BIOINFORMATICS APPLICATIONS NOTE Vol. 28 no. 11 2012, pages 1551–1552
doi:10.1093/bioinformatics/bts151

Databases and ontologies

Advance Access publication March 30, 2012

DeOri: a database of eukaryotic DNA replication origins

Feng Gao, Hao Luo and Chun-Ting Zhang*

Department of Physics, Tianjin University, Tianjin 300072, China

Associate Editor: Jonathan Wren

ABSTRACT

Summary: DNA replication, a central event for cell proliferation, is the basis of biological inheritance. The identification of replication origins helps to reveal the mechanism of the regulation of DNA replication. However, only few eukaryotic replication origins were characterized not long ago; nevertheless, recent genome-wide approaches have boosted the number of mapped replication origins. To gain a comprehensive understanding of the nature of eukaryotic replication origins, we have constructed a Database of Eukaryotic ORIs (DeOri), which contains all the eukaryotic ones identified by genome-wide analyses currently available. A total of 16 145 eukaryotic replication origins have been collected from 6 eukaryotic organisms in which genome-wide studies have been performed, the replication-origin numbers being 433, 7489, 1543, 148, 348 and 6184 for humans, mice, *Arabidopsis thaliana*, *Kluyveromyces lactis*, *Schizosaccharomyces pombe* and *Drosophila melanogaster*, respectively.

Availability: Database of Eukaryotic ORIs (DeOri) can be accessed from http://tubic.tju.edu.cn/deori/

Contact: ctzhang@tju.edu.cn

Received on December 3, 2011; revised on March 18, 2012; accepted on March 24, 2012

1 INTRODUCTION

The initiation of replication is the central event in the cell cycle. In all three domains of life, DNA replication begins at specialized loci termed replication origins. In bacteria, replication initiates from a single replication origin, whereas eukaryotic organisms exploit many replication origins (Costas *et al*., 2011). In eukaryotes, the origins of DNA replication in the budding yeast *Saccharomyces cerevisiae* and fission yeast *Schizosaccharomyces pombe* have been well-characterized (Hayashi *et al*., 2007; Heichinger *et al*., 2006; Nieduszynski *et al*., 2007; Segurado *et al*., 2003). Recently, a high-resolution genome-wide map of DNA replication origins in *Kluyveromyces lactis* has also been generated (Liachko *et al*., 2010). However, only few origins were characterized in the genomes of higher eukaryotes not long ago. Fortunately, recent development of genome-wide approaches has led to a boost in the number of replication origins that have been mapped in eukaryotic genomes, including those in human (Cadoret *et al*., 2008; Karnani *et al*., 2010), mouse (Cayrou *et al*., 2011; Sequeira-Mendes *et al*., 2009), *Arabidopsis thaliana* (Costas *et al*., 2011) and *Drosophila melanogaster* (Cayrou *et al*., 2011). The availability of increasing origins throughout the genomes has created opportunities for analysis of the origins on a genome scale. Up to date, the DNA replication origin database for budding yeast (Nieduszynski *et al*., 2007) has been built, whereas the database of other eukaryotic origins has not been available. In order to gain a comprehensive understanding of the nature of eukaryotic replication origins, we have constructed a Database of Eukaryotic ORIs (DeOri), which contains all the eukaryotic DNA replication origins identified by genome-wide analyses currently available.

*To whom correspondence should be addressed.

2 DATABASE CONTENT

All information in DeOri is stored and operated using MySQL relational database management system. One entry in the database corresponds to an origin in a certain organism. The corresponding organism, chromosome, cell type (if any), genomic location, length, GC content, origin sequence and some related links have been displayed. In order to facilitate the comparative genomic analysis and visualize the chromosome context of the origins, the URLs that link to UCSC Genome Browser (Fujita *et al*., 2011) or NCBI Map Viewer (Sayers *et al*., 2011) have also been provided. The database access is via a web interface based on PHP script and provides various ways to search for DeOri records, such as DeOri ID, cell type, species and chromosome. In addition, users can also BLAST (Altschul *et al*., 1997) a query sequence against DeOri to find a homologous one. DeOri will be updated timely and a total of 16 145 eukaryotic DNA replication origins were collected from 6 eukaryotic organisms in genome-wide studies, including 433 human origins, 7489 mouse origins, 1543 *A.thaliana* origins, 148 *K.lactis* origins, 348 *S.pombe* origins and 6184 *D.melanogaster* origins in the current release (Table 1), which can be accessed from http://tubic.tju.edu.cn/deori/.

Table 1. Contents of DeOri version 1.0

No.	Organism	ORI no.	References
1	*Arabidopsis thaliana*	1543	Costas *et al*., 2011
2	*Drosophila melanogaster*	6184	Cayrou *et al*., 2011
3	Human_1	283	Cadoret *et al*., 2008
4	Human_2	150	Karnani *et al*., 2010
5	*Kluyveromyces lactis*	148	Liachko *et al*., 2010
6	Mouse_ES_1	2412	Cayrou *et al*., 2011
7	Mouse_ES_2	98	Sequeira-Mendes *et al*., 2009
8	Mouse_MEF	2231	Cayrou *et al*., 2011
9	Mouse_P19	2748	Cayrou *et al*., 2011
10	*Schizosaccharomyces pombe*	348	Hayashi *et al*., 2007; Heichinger *et al*., 2006; Segurado *et al*., 2003

3 DATA RETRIEVAL

The replication origins stored in DeOri include those identified in the genomes of human (Cadoret *et al*., 2008; Karnani *et al*., 2010), mouse (Cayrou *et al*., 2011; Sequeira-Mendes *et al*., 2009), *A.thaliana* (Costas *et al*., 2011), *D.melanogaster* (Cayrou *et al*., 2011), *K.lactis* (Liachko *et al*., 2010) and *S.pombe* (Hayashi *et al*., 2007; Heichinger *et al*., 2006; Segurado *et al*., 2003). In the human genome, 283 replication origins have been systematically mapped using the HeLa S3 suspension cell line (Cadoret *et al.*, 2008) and in an independent study, 150 new origins have been identified in adherent HeLa cells in the ENCODE area (Karnani *et al.*, 2010). In mouse, identification of preferential sites of DNA replication initiation in 0.4% of the mouse genome has resulted in 97 new ORIs (Sequeira-Mendes *et al*., 2009). By performing a genome-wide analysis in both *Drosophila* and mouse cell lines, up to 2748 ORIs on mouse Chromosome 11 (P19 cells) and 6184 ORIs in the *Drosophila* genome have been characterized (Cayrou *et al*., 2011). In *A.thaliana*, by high-throughput sequencing of newly synthesized DNA, ∼1500 putative have been identified at the genome-wide level (Costas *et al*., 2011). In *K.lactis*, a total of 148 ARSs have been identified by a predict-and-verify approach (Liachko *et al*., 2010). In the *S.pombe* genome, 385 ORIs have been predicted initially from AT content calculation (Segurado *et al*., 2003). Subsequent experiments have also confirmed most of the prediction (Hayashi *et al*., 2007; Heichinger *et al*., 2006). In addition, the whole genome sequences of human (hg17), mouse (mm8, mm9), *D.melanogaster* (dm2) were downloaded from http://hgdownload.cse.ucsc.edu/downloads.html, and the whole genome sequences of *A.thaliana*, *S.pombe* and *K.lactis* were downloaded from the NCBI FTP server (ftp://ftp.ncbi.nih.gov/genomes/). Based on the position information provided in the literatures or by personal communication and the corresponding genome sequences, we have obtained the sequences and other information of the origins. It should be noted that the positions of the origins in *A.thaliana* and *S.pombe* have been relocated and mapped to the newest releases of genome sequences from GenBank in order to remove the sequencing errors in the older versions, and the positions of the origins in Mouse_ES_2 have also been relocated and mapped to the genome sequences of mouse (mm9) in order to make the UCSC Genome Browser available.

4 CONCLUSION

Consequently, we have constructed a database DeOri, which contains all the eukaryotic DNA replication origins identified by genome-wide analyses currently available. With the availability of the origins newly identified by genome-wide analyses, we will update DeOri constantly to include more entries and integrate more information for each entry. This database will facilitate the comparative genomic analysis of replication origins, and provide some insight into the nature of replication origins on a genome scale. One of the applications is to predict replication origins based on homologous sequence search against the origins of closely related species in DeOri. For example, if query sequences in the rice genome using BLAST have homologous origins of *A.thaliana* in DeOri, it is likely that the query sequences are also served as origins in the rice genome. Another application is to find some principles for specific organisms by the genome-wide characterization of the origins in DeOri, which will be useful to develop new algorithms to predict replication origins. For example, we can use some tools for motif discovery to find the conserved elements within the origins of a specific organism.

ACKNOWLEDGEMENTS

We are indebted to Prof. Méchali and Dr Cayrou for providing the positions of the replication origins in mice (three different lines: P19, MEFs and ES cells) and *Drosophila* (one single line), and Prof. Dutta and Dr Kumar for providing the positions of 150 replication origins in human. We would also like to thank Prof. Masukata and Prof. Gómez for answering our questions about the DNA replication origins identified in fission yeast or mouse by them and Dr Ren Zhang for critical revision of manuscript. Technical supports from Dr Yan Lin are gratefully acknowledged.

Funding: The present work was supported in part by the National Natural Science Foundation of China (Grant Nos. 90408028, 31171238, 30800642 and 10747150).

Conflict of Interest: none declared.

REFERENCES

Altschul,S.F. *et al.* (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.*, **25**, 3389–3402.

Cadoret,J.C. *et al.* (2008) Genome-wide studies highlight indirect links between human replication origins and gene regulation. *Proc. Natl Acad. Sci. USA*, **105**, 15837–15842.

Cayrou,C. *et al.* (2011) Genome-scale analysis of metazoan replication origins reveals their organization in specific but flexible sites defined by conserved features. *Genome Res.*, **21**, 1438–1449.

Costas,C. *et al.* (2011) Genome-wide mapping of Arabidopsis thaliana origins of DNA replication and their associated epigenetic marks. *Nat. Struct. Mol. Biol.*, **18**, 395–400.

Fujita,P.A. *et al.* (2011) The UCSC Genome Browser database: update 2011. *Nucleic Acids Res.*, **39**, D876–D882.

Hayashi,M. *et al.* (2007) Genome-wide localization of pre-RC sites and identification of replication origins in fission yeast. *EMBO J.*, **26**, 1327–1339.

Heichinger,C. *et al.* (2006) Genome-wide characterization of fission yeast DNA replication origins. *EMBO J.*, **25**, 5171–5179.

Karnani,N. *et al.* (2010) Genomic study of replication initiation in human chromosomes reveals the influence of transcription regulation and chromatin structure on origin selection. *Mol. Biol. Cell.*, **21**, 393–404.

Liachko,I. *et al.* (2010) A comprehensive genome-wide map of autonomously replicating sequences in a naive genome. *PLoS Genet.*, **6**, e1000946.

Nieduszynski,C.A. *et al.* (2007) OriDB: a DNA replication origin database. *Nucleic Acids Res.*, **35**, D40–D46.

Sayers,E.W. *et al.* (2011) Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res.*, **39**, D38–D51.

Segurado,M. *et al.* (2003) Genome-wide distribution of DNA replication origins at A+T-rich islands in Schizosaccharomyces pombe. *EMBO Rep.*, **4**, 1048–1053.

Sequeira-Mendes,J. *et al.* (2009) Transcription initiation activity sets replication origin efficiency in mammalian cells. *PLoS Genet.*, **5**, e1000446.