Published online 28 October 2022 *Nucleic Acids Research, 2023, Vol. 51, Database issue* ***D117–D120***
https://doi.org/10.1093/nar/gkac964

DoriC 12.0: an updated database of replication origins in both complete and draft prokaryotic genomes

Mei-Jing Dong[1,†], Hao Luo [1,†] and 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 and [3]SynBio Research Platform, Collaborative Innovation Center of Chemical Science and Engineering (Tianjin), Tianjin 300072, China

Received September 15, 2022; Revised October 08, 2022; Editorial Decision October 10, 2022; Accepted October 13, 2022

ABSTRACT

DoriC was first launched in 2007 as a database of replication origins (*oriC*s) in bacterial genomes and has since been constantly updated to integrate the latest research progress in this field. The database was subsequently extended to include the *oriC*s in archaeal genomes as well as those in plasmids. This latest release, DoriC 12.0, includes the *oriC*s in both draft and complete prokaryotic genomes. At the same time, the number of *oriC*s in the database has also increased significantly and currently contains over 200 000 bacterial entries distributed in more than 40 phyla. Among them, a large number are from bacteria in new phyla whose *oriC*s were not explored before. Additionally, new *oriC* features and improvements have been introduced, especially in the visualization and analysis of *oriC*s. Currently, DoriC is considered as an important database in the fields of bioinformatics, microbial genomics, and even synthetic biology, providing a valuable resource as well as a comprehensive platform for the research on *oriC*s. DoriC 12.0 can be accessed at https://tubic.org/doric/ and http://tubic.tju.edu.cn/doric/.

INTRODUCTION

In general, DNA replication is initiated at replication origins, which encode the information instructions of replication initiation and regulation (1,2). The bacterial replication origin (*oriC*) usually includes DnaA-binding sites, binding motifs of replication regulatory proteins, and an AT-rich DNA unwinding element (DUE). Similarly, the archaeal *oriC*s consist of a DUE and a number of conserved repeats known as origin recognition boxes (ORBs), which act as binding sites for the origin recognition proteins (3). With the exception of the conserved core *oriC* functional elements (e.g. DnaA box and DnaA-trios), the functional elements within *oriC*s are highly diverse, revealing the complexity of the replication initiation and regulation mechanisms. Thus, the identification of *oriC*s is essential for understanding their structure and functions, which would provide further insights into the regulatory mechanisms of the initiation step in DNA replication. Compared with experimental means, the *in silico* prediction of *oriC*s is more efficient and economical, rendering the method especially suitable for dealing with large-scale genome sequences. However, a variety of factors make the prediction of *oriC*s more complicated than expected, such as atypical GC skews, species-specific DnaA boxes, and loss of the *dnaA* gene by genome reduction. Consequently, the *oriC*s in a number of sequenced genomes could not be identified and some are even annotated incorrectly in the genome reports. To address this problem, we had carried out systematic research on *oriC*s over the last two decades and developed the Ori-Finder system in 2008 using the Z-curve theory and comparative genomics method (4,5). At the same time, we also launched DoriC, a database of *oriC*s in prokaryotic genomes based on the prediction of the Ori-Finder system. Initially, the DoriC database only contained predicted or experimentally confirmed *oriC*s in hundreds of bacterial chromosomes (6). Subsequently, DoriC 5.0 provided more than 1000 *oriC*s in both bacterial and archaeal chromosomes (7), and DoriC 10.0 included over 10 000 *oriC*s of chromosomes and plasmids in prokaryotic genomes (8). In addition to the information on *oriC*s, graphical views of GC, AT, RY and MK disparity curves and relevant information about the species or genome (e.g. organism, lineage, chromosome topology, *dnaA* gene position) have also been supplied to users. Moreover, the DoriC database provides the basic search and BLAST functions. Currently, the DoriC database is considered as an important database in the fields of bioinformatics, microbial genomics, and even synthetic biology, providing a valuable resource for relevant research on *oriC*s (9–11). The DoriC database has facilitated research on *oriC* functional elements (12), the replication mechanism (13), strand-biased analysis (14,15) and

*To whom correspondence should be addressed. Tel: +86 22 27404118; Fax: +86 22 27404118; Email: fgao@tju.edu.cn
†The authors wish it to be known that, in their opinion, the first two authors should be regarded as Joint First Authors.

large-scale *oriC* data mining (16). Additionally, the database is also used in research involving metagenomics, such as the study of bacterial growth dynamics (17–19) and *oriC* prediction (20).

With the rapid development of genome sequencing technology and the continuous reduction in sequencing costs, increasingly more prokaryotic genomes are being sequenced. However, as of 9 April 2022, complete genomes only accounted for ∼8% of all available bacterial genomes deposited in the National Center for Bioinformatics Information (NCBI) databases. To make full use of these genome data and facilitate user access to more comprehensive information about *oriC*s, DoriC has been updated to the new version, DoriC 12.0, which presents with updates in *oriC* numbers as well as contents and functions.

DATABASE UPDATES

Significant increase in *oriC* data

Complete and draft genomes together with their available annotation files were downloaded from the NCBI RefSeq database (https://www.ncbi.nlm.nih.gov/refseq/) on 9 April 2022 (21,22). First, the *oriC*s in bacterial genomes were predicted using Ori-Finder 2022 (https://tubic.org/Ori-Finder2022/) (23), an updated version of Ori-Finder, which adopts a new scoring system to consider the currently available *oriC* characteristics comprehensively and shows better prediction performance compared with the original version. Then, the predicted *oriC*s were added to the DoriC database after manual curation. Consequently, the DoriC database has been substantially expanded by including the *oriC*s predicted by Ori-Finder 2022. Now, the bacterial chromosomes included in DoriC 12.0 have increased from 7580 to 205,798 distributed in more than 40 phyla.

Updated functions

New search functions. DoriC 10.0 supported the search of RefSeq ID, DoriC ID, organism, lineage and other fields. The extended search functions in DoriC 12.0 make it convenient for users to search and use the database more efficiently. The search interfaces are designed according to the different characteristics of the bacterial, archaeal and plasmid data (Figure 1A), and the search results can be downloaded directly. For example, the search fields provided by the search interface of bacterial data include genome assembly level, chromosome topology, chromosome type, *oriC* type, regulatory protein, and gene flanking the *oriC*.

New tool. Although the lengths and nucleotide sequences of different *oriC*s vary considerably, bacterial *oriC*s share similar functional elements (24). Therefore, DoriC 12.0 provides a tool for comparative analysis of the distribution of functional elements in bacterial *oriC*s. Users can enter a comma-separated DoriC ID list, extract the *oriC* maps recorded in DoriC 12.0 for comparison, and thereby determine if there are any similarities or differences. Additionally, users can also upload one or multiple sequences in a FASTA file to generate the corresponding *oriC* maps. For example, although the lengths and sequences of the *oriC*s of *Caulobactor crescentus* and *Caulobactor segnis* are not completely consistent, the number and relative position of CtrA-binding motifs seem to be conserved (Figure 1B) (25). These *oriC* maps may also be helpful in the analysis of *oriC* sequences acquired by users through experiments, metagenomic data analysis, or other methods.

Updated contents

Redesigned web page. To enhance user experience, the web page has been redesigned to be more intuitive. The home page displays statistical information (e.g. the number of bacterial chromosomes at different assembly levels) through which users can gain a general understanding of the DoriC database. Each record in DoriC 12.0 usually contains two page: the primary page (Figure 1C) showing information about the genome and the secondary page (Figure 1E) showing information about each *oriC*.

Interactive Z-curve figure. The Z-curve figure shows the four disparity curves, the distribution of DnaA boxes, indicator genes, potential *oriC*s and replication terminus, making it convenient for users to observe the relative positions of these features (Figure 1D). The redesigned Z-curve figure is interactive, allowing users to zoom into the figure to view more details on the changes in the base composition, select all or only a few sets of data for analysis, and hover over a particular site to see the details of indicator genes, potential *oriC*s, and replication terminus.

Characteristic visualization of oriC sequences. The characteristic visualization of *oriC* sequences ('*oriC* map' and '*oriC* sequences' in Figure 1E) is designed to help users observe the distribution of functional elements within the *oriC*s and to facilitate exploration of the binding mechanism of the initiator proteins and regulatory proteins. At present, three main types of bacterial functional elements can be visualized in DoriC 12.0: (i) binding motifs of initiator protein DnaA, such as DnaA boxes, DnaA-trios, and the ATP-DnaA boxes (26); (ii) binding motifs of the replication regulatory proteins SeqA, CtrA, Fis and IHF (27–30); and (iii) AT-rich sequences that might serve as DUEs (31). To facilitate potential new discoveries, all possible functional elements have been searched for within the *oriC*s recorded in DoriC 12.0. However, it should be noted that the predicted binding motifs may not always be functional as real binding sites, since many of these are restricted to specific lineages (e.g. Fis, IHF and CtrA binding sites are usually found in a subset of Proteobacteria) and are not present throughout the bacterial domain (same GATC methylation sites). The default display of these sites is also set according to the lineage of species. Additionally, the characteristic visualization of *oriC* sequences is interactive, and users can select functional elements for observation by clicking corresponding buttons. For the archaeal *oriC*s, the annotated functional elements are mainly ORBs and AT-rich sequences.

More information on each record. More information has been added to each record to assist users in analyzing the *oriC* comprehensively. A list of genes encoding the regulatory proteins of chromosome replication is provided. Combined with the different binding motifs annotated in the

Figure 1. Overview of the DoriC 12.0 interface. (**A**) New search functions. Different search interfaces are designed for search of bacterial, archaeal and plasmid data. (**B**) New tool for comparing the distribution of functional elements in the *oriC*s. The *oriC* maps of *Caulobacter crescentus* and *Caulobacter segnis* generated by the new tool have been presented here as an example. (**C**) Primary page of each record. The primary page presents basic information about the genome, genes encoding the regulatory proteins of chromosome replication, and interactive Z-curve figures. (**D**) Interactive Z-curve figures. (**E**) Secondary page of each record. The secondary page presents information of each *oriC*, such as characteristic visualization of *oriC* sequence, repeat sequences discovered by MEME, strand-biased analysis and homologous *oriC*s searched by BLAST.

'characteristic visualization of *oriC* sequences', some new insights related to regulatory mechanisms may be gained. The repeat sequences discovered by the Multiple EM for Motif Elicitation (MEME) tool in predicted *oriC*s may reveal new functional elements (32). The differences in gene distributions and base compositions between the leading and lagging strands are revealed by the strand-biased analysis. Additionally, the homology search results of each *oriC* are provided in this release, making comparative analyses of the *oriC* sequences convenient for users (33).

CONCLUDING REMARKS

The recent development of high-throughput sequencing technology has resulted in a rapidly growing number of prokaryotic genome sequences, especially of bacterial genomes. However, a large number of the *oriC*s among these genomes remain unknown, necessitating the development of bioinformatics algorithms to address this issue. Based on the prediction results of Ori-Finder 2022, over 200 000 bacterial *oriC*s have been predicted reliably. This latest release of the DoriC database provides a substantial number of novel *oriC*s of the bacteria belonging to the new phyla, which have not been discovered before. This will help researchers to further reveal the conserved and diverse features of *oriC*s, such as their DnaA boxes or other functional elements, which will comprehensively enrich our knowledge about these sequences. Artificial intelligence, especially deep learning, has changed the paradigm of scientific research profoundly. However, it requires massive amounts of training data, something which DoriC 12.0 can provide, such as high-quality *oriC* data that can be used for data mining via deep learning. New discoveries based on these *oriC*s are expected, leading to a better understanding of the DNA replication mechanisms. In the future, we will carry out further systematic studies of different types of *oriC*s to extract their respective features and provide more relevant services for their analysis, making the DoriC database a universal knowledge base on *oriC*s. The elucidation and application of principles of *oriC*s based on over 200 000 *oriC* regions in prokaryotic genomes will contribute to the development of *oriC* prediction algorithms as well as experimental confirmation and functional analysis of *oriC*s in prokaryotic genomes.

DATA AVAILABILITY

DoriC 12.0 is freely available to the public without registration or login requirements (https://tubic.org/doric/ and http://tubic.tju.edu.cn/doric/).

ACKNOWLEDGEMENTS

We thank Professor Chun-Ting Zhang for the invaluable assistance and inspiring discussions.

FUNDING

National Key Research and Development Program of China [2018YFA0903700 to F.G.]; National Natural Science Foundation of China [21621004 to F.G., 31801104 to H.L.]. Funding for open access charge: National Key Research and Development Program of China.
Conflict of interest statement. None declared.

REFERENCES

1. Hill,N.S., Kadoya,R., Chattoraj,D.K. and Levin,P.A. (2012) Cell size and the initiation of DNA replication in bacteria. *PLoS Genet.,***8**, e1002549.
2. Wolanski,M., Donczew,R., Zawilak-Pawlik,A. and Zakrzewska-Czerwinska,J. (2015) *oriC*-encoded instructions for the initiation of bacterial chromosome replication. *Front. Microbiol.*, **5**, 735.
3. Ekundayo,B. and Bleichert,F. (2019) Origins of DNA replication. *PLoS Genet.,***15**, e1008320.
4. Gao,F. and Zhang,C.T. (2008) Ori-Finder: a web-based system for finding *oriC*s in unannotated bacterial genomes. *BMC Bioinf.,***9**, 79.
5. Luo,H., Zhang,C.T. and Gao,F. (2014) Ori-Finder 2, an integrated tool to predict replication origins in the archaeal genomes. *Front. Microbiol.*, **5**, 482.
6. Gao,F. and Zhang,C.T. (2007) DoriC: a database of *oriC* regions in bacterial genomes. *Bioinformatics*, **23**, 1866–1867.
7. 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.
8. Luo,H. and Gao,F. (2019) DoriC 10.0: an updated database of replication origins in prokaryotic genomes including chromosomes and plasmids. *Nucleic Acids Res.,***47**, D74–D77.
9. Pei,L., Schmidt,M. and Wei,W. (2011) Synthetic biology: an emerging research field in china. *Biotechnol. Adv.*, **29**, 804–814.
10. Zhulin,I.B. (2015) Databases for microbiologists. *J. Bacteriol.*, **197**, 2458–2467.
11. Lioy,V.S., Junier,I. and Boccard,F. (2021) Multiscale dynamic structuring of bacterial chromosomes. *Annu. Rev. Microbiol.*, **75**, 541–561.
12. Pelliciari,S., Dong,M.J., Gao,F. and Murray,H. (2021) Evidence for a chromosome origin unwinding system broadly conserved in bacteria. *Nucleic Acids Res.,***49**, 7525–7536.
13. Sankar,T.S., Wastuwidyaningtyas,B.D., Dong,Y., Lewis,S.A. and Wang,J.D. (2016) The nature of mutations induced by replication-transcription collisions. *Nature*, **535**, 178–181.
14. Sobetzko,P., Travers,A. and Muskhelishvili,G. (2012) Gene order and chromosome dynamics coordinate spatiotemporal gene expression during the bacterial growth cycle. *Proc. Natl. Acad. Sc. U.S.A.*, **109**, E42–E50.
15. Chen,W.H., Lu,G., Bork,P., Hu,S. and Lercher,M.J. (2016) Energy efficiency trade-offs drive nucleotide usage in transcribed regions. *Nat. Commun.*, **7**, 11334.
16. Luo,H., Quan,C.L., Peng,C. and Gao,F. (2019) Recent development of Ori-Finder system and DoriC database for microbial replication origins. *Brief. Bioinf.,***20**, 1114–1124.
17. Gao,Y. and Li,H. (2018) Quantifying and comparing bacterial growth dynamics in multiple metagenomic samples. *Nat. Methods*, **15**, 1041–1044.
18. Joseph,T.A., Chlenski,P., Litman,A., Korem,T. and Pe'er,I. (2022) Accurate and robust inference of microbial growth dynamics from metagenomic sequencing reveals personalized growth rates. *Genome Res.,***32**, 558–568.
19. Korem,T., Zeevi,D., Suez,J., Weinberger,A., Avnit-Sagi,T., Pompan-Lotan,M., Matot,E., Jona,G., Harmelin,A., Cohen,N. *et al.* (2015) Microbiome growth dynamics of gut microbiota in health and disease inferred from single metagenomic samples. *Science*, **349**, 1101–1106.
20. Ionescu,D., Bizic-Ionescu,M., De Maio,N., Cypionka,H. and Grossart,H.-P. (2017) Community-like genome in single cells of the sulfur bacterium *Achromatium oxaliferum*. *Nat. Commun.*, **8**, 455.
21. Sayers,E.W., Bolton,E.E., Brister,J.R., Canese,K., Chan,J., Comeau,D.C., Connor,R., Funk,K., Kelly,C., Kim,S. *et al.* (2022) Database resources of the national center for biotechnology information. *Nucleic Acids Res.,***50**, D20–D26.
22. Li,W., O'Neill,K.R., Haft,D.H., DiCuccio,M., Chetvernin,V., Badretdin,A., Coulouris,G., Chitsaz,F., Derbyshire,MyraK., Durkin,A.S. *et al.* (2020) RefSeq: expanding the prokaryotic genome annotation pipeline reach with protein family model curation. *Nucleic Acids Res.,***49**, D1020–D1028.
23. Dong,M.J., Luo,H. and Gao,F. (2022) Ori-Finder 2022: a comprehensive web server for prediction and analysis of bacterial replication origins. *Genomics, Proteomics Bioinf.*, doi: 10.1016/j.gpb.2022.10.002
24. Leonard,A.C. and Méchali,M. (2013) DNA replication origins. *Cold Spring Harbor Perspect. Biol.*, **5**, a010116.
25. Marczynski,G.T. and Shapiro,L. (2002) Control of chromosome replication in *Caulobacter crescentus*. *Annu. Rev. Microbiol.*, **56**, 625–656.
26. Speck,C. and Messer,W. (2001) Mechanism of origin unwinding: sequential binding of DnaA to double- and single-stranded DNA. *EMBO J.,***20**, 1469–1476.
27. Chung,Y.S., Brendler,T., Austin,S. and Guarne,A. (2009) Structural insights into the cooperative binding of SeqA to a tandem GATC repeat. *Nucleic Acids Res.,***37**, 3143–3152.
28. Brassinga,A.K.C., Siam,R., McSween,W., Winkler,H., Wood,D. and Marczynski,G.T. (2002) Conserved response regulator CtrA and IHF binding sites in the alpha-proteobacteria *Caulobacter crescentus* and *Rickettsia prowazekii* chromosomal replication origins. *J. Bacteriol.*, **184**, 5789–5799.
29. Shao,Y., Feldman-Cohen,L.S. and Osuna,R. (2008) Functional characterization of the *Escherichia coli* Fis–DNA binding sequence. *J. Mol. Biol.*, **376**, 771–785.
30. Hales,L.M., Gumport,R.I. and Gardner,J.F. (1994) Determining the DNA sequence elements required for binding integration host factor to two different target sites. *J. Bacteriol.*, **176**, 2999–3006.
31. Zhabinskaya,D., Madden,S. and Benham,C.J. (2014) SIST: stress-induced structural transitions in superhelical DNA. *Bioinformatics*, **31**, 421–422.
32. Bailey,T.L., Johnson,J., Grant,C.E. and Noble,W.S. (2015) The MEME suite. *Nucleic Acids Res.,***43**, W39–W49.
33. Boratyn,G.M., Camacho,C., Cooper,P.S., Coulouris,G., Fong,A., Ma,N., Madden,T.L., Matten,W.T., McGinnis,S.D., Merezhuk,Y. *et al.* (2013) BLAST: a more efficient report with usability improvements. *Nucleic Acids Res.,***41**, W29–W33.