Ori-Finder 2022: A Comprehensive Web Server for Prediction and Analysis of Bacterial Replication Origins

Meijing Dong[1,#], Hao Luo[1,#], Feng Gao[1,2,3,*]

[1] *Department of Physics, School of Science, Tianjin University, Tianjin 300072, China*

[2] *Frontiers Science Center for Synthetic Biology and Key Laboratory of Systems Bioengineering (Ministry of Education), Tianjin University, Tianjin 300072, China*

[3] *SynBio Research Platform, Collaborative Innovation Center of Chemical Science and Engineering (Tianjin), Tianjin 300072, China*

[#] Equal contribution.

[*] Corresponding author.

E-mail: fgao@tju.edu.cn (Gao F).

Running title: *Dong M et al / Web Server Ori-Finder 2022*

Total word counts (from “Introduction” to “Conclusions”): 2898

Total figures: 2

Total supplementary tables: 2

Total references: 61

Total characters in the article tile: 91

Total characters in the running title: 33

Total keywords: 5

Total words in Abstract: 201

Abstract

The replication of DNA is a complex biological process that is essential for life. Bacterial DNA replication is initiated at genomic loci referred to as replication origins (*oriC*s). Integrating the Z-curve method, DnaA box distribution, and comparative genomic analysis, we developed a web server to predict bacterial *oriC*s in 2008 called Ori-Finder, which contributes to clarify the characteristics of bacterial *oriC*s. The *oriC*s of hundreds of sequenced bacterial genomes have been annotated in their genome reports using Ori-Finder and the predicted results have been deposited in DoriC, a manually curated database of *oriC*s. This has facilitated large-scale data mining of functional elements in *oriC*s and strand-biased analysis. Here, we describe Ori-Finder 2022 with updated prediction framework, interactive visualization module, new analysis module, and user-friendly interface. More species-specific indicator genes and functional elements of *oriC*s are integrated into the updated framework, which has also been redesigned to predict *oriC*s in draft genomes. The interactive visualization module displays more genomic information related to *oriC*s and their functional elements. The analysis module includes regulatory protein annotation, repeat sequence discovery, homologous *oriC* search, and strand-biased analyses. The redesigned interface provides additional customization options for *oriC* prediction. Ori-Finder 2022 is freely available at http://tubic.tju.edu.cn/Ori-Finder2022 and https://tubic.org/Ori-Finder2022.

KEYWORDS: Bacteria; DNA replication; Replication origin; Z-curve; DnaA-trio

Introduction

As a complex and essential process of cell life, DNA replication is strictly regulated to ensure the accurate transfer of genetic material from parents to offspring. Identification and characterization of replication origins (*oriC*s) can provide new insights into the mechanisms of DNA replication as well as cell cycle regulation and facilitate drug development [1], genome design [2], plasmid construction [3] etc. Therefore, various experimental approaches such as two-dimensional agarose gel electrophoresis [4], assay of autonomously replicating sequence activity [5], and marker frequency analysis (MFA) [6] have been developed to identify bacterial *oriC*s. Microarray based whole-genome MFA [7] as well as high-throughput sequencing-based MFA [8] with higher resolution have been proposed to generate the replication maps of genomes and to locate *oriC*s. Detecting interactions between DNA and proteins can also provide evidence for predicted *oriC*s [9].

However, the rapid accumulation of sequenced genomes has rendered identifying *oriC*s in all of them impossible using experimental methods. Therefore, the development of bioinformatics algorithms to predict *oriC*s on a large scale is particularly important. Classical *in silico* methods, such as GC skew [10], cumulative GC skew [11], and oligomer skew [12] have been proposed based on DNA asymmetry. Furthermore, Oriloc was developed to predict bacterial *oriC*s by analyzing local and systematic deviations of base composition within each strand [13]. However, these methods only provided the approximate location without the precise boundary of predicted *oriC*s. In addition, they cannot accurately predict *oriC*s in bacterial genomes without a typical GC skew, which is universal sometimes for genomes in certain phylum, such as Cyanobacteria [14, 15]. Although DNA asymmetry is the most common characteristic used for predicting *oriC*s, Mackiewicz *et al.* also found that the prediction could be improved more by considering *dnaA* and DnaA box clusters [16]. However, DnaA box motifs are often species-specific, and the *oriC* is not always close to the *dnaA* gene in some species. Considering these factors, the Ori-Finder web server was developed to provide users with a more convenient and accurate tool for predicting

*oriC*s [17].

Since it was introduced in 2008, Ori-Finder has been widely used to help investigators identify *oriC*s. To date, Ori-Finder has been used to identify *oriC*s in hundreds of sequenced bacterial genomes in their genome reports [18-20], and dozens of the predicted *oriC*s have been experimentally confirmed [21-24]. Furthermore, Ori-Finder predictions have led to new discoveries. For example, each bacterial chromosome is generally considered to carry a single *oriC*. However, Ori-Finder predictions indicate that multiple *oriC*s may occur on a bacterial chromosome [25, 26], and this opinion has been used to explain the experimental results of investigations into single *Achromatium* cells [27]. Naturally occurring single chromosome in *Vibrio cholerae* strain harbor two functional *oriC*s, which provides strong support for our opinion [28]. Ori-Finder provides a large number of *oriC*s as resources for data mining. Particularly, the *oriC*s identified by Ori-Finder, including those confirmed by experiments *in vivo* and *in vitro*, have been organized into the DoriC database [29-31] available at http://tubic.org/doric. Therefore, the data for *oriC* characteristics can be mined on a large scale [32, 33]. For example, vast amounts of *oriC* data can be used to identify and analyze functional elements, such as DnaA boxes and DnaA-trios [34, 35]. Finally, Ori-Finder facilitates analyses of strand-biased biological characteristics that are closely associated with DNA replication, transcription, and other biological processes [10, 36]. The Ori-Finder web server and DoriC database have been extensively applied to in strand-biased analyses, such as base composition [37, 38], gene orientation [39], and codon usage [40]. Ori-Finder has also been recommended as a software tool to identify replichores [41].

Bacterial *oriC*s generally contain several functional elements, such as DnaA-binding sites, AT-rich DNA unwinding elements (DUEs), and binding sites for proteins that regulate replication initiation [42]. These functional elements play important roles in the initiation of DNA replication, which should be considered in the prediction of *oriC*s. Most of bacterial *oriC*s contain DnaA box clusters that are recognized and bound by DnaA proteins. Therefore, the DnaA box cluster is considered as an important characteristic for predicting *oriC*s [16]. DnaA box is usually a 9-bp non-palindromic

109 motif, such as the perfect *Escherichia coli* DnaA box TTATCCACA. Species-specific
110 DnaA box motifs, such as TTTTCCACA in Cyanobacteria and AAACCTACCACC in
111 *Thermotoga maritima* have been identified [43]. In addition, degenerated DnaA boxes
112 have also been identified within *oriC*s in some species, such as 6mer ATP-DnaA boxes
113 (AGATCT) in *E. coli* [44]. Although degenerate DnaA boxes can also bind DnaA
114 protein, only the broadly conserved DnaA box is considered for *oriC* prediction here.
115 The DnaA protein not only interacts with the double-stranded DnaA box, but also
116 binds to the single-stranded DNA to promote unwinding. For example, DnaA protein
117 can bind to single-stranded ATP-DnaA boxes mentioned above. The two-state and loop-
118 back models can explain how DnaA protein melts DNA and stabilizes the unwound
119 region by DnaA-ssDNA interaction [42]. In two-state model, DnaA protein guided from
120 double-stranded DnaA boxes to the adjacent single-stranded DNA changes from a
121 double- to a single-stranded binding mode. A new *oriC* element comprising repeated 3-
122 mer motif (DnaA-trio), found in *Bacillus subtilis*, promotes DNA unwinding by
123 stabilizing DnaA filaments on a single DNA strand [45]. Consequently, a basal
124 unwinding system (BUS) comprised DnaA boxes and DnaA-trios in bacterial *oriC*s has
125 been proposed [46]. Subsequent bioinformatic analyses of *oriC*s from over 2000
126 bacterial species, together with molecular biology studies of six representative species,
127 found that the BUS is broadly conserved in bacteria [35]. Integration host factor (IHF)
128 induces DNA to bend backwards in the loop-back model, bringing the DUE close to
129 the DnaA protein bound to the DnaA box and thus simultaneously facilitating protein
130 binding to double- and single-stranded DNA sequences. This mechanism has been
131 identified in *E. coli* [47], and a similar mechanism might be also found in *Helicobacter*
132 *pylori* [48] with a bipartite *oriC* and *V. cholerae* chromosome 2 whose replication
133 initiator requires RctB protein other than DnaA protein [49].
134 In addition to binding sites for the DnaA protein, *oriC* has other binding sites for
135 proteins that regulate replication initiation. Factor for inversion stimulation (Fis) and
136 IHF bind to specific sites and bend *oriC* DNA to inhibit or facilitate DnaA binding in
137 *E. coli* [47]. SeqA blocks *oriC* recognition of DnaA by binding to the transiently
138 hemimethylated GATC sequence cluster [50]. The regulatory mechanisms might differ

because of the diversity of regulatory proteins and their binding motifs among species. For example, CtrA in *Caulobacter crescentus* plays a similar role to SeqA and inhibit replication initiation by binding motifs (TTAA-N7-TTAA) [51, 52]. Wolanski *et al.* comprehensively summarized the detailed information about the proteins that regulate DNA replication initiation and their binding sites [53].

To facilitate a comprehensive understanding of the replication mechanism and sequence characteristics related to *oriC*s, Ori-Finder 2022 annotates various regulatory proteins and functional elements within *oriC*s. Updated information about the user interface, prediction framework, visualization, and analysis modules are described in detail below.

Method

Software implementation

Ori-Finder 2022 was deployed using a Linux-Apache-MySQL-PHP structure and mainly developed using Python and C++ languages. We packaged the pipeline into a container using Docker to ensure reproducible and reliable execution. We also integrated the third-party tools BLAST+ 2.11.0 [54], Prodigal [55], Stress-Induced Structural Transitions (SIST) [56], and MEME 5.4.1 [57], into Ori-Finder 2022 and tested the updated server on the web browsers, Firefox, Chrome, Safari, and Microsoft Edge.

Input file

By December 28, 2021, 91.5% of 362,223 bacterial genomes in the NCBI Genome database were draft genomes with scaffold or contig assembly levels. We updated Ori-Finder to enable *oriC* prediction to meet the imperative need to annotate *oriC*s in these genomes (**Figure 1A**). The updated web server can consequently handle complete or draft bacterial genomes with or without annotations. Ori-Finder 2022 integrates the gene-finding algorithm Prodigal [55] to predict protein-coding genes in unannotated genomes in the FASTA format. If an annotated genome file is uploaded in the GBK

format, the annotation information is automatically extracted by parsing text.

Updated prediction framework

Ori-Finder was originally developed with DNA asymmetry analysis using the Z-curve method, the distribution of DnaA boxes, and indicator genes close to *oriC*s [17]. Considering more *oriC* characteristics, the updated prediction framework of Ori-Finder 2022 adopts a new scoring criterion to quantitatively reflect these *oriC* characteristics of each intergenic sequence (IGS), and the IGSs with highest score are predicted as potential *oriC*s (**Figure 1B** and Table S1). As a characteristic of base composition, GC asymmetry is widely used for predicting *oriC*s. Ori-Finder 2022 scores the characteristics of base composition according to the distance to the minimum of the GC disparity (Table S1). Bacterial *oriC*s are usually adjacent to a *dnaA* gene, which can sever as an indicator for *oriC*s, but such genes are often different among bacterial species. Ori-Finder 2022 scores the characteristics of indicator genes, which can be adjusted based on the lineage and chromosome type entered by the user (Table S2). Ori-Finder 2022 scores DnaA boxes according to their numbers and mismatches. In addition, Ori-Finder 2022 identifies other functional elements of *oriC*, such as the Dam methylation site (GATC), and DnaA-trio, to screen prediction results if several IGSs with the same highest scores occur during the prediction process. For draft genomes, each sequence fragment will be predicted using Ori-Finder 2022, and all results will be considered together using the same prediction framework. Unlike the complete genome, the GC disparity minimum of each sequence fragment was used when scoring base composition.

Updated user interface

According to the updated prediction framework, the user interface for data submission was redesigned to enhance user experience (**Figure 2A**). Ori-Finder 2022 only requires users to upload the genome file in FASTA or GBK format to deliver a default *oriC* prediction; moreover, it provides some customization parameters. In Ori-Finder 2022, the principal indicator gene is *dnaA* by default and will be adjusted according to the lineage and chromosome type entered by users (Table S2). The default DnaA box is the standard motif (TTATCCACA) of *E. coli*, while the built-in DnaA box motif can be

selected according to the organism or lineage of the input genome. The drop-down checkboxes of the DnaA box motif and *dif* motif can achieve certain linkages for user convenience. Because of the diversity of DnaA boxes, Ori-Finder 2022 allows users to define their own DnaA box motifs. Users can select or define the *dif* motif in a similar way. Users can also choose strand-biased analysis for complete genomes.

Updated visualization module

The updated visualization module in Ori-Finder 2022 contains interactive Z-curve graph and characteristic visualization of *oriC* sequences. Global or local information of the genome can be grasped at a glance from the interactive Z-curve graph that displays the four disparity curves and the distribution of DnaA boxes, indicator genes, potential *oriC*s, and replication terminus (**Figure 2B**). The red, green, blue, and yellow line graphs indicate the AT, GC, RY, and MK disparity curves, respectively, calculated according to the Z-curve method. The purple vertical line displays the density of DnaA boxes, which is used to indicate the existence of DnaA box cluster. Red, dark blue, and light blue dotted lines indicate locations of the indicator genes, *oriC*s, and replication terminus, respectively. The indicator genes were identified by parsing the annotation information of the genome or BLAST with known protein sequences encoding indicator genes. Ori-Finder 2022 can also predict the replication terminus of a complete genome according to the *dif* motif or the maximum of GC disparity. Users can select all the information or only several datasets to analyze according to their requirements. The graph also supports the zoom function for analyzing the details. Moreover, when users hover the cursor over the dotted lines marking predicted *oriC*s, indicator genes or replication terminus, the exact location and other information are automatically displayed.

The other visualization result provided by Ori-Finder 2022 is the characteristic visualization of *oriC* sequence, which displays the distribution of its functional elements. The first part is the line graph (**Figure 2C**, top), which shows the transition probability of each base pair in the sequences calculated using Stress-Induced Duplex Destabilization method [56] that analyzes stress-driven DNA strand separation. Five lines with gradient colors were calculated using different negative superhelicity values,

and the peaks corresponded to the AT-rich sequence that might serve as a DUE. The second part is an *oriC* sequence schematic diagram showing the distribution of functional elements, such as DnaA box, DnaA-trio, ATP-DnaA box, GATC motif, and binding sites of CtrA, Fis, and IHF found in the predicted *oriC*s (Figure 2C, middle). The third part is the sequence of the predicted *oriC* in which the different elements are labeled with colors or symbols (Figure 2C, bottom). Indicator genes upstream and downstream of the predicted *oriC* are also labeled. In order to display the possible functional elements as comprehensively as possible, all possible DnaA trios are labeled, and a less conserved DnaA box with ≤ 4 mismatches from the standard DnaA box motif adjacent to a potential DnaA-trio will also be labeled, although its mismatch might be more than that entered by users.

Updated analysis module

Ori-Finder 2022 was expanded to include the new analysis modules. Combined with the different elements labeled in *oriC* sequence (Figure 2C), the annotation of corresponding regulatory proteins, such as Fis, SeqA, and CtrA (**Figure 2D**), by Ori-Finder 2022 might provide new insights into the related regulatory mechanisms. In addition, the repeat sequences in predicted *oriC*s were discovered by MEME are listed in a HTML table to reveal possible new motifs (**Figure 2E**). A strand-biased analysis can reveal the distribution of genes and bases on the leading and lagging strands of a complete genome (**Figure 2F**). Sequences homologous to predicted *oriC*s were searched using BLAST against the DoriC database [31], and the BLAST results linked to the corresponding entry in the DoriC database are also provided (**Figure 2G**).

Result and discussion

Here, *Yersinia pestis* KIM+ is presented to illustrate details of the predicted results of Ori-Finder 2022. The structure of *oriC* in *Y. pestis* KIM+ is similar to that in *E. coli* [58]. Figure 2 shows the main visualization and analytical results of the *oriC* predicted by Ori-Finder 2022 and the complete predicted results are available as a Sample result

at our website (http://tubic.tju.edu.cn/Ori-Finder2022/public/index.php/retrieve/sample_result). Due to possible rearrangement, the four disparity curves of this genome fluctuate at their extrema [58], which does not seem to provide sufficient evidence to identify an *oriC* (Figure 2B). Ori-Finder 2022 identified an IGS of 380 base pairs as the potential *oriC* by taking more characteristics into consideration, such as indicator genes, DnaA box cluster, and other functional elements. Like that in *E. coli*, the predicted *oriC* in *Y. pestis* KIM+ was located between *gidA* and *mioC*. The sequence corresponding to the peak of the lines calculated by SIST also contained DnaA-trios and three ATP-DnaA boxes (AGATCT), which was likely to contain a site of DNA duplex unwinding (Figure 2C). The genome of *Y. pestis* KIM+ encodes regulatory proteins such as Fis, SeqA, IHF, and Dam (Figure 2D), and the possible binding sites for corresponding proteins are also found in the predicted *oriC*. Although the genome of *Y. pestis* KIM+ does not appear to encode CtrA proteins, two possible CtrA binding sites were identified within the predicted *oriC*. The repeat sequences in the predicted *oriC*s were discovered using MEME, which might reveal new *oriC* motifs. For example, two of the five motifs in the first set (ARGATC) overlapped with predicted ATP-DnaA boxes. In the second set (GTTATGCACAT), three of the five motifs overlapped with the predicted DnaA boxes, and the other two contained DnaA box-like motifs with three and four mismatches, respectively, from the perfect DnaA box (TTATCCACA) in *E.coli*. A *dif* site was located near the top of the GC disparity curve. Strand-biased analysis reveals the differences in some features between the leading and lagging strands. The lengths of the leading (50.66%) and lagging (49.34%) strands were almost identical. The leading strand included 2296 (56.76%) genes, probably a result of rearrangement during which the strand-biased phenomenon of genes is not obvious. The adenine (A), thymine (T), guanine (G), cytosine (C), purine (A+G), and pyrimidine (C+T) contents of the leading and lagging strands are also calculated (Figure 2F). The predicted result was considered reliable because homologous sequences were found in the DoriC database (Figure 2G).

Conclusion

Ori-Finder has been widely applied by biologists over the past decade to predict bacterial *oriC*s, and some predictions have been experimentally confirmed [21-24] or supported by various studies [45, 59]. For example, the *oriC*s of 132 gut microbes in metagenomic samples predicted by metagenomic analyses and Ori-Finder were consistent ($R^2 = 0.98$, $P < 10^{-30}$) [59]. The bacterial *oriC* element, DnaA-trio, was found in 85% of *oriC*s predicted or confirmed from > 2000 species. Numerous bacterial *oriC*s generated by Ori-Finder have been used for large-scale data mining and analysis. Ori-Finder 2022 can now predict *oriC*s in complete or draft genomes based on an updated prediction framework and provides an interactive visualization module as well as a new analysis module. Ori-Finder will be continuously improved by incorporating state-of-the-art research results and integrating additional analysis modules. We plan to provide users with an integrated platform for comprehensive prediction, analysis, and knowledge mining to determine microbial replication origins. This will be achieved by integrating Ori-Finder 2 [60] that predicts archaeal replication origins, and Ori-Finder 3 [61], an online service for predicting replication origins in *Saccharomyces cerevisiae* in the future.

Availability

Ori-Finder 2022 is freely available at:

http://tubic.tju.edu.cn/Ori-Finder2022 and https://tubic.org/Ori-Finder2022.

CRediT author statement

Meijing Dong: Software, Writing-original draft, Visualization. **Hao Luo:** Software, Writing-original draft, Visualization. **Feng Gao:** Conceptualization, Writing-review & editing, Supervision, Funding acquisition. All authors read and approved the final manuscript.

Competing interests

The authors have declared no competing interests.

Acknowledgments

This work was supported by the National Key R&D Program of China (Grant No. 2018YFA0903700); and the National Natural Science Foundation of China [Grant Nos 21621004 and 31571358]. We thank Professor Chun-Ting Zhang for the invaluable assistance and inspiring discussions.

ORCID

0000-0002-4833-1939 (Meijing Dong)

0000-0003-2714-8817 (Hao Luo)

0000-0002-9563-3841 (Feng Gao)

References

[1] Grimwade JE, Leonard AC. Blocking the trigger: inhibition of the initiation of bacterial chromosome replication as an antimicrobial strategy. Antibiotics 2019;8:111.

[2] Yoneji T, Fujita H, Mukai T, Su'etsugu M. Grand scale genome manipulation via chromosome swapping in *Escherichia coli* programmed by three one megabase chromosomes. Nucleic Acids Res 2021;49:8407–18.

[3] Yue H, Ling C, Yang T, Chen X, Chen Y, Deng H, et al. A seawater-based open and continuous process for polyhydroxyalkanoates production by recombinant *Halomonas campaniensis* LS21 grown in mixed substrates. Biotechnol Biofuels 2014;7:108.

[4] Moriya S, Ogasawara N. Mapping of the replication origin of the *Bacillus subtilis* chromosome by the two-dimensional gel method. Gene 1996;176:81–4.

[5] Oka A, Sugimoto K, Takanami M, Hirota Y. Replication origin of the *Escherichia coli* K-12 chromosome: the size and structure of the minimum DNA segment carrying the information for autonomous replication. Mol Gen Genet 1980;178:9–20.

[6] Bird RE, Louarn J, Martuscelli J, Caro L. Origin and sequence of chromosome replication in *Escherichia coli*. J Mol Biol 1972;70:549–66.

[7] Khodursky AB, Peter BJ, Schmidt MB, DeRisi J, Botstein D, Brown PO, et al. Analysis of topoisomerase function in bacterial replication fork movement: use of DNA microarrays. Proc Natl Acad Sci U S A 2000;97:9419–24.

[8] Srivatsan A, Han Y, Peng J, Tehranchi AK, Gibbs R, Wang JD, et al. High-precision, whole-genome sequencing of laboratory strains facilitates genetic studies. PLoS Genet 2008;4:e1000139.

[9] Song CC, Zhang SC, Huang H. Choosing a suitable method for the identification of replication origins in microbial genomes. Front Microbiol 2015;6:1049.

[10] Lobry JR. Asymmetric substitution patterns in the two DNA strands of bacteria. Mol Biol Evol 1996;13:660–5.

[11] Grigoriev A. Analyzing genomes with cumulative skew diagrams. Nucleic Acids Res 1998;26:2286–90.

[12] Salzberg SL, Salzberg AJ, Kerlavage AR, Tomb JF. Skewed oligomers and origins of replication. Gene 1998;217:57 – 67.

[13] Frank AC, Lobry JR. Oriloc: prediction of replication boundaries in unannotated bacterial chromosomes. Bioinformatics 2000;16:560–1.

[14] Ohbayashi R, Hirooka S, Onuma R, Kanesaki Y, Hirose Y, Kobayashi Y, et al. Evolutionary changes in DnaA-dependent chromosomal replication in Cyanobacteria. Front Microbiol 2020;11:786.

[15] Gao F, Zhang CT. Origins of replication in *Cyanothece* 51142. Proc Natl Acad Sci U S A 2008;105:E125.

[16] Mackiewicz P, Zakrzewska-Czerwinska J, Zawilak A, Dudek MR, Cebrat S. Where does bacterial replication start? Rules for predicting the *oriC* region. Nucleic Acids Res 2004;32:3781–91.

[17] Gao F, Zhang CT. Ori-Finder: a web-based system for finding *oriC*s in unannotated bacterial genomes. BMC Bioinformatics 2008;9:79.

[18] Graf JS, Schorn S, Kitzinger K, Ahmerkamp S, Woehle C, Huettel B, et al. Anaerobic endosymbiont generates energy for ciliate host by denitrification. Nature 2021;591:445–50.

[19] Lawrence D, Campbell DE, Schriefer LA, Rodgers R, Walker FC, Turkin M, et al. Single-cell genomics for resolution of conserved bacterial genes and mobile genetic elements of the human intestinal microbiota using flow cytometry. Gut Microbes 2022;14:2029673.

[20] Park J, Kim M, Shin B, Kang M, Yang J, Lee TK, et al. A novel decoy strategy for polymyxin resistance in *Acinetobacter baumannii*. eLife 2021;10:e66988.

[21] Huang H, Song CC, Yang ZL, Dong Y, Hu YZ, Gao F. 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.

[22] Li P, Zhang J, Deng Z, Gao F, Ou HY. Identification and characterization of a central replication origin of the mega-plasmid pSCATT of *Streptomyces cattleya*. Microbiol Res 2022;257:126975.

[23] Chen AH, Afonso B, Silver PA, Savage DF. Spatial and temporal organization of chromosome duplication and segregation in the cyanobacterium *Synechococcus elongatus* PCC 7942. PLoS One 2012;7:e47837.

[24] Blötz C, Lartigue C, Valverde Timana Y, Ruiz E, Paetzold B, Busse J, et al. Development of a replicating plasmid based on the native *oriC* in *Mycoplasma pneumoniae*. Microbiology 2018;164:1372–82.

[25] Gao F. Bacteria may have multiple replication origins. Front Microbiol 2015;6:324.

[26] Zhang Y, Dong MJ, Gao F. Recent advances in multiple replication origins within a single prokaryotic chromosome. Chem. Life 2021;11:2394–400 (in Chinese with an English abstract).

[27] Ionescu D, Bizic-Ionescu M, De Maio N, Cypionka H, Grossart H-P. Community-like genome in single cells of the sulfur bacterium *Achromatium oxaliferum*. Nat Commun 2017;8:455.

[28] Bruhn M, Schindler D, Kemter FS, Wiley MR, Chase K, Koroleva GI, et al. Functionality of two origins of replication in *Vibrio cholerae* strains with a single chromosome. Front Microbiol 2018;9:2932.

[29] Gao F, Zhang CT. DoriC: a database of *oriC* regions in bacterial genomes.

Bioinformatics 2007;23:1866–7.

[30] Gao F, Luo H, Zhang CT. DoriC 5.0: an updated database of *oriC* regions in both bacterial and archaeal genomes. Nucleic Acids Res 2013;41:D90–3.

[31] Luo H, Gao F. DoriC 10.0: an updated database of replication origins in prokaryotic genomes including chromosomes and plasmids. Nucleic Acids Res 2019;47:D74–7.

[32] Gao F. Recent advances in the identification of replication origins based on the Z-curve method. Curr Genomics 2014;15:104–12.

[33] Luo H, Quan CL, Peng C, Gao F. Recent development of Ori-Finder system and DoriC database for microbial replication origins. Brief Bioinform 2019;20:1114–24.

[34] Barzantny H, Schröder J, Strotmeier J, Fredrich E, Brune I, Tauch A. The transcriptional regulatory network of *Corynebacterium jeikeium* K411 and its interaction with metabolic routes contributing to human body odor formation. J Biotechnol 2012;159:235–48.

[35] Pelliciari S, Dong MJ, Gao F, Murray H. Evidence for a chromosome origin unwinding system broadly conserved in bacteria. Nucleic Acids Res 2021;49:7525–36.

[36] Necşulea A, Lobry JR. A new method for assessing the effect of replication on DNA base composition asymmetry. Mol Biol Evol 2007;24:2169–79.

[37] Zhang G, Gao F. Quantitative analysis of correlation between AT and GC biases among bacterial genomes. PLoS One 2017;12:e0171408.

[38] Chen WH, Lu G, Bork P, Hu S, Lercher MJ. Energy efficiency trade-offs drive nucleotide usage in transcribed regions. Nat Commun 2016;7:11334.

[39] Quan CL, Gao F. Quantitative analysis and assessment of base composition asymmetry and gene orientation bias in bacterial genomes. FEBS Lett 2019;593:918–25.

[40] 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;16:91–104.

[41] Wannier TM, Ciaccia PN, Ellington AD, Filsinger GT, Isaacs FJ, Javanmardi K, et al. Recombineering and MAGE. Nat Rev Methods Primers 2021;1:7.

[42] Ekundayo B, Bleichert F. Origins of DNA replication. Plos Genet 2019;15:e1008320.

[43] Ozaki S, Fujimitsu K, Kurumizaka H, Katayama T. The DnaA homolog of the hyperthermophilic eubacterium *Thermotoga maritima* forms an open complex with a minimal 149-bp origin region in an ATP-dependent manner. Genes Cells 2006;11:425–38.

[44] Speck C, Messer W. Mechanism of origin unwinding: sequential binding of DnaA

to double- and single-stranded DNA. EMBO J 2001;20:1469–76.

[45] Richardson TT, Harran O, Murray H. The bacterial DnaA-trio replication origin element specifies single-stranded DNA initiator binding. Nature 2016;534:412–6.

[46] Richardson TT, Stevens D, Pelliciari S, Harran O, Sperlea T, Murray H. Identification of a basal system for unwinding a bacterial chromosome origin. EMBO J 2019;38:e101649.

[47] Grimwade JE, Leonard AC. Blocking, bending, and binding: regulation of initiation of chromosome replication during the *Escherichia coli* cell cycle by transcriptional modulators that interact with origin DNA. Front Microbiol 2021;12:732270.

[48] Jaworski P, Zyla-Uklejewicz D, Nowaczyk-Cieszewska M, Donczew R, Mielke T, Weigel C, et al. Putative cooperative ATP–DnaA binding to double-stranded DnaA box and single-stranded DnaA-trio motif upon *Helicobacter pylori* replication initiation complex assembly. Int J Mol Sci 2021;22:6643.

[49] Chatterjee S, Jha JK, Ciaccia P, Venkova T, Chattoraj DK. Interactions of replication initiator RctB with single- and double-stranded DNA in origin opening of *Vibrio cholerae* chromosome 2. Nucleic Acids Res 2020;48:11016–29.

[50] Chung YS, Brendler T, Austin S, Guarne A. Structural insights into the cooperative binding of SeqA to a tandem GATC repeat. Nucleic Acids Res 2009;37:3143–52.

[51] Marczynski GT, Shapiro L. Control of chromosome replication in *Caulobacter crescentus*. Annu Rev Microbiol 2002;56:625–56.

[52] Brassinga AKC, Siam R, McSween W, Winkler H, Wood D, Marczynski GT. Conserved response regulator CtrA and IHF binding sites in the alpha-proteobacteria *Caulobacter crescentus* and *Rickettsia prowazekii* chromosomal replication origins. J Bacteriol 2002;184:5789–99.

[53] Wolański M, Donczew R, Zawilak-Pawlik A, Zakrzewska-Czerwińska J. *oriC*-encoded instructions for the initiation of bacterial chromosome replication. Front Microbiol 2015;5:735.

[54] Boratyn GM, Camacho C, Cooper PS, Coulouris G, Fong A, Ma N, et al. BLAST: a more efficient report with usability improvements. Nucleic Acids Res 2013;41:W29–33.

[55] Hyatt D, Chen GL, LoCascio PF, Land ML, Larimer FW, Hauser LJ. Prodigal: prokaryotic gene recognition and translation initiation site identification. BMC Bioinformatics 2010;11:119.

[56] Zhabinskaya D, Madden S, Benham CJ. SIST: stress-induced structural transitions in superhelical DNA. Bioinformatics 2014;31:421–2.

[57] Bailey TL, Elkan C. Fitting a mixture model by expectation maximization to

485 discover motifs in biopolymers. Proc Int Conf Intell Syst Mol Biol 1994;2:28–36.

486 [58] Deng W, Burland V, Plunkett G, Boutin A, Mayhew GF, Liss P, et al. Genome
487 sequence of *Yersinia pestis* KIM. J Bacteriol 2002;184:4601–11.

488 [59] Korem T, Zeevi D, Suez J, Weinberger A, Avnit-Sagi T, Pompan-Lotan M, et al.
489 Microbiome growth dynamics of gut microbiota in health and disease inferred
490 from single metagenomic samples. Science 2015;349:1101–6.

491 [60] Luo H, Zhang CT, Gao F. Ori-Finder 2, an integrated tool to predict replication
492 origins in the archaeal genomes. Front Microbiol 2014;5:482.

493 [61] Wang D, Lai FL, Gao F. Ori-Finder 3: a web server for genome-wide prediction of
494 replication origins in *Saccharomyces cerevisiae*. Brief Bioinform
495 2020;22:bbaa182.

496

497

Figure legends

Figure 1 Workflow of Ori-Finder 2022

A. Input file of Ori-Finder 2022. Users can submit complete or draft genome in FASTA or GBK format. **B**. Prediction framework of Ori-Finder 2022. Ori-Finder 2022 predicts *oriC*s by comprehensively assessing DNA asymmetry, indicator genes, and *oriC* functional elements. **C**. Visualization module of Ori-Finder 2022. **D**. Analysis module of Ori-Finder 2022. *oriC*, replication origin; IGS, intergenic sequence.

Figure 2 User interface and predicted result of Ori-Finder 2022 for *Yersinia pestis* KIM+

A. User interface of Ori-Finder 2022. **B.** Interactive Z-curve graph (original Z-curve graph and the one rotated at maximum of GC disparity) with *oriC* related information. **C**. Sequence characteristic visualization of the predicted *oriC*. **D**. Table in HTML providing basic information of *Y. pestis* KIM+ genome and its potential *oriC*. Genes encoding regulatory proteins of chromosome replication in the genome is provided here. **E**. Repeat sequence discovery by MEME. **F**. Strand-biased analysis. Pie and bar charts respectively show distribution of genes and bases in leading and lagging strands. **G**. Homologous *oriC* search by BLAST. *oriC*, replication origin.

Supplementary material

Table S1 Scoring criterion of Ori-Finder 2022

Table S2 Indicator genes list for different chromosome type and lineage

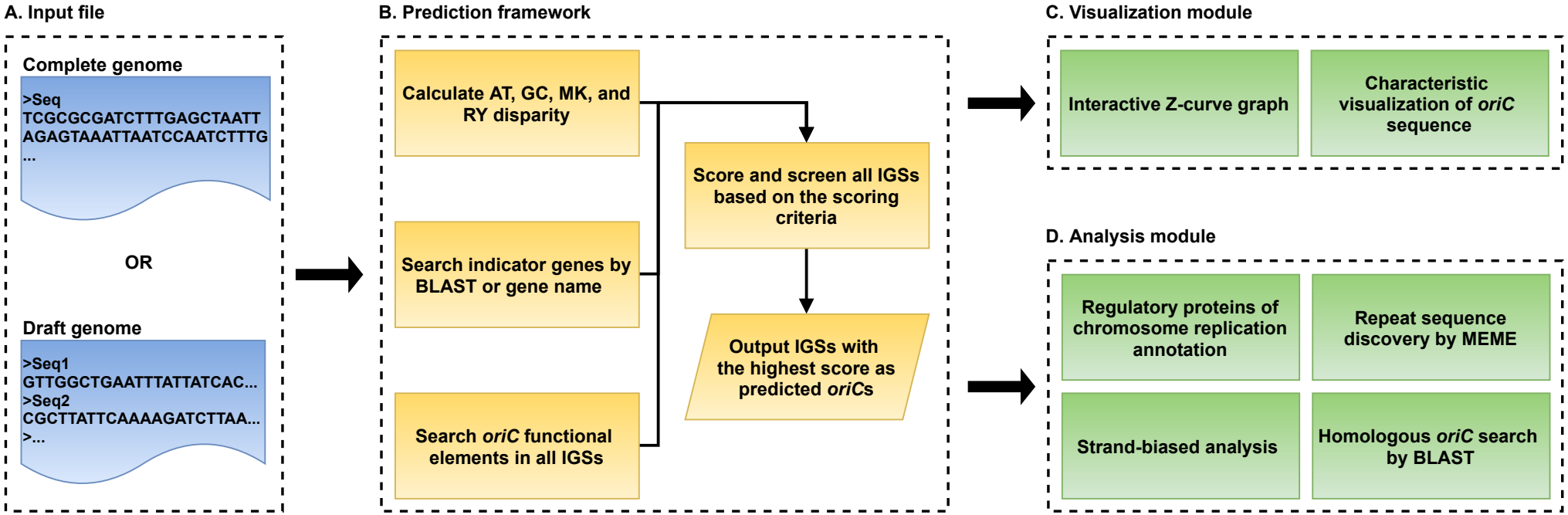
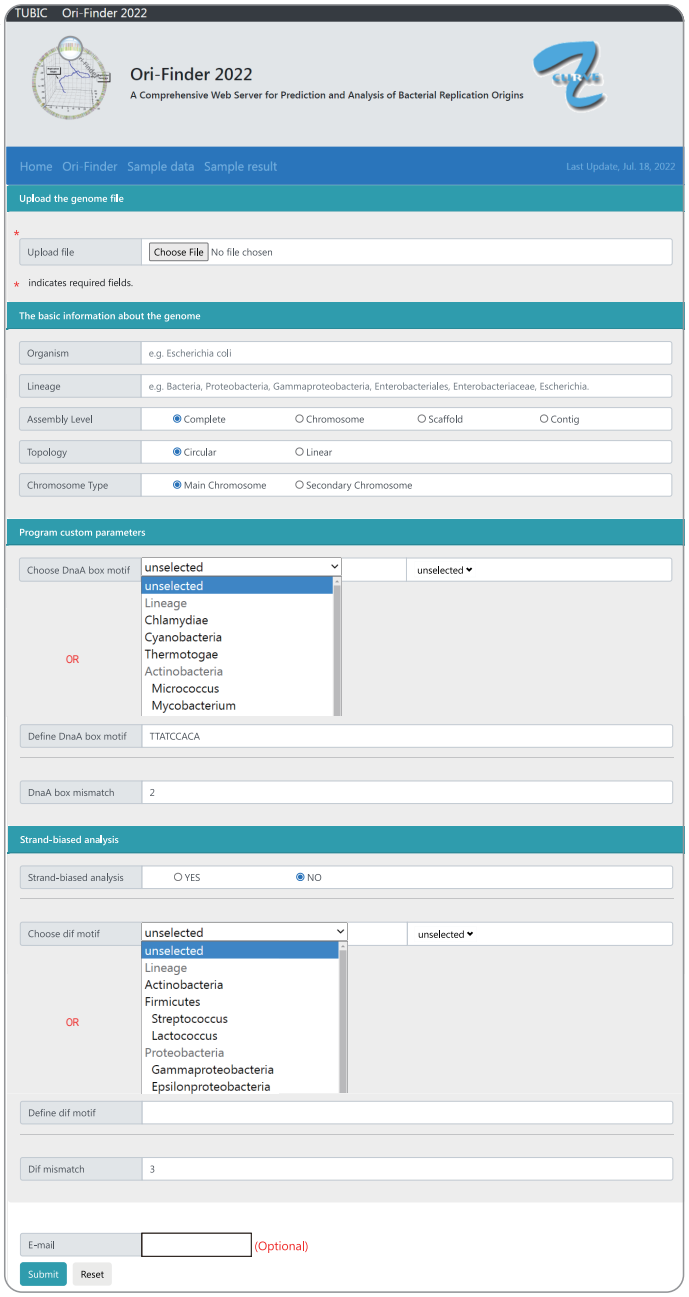
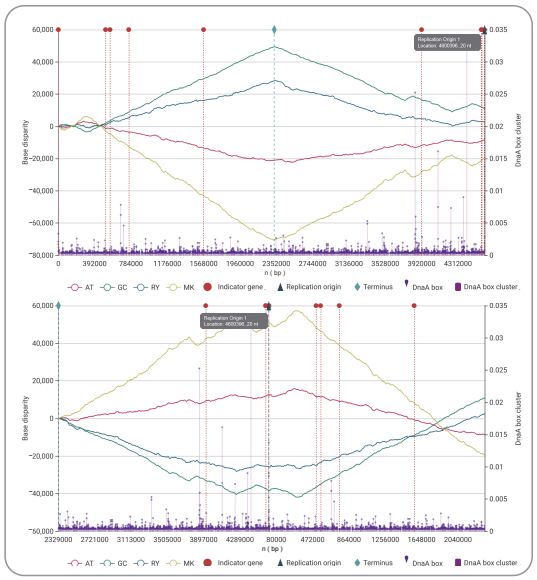

A. Input file
Complete genome
>Seq
TCGCGCGATCTTTGAGCTAATT
AGAGTAAATTAATCCAATCTTTG
...
OR
Draft genome
>Seq1
GTTGGCTGAATTTATTATCAC...
>Seq2
CGCTTATTCAAAAGATCTTAA...
>...
B. Prediction framework
Calculate AT, GC, MK, and RY disparity
Search indicator genes by BLAST or gene name
Search oriC functional elements in all IGSs
Score and screen all IGSs based on the scoring criteria
Output IGSs with the highest score as predicted oriCs
C. Visualization module
Interactive Z-curve graph
Characteristic visualization of oriC sequence
D. Analysis module
Regulatory proteins of chromosome replication annotation
Repeat sequence discovery by MEME
Strand-biased analysis
Homologous oriC search by BLAST

Figure 1

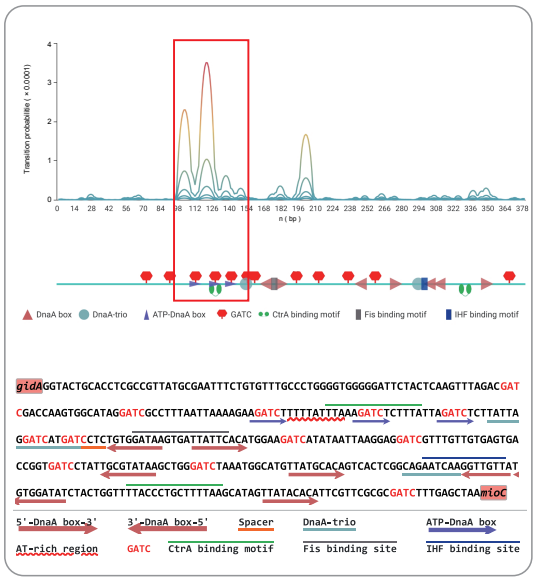
A



B



C



D

Organism	Yersinia pestis KIM10+
RefSeq	NC_004088.1
Topology	Circular
Lineage	Bacteria, Proteobacteria, Gammaproteobacteria, Enterobacteriales, Enterobacteriaceae, Yersinia.
Assembly level	Complete
Chromosome size	4600755 nt
Chromosome GC content	0.48
The extremes of GC disparity	307442 nt (minimum), 2328783 nt (maximum)
Gene encoding regulatory protein of chromosome replication	*fis* 229,845 ... 230,142 nt *seqA* 1,404,052 ... 1,404,580 nt *ihfA* 2,106,185 ... 2,106,482 nt *ihfB* 3,073,509 ... 3,073,794 nt *dam* 4,391,550 ... 4,392,366 nt
Replication origins	You can click the link here to view the detailed information of each replication origin. • Replication Origin 1 4,600,396 ... 20 nt

E

Motif ARGATC

Start	P-value		Site	
210	2.04e-04	ATAATTAAGG	AGGATC	GTTTGTTGTG
152	2.04e-04	ATCTCTTATT	AGGATC	ATGATCCTCT
89	2.04e-04	CAAGTGGCAT	AGGATC	GCCTTTAATT
192	4.27e-04	ATTCACATGG	AAGATC	ATATAATTAA
126	4.27e-04	TTTTTATTTA	AAGATC	TCTTTATTAG

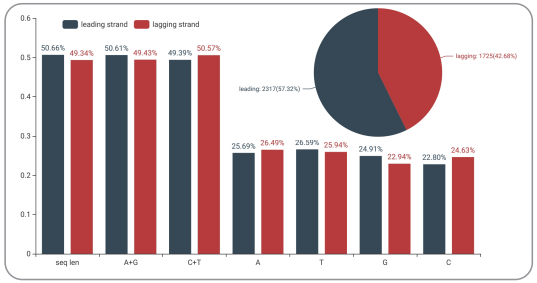
Motif GTTATGCACAT

Start	P-value		Site	
272	1.09e-06	TAAATGGCAT	GTTATGCACAG	TCACTCGGCA
346	1.74e-06	TTTAAGCATA	GTTATACACAT	TCGTTCGCGC
179	1.33e-05	TGGATAAGTG	ATTATTCACAT	GGAAGATCAT
17	1.39e-05	GCACCTCGCC	GTTATGCGAAT	TTCTGTGTTT
2	1.43e-05	G	GTACTGCACCT	CGCCGTTATG

Motif CCCTGS

Start	P-value		Site	
329	7.25e-05	ACTGGTTTTA	CCCTGC	TTTTAAGCAT
39	1.69e-04	TCTGTGTTTG	CCCTGG	GGTGGGGGAT

F



G

subject_id	identity	aln_length	mismatches	gaps	q_start	q_end	s_start	s_end	e_value	bit_score
ORI96000376	100.0	380	0	0	1	380	380	1	0.0	702
ORI96000375	100.0	380	0	0	1	380	1	380	0.0	702
ORI96000374	100.0	380	0	0	1	380	1	380	0.0	702
ORI93020117	100.0	380	0	0	1	380	1	380	0.0	702
ORI90020045	100.0	380	0	0	1	380	380	1	0.0	702
ORI20020031	100.0	380	0	0	1	380	1	380	0.0	702
ORI20020030	100.0	380	0	0	1	380	1	380	0.0	702
ORI10020019	99.74	380	1	0	1	380	1	380	0.0	697
ORI92520061	99.47	380	2	0	1	380	380	1	0.0	691

Figure 2