Briefings in Bioinformatics, 20(4), 2019, 1114–1124

doi: 10.1093/bib/bbx174
Advance Access Publication Date: 9 January 2018
Paper

Recent development of Ori-Finder system and DoriC database for microbial replication origins

Hao Luo, Chun-Lan Quan, Chong Peng and Feng Gao

Corresponding author: Feng Gao, Department of Physics, School of Science, Key Laboratory of Systems Bioengineering (Ministry of Education) and SynBio Research Platform, Collaborative Innovation Center of Chemical Science and Engineering (Tianjin), Tianjin University, China. Fax: +86(0)22 27402697; E-mail: fgao@tju.edu.cn

Abstract

DNA replication begins at replication origins in all three domains of life. Identification and characterization of replication origins are important not only in providing insights into the structure and function of the replication origins but also in understanding the regulatory mechanisms of the initiation step in DNA replication. The Z-curve method has been used in the identification of replication origins in archaeal genomes successfully since 2002. Furthermore, the Web servers of Ori-Finder and Ori-Finder 2 have been developed to predict replication origins in both bacterial and archaeal genomes based on the Z-curve method, and the replication origins with manual curation have been collected into an online database, DoriC. Ori-Finder system and DoriC database are currently used in the research field of DNA replication origins in prokaryotes, including: (i) identification of *oriC* regions in bacterial and archaeal genomes; (ii) discovery and analysis of the conserved sequences within *oriC* regions; and (iii) strand-biased analysis of bacterial genomes.
Up to now, more and more predicted results by Ori-Finder system were supported by subsequent experiments, and Ori-Finder system has been used to identify the replication origins in >100 newly sequenced prokaryotes in their genome reports. In addition, the data in DoriC database have been widely used in the large-scale analyses of replication origins and strand bias in prokaryotic genomes. Here, we review the development of Ori-Finder system and DoriC database as well as their applications. Some future directions and aspects for extending the application of Ori-Finder and DoriC are also presented.

Key words: replication origin; DNA replication; Z-curve; prokaryotic genome

Introduction

DNA replication is one of the basic processes in all three domains of cellular life. The duplication of the genetic information in a cell begins at specific sites on the chromosomes, termed DNA replication origins. The replication origin regions play significant roles in the DNA replication and serve as recognition sites for initiator proteins and assembly of replication forks [1]. The characteristics of replication origin regions are various among prokaryote and eukaryote, and their nucleotide sequences also present diversities in different organisms [2, 3]. In most bacteria, two replication forks assemble at the replication origin of chromosomes (*oriC*s) and move in opposite directions, and then leading to bidirectional growth of both daughter strands. The *oriC* regions contain several DnaA box motifs, which are 9 bp highly conserved consensus sequences. DnaA boxes are the recognition sites for the DnaA protein, which is essential for the initiation of chromosome replication. Moreover, the *oriC* regions are

Hao Luo is an assistant professor in the Department of Physics, School of Science, Tianjin University, China. His research focuses on DNA replication, gene essentiality and bioinformatics.
Chun-Lan Quan is a graduate student in the Department of Physics, School of Science, Tianjin University, China. Her research interests are bioinformatics and microbial genomics.
Chong Peng is a PhD candidate in the Department of Physics, School of Science, Tianjin University, China. Her research interests are bioinformatics and gene essentiality.
Feng Gao is a professor in the Department of Physics, School of Science, Key Laboratory of Systems Bioengineering (Ministry of Education) and SynBio Research Platform, Collaborative Innovation Center of Chemical Science and Engineering (Tianjin), Tianjin University, China. His research studies are performed in the fields of computational biology and bioinformatics with a special focus on microbial genomics and functional genomics.
Submitted: 20 September 2017; **Received (in revised form):** 4 December 2017

frequently located next to the replication-related genes [4, 5]. Detailed analyses have shown that the consensus sequences of DnaA boxes and distributed genes adjacent to *oriC*s are highly conserved in different phyla [6]. In addition, because of the asymmetric nucleotide composition of prokaryotic chromosomes, the replication origins have been identified around the boundaries of GC or AT skew [7]. In eukaryote, because of the huge size of genomes, the chromosomes use multiple dispersed replication origins to initiate the DNA replication, ranging from hundreds in yeast to tens of thousands in human [8–10]. Minichromosome maintenance (MCM) complexes are first loaded at replication origins in G1 phase of the cell cycle. Then, the origin-bound MCM complexes unwind the double-stranded DNA at the origins, recruit DNA polymerases and initiate DNA synthesis in S phase [11]. The replication origins in unicellular eukaryotes *Saccharomyces cerevisiae* and *Schizosaccharomyces pombe* have been well characterized by microarray and deep sequencing techniques [12, 13]. However, the selection and regulation of DNA replication origins in the higher eukaryotes are more complex and diverse [14–16]. As a separate domain in the three-domain system, archaea share some similar features with both bacteria and eukaryotes. The locations of *oriC* regions in archaea are also adjacent to the replication-related genes, and the origin recognition boxes (ORBs) are distributed within *oriC* sequences [17]. In some archaea, single chromosome could adopt more than one *oriC*s in initiation of DNA replication as eukaryotes. Furthermore, the origin binding proteins in archaea are homologous to the corresponding eukaryotic Orc1/Cdc6 proteins [18], and a homologous MCM hexameric ring serving as replicative helicase is loaded by origin-bound initiator proteins [19].

Since the first complete bacterial genome was sequenced in 1995 [20], the available microbial genomic data have been increasing exponentially with the rapid development of sequencing techniques. Recently, the White House launched the National Microbiome Initiative (NMI) in 2016, which aimed to deepen the understanding of the microbes that live in humans, animals, crops, soils, oceans, etc. [21], and some other microbial programs, such as MGP, MetaHIT and HMP, have also been carried out in the past decade [22–24]. These microbial projects have produced a large amount of sequence data, thereby creating an opportunity for exploration of the molecular mechanisms for initiating cellular DNA replication by *in vivo* experiments as well as *in silico* analysis at the genome level. Development of bioinformatics tools to mine useful biological information in microbial genomes will contribute to bridge the gap between genomic data and knowledge discovery.

On the other hand, the accumulation of genomic data has created great challenges and opportunities for identification and characterization of the replication origins on a large scale. Identification of the *oriC* regions will not only provide insights into the structure and functions of replication origins but also facilitate the studies in regulatory mechanisms of DNA replication initiation [25]. Our laboratory has developed the Web service and database in this field based on the Z-curve method. The Z-curve is a unique three-dimensional curve used to transform the DNA sequence in a three-dimensional space, and the three components of the Z-curve could represent three independent distributions, including purine/pyrimidine (R/Y) bases, amino/keto (M/K) bases and strong-H bond/weak-H bond (S/W) along the sequence. The Z-curve method could be used in the detection of the asymmetric nucleotide distribution around replication origins [26]. With this method, the *oriC* regions in *Methanocaldococcus jannaschii*, *Methanosarcina mazei*, *Halobacterium* sp. strain NRC-1 and *Sulfolobus solfataricus* P2 have been identified successfully, and the predicted results are also consistent with the subsequent experiments [27, 28]. To facilitate the prediction of *oriC* regions, a Web-based system, Ori-Finder (http://tubic.tju.edu.cn/Ori-Finder/), was developed to find replication origins in bacterial genomes with high accuracy and reliability [29]. Up to now, more and more predicted results by Ori-Finder system were supported by subsequent experiments, and Ori-Finder system has been used to identify the replication origins in >100 newly sequenced prokaryotes in their genome reports [30–32]. Furthermore, we also designed Ori-Finder 2, a new Web-based tool to identify the *oriC* regions in the archaeal genomes [33]. Then, the *oriC* regions identified by *in silico* analyses, as well as *in vivo* and *in vitro* experiments have been organized into DoriC (http://tubic.tju.edu.cn/doric/), a database of *oriC* regions in bacterial and archaeal genomes [34, 35]. DoriC has provided insights into the regulatory mechanisms of the initiation step in DNA replication as well as the molecular mechanisms of strand bias in genomes. The application of the rules derived from the database will also be helpful to develop new prediction algorithms of replication origins and speed up the experimental confirmation and functional analysis of *oriC*s in bacterial or archaeal genomes.

In this review, we briefly introduce the development of the Ori-Finder system and DoriC database and review some applications with those tools. Additionally, some future directions and aspects for extending the application of Ori-Finder and DoriC are also presented.

The development of Ori-Finder and DoriC

Ori-Finder: an online tool for *oriC* prediction in bacteria

DNA replication is a precise and complex process in the cell life, during which the cell uses a great deal of enzymes and proteins to synthesize the nucleotides. So that the perfect prediction algorithms of *oriC* regions should take as many factors as possible into consideration, which are concerned with the DNA replication process. It is well known that the DNA replication asymmetry gives rise to compositional deviations between the leading and lagging strands [36]. As the pioneer work to identify bacterial *oriC* regions *in silico*, the GC skew analysis is mainly based on the asymmetric nucleotide composition [7]. Later, other skew methods, such as the cumulative GC skew without sliding windows method and oligomer-skew, were proposed to predict *oriC* regions in bacterial and archaeal genomes [37]. Nevertheless, with the analysis of asymmetric nucleotide composition, scientists could only predict the approximate location, but not the exact boundary of replication origins. Meanwhile, the bacterial replication origins are frequently located in the intergenic regions that are adjacent to the replication-related genes, such as *dnaA*, *gidA* and so on. Hence, a GC skew analysis together with the location of *dnaA* gene and distribution of DnaA boxes led to the more accurate predictions of *oriC* regions [38, 39]. However, it is inconvenient for the biologists who sequenced the bacterial genomes to take all the possible characteristics of *oriC* regions into account, such as the effects of 'species-specific' DnaA box motif, thereby leading to the wrong prediction results of *oriC* regions [40, 41], and the pipelines or Web servers that could predict and visualize the related data of *oriC* regions in complete bacterial genome automatically were in great need. Therefore, we developed Ori-Finder to predict *oriC* regions in the complete bacterial genomes, which integrated gene prediction, analysis of base composition asymmetry, distribution of DnaA boxes, occurrence of genes frequently close

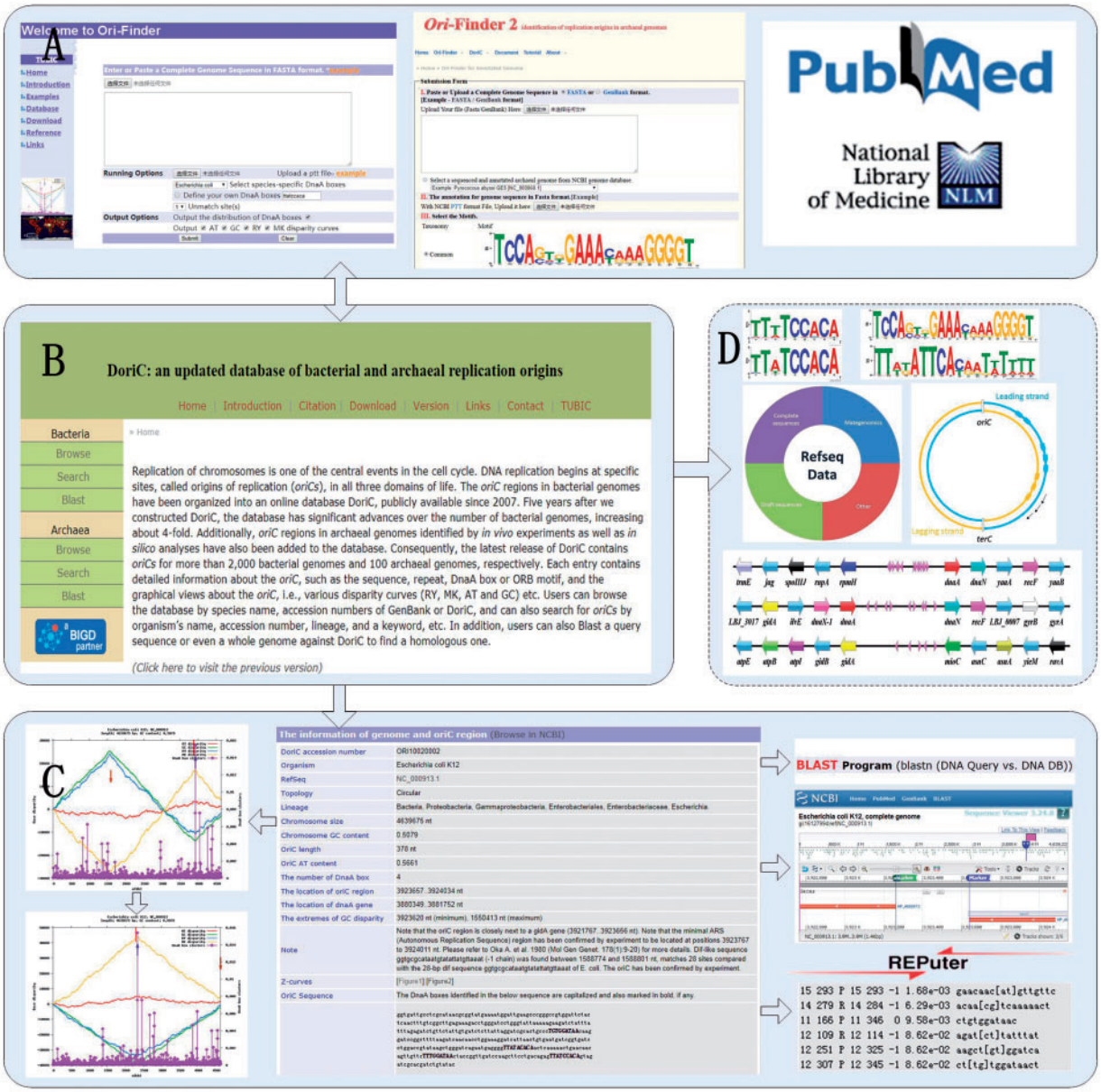

Figure 1. The architecture of Ori-Finder system and DoriC database. (**A**) Ori-Finder, Ori-Finder 2 and PubMed used as the data source of DoriC database. (**B**) The screenshot of DoriC main page. (**C**) A representative record in DoriC database. The left part shows the Z-curves for the genome sequences, and the right part presents some tools used by DoriC, including BLAST, NCBI genome viewer and REPuter. (**D**) Future perspectives of Ori-Finder and DoriC including the newfound characteristics as well as the extended *oriC* prediction and collection.

to *oriC* regions and phylogenetic relationships [29] (Figure 1A). In addition, Ori-Finder can also predict *oriC* regions in some draft genomes only with contigs or scaffolds. Owing to integration of Z-curve method, Ori-Finder is also used to separate the leading and lagging strands to perform the strand bias analysis of biological characteristics.

DoriC: a database of *oriC*s

DoriC is a database of manually curated *oriC* regions, which was initially publicly available in 2007 (Figure 1B). At that time, the complete bacterial genome data were accumulated rapidly because of the advance in high-throughput sequencing technology. However, the experimental method is impossible to identify all the replication origins in the sequenced genomes extensively. Furthermore, some large-scale analyses of the bacterial genomes, such as those of replication origins and strand bias in genomes, were restricted by the absence of *oriC* data. Before the construction of DoriC, the Z-curve method has been used in the prediction of *oriC*s in several archaeal genomes, and some of the results were confirmed by experimental data subsequently. To extensively identify *oriC*s with high accuracy and reliability, our laboratory developed an integrated *in silico* method to predict *oriC* regions of bacterial genomes, and the predicted *oriC*s as well as those identified by *in vivo* or *in vitro* experiments were manually curated and collected into the DoriC database. The first public release of the database only collected 478 predicted *oriC*s in 425 bacterial genomes, and 72.2% of the predicted *oriC*s have consistent features with each other, including typical base composition asymmetry, DnaA box

distribution and indicator gene positions [34]. Furthermore, DoriC database presents the detailed information of *oriC* regions, including the experimental evidences of the replication origins, the number of DnaA boxes, disparity curves and replication-related genes, and enables the retrieve of DoriC entries and BLAST search for *oriC* regions. The putative *dif* (deletion-induced filamentation) sequences, which are associated with DNA replication terminus, are also added to the *oriC* records. The accumulation of *oriC* records in the database would provide the possibility to explore the characteristics in the *oriC* regions.

With the increasing availability of completely sequenced prokaryotic genomes and experimental evidences, we presented an updated version of the database DoriC 5.0 in 2013. Compared with the initial release, the number of *oriC* regions in bacterial genomes has been increased from 425 to 1528 in DoriC 5.0, and the database provides more information of the *oriC* regions, including repeats by REPuter, URLs to NCBI or UCSC genome browsers, which are useful to explore the characteristics of the *oriC* regions [42–44]. In addition, the 86 *oriC* regions in 83 archaeal genomes identified by *in vivo* experiments, as well as *in silico* analyses, were also added to the database in this version. Currently, the DoriC database has collected 3423 *oriC* records in >2700 complete RefSeq bacterial genomes and 257 *oriC* regions in over 200 archaeal genomes with manual curation. Figure 1C displays a representative record in DoriC database.

Ori-Finder 2: an online tool for *oriC* prediction in archaea

Archaea are classified as a separate domain in the three-domain system, and some of them exist in various extreme environments on earth, such as hot spring and salt lake [45]. Their special habits make it difficult in strain collection and cell cultivation, leading to slow progress in the genome sequencing for a long time. The first *oriC* of archaea was predicted in *Halobacterium* sp. strain NRC-1 with GC skew method and then confirmed by cloning into a nonreplicating plasmid [46]. With the Z-curve method, the *oriC*s in *M. jannaschii*, *M. mazei*, *Halobacterium* sp. strain NRC-1 and *S. solfataricus* P2 were identified, and some predicted results were consistent with subsequent experiments. In 2014, Wu *et al.* [47] also predicted putative multiple *orc1*/*cdc6*-associated *oriC*s in all the available Haloarchaeal genomes. In recent years, the development of the high-throughput sequencing technology results in the rapid increase of the archaeal genome projects. Therefore, we further developed a Web-based tool Ori-Finder 2 to predict the *oriC* regions in archaeal genomes automatically, based on the frame of Ori-Finder. The *oriC*s in archaea have significant differences with those in bacteria. For example, in contrast to the DnaA boxes in bacteria, the ORB sequences present more diversities in different species. With the archaeal *oriC*s in DoriC database, the consensus sequences of ORB motifs were calculated for different taxonomies, including Methanobacteriaceae, Methanomicrobia, Methanococcaceae, Sulfolobaceae and Thermococcaceae, by Multiple EM for Motif Elicitation (MEME) program [48]. Based on Ori-Finder 2, the intergenic sequences with the putative ORB sequences and adjacent to the replication-related genes are predicted as *oriC* regions. Because this method may fail to identify the *oriC*s adjacent to the uncharacterized genes that might be involved in DNA replication, the intergenic sequences, which contain more than two putative ORB motifs, are also predicted as *oriC*s. Currently, Ori-Finder 2 has been used to identify the *oriC*s in *Pyrococcus chitonophagus* DSM 10152, *Thermococcus* sp. strain 2319x1, *Haloarculahispanica pleomorphic* virus 3, *Natrinema* sp. J7 and *Methanobrevibacter ruminantium* M1 [49–53]. However, Ori-Finder 2 could not find all the potential origins of replication for the genomes with multiple *oriC*s currently. With the increase of the experimentally confirmed *oriC*s in archaea, it will become more accurate and sensitive by the continuous improvement.

Prediction of replication origins by Ori-Finder system

Ori-Finder and Ori-Finder 2 have a friendly and intuitive input interface, and use an integrated method to predict replication origins in prokaryotic genomes, which are available at http://tubic.tju.edu.cn/Ori-Finder/ and http://tubic.tju.edu.cn/Ori-Finder2/, respectively. Figure 1A presents the submission Web pages of the Web servers. Both the Web servers integrate the gene predicting pipelines, ZCURVE1.02 or Glimmer3, to perform the gene prediction in the unannotated genomes [54, 55]. Users can submit the annotated genome sequence by uploading the sequence file in FASTA format together with its protein table (PTT) file to the Web servers, and the Ori-Finder 2 can also accept the annotated genome file in GenBank format. BLAST program has been installed for functional annotation of genes by search against indicator genes (such as *dnaA*, *dnaN*, *hemE* and *gidA* in bacteria, or *cdc6*, *orc1* and Mc-pRIP in archaea) throughout the genome. Both the Web servers enable users to select or type the motif sequences of DnaA boxes or ORBs to predict replication origin. The DnaA boxes or ORBs are 'species-specific' conserved sequences within the *oriC* regions and recognition sites for the DnaA proteins or Cdc6/Orc1 proteins. Ori-Finder provides 16 different types of DnaA box, and also allows the users to define some unique DnaA boxes by themselves. Whereas, FIMO (Find Individual Motif Occurrences) is used to obtain ORB sequences with the PSPM of five taxonomic clusters in Ori-Finder 2, and Weblogos are provided to facilitate the selection of ORBs [56]. Finally, the intergenic regions with the required characteristics are predicted as *oriC* regions. Note that, as Ori-Finder 2 used text search in the annotated genome file, some replication-related genes with unclear annotation might be ignored. So that we recommend users to compare the result with that based on the unannotated sequence. Figure 2 displays the workflow of both Ori-Finder and Ori-Finder 2.

In the result Web page, the information, including the genome size, GC content, the locations of the indicator genes and predicted *oriC* regions, as well as the Z-curve (AT, GC, RY and MK disparity curves) of the input sequence, is presented as an html table. In addition, the location of DnaA boxes or ORB sequences within the *oriC* regions and the distribution of them in the whole genomes are also available for download from the provided URL. For Ori-Finder 2, the repeats identified by REPuter and the homologs in DoriC are also displayed in the result table.

By comparison, the two Web servers share a common procedure to predict *oriC* regions (Figure 2). However, Ori-Finder could identify *oriC* regions in most of the bacterial chromosomes. In contrast, the sensitivity and precision of the predictions by Ori-Finder 2 are only 66.7 and 62.1%, respectively [33]. There are three main reasons that result in the significantly different performances. First, it is well known that most of the *oriC* regions in prokaryotic genomes are adjacent to the replication-related genes. In bacteria, *dnaA* gene plays a key role in the initiation of DNA replication for most bacteria. In addition, some other replication-related genes, such as *dnaN*, *hemE* and *gidA*, are also

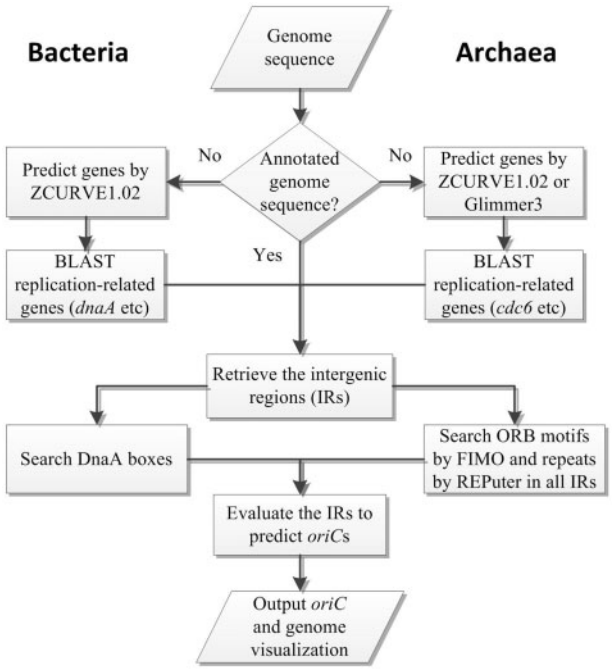

Figure 2. The workflow of Ori-Finder system for the bacterial and archaeal genomes.

considered as the indicators in the *oriC* prediction. Besides the *orc1/cdc6* gene, there are still many unknown genes involved in DNA replication in archaea. It is the reason that the intergenic sequences with more than two putative ORBs are also predicted as *oriC*s by Ori-Finder 2. Second, the binding sites of replication proteins, such as DnaA box and ORB, are essential for initiation of chromosome replication. Although the DNA boxes in bacteria show some differences throughout the bacterial kingdom, the sequences of them are considerably more conserved in comparison with the ORB sequences in archaea. Most of DnaA boxes are 9 bp sequences and the derivatives are based on the *Escherichia coli* perfect DnaA box 'TTATCCACA' with one or more mismatches. In archaea, the consensus sequences 'TCCA—GAAAC' were found by scanning DoriC database with MEME, and a 'G-string' (GGGGT) is observed obviously at the end of ORB motifs in Methanomicrobia and Sulfolobacea. Furthermore, some other conserved motifs are also found in Sulflobacea and Thermococcaceae. Nevertheless, these motifs are more degenerative compared with DnaA boxs. Finally, the majority of bacteria use one *oriC* to start the DNA replication, whereas some archaea could adopt multiple *oriC*s. Moreover, the location of typical *oriC* in bacteria is next to the extreme of GC disparity curve, and the curve is shaped like 'V' graph clearly. However, the GC disparity curves in archaea are more irregular, so that the *oriC*s in archaea are not always near the extremes of GC disparity curves. Despite of some difficulties in the prediction of archaeal *oriC*s, Ori-Finder 2 will be improved to be more accurate with the increase of the experimental *oriC*s data.

In addition to Ori-Finder system, Oriloc is the alternative tool to predict the *oriC* regions in bacterial chromosomes, which is mainly based on the GC skew method [57]. The shift points of relative GC skew defined as $(G-C)/(G+C)$ have been used in the identification of the replication origins and termini. To improve the accuracy of prediction, the cumulative GC skew calculated as $(G-C)$ was used to eliminate the effect of the window size. However, the relative GC skew could predict the turning point in the skew graphs as the location of *oriC* region, which corresponds to the extreme value in the cumulative GC skew as well as GC disparity curve by Z-curve method. However, Oriloc mainly used the cumulative GC skew method to identify *oriC* regions and could not provide the exact boundaries of *oriC* regions. Therefore, we compared the location of *oriC* and minimum of GC disparity curve using the records in DoriC database and found that the most of *oriC*s are close to the minimum of GC disparity curve, but there are still about 7% (246) of total *oriC*s, which are of over the one-tenth of the distance to the whole chromosomes. This also confirmed that the *oriC* prediction in bacteria should take other biological characteristics into consideration, such as the distribution of indictor genes and DnaA boxes. Furthermore, the experimentally confirmed *oriC*s together with the corresponding predictions by Ori-Finder are summarized at http://tubic.tju.edu.cn/doric/supplementary.php.

Exploration of replication origins with DoriC database

Figure 1A presents the main source of DoriC database including the predicted results by Ori-Finder system and experimentally confirmed *oriC*s from literature in PubMed. The *oriC*s in DoriC are distributed across all the phyla of bacteria and four archaeal phyla, including Crenarchaeota, Euryarchaeota, Korarchaeota and Thaumarchaeota. Figure 3 presents the main distribution of *oriC* records in DoriC database by phyla in both bacteria and archaea. The *oriC*s from the phyla Proteobacteria and Firmicutes constitute the primary component in bacteria, while for the archaea, the *oriC*s from the phylum Euryachaeota are in the majority. DoriC database also included seven archaeal species from Thaumarchaeota and Korarchaeota, and one unclassified archaeal species, whose *oriC* regions were identified by Ori-Finder 2. The predicted results are consistent with the typical characteristics of archaeal *oriC*s. Different from the bacteria, a substantial proportion of *oriC*s in archaea could not be identified *in vivo* or *in silico*. Currently, DoriC database has covered all the RefSeq bacterial complete genomes, so that we will focus on the bacterial *oriC*s in this section.

In the previous section, we have described several typical characteristics of bacterial *oriC*s, such as asymmetrical nucleotide distributions, the replication-related genes and DnaA boxes. Now, these conserved features of *oriC*s could be summarized based on DoriC database, which were calculated by python scripts. The DnaA boxes are essential for DNA replication and enriched in the *oriC* regions. We retrieved the DnaA boxes in all the *oriC*s with no more than two mismatches from *E. coli* perfect DnaA box (TTATCCACA), and calculated the GC contents of those DnaA boxes and the corresponding chromosomes. Consequently, the statistically significant correlation is observed in Figure 4A ($R = 0.71$), and the result is consistent with the previous reports [38]. It suggests that the GC content of the chromosome could affect that of DnaA boxes. In particular, the position of cytosine in TTATCCACA is more conserved in the high GC content genomes. Besides that, adenine or thymine in some other positions is also offset to guanine or cytosine in the bacteria with high GC content, and *vice versa*. A cluster of DnaA boxes could facilitate DnaA proteins to bind the *oriC* regions and participate in the regulation of chromosome replication. Figure 4B presents the percentage of *oriC* regions with different number of DnaA boxes and displays that most of *oriC*s contain more than four DnaA boxes. In addition, the

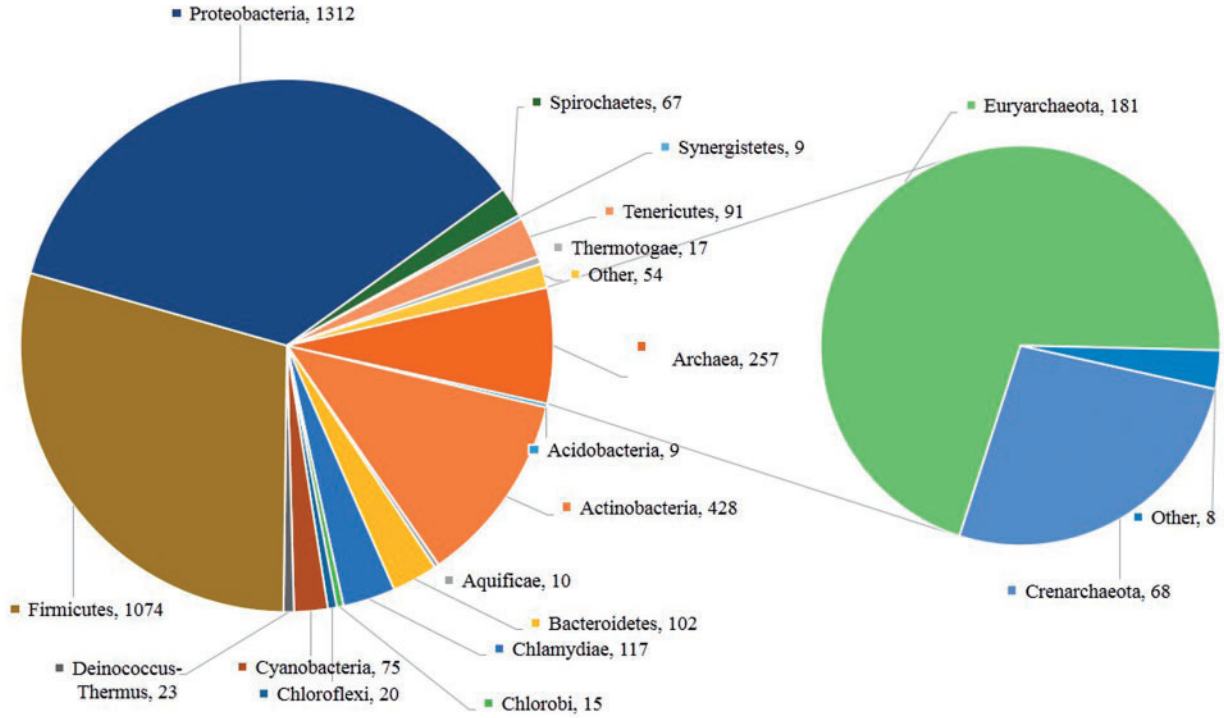

Figure 3. The taxonomic distribution of *oriC* records in DoriC database. The phyla with the *oriC* records less than nine are classified into the 'other' subcategory in the pie chart.

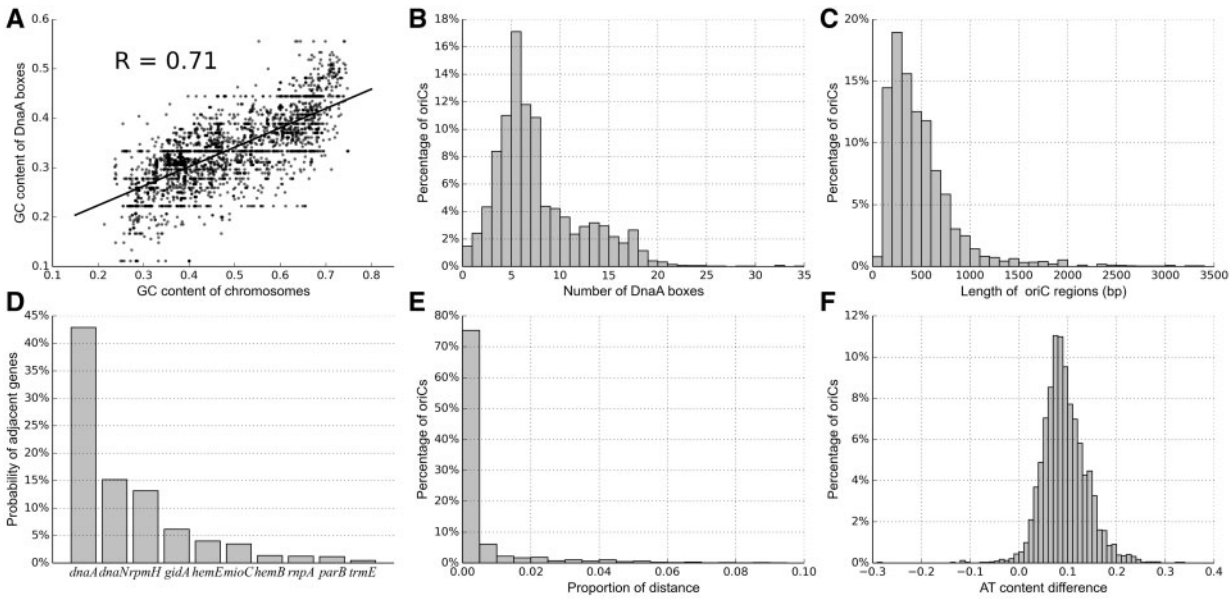

Figure 4. Several characteristics of replication origins. (**A**) The relationship between the GC content of DnaA boxes and that of chromosome. (**B**) The percentages of bacterial *oriC*s with different number of DnaA boxes. (**C**) The percentages of bacterial *oriC*s with different length. (**D**) The probabilities of top 10 replication-related indicator genes adjacent to the *oriC*s. Note that some *oriC*s are next to two indicator genes, so that the sum of probabilities is not equal to 100%. (**E**) The distribution of relative distance between *oriC* and minimum GC disparity. It should be noted that the *x* axis indicates the proportion of the chromosome size. (**F**) The difference between the AT content of *oriC* region and that of chromosome.

distribution of DnaA boxes throughout the chromosomes is also presented in the Z-curve figures provided by DoriC, and the significant abundance in the *oriC* regions could be observed in most of the records. The replication-related gene is another important indicator of *oriC* region. Figure 4D displays the probabilities of the top 10 genes adjacent to the *oriC* regions. Nearly half of the *oriC*s are next to the *dnaA* genes. Moreover, *dnaN*, *rpmH* and *gidA* also appear on the sides of the *oriC* regions frequently. It should be noted that some *oriC*s were bipartite origins [58], which were split into two subregions by the *dnaA* gene and located in the intergenic regions of *dnaA-dnaN* and *rpmH-dnaA* (*Nocardia farcinica* IFM 10152 and *Chlorobium chlorochromatii*

CaD3 in Table 1). For the phylum Cyanobacteria, most of *oriC* regions are adjacent to *dnaN* gene instead of *dnaA* gene in case of the separation of the two genes in the chromosomes, which is supported by a series of experiments [59–61]. The different mechanism of DNA replication between the leading and lagging strands leads to the asymmetric nucleotide distribution. As a result, the *oriC* regions are usually near the switch of GC skew or extreme of GC disparity. The distances between *oriC* and minimum GC disparity relative to the whole chromosome length were presented in Figure 4E. The majority of *oriC*s are close to the location of minimum GC disparity. In addition, the length of *oriC*s and the difference of AT content in *oriC* to that in chromosome are presented in Figure 4C and 4F. Most of the *oriC*s are about 500 bp long with >10% higher AT content, which facilitate the DNA melting.

However, the statistical analyses could only reflect some common characteristics of bacterial *oriC*s, but exceptions always exist in individuals. Consequently, we analyzed the indicator genes and DnaA boxes by phyla based on the records in DoriC database and found the gene clusters frequently around *oriC*s and the 'species-specific' DnaA boxes within *oriC*s. Table 1 displays the *oriC* information of some bacterial and archaeal chromosomes in DoriC, which are used to illuminate the common features in their phyla. Besides the genes (*dnaA*, *dnaN*, *gidA*, *hemE* and *rpmH*) listed above, some other genes, such as *rnpA*, *gyrB* and *recF*, seem to be near the replication origins. As for the bacteria with multiple chromosomes, the replication initiation genes and plasmid partition genes, such as *repA*, *repC*, *parA* and *parB*, are often the indicators for the *oriC*s in the extra chromosomes, and those genes have also been verified to be relevant with plasmid replication [62]. This suggests that the microbial extra chromosomes may originate from megaplasmids. Moreover, the 'species-specific' DnaA boxes different from the *E. coli* perfect DnaA box 'TTATCCACA' are also outlined in this table. For example, the motif 'TTTTCCACA' was found in most species of the phylum Cyanobacteria. Owing to the high GC content of chromosome, a cluster of 'TTGTCCACA' was discovered in the *oriC* of *N. farcinica* IFM 10152. In the extra chromosomes, some other motifs entirely different from classic DnaA box are also listed in this table, and these motifs usually appear as repeats in *oriC* regions.

Currently, the majority of archaeal *oriC*s in DoriC are the *oric1/cdc6*-associated *oriC*s, and a distant homolog of the *cdc6* gene, named Mc-pRIP for the putative replication initiator protein, was found next to the *oriC* regions in the order Methanococcales during the update of DoriC database. Beyond that, some archaea with multiple *oriC*s are also summarized, such as *Sulfolobus* and *Halobacteria*. For more details, please refer to the article of Ori-Finder 2 [33].

Applications with DoriC database

With the accumulation of the prokaryotic *oriC* records in DoriC database, it is possible to determine the conserved features of *oriC*s, analyze the strand bias and search the homologous *oriC*s. Figure 5 outlines the main applications, and several examples will be presented in this section.

Data mining with DoriC database

By exploring *oriC*s in DoriC database, some newfound features associated with *oriC* regions, such as motif sequences, and multiple replication origins in single bacterial chromosome have been discovered and supported by other studies. For example, Murray *et al.* identified a new indispensable bacterial replication origin element, repeating trinucleotide motif, named DnaA-trio, and demonstrated these elements play an important role in the stabilization of DNA filaments by experiments. Then, the new elements have been detected throughout the bacterial kingdom by bioinformatics analysis with DoriC database, indicating that DnaA-trio is another core *oriC* element [63]. DNA methylation is an epigenetic mechanism, which is involved in various biological processes in bacteria including DNA replication. Bendall *et al.* [64] performed the single-molecule real-time sequencing in *Shewanella oneidensis* MR-1 to reveal methylation of adenine (N6mA) throughout the genome, and the methylated GATC motifs are found enriched in the *oriC* region. The further comparative analysis of the Gammaproteobacteria genomes including those in DoriC database revealed that the *oriC*s are enriched for GATC motifs with the presence of *dam* and *seqA*. It is well known that bacteria typically have single replication origin in a chromosome. However, double replication origins could exist in some artificial biological systems. Several bacterial genomes were reformed by synthetic biology methodologies, and more than one WT origins have been extensively characterized in those chromosomes [65–67]. This discovery indicated that multiple origins could occur on a bacterial chromosome, and several bacterial chromosomes with putative double origins of replication, including *Acidaminococcus fermentans* DSM 20731, *Dehalobacter* sp. CF, *Ralstonia pickettii* 12 D chromosome 1 and *Ochrobactrum anthropi* ATCC 49188 chromosome I, were found indeed in DoriC [68]. Recent work reported that *Achromatium oxaliferum* and *Synechocystis* may harbor different replication origins, which also supported our hypothesis [69, 70].

The strand-biased analysis with DoriC database

The strand-biased analyses of the biological features are important to understand the mechanisms of many biological processes. Consequently, the data of *oriC* regions in DoriC database have been widely used in a series of comparative genomics studies focused on the strand bias, such as nucleotide composition [71–73], codon usage [74, 75], substitution rate [76], gene expression [77] and genes distributions [78, 79]. Here, we introduce a few examples about strand-biased analyses. It is well known that majority of genes in bacterial chromosomes tend to locate at the leading strand. A number of studies have been carried out with DoriC database aiming to provide explanations for such observations. Mao *et al.* [80] performed a computational study on 725 bacterial genomes, and found the genes with different functional categories have a various performances of strand bias. The preference for genes on the leading strand in certain functional categories could enhance the survivability of the host and keep them moving to the more efficient leading strand. The expression level of genes was once considered as the main force to cause the strand bias. However, the analysis of the gene distributions in *Bacillus subtilis* and *E. coli* showed that essentiality, not expressiveness, is the basis of gene strand bias [81], and our laboratory also confirmed the previous findings that essential genes are more frequently situated at the leading strand with DoriC and DEG database [78]. Furthermore, only the essential genes with certain COG subcategories showed the preference. These results are helpful to understand the architecture of bacterial chromosomes. This property was also used in the prediction of gene essentiality in bacterial genomes [82, 83].

Table 1. The *oriC* information of some representative bacterial and archaeal chromosomes in DoriC

Organism	Refseq	Phylum	GC (%)	Adjacent gene cluster	DnaA box or ORB sequence
Nocardia farcinica IFM 10152	NC_006361	Actinobacteria	70.83	*rnpA_rpmH_***oriC**1*_dnaA_***oriC2***_dnaN_recF_gyrB*	TTGTCCACA
Bacteroides helcogenes P 36-108	NC_014933	Bacteroidetes	44.72	*gidA_***oriC**	TTATACACA
Chlamydia trachomatis Sweden2	NC_017441	Chlamydiae	41.31	*ispA_glmU_***oriC**_hemB_nqrA_greA_aspC*	TTATCAACA
Chlorobium chlorochromatii CaD3	NC_007514	Chlorobi	44.28	*rnpA_rpmH_***oriC1**_dnaA_***oriC2**_dnaN_ recF_*	TTATCCACA
Dehalococcoides mccartyi DCMB5	NC_020386	Chloroflexi	47.07	*rpsT_lexA_fmt_dnaA_***oriC**_obgE_nadD_gyrB_relA_hisS*	TTATCCAAA
Synechococcus sp. WH 7803	NC_009481	Cyanobacteria	60.24	*uvrA_recN_thrC_***oriC**_dnaN_purL_purF_gyrA*	TTTTCCACA
Deinococcus radiodurans R1 chr. I and II	NC_001263	Deinococcus-Thermus	67.01	*eno_dnaN_***oriC**_dnaA*	TT[AT]TCCACA
	NC_001264		66.69	***oriC**_parA_parB*	
Bacillus subtilis subsp. subtilis str. 168	NC_000964	Firmicutes	43.52	*jag_spoIIIJ_rnpA_rpmH_***oriC1**_dnaA_***oriC2**_dnaN_yaaA_recF_yaaB_gyrB*	TT[AT]TCCACA
Brucella abortus biovar 1 str. 9-941 chr. I and II	NC_006932	Proteobacteria	57.16	*trmE_rho_hemE_***oriC**_maf_aroE_coaE_dnaQ*	TTATCCACA
	NC_006933		57.34	*hemN-2_repA_repB_***oriC**_repC*	
Escherichia fergusonii ATCC 35469	NC_011740	Proteobacteria	49.94	*atpE_atpB_atpI_gidB_gidA_***oriC**_mioC_asnC_asnA_yieM_ravA*	TTATCCACA
Burkholderia cenocepacia J2315 chr. I, II and III	NC_011000	Proteobacteria	66.68	*arsC_***oriC**_parA_parB_repC*	TTATCCACA
	NC_011001		67.28	***oriC**_repC_parB_repA*	TTATGCGCATAA
	NC_011002		66.92	***oriC**_repC*	
Geobacter sulfurreducens PCA	NC_002939	Proteobacteria	60.94	*yidC_rnpA_rpmH_***oriC1**_dnaA_***oriC2**_dnaN_recF_gyrB_gyrA*	TTATCCACA
Helicobacter pylori 26695	NC_000915	Proteobacteria	38.87	*dnaA_***oriC**_ glmS_thyX*	TTATTCACA
Vibrio cholerae O1 biovar eltor str. N16961 chr. I and II	NC_002505	Proteobacteria	47.7	*parA_parB_gidB_gidA_***oriC**_yidC_trmE*	TTATCCACA
	NC_002506		46.91	*parB_parA_***oriC**	ATGATCAAGAG
Leptospira borgpetersenii serovar Hardjo-bovis JB197 chr. I and II	NC_008510	Spirochaetes	40.23	*gidA_ilvE_dnaX1_dnaA_***oriC**_dnaN_recF_gyrB_gyrA*	TTTTCCACA
	NC_008511		40.43	***oriC**_parA_parB*	
Thermotoga sp. RQ2	NC_010483	Thermotogae	46.18	***oriC**_rpmF*	AAACCTACCACC
Sulfolobus solfataricus P2	NC_002754	Crenarchaeota	35.79	*cdc6_***oriC-I**, **oriC-II**,***oriC-III**_cdc6*	TCCA[AG][AT][TG]GAA[CA][CT][GA]AAGGGGT
Pyrococcus abyssi GE5	NC_000868	Euryarchaeota	44.71	*cdc6_***oriC**	TCCA[CG]T[TG]GAAA[TC][GA]AAGGGGT
Methanococcus maripaludis X1	NC_015847	Euryarchaeota	32.94	***oriC**_Mc-pRIP*	TT[TA][GT] ATTCA[TC][GA]AT[AT]T[AT]T[AT]

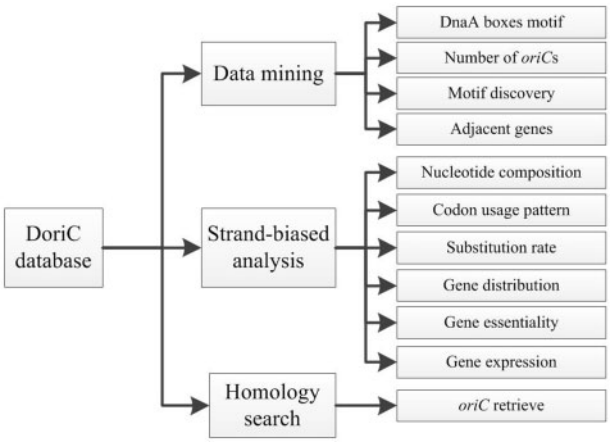

Figure 5. Main applications based on DoriC database including data mining, strand-biased analysis and homology search.

Conclusion and future perspectives

In this article, we briefly reviewed the history of Ori-Finder system and DoriC database and then outlined the main methodology and applications associated with them. Ori-Finder system used an integrated method to identify *oriC*s in prokaryotic genome sequences, and the *oriC*s predicted by *in silico* methods as well as those identified by *in vivo* or *in vitro* experiments were collected to DoriC database after manual curation. Currently, Ori-Finder system becomes a popular software tool to predict prokaryotic replication origins, and some of the predictions were confirmed by experiments. DoriC database has stored about 3600 records of *oriC* regions in both bacterial and archaeal genomes, which would facilitate the research of the large-scale data mining and strand-biased analyses associated with the replication origins. Furthermore, we also explored the *oriC* records in DoriC database and displayed the statistical results as well as the representative organisms here. However, next-generation sequencing technologies have created a new challenge for the identification of replication origins in different types of genomic data. To address this challenge, Ori-Finder system will be extended to predict *oriC* regions in metagenomic sequences in future, and the new version of DoriC database will also include the information of the strand-biased analyses for nucleotide asymmetry, codon usage, gene distribution, etc. The *oriC* prediction in the RefSeq genomes has laid a firm foundation for the further development of Ori-Finder system and DoriC database, which will serve as the critical tools in the prokaryotic genomics.

Key Points

- Ori-Finder system is designed for the *oriC* prediction in bacterial and archaeal genomes with high accuracy and reliability, which integrates gene prediction, analysis of base composition asymmetry, distribution of DnaA boxes or ORBs, occurrence of genes frequently close to *oriC* regions and phylogenetic relationships.
- DoriC database contains 3423 *oriC* records in > 2700 complete RefSeq bacterial chromosomes and 257 *oriC* regions in over 200 archaeal genomes with manual curation. Detailed information about *oriC* regions, such as DNA boxes or ORBs, repeat sequences, replication-related genes and URLs to genome browser, are provided in the database.
- Ori-Finder system and DoriC database have been widely used in the research field of DNA replication origins in prokaryotes, including the *oriC* retrieve, motif sequences discovery and strand-biased analysis.

Acknowledgements

The authors would like to thank Professor Chun-Ting Zhang for the invaluable assistance and inspiring discussions.

Funding

This work was supported by the National Natural Science Foundation of China (grant numbers 31571358, 21621004 and 31171238) and the National High-Tech Research and Development Program (863) of China (grant number 2015AA020101).

Reference

1. Costa A, Hood IV, Berger JM. Mechanisms for initiating cellular DNA replication. *Annu Rev Biochem* 2013;**82**:25–54.
2. O'Donnell M, Langston L, Stillman B. Principles and concepts of DNA replication in bacteria, archaea, and eukarya. *Cold Spring Harb Perspect Biol* 2013;**5**:a010108.
3. Stillman B. Origin recognition and the chromosome cycle. *FEBS Lett* 2005;**579**(4):877–84.
4. Mott ML, Berger JM. DNA replication initiation: mechanisms and regulation in bacteria. *Nat Rev Microbiol* 2007;**5**(5):343–54.
5. Skarstad K, Katayama T. Regulating DNA replication in bacteria. *Cold Spring Harb Perspect Biol* 2013;**5**(4):a012922.
6. Gao F. Recent advances in the identification of replication origins based on the Z-curve method. *Curr Genomics* 2014;**15**(2):104–12.
7. Grigoriev A. Analyzing genomes with cumulative skew diagrams. *Nucleic Acids Res* 1998;**26**(10):2286–90.
8. Gao F, Luo H, Zhang CT. DeOri: a database of eukaryotic DNA replication origins. *Bioinformatics* 2012;**28**(11):1551–2.
9. Leonard AC, Méchali M. DNA replication origins. *Cold Spring Harbor Perspect Biol* 2013;**5**(10):a010116.
10. Liu F, Ren C, Li H, *et al*. *De novo* identification of replication-timing domains in the human genome by deep learning. *Bioinformatics* 2016;**32**(5):641–9.
11. Lei M. The MCM complex: its role in DNA replication and implications for cancer therapy. *Curr Cancer Drug Targets* 2005;**5**(5):365–80.
12. Peng C, Luo H, Zhang X, *et al*. Recent advances in the genome-wide study of DNA replication origins in yeast. *Front Microbiol* 2015;**6**:117.
13. Xu J, Yanagisawa Y, Tsankov AM, *et al*. Genome-wide identification and characterization of replication origins by deep sequencing. *Genome Biol* 2012;**13**(4):R27.
14. Sasaki T, Gilbert DM. Unearthing worm replication origins. *Nat Struct Mol Biol* 2017;**24**(3):195–6.
15. Cayrou C, Ballester B, Peiffer I, *et al*. The chromatin environment shapes DNA replication origin organization and defines origin classes. *Genome Res* 2015;**25**(12):1873–85.
16. Costas C, Sanchez MD, Stroud H, *et al*. Genome-wide mapping of Arabidopsis thaliana origins of DNA replication and their associated epigenetic marks. *Nat Struct Mol Biol* 2011;**18**(3):395–400.

17. Wu ZF, Liu JF, Yang HB, *et al*. DNA replication origins in archaea. *Front Microbiol* 2014;**5**:179.
18. Barry ER, Bell SD. DNA replication in the archaea. *Microbiol Mol Biol Rev* 2006;**70**(4):876–87.
19. Samson RY, Abeyrathne PD, Bell SD. Mechanism of archaeal MCM helicase recruitment to DNA replication origins. *Mol Cell* 2016;**61**(2):287–96.
20. Fleischmann RD, Adams MD, White O, *et al*. Whole-genome random sequencing and assembly of *Haemophilus influenzae* Rd. *Science* 1995;**269**(5223):496–512.
21. Bouchie A. White house unveils national microbiome initiative. *Nat Biotechnol* 2016;**34**(6):580–1.
22. Ommen B, El-Sohemy A, Hesketh J, *et al*. The Micronutrient Genomics Project: a community-driven knowledge base for micronutrient research. *Genes Nutr* 2010;**5**:285.
23. Ehrlich SD, Consortium M. MetaHIT: the European Union Project on metagenomics of the human intestinal tract. In: *Metagenomics of the Human Body*. New York, NY: Springer, 2011, 307–16.
24. Collins FS, Morgan M, Patrinos A. The Human Genome Project: lessons from large-scale biology. *Science* 2003; **300**(5617):286–90.
25. Gao F. Editorial: DNA replication origins in microbial genomes. *Front Microbiol* 2015;**6**:1545.
26. Zhang R, Zhang C-T. A brief review: the Z-curve theory and its application in genome analysis. *Curr Genomics* 2014;**15**(2):78–94.
27. Zhang R, Zhang CT. Identification of replication origins in archaeal genomes based on the Z-curve method. *Archaea* 2005;**1**(5):335–46.
28. Soppa J. From genomes to function: haloarchaea as model organisms. *Microbiology* 2006;**152**(Pt 3):585–90.
29. Gao F, Zhang CT. Ori-Finder: a web-based system for finding *oriC*s in unannotated bacterial genomes. *BMC Bioinformatics* 2008;**9**(1):79.
30. Korem T, Zeevi D, Suez J, *et al*. Microbiome growth dynamics of gut microbiota in health and disease inferred from single metagenomic samples. *Science* 2015;**349**(6252):1101–6.
31. Xu YX, Ji XF, Chen N, *et al*. Development of replicative *oriC* plasmids and their versatile use in genetic manipulation of Cytophaga hutchinsonii. *Appl Microbiol Biotechnol* 2012;**93**(2): 697–705.
32. Makowski L, Donczew R, Weigel C, *et al*. Initiation of chromosomal replication in predatory bacterium bdellovibrio bacteriovorus. *Front Microbiol* 2016;**7**:1898.
33. Luo H, Zhang CT, Gao F. Ori-Finder 2, an integrated tool to predict replication origins in the archaeal genomes. *Front Microbiol* 2014;**5**:482.
34. Gao F, Zhang CT. DoriC: a database of *oriC* regions in bacterial genomes. *Bioinformatics* 2007;**23**(14):1866–7.
35. Gao F, Luo H, Zhang C-T. DoriC 5.0: an updated database of *oriC* regions in both bacterial and archaeal genomes. *Nucleic Acids Res* 2013;**41**:D90–3.
36. Xia X. DNA replication and strand asymmetry in prokaryotic and mitochondrial genomes. *Curr Genomics* 2012;**13**(1):16–27.
37. Salzberg SL, Salzberg AJ, Kerlavage AR, *et al*. Skewed oligomers and origins of replication. *Gene* 1998;**217**(1–2):57–67.
38. Mackiewicz P, Zakrzewska CJ, Zawilak A, *et al*. Where does bacterial replication start? Rules for predicting the *oriC* region. *Nucleic Acids Res* 2004;**32**(13):3781–91.
39. Sernova NV, Gelfand MS. Identification of replication origins in prokaryotic genomes. *Brief Bioinform* 2008;**9**(5):376–91.
40. Gao F, Zhang C-T. Origins of replication in *Sorangium cellulosum* and *Microcystis aeruginosa*. *DNA Res* 2008;**15**(3):169–71.
41. Gao F, Zhang C-T. Origins of replication in Cyanothece 51142. *Proc Natl Acad Sci USA* 2008;**105**(52):E125.
42. Kurtz S, Choudhuri JV, Ohlebusch E, *et al*. REPuter: the manifold applications of repeat analysis on a genomic scale. *Nucleic Acids Res* 2001;**29**(22):4633–42.
43. Tyner C, Barber GP, Casper J, *et al*. The UCSC genome browser database: 2017 update. *Nucleic Acids Res* 2017;**45**(D1):D626–34.
44. Coordinators NR. Database resources of the national center for biotechnology information. *Nucleic Acids Res* 2017;**45**: D12–17.
45. Woese CR, Kandler O, Wheelis ML. Towards a natural system of organisms: proposal for the domains Archaea, Bacteria, and Eucarya. *Proc Natl Acad Sci USA* 1990;**87**(12):4576–9.
46. Myllykallio H, Lopez P, Lopez GP, *et al*. Bacterial mode of replication with eukaryotic-like machinery in a hyperthermophilic archaeon. *Science* 2000;**288**(5474):2212–15.
47. Wu Z, Liu J, Yang H, *et al*. Multiple replication origins with diverse control mechanisms in *Haloarcula hispanica*. *Nucleic Acids Res* 2014;**42**(4):2282–94.
48. Bailey TL, Boden M, Buske FA, *et al*. MEME SUITE: tools for motif discovery and searching. *Nucleic Acids Res* 2009;**37**: W202–8.
49. Wang Y, Chen B, Sima L, *et al*. Construction of expression shuttle vectors for the *Haloarchaeon Natrinema* sp. *J7* based on its chromosomal origins of replication. *Archaea* 2017;**2017**: 4237079.
50. Papadimitriou K, Baharidis PK, Georgoulis A, *et al*. Analysis of the complete genome sequence of the archaeon *Pyrococcus chitonophagus* DSM 10152 (formerly *Thermococcus chitonophagus*). *Extremophiles* 2016;**20**(3):351–361.
51. Gavrilov SN, Stracke C, Jensen K, *et al*. Isolation and characterization of the first xylanolytic hyperthermophilic *Euryarchaeon thermococcus* sp strain 2319x1 and its unusual multidomain glycosidase. *Front Microbiol* 2016;**7**:552.
52. Demina TA, Atanasova NS, Pietila MK, *et al*. Vesicle-like virion of *Haloarcula hispanica pleomorphic* virus 3 preserves high infectivity in saturated salt. *Virology* 2016;**499**:40–51.
53. Bharathi M, Chellapandi P. Intergenomic evolution and metabolic cross-talk between rumen and thermophilic autotrophic methanogenic archaea. *Mol Phylogenet Evol* 2017;**107**: 293–304.
54. Guo FB, Ou HY, Zhang CT. ZCURVE: a new system for recognizing protein-coding genes in bacterial and archaeal genomes. *Nucleic Acids Res* 2003;**31**(6):1780–9.
55. Delcher AL, Bratke KA, Powers EC, *et al*. Identifying bacterial genes and endosymbiont DNA with Glimmer. *Bioinformatics* 2007;**23**(6):673–9.
56. Grant CE, Bailey TL, Noble WS. FIMO: scanning for occurrences of a given motif. *Bioinformatics* 2011;**27**(7):1017–18.
57. Frank AC, Lobry JR. Oriloc: prediction of replication boundaries in unannotated bacterial chromosomes. *Bioinformatics* 2000;**16**(6):560–1.
58. Wolanski M, Donczew R, Zawilak-Pawlik A, *et al*. *oriC*-encoded instructions for the initiation of bacterial chromosome replication. *Front Microbiol* 2015;**5**:735.
59. Zhou Y, Chen WL, Wang L, *et al*. Identification of the *oriC* region and its influence on heterocyst development in the filamentous *Cyanobacterium anabaena* sp. strain PCC 7120. *Microbiology* 2011;**157**:1910–19.
60. Watanabe S, Ohbayashi R, Shiwa Y, *et al*. Light-dependent and asynchronous replication of cyanobacterial multi-copy chromosomes. *Mol Microbiol* 2012;**83**(4):856–65.
61. Huang H, Song CC, Yang ZL, *et al*. Identification of the replication origins from cyanothece ATCC 51142 and their

interactions with the DnaA protein: from *in silico* to *in vitro* studies. *Front Microbiol* 2015;**6**:1370.
62. Stolz A. Degradative plasmids from sphingomonads. *FEMS Microbiol Lett* 2014;**350**(1):9–19.
63. Richardson TT, Harran O, Murray H. The bacterial DnaA-trio replication origin element specifies single-stranded DNA initiator binding. *Nature* 2016;**534**(7607):412–16.
64. Bendall ML, Luong K, Wetmore KM, *et al*. Exploring the roles of DNA methylation in the metal-reducing bacterium *Shewanella oneidensis* MR-1. *J Bacteriol* 2013;**195**(21):4966–74.
65. Wang X, Lesterlin C, Reyes-Lamothe R, *et al*. Replication and segregation of an *Escherichia coli* chromosome with two replication origins. *Proc Natl Acad Sci USA* 2011;**108**(26):E243–50. 2
66. Liang X, Baek CH, Katzen F. *Escherichia coli* with two linear chromosomes. *ACS Synth Biol* 2013;**2**(12):734–40.
67. Messerschmidt SJ, Kemter FS, Schindler D, *et al*. Synthetic secondary chromosomes in Escherichia coli based on the replication origin of chromosome II in *Vibrio cholerae*. *Biotechnol J* 2015;**10**:302–14.
68. Gao F. Bacteria may have multiple replication origins. *Front Microbiol* 2015;**6**:324.
69. Ionescu D, Bizic-Ionescu M, De Maio N, *et al*. Community-like genome in single cells of the sulfur bacterium *Achromatium oxaliferum*. *Nat Commun* 2017;**8**(1):455.
70. Ohbayashi R, Watanabe S, Ehira S, *et al*. Diversification of DnaA dependency for DNA replication in cyanobacterial evolution. *ISME J* 2016;**10**(5):1113–21.
71. Zhao HL, Xia ZK, Zhang FZ, *et al*. Multiple factors drive replicating strand composition bias in bacterial genomes. *Int J Mol Sci* 2015;**16**(9):23111–26.
72. Zhang G, Gao F. Quantitative analysis of correlation between AT and GC biases among bacterial genomes. *PLoS One* 2017; **12**(2):e0171408.
73. Chen WH, Lu G, Bork P, *et al*. Energy efficiency trade-offs drive nucleotide usage in transcribed regions. *Nat Commun* 2016;**7**: 11334.
74. Banerjee R, Roy D. Codon usage and gene expression pattern of *Stenotrophomonas maltophilia* R551-3 for pathogenic mode of living. *Biochem Biophys Res Commun* 2009;**390**(2):177–81.
75. Guo FB, Yuan JB. Codon usages of genes on chromosome, and surprisingly, genes in plasmid are primarily affected by strand-specific mutational biases in *Lawsonia intracellularis*. *DNA Res* 2009;**16**(2):91–104.
76. Khrustalev VV, Barkovsky EV. The probability of nonsense mutation caused by replication-associated mutational pressure is much higher for bacterial genes from lagging than from leading strands. *Genomics* 2010;**96**(3):173–180.
77. Sobetzko P, Travers A, Muskhelishvili G. Gene order and chromosome dynamics coordinate spatiotemporal gene expression during the bacterial growth cycle. *Proc Natl Acad Sci USA* 2012;**109**(2):E42–50.
78. Lin Y, Gao F, Zhang CT. Functionality of essential genes drives gene strand-bias in bacterial genomes. *Biochem Biophys Res Commun* 2010;**396**(2):472–76.
79. Gao N, Lu G, Lercher MJ, *et al*. Selection for energy efficiency drives strand-biased gene distribution in prokaryotes. *Sci Rep* 2017;**7**:10572.
80. 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**(17):8210–18.
81. Rocha EP, Danchin A. Essentiality, not expressiveness, drives gene-strand bias in bacteria. *Nat Genet* 2003;**34**(4):377.
82. Wei W, Ning LW, Ye YN, *et al*. Geptop: a gene essentiality prediction tool for sequenced bacterial genomes based on orthology and phylogeny. *PLoS One* 2013;**8**(8):e72343.
83. Lin Y, Zhang RR. Putative essential and core-essential genes in Mycoplasma genomes. *Sci Rep* 2011;**1**(1):53.