OXFORD

Briefings in Bioinformatics, 22(3), 2021, 1–11

doi: 10.1093/bib/bbaa172
Method Review

The functional determinants in the organization of bacterial genomes

Zhaoqian Liu, Jingtong Feng, Bin Yu, Qin Ma and Bingqiang Liu

Corresponding authors: Qin Ma, Department of Biomedical Informatics, College of Medicine, The Ohio State University, Columbus, OH 43210, USA. Tel.: +1-614-688-9857. E-mail: qin.ma@osumc.edu; Bingqiang Liu, School of Mathematics, Shandong University, Jinan, Shandong 250100, China. Tel.: +86-531-88363455. E-mail: bingqiang@sdu.edu.cn

Abstract

Bacterial genomes are now recognized as interacting intimately with cellular processes. Uncovering organizational mechanisms of bacterial genomes has been a primary focus of researchers to reveal the potential cellular activities. The advances in both experimental techniques and computational models provide a tremendous opportunity for understanding these mechanisms, and various studies have been proposed to explore the organization rules of bacterial genomes associated with functions recently. This review focuses mainly on the principles that shape the organization of bacterial genomes, both locally and globally. We first illustrate local structures as operons/transcription units for facilitating co-transcription and horizontal transfer of genes. We then clarify the constraints that globally shape bacterial genomes, such as metabolism, transcription and replication. Finally, we highlight challenges and opportunities to advance bacterial genomic studies and provide application perspectives of genome organization, including pathway hole assignment and genome assembly and understanding disease mechanisms.

Key words: bacterial genome organization; transcription unit; chromosomal folding structure; gene strand bias; computational model

Introduction

As essential life forms on earth, bacteria have been around for at least 3.5 billion years [1]. To date, bacteria have been found within various environments, such as mountaintops [2], marine and terrestrial subsurface [3,4], the guts of animals [5] and even the frozen rocks and ice [5]. Different from eukaryotic genomes, most bacterial genomes consist of a circular DNA molecule of smaller size ranging from 100 to 15 000 kbp [6]. One of the key reasons for bacteria with the small genome size is that most bacterial species have undergone genomic degradation by the deletion of superfluous sequences during evolution [7] and show a relatively small number of junk DNA [7]. Furthermore, a third of the annotated genes of bacteria overlap with others [8], greatly expanding the functional potential from a small set of genetic information. The property of a small size genome enables DNA

Zhaoqian Liu is a PhD student in School of Mathematics, Shandong University, and now she is a visiting scholar at the Ohio State University. Her research interest is microbiome studies.
Jingtong Feng is a master student in School of Mathematics, Shandong University. Her research interest is the bacterial structure.
Bin Yu, PhD, is an associate professor in College of Mathematics and Physics, Qingdao University of Science and Technology. His primary research interests are artificial intelligence, data mining and biomedical image processing.
Qin Ma, PhD, is an associate professor in the Department of Biomedical Informatics, the Ohio State University. Dr Ma has over 10 years of research experience in studying how functional machinery encoded in a (meta-)genome.
Bingqiang Liu, PhD, is a professor in School of Mathematics, Shandong University. His primary research strength is microbial regulatory network construction.

Submitted: 30 May 2020; **Received (in revised form):** 30 June 2020

transactions (e.g. replication and repair) at an appropriate time [9,10], making bacteria highly adaptable to diverse environments [11].

The insight into the plasticity of bacterial genomes introduces a fundamental question: are the genomes organized in arbitrary manners or following some rules during evolution? The initial understanding of genome structures was proposed in the 1960s. Jacob and Monod observed that genes involving in uptake and metabolism of lactose are arranged together in the genome (i.e. *lac* operon) [12,13], which are controlled by a single promoter and co-transcribe as a single unit [12]. This finding indicates that the genome arrangement is not arbitrary but follows certain principles with important functional implications [14]. Based on this landmark work, more effort has been made for a full understanding of underlying organization principles. Yet the progress was stalled by the lack of experimental data. Until 1995, the first two complete bacterial genome sequences were published [15–17]. Since then, the increasing availability of genome data has greatly improved the understanding of genome arrangement.

Compelling evidence suggests that genome structures typically have direct or indirect associations with significant cellular processes. For example, the existence of operon structure on genomes not only coordinate expression and regulation of the constituent genes [12] but also improve the persistence of genes by horizontal transfer within and among taxa (the selfish operon model) [18]. Specifically, due to the heterogeneity of natural selection on both temporal and spatial scales, the genes encoding unimportant or rarely employed functions are more likely to evolutionary extinction with genetic drift [18]. To escape such loss, genes with neighborly functions tend to form clusters for the increasing probabilities of horizontal transfer to novel genomes [18]. This model explicated the existence of gene clusters from a new perspective, providing mechanistic insights into the local structures of bacterial genomes. Furthermore, another study observed that the locations of operons are under strong constrains of metabolic pathways, where operons participating in frequently activated pathways are clustered more tightly along genomes [19].

Beyond the local gene arrangements, there are discernible patterns of global genome organization [20,21]. For example, there is a unique origin of replication for most bacteria. During replication, DNA will open up at this origin and form two Y-shaped structures (a.k.a., replication forks) [22]. The replication forks move in opposite directions until they arrive at the terminus [22]. Based on the directions of DNA chain elongation and fork movement, the strands can be classified as leading strands (the one in which the directions of these two are the same) and lagging strands (the one in which they are opposite) [22]. The general observation suggested that the majority of protein-encoding genes are located on the leading strands, ranging from 50% to 90% across different bacteria [23,24]. Intuitively this makes sense to avoid head-on collisions in terms of DNA replication and transcription [25]. Additionally, the spatial organization, i.e. chromosome, has gained much interest in genomic studies. A large number of approaches have been available to measure chromosome-wide patterns, such as Hi-C [26], fluorescent repressor operator system (FROS) [21] and chromatin immunoprecipitation sequencing (ChIP-seq) [21], offering a strong opportunity to understand the overall layout of genome organization. Globally, bacterial chromosomes are several orders of magnitude longer than the cells. They fold compactly to fit the volume of cells and form complex structures [22]. For example, the chromosome of *Escherichia coli* (*E. coli*) is folded into a series of consecutive loops, i.e. supercoils [27]. Imaging data indicated that the distribution of supercoils is uneven and variable under different physiological conditions. There are fewer supercoils of *E. coli* during the stationary phase, while there are more of those during the exponential growth phase [28]. Additionally, the highly expressed genes (HEGs) are distributed at the folding-domain boundaries [21]. These findings demonstrate that the locations of genes and chromosome conformations have important impacts on gene expression [29] and copy number [30]. On this basis, an interesting computational model was proposed by Ma *et al.* [31] to infer supercoil structure, and the predicted results showed high consistency with the experimentally verified DNA binding sites of the nucleoid-associated proteins (NAPs) that assist the structure. Furthermore, the folding chromosome was organized into large size domains, such as chromosome interaction domains (CIDs) and macrodomains (MDs) [32,33]. Thirty-one CIDs and four MDs of the *E. coli* genomes have been reported. Nevertheless, the understanding of functions associated with such domains is not sufficient, and more effort is being made for a deep insight into the underlying mechanisms.

Taken together, given the structures of bacterial genomes, one cannot simply understand them as an evolutionary accident. Instead, they are closely related to functional activities. Understanding the principle of the genome organization can contribute to functional insights into diverse bacteria, providing favorable strategies for the full design of a new-to-nature genome layout in synthetic biology. In this review, our focus is on the bacterial organization and current understanding, both experimentally and com putationally, of the essential constraints on the genomes.

Local organization as operons, transcription units and alternative transcription units facilitates gene co-transcription

In 1961, Jacob and Monod showed that one or multiple continuous genes in a bacterial genome that encode proteins with related biological functions were transcribed in a single polycistronic unit, and these genes shared the same regulation signals by common promoters and terminators [13]. Then, the concept ‘operon’ was proposed as a local organizational feature of bacterial genomes [12] (Figure 1). Such a structure facilitates efficient co-transcription [34] and increases the fitness of genes during evolution [22]. Operons have since then been extensively studied as the basic transcriptional and functional units of bacterial organisms, laying a foundation for a holistic viewpoint of genome organization.

Initially, operons were assumed to be conserved across different genomes of a species and not overlapped with each other along the genome [35–37], which served as the basis for the identification of operons. A previous study used the conservation of gene pairs across genomes to identify operons, and its application in *E. coli* demonstrated the effectiveness of the proposed model [36]. Further, Dam *et al.* [38] integrated five available features of genomic sequences (the intergenic distance, the conserved gene neighborhood, the distances between adjacent genes' phylogenetic profiles, the ratio between the lengths of two adjacent genes and the frequencies of specific DNA motifs in the intergenic regions) into a new model, which substantially improved the accuracy of operon prediction. These methods present exciting opportunities for understanding the local structures of various genomes.

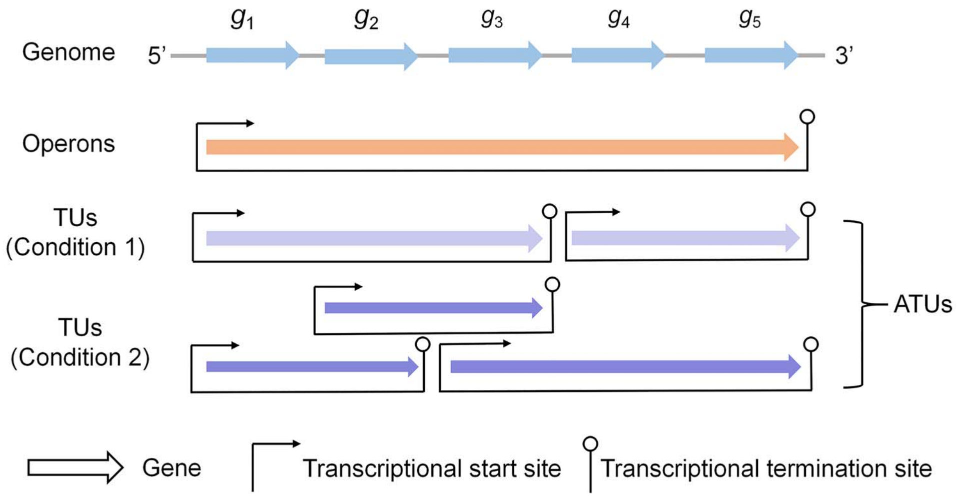

Figure 1. A diagram of operons, TUs and ATUs of the bacterial genome. The genome consists of five genes (blue arrows). These genes form an operon, i.e. (g_1, g_2, g_3, g_4, g_5) (orange arrow). Actually, genes are dynamically co-transcribed with different gene sets, and each gene set is called a TU (all purple arrows under both two conditions). In some cases, such as under condition 1, TUs are nonoverlapping (light purple arrows). However, TUs are overlapped with each other under certain conditions, such as condition 2 (dark purple arrows). Given the different transcription structures under multiple conditions, the TUs under condition 1 and condition 2 are specifically called ATUs.

Subsequently, a large amount of computational predicted and experimentally validated operons have become available in public databases, such as RegulonDB [39], DBTBS [40], OperonDB [41], MicrobesOnline [42], ODB [43], ProOpDB [44] and DOOR [45]. These databases are different from each other in terms of reliability and emphases. Specifically, RegulonDB contains highly reliable operons of *E. coli* from both experiments and computationally prediction, while DBTBS contains experimentally characterized operons of *Bacillus subtilis* (*B. subtilis*) [39,40]. OperonDB and MicrobesOnline have computationally predicted operons for 550 genomes and 620 genomes, respectively [41,42]. Additionally, ODB provides both experimental and predicted operons for 203 genomes [43]. Of particular note, the predicted operons used in the above several databases are with relatively low accuracies (~80%). Several databases were constructed thereafter, including ProOpDB and DOOR [44,45], and provided high-quality predicted operons (with accuracy more than 90%) besides those from experiments. Notably, the predicted operons of more than 2000 genomes are also available at DMINDA 2.0 in support of *cis*-regulatory motif discovery for prokaryote [46]. The information of operons serves as a basis of higher-level genome organization as well as underlying cellular functions [19], such as metabolic pathways and regulation systems [47].

Recently, researchers found that genes within the predicted single operon are not always co-transcribed but dynamically transcribed with different gene sets under different triggering conditions [48,49]. To cope with this discrepancy, the concept 'transcription unit (TU)' was specifically used to refer to the conditionally dependent co-transcribed gene sets (Figure 1). At least three TUs, including (*pdhR*, *aceE*, *aceF*, *lpd*), (*aceE*, *aceF*) and (*ldp*), of the *pdhR-aceEF-lpd* operon in *E. coli* have been found in different growth environments [48]. Additionally, a critical study suggested that TUs may overlap with each other by sharing the common genes, forming TU clusters [50]. It has been demonstrated that the TUs in a cluster are controlled by two different bacterial mechanisms for terminating transcription [51]: TUs that end inside the TU cluster tend to use Rho-dependent terminations, while the terminal TUs (i.e. those end at the last gene of TU cluster) more likely link with Rho-independent terminations [50].

Currently, TUs have been shown as a reliable structure of the bacterial transcriptional profiles that associate with genome organization [49]. Cho *et al*. [52] identified 4661 TUs of *E. coli* based on the organizational elements of TUs, such as transcription start and termination sites. A recent study by Yan *et al*. [49] applied SMRT-Cappable-seq to reidentify the TUs of *E. coli* under multiple growth conditions, which avoids information missing on the large transcriptional contexts and provides a more accurate visualization of genomic transcriptional profiles. Yet little is known about complete TU maps of various bacteria due to the time-consuming, laborious and costly experiment technologies. Fortunately, a variety of computational methods/tools based on available transcriptional data have been developed to advance the deep understanding of TUs, such as DOOR2 [53], BAC-BROWSER [54], SeqTU [55,56] and rSeqTU [57]. Among them, the first three are devoted to capturing the features of transcription intervals based on RNA-seq reads count signals [56–58] for TU identification of bacteria. rSeqTU applies random forest-based feature selection for TU inference [57], and it has improved into a comprehensive pipeline with high performance, from data quality control to TU visualization [57]. Overall, both experimental and computational techniques enhance the understanding of genome arrangement at the transcriptional level. Notably, due to the dynamic nature of gene co-transcription, TUs are of high diversity under different conditions. The different TU structures of a genome are specifically called alternative TUs (ATUs) (Figure 1) [57].

Metabolic pathways constrain operon locations on genomes

While local structures of genome arrangement have been well studied, researchers began to explore the rules that control the global arrangement of operons, not simply to push the limits of their specialties but rather to provide insights into cellular functions (e.g. metabolism and regulatory) at the genome scale [22,59]. By analogy with the operon organization within a regulon

[60,61], it was originally speculated that operons working in the same metabolic pathway would cluster together along the genome. However, this speculation is generally not correct. A seminal study, focusing on both experimental [40,62] and computationally predicted [45] operons of *E. coli* K12 and *B. subtilis* str. 168, showed that more than 40% operons are involved in multiple pathways [19] (the definition of the pathway is based on three popular pathway databases, the SEED database [63], the Kyoto Encyclopedia of Genes and Genomes database [64] and the BioCyc database [65]). The component-sharing property among pathways makes each pathway, especially those encoded by multiple operons, tend to compose of a few clusters rather than one [66]. Inspired by this observation, researchers have put more effort into the studies about the potential mechanisms of global genomic locations.

The great advances of available genomic and transcriptomic data in the last decades have greatly improved the development of computational models. The first influential model was developed by Yin *et al.* [19], and they found that the position of an operon is determined by the relevant pathways in which it is involved. Particularly, the operons involved in pathways of higher activation frequencies (where the activation frequency is assessed based on the available microarray data) tend to form tighter clusters in a genome [19]. This finding motivates the development of the mathematical model, *C* value, to quantify the relationships between metabolic pathways and relative locations of operons in a bacterial genome [19]. In the proposed model, the *C* value of a genome is assessed by the compactness of operons in all pathways ($C = \sum_{i=1}^{N} \sum_{j=1}^{M_i} d_{ij}$ where *N* indicates the number of pathways, M_i indicates the number of operons of the *i*th pathway and d_{ij} indicates the distance between the j^{th} operon and the center operon in the *i*th pathway). We recalculated *C* values of the actual *E. coli* K12 genome and of a large number of reshuffled genomes (permuting the locations of operons in the genome *in silico*) based on the available genome data of *E. coli* K12. The result suggests that the actual genome typically has a smaller *C* value than most reshuffled genomes (Figure 2A), demonstrating that a bacterium keeps operons in individual pathways as compact as possible. Specifically, the operons that participate in multiple pathways will tend to minimize the overall compactness of their participating pathways, while operons involved in single pathway tend to minimize the compactness of the individual pathway.

Despite much advance, the *C* value model neglects the transcriptional differences in pathways when quantifying the compactness of genomes, whereas the transcriptional frequency has been proven with a significant impact on global operon arrangement [19]. With this in mind, another extended model, C^+ value was proposed to infer the locations of operons, where $C^+ = \sum_{i=1}^{N} f_i \times \left(t_i + \alpha \sum_{j=1}^{M} \omega_{ij}\right)$, *N* is the number of pathways, *M* is the number of partitioned segments to be determined, f_i is the transcriptional frequency of the *i*th pathway, t_i is the number of partitioned segments containing operons of the *i*th pathway, ω_{ij} is the ratio of operons involved in the *i*th pathway but not covered by the *j*th partitioned segment and α is a scaling factor. This model not only quantifies the effect of transcriptional frequencies but also assesses multiple clusters in each pathway (based on the hypothesis that a bacterial genome is partitioned to a set of genomic segments) [66]. By minimizing C^+ value of a target genome, the operon locations and optimal partitions of a genome can be determined simultaneously with dynamic programming procedures [66]. Similar to the application of the *C* value model, C^+ value was recalculated for both actual and reshuffled genomes of *E. coli* K12 to evaluate the performance of the model (both genome and transcriptome data were used here). Not surprisingly, the actual genome arrangement tends to minimize C^+ value in comparison with those of alternatives (Figure 2B). The result showed that, compared with the previous *C* value model, the actual genome stands out more significantly from those artificially generated [66], suggesting a better performance of the improved C^+ value model. The study substantially demonstrates that the pathway activation frequencies impose strong constraints on operon locations along the genome, contributing to a deep understanding of global arrangement of operons.

Note that both the *C* value and the C^+ value are computational models based on the hypothesis from a large amount of data, and there is a lack of biological interpretation on the results. For example, what is the biological meaning of the predicted partitioned segments in the C^+ value model? Fortunately, the subsequent study has proven that the predicted segments are consistent with the supercoiled structure of the folded chromosome (next section). Nevertheless, more work associated with biological explanations is required for a comprehensive viewpoint of genome organization.

Gene expression and replication affect chromosome structures

In bacteria, the double-strand DNA with a helical structure usually keeps a twisting state and folds compactly within cells [67,68], because if fully stretched out, the chromosome is approximately 1000 times longer than the length of the cell. Researchers have found that the folded conformations of chromosomes are highly variable under different conditions [69] (Figure 3A). More intriguingly, the varying conformations are not random but formed in a way that is compatible with multiple cellular processes [69]. Here, we discuss the functional factors (gene expression and DNA replication) that affect chromosome structures at different scales (supercoils, CIDs and MDs).

Supercoils

Generally, the helix DNA strands undergo underwound or overwound conditions, resulting in turns of the entire DNA molecule and forming a tertiary structure, called supercoil [70]. Due to the distinct directions of twisting, supercoils are classified as positive supercoils and negative supercoils (twisting in right-handed and left-handed directions, respectively). Bacterial DNA strands are often found to be underwound and introduce more negative supercoils with the help of various topoisomerases [71,72]. It has been demonstrated that the variability of negative supercoils (in both number and structure [70]) is closely related to the implementation of cell functions and can contribute to localizing conformational changes (e.g. DNA breakage) [73]. For example, when the bacteria immunity system (clustered regularly interspaced short palindromic repeats associated system) is stimulated [74–80], the number of negative supercoils will increase to accelerate the R-loop formation between guide RNA and target DNA, thereby preventing foreign DNA transcription [81,82]. Additionally, the density of negative supercoils is associated with varying levels of gene expression and metabolism processes. Studies have found that negatively supercoiled genes are generally more efficiently transcribed than compact positively supercoiled genes [83,84]. Therefore, bacterial chromosomes in the exponential growth stage are highly negatively supercoiled,

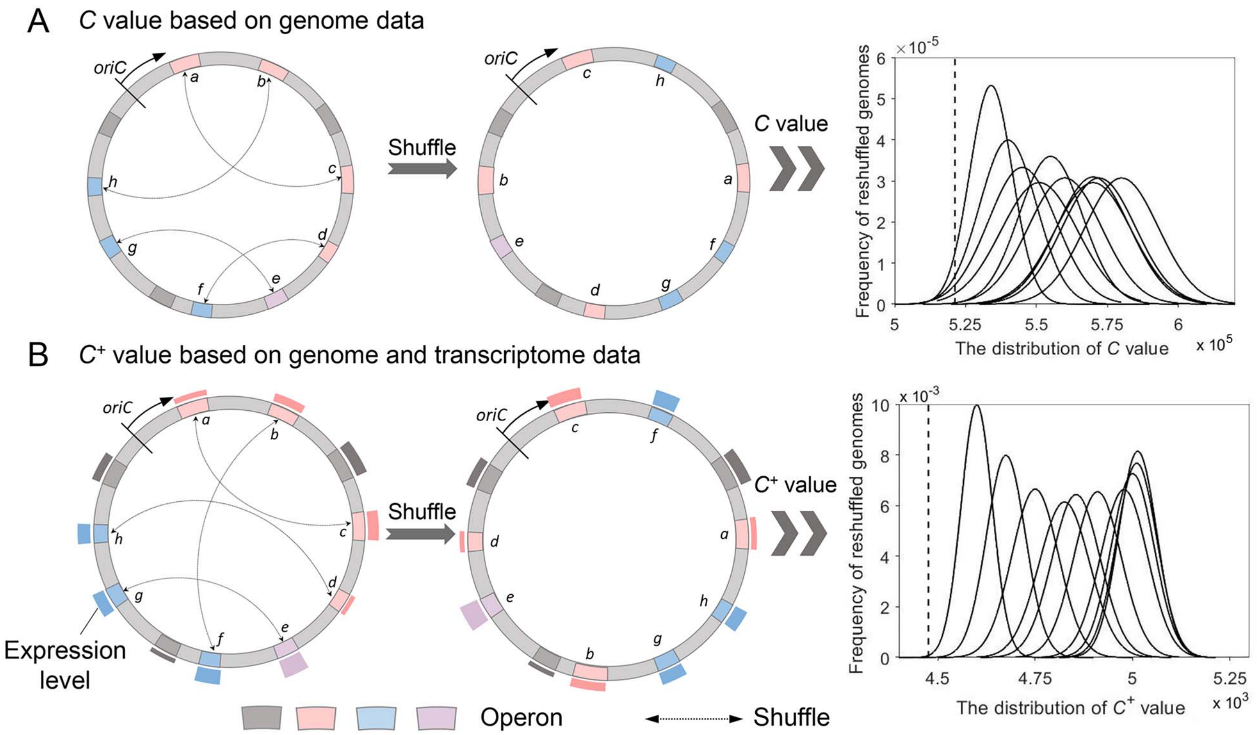
A C value based on genome data
oriC
a
b
c
d
e
f
g
h
Shuffle
C value
Frequency of reshuffled genomes
×10^-5
The distribution of C value
x 10^5
5
5.25
5.5
5.75
6
0
1
2
3
4
5
6
B C+ value based on genome and transcriptome data
C+ value
x 10^-3
The distribution of C+ value
x 10^3
4.5
4.75
5
5.25
0
2
4
6
8
10
Expression level
Operon
Shuffle

Figure 2. The calculation of C value and C^+ value for the actual genome and reshuffled genomes. The blocks (pink, blue, gray and purple) represent operons. The two pathways that pink and blue operons participate in (the same color means those are involved in the same pathway) are used here to show the calculation of C and C^+ value. Operon *e* participates in both these two pathways, and gray operons represent other operons on the genome. The bars outside the circle of B indicate the expression level of genes (operons), which can be assessed from transcriptomic data. *oriC* represents the origin of chromosomal replication. (**A**) The C value calculation of *E. coli* K12. (**B**) The C^+ value calculation of *E. coli* K12. In each coordinate system, the *x*-axis represents the C or C^+ value, and the *y*-axis represents the frequency of reshuffled genomes for a specific C or C^+ value. We use 1 million permutations by randomly shuffling X% of operons ($X = 10, 20, \ldots, 100$, respectively) and calculate the distribution of C or C^+ value. In this way, 10 curves are generated corresponding to the specific X value from left to right. The dashed line represents the calculated C or C^+ value of the actual *E. coli* K12 genome.

which has more looped domains compared to those in the stationary stage [83,84]. Notably, the density of negative and positive supercoils is changeable. However, overall supercoiling levels would be controlled to ensure primary bacterial life via topoisomerases due to the impact of supercoils on replication processes (accumulation of negative supercoils lead to the catenation of newly synthesized DNA strands and hinder chromosome segregation, while excess positive supercoils block the replisome progression) [85].

Based on such general understanding, more attention is given to the specific folded conformations of bacterial chromosomes. Considering the experimental data of chromosomal conformations are limited, several computational models are carried out aiming at a deep understanding of genome organization [86,87]. In view of the high correlation of chromosome conformations (the structure of supercoils) with bacterial transcription, Ma *et al.* inferred that the supercoiled structure would form in a way that facilitates the realization of transcription (i.e. keep energy consumption minimum for the folding and unfolding of supercoils during transcription) [31,88]. Based on this idea, a function called E value ($E = \sum_{i=1}^{N} \omega_i s_i$, where s_i is the number of supercoils containing the operons of the i^{th} pathway and ω_i is the transcriptional frequencies of the *i*th pathway) was developed to measure the energy consumption of transcription [31]. In this function, the energy is quantified proportional to the number of supercoils containing genes/operons of the pathway multiplied by its transcription frequency. It is noteworthy that the transcription frequency can be assessed by gene expression data [19], and a pathway is considered as activated if and only if at least a certain number of its operons/genes are activated [31]. Ten *E. coli* strains were used to validate the usefulness of the function [89,90]. Results showed that the inferred dynamic structures of supercoil folding domains are consistent with the experimental data that assist the folding of chromosomal domains [31]. Additionally, the analysis suggested that E value is highly correlated to the cell growth rate (with correlation coefficient −0.8) [89,90], which is intuitively reliable with the fact that less consumption accompanies faster transcription.

Chromosome interaction domains

The progress in experimental technologies, such as 3C [91], 4C [92], 5C [91], Hi-C [91] and ChIP-seq [21], has greatly improved the understanding of the conformations of bacterial chromosomes, in which 4C, 5C and Hi-C are advanced versions of 3C to capture chromosome structures [91], and ChIP-seq contributes to detecting focused protein binding on the wide genome [21] (Box 1). Based on these technologies, it has been found that the bacteria DNA conformation relies heavily on various NAPs [93], which have various structural effects on DNA. For example, the histone-like nucleoid-structuring proteins form filaments [94,95] and bridges between two DNA segments [21,96], while structural maintenance of chromosomes proteins extrudes DNA loops [97–103]. Additionally, the inversion stimulation proteins collaborated with heat-stable proteins which help DNA bending

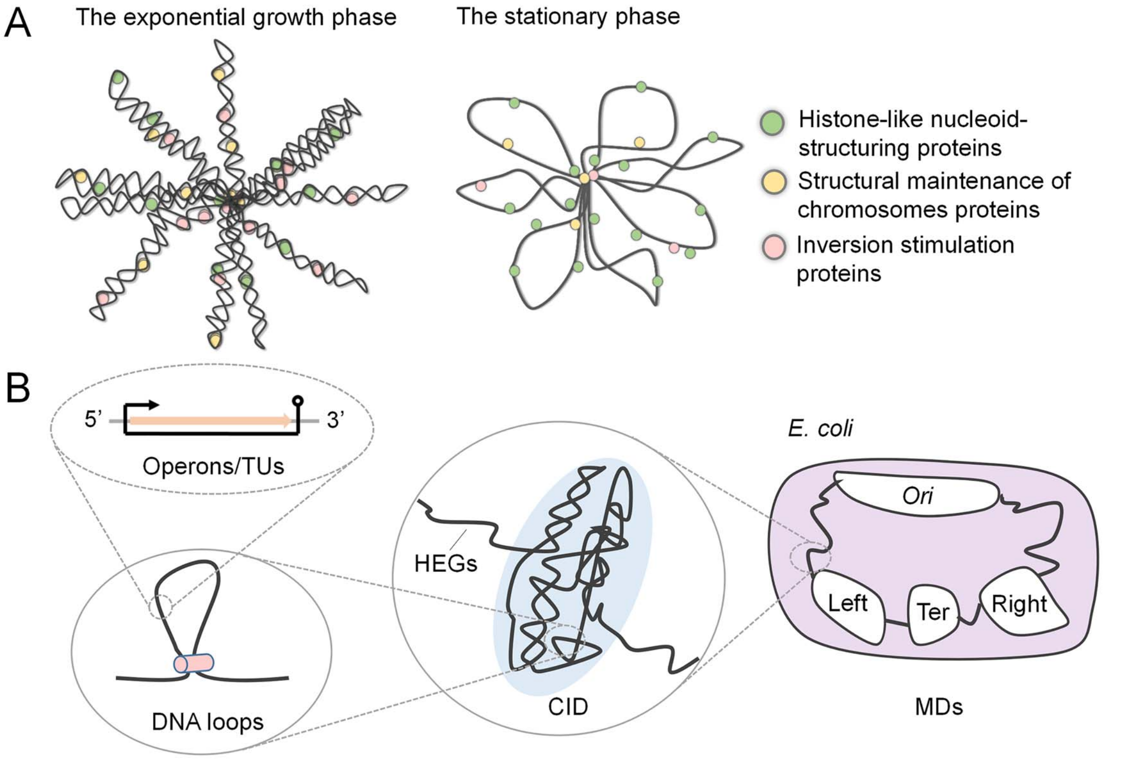
A
The exponential growth phase
The stationary phase
● Histone-like nucleoid-structuring proteins
● Structural maintenance of chromosomes proteins
● Inversion stimulation proteins
B
5'
3'
Operons/TUs
DNA loops
HEGs
CID
E. coli
Ori
Left
Ter
Right
MDs

Figure 3. The folded chromosome in cells. The bacterial chromosomes are compactly organized into complex structures. (**A**) The chromosomal structures of *E. coli* are different in the exponential growth phase and the stationary phase. (**B**) The bacterial chromosome is organized at different scales. The operons/TUs along the genome will form DNA loops and large chromosome substructures (i.e. CIDs). Further, independently organized domains (MDs) can be observed.

and twisting [104,105]. Through Hi-C contact maps, researchers observed that the small DNA loops by NAPs form larger chromosome substructures along the genomes (i.e. CIDs with sizes ranging from 30 to ∼500kbp) [106,107] (Figure 3B). These CIDs are highly self-interacting genomic regions and relatively insulated from others [21,108]. Current studies have suggested that HEGs play a direct role in defining CID boundaries [109,110]. The *E. coli* chromosome is organized into 31 CIDs, in which 22 are consistent with locations of HEGs [111]. Moreover, HEGs are clustered to varying degrees under different conditions, contributing to the CID boundaries with distinct characteristics, such as sharpening boundaries while undergoing exponential growth and diminishing boundaries in the starvation state [67]. This phenomenon indicates that transcription is closely associated with genome organization at a higher level.

Box 1 Experimental technologies used to study chromosome structures

3C [91]: 3C is a technology to capture the interaction between a paired genomic loci. After a series of restricted digesting, religating, reverse cross-linking and polymerase chain reaction (PCR), an interaction between two gene loci exists when ligation products exist. It is a 'one-to-one' approach and has enormous difficulties in designing quantified specified ligation products.

4C [92]: 4C is a technology to quantify the interactions between one locus and all other genomic loci. Different from 3C and 5C (see below), it does not need the prior knowledge of interacting chromosomal regions, and it involves a second ligation step to create self-circularized DNA fragments for inverse PCR. Inverse PCR allows the known sequence to be used to amplify the unknown sequence ligated to it.

5C [91]: 5C is another improvement of 3C to capture interactions between all fragments within a given region. It is a 'many-to-many' approach and can be performed by ligating universal primers to all fragments.

Hi-C [91]: Hi-C represents the technology to analyze the structure of genomes. The most distinguishing feature is that Hi-C is an 'all-to-all' approach. It can offer a complete spatial view of the DNA information. Even genomic elements far away from each other, like HEGs and regulatory factors, can find the interactions among them thanks to the huge library.

ChIP-seq [21]: ChIP-seq is a widely used technology and is convenient in identifying the binding sequence combined with PCR. It is a technology that combines Hi-C with chromatin immunoprecipitation by using a specific antibody. It is a 'many-to-many' approach. The advantage is the capability of identifying a wide range of interactions even if it is rare, while the disadvantage is the difficulties in interpreting data.

FROS [21]: FROS is a single-cell technology offering the dynamics of chromosomes. Specific loci can be tagged and tracked in live cells to define the spatial position and interrelationship of positions, providing the dynamics of interesting loci. However, it is limited in its throughput.

Macrodomains

More advances in genome technologies, such as FROS [21] (Box 1), enable researchers to determine the positions of genomic loci and track the dynamics of specific loci, improving the understanding of genome organization at the macro scale. By visualizing the locations of up to 100 genomic loci, the origin and termination sites of bacterial replication were revealed not located randomly but in specific positions [112]. This fundamental feature of chromatin organization has been found associated with the macroscale domains of bacterial chromosomes [33]. Specifically, *E. coli* has the well-defined structure with four MDs (Ori MD, Ter MD, Right and Left MDs) (Figure 3B) and two nonstructured (NS) regions based on 3C and DNA–DNA interaction studies [113,114], in which Ori MD contains the origin site of DNA replication [115], while Ter MD contains the terminators of the replication process. Further studies suggested that the positions of loci on the chromosome are not stable. The loci in NS regions are under fewer constraints than loci in MDs [114,115]. Loci in Ter MD are the least mobile, while mobility increases along the chromosomal arms towards Ori MD [116]. Nevertheless, the mechanisms of the positions of genomic loci and chromosome separate processes are far from clear. Although several proteins related to the organization of MDs, such as the gene product named MaoP and MatP, have been studied [117], more attention is required to explore the structural relations with function.

Collectively, the organization of chromosomes is not random but associated with critical cellular functions, from supercoils to CIDs and MDs. Further studies are expected to deepen the understanding of how cellular processes shape the chromosomes, which will, in turn, help reveal the mechanisms of complex cellular processes.

DNA replication, transcription and gene products direct gene strand bias

In addition to the organization of operons and chromosomal conformation, another significant pattern within bacterial genomes is the preferential location on the leading strand. It has been observed that the majority of bacterial genes, with percentages ranging from 50% to 90% across different species, are preferably located on the leading strand in a genome [23,24]. This observation is well known as strand-biased gene distribution (SGD) [118]. Currently, a large number of studies have been performed, aiming at a mechanistic explanation of this observation. A key finding revealed in these studies is that the collisions between replication and transcription result in SGD. Since replication and transcription occur simultaneously based on the same DNA template but proceed at different rates, collisions are inevitable between them (head-on collisions occur if the gene is in the lagging strand, while co-oriented collisions occur if the gene is in the leading strand) [119]. However, the collisions are deleterious, such as leading to abortive transcription or stalling replication [118]. Recent experimental results suggested that head-on collisions are more deleterious than co-oriented collisions, making the genes more prone to exert negative mutations when performing specific functions [118]. Therefore, the genes with essential functions (essentiality-driven hypotheses [119]) or high expression (expression-driven hypotheses [120]) are preferably located on leading strands [22].

Despite these insights, neither essentiality-driven hypotheses nor expression-driven hypotheses can explain SGD thoroughly. By removing highly expressed genes and essential genes from multiple bacterial species, the phenomena of SGD were decreased in varying degrees of different bacteria but not completely removed [118]. For a more comprehensive understanding of SGD, Mao *et al.* [121] explored other factors that may affect the strand bias in a bacterium. They found that genes with specific functions prefer different strands and consequently cause SGD, such as genes of ribosome prefer to locate on the leading strands, while genes of transcription factor prefer the lagging strands [121]. Additionally, Gao *et al.* [118] proposed the energy efficiency model, and they argued that the locations of genes tend to under strong selection for efficient energy usage. Since lagging-strand genes encode more expensive protein products (the proteins whose synthesis requires cost more energy) than leading-strand ones, genes prefer to located on the leading strand [118], especially those highly expressed. Nevertheless, the understanding of SGD is not comprehensive yet. If more sophisticated analyses, including the living environments and the survivability requirements of bacteria, could be considered, the quantitative models that integrated more parameters will be established, contributing to a comprehensive insight into the factors that drive SGD [121].

Conclusions and discussions

Recent advances in genomic studies have revealed the bacterial genomes are highly organized with functional implications. Here, we overviewed the well-studied organization during bacterial evolution, including operons, TUs, supercoils, CIDs, MDs and SGD. Meanwhile, there are several other structures, such as overlapping genes, playing important roles in improving genome compaction and being associated with translational coupling [122], which will not be extended into more details here.

The summarized knowledge in this manuscript will contribute to the studies of functional activities of bacteria [123]. For example, there are some missing enzymes to catalyze reactions in a pathway (i.e. pathway holes) [124], and we can apply the understanding of global gene arrangement to fill the holes. By identifying all the metabolic pathways that involve an unassigned enzyme based on its EC number and its substrates and products, we can predict the supercoils containing operons for each of the involved pathways. In this way, the candidate gene would lie in a supercoil covered by all these pathways, i.e. within a narrow genomic range, for researchers to detect.

Additionally, the understanding of genome organization improves the genome assembly process, which in turn benefits the genomic studies. Advances in next-generation sequencing offer a large number of raw datasets consisting of short reads, which make genome assembly necessary [125]. Most current approaches depend heavily on phylogenetic relationships among species to select close-related reference genome for the separate contigs ordering during assembly [126–128]. However, due to the intensive rearrangement during bacterial evolution, the same genome organization may present in remotely related strains rather than closely related strains [11,129], which would lead to systemic errors when assembly contigs based on the closely related reference. Fortunately, focusing on the genome organization, the high-order conservation of core genes along the genome is generalized [130]. Therefore, one can envision that ordering the contigs, by extracting the order-conserved genes from known complete genomes [131], provides a reliable draft genome for a full understanding of bacteria at both taxonomic and functional levels.

More importantly, the study of bacterial genomes allows the understanding of diseases in a new perspective. The explosion of meta-omics data of human microbiomes provides an unprecedented opportunity to study the relations between microbes and diseases in species levels [132,133]. Microbes have been demonstrated with significant associations with diseases [134]. Based on the well-studied genome organization, we can explore markers of diseases that go beyond species classification, such as the operon/TU structure [135,136]. These information will provide deep insights into the occurrence of specific diseases, such as tuberculosis and gastric cancer [135,136].

Notably, current studies remain to be improved for a system-level understanding of bacterial genomes. Firstly, the findings based solely on data from informatics analyses or predictions can be misleading. Extensive experimental validation of genome structures, from operons to chromosomes, will be significant to iteratively construct, test and improve models for understanding genome organization. Secondly, the computational model should be developed to keep up with advanced experimental data, such as the increasing data from chromosome capture technologies. One can apply these data to study the mechanisms of genome organization and validate the effectiveness of proposed models. We expect the field of the genomic organization of bacteria, more broadly of all microbes, would be active and exciting in the coming years.

Key Points

- Bacterial genomes interact intimately with multiple cellular processes. These interactions impose constraints on genome organization.
- The genome organization of bacteria, such as the local structure as operons, transcription units and alternative transcription units, and global chromosome architecture as supercoils, chromosome interaction domains and macrodomains, has close associations with functional activities (e.g. metabolism, transcription and replication). Additionally, gene strand bias is another significant pattern within bacterial genomes, which can be explained by multiple cellular processes.
- Bacterial genome organization contributes to the studies of functional activities of bacteria. One can apply the insights of genome organization to multiple contexts, such as pathway holes, genome assembly and the mechanistic understanding of human diseases.
- Despite much progress, current studies of bacterial genome organization remain to be improved from both experimental and computational perspectives.

Authors' contributions

Q.M. and B.L. conceived the project. Z.L., J.F. and B.Y. collected materials. J.F. drafted the chromosome and gene strand bias sections, as well as data analysis of C value and C^+ value. Z.L. drafted other sections and prepared all figures. All the authors revised the manuscript. The authors declare that they have no competing interests.

Acknowledgments

The authors would like to thank Xianquan Wei and Yu Gan from Shandong University for their effort in collecting literature. The authors want to thank Anjun Ma from the Ohio State University for his valuable suggestions on figures.

Funding

This work was supported by the National Nature Science Foundation of China (61772313, 11931008), Interdisciplinary Science Innovation Group Project of Shandong University (2019) and the Innovation Method Fund of China (Ministry of Science and Technology of China) (2018IM020200).

Conflict of interest

None.

References

1. Chater KF. Streptomyces inside-out: a new perspective on the bacteria that provide us with antibiotics. *Philos Trans R Soc B Biol Sci* 2006;**361**:761–8.
2. Bier RL, Voss KA, Bernhardt ES. Bacterial community responses to a gradient of alkaline mountaintop mine drainage in central Appalachian streams. *ISME J* 2015; **9**:1378–90.
3. Gao C, Fernandez VI, Lee KS, *et al*. Single-cell bacterial transcription measurements reveal the importance of dimethylsulfoniopropionate (DMSP) hotspots in ocean sulfur cycling. *Nat Commun* 2020;**11**:1–11.
4. Brown MG, Balkwill DL. Antibiotic resistance in bacteria isolated from the deep terrestrial subsurface. *Microb Ecol* 2009;**57**:484.
5. Fetissov SO. Role of the gut microbiota in host appetite control: bacterial growth to animal feeding behaviour. *Nat Rev Endocrinol* 2017;**13**:11.
6. Ochman H, Caro-Quintero. *Genome Size and Structure: Bacterial*. Oxford: Academic Press, 2016.
7. Bobay L-M, Ochman H. The evolution of bacterial genome architecture. *Front Genet* 2017;**8**:72.
8. Huvet M, Stumpf MPH. Overlapping genes: a window on gene evolvability. *BMC Genomics* 2014;**15**:721.
9. Blair JMA, Webber MA, Baylay AJ, *et al*. Molecular mechanisms of antibiotic resistance. *Nat Rev Microbiol* 2015;**13**:42–51.
10. Spellberg B, Gilbert DN. The future of antibiotics and resistance: a tribute to a career of leadership by John Bartlett. *Clin Infect Dis* 2014;**59**:S71–5.
11. Darmon E, Leach DRF. Bacterial genome instability. *Microbiol Mol Biol Rev* 2014;**78**:1–39.
12. Jacob F, Perrin D, Sánchez C, *et al*. Operon: a group of genes with the expression coordinated by an operator. *C R Hebd Seances Acad Sci* 1960;**250**:1727–9.
13. Jacob F, Monod J. Genetic regulatory mechanisms in the synthesis of proteins. *J Mol Biol* 1961;**3**:318–56.
14. Brocken DJW, Tark-Dame M, Dame RT. The organization of bacterial genomes: towards understanding the interplay between structure and function. *Curr Opin Syst Biol* 2018;**8**:137–43.
15. Land M, Hauser L, Jun S-R, *et al*. Insights from 20 years of bacterial genome sequencing. *Funct Integr Genomics* 2015;**15**:141–61.
16. Fraser CM, Gocayne JD, White O, *et al*. The minimal gene complement of *Mycoplasma genitalium*. *Science* 1995;**270**:397–404.
17. Fleischmann RD, Adams MD, White O, *et al*. Whole-genome random sequencing and assembly of *Haemophilus influenzae* Rd. *Science* 1995;**269**:496–512.

18. Lawrence JG, Roth JR. Selfish operons: horizontal transfer may drive the evolution of gene clusters. *Genetics* 1996;**143**:1843–60.
19. Yin Y, Zhang H, Olman V, *et al.* Genomic arrangement of bacterial operons is constrained by biological pathways encoded in the genome. *Proc Natl Acad Sci USA* 2010;**107**:6310–5.
20. Cavalli G, Misteli T. Functional implications of genome topology. *Nat Struct Mol Biol* 2013;**20**:290–9.
21. Dame RT, Rashid F-ZM, Grainger DC. Chromosome organization in bacteria: mechanistic insights into genome structure and function. *Nat Rev Genet* 2019;**12**:1–16.
22. Rocha EPC. The organization of the bacterial genome. *Annu Rev Genet* 2008;**42**:211–33.
23. Koonin E. Evolution of genome architecture. *Int J Biochem Cell Biol* 2009;**41**:298–306.
24. Zivanovic Y, Lopez P, Philippe H, *et al.* Pyrococcus genome comparison evidences chromosome shuffling-driven evolution. *Nucleic Acids Res* 2002;**30**:1902–10.
25. French S. Consequences of replication fork movement through transcription units in vivo. *Science* 1992;**258**:1362–5.
26. Van Berkum NL, Lieberman-Aiden E, Williams L, *et al.* Hi-C: a method to study the three-dimensional architecture of genomes. *JoVE* 2010;**39**:e1869.
27. Dillon SC, Dorman CJ. Bacterial nucleoid-associated proteins, nucleoid structure and gene expression. *Nat Rev Microbiol* 2010;**8**:185–95.
28. Stein RA, Deng S, Higgins NP. Measuring chromosome dynamics on different time scales using resolvases with varying half-lives. *Mol Microbiol* 2005;**56**:1049–61.
29. Bryant JA, Sellars LE, Busby SJW, *et al.* Chromosome position effects on gene expression in *Escherichia coli* K-12. *Nucleic Acids Res* 2014;**42**:11383–92.
30. Dryselius R, Izutsu K, Honda T, *et al.* Differential replication dynamics for large and small vibrio chromosomes affect gene dosage, expression and location. *BMC Genomics* 2008;**9**:559.
31. Ma Q, Yin Y, Schell MA, *et al.* Computational analyses of transcriptomic data reveal the dynamic organization of the *Escherichia coli* chromosome under different conditions. *Nucleic Acids Res* 2013;**41**:5594–603.
32. Niki H, Yamaichi Y, Hiraga S. Dynamic organization of chromosomal DNA in *Escherichia coli*. *Genes Dev* 2000;**14**:212–23.
33. Dame RT, Kalmykowa OJ, Grainger DC. Chromosomal macrodomains and associated proteins: implications for DNA organization and replication in gram negative bacteria. *PLoS Genet* 2011;**7**:e1002123.
34. Zheng Y, Szustakowski JD, Fortnow L, *et al.* Computational identification of operons in microbial genomes. *Genome Res* 2002;**12**:1221–30.
35. Craven M, Page D, Shavlik JW, *et al.* A probabilistic learning approach to whole-genome operon prediction. *ISMB* 2000;**8**:116–27.
36. Ermolaeva MD. Prediction of operons in microbial genomes. *Nucleic Acids Res* 2001;**29**:1216–21.
37. Salgado H, Moreno-Hagelsieb G, Smith TF, *et al.* Operons in Escherichia coli: genomic analyses and predictions. *Proc Natl Acad Sci USA* 2000;**97**:6652–7.
38. Dam P, Olman V, Harris K, *et al.* Operon prediction using both genome-specific and general genomic information. *Nucleic Acids Res* 2007;**35**:288–98.
39. Salgado H, Peralta-Gil M, Gama-Castro S, *et al.* RegulonDB v8. 0: omics data sets, evolutionary conservation, regulatory phrases, cross-validated gold standards and more. *Nucleic Acids Res* 2012;**41**:D203–13.
40. Sierro N, Makita Y, de Hoon M, *et al.* DBTBS: a database of transcriptional regulation in *Bacillus subtilis* containing upstream intergenic conservation information. *Nucleic Acids Res* 2007;**36**:D93–6.
41. Pertea M, Ayanbule K, Smedinghoff M, *et al.* OperonDB: a comprehensive database of predicted operons in microbial genomes. *Nucleic Acids Res* 2008;**37**:D479–82.
42. Dehal PS, Joachimiak MP, Price MN, *et al.* MicrobesOnline: an integrated portal for comparative and functional genomics. *Nucleic Acids Res* 2010;**38**:D396–400.
43. Okuda S, Katayama T, Kawashima S, *et al.* ODB: a database of operons accumulating known operons across multiple genomes. *Nucleic Acids Res* 2006;**34**:D358–62.
44. Taboada B, Ciria R, Martinez-Guerrero CE, *et al.* ProOpDB: prokaryotic operon data base. *Nucleic Acids Res* 2011;**40**:D627–31.
45. Mao F, Dam P, Chou J, *et al.* DOOR: a database for prokaryotic operons. *Nucleic Acids Res* 2008;**37**:D459–63.
46. Yang J, Chen X, McDermaid A, *et al.* DMINDA 2.0: integrated and systematic views of regulatory DNA motif identification and analyses. *Bioinformatics* 2017;**33**:2586–8.
47. Cao H, Ma Q, Chen X, *et al.* DOOR: a prokaryotic operon database for genome analyses and functional inference. *Brief Bioinform* 2019;**20**:1568–77.
48. Quail MA, Haydon DJ, Guest JR. The pdhR-aceEF-lpd operon of *Escherichia coli* expresses the pyruvate dehydrogenase complex. *Mol Microbiol* 1994;**12**:95–104.
49. Yan B, Boitano M, Clark TA, *et al.* SMRT-cappable-seq reveals complex operon variants in bacteria. *Nat Commun* 2018;**9**:3676.
50. Mao X, Ma Q, Liu B, *et al.* Revisiting operons: an analysis of the landscape of transcriptional units in *E. coli*. *BMC Bioinformatics* 2015;**16**:1–9.
51. Chen J, Morita T, Gottesman S. Regulation of transcription termination of small RNAs and by small RNAs: molecular mechanisms and biological functions. *Front Cell Infect Microbiol* 2019;**9**:201.
52. Cho B-K, Zengler K, Qiu Y, *et al.* The transcription unit architecture of the *Escherichia coli* genome. *Nat Biotechnol* 2009;**27**:1043.
53. Mao X, Ma Q, Zhou C, *et al.* DOOR 2.0: presenting operons and their functions through dynamic and integrated views. *Nucleic Acids Res* 2013;**42**:D654–9.
54. Garanina IA, Fisunov GY, Govorun VM. BAC-BROWSER: the tool for visualization and analysis of prokaryotic genomes. *Front Microbiol* 2018;**9**:2827.
55. Chou W-C, Ma Q, Yang S, *et al.* Analysis of strand-specific RNA-seq data using machine learning reveals the structures of transcription units in *Clostridium thermocellum*. *Nucleic Acids Res* 2015;**43**:e67–7.
56. Chen X, Chou WC, Ma Q, *et al.* SeqTU: a web server for identification of bacterial transcription units. *Sci Rep* 2017;**7**:1–6.
57. Niu S-Y, Liu B, Ma Q, *et al.* rSeqTU—a machine-learning based R package for prediction of bacterial transcription units. *Front Genet* 2019;**10**:374.
58. Mazin PV, Fisunov GY, Gorbachev AY, *et al.* Transcriptome analysis reveals novel regulatory mechanisms in a genome-reduced bacterium. *Nucleic Acids Res* 2014;**42**:13254–68.
59. Képès F, Jester BC, Lepage T, *et al.* The layout of a bacterial genome. *FEBS Lett* 2012;**586**:2043–8.

60. Liu B, Zhou C, Li G, *et al*. Bacterial regulon modeling and prediction based on systematic cis regulatory motif analyses. *Sci Rep* 2016;**6**:23030.
61. Zhang H, Yin Y, Olman V, *et al*. Genomic arrangement of regulons in bacterial genomes. *PLoS One* 2012;**7**:e29496.
62. Salgado H, Gama-Castro S, Peralta-Gil M, *et al*. RegulonDB (version 5.0): *Escherichia coli* K-12 transcriptional regulatory network, operon organization, and growth conditions. *Nucleic Acids Res* 2006;**34**:D394–7.
63. Overbeek R, Begley T, Butler RM, *et al*. The subsystems approach to genome annotation and its use in the project to annotate 1000 genomes. *Nucleic Acids Res* 2005;**33**: 5691–702.
64. Kanehisa M, Araki M, Goto S, *et al*. KEGG for linking genomes to life and the environment. *Nucleic Acids Res* 2007;**36**: D480–4.
65. Karp PD, Ouzounis CA, Moore-Kochlacs C, *et al*. Expansion of the BioCyc collection of pathway/genome databases to 160 genomes. *Nucleic Acids Res* 2005;**33**:6083–9.
66. Ma Q, Xu Y. Global genomic arrangement of bacterial genes is closely tied with the total transcriptional efficiency. *Genomics Proteomics Bioinformatics* 2013;**11**:66–71.
67. Le TBK, Imakaev MV, Mirny LA, *et al*. High-resolution mapping of the spatial organization of a bacterial chromosome. *Science* 2013;**342**:731–4.
68. Messerschmidt SJ, Waldminghaus T. Dynamic organization: chromosome domains in *Escherichia coli*. *J Mol Microbiol Biotechnol* 2014;**24**:301–15.
69. Toro E, Shapiro L. Bacterial chromosome organization and segregation. *Cold Spring Harb Perspect Biol* 2010;**2**:a000349.
70. Dorman CJ. DNA supercoiling and transcription in bacteria: a two-way street. *BMC Mol cell Biol* 2019; **20**:26.
71. Brown P, Cozzarelli N. A sign inversion mechanism for enzymatic supercoiling of DNA. *Science* 1979;**206**:1081–3.
72. Vinograd J, Lebowitz J, Radloff R, *et al*. The twisted circular form of polyoma viral DNA. *Proc Natl Acad Sci USA* 1965;**53**:1104.
73. Dorman CJ. Genome architecture and global gene regulation in bacteria: making progress towards a unified model? *Nat Rev Microbiol* 2013;**11**:349–55.
74. Hille F, Richter H, Wong SP, *et al*. The biology of CRISPR-Cas: backward and forward. *Cell* 2018;**172**:1239–59.
75. Garneau JE, Dupuis M-È, Villion M, *et al*. The CRISPR/Cas bacterial immune system cleaves bacteriophage and plasmid DNA. *Nature* 2010;**468**:67–71.
76. Edgar R, Qimron U. The *Escherichia coli* CRISPR system protects from λ lysogenization, lysogens, and prophage induction. *J Bacteriol* 2010;**192**:6291–4.
77. Bikard D, Hatoum-Aslan A, Mucida D, *et al*. CRISPR interference can prevent natural transformation and virulence acquisition during in vivo bacterial infection. *Cell Host Microbe* 2012;**12**:177–86.
78. Levy A, Goren MG, Yosef I, *et al*. CRISPR adaptation biases explain preference for acquisition of foreign DNA. *Nature* 2015;**520**:505–10.
79. Stolz R, Sulthana S, Hartono SR, *et al*. Interplay between DNA sequence and negative superhelicity drives R-loop structures. *Proc Natl Acad Sci USA* 2019;**116**: 6260–9.
80. Westra ER, van Erp PBG, Künne T, *et al*. CRISPR immunity relies on the consecutive binding and degradation of negatively supercoiled invader DNA by Cascade and Cas3. *Mol Cell* 2012;**46**:595–605.
81. Leng F, Amado L, Mcmacken R. Coupling DNA supercoiling to transcription in defined protein systems. *J Biol Chem* 2004;**279**:47564–71.
82. Drolet M, Bi X, Liu LF. Hypernegative supercoiling of the DNA template during transcription elongation in vitro. *J Biol Chem* 1994;**269**:2068–74.
83. Westra ER, Nilges B, van Erp PBG, *et al*. Cascade-mediated binding and bending of negatively supercoiled DNA. *RNA Biol* 2012;**9**:1134–8.
84. Bordes P, Conter A, Morales V, *et al*. DNA supercoiling contributes to disconnect σS accumulation from σS-dependent transcription in *Escherichia coli*. *Mol Microbiol* 2003;**48**:561–71.
85. Keszthelyi A, Minchell NE, Baxter J. The causes and consequences of topological stress during DNA replication. *Genes (Basel)* 2016;**7**:134.
86. Fritsche M, Li S, Heermann DW, *et al*. A model for *Escherichia coli* chromosome packaging supports transcription factor-induced DNA domain formation. *Nucleic Acids Res* 2012;**40**:972–80.
87. Buenemann M, Lenz P. Geometrical ordering of DNA in bacteria. *Commun Integr Biol* 2011;**4**:291–3.
88. Szczelkun MD, Tikhomirova MS, Sinkunas T, *et al*. Direct observation of R-loop formation by single RNA-guided Cas9 and Cascade effector complexes. *Proc Natl Acad Sci USA* 2014;**111**:9798–803.
89. Rajaraman E, Agarwal A, Crigler J, *et al*. Transcriptional analysis and adaptive evolution of *Escherichia coli* strains growing on acetate. *Appl Microbiol Biotechnol* 2016;**100**:7777–85.
90. Ma Q, Chen X, Liu C, *et al*. Understanding the commonalities and differences in genomic organizations across closely related bacteria from an energy perspective. *Sci China Life Sci* 2014;**57**:1121–30.
91. Sati S, Cavalli G. Chromosome organization in bacteria: mechanistic understanding genome function. *Chromosoma* 2017;**126**:33–44.
92. Zhao Z, Tavoosidana G, Sjölinder M, *et al*. Circular chromosome conformation capture (4C) uncovers extensive networks of epigenetically regulated intra-and interchromosomal interactions. *Nat Genet* 2006;**38**:1341–7.
93. Santos-Zavaleta A, Salgado H, Gama-Castro S, *et al*. RegulonDB v 10.5: tackling challenges to unify classic and high throughput knowledge of gene regulation in *E. coli* K-12. *Nucleic Acids Res* 2019;**47**:D212–20.
94. Grainger DC. Structure and function of bacterial H-NS protein. *Biochem Soc Trans* 2016;**44**:1561–9.
95. Kahramanoglou C, Seshasayee ASN, Prieto AI, *et al*. Direct and indirect effects of H-NS and Fis on global gene expression control in *Escherichia coli*. *Nucleic Acids Res* 2011;**39**:2073–91.
96. Grainger DC, Hurd D, Goldberg MD, *et al*. Association of nucleoid proteins with coding and non-coding segments of the *Escherichia coli* genome. *Nucleic Acids Res* 2006;**34**:4642–52.
97. Schleiffer A, Kaitna S, Maurer-Stroh S, *et al*. Kleisins: a superfamily of bacterial and eukaryotic SMC protein partners. *Mol Cell* 2003;**11**:571–5.
98. Nolivos S, Sherratt D. The bacterial chromosome: architecture and action of bacterial SMC and SMC-like complexes. *FEMS Microbiol Rev* 2014;**38**:380–92.
99. Palecek JJ, Gruber S. Kite proteins: a superfamily of SMC/Kleisin partners conserved across bacteria, archaea, and eukaryotes. *Structure* 2015;**23**:2183–90.

100. Yu W, Herbert S, Graumann PL, *et al*. Contribution of SMC (structural maintenance of chromosomes) and spoIIIE to chromosome segregation in staphylococci. *J Bacteriol* 2010;**192**:4067–73.
101. Alipour E, Marko JF. Self-organization of domain structures by DNA-loop-extruding enzymes. *Nucleic Acids Res* 2012;**40**:11202–12.
102. Melby TE, Ciampaglio CN, Briscoe G, *et al*. The symmetrical structure of structural maintenance of chromosomes (SMC) and MukB proteins: long, antiparallel coiled coils, folded at a flexible hinge. *J Cell Biol* 1998;**142**:1595–604.
103. Uhlmann F. SMC complexes: from DNA to chromosomes. *Nat Rev Mol Cell Biol* 2016;**17**:399–412.
104. Hancock SP, Stella S, Cascio D, *et al*. DNA sequence determinants controlling affinity, stability and shape of DNA complexes bound by the nucleoid protein Fis. *PLoS One* 2016;**11**:1–24.
105. Ghosh S, Mallick B, Nagaraja V. Direct regulation of topoisomerase activity by a nucleoid-associated protein. *Nucleic Acids Res* 2014;**42**:11156–65.
106. Vietri Rudan M, Barrington C, Henderson S, *et al*. Comparative hi-C reveals that CTCF underlies evolution of chromosomal domain architecture. *Cell Rep* 2015;**10**:1297–309.
107. Lal A, Dhar A, Trostel A, *et al*. Genome scale patterns of supercoiling in a bacterial chromosome. *Nat Commun* 2016;**7**:1–8.
108. Dixon JR, Selvaraj S, Yue F, *et al*. Topological domains in mammalian genomes identified by analysis of chromatin interactions. *Nature* 2012;**485**:376–80.
109. Le TB, Laub MT. Transcription rate and transcript length drive formation of chromosomal interaction domain boundaries. *EMBO J* 2016;**35**:1582–95.
110. Ulianov SV, Khrameeva EE, Gavrilov AA, *et al*. Active chromatin and transcription play a key role in chromosome partitioning into topologically associating domains. *Genome Res* 2016;**26**:70–84.
111. Lioy VS, Cournac A, Marbouty M, *et al*. Multiscale structuring of the *E. coli* chromosome by nucleoid-associated and condensin proteins. *Cell* 2018;**172**:771–83.
112. Webb CD, Teleman A, Gordon S, *et al*. Bipolar localization of the replication origin regions of chromosomes in vegetative and sporulating cells of *B. subtilis*. *Cell* 1997;**88**:667–74.
113. Thiel A, Valens M, Vallet-Gely I, *et al*. Long-range chromosome organization in *E. coli*: a site-specific system isolates the ter macrodomain. *PLoS Genet* 2012;**8**:e1002672.
114. Espeli O, Mercier R, Boccard F. DNA dynamics vary according to macrodomain topography in the *E. coli* chromosome. *Mol Microbiol* 2008;**68**:1418–27.
115. Duigou S, Boccard F. Long range chromosome organization in *Escherichia coli*: the position of the replication origin defines the non-structured regions and the right and left macrodomains. *PLoS Genet* 2017;**13**:e1006758.
116. Javer A, Long Z, Nugent E, *et al*. Short-time movement of *E. coli* chromosomal loci depends on coordinate and subcellular localization. *Nat Commun* 2013;**4**:3003.
117. Valens M, Thiel A, Boccard F. The MaoP/maoS site-specific system organizes the Ori region of the *E. coli* chromosome into a macrodomain. *PLoS Genet* 2016;**12**:1–23.
118. Gao N, Lu G, Lercher MJ, *et al*. Selection for energy efficiency drives strand-biased gene distribution in prokaryotes. *Sci Rep* 2017;**7**:1–10.
119. Rocha EPC, Danchin A. Essentiality, not expressiveness, drives gene-strand bias in bacteria. *Nat Genet* 2003;**34**:377–8.
120. Brewer BJ. When polymerases collide: replication and the transcriptional organization of the *E. coli* chromosome. *Cell* 1988;**53**:679–86.
121. Mao X, Zhang H, Yin Y, *et al*. The percentage of bacterial genes on leading versus lagging strands is influenced by multiple balancing forces. *Nucleic Acids Res* 2012;**40**:8210–8.
122. Inokuchi Y, Hirashima A, Sekine Y, *et al*. Role of ribosome recycling factor (RRF) in translational coupling. *EMBO J* 2000;**19**:3788–98.
123. Li G, Ma Q, Mao X, *et al*. Integration of sequence-similarity and functional association information can overcome intrinsic problems in orthology mapping across bacterial genomes. *Nucleic Acids Res* 2011;**39**:e150.
124. Hanson AD, Pribat A, Waller JC, *et al*. Unknown'proteins and 'orphan'enzymes: the missing half of the engineering parts list–and how to find it. *Biochem J* 2010;**425**: 1–11.
125. Alkan C, Sajjadian S, Eichler EE. Limitations of next-generation genome sequence assembly. *Nat Methods* 2011;**8**:61.
126. Bosi E, Donati B, Galardini M, *et al*. MeDuSa: a multi-draft based scaffolder. *Bioinformatics* 2015;**31**:2443–51.
127. Bao E, Jiang T, Girke T. AlignGraph: algorithm for secondary de novo genome assembly guided by closely related references. *Bioinformatics* 2014;**30**:i319–28.
128. Kim J, Larkin DM, Cai Q, *et al*. Reference-assisted chromosome assembly. *Proc Natl Acad Sci USA* 2013;**110**: 1785–90.
129. Koren S, Harhay GP, Smith TPL, *et al*. Reducing assembly complexity of microbial genomes with single-molecule sequencing. *Genome Biol* 2013;**14**:R101.
130. Kang Y, Gu C, Yuan L, *et al*. Flexibility and symmetry of prokaryotic genome rearrangement reveal lineage-associated core-gene-defined genome organizational frameworks. *MBio* 2014;**5**:e01867–14.
131. Yuan L, Yu Y, Zhu Y, *et al*. GAAP: genome-organization-framework-assisted assembly pipeline for prokaryotic genomes. *BMC Genomics* 2017;**18**:1–8.
132. Carr VR, Shkoporov A, Hill C, *et al*. Probing the mobilome: discoveries in the dynamic microbiome. *Trends Microbiol* 2020. doi: doi.org/10.1016/j.tim.2020.05.003.
133. Liu Z, Ma A, Mathé E, *et al*. Network analyses in microbiome based on high-throughput multi-omics data. *Brief Bioinform* 2020; doi.org/10.1093/bib/bbaa005.
134. Zhang X, Zhao L, Li H. The gut microbiota: emerging evidence in autoimmune diseases. *Trends Mol Med* 2020. doi: doi.org/10.1016/j.molmed.2020.04.001.
135. Bhat AH, Pathak D, Rao A. The alr-groEL1 operon in mycobacterium tuberculosis: an interplay of multiple regulatory elements. *Sci Rep* 2017;**7**:43772.
136. Sharma CM, Hoffmann S, Darfeuille F, *et al*. The primary transcriptome of the major human pathogen *Helicobacter pylori*. *Nature* 2010;**464**:250–5.