Ori-Finder 2, an integrated tool to predict replication origins in the archaeal genomes

Hao Luo1, Chun-Ting Zhang1* and Feng Gao1,2,3*

1 Department of Physics, Tianjin University, Tianjin, China
2 Key Laboratory of Systems Bioengineering (Ministry of Education), Tianjin University, Tianjin, China
3 SynBio Research Platform, Collaborative Innovation Center of Chemical Science and Engineering, Tianjin, China

INTRODUCTION
DNA replication is one of the most basic processes in all three domains of cellular life. With the advent of the post-genomic era, the increasing number of complete archaeal genomes has created an opportunity for exploration of the molecular mechanisms for initiating cellular DNA replication by in vivo experiments as well as in silico analysis. However, the location of replication origins (oriCs) in many sequenced archaeal genomes remains unknown. We present a web-based tool Ori-Finder 2 to predict oriCs in the archaeal genomes automatically, based on the integrated method comprising the analysis of base composition asymmetry using the Z-curve method, the distribution of origin recognition boxes identified by FIMO tool, and the occurrence of genes frequently close to oriCs. The web server is also able to analyze the unannotated genome sequences by integrating with gene prediction pipelines and BLAST software for gene identification and function annotation. The result of the predicted oriCs is displayed as an HTML table, which offers an intuitive way to browse the result in graphical and tabular form. The software presented here is accurate for the genomes with single oriC, but it does not necessarily find all the origins of replication for the genomes with multiple oriCs. Ori-Finder 2 aims to become a useful platform for the identification and analysis of oriCs in the archaeal genomes, which would provide insight into the replication mechanisms in archaea. The web server is freely available at http://tubic.tju.edu.cn/Ori-Finder2/.

Keywords: archaea, replication origins, Z-curve, origin recognition box, DNA replication

The first putative oriC of archaea was identified in Halobacterium sp. strain NRC-1 by GC-skew method and demonstrated by cloning into a non-replicating plasmid (Myllykallio et al., 2000). The Z-curve method is an alternative technique that detects the asymmetrical nucleotide distribution around replication origins. The three components of the Z-curve, \(x_n\), \(y_n\), 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 along the sequence, respectively. The \(x_n\) and \(y_n\) components are termed the RY and MK disparity curves, respectively. The AT and GC disparity curves are defined as \((x_n + y_n)/2 - (x_n - y_n)/2\), which shows the excess of A over T and G over C, respectively, along the sequence (Zhang and Zhang, 2005; Gao, 2014). Based on the Z-curve analysis, we have identified single oriC in Methanocaldococcus jannaschii and Methanosarcina mazei, double oriCs in Halobacterium sp. strain NRC-1, and three oriCs in Sulfolobus solfataricus P2, which are consistent with the subsequent experiments (Soppa, 2006). Recently, multiple oric/cdc6-associated oriCs in all the available haloarchaeal genomes have been predicted by identification of putative ORBs (Wu et al., 2012). Based on these discoveries, several basic features of the oriCs could be summarized in archaea. Firstly, most oriCs are located in proximity to the genes encoding archaeal replication-related proteins, such as archaeal Orc/Cdc6 protein, Whip (Winged-Helix Initiator...
Protein) and DNA primase. Secondly, oriCs are often located around the extremes of disparity curves. Finally, most of the oriC contains the AT-rich unwinding elements and conserved ORBs (Zhang and Zhang, 2005; Barry and Bell, 2006; Wu et al., 2014a).

Our group has developed a web-based system Ori-Finder 1 to find oriCs in the bacterial genomes based on the Z-curve method with high accuracy and reliability (Gao and Zhang, 2008). Now with the knowledge of oriCs in the archaeal genomes, we present an online tool, Ori-Finder 2, to identify the oriCs in the archaeal genomes, based on the integrated method comprising the analysis of base composition asymmetry using the Z-curve method, the distribution of ORB elements identified by FIMO tool, and the occurrence of genes frequently close to replication origins, which is available at http://tubic.tju.edu.cn/Ori-Finder2/.

**METHODS AND IMPLEMENTATION**

Ori-Finder 2 utilizes an integrated approach to predict oriCs in the user-supplied archaeal genomes automatically. Figure 1 presents the workflow of Ori-Finder 2. Users submit an annotated or unannotated genome sequence to the web server. For the annotated genome, we recommend that users submit the sequence file in GenBank format or upload the sequence file in FASTA format as well as its corresponding protein table (PTT) file. The web server is also able to analyze the unannotated genomes by integrating two gene prediction pipelines, ZCURVE1.02 and Glimmer3 (Guo et al., 2003; Delcher et al., 2007), for gene identification and BLAST program for functional annotations of genes. Then all the intergenic sequences are scanned by Find Individual Motif Occurrences (FIMO), a software tool for scanning DNA or protein sequences with motifs described as position-specific scoring matrices (Grant et al., 2011), to obtain the ORB sequences, and also by REPuter program, a classic pipeline to compute exact repeats and palindromes in complete genomes (Kurtz et al., 2001), to identify the repeats. Finally, all the intergenic sequences adjacent to the replication-related genes with the ORB sequences are predicted as oriCs. Since the approach relies on the prior knowledge of oriCs in archaea, it may fail to identify the oriCs adjacent to the unknown genes which might be involved in DNA replication. In order to overcome the drawback, the intergenic sequences, which contain more than two conserved motifs, will be also predicted as oriCs. BLAST searches are performed against DoriC, a database of bacterial and archaeal replication origins, to search the homologs (Gao and Zhang, 2007; Gao et al., 2013). Here, the conserved motifs of ORB sequences used in FIMO were obtained from DoriC. All the records in DoriC were organized into several taxonomic clusters, including Methanobacteriaceae, Methanomicrobia, Methanococcaceae, Sulfolobaceae and Thermococcaceae. And the conserved ORB motifs were calculated from the corresponding clusters by Multiple EM for Motif Elicitation (MEME) program, a tool used to discover motifs in a group of related DNA or protein sequences (Bailey et al., 2009). Table 1 displays the regular expressions of ORB motifs. Note that the common motif is calculated from all the records in DoriC. The motif logos are shown in the submission form, and the position specific probability matrix (PSPM) is available in the document webpage. Each job of Ori-Finder 2 is assigned a unique ID, and the whole process will take several minutes to complete. Users could retrieve their results with the job ID or be notified by email if specified in the submission page.

In the result, the information including genome size, GC content, the locations of replication-related genes and the predicted oriCs, as well as the Z-curve (AT, GC, RY, and MK disparity curves) for the input genome is displayed as an HTML table. In addition, the detailed information about the repeats identified by REPuter program, ORBs recognized by FIMO and the homologs in DoriC are also presented in the corresponding subtable. The ORB motifs in all the intergenic regions are also available for download from the provided URL. Users could also click to enlarge the embedded figure to obtain the high
### Table 1 | The regular expressions of the ORB motifs identified by MEME.

| Taxonomy                   | E-value | Regular expressions                                                                 |
|----------------------------|---------|-------------------------------------------------------------------------------------|
| Halobacteriaceae           | 4.5E-180| TT[TC]CACC[GCT]GAAAC[GA][AC][GA][GT][GT]                                           |
| Methanobacteriaceae        | 4.20E-68| TT[TA]CACCTTGAAT[GTAT][CT][GT]                                                  |
| Methanomicrobia            | 1.50E-202| TCCA[GC][GT][GAA][CA][AT][GT]                                                  |
| Methanococcaceae           | 4.80E-90| TT[TA][GT]ATTCA[TC][GA][AT][AT][AT]                                              |
| Sulfolobaceae              | 7.50E-296| [GC][GCC][AAG][GT][CT][GT][T][CG][CA][CC][TC][GG]                            |
|                            | 2.40E-286| TCCA[AG][AT][TG][GAA][CA][CT][GA][AG][GGGT]                                      |
|                            | 8.20E-120| GAGTG[CT][CGGT][CT][GCA][ATC][CC][AG]                                           |
|                            | 2.40E-286| TCCA[AG][AT][TG][GAA][CA][CT][GA][AG][GGGT]                                      |
|                             | 8.20E-120| GAGTG[CT][CGGT][CT][GCA][ATC][CC][AG]                                           |
| Thermococcaceae            | 9.10E-223| [TC][TCCAGTTGAAAT][GTA][CT][GT]                                                  |
|                            | 6.90E-56 | [CAG][TTCCA[CT][TA][GGA][AT][CGA][CT]                                           |
|                            | 2.00E-52 | AATG[ACT][AC][AT][AT][GT][AT][T][T]                                             |
| Common\(^a\)               | 1.20E-225| TCCA[CG][TG][GAA][AT][G][GGGT]                                                   |

\(^a\) Note that the Common motif is calculated from all the records in DoriC by MEME. In Halobacteriaceae, Methanobacteriaceae, Methanomicrobia, Sulfolobaceae, and Thermococcaceae, they share the consensus sequences “TCCA—GAAAC” similar to the common motif. In Methanomicrobia and Sulfolobaceae, “G-string” (GGGGT) is observed obviously at the end of ORB motifs.

### Table 2 | The prediction results of 13 archaeal chromosomes\(^a\).

| Organism                                      | Refseq     | OriCs in DoriC | OriCs predicted by Ori-Finder 2 | True positive |
|-----------------------------------------------|------------|----------------|---------------------------------|---------------|
| Aeropyrum pernix K1                           | NC_000854  | 2              | 2                              | 1             |
| Pyrococcus abyssi GE5                         | NC_000868  | 1              | 1                              | 1             |
| Methanothermobacter thermotrophicus str. Delta H chromosome | NC_000916  | 1              | 2                              | 1             |
| Archaeoglobus fulgidus DSM 4304              | NC_000917  | 1              | 1                              | 0             |
| Pyrococcus horikoshii OT3                     | NC_000961  | 1              | 1                              | 1             |
| Halobacterium sp. NRC-1                       | NC_002607  | 2              | 4                              | 2             |
| Pyrococcus furiosus DSM 3638                  | NC_003413  | 1              | 1                              | 1             |
| Hyperthermus butylicus DSM 5456               | NC_008818  | 2              | 1                              | 1             |
| Pyrobaculum calidontis JCM 11548              | NC_009073  | 1              | 1                              | 0             |
| Haloferax volcanii DS2                        | NC_013967  | 5              | 6                              | 5             |
| Haloarcula hispanica ATCC 33960 chromosome II | NC_015943  | 4              | 7                              | 3             |
| Haloarcula hispanica ATCC 33960 chromosome I  | NC_015948  | 5              | 1                              | 1             |
| Nitrosopumilus maritimus SCM1                  | NC_010085  | 1              | 1                              | 1             |
| Total                                         | –          | 27             | 29                             | 18            |

\(^a\) Note that the detailed information is available at http://hubic.tju.edu.cn/Ori-Finder2/doc.php#9.

Resolution one which displays the RY, MK, GC, AT disparity curves, replication-related proteins, and the predicted oriCs. The result webpage and figures will be stored in 7 days on the web server.

Ori-Finder 2 is developed using Python and PHP on a Unix platform with an Apache web-server. The web interface is implemented using Common Gateway Interface (CGI) python scripts, and the webpage is designed with HTML, CSS, and JavaScript. The pipeline of Ori-Finder 2 uses the Biopython library, and the output graphs are generated by the Python module Matplotlib (Hunter, 2007; Cock et al., 2009).

**RESULTS AND DISCUSSION**

Based on this online system, we predicted the oriCs for all the available complete archaeal genomes in GenBank. For example, *Pyrococcus abyssi* is a classical model of DNA replication in the archaeal organisms. Similar to bacteria, there is only one oriC in its circular chromosome, which has been identified by
FIGURE 2 | Example of Ori-Finder 2 result for *Pyrococcus abyssi* GE5.

(A) The information of genome size, GC content, the locations of replication-related genes and the predicted oriC.

(B) The detailed information of the predicted oriC region including size, GC content, homologs in DoriC and sequence, as well as the information of the identified ORBs including the ORB motif (also referred to as “Pattern name”), location, strand, the associated log-likelihood ratio score, P value and the matched sequences. Note that the log-likelihood ratio score and P value are computed by FIMO to measure the similarity between the ORB motif and the matched sequence, and the P value cutoff for FIMO motif searching is $10^{-4}$. The ORB motif used here is the common motif.

(C) The left figure shows the Z-curves (AT, GC, RY, and MK disparity curves) for the original sequence, and the right figure shows the Z-curves (AT, GC, RY, and MK disparity curves) for the rotated sequence beginning and ending in the maximum of the GC disparity curve. The short vertical red line indicates the location of replication-related protein. The black arrow is the predicted oriC region.
cumulative oligomer skew and confirmed by in vivo method. With the annotated genome file, the oriC predicted by Ori-Finder 2 is in accordance with the experimental result and located at the peak of the MK disparity curve. Several ORB sequences are recognized in the oriC. Figure 2 is a screenshot of the result by Ori-Finder 2. In addition, some archaea adopt more than one oriC during the DNA replication. For this situation, Ori-Finder 2 also predicted multiple oriCs in their genomes. Haloferax volcanii DS2 has a chromosome with multiple oriCs. Five oriCs were identified in silico, and three of them have been confirmed in vitro (Norais et al., 2007; Wu et al., 2012; Hawkins et al., 2013). With the annotated genome file, all the five oriCs mentioned above have been predicted by Ori-Finder 2 successfully, and another oriC with three ORB motifs is also found, which is adjacent to the genes purO and ceg. Besides that, the oriCs identified in the unannotated genomes are consistent with the previous results. In order to estimate the performance of Ori-Finder 2, we used 13 annotated archaeal chromosomes, whose oriCs have been confirmed by experimental method or identified in silico by other groups (Table 2). Compared with the records in DorIC, the sensitivity and precision are 66.7% and 62.1%, respectively. The reason of the lower precision and sensitivity compared with the programs to detect bacterial origins, such as Ori-Finder 1, is that bacteria have only one oriC in their chromosomes, but archaea tend to have more than one. Furthermore, oriCs in archaea show more diversity than those in bacteria, such as more complex ORBs in comparison with the DnaA boxes, and more unknown species-specific replication-related genes. It is difficult to predict the oriCs in archaea with high precision and sensitivity due to the limited amount of experimental data. For example, not all the oriCs in the genomes with multiple oriCs are found, and the ORBs with unique features need to be further explored by experimental methods. For the convenience of users' query, the oriCs confirmed by in vivo or in silico methods have been collected into DorIC, which is freely available at http://tubic.tju.edu.cn/doric/.

CONCLUSION
Here, we presented a user-friendly interactive web-based platform Ori-Finder 2 to predict the oriCs in the archaeal genomes. The tool integrates several genomic pipelines, including FIMO, BLAST, ZCURVE, Glimmer, and REPuter, to comprehensively annotate and analyze the oriCs. Moreover, the ORB motifs are also calculated by MEME and organized by taxonomy. The software presented here does not necessarily find all the origins of replication in cases where there are multiple ones in a genome. However, we will continually strive to improve our approach to make it more accurate and sensitive with the increase of the oriCs confirmed experimentally in archaea. As the only currently available auto-annotation system for the archaeal replication origins at the sequence level, we believe that Ori-Finder 2 will be helpful to predict the archaeal replication origins and provide insight into DNA replication in archaea.

AUTHOR CONTRIBUTIONS
Hao Luo designed the computer program and drafted the manuscript. Chun-Ting Zhang and Feng Gao supervised the study and revised the manuscript. All authors read and approved the final manuscript.

ACKNOWLEDGMENTS
The authors thank Dr. Kurtz for providing the REPuter binaries. They also would like to thank Dr. Ren Zhang for invaluable assistance. The present work was supported in part by National Natural Science Foundation of China (Grant Nos. 31172128 and 30800642), and Program for New Century Excellent Talents in University (No. NCET-12-0396).

REFERENCES
Bailey, T. L., Boden, M., Buske, F. A., Frith, M., Grant, C. E., Clementi, L., et al. (2009). MEME SUITE: tools for motif discovery and searching. Nucleic Acids Res. 37, W202–W208. doi: 10.1093/nar/gkp335
Barry, E. R., and Bell, S. D. (2006). DNA replication in the archaea. Microbiol. Mol. Biol. Rev. 70, 876–887. doi: 10.1128/MMBR.00029-06
Cock, P. J., Antao, T., Chang, J. T., Chapman, B. A., Cox, C. J., Dalke, A., et al. (2009). Biopython: freely available Python tools for computational molecular biology and bioinformatics. Bioinformatics 25, 1422–1423. doi: 10.1093/bioinformatics/btp163
Delcher, A. L., Bratke, K. A., Powers, E. C., and Salzberg, S. L. (2007). Identifying bacterial genes and endosymbiont DNA with Glimmer. Bioinformatics 23, 673–679. doi: 10.1093/bioinformatics/btm069
Gao, F. (2014). Recent advances in the identification of replication origins based on the Z-curve method. Curr. Genomics 15, 104–112. doi: 10.2174/1389202915999140328162938
Gao, F., Luo, H., and Zhang, C. T. (2012). DeOri: a database of eukaryotic DNA replication origins. Bioinformatics 28, 1551–1552. doi: 10.1093/bioinformatics/bts151
Gao, F., Luo, H., and Zhang, C. T. (2013). DorIC 5.0: an updated database of oriC regions in both bacterial and archaeal genomes. Nucleic Acids Res. 41, D90–D93. doi: 10.1093/nar/gks990
Gao, F., and Zhang, C. T. (2007). DorIC: a database of oriC regions in bacterial genomes. Bioinformatics 23, 1866–1867. doi: 10.1093/bioinformatics/btm255
Gao, F., and Zhang, C. T. (2008). Ori-Finder: a web-based system for finding oriCs in unannotated bacterial genomes. BMC Bioinformatics 9:79. doi: 10.1186/1471-2105-9-79
Grant, C. E., Bailey, T. L., and Noble, W. S. (2011). FIMO: scanning for occurrences of a given motif. Bioinformatics 27, 1017–1018. doi: 10.1093/bioinformatics/btr064
Guo, F. B., Ou, H. Y., and Zhang, C. T. (2003). ZCURVE: a new system for recognizing protein-coding genes in bacterial and archaeal genomes. Nucleic Acids Res. 31, 1780–1789. doi: 10.1093/nar/gkg234
Hawkins, M., Malla, S., Blythe, M. J., Nieduszynski, C. A., and Allers, T. (2013). Accelerated growth in the absence of DNA replication origins. Nature 503, 544–547. doi: 10.1038/nature12650
Hunter, J. D. (2007). Matplotlib: a 2D graphics environment. Comput. Sci. Eng. 9, 90–95. doi: 10.1109/MCSE.2007.53
Kurtz, S., Choudhuri, J. V., Ohebusch, E., Schleiermacher, C., Stoeck, I., and Giegerich, R. (2007). REPuter: the manifold applications of repeat analysis on a genomic scale. Nucleic Acids Res. 35, 2549–2565. doi: 10.1093/nar/gkm254
Myllykallio, H., Lopez, P., Lopez-Garcia, P., Heilig, R., Saurin, W., Zivanovic, Y., et al. (2000). Bacterial mode of replication with eukaryotic-like machinery in a hyperthermophilic archaeon. Science 288, 2212–2215. doi: 10.1126/science.288.5474.2212
Norais, C., Hawkins, M., Hartman, A. L., Eisen, J. A., Myllykallio, H., and Allers, T. (2007). Genetic and physical mapping of DNA replication origins in Haloferax volcanii. PLoS Genet. 3:e77. doi: 10.1371/journal.pgen.0030077
Soppa, J. (2006). From genomes to function: haloarchaea as model organisms. Microbiology 152(Pt 3), 585–590. doi: 10.1099/mic.0.28504-0
Woese, C. R., and Fox, G. E. (1977). Phylogenetic structure of the prokaryotic domain: the primary kingdoms. Proc. Natl. Acad. Sci. U.S.A. 74, 5088–5090. doi: 10.1073/pnas.74.11.5088
Wu, Z., Liu, H., Liu, J., Liu, X., and Xiang, H. (2012). Diversity and evolution of multiple orc/cdc6-adjacent replication origins in haloarchaea. BMC Genomics 13:478. doi: 10.1186/1471-2164-13-478
Wu, Z., Liu, J., Yang, H., Liu, H., and Xiang, H. (2014a). Multiple replication origins with diverse control mechanisms in *Haloarcula hispanica*. *Nucleic Acids Res.* 42, 2282–2294. doi: 10.1093/nar/gkt1214

Wu, Z., Liu, J., Yang, H., and Xiang, H. (2014b). DNA replication origins in archaea. *Front. Microbiol.* 5:179. doi: 10.3389/fmicb.2014.00179

Zhang, R., and Zhang, C. T. (2005). Identification of replication origins in archaeal genomes based on the Z-curve method. *Archaea* 1, 335–346. doi: 10.1155/2005/509646

**Conflict of Interest Statement:** The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.