Recent Advances in the Identification of Replication Origins Based on the Z-curve Method

Feng Gao*

Department of Physics, Tianjin University, Tianjin 300072, China

Abstract: Precise DNA replication is critical for the maintenance of genetic integrity in all organisms. In all three domains of life, DNA replication starts at a specialized locus, termed as the replication origin, oriC or ORI, and its identification is vital to understanding the complex replication process. In bacteria and eukaryotes, replication initiates from single and multiple origins, respectively, while archaea can adopt either of the two modes. The Z-curve method has been successfully used to identify replication origins in genomes of various species, including multiple oriC regions in some archaea. Based on the Z-curve method and comparative genomics analysis, we have developed a web-based system, Ori-Finder, for finding oriC in bacterial genomes with high accuracy. Predicted oriC regions in bacterial genomes are organized into an online database, DoriC. Recently, archaeal oriC regions identified by both in vivo and in silico methods have also been included in the database. Here, we summarize the recent advances of in silico prediction of oriC in bacterial and archaeal genomes using the Z-curve based method.

Received on: September 25, 2013- Revised on: November 04, 2013- Accepted on: November 05, 2013

Keywords: Archaea, Bacteria, Comparative genomics, Origins of replication, Z-curve.

1. INTRODUCTION

In 1963, Jacob, Brenner, and Cuzin proposed the replicon model, in which the replicon was defined as the fundamental unit of replication [1]. The initiator protein (bacterial DnaA or archaeal Orc1/Cdc6) binds a sequence (bacterial DnaA box or archaeal ORB element) within a replicon called a replicator, and then DNA synthesis initiates from a specific site, called origin of replication [1]. The events that occur at the replication origin (oriC or ORI) are central to the process of regulating DNA replication and the cell cycle. Therefore, it is important to precisely identify the replication origins within the analyzed genomes. This critical information allows us to better understand not only the structure and function of the replication origins, but also the mechanisms of DNA replication [2, 3].

The oriC regions can be identified by several experimental methods including construction of replicative oriC plasmids [4, 5], microarray-based [6] or high-throughput sequencing-based [7] marker frequency analysis, and two-dimensional gel electrophoresis analysis [8]. The experimental methods for identifying replication origins in vivo are reliable, but time-consuming and labor-intensive. The identification of replication origins based on in situ analysis has been the subject of intensive study in the last two decades. The pioneer work to identify oriC in vivo is the GC-skew analysis [9, 10], and the cumulative GC-skew was later proposed to provide better resolution [11]. An oligomer-skew method was also proposed to predict oriC regions in bacterial genomes [12]. The same method was later used to identify oriCs in more than 200 prokaryotic chromosomes [13]. Use of GC-skew analysis, together with the location of the dnaA gene and distribution of DnaA boxes led to more accurate prediction of oriC regions [14].

The Z-curve method was developed in 1994 as a way to display base composition distributions along DNA sequences [15]. The x and y components of the Z-curve are related to distributions of RY (purine/pyrimidine) and MK (amino/keto), as well as GC and AT bases, and can be used to identify oriC regions in bacterial and archaeal genomes [16]. For instance, Z-curve analysis predicted single oriC in the archaeal genomes of Methanosarcina mazei Go1 [17] and Methanocaldococcus jannaschii DSM 2661 [18], two oriCs in Halobacterium species NRC-1 [19], and three oriCs in Sulfolobus solfataricus P2 genome [19], and these prediction were consistent with later in vivo experimental evidence, e.g., that obtained in studies of Halobacterium species NRC-1 [20, 21] and Sulfolobus solfataricus P2 genome [6, 8].

Based on the Z-curve method, a web-based system, Ori-Finder [22], has been developed to find oriCs in over 2,000 bacterial genomes including Sorangium cellulosum 'So ce 56', Micocystis aeruginosa NIES-843 [23] and Cyanobacterium 51142 [24]. The predicted oriC regions have been organized into DoriC [25], a database of oriC regions in bacterial genomes. Recently, the database has been updated to include the oriC regions in archaeal genomes [26].

With the advent of the post-genomic era, genomic data accumulation has been increasing exponentially [27]. However, locations of a large number of oriCs in sequenced bacterial and archaeal genomes still remain unknown. This has created challenges as well as opportunities for identifying
these oriCs by in silico analysis. Clarification of the archaeal replication mechanism is particularly important, as it may provide insight into the replication mechanisms of eukary.

2. METHODS

2.1. Z-curve and RY, MK, AT or GC Disparity Curves

The Z-curve is a three-dimensional curve that constitutes a unique representation of a DNA sequence, such that the Z-curve and the given DNA sequence can each be uniquely reconstructed from the other [15]. The three components of the Z-curve, $x_n$, $y_n$ and $z_n$, represent three independent distributions that completely describe the DNA sequence being studied. The components $x_n$, $y_n$ and $z_n$ display the distributions of purine versus pyrimidine (R vs. Y), amino versus keto (M vs. K) and strong H-bond versus weak H-bond (S vs. W) bases, respectively, along the DNA sequence. The $x_n$ and $y_n$ components are termed RY and MK disparity curves, respectively. The AT and GC disparity curves are defined by $(x_n + y_n)/2$ and $(x_n - y_n)/2$, which show the excess of A over T and G over C respectively, along the genome. The RY and MK disparity curves, as well as the AT and GC disparity curves, can be used to predict replication origins [16]. For instance, Z-curves (that is, RY, MK, AT and GC disparity curves) show a single oriC in the genome of the bacterium of Cyanothece sp. PCC 7425 (Fig. 1A) and one, two, three oriCs in genomes of the archaea of Pyrococcus abyssi GE5, Halobacterium sp. NRC-1, and Sulfolobus acidocaldarius DSM 639, respectively (Fig. 1B-D).

2.2. Ori-Finder and DoriC

Ori-Finder is an online system for finding oriCs in bacterial genomes based on an integrated method involving the analysis of base composition asymmetry using the Z-curve method, distribution of DnaA boxes, and the occurrence of genes frequently adjacent to oriCs. Currently, Ori-Finder version 1.0 is designed only for the identification of oriCs in bacterial genomes, which is available at http://tubic.tju.edu.cn/Ori-Finder/. Ori-Finder has been used to analyze roughly 50 newly sequenced bacterial genomes, such as Corynebacterium pseudotuberculosis FRC41 [28], Orientia tsutsugamushi Ikeda [29], Bacillus pseudofirmus OF4 [30], Klebsiella pneumoniae subsp. pneumoniae HS11286 [31], Streptococcus parasanguinis FW213 [32],

![Fig. (1)](image_url)

Fig. (1). RY, MK, AT and GC disparity curves reveal oriC locations in bacterial and archaeal genomes. Z-curves show a single oriC in the genome of the bacterium of Cyanothece sp. PCC 7425 (A) and one, two, three oriCs in genomes of the archaea of Pyrococcus abyssi GE5 (B), Halobacterium sp. NRC-1 (C), and Sulfolobus acidocaldarius DSM 639 (D), respectively. Note that the Z-curves have been drawn for the rotated sequences beginning and ending in the maximum of the GC disparity curves. Short vertical line indicates dnaA or cdc6 gene location, and short up vertical arrow indicates the identified oriC location.
Acinetobacter baumannii MDR-TJ [33], Streptococcus infantarius subsp. infantarius CJ18 [34] and Streptococcus equi ssp. zooepidemicus strain ATCC35246 [35].

The oriC regions predicted by Ori-Finder in bacterial genomes have been organized into an online database, DoriC, which has been publicly available at http://tubic.tju.edu.cn/doric since 2007. Six years after we constructed DoriC, the database has made significant advances in the number of bacterial genomes available, increasing about four-fold. Additionally, oriC regions in archaeal genomes identified by in vivo experiments as well as in silico analyses have been added to the database. Consequently, the latest release of DoriC 6.5 contains oriCs for more than 2,000 bacterial genomes and 100 archaeal genomes. Each entry contains detailed information about the oriC, such as the sequence, repeat, DnaA box or ORB motif, and graphical representations of the oriC, such as the various disparity curves (RY, MK, AT and GC). Users can browse the database by species name, or accession numbers of GenBank or DoriC, can search for oriCs by the organism’s name, accession number, lineage, or a keyword, and can also explore the genomic context around the oriC regions via NCBI Map Viewer or UCSC Archaeal Genome Browser by clicking the corresponding links provided by DoriC. In addition, users can select the ‘BLAST’ option to compare a query sequence or even a whole genome against DoriC to find homologous oriCs. DoriC has been widely used as a source of data in comparative genomics analysis [36-42].

3. RESULTS AND DISCUSSION

3.1. Replication Origins in Cyanobacteria

Based on DoriC, the relationships between the conserved features associated with the oriC regions, such as adjacent genes and DnaA boxes, and the taxonomic levels of the corresponding bacteria can be summarized. For example, detailed analyses have shown that the consensus sequence of the DnaA boxes in oriC regions, and the distribution of genes around oriCs, are strongly conserved among the bacteria in the phylum cyanobacteria [24]. The position of the oriC, adjacent to dnaN gene which encodes the beta clamp processivity factor, has been found to be universal among the bacteria within the phylum cyanobacteria. The ‘species-specific’ DnaA box motif for the phylum cyanobacteria is ‘TTTTCCACA’ instead of ‘TTATCCACA’, the DnaA box motif of Escherichia coli [43]. These strongly conserved features indicate that the in silico identified oriCs are reliable, as they have been confirmed by comparative genomics approaches. As we expected, the experimentally confirmed replication origins of Anabaena sp. PCC 7120 [4] and S. elongatus PCC 7942 [44] in the phylum cyanobacteria are all adjacent to the dnaN gene.

Recently, coverage of the cyanobacterial phylum has improved significantly using diversity-driven genome sequencing [45], and some exceptions to the proposed rules have been uncovered in the process. For example, a cluster of DnaA boxes with perfect matches to the motif ‘TTTTCCACA’ has been found adjacent to dnaA gene instead of dnaN gene in Anabaena sp. 90, Geitlerinema sp. PCC 7407 and Synechococcus sp. PCC 6312. For Dactylococcopsis salina PCC 8305, Halothece sp. PCC 7418, Lepolyngbya sp. PCC 7376 and Thermosynechococcus elongatus BP-1, a cluster of DnaA boxes with perfect matches to the motif ‘TTTCCACA’ has been found adjacent to neither dnaA nor dnaN (Table 1). Perhaps the ancestral position of the replication origins in the phylum cyanobacteria was within the dnaA-dnaN intergenic region, and the translocation of the dnaA or dnaN gene from the putative origin of replication to another place on the chromosome has led to some origins linked only to dnaN or dnaA gene. If the oriC region instead of dnaA or dnaN gene had translocated away from its ancestral position, origins would be linked to neither dnaA nor dnaN genes.

3.2. Replication Origins in Some Intracellular Bacteria

Some bacteria are intracellular parasites or symbionts. Recently, the genome of Blattabacterium cuenoti, primary endosymbiont of the omnivorous cockroach Blatta orientalis, has been completely sequenced [46]. In their report, Patiño-Navarrete et al. concluded that ‘Similar to previously sequenced Blattabacterium strains, the strain from Blatta orientalis does not possess any features determining replication origin.’ Based on the results of Ori-Finder and DoriC in the genomes of Blattabacterium strains, we have identified candidate oriC regions which are adjacent to the gidA gene encoding glucose-inhibited division protein A. They contain putative DnaA boxes and repeat elements. The location of oriCs adjacent to the gidA gene, is common among intracellular bacteria such as secondary endosymbiont of Heterosylly cubana, secondary endosymbiont of Cienaraytna eucalypti, Wigglesworthia glossinidia endosymbiont of Glossina morsitans morsitans (Yale colony), and Wigglesworthia glossinidia endosymbiont of Glossina brevipalpis. However, for Wolbachia endosymbionts (Wolbachia endosymbiont of Drosophila melanogaster, Wolbachia endosymbiont strain TRS of Brugia malayi, Wolbachia pipiensis, Wolbachia sp. wr1, and Wolbachia endosymbiont of Onchocerca ochengi), we have identified candidate oriC regions which are adjacent to the hemE gene encoding uroporphyrinogen decarboxylase.

The replication origin of Orienteria tsutsugamushi, an obligate intracellular bacterium belonging to the family Rickettsiaceae, is also predicted to be adjacent to the hemE gene by Ori-Finder. For Mollicutes whose genomes underwent considerable reduction because of a parasitic style of life, the oriCs are adjacent to dnaA gene. Interestingly, for Chlamydiae, a phylum of bacteria whose members are obligate intracellular pathogens, oriCs are adjacent to the hemB gene encoding delta-aminolevulinic acid dehydratase instead of dnaA gene, although two dnaA genes are contained in their genomes according to annotations in GenBank.

3.3. Multiple Replication Origins in Pyrobaculum calidifontis JCM 11548

The number of oriCs in archaea has been found to correlate with the phylogeny. For example, all the archaea within the phylum Crenarchaeota examined to date contain multiple origins [7]. Recently, four chromosome replication origins in the archaean Pyrobaculum calidifontis JCM 11548 have been mapped by using high-throughput sequencing-based marker frequency analysis [7]. However, only one
Table 1. The statistics of adjacent genes for the bacteria in the phylum Cyanobacteria.

| RefSeq   | Organism                             | Lineage                      | Adjacent Genes |
|----------|--------------------------------------|------------------------------|----------------|
| NC_009925 | Acaryochloris marina MBIC11017       | Acaryochloris                | dnaA, dnaN     |
| NC_008312 | Trichodesmium erythroaum IMS101      | Oscillatoriales, Trichodesmium | dnaA, dnaN     |
| NC_019776 | Cyanobacterium aponinum PCC 10605<sup>a</sup> | Chroococcales, Cyanobacterium | dnaN           |
| NC_019778 | Cyanobacterium staniier PCC 7202<sup>b</sup> | Chroococcales, Cyanobacterium | dnaN           |
| NC_013771 | Cyanobacterium UCYN-A<sup>c</sup>     | Chroococcales                 | dnaN           |
| NC_019675 | Cyanobium gracile PCC 6307           | Chroococcales, Cyanobium     | dnaN           |
| NC_010546 | Cyanotheta sp. ATCC 51142<sup>a</sup> | Chroococcales, Cyanotheta    | dnaN           |
| NC_011729 | Cyanotheta sp. PCC 7424              | Chroococcales, Cyanotheta    | dnaN           |
| NC_011884 | Cyanotheta sp. PCC 7425              | Chroococcales, Cyanotheta    | dnaN, dnaA     |
| NC_014501 | Cyanotheta sp. PCC 7822              | Chroococcales, Cyanotheta    | dnaN           |
| NC_011726 | Cyanotheta sp. PCC 8801              | Chroococcales, Cyanotheta    | dnaN           |
| NC_013161 | Cyanotheta sp. PCC 8802              | Chroococcales, Cyanotheta    | dnaN           |
| NC_019780 | Dactylococcopsis salina PCC 8305     | Chroococcales, Dactylococcopsis | others        |
| NC_019779 | Halotheca sp. PCC 7418               | Chroococcales, Halotheca cluater, Halotheca | others        |
| NC_010296 | Microcystis aeruginosa NIES-843      | Chroococcales, Microcystis   | dnaN           |
| NC_006576 | Synechococcus elongatus PCC 6301     | Chroococcales, Synechococcus | dnaN           |
| NC_007604 | Synechococcus elongatus PCC 7942     | Chroococcales, Synechococcus | dnaN           |
| NC_008319 | Synechococcus sp. CC9311             | Chroococcales, Synechococcus | dnaN           |
| NC_007516 | Synechococcus sp. CC9605             | Chroococcales, Synechococcus | dnaN           |
| NC_007513 | Synechococcus sp. CC9902             | Chroococcales, Synechococcus | dnaN           |
| NC_007776 | Synechococcus sp. JA-2-3B'a(2-13)    | Chroococcales, Synechococcus | dnaN           |
| NC_007775 | Synechococcus sp. JA-3-3Ab            | Chroococcales, Synechococcus | dnaN           |
| NC_019680 | Synechococcus sp. PCC 6312           | Chroococcales, Synechococcus | dnaA           |
| NC_010475 | Synechococcus sp. PCC 7002           | Chroococcales, Synechococcus | dnaN           |
| NC_019702 | Synechococcus sp. PCC 7502           | Chroococcales, Synechococcus | dnaN           |
| NC_009482 | Synechococcus sp. RCC307             | Chroococcales, Synechococcus | dnaA           |
| NC_009481 | Synechococcus sp. WH 7803            | Chroococcales, Synechococcus | dnaN           |
| NC_005070 | Synechococcus sp. WH 8102            | Chroococcales, Synechococcus | dnaN           |
| NC_000911 | Synechocystis sp. PCC 6803           | Chroococcales, Synechocystis | dnaN           |
| NC_017277 | Synechocystis sp. PCC 6803           | Chroococcales, Synechocystis | dnaN           |
| NC_017038 | Synechocystis sp. PCC 6803 substr. GT-I | Chroococcales, Synechocystis | dnaN           |
| NC_017052 | Synechocystis sp. PCC 6803 substr. PCC-N | Chroococcales, Synechocystis | dnaN           |
| NC_017039 | Synechocystis sp. PCC 6803 substr. PCC-P | Chroococcales, Synechocystis | dnaN           |
| NC_004113 | Thermosynechococcus elongatus BP-1   | Chroococcales, Thermosynechococcus | others        |
| NC_005125 | Gloeobacter violaceus PCC 7421      | Gloeobacteria, Gloeobacterales, Gloeobacter | dnaN           |
| NC_019427 | Anabaena sp. 90<sup><s>9</s></sup>   | Nostocales, Nostocaceae, Anabaena | dnaA           |
| NC_007413 | Anabaena variabilis ATCC 29413      | Nostocales, Nostocaceae, Anabaena | dnaA, dnaN     |
(Table 1) contd....

| RefSeq   | Organism                                   | Lineage                                  | Adjacent Genes |
|----------|--------------------------------------------|------------------------------------------|----------------|
| NC_014248 | ‘Nostoc azollae’ 0708                      | Nostocales, Nostocaceae, Anabaena         | dnaA, dnaN     |
| NC_010628 | Nostoc punctiforme PCC 73102               | Nostocales, Nostocaceae, Nostoc          | dnaA          |
| NC_019676 | Nostoc sp. PCC 7107                       | Nostocales, Nostocaceae, Nostoc          | dnaA, dnaN    |
| NC_003272 | Nostoc sp. PCC 7120                       | Nostocales, Nostocaceae, Nostoc          | dnaA, dnaN    |
| NC_019684 | Nostoc sp. PCC 7524                       | Nostocales, Nostocaceae, Nostoc          | dnaA, dnaN    |
| NC_019751 | Calothrix sp. PCC 6303                     | Nostocales, Rivulariaceae, Calothrix     | dnaA          |
| NC_019682 | Calothrix sp. PCC 7507                     | Nostocales, Rivulariaceae, Calothrix     | dnaA, dnaN    |
| NC_019678 | Rivularia sp. PCC 7116                     | Nostocales, Rivulariaceae, Rivularia     | dnaA, dnaN    |
| NC_019753 | Crinalium epipsammum PCC 9333             | Oscillatoriales, Crinalium               | dnaA, dnaN    |
| NC_019703 | Geitlerinema sp. PCC 7407                  | Oscillatoriales, Geitlerinema             | dnaA          |
| NC_019683 | Leptolyngbya sp. PCC 7376                 | Oscillatoriales, Leptolyngbya            | others        |
| NC_019738 | Microcoleus sp. PCC 7113                  | Oscillatoriales, Microcoleus             | dnaA, dnaN    |
| NC_019693 | Oscillatoria acuminata PCC 6304           | Oscillatoriales, Oscillatoria            | dnaA, dnaN    |
| NC_019729 | Oscillatoria nigro-viridis PCC 7112       | Oscillatoriales, Oscillatoria            | dnaA, dnaN    |
| NC_019701 | Pseudanabaena sp. PCC 7367                | Oscillatoriales, Pseudanabaena           | dnaN          |
| NC_019695 | Chroococcidiopsis thermalis PCC 7203      | Pleurocapsales, Chroococcidiopsis        | dnaA          |
| NC_019689 | Pleurocapsa sp. PCC 7327                  | Pleurocapsales, Pleurocapsa              | dnaA, dnaN    |
| NC_008816 | Prochlorococcus marinus str. AS9601       | Prochlorales, Prochlorococcaceae, Prochlorococcus | dnaN |
| NC_009976 | Prochlorococcus marinus str. MIT 9211     | Prochlorales, Prochlorococcaceae, Prochlorococcus | dnaA |
| NC_009840 | Prochlorococcus marinus str. MIT 9215     | Prochlorales, Prochlorococcaceae, Prochlorococcus | dnaA |
| NC_009901 | Prochlorococcus marinus str. MIT 9301     | Prochlorales, Prochlorococcaceae, Prochlorococcus | dnaA |
| NC_008820 | Prochlorococcus marinus str. MIT 9303     | Prochlorales, Prochlorococcaceae, Prochlorococcus | dnaA |
| NC_007577 | Prochlorococcus marinus str. MIT 9312     | Prochlorales, Prochlorococcaceae, Prochlorococcus | dnaA |
| NC_005071 | Prochlorococcus marinus str. MIT 9313     | Prochlorales, Prochlorococcaceae, Prochlorococcus | dnaA |
| NC_008817 | Prochlorococcus marinus str. MIT 9515     | Prochlorales, Prochlorococcaceae, Prochlorococcus | dnaA |
| NC_008819 | Prochlorococcus marinus str. NATL1A       | Prochlorales, Prochlorococcaceae, Prochlorococcus | dnaA |
| NC_007335 | Prochlorococcus marinus str. NATL2A       | Prochlorales, Prochlorococcaceae, Prochlorococcus | dnaA |
| NC_005042 | Prochlorococcus marinus subspp. marinus str. CCMP1375 | Prochlorales, Prochlorococcaceae, Prochlorococcus | dnaA |
| NC_005072 | Prochlorococcus marinus subspp. pastoris str. CCMP1986 | Prochlorales, Prochlorococcaceae, Prochlorococcus | dnaA |

*Note that no dnaA gene is annotated in these genomes.

*Note that only the chromosome 1 (I) or chromosome circular was counted if the bacterium has multiple chromosomes.

*Note that the oriC region is about 5 kb away from the dnaN gene.

The location of oriC, flanked by tRNA genes, is universal among the archaea in the class Thermoprotei within the phylum Crenarchaeota. For example, we found that the origins were adjacent to tRNA genes in Sulfolobus solfataricus P2, Sulfolobus tokodaii str. 7, Sulfolobus acidocaldarius DSM 639, Sulfolobus islandicus Y.N.15.51, Sulfolobus solfataricus 98/2, Metallosphaera cuprina Ar-4, Acidianus hospitalis W1, and Thermofilum pendens Hrk 5. Based on this conserved feature, the other three putative origins of replication in Pyrococulium caldiphon JCM 11548 have been identified at the sequence level (Fig. 2).
The putative oriC2 is within an intergenic region between the gene Pcal_0541 and Pcal_0542, from 514,406 nt to 514,741 nt (Fig. 2B). The putative oriC3 is within an intergenic region between the gene Pcal_1006 and Pcal_1007, from 950,832 nt to 951,332 nt (Fig. 2C). The putative oriC4 is within an intergenic region between the gene Pcal_1820 and Pcal_1821, from 1,687,883 nt to 1,688,541 nt (Fig. 2D).

Among the predicted oriCs, the putative oriC2 shares a long sequence, ‘atcggcgccccgggcttgctgaatccggttcggtc’, with the putative oriC3. These three putative oriC regions all contain a 13-mer consensus element, ‘GGGTTCAAATCCC’, which has also been found in the oriCs of closely-related species such as Sulfolobus solfataricus P2, Acidimianus hospitalis W1, and Metallosphaera cuprina Ar-4. We also found that the putative oriC2 and oriC3 share a common sequence, ‘gccggggtggccgagcggcccaaggcg’, with the putative origin of Thermofilum pendens Hrk 5, and the putative oriC4 shares a sequence, ‘atcggcggggggttcggtc’, with the origins of Sulfolobus solfataricus P2, Acidimianus hospitalis W1, and Metallosphaera cuprina Ar-4.

Some conserved genes associated with oriCs, such as copG gene encoding plasmid copy number control protein, were also found around the predicted oriCs. The replication origin was flanked by tRNA gene and copG gene, which could play a fundamental role in shaping the origin-containing loci [47]. Around the putative oriC2, there is a tRNA-Ser gene (514,425..514,522 nt) recognizing UCA codons and a gene Pcal_0536 (510,064..510,240 nt) encoding CopG/Arc/MetJ family transcriptional regulator. Around the putative oriC3, there is a tRNA-Ser gene (951,001..951,098) recognizing UCC codons and a gene Pcal_1012 (953,989..954,357 nt) encoding CopG family transcriptional regulator. Around the putative oriC4, there is a tRNA-Cys gene (1,687,978..1,688,071 nt) recognizing UGC codons. Therefore, these origins may also be introduced by an extrachromosomal element.

In addition, we found an intergenic region, (1,957,398..1,957,754 nt), which also contains a 13-mer consensus element, ‘GGGTTCAAATCCC’, and a tRNA gene. However, this region is in close proximity to the putative oriC1,

---

**Fig. (2).** Schematic diagram of the replication origins of *P. calidifontis* JCM 11548. Within the oriC1 (A), there are two 12-mer palindromic sequences (blue) annotated as Orb-1 elements in Pelve et al., 2012. Within the oriC2 (B), oriC3 (C) and oriC4 (D), there is a 13-mer consensus element (yellow) and a tRNA gene (blue).
so it is not believed to function as a replication origin. Furthermore, the locations of all the predicted replication origins are in accordance with those determined by using the high-throughput sequencing-based marker frequency analysis (Fig. 3). Therefore, the predicted replication origins would be useful to further the experimental study of the replication origins in *Pyrobaculum calidifontis* JCM 11548.

### 3.4. Mc-pRIP-adjacent Replication Origins in Methanococcales

While formulating our hypothesis, we found that the locations of other putative replication initiator genes would be helpful in predicting oriC. For example, in the genome of *M. jannaschii*, an ORF (MJ 0774), annotated as a ‘hypothetical protein’, is in fact a distant homolog of the Cdc6 protein [18]. The name Mc-pRIP for the putative replication initiator protein in Methanococcales has been used for MJ0774 and related proteins to distinguish it from bona fide orthologous Cdc6 [26]. We also found the genes, which encode Mc-pRIP in the other thirteen genomes within the order Methanococcales (*Methanococcus aeolicus* Nankai-3, *Methanocaldococcus fervens* AG86, *Methanococcus maripaludis* C5, *M. maripaludis* C6, *M. maripaludis* C7, *M. maripaludis* S2, *M. maripaludis* X1, *Methanococcus vannielii* SB, *Methanococcus voltae* A3, *Methanocaldococcus vulcanus* M7, *Methanocaldococcus* sp. FS406-22, *Methanothermococcus okinawensis* IH1, *Methanocaldococcus infernus* ME), were annotated as ‘LysR family protein’, ‘regulatory protein ArsR’, ‘MarR family transcriptional regulator’, etc. No *cdc6* gene was annotated in the above genomes.

All of the Mc-pRIP genes have been assigned COG identification number COG1474 (Cdc6-related protein, AAA superfamily ATPase), and belong to the COG functional categories L (Replication, recombination and repair) and O (Posttranslational modification, protein turnover, chaperones). In addition, helix-turn-helix domains were found in Mc-pRIP genes, which are believed to be involved in the DNA binding. Conserved domain annotation on the Mc-pRIP protein sequence in *M. jannaschii*, using the CD-Search web-service [48], also confirms the above results (Fig. 4). The AAA+ (ATPases Associated with a wide variety of cellular Activities) superfamily, multi-domains of Arch_ATPase (pfam01637, Archaeal ATPase), CDC6 (COG1474, Cdc6-related protein, AAA superfamily ATPase), TIGR02928 (orc1/cdc6 family replication initiation protein), and putative DNA binding sites have been found on Mc-pRIP protein in *M. jannaschii*. Similar results have also been obtained for the other Mc-pRIP proteins. Consequently, based on the locations of Mc-pRIP genes, the *oriC*s in the aforementioned genomes were predicted reliably and contain almost all the features of known replication origins in archaean genomes.

![Fig. (3). Graphical circular map of the archaeon *P. calidifontis* JCM 11548. The filled circles indicate the locations of four chromosome replication origins in *P. calidifontis* JCM 11548, determined by using the high-throughput sequencing-based marker frequency analysis. The lines indicate the locations of the predicted replication origins and some conserved genes related to the origin regions, such as tRNA gene and *copG* gene.](image-url)
Fig. (4). Conserved domain annotation on the protein sequence of Mc-pRIP in *M. jannaschii* DSM 2661. Shown here is the full view generated by the CD-Search tool, and the default values are used for the BLAST search parameters.

CONFLICT OF INTEREST

The authors confirm that this article content has no conflicts of interest.

ACKNOWLEDGEMENTS

The author would like to thank Prof. Chun-Ting Zhang and Dr. Ren Zhang for their invaluable assistance and inspiring discussions. The present work was supported in part by the National Natural Science Foundation of China (Grant Nos. 31171238, 30800642 and 10747150) and the Program for New Century Excellent Talents in University (No. NCET-12-0396).

REFERENCES

[1] Jacob, F.; Brenner, S.; Cuzin, F. On the regulation of DNA replication in bacteria. *Cold Spring Harbor Symposia on Quant. Biol.*, 1963, 28, 329-348.

[2] Katayama, T.; Ozaki, S.; Keyamura, K.; Fujimitsu, K. Regulation of the replication cycle: conserved and diverse regulatory systems for DNA and oriC. *Nat. Rev. Microbiol.*, 2010, 8(3), 163-170.

[3] Mott, M.L.; Berger, J.M. DNA replication initiation: mechanisms and regulation in bacteria. *Nat. Rev. Microbiol.*, 2007, 5(5), 343-354.

[4] Zhou, Y.; Chen, W.L.; Wang, L.; Zhang, C.C. Identification of the oriC region and its influence on heterocyst development in the filamentous cyanobacterium *Anabaena* sp. strain PCC 7120. *Microbiol.*, 2011, 157(Pt 7), 1910-1919.

[5] Xu, Y.; Ji, X.; Chen, N.; Li, P.; Liu, W.; Lu, X. Development of replicative oriC plasmids and their versatile use in genetic manipulation of *Cytophaga* hatchinsonii. *Appl. Microbiol. Biotech.*, 2012, 93(2), 697-705.

[6] Lundgren, M.; Andersson, A.; Chen, L.; Nilsson, P.; Bernardar, R. Three replication origins in *Sulfolobus* species: synchronous initiation of chromosome replication and asynchrony termination. *Proc. Natl. Acad. Sci. U S A.*, 2004, 101(18), 7046-7051.

[7] Pelve, E.A.; Lindas, A.C.; Knoppel, A.; Mira, A.; Bernardar, R. Four chromosome replication origins in the archaeon *Pyrobaculum calidifontis*. *Mol. Microbiol.*, 2012, 83(5), 986-995.

[8] Robinson, N.P.; Dionne, I.; Lundgren, M.; Marsh, V.L.; Bernardar, R.; Bell, S.D. Identification of Two Origins of Replication in the Single Chromosome of the Archaeon *Sulfolobus solfataricus*. *Cell, 2004*, 116(1), 25-38.

[9] Lobry, J.R. A simple vectorial representation of DNA sequences for the detection of replication origins in bacteria. *Biocliniche*, 1996, 78(5), 323-326.

[10] Lobry, J.R. Asymmetric substitution patterns in the two DNA strands of bacteria. *Mol Biol Evol.*, 1996, 13(5), 660-665.

[11] Gierigoriev, A. Analyzing genomes with cumulative skew diagrams. *Nucleic Acids Res.*, 1998, 26(10), 2286-2290.

[12] Salzberg, S.L.; Salzberg, A.J.; Kerlavage, A.R.; Tomb, J.F. Skewed oligomers and origins of replication. *Gene, 1998*, 217(1-2), 57-67.

[13] Worning, P.; Jensen, L.J.; Hallin, P.F.; Staerfeldt, H.H.; Ussery, D.W. Origin of replication in circular prokaryotic chromosomes. *Environ. Microbiol.*, 2006, 8(2), 353-361.

[14] Mackiewicz, P.; Zakrzewska-Czerwinska, J.; Zawilak, A.; Dudek, M.R.; Cebrat, S. Where does bacterial replication start? Rules for predicting the oriC region. *Nucleic Acids Res.*, 2004, 32(13), 3781-3791.

[15] Zhang, R.; Zhang, C.T. Z-curves, an intuitive tool for visualizing and analyzing the DNA sequences. *J. Biomol. Struct. Dyn.*, 1994, 11(4), 767-782.

[16] Zhang, R.; Zhang, C.T. Identification of replication origins in archaeal genomes based on the Z-curve method. *Archaeaa*, 2005, 1(5), 335-346.

[17] Zhang, R.; Zhang, C.T. Single replication origin of the archaeon *Methanosarcina mazei* revealed by the Z-curve method. *Biochem. Biophys. Res. Commun.*, 2002, 297(2), 396-400.

[18] Zhang, R.; Zhang, C.T. Identification of replication origins in the genome of the methanogenic archaeon, *Methanolocococcus jannaschii*. *Extremophiles, 2004*, 8(3), 253-258.

[19] Zhang, R.; Zhang, C.T. Multiple replication origins of the archaeon *Halobacterium* species NRC-1. *Biochem. Biophys. Res. Commun.*, 2003, 302(4), 728-734.

[20] Berquist, B.R.; DasSarma, S. An archaeal chromosomal autonomously replicating sequence element from an extreme halophile, *Halobacterium* sp. strain NRC-1. *J. Bacteriol.*, 2000, 182(20), 5959-5966.

[21] Gao, F.; Zhang, C.T.-F. Ori-Finder: a web-based system for finding oriC’s in unannotated bacterial genomes. *BMC Bioinform.*, 2008, 9(1), 79.

[22] Gao, F.; Zhang, C.T. Origins of replication in *Sorangium cellulovorum* and *Microcystis aeruginosa*. *DNA Res.*, 2008, 15(3), 169-171.
Hu, S. Insight into the specific virulence related genes and toxin-antitoxin virulent pathogenicity islands in swine streptococcosis pathogen Streptococcus equi ssp. zooepidemicus strain ATCC35246. *BMC Genom.* 2013, 14(1), 377.

[36] Lin, Y.; Gao, F.; Zhang, C.T. Functionality of essential genes drives gene strand-bias in bacterial genomes. *Biochem. Biophys. Res. Commun.* 2010, 396(2), 472-476.

[37] Qu, H.; Wu, H.; Zhang, T.; Zhang, Z.; Hu, S.; Yu, J. Nucleotide compositional asymmetry between the leading and lagging strands of eubacterial genomes. *Res. Microbiol.*, 2010, 161(10), 838-846.

[38] Vieira-Silva, S.; Rocha, E.P. The systemic imprint of growth and its use in ecological (meta)genomics. *PLoS Genet.*, 2010, 6(1), e1000008.

[39] Lin, Y.; Zhang, R.R. Putative essential and core-essential genes in Mycoplasma genomes. *Sci. Rep.*, 2011, 1, 53.

[40] Sobetzko, P.; Travers, A.; Muskeshilvili, G. Gene order and chromosome dynamics coordinate spatiotemporal gene expression during the bacterial growth cycle. *Proc. Natl. Acad. Sci. U S A.*, 2012, 109(2), E42-50.

[41] Mao, X.; Zhang, H.; Yin, Y.; Xu, Y. The percentage of bacterial genes on leading versus lagging strands is influenced by multiple balancing forces. *Nucleic Acids Res.*, 2012, 40(17), 8210-8218.

[42] Sobetzko, P.; Glinkowska, M.; Travers, A.; Muskeshilvili, G. DNA thermodynamic stability and supercoil dynamics determine the gene expression program during the bacterial growth cycle. *Mol. Biosyst.*, 2013, 9(7), 1643-1651.

[43] Fuller, R.S.; Funnell, B.E.; Kornberg, A. The dnaA protein drives gene strand-bias in bacterial genomes. *Biochem. Biophys. Res. Commun.*, 2011, 413(1), 377.

[44] Robinson, N.P.; Bell, S.D. Extrachromosomal element capture and its uses in ecological (meta)genomics. *PLoS Genet.*, 2010, 6(12), e1001309.

[45] Robinson, N.P.; Bell, S.D. Extrachromosomal element capture and its uses in ecological (meta)genomics. *PLoS Genet.*, 2010, 6(12), e1001309.

[46] Robinson, N.P.; Bell, S.D. Extrachromosomal element capture and its uses in ecological (meta)genomics. *PLoS Genet.*, 2010, 6(12), e1001309.

[47] Robinson, N.P.; Bell, S.D. Extrachromosomal element capture and its uses in ecological (meta)genomics. *PLoS Genet.*, 2010, 6(12), e1001309.

[48] Robinson, N.P.; Bell, S.D. Extrachromosomal element capture and its uses in ecological (meta)genomics. *PLoS Genet.*, 2010, 6(12), e1001309.

[49] Robinson, N.P.; Bell, S.D. Extrachromosomal element capture and its uses in ecological (meta)genomics. *PLoS Genet.*, 2010, 6(12), e1001309.

[50] Robinson, N.P.; Bell, S.D. Extrachromosomal element capture and its uses in ecological (meta)genomics. *PLoS Genet.*, 2010, 6(12), e1001309.

[51] Robinson, N.P.; Bell, S.D. Extrachromosomal element capture and its uses in ecological (meta)genomics. *PLoS Genet.*, 2010, 6(12), e1001309.

[52] Robinson, N.P.; Bell, S.D. Extrachromosomal element capture and its uses in ecological (meta)genomics. *PLoS Genet.*, 2010, 6(12), e1001309.

[53] Robinson, N.P.; Bell, S.D. Extrachromosomal element capture and its uses in ecological (meta)genomics. *PLoS Genet.*, 2010, 6(12), e1001309.

[54] Robinson, N.P.; Bell, S.D. Extrachromosomal element capture and its uses in ecological (meta)genomics. *PLoS Genet.*, 2010, 6(12), e1001309.

[55] Robinson, N.P.; Bell, S.D. Extrachromosomal element capture and its uses in ecological (meta)genomics. *PLoS Genet.*, 2010, 6(12), e1001309.

[56] Robinson, N.P.; Bell, S.D. Extrachromosomal element capture and its uses in ecological (meta)genomics. *PLoS Genet.*, 2010, 6(12), e1001309.

[57] Robinson, N.P.; Bell, S.D. Extrachromosomal element capture and its uses in ecological (meta)genomics. *PLoS Genet.*, 2010, 6(12), e1001309.

[58] Robinson, N.P.; Bell, S.D. Extrachromosomal element capture and its uses in ecological (meta)genomics. *PLoS Genet.*, 2010, 6(12), e1001309.

[59] Robinson, N.P.; Bell, S.D. Extrachromosomal element capture and its uses in ecological (meta)genomics. *PLoS Genet.*, 2010, 6(12), e1001309.

[60] Robinson, N.P.; Bell, S.D. Extrachromosomal element capture and its uses in ecological (meta)genomics. *PLoS Genet.*, 2010, 6(12), e1001309.

[61] Robinson, N.P.; Bell, S.D. Extrachromosomal element capture and its uses in ecological (meta)genomics. *PLoS Genet.*, 2010, 6(12), e1001309.

[62] Robinson, N.P.; Bell, S.D. Extrachromosomal element capture and its uses in ecological (meta)genomics. *PLoS Genet.*, 2010, 6(12), e1001309.

[63] Robinson, N.P.; Bell, S.D. Extrachromosomal element capture and its uses in ecological (meta)genomics. *PLoS Genet.*, 2010, 6(12), e1001309.

[64] Robinson, N.P.; Bell, S.D. Extrachromosomal element capture and its uses in ecological (meta)genomics. *PLoS Genet.*, 2010, 6(12), e1001309.

[65] Robinson, N.P.; Bell, S.D. Extrachromosomal element capture and its uses in ecological (meta)genomics. *PLoS Genet.*, 2010, 6(12), e1001309.

[66] Robinson, N.P.; Bell, S.D. Extrachromosomal element capture and its uses in ecological (meta)genomics. *PLoS Genet.*, 2010, 6(12), e1001309.

[67] Robinson, N.P.; Bell, S.D. Extrachromosomal element capture and its uses in ecological (meta)genomics. *PLoS Genet.*, 2010, 6(12), e1001309.

[68] Robinson, N.P.; Bell, S.D. Extrachromosomal element capture and its uses in ecological (meta)genomics. *PLoS Genet.*, 2010, 6(12), e1001309.

[69] Robinson, N.P.; Bell, S.D. Extrachromosomal element capture and its uses in ecological (meta)genomics. *PLoS Genet.*, 2010, 6(12), e1001309.

[70] Robinson, N.P.; Bell, S.D. Extrachromosomal element capture and its uses in ecological (meta)genomics. *PLoS Genet.*, 2010, 6(12), e1001309.

[71] Robinson, N.P.; Bell, S.D. Extrachromosomal element capture and its uses in ecological (meta)genomics. *PLoS Genet.*, 2010, 6(12), e1001309.

[72] Robinson, N.P.; Bell, S.D. Extrachromosomal element capture and its uses in ecological (meta)genomics. *PLoS Genet.*, 2010, 6(12), e1001309.

[73] Robinson, N.P.; Bell, S.D. Extrachromosomal element capture and its uses in ecological (meta)genomics. *PLoS Genet.*, 2010, 6(12), e1001309.

[74] Robinson, N.P.; Bell, S.D. Extrachromosomal element capture and its uses in ecological (meta)genomics. *PLoS Genet.*, 2010, 6(12), e1001309.

[75] Robinson, N.P.; Bell, S.D. Extrachromosomal element capture and its uses in ecological (meta)genomics. *PLoS Genet.*, 2010, 6(12), e1001309.

[76] Robinson, N.P.; Bell, S.D. Extrachromosomal element capture and its uses in ecological (meta)genomics. *PLoS Genet.*, 2010, 6(12), e1001309.