Long-Range Periodic Patterns in Microbial Genomes Indicate Significant Multi-Scale Chromosomal Organization

Timothy E. Allen[1¤a], Nathan D. Price[1¤b], Andrew R. Joyce[2], Bernhard Ø. Palsson[1*]

1 Department of Bioengineering, University of California San Diego, La Jolla, California, United States of America, **2** Bioinformatics Program, University of California San Diego, La Jolla, California, United States of America

Genome organization can be studied through analysis of chromosome position-dependent patterns in sequence-derived parameters. A comprehensive analysis of such patterns in prokaryotic sequences and genome-scale functional data has yet to be performed. We detected spatial patterns in sequence-derived parameters for 163 chromosomes occurring in 135 bacterial and 16 archaeal organisms using wavelet analysis. Pattern strength was found to correlate with organism-specific features such as genome size, overall GC content, and the occurrence of known motility and chromosomal binding proteins. Given additional functional data for *Escherichia coli*, we found significant correlations among chromosome position dependent patterns in numerous properties, some of which are consistent with previously experimentally identified chromosome macrodomains. These results demonstrate that the large-scale organization of most sequenced genomes is significantly nonrandom, and, moreover, that this organization is likely linked to genome size, nucleotide composition, and information transfer processes. Constraints on genome evolution and design are thus not solely dependent upon information content, but also upon an intricate multi-parameter, multi-length-scale organization of the chromosome.

Citation: Allen TE, Price ND, Joyce AR, Palsson BØ (2006) Long-range periodic patterns in microbial genomes indicate significant multi-scale chromosomal organization. PLoS Comput Biol 2(1): e2.

Introduction

Genomes in prokaryotic organisms typically are packed tightly into a nucleoid where they carry out multiple functions simultaneously [1,2]. The condensed DNA within the bacterial nucleoid must not only be efficiently replicated and segregated during cell division [3], but it must also simultaneously participate in the information transfer processes of transcription and translation [4]. Recent studies have significantly advanced our understanding of the ultrastructural and multifunctional organization of prokaryotic chromosomes. DNA in *Escherichia coli* has been found to be packed into supercoiled domains ranging 2–66 kb and averaging ~10 kb [5]. At a slightly longer length-scale, studies using fluorescence in situ hybridization have revealed that the origin and terminus of replication in *E. coli* gravitate toward the poles of the cell throughout replication, but both migrate to the mid-cell region just prior to the initiation of chromosome replication [6]. Fluorescence experiments in synchronized cultures of the aquatic bacterium *Caulobacter crescentus* have revealed the cellular location of 112 individual chromosomal loci throughout replication and cell division [7]. In addition to these imaging techniques, genetic dissection has been used to identify four macrodomains and two less-structured regions in the *E. coli* chromosome [8]. Two of these macrodomains were consistent with those found near the origin and terminus of replication using fluorescence in situ hybridization [6]. However, many issues remain unresolved regarding the intricacies of this arrangement, and particularly the relationship between chromosomal ultrastructure and the processes of transcriptional regulation and protein synthesis [4,9].

Several studies have revealed that genes in bacterial nucleoids tend to be arranged along the long axis of the cell (in the case of rod-shaped bacteria) so as to preserve the linear order of the genes along the chromosome [6,7,10,11]. Given this linear arrangement, prokaryotic genome sequences inherently contain useful information relating to chromosomal ultrastructure since they provide numerous properties as a function of chromosome position [12]. However, the inference of 3-D genome-packing from direct examination of the raw sequence is somewhat challenging at the short length-scales of the nucleotide, gene, or operon (1 bp–10 kb) due to the inherently one-dimensional nature of sequence data and hence the considerable sequence noise over shorter scales. Accordingly, various averaging and filtering methods have been used to identify long-range (i.e.,

Editor: Philip Bourne, University of California San Diego, United States of America

Received July 25, 2005; **Accepted** December 7, 2005; **Published** January 13, 2006

A previous version of this article appeared as an Early Online Release on December 7, 2005 (DOI: 10.1371/journal.pcbi.0020002.eor).

DOI: 10.1371/journal.pcbi.0020002

Abbreviations: CAI, codon adaptation index; FDR, false discovery rate; SD, standard deviation

* To whom correspondence should be addressed. E-mail: palsson@ucsd.edu

¤a Current address: Department of Biomedical Engineering, University of Virginia, Charlottesville, Virginia, United States of America

¤b Current address: The Institute for Systems Biology, Seattle, Washington, United States of America

Synopsis

For more than a decade, the genetic material for a growing number of microbial organisms has been determined experimentally using genome sequencing techniques. These sequenced genomes provide researchers with an abundance of information regarding the composition and capabilities of each organism since they serve as "parts lists" that specify the protein machinery that each cell generates. However, genomes are not merely "lists" but also are typically arranged in nonrandom order. It is thought that this order may be related to some extent to the way in which each genome is packed into the tiny confines of a cell (often more than 1,000-fold packing). The authors have used signal processing methods to identify long-range spatial patterns in the arrangement of most sequenced microbial genomes, and they have related the degree of organization in each genome to various characteristics specific to the corresponding organisms. They have also analyzed in detail the degree of overlap among patterns in numerous different kinds of data for a model bacterial organism, *Escherichia coli*. Their results conclusively demonstrate that there are significant evolutionary constraints that act upon genome organization as well as genome content, and that the interplay between organization and function cannot be ignored in understanding fundamentally how a microbial cell works.

>10-kb) position-dependent patterns in genome-associated properties [12–14]. In order to detect such long-range periodic patterns in inherently noisy chromosome position-dependent data, wavelet analysis has been used in several studies [13,15] (Figure 1). This method has previously been used to detect patterns in gene orientation [14], DNA-bending profiles [16], and gene expression data [17,18] in prokaryotes, as well as GC/AT skew oscillations in human chromosomes [19]. These studies have revealed that genome sequences are generally nonrandom with respect to chromosome position, and that long-range correlations in certain properties (e.g., gene orientation; [14]) exist across many different length-scales.

As more prokaryotic genome sequences become available, it should be increasingly possible to relate the quantitative degree of genome organization to global properties of each organism, including the presence of known nucleoid-binding proteins [20], organism taxa, and genome size and composition. Observed correlations may indicate constraints that affect (or are affected by) genome organization. Furthermore, a study of genome position-dependent patterns in heterogeneous data types in a well-studied model organism such as *E. coli* (e.g., gene expression versus specific codon preferences) may reveal properties that are spatially linked. Therefore, the need exists to define an unbiased, quantitative measure of genome organization from sequence-derived data, compute this quantity for numerous sequenced prokaryotic genomes, relate this quantity to global properties of each organism, and determine the spatial coupling of multiple heterogeneous properties for a well-studied model organism.

In this study, we address these needs by employing wavelet analysis in concert with a bootstrap significance test (Materials and Methods) to compute the pattern strengths of chromosome position-associated datasets derived from 163 sequenced prokaryotic chromosomes. This pattern strength provides a measure of the nonrandom nature of sequence-derived data that is independent of genome length. We then computed the pattern strength of genome position-dependent properties for nearly every sequenced prokaryotic genome, and we related this measure to taxonomic and physiological characteristics of each organism. Finally, we examined disparate genome position-dependent data available for *E. coli* to determine properties that are spatially correlated over multiple length-scales. Our results demonstrate that the degree of organization in bacterial genomes is highly variable and correlates with specific properties, and our analysis of patterns in multiple *E. coli* datasets supports the notion that the overall organization of the bacterial chromosome results from the simultaneous optimization of functional and structural constraints.

Results/Discussion

Pattern Strengths of Sequenced Prokaryotic Organisms

Using the pattern detection method described (Materials and Methods), we computed the pattern strengths for the GC/AT content, fractional gene density, and codon adaptation index (CAI) derived from 163 sequenced prokaryotic chromosomes (Figure 2). The average pattern strength for GC/AT content was 40% (standard deviation [SD] = 20%), 19% (SD = 14%) for gene density, and 37% (SD = 22%) for CAI. (The descriptive statistics for these distributions are summarized in Table 1.) The high SDs indicate that significant chromosome position-dependent patterns vary extensively for different organisms. The relative lack of patterning in gene density is a result of the low positional variability due to the short intergenic regions found in the generally gene-dense prokaryotic organisms. Rank-ordering the genomes by pattern strength revealed the variation in the degree of patterning in sequence-derived parameters in these chromosomes (Figure 2, right column). Table 2 lists the chromosomes containing the strongest and weakest patterns for each parameter, and the scalograms corresponding to the strongest patterns are indicated in the left column of Figure 2. The scalograms for *E. coli* are provided for reference (Figure 2, middle column). Significant patterns were also detected in gene orientation (i.e., strand) for all but one of the chromosomes (Table S1).

Correlation of Pattern Strengths to Organism-Specific Properties

Pattern strengths in the sequence-derived parameters for each chromosome were compared with global properties such as genome length, total AT composition, organism taxon, and the presence of specific nucleoid-binding proteins. Pattern strengths in CAI and GC/AT content were found to be weakly but significantly correlated with genome size ($r = 0.60$, $p = 2.4 \times 10^{-17}$; Figure 3A) and anti-correlated with total AT composition ($r = -0.51$, $p = 2.0 \times 10^{-12}$; Figure 3B). These correlations are consistent with previously observed correlation between genome size and GC-content [21] and suggest an evolutionary requirement for greater genome organization in larger and more GC-rich organisms. However, a causal relationship among these three parameters is impossible to determine at this point. The potential evolutionary constraint regarding genome size may simply be the function of a requirement for a higher-level organization necessary to pack larger genomes into the bacterial cell. The tendency of GC-rich genomes to be more highly patterned is likely linked to physical constraints

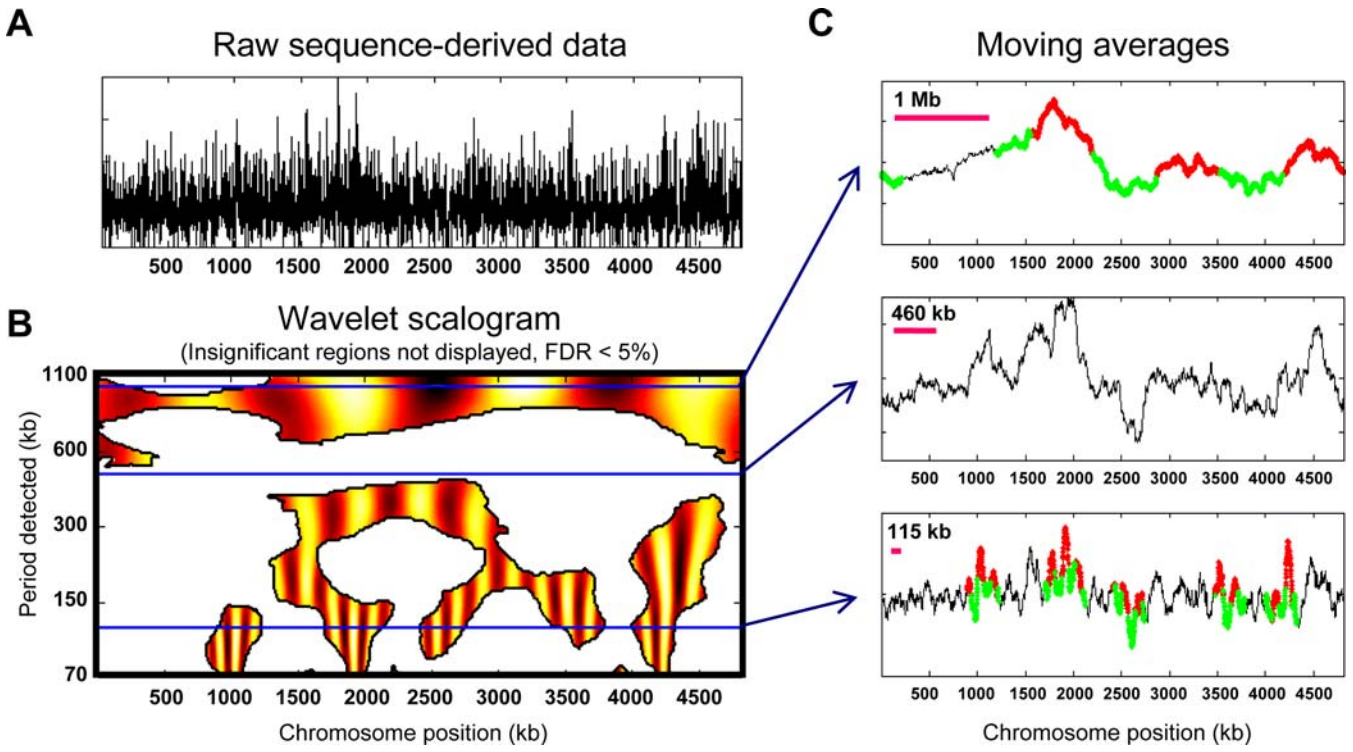

Figure 1. Approach for Detecting Genome Position-Dependent Patterns

(A) Raw sequence-derived data often contain patterns with respect to chromosome position that are not obvious from casual observance. (This example is for the fractional gene density per kilobase for *Salmonella enterica* serovar Typhi strain CT18.)

(B) Wavelet analysis was used to generate a scalogram showing significant chromosome position-dependent patterns in gene density over varying periodicities. The level of significance of the patterns was determined by randomizing the order of the raw sequence data 200× and recomputing the real and imaginary portions of the Morlet wavelet transform values at each point in the scalogram for each randomization. Regions having an FDR greater than 5% are not displayed (white). The pattern strength for this dataset is 33%.

(C) To facilitate the interpretation of the wavelet scalogram, three examples are shown for the moving averages of the raw data at three different length scales: 1 Mb, 460 kb, and 115 kb. Regions highlighted in red/green indicate significant regions of the scalogram at that scale that lie above/below the mean real transform value.

DOI: 10.1371/journal.pcbi.0020002.g001

imposed by the more rigid DNA resulting from the triple hydrogen bond between guanine and cytosine.

We then examined correlation of pattern strength with particular organism-specific characteristics relating to taxon, gram stain, cell shape, and the presence of particular classes of proteins in each organism (summarized in Table 3). The Wilcoxon rank-sum test ($p < 0.05$) was used to assess significance. With respect to organism taxa, patterns in CAI were found to be stronger among the proteobacteria and weaker among the mollicutes and spirochetes. Cell-shape biases in pattern strength included a preference for stronger patterns in rod-shaped bacteria and weaker patterns in spiral-shaped bacteria. No other correlations relating to organism taxa, staining characteristics, or cell shape were observed. However, this analysis is inherently biased by the particular genomes that have been sequenced to date and are thus somewhat skewed toward enteric bacteria and pathogens. As the physiological and morphological diversity of sequenced prokaryotes increases, more definitive conclusions can be drawn regarding possible correlation between genome patterning and such properties as organism lifestyle and cell shape.

Genomes exhibiting the strongest patterns in CAI and GC/AT content had a higher likelihood (Wilcoxon rank-sum $p < 0.05$) of containing genes for flagella and pili than would be expected if the existence of these structures were uncorrelated with pattern strength. As shown in Table 3, the presence of genes encoding the specific nucleoid binding proteins H-NS, Fis, CbpB, Hfq, IciA, Lrp, and Muk was also found to be correlated with overall patterning in CAI. Comparisons of pattern strengths for each sequence-derived parameter revealed no significant correlations, with the exception of GC/AT content versus CAI ($r = 0.74$, $p = 5.6 \times 10^{-29}$). This correlation reflects the fact that CAI and GC/AT content are not actually independent properties, since GC-rich stretches of DNA will favor synonymous codons containing G and C.

Overlap of Patterns in Heterogeneous Datasets in *E. coli*

Since a 600–650 kb periodic pattern has previously been detected in *E. coli* gene expression [17,18,22], the above results motivated an assessment of chromosome position-dependent patterns in functional properties specifically in *E. coli* (in addition to the patterns in GC/AT content, CAI, gene density, and gene orientation discussed above). Correlation of similar patterns in these heterogeneous datasets allows for an evaluation of the structural and functional organization of the *E. coli* genome. Binary matrices of significant pattern density regions were generated for a *p*-value cutoff corresponding to a specified false discovery rate (FDR) [23] (FDR < 5% for our analysis). Unity was assigned to regions in the scalogram deemed to have statistically significant patterning

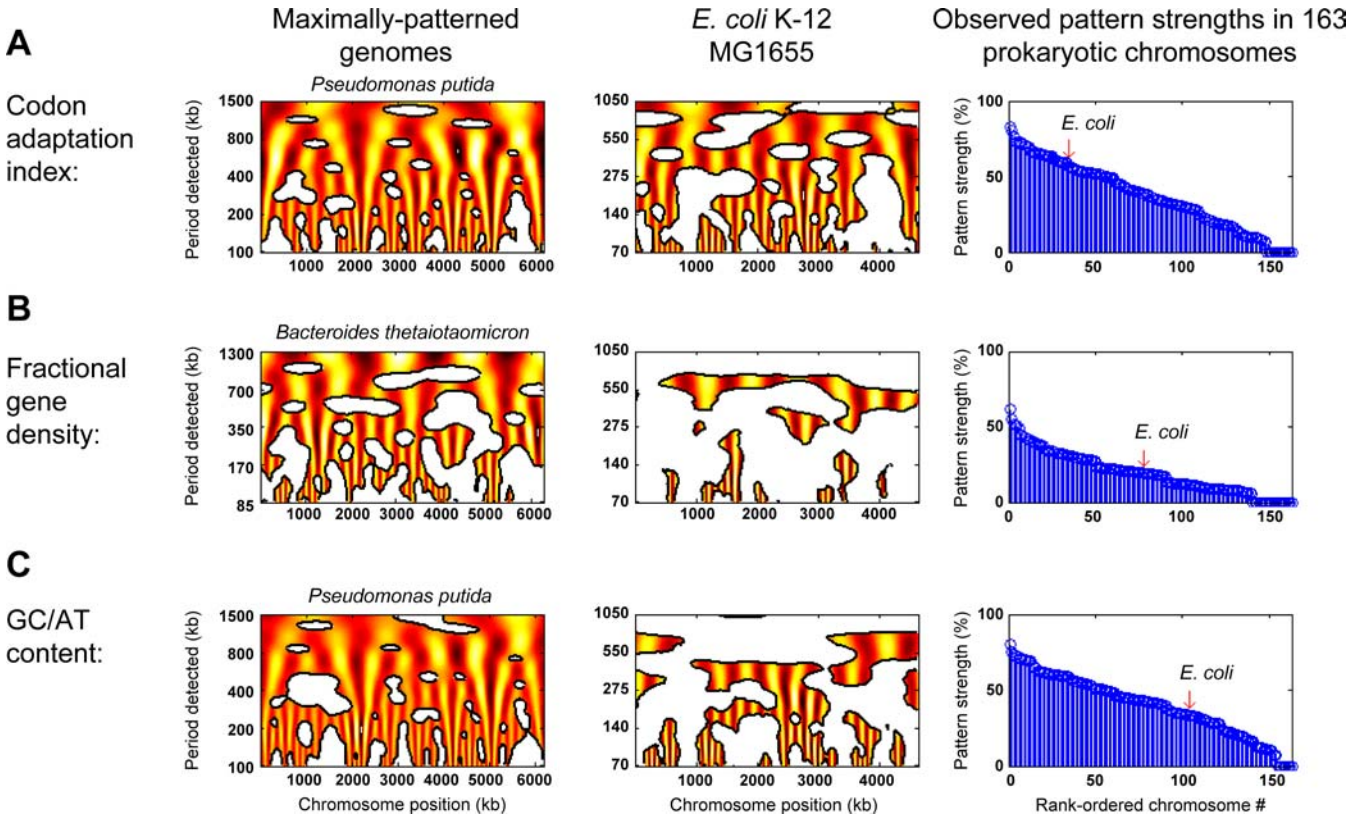

Figure 2. Generality of Chromosome Position-Dependent Patterns in Sequence Properties for 163 Prokaryotic Chromosomes

Continuous wavelet scalograms were computed for most prokaryotic chromosomes sequenced through January 2005 to identify patterns in CAI per gene (A), fractional gene density per kilobase (B), and GC/AT content per kilobase (C). The colored portions of the scalogram indicate significant periodic patterns (FDR $< 5\%$). The degree of patterning for each prokaryotic sequence and each parameter (called the fractional pattern strength) was taken as the percentage of the area of the scalogram containing significant patterns. The first column shows the scalograms for the maximally patterned chromosome found for each sequence property. For reference, the second column shows these scalograms for *E. coli* K-12 MG1655. The third column shows the rank-ordered fractional pattern strengths for the 163 sequenced prokaryotic chromosomes that were analyzed, with *E. coli* indicated relative to the other chromosomes on each plot.

DOI: 10.1371/journal.pcbi.0020002.g002

and zeros assigned elsewhere (Figure 4; Materials and Methods). For any given collection of datasets, the corresponding binary pattern-significance matrices can then be collated and visualized as a contour plot to reveal the extent of overlap in regions of the wavelet scalograms sharing significant *p*-values (Figure 4A).

In analyzing the overlap of patterns in functional genome position-dependent data in *E. coli*, we observed that gene expression [17], gene essentiality [24], and the evolutionary retention index [24] contain significant periodic patterns overlapping at the 650-kb length-scale (Figure 4B) and are strongly (positively) correlated (Figure 4C). Significant patterns in gene expression at the 600–650-kb period were also found to overlap with patterns in fractional gene density and CAI over most of the genome (Figure 5A). This observation is consistent with the known coupling of transcription and translation in prokaryotes [25], since shared positional biases in CAI and expression imply that codon usage (which affects translation) is spatially coupled to gene expression (transcription). Additionally, large-scale periodic patterns (most at the ~650-kb length-scale) in the intragenic preference of specific synonymous codons were detected in *E. coli*, implying consequent positional biases in the corresponding tRNA species. Thus, certain tRNA species will be preferentially demanded over specific regions of the chromosome; e.g., different tRNAs for arginine and lysine will be demanded at regions of either high or low gene expression at the 600–650-kb length-scale (Figure 4D). The observed chromosome-position biases in gene expression and specific codon preferences in *E. coli*, along with the codon adaptation patterns observed in most of the 163 prokaryotic chromosomes analyzed in this study, suggest the existence of spatial gradients in the functional state of specific domains within

Table 1. Descriptive Statistics for Pattern Strengths in GC/AT Content, Gene Density, and CAI across 163 Prokaryotic Chromosomes

Property	Mean	SD	Min	Max
GC/AT content	39.6	19.7	0	80.0
Gene density	19.1	13.9	0	62.0
CAI	36.6	22.0	0	82.4

All values are percentages (%).
SD = standard deviation.

DOI: 10.1371/journal.pcbi.0020002.t001

Table 2. Organisms Exhibiting Either very High or very Low Chromosome Position-Dependent Patterns in Sequence-Derived Data

Rank	GC/AT Content	CAI	Gene Density
1	*Pseudomonas putida* (80.0)	*Pseudomonas putida* (82.4)	*Bacteroides thetaiotaomicron* (62.0)
2	*Xylella fastidiosa* (75.4)	*Mesorhizobium loti* (80.4)	*Nocardia farcinica* (55.2)
3	*Mesorhizobium loti* (73.7)	*Bordetella bronchiseptica* (76.0)	*Synechocystis* sp. PCC6803 (55.2)
4	*Acinetobacter species* (72.6)	*Pseudomonas syringae* (72.7)	*Pseudomonas putida* (50.8)
5	*Burkholderia pseudomallei* chr. 1 (72.2)	*Erwinia carotovora* (72.6)	*Photorhabdus luminescens* (50.6)
6	*Bacillus subtilis* (71.8)	*Salmonella enterica* typhi (72.6)	*Bacteroides fragilis* (49.2)
7	*Bordetella bronchiseptica* (71.2)	*Bacteroides thetaiotaomicron* (72.3)	*Bifidobacterium longum* (44.7)
8	*Bacteroides thetaiotaomicron* (70.3)	*Burkholderia pseudomallei* chr. 1 (71.3)	*Bordetella bronchiseptica* (44.2)
9	*Pseudomonas aeruginosa* (70.2)	*Ralstonia solanacearum* chr. 1 (71.1)	*Burkholderia pseudomallei* chr. 2 (43.2)
10	*Geobacillus kaustophilus* (69.9)	*Nocardia farcinica* (70.4)	*Xanthomonas axonopodis* (42.6)
154	*Tropheryma whippelii* (7.3)	*Rickettsia typhi* (0)	*Leptospira interrogans* chr. 1 (0)
155	*Nanoarchaeum equitans* (0)	*Rickettsia prowazekii* (0)	*Helicobacter hepaticus* (0)
156	*Haloarcula marismortui* chr. 2 (0)	*Mycoplasma pulmonis* (0)	*Deinococcus radiodurans* chr. 1 (0)
157	*Wolbachia pipientis* (0)	*Mycoplasma genitalium* (0)	*Coxiella burnetii* (0)
158	*Thermosynechococcus elongatus* (0)	*Leptospira interrogans* chr. 2 (0)	*Corynebacterium efficiens* (0)
159	*Rickettsia typhi* (0)	*Helicobacter pylori* (0)	*Chlorobium tepidum* (0)
160	*Rickettsia prowazekii* (0)	*Fusobacterium nucleatum* (0)	*Campylobacter jejuni* (0)
161	*Parachlamydia species* (0)	*Coxiella burnetii* (0)	*Brucella Suis* chr. 1 (0)
162	*Ehrlichia ruminantium* (0)	*Borrelia garinii* (0)	*Brucella melitensis* chr. 2 (0)
163	*Anabaena nostoc* (0)	*Borrelia burgdorferi* (0)	*Brucella melitensis* chr. 1 (0)

The numbers shown in parentheses refer to fractional pattern strengths for each organism. See Table S1 for complete listings, including strain names. All fractional pattern strengths are displayed as percentages (%).
DOI: 10.1371/journal.pcbi.0020002.t002

each folded nucleoid [26]. These gradients may lead to spatial gradients in tRNA concentration that result from differential local demands for specific tRNA species [27].

Analysis of all 163 chromosomes revealed that long-range patterns in synonymous codon usage (CAI) are not strictly independent from those in GC/AT composition. However, patterns in sequence-derived DNA-bending parameters for *E. coli* (e.g., intrinsic curvature, propeller twist, stacking energy, etc.) almost completely overlap with patterns in GC/AT content (Figure 5B). As described previously, the GC/AT content reflects the average bendability of the chromosome over multiple length-scales [12]. Thus, the observed correlation of pattern strengths in CAI and GC/AT content implies a general coupling of information storage with chromosomal bending. The strongest overlap in nucleotide sequence content and sequence-derived bending parameters in *E. coli* consists of a 600–650-kb periodic pattern near the origin of replication between the 3,800–250-kb nucleotide coordinates (82′ to 5′). This region closely coincides with the *E. coli*–origin macrodomain detected in previous studies cited [8], and the structural regularity at the 600-kb length scale may facilitate localization of the origin to one of the cell poles during replication [6]. These DNA-bending associated datasets also contain localized periodic patterns at length scales on the order of 80–100 kb that occur in specific regions of the chromosome. The maximum pattern density in GC/AT content in this range occurred at the 74-kb period, containing eight localized patterns. Six of these eight pattern-rich segments were found to be significantly enriched (hypergeometric $p < 0.001$) with genes belonging to particular functional classes [28], which included prophage-related genes and genes encoding membrane-associated proteins

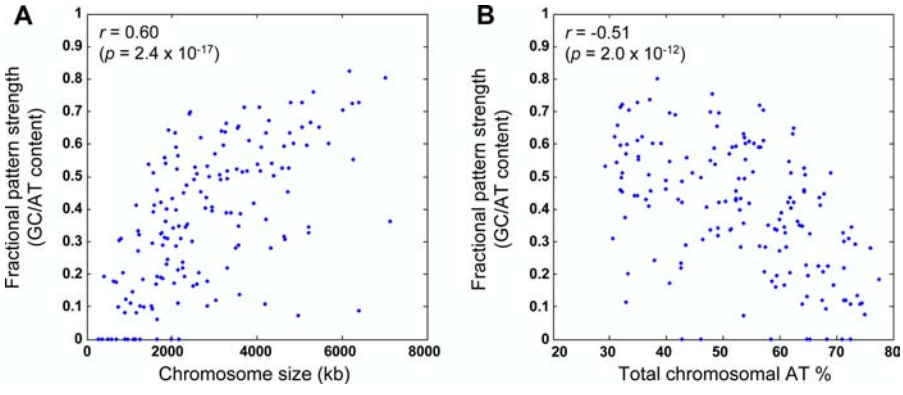

Figure 3. Correlations between Sequence-Derived Properties for 163 Prokaryotic Chromosomes
(A) Correlation of fractional pattern strength in GC/AT content with chromosome length.
(B) Anti-correlation of fractional pattern strength in GC/AT content with total chromosomal AT%. The correlation coefficients and associated *p*-values are indicated on each graph.
DOI: 10.1371/journal.pcbi.0020002.g003

Table 3. Correlation between Pattern Strength in CAI and Organism Taxon, Gram Staining, Cell Shape, and the Presence of Known Motility and Nucleoid Proteins

	Average Pattern Strength[1]		
	Selected	**Remainder**	***p*-Value[2]**
Proteobacteria	**40.40**	**27.28**	**0.006**
— gamma	**41.96**	**30.16**	**0.025**
— beta	**58.29**	**32.08**	**0.045**
— d/e	8.37	33.66	0.108
— alpha	38.52	32.19	0.446
Firmicutes	31.68	33.60	0.771
Bacillales	35.04	32.79	0.755
Lactobacillales	42.57	32.29	0.222
Clostridia	34.73	32.95	0.755
Mollicutes	**10.14**	**34.53**	**0.009**
Actinobacteria	40.94	32.42	0.323
Fusobacteria	0.00	33.45	0.128
Chlamydia	16.90	33.65	0.174
Spirochete	**6.97**	**34.38**	**0.011**
Cyanobacteria	18.09	33.61	0.190
Green sulfur bacteria	10.82	33.32	0.331
Radioresistant bacteria	19.77	33.37	0.408
Hyperthermophilic bacteria	32.79	33.05	0.940
gram +	33.92	32.59	0.656
gram -	33.95	31.69	0.712
cocci	32.38	33.13	0.994
rods	**41.40**	**23.82**	**0.000**
spirals	**5.58**	**34.82**	**0.003**
flagellum	**43.17**	**32.31**	**0.007**
pilus	**45.58**	**34.48**	**0.007**
Hu/IHF	33.08	32.24	0.855
H-NS/StpA	**46.15**	**28.53**	**0.001**
Dps	35.16	27.60	0.121
Fis	**44.09**	**28.73**	**0.003**
CbpA	35.71	32.44	0.565
DnaA	33.45	0.00	0.128
CbpB	**49.25**	**26.72**	**0.000**
Hfq	**41.65**	**21.47**	**0.000**
IciA/LysR	**40.78**	**11.93**	**0.000**
Lrp/AsnC	**42.30**	**20.60**	**0.000**
Smc (muk)	**49.89**	**31.71**	**0.045**

[1]The "selected" column indicates the average pattern strength in the organisms meeting each criterion in the leftmost column, and the "remainder" column shows the average pattern strength of all the remaining organisms.
[2]The *p*-values were computed from the Wilcoxon rank-sum test, and the bolded rows met a cutoff of $p < 0.05$.
DOI: 10.1371/journal.pcbi.0020002.t003

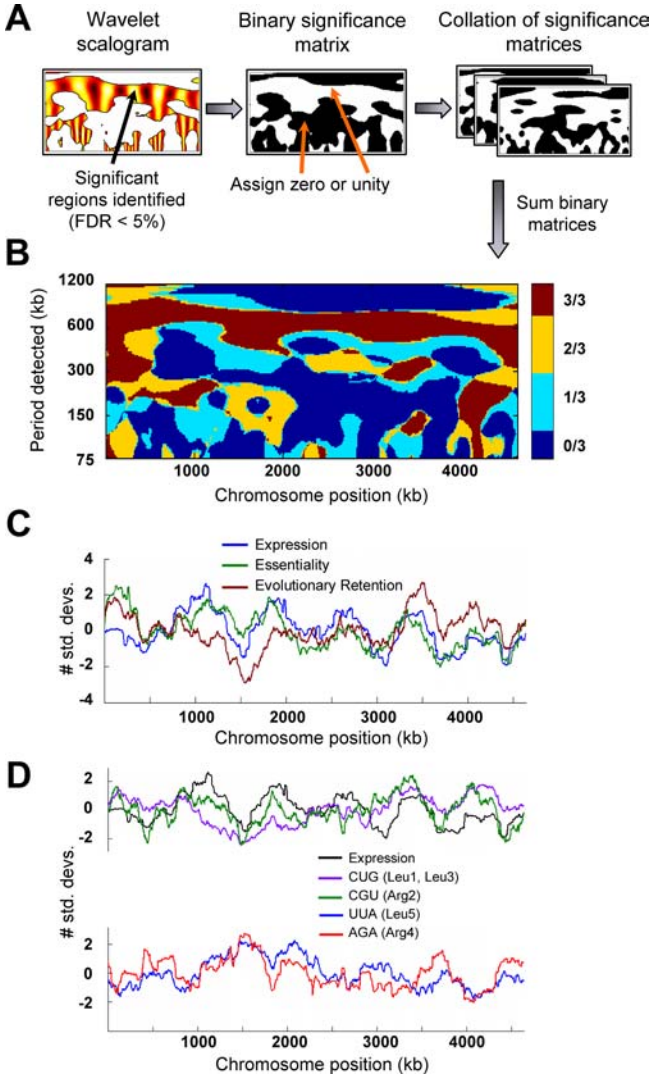

Figure 4. Correlation of Specific Chromosome Position-Dependent Patterns in *E. coli* Functional Properties

(A) Wavelet scalograms calculated for gene expression, gene essentiality, and evolutionary retention index were converted to a binary significance matrix by setting each significant point in a scalogram (FDR $<$ 5%) to unity and each non-significant point to zero.
(B) These binary matrices were summed across the three properties listed above to determine chromosome position-dependent patterns that were consistent across the different properties, and the resulting map was color-coded according to how many of the properties shared significant patterns. The red-colored segments indicate the periods and chromosome positions at which all three properties exhibited significant patterns. The averaged data have been normalized such that the mean is zero and the tick marks indicate SDs from the mean value.
(C) Correlation of gene expression, essentiality, and evolutionary retention averaged at a window of 325 kb (650-kb period).
(D) Correlation of gene expression with intragenic codon preferences for two of the major codons encoding leucine (CUG) and arginine (CGU), and anti-correlation of these with preferences for the corresponding minor codons, UUA and AGA, at a moving average of 325 kb. The labels are as described above in (C).
DOI: 10.1371/journal.pcbi.0020002.g004

(flagellar, energy production and transport, and cell-surface antigens). The enrichment of patterned regions with genes of extrachromosomal origin implies a preferred regularity in chromosome structure and nucleotide content that facilitates foreign DNA incorporation. In the case of the regions enriched in membrane-associated proteins (flagellar, cell surface, etc.), the translocation of these proteins [29] may be enhanced by regular structure at the 80–100-kb length-scale.

Genome topology has been shown to be a selection target in the long-term evolution of *E. coli* [30]. Our results demonstrate that prokaryotic genomes generally possess significant organization that increases with genome size, overall GC composition, and the presence of several known nucleoid-binding proteins. Thus, genome composition and size may impose additional constraints on the evolution of gene order and chromosomal arrangement in prokaryotes. Given that the spatial organization of chromosomal loci within a replicating *E. coli* cell is linearly ordered along the cellular axis [6,11], the analysis presented here would imply the existence of six subchromosomal functional domains in

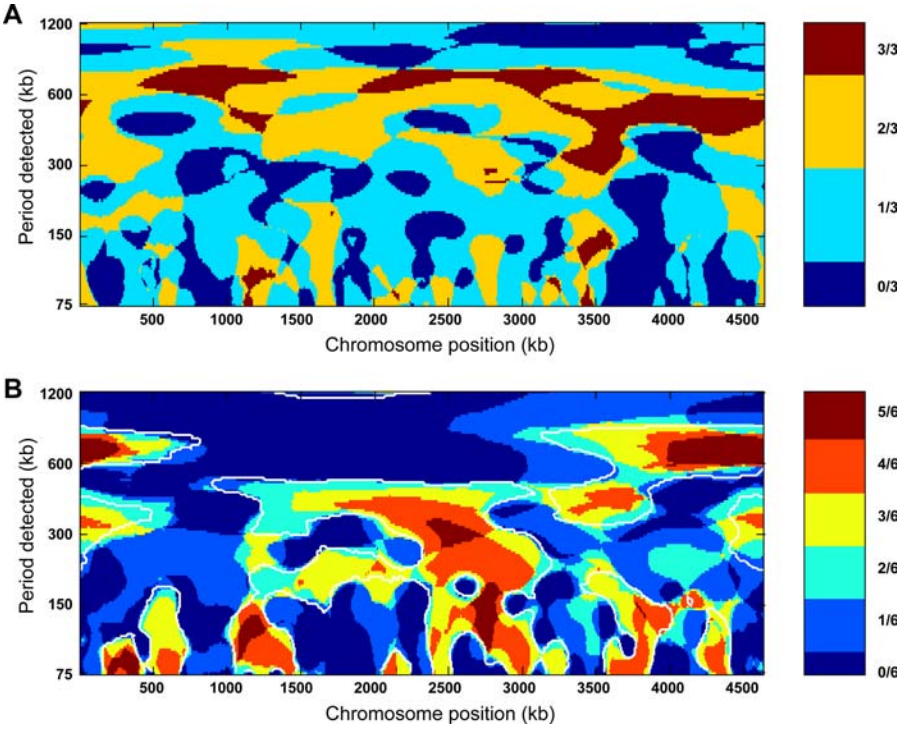

Figure 5. Overlay Plots of Significant Regions of Wavelet Scalograms for Various *E. coli* Parameters

(A) Degree of significant pattern overlap in expression, gene density, and codon adaptation in *E. coli*. Binary matrices corresponding to significant regions of wavelet scalograms (FDR < 5%) for gene expression, CAI, and fractional gene density in *E. coli* were summed as described in Materials and Methods. A periodic pattern of 600–650 kb can be seen across nearly three-quarters of the chromosome.

(B) Degree of significant pattern overlap sequence-derived DNA-bending parameters in *E. coli*. Binary matrices corresponding to significant regions of wavelet scalograms (FDR < 5%) for intrinsic curvature, DNAseI sensitivity, protein-induced deformability, propeller twist, stacking energy, and nucleosome position preference in *E. coli* [12] were summed as described in the text. The white contour lines outline the significant regions of the wavelet scalogram for GC/AT content, thus demonstrating that these parameters are not independent.

DOI: 10.1371/journal.pcbi.0020002.g005

the *E. coli* genome [22]. This notion of highly expressed topological domains has been suggested before [31] and is consistent with the macrodomains elucidated by genetic dissection of *E. coli* [8]. The boundaries of those four domains and two less-structured regions [8] align with the boundaries of the regions of high and low gene expression, gene essentiality, and evolutionary retention in *E. coli* at the 600–650-kb length-scale (Figure 6). The observed patterns reveal that information transfer and chromosomal organization within the *E. coli* nucleoid are spatially interlinked.

Implications and Conclusions

As demonstrated in the analyses described above, genome sequences and sequence-derived properties are significantly patterned (i.e., non-randomly distributed) with respect to chromosome position in most of the prokaryotic genomes sequenced to date (Figure 2). The degree of patterning in a bacterial organism is positively correlated with genome size, overall GC-content, the presence of several known nucleoid-binding proteins, and the presence of flagellar proteins (Figure 3; Table 3). These results strongly suggest the existence of structural constraints imposed by organism-specific features on the evolution of genome organization and base-pair composition in each organism.

In *E. coli*, a more detailed analysis of available data demonstrates that patterns in multiple disparate properties are interlinked (Figures 4 and 5). The consistency of the 650-

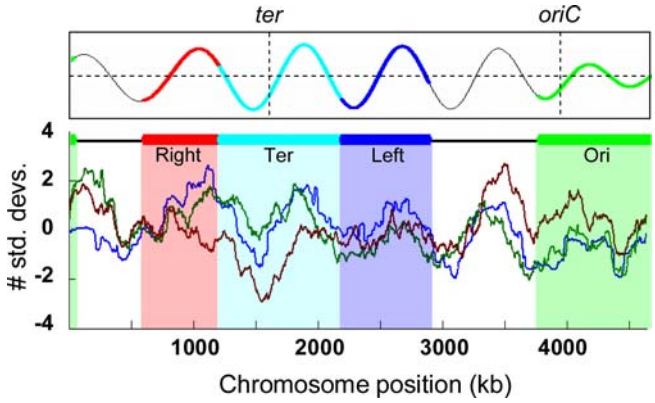

Figure 6. Comparison of *E. coli* Gene Expression, Essentiality, and Evolutionary Retention at 600–650-kb Length Scale with Experimentally Identified Chromosome Macrodomains [8]

The four shaded regions correspond to four macrodomains identified previously based upon the frequency of recombination events following genetic dissection of the *E. coli* chromosome. The two unshaded regions correspond to less-structured macrodomains. The traces in the lower panel are exactly as described in Figure 4C. The upper panel is a section of the wavelet scalogram for *E. coli* gene expression at a 650-kb period. Segments of this wavelet transform trace have been colored to correspond to the experimentally identified chromosome macrodomains.

DOI: 10.1371/journal.pcbi.0020002.g006

kb chromosome macrodomains identified using wavelet analysis of expression data [17] with those identified from genetics experiments [8] indicates that large-scale genome packing is indeed linked to transcription, as has been previously hypothesized [4] (Figure 6). This work has additional implications for de novo genome design [32], in that gene order and composition—and the resulting chromosomal ultrastructure—are significant design variables that will likely need to be taken into account. Given the non-random distribution of these parameters in nearly all sequenced prokaryotes, as well as the linked nature of disparate parameters in *E. coli*, it is clear that any genome design endeavor will involve a multivariable, multidimensional optimization problem. The present study constitutes an early step in the evolution of systems biology from analyses of component (1-D) and systemic (2-D) annotations [33] toward the systems analysis of 3-D genome organization.

Materials and Methods

Chromosome position-associated datasets. Datasets were analyzed from most prokaryotic genome sequences published through January 2005 and were downloaded from the CBS Genome Atlas Database [34] (http://www.cbs.dtu.dk/services/GenomeAtlas). Four types of chromosome position-dependent data were analyzed for 151 prokaryotic organisms (corresponding to 163 chromosomes in 16 archaeal and 135 bacterial organisms): 1) GC/AT content averaged in kilobase bins, 2) gene orientation (i.e., strand), 3) fractional gene density (defined as the number of genes—or fractions of genes—per kilobase), and 4) CAI [35] per gene. For the CAI, we used the global codon usage as the reference set to maintain consistency, since the highly expressed genes for some of the organisms may not be predictable a priori. GC and AT content are by definition inversions of one another and are strictly anti-correlated, so any patterns present in either property will be identical. Thus, patterns in these properties are simply referred to as patterns in GC/AT content. The analysis of additional data from *E. coli* K-12 MG1655 included sequence-derived biophysical parameters averaged across 1-kb segments [12], gene classifications and product locations [28], gene expression [17], gene essentiality [24], and evolutionary retention indices computed based upon homology with 32 representative bacterial sequences [24].

Pattern detection by wavelet analysis and significance testing. Wavelet analysis, reviewed in detail elsewhere [36], is an approach whereby irregular patterns in biological data may be elucidated [14,15,17–19,37]. In short, each genome-scale dataset was ordered according to position along the chromosome. These ordered data, $f(x)$ (where x is defined as the nucleotide position along the chromosome), were then continuously integrated using a family of filter functions to obtain a transform value for numerous filter widths (i.e., scales, designated a) centered at each position x in the dataset:

$$W(x,a) = \frac{1}{\sqrt{a}} \int_{-\infty}^{\infty} g\left(\frac{x'-x}{a}\right) f(x')\,dx' \qquad (1)$$

The filter function used in this study was the Morlet wavelet, defined as:

$$g(x) = e^{i5x} e^{-x^2/2}. \qquad (2)$$

This particular wavelet was chosen because the length scale of the transform corresponds approximately to the period of any localized pattern [36]. The resulting transform values may be plotted in the form of a scalogram (Figure 1B), comprised of a contour plot in which the *x*-axis is the position along the genome *(x),* and the *y*-axis is the length scale *(a)* at which the transform is computed. Given that we employed the Morlet wavelet, this scalogram is useful for elucidating the strength of a range of periodicities localized at each point in time-series data (or, in this case, chromosome position-associated data). The particular voices (i.e., length scales) assessed in the transform for each genome were chosen such that the length scales presented on each scalogram correspond to periods between approximately 1.5% and 20% of the overall genome size.

Currently, no standard statistical methods of verifying patterns identified using continuous wavelet transforms are in common use. Thus, the significance of each transform value was ascertained by a bootstrap approach in which the order of the data points along the chromosome was randomized 200×, and the real and imaginary portions of the Morlet wavelet transform were recomputed for each randomized dataset (described previously for the real portion of the Morlet wavelet [17]). As described in Protocol S1, the randomization of each genome position-associated dataset was performed on either a gene-by-gene basis (for annotation-derived data) or on a kilobase-by-kilobase basis (for annotation-independent properties such as GC content). Thus, the null hypothesis against which each wavelet scalogram was tested consisted of the wavelet transform of a "scrambled" dataset, where the unit of chromosome which was scrambled was either the gene or a kilobase segment. A p-value was then computed for each point in the scalogram based upon the number of times the magnitude of the transform value from each randomization exceeded that of the original transform.

The p-value cutoff corresponding to a selected FDR [23] (FDR $<$ 5%) was then determined from the distribution of p-values computed for each scalogram from the randomization tests. Given this cutoff, one can generate a binary matrix (the same size as the scalogram) containing unity for each point in the scalogram for which FDR $<$ 0.05, and zeroes elsewhere. The ratio of the sum of the non-zero elements in this binary matrix to the total matrix size is taken to be the pattern strength of a given dataset (colored areas in Figure 1B). For matrices of the same size (as for *E. coli* gene expression, essentiality, and evolutionary retention), the sum of the binary significance matrices yields the degree of pattern overlap, as illustrated schematically in Figure 4A.

Controls. Presented in the Protocol S1 are a set of positive and negative controls for the wavelet transform and bootstrap procedure described above. The negative controls showed that no significant patterns were detected in trivial or randomly ordered datasets (for which no pattern would be expected a priori), thus effectively ruling out the possibility that the observed periodic patterns are simply artifacts inherent either in the wavelet filter used or spurious cyclic patterns caused by outliers in otherwise random data (called the Slutzky-Yule effect when observed in moving averages). Wavelet analysis was performed for a 1-Mb subset of the *Pseudomonas putida* GC/AT dataset in order to rule out the possibility that the correlation shown in Figure 3A was due to an artifact of the wavelet voices chosen for the varying genome sizes. No significant decrease in fractional pattern strength was detected for the smaller subset.

Supporting Information

Protocol S1. Additional Material Providing More Detailed Experimental Methods and Positive and Negative Controls

Found at DOI: 10.1371/journal.pcbi.0020002.sd001 (195 KB DOC).

Table S1. Complete List of Computed Pattern Strengths for Each Chromosome in this Study, along with Associated Organismal Data

Found at DOI: 10.1371/journal.pcbi.0020002.st001 (53 KB XLS).

Acknowledgments

We thank Chris Herring, Markus Herrgård, Jennifer Reed, Steve Fong, and Jason Papin for stimulating discussions and comments on the manuscript, Adam Feist for assistance with data pre-processing, and the National Institutes of Health (GM57089) and National Science Foundation (BES 03–31342) for funding and support.

Author contributions. TEA and BØP conceived and designed the experiments. TEA performed the experiments. TEA, NDP, and ARJ analyzed the data. TEA and ARJ contributed reagents/materials/analysis tools. TEA, NDP, ARJ, and BØP wrote the paper.

Competing interests. BØP serves on the Scientific Advisory Board of Genomatica. ■

References

1. Woldringh CL, Odijk T (1999) Structure of DNA within the bacterial cell: Physics and physiology. In: Charlebois RL, editor. Organization of the prokaryotic genome. Washington (DC): ASM Press. pp. 171–187.
2. Zimmerman SB (2003) Underlying regularity in the shapes of nucleoids of *Escherichia coli*: Implications for nucleoid organization and partition. J Struct Biol 142: 256–265.
3. Sherratt DJ (2003) Bacterial chromosome dynamics. Science 301: 780–785.

4. Cook PR (2002) Predicting three-dimensional genome structure from transcriptional activity. Nat Genet 32: 347–352.
5. Postow L, Hardy CD, Arsuaga J, Cozzarelli NR (2004) Topological domain structure of the *Escherichia coli* chromosome. Genes Dev 18: 1766–1779.
6. Niki H, Yamaichi Y, Hiraga S (2000) Dynamic organization of chromosomal DNA in *Escherichia coli*. Genes Dev 14: 212–223.
7. Viollier PH, Thanbichler M, McGrath PT, West L, Meewan M, et al. (2004) Rapid and sequential movement of individual chromosomal loci to specific subcellular locations during bacterial DNA replication. Proc Natl Acad Sci U S A 101: 9257–9262.
8. Valens M, Penaud S, Rossignol M, Cornet F, Boccard F (2004) Macrodomain organization of the *Escherichia coli* chromosome. EMBO J 23: 4330–4341.
9. Thanbichler M, Viollier PH, Shapiro L (2005) The structure and function of the bacterial chromosome. Curr Opin Genet Dev 15: 153–162.
10. Wu LJ, Errington J (1998) Use of asymmetric cell division and spoIIIE mutants to probe chromosome orientation and organization in *Bacillus subtilis*. Mol Microbiol 27: 777–786.
11. Breier AM, Cozzarelli NR (2004) Linear ordering and dynamic segregation of the bacterial chromosome. Proc Natl Acad Sci U S A 101: 9175–9176.
12. Pedersen AG, Jensen LJ, Brunak S, Staerfeldt HH, Ussery DW (2000) DNA structural atlas for *Escherichia coli*. J Mol Biol 299: 907–930.
13. Arneodo A, Bacry E, Graves PV, Muzy JF (1995) Characterizing long-range correlations in DNA sequences from wavelet analysis. Phys Rev Lett 74: 3293–3296.
14. Audit B, Ouzounis CA (2003) From genes to genomes: Universal scale-invariant properties of microbial chromosome organisation. J Mol Biol 332: 617–633.
15. Lio P (2003) Wavelets in bioinformatics and computational biology: State of art and perspectives. Bioinformatics 19: 2–9.
16. Audit B, Vaillant C, Arneodo A, D'Aubenton-Carafa Y, Thermes C (2004) Wavelet analysis of DNA-bending profiles reveals structural constraints on the evolution of genomic sequences. (Germany) J Biol Phys 30: 33–81.
17. Allen TE, Herrgard MJ, Liu M, Qiu Y, Glasner JD, et al. (2003) Genome-scale analysis of the uses of the *Escherichia coli* genome: Model-driven analysis of heterogeneous data sets. J Bacteriol 185: 6392–6399.
18. Jeong KS, Ahn J, Khodursky AB (2004) Spatial patterns of transcriptional activity in the chromosome of *Escherichia coli*. Genome Biol 5: R86.
19. Nicolay S, Argoul F, Touchon M, d'Aubenton-Carafa Y, Thermes C, et al. (2004) Low frequency rhythms in human DNA sequences: A key to the organization of gene location and orientation? Phys Rev Lett 93: 108101.
20. Kim J, Yoshimura SH, Hizume K, Ohniwa RL, Ishihama A, et al. (2004) Fundamental structural units of the *Escherichia coli* nucleoid revealed by atomic force microscopy. Nucleic Acids Res 32: 1982–1992.
21. Bentley SD, Parkhill J (2004) Comparative genomic structure of prokaryotes. Annu Rev Genet 38: 771–791.
22. Lobner-Olesen A, Marinus MG, Hansen FG (2003) Role of SeqA and Dam in *Escherichia coli* gene expression: A global/microarray analysis. Proc Natl Acad Sci U S A 100: 4672–4677.
23. Benjamini Y, Hochberg Y (1995) Controlling the false discovery rate: A practical and powerful approach to multiple testing. J R Stat Soc B 57: 289–300.
24. Gerdes SY, Scholle MD, Campbell JW, Balazsi G, Ravasz E, et al. (2003) Experimental determination and system level analysis of essential genes in *Escherichia coli* MG1655. J Bacteriol 185: 5673–5684.
25. Neidhardt FC, Ingraham JL, Schaechter M (1990) Physiology of the Bacterial Cell. Sunderland (Massachusetts): Sinauer. 506 p.
26. Rocha EPC, Guerdoux-Jamet P, Moszer I, Viari A, Danchin A (2000) Implication of gene distribution in the bacterial chromosome for the bacterial cell factory. J Biotechnol 78: 209–219.
27. Danchin A, Guerdoux-Jamet P, Moszer I, Nitschké P (2000) Mapping the bacterial cell architecture into the chromosome. Philos Trans R Soc Lond B Biol Sci 355: 179–190.
28. Serres MH, Riley M (2000) MultiFun, a multifunctional classification scheme for *Escherichia coli* K-12 gene products. Microb Comp Genomics 5: 205–222.
29. Woldringh CL (2002) The role of co-transcriptional translation and protein translocation (transertion) in bacterial chromosome segregation. Mol Microbiol 45: 17–29.
30. Crozat E, Philippe N, Lenski RE, Geiselmann J, Schneider D (2005) Long-term experimental evolution in *Escherichia coli*. XII. DNA topology as a key target of selection. Genetics 169: 523–532.
31. Ussery D, Larsen TS, Wilkes KT, Friis C, Worning P, et al. (2001) Genome organisation and chromatin structure in *Escherichia coli*. Biochimie 83: 201–212.
32. Smith HO, Hutchison CA III, Pfannkoch C, Venter JC (2003) Generating a synthetic genome by whole genome assembly: phiX174 Bacteriophage from synthetic oligonucleotides. Proc Natl Acad Sci U S A 100: 15440–15445.
33. Palsson BO (2004) Two-dimensional annotation of genomes. Nat Biotechnol 22: 1218–1219.
34. Hallin PF, Ussery DW (2004) CBS Genome Atlas Database: A dynamic storage for bioinformatic results and sequence data. Bioinformatics 20: 3682–3686.
35. Sharp PM, Li WH (1987) The Codon Adaptation Index—A measure of directional synonymous codon usage bias, and its potential applications. Nucleic Acids Res 15: 1281–1295.
36. Torrence C, Compo GP (1998) A practical guide to wavelet analysis. Bull Amer Meteor Soc 79: 61–78.
37. Murray KB, Gorse D, Thornton JM (2002) Wavelet transforms for the characterization and detection of repeating motifs. J Mol Biol 316: 341–363.