiently distant from the mean of the distribution. Namely, given a periodogram, we consider all spectra that are local maxima, compute the distribution of these spectra, and extract those periods that are associated to spectra lying on the queue of the distribution, that is at least c s.d. away from the mean $\mu_{spectra}$. In the text, c=3 if not specified otherwise. Given a maximal spectrum M selected as above, we associate to its corresponding period a Z-score defined as $M - \mu_{spectra}/\sigma_{spectra}$. The statistical significance of a period is determined with respect to P-values computed from the null model described below, whereas the notion of Z-score is interesting for comparing different periodograms as in Supplementary Figure 7.

Generation of random genomes

To generate a random genome close enough to the *E. coli K12* genome, we tag core genes in the *E. coli K12* genome and define two sets: one contains *all* gene lengths with a tag on those lengths originally associated to core genes, and the other contains the lengths of all intergenic regions. To construct a random genome, we shuffle the two sets independently, and put them together by randomly inserting a gene between two intergenic regions along a linear arrangement. The starting and ending points of this random arrangement coincide with the ORI of the randomized circular chromosome and the TER is fixed at the same ORI-TER distance, which is characteristic of the *E. coli K12* genome. Gene (leading or lagging) strand is determined by respecting the original strand in the *E. coli K12* genome. By construction, the distribution of intergenic region lengths, the distribution of gene lengths, and the number of core genes in the randomized genomes are the same as for the *E. coli K12* genome. Notice, however, that the

number of genes in each arc is not preserved. The methodology we describe can be applied to any set X of selected genes, not only to core genes. P-values computed for sets in Supplementary Table X are evaluated through randomly generated genomes as described.

Comparison of periodogram peaks against the null model

Given a random genome, we calculate its periodograms on S_{all}, $S_{leading}$, $S_{lagging}$, and S_{circ} as done for the original genome (see Supplementary Figure 10 for instance). We count how many maximal spectral values are higher than the maximal spectral value of the corresponding periodogram on the original genome. We randomly generate 10 000 genomes and perform 10 000 times the analysis. A P-value is defined by counting the number of randomized genomes displaying a spectral value for the maximal peak, which is as strong as or stronger than the spectral value of the maximal peak determined for the *E. coli K12* genome, and by dividing this number by 10 000 (Supplementary Tables III–VI and X). A P-value $< 10^{-4}$ says that 10 000 randomizations did not generate peaks, which were stronger than the *E. coli K12* genome strongest peak. In Supplementary Table VI, we consider peaks with a spectral value $> \mu_r + 3\sigma_r$, where $> \mu_r, \sigma_r$ are mean and s.d. of the distribution of peaks for the randomized genome, this condition guaranting the peak to be exceptional.

Symmetrization of distribution diagrams

FFT is realized on symmetrized distance distributions. Symmetrization of right triangle shapes has two main effects: (1) the suppression of noise introduced by the artificial right triangular shape into Fourier analysis, as illustrated in Supplementary Figure 6A and C compared with Figure 2B and Supplementary Figure 6D, respectively; (2) the amplification of the signal as shown by the variation of spectral values in Supplementary Figure 5 when compared with Figure 2B.

Prediction of the replication origin

Prediction of origin and terminus has been done with the program Oriloc (Frank and Lobry, 2000) accessible at http://pbil.univ-lyon1.fr/software/SeqinR/SEQINR_CRAN/DOC/html/oriloc.html.

Comparison with Wright's data

We consider 2247 of the 2255 genes in the Supporting Information of Wright *et al* (2007). Missing genes are as follows: three gene names do not appear in the GenBank file, one is a tRNA, four are synonymous genes and we count them only once.

Transcriptional data

As in Wright *et al* (2007), we computed an average of the absolute transcript level for wild-type standard growth conditions (4-morpholinepropanesulfonic acid minimal glucose, doubling time 2–8 h) using 5 Affimetrix (Santa Clara, CA) microarray data sets. The same has been done for two other data sets: for wild-type growth under heat shock (at 50 degrees, glucose as carbon source) and for wild-type growth under acid shock (pH adjusted to 2, glucose as carbon source). Microarray data sets are extracted from the ASAP database at https://asap.ahabs.wisc.edu/asap/experiment_data.php and refer to (Allen *et al*, 2003). The Expression* curve in Figure 3A is computed based on 3982 genes, that is a subset of the set defining the SCCI* curve.

Correlation between expression levels and SCCI curve

Transcriptional data were renormalized by dividing all values by the maximum (where max=23 840.88). Then, they were smoothed by using a Gaussian smoothing window (σ=6 kb). The same smoothing (following normalization by a max value of 0.83) has been applied to the SCCI curve. The curves (i.e. SCCI* and Expression*) were translated to identify the minimal y value to 0. The Pearson correlation coefficient R between the smoothed transcriptional data and the smoothed SCCI values has been computed by sampling the *E. coli K12* genome once every 12 kb to avoid smoothing artifacts, and by averaging the Pearson correlation coefficients obtained over all choices of the sampling phase. The resulting correlation value is assigned to the entire *E. coli K12* genome. P was computed by using Student's t test as implemented in the function t test in the R-package.

Pearson correlation curves in Figure 3B and Supplementary Figures 14 and 16 were computed using the same idea. Given the smoothed SCCI curve and the smoothed expression level curve, the Pearson correlation curve was computed by sliding windows (from ORI to ORI, by reading the left arc first) of 100 kb along the *E. coli K12* genome with a step of 1 kb. A Pearson correlation coefficient was computed for each 100 kb long window (by sampling once every 1 kb and by averaging Pearson correlation coefficients on all sampled values), and the Pearson correlation curve was defined to be the curve joining all points determined by the starting position of a 100 kb window (x axis) and the Pearson correlation coefficient of that window (y axis).

Chromosomal sectors

Chromosomal sectors are maximal intervals $[X,Y]$ where a Pearson correlation curve $P(x)$ takes positive values. Formally, $[X,Y]$ should fulfill the following two properties: (1) $P(X-1000) < 0$ and $P(Y+1000) < 0$; (2) $P(x) \geqslant 0$ for all $x \in [X,Y]$. Condition 1 follows from the fact that points defining $P(x)$ are spaced of 1 kb by construction.

Chromosomal strips

Given a chromosomal sector defined by the interval $[X,Y]$, let S be the set of x coordinates of SCCI* peaks falling in the interval. For each integer $i \in [X, Y + 33\,565]$, we slide a 11 189 kb (i.e. a third of 33 566 kb) long window centered at i and count the number of peaks in S that overlap the window or its periodic instances centered at $I + k$ 33 566, for $k \geqslant 1$, over $[X,Y]$. The ensemble of the periodically spaced sliding windows that maximizes the number of peaks in S defines the *strip* of the chromosomal sector (see violet strips in Figure 3C and D).

Random genomes for the evaluation of sectors and strips

In all, 1000 random genomes have been constructed as described above, based on the set of all genes and intergenic regions in *E. coli K12*. For each generated genome, 3982 genes (corresponding to those used to compute the curve Expression* for *E. coli K12*) have been identified to compute a curve Expression* for each randomly generated genome.

Supplementary information

Supplementary information is available at the *Molecular Systems Biology* website (www.nature.com/msb).

Acknowledgements

We thank Yann Ponty for providing the software to draw spirals. This work was performed with the financial support from MESR (for AM doctoral fellowship) and from INSERM.

Conflict of interest

The authors declare that they have no conflict of interest.

References

Allen TE, Herrgård MJ, Liu M, Qiu Y, Glasner JD, Blattner FR, Palsson BØ (2003) Genome-scale analysis of the uses of the *Escherichia coli* genome: model-driven analysis of heterogeneous data sets. *J Bacteriol* **185:** 6392–6399

Azam TA, Ishihama A (1999) Twelve species of the nucleoid-associated protein from Escherichia coli. Sequence recognition specificity and DNA binding affinity. *J Biol Chem* **274:** 33105–33113

Azam TA, Iwata A, Nishimura A, Ueda S, Ishihama A (1999) Growth phase-dependent variation in protein composition of the *Escherichia coli* nucleoid. *J Bacteriol* **181:** 6361–6370

Baba T, Ara T, Hasegawa M, Takai Y, Okumura Y, Baba M, Datsenko KA, Tomita M, Wanner BL, Mori H (2006) Construction of *Escherichia coli K-12* in-frame, single-gene knockout mutants: the Keio collection. *Mol Syst Biol* **2:** 2006.0008

Balke VL, Gralla JD (1987) Changes in the linking number of DNA accompany growth transitions in *Escherichia coli*. *J Bacteriol* **169:** 4499–4506

Blot N, Mavathur R, Geertz M, Travers A, Muskhelishvili G (2006) Homeostatic regulation of supercoiling sensitivity coordinates transcription of the bacterial genome. *EMBO Rep* **7:** 710–715

Brown JR, Douady CJ, Italia MJ, Marshall WE, Stanhope MJ (2001) Universal trees based on large combined protein sequence data sets. *Nat Genet* **28:** 281–285

Carbone A (2006) Computational prediction of genomic functional cores specific to different microbes. *J Mol Evol* **63:** 733–746

Cabrera JE, Ding JJ (2003) The distribution of RNA polymerase in *Escherichia coli* is dynamic and sensitive to environmental cues. *Mol Microbiol* **50:** 1493–1505

Carbone A, Madden R (2005) Insights on the evolution of metabolic networks of unicellular translationally biased organisms from transcriptomic data and sequence analysis. *J Mol Evol* **61:** 456–469

Carbone A, Zinovyev A, Képès F (2003) Codon Adaptation Index as a measure of dominating codon bias. *Bioinformatics* **19:** 2005–2015

Charlebois RL, Doolittle WF (2004) Computing prokaryotic gene ubiquity: rescuing the core from extinction. *Genome Res* **14:** 2469–2477

Cheung KJ, Badarinarayana V, Selinger DW, Janse D, Church GM (2003) A microarray-based antibiotic screen identifies a regulatory role for supercoiling in the osmotic stress response of *Escherichia coli*. *Genome Res* **13:** 206–215

Cook PR (2002) Predicting three-dimensional genome structure from transcriptional activity. *Nat Genet* **32:** 347–352

Dorman CJ (1996) Flexible response: DNA supercoiling, transcription and bacterial adaptation to environmental stress. *Trends Microbiol* **4:** 214–216

Esnault E, Valens M, Espéli O, Boccard F (2007) Chromosome structuring limits genome plasticity in *Escherichia coli*. *PLoS Genet* **3:** e226

Espéli O, Boccard F (2006) Organization of the *Escherichia coli* chromosome into macrodomains and its possible functional implications. *J Struct Biol* **156:** 304–310

Frank AC, Lobry JR (2000) Oriloc: prediction of replication boundaries in unannotated bacterial chromosomes. *Bioinformatics* **16:** 566–567

French S (1992) Consequences of replication fork movement through transcription units *in vivo*. *Science* **258:** 1362–1365

Gerdes SY, Scholle MD, Campbell JW, Balázsi G, Ravasz E, Daugherty MD, Somera AL, Kyrpides NC, Anderson I, Gelfand MS, Bhattacharya A, Kapatral V, D'Souza M, Baev MV, Grechkin Y, Mseeh F, Fonstein MY, Overbeek R, Barabási AL, Oltvai ZN *et al* (2003) Experimental determination and system level analysis of essential genes in Escherichia coli MG1655. *J Bacteriol* **185:** 5673–5684

Grantham R, Gautier C, Gouy M, Mercier R, Pave A (1980) Codon catalog usage and the genome hypothesis. *Nucleic Acid Res* **8:** r49–r62

Hardy CD, Cozzarelli NR (2005) A genetic selection for supercoiling mutants of *Escherichia coli* reveals proteins implicated in chromosome structure. *Mol Microbiol* **57:** 1636–1652

Harris JK, Kelley ST, Spiegelman GB, Pace NR (2003) The genetic core of the universal ancestor. *Genome Res* **13:** 407–412

Jeong KS, Ann J, Khodursky AB (2004) Spatial patterns of transcriptional activity in the chromosome of *Escherichia coli*. *Genome Biol* **5:** R86

Koonin EV (2003) Comparative genomics, minimal gene sets and the last common ancestor. *Nat Rev Microbiol* **1:** 127–136

Képès F (2004) Periodic transcriptional organization of the *E. coli* genome. *J Mol Biol* **340:** 957–964

Képès F, Valliant C (2003) Transcription-based solenoidal model of chromosomes. *ComPlexUs* **1:** 171–180

Louarn JM, Bouche JP, Legendre F, Louarn J, Patte J (1985) Characterization and properties of very large inversions of the *E. coli* chromosome along the origin-to-terminus axis. *Mol Gen Genet* **201:** 467–476

Mirkin EV, Mirkin SM (2005) Mechanisms of transcription-replication collisions in bacteria. *Mol Cell Biol* **25:** 888–895

Mushegian AR, Koonin EV (1996) A minimal gene set for cellular life derived by comparison of complete bacterial genomes. *Proc Natl Acad Sci USA* **93:** 10268–10273

Omont N, Képès F (2004) Transcription/replication collisions cause bacterial transcription units to be longer on the leading strand of replication. *Bioinformatics* **20:** 2719–2725

Peter BJ, Arsuaga J, Breier AM, Khodursky AB, Brown PO, Cozzarelli NR (2004) Genomic transcriptional response to loss of chromosomal supercoiling in *Escherichia coli*. *Genome Biol* **5:** R87

Postow L, Hardy CD, Arsuaga J, Cozzarelli NR (2004) Topological domain structure of the *Escherichia coli* chromosome. *Genes Dev* **18:** 1766–1779

R Development Core Team (2008) *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. ISBN 3-900051-07-0, http://www.R-project.org

Segall A, Mahan MJ, Roth JR (1988) Rearrangement of the bacterial chromosome: forbidden inversions. *Science* **241:** 1314–1318

Sharp PM, Li WH (1987) The codon adaptation index—a measure of directional synonymous codon usage bias, and its potential applications. *Nucleic Acid Res* **15:** 1281–1295

Sinden RR, Pettijohn DE (1981) Chromosomes in living *Escherichia coli* cells are segregated into domains of supercoiling. *Proc Natl Acad Sci USA* **78:** 224–228

Tatusov RL, Fedorova ND, Jackson JD, Jacobs AR, Kiryutin B, Koonin EV, Krylov DM, Mazumder R, Mekhedov SL, Nikolskaya AN, Rao BS, Smirnov S, Sverdlov AV, Vasudevan S, Wolf YI, Yin JJ, Natale DA (2003) The COG database: an updated version includes eukaryotes. *BMC Bioinformatics* **4:** 41

Thanbichler M, Shapiro L (2006) Chromosome organization and segregation in bacteria. *J Struct Biol* **156:** 292–303

Travers A, Muskhelishvili G (2005a) Bacterial chromatin. *Curr Opin Genet Dev* **15:** 507–514

Travers A, Muskhelishvili G (2005b) DNA supercoiling—a global transcriptional regulator for enterobacterial growth? *Nat Rev Microbiol* **3:** 157–169

Tse-Dinh YC, Qi H, Menzel R (1997) DNA supercoiling and bacterial adaptation: thermotolerance and thermoresistance. *Trends Microbiol* **5:** 323–326

Willenbrock H, Ussery DW (2004) Chromatin architecture and gene expression in *Escherichia coli*. *Genome Biol* **5:** 252

Wright MA, Kharchenko P, Church GM, Segré D (2007) Chromosomal periodicity of evolutionarily conserved gene pairs. *Proc Natl Acad Sci USA* **104:** 10559–10564