oriTfinder: a web-based tool for the identification of origin of transfers in DNA sequences of bacterial mobile genetic elements

Xiaobin Li¹, Yingzhou Xie¹, Meng Liu¹, Cui Tai¹, Jingyong Sun², Zixin Deng¹ and Hong-Yu Ou¹,3,*

¹State Key Laboratory of Microbial Metabolism, Joint International Laboratory on Metabolic & Developmental Sciences, School of Life Sciences & Biotechnology, Shanghai Jiao Tong University, Shanghai, China, ²Department of Clinical Microbiology, Ruijin Hospital, Shanghai Jiao Tong University School of Medicine, Shanghai 200025, China and ³Shanghai Key Laboratory of Veterinary Biotechnology, Shanghai Jiao Tong University, Shanghai 200240, China

Received February 15, 2018; Revised April 12, 2018; Editorial Decision April 21, 2018; Accepted April 24, 2018

ABSTRACT

oriTfinder is a web server that facilitates the rapid identification of the origin of transfer site (oriT) of a conjugative plasmid or chromosome-borne integrative and conjugative element. The utilized backend database oriTDB was built upon more than one thousand known oriT regions of bacterial mobile genetic elements (MGEs) as well as the known MGE-encoding relaxases and type IV coupling proteins (T4CP). With a combination of similarity searches for the oriTDB-archived oriT nucleotide sequences and the co-localization of the flanking relaxase homologous genes, the oriTfinder can predict the oriT region with high accuracy in the DNA sequence of a bacterial plasmid or chromosome in minutes. The server also detects the other transfer-related modules, including the potential relaxase gene, T4CP gene and the type IV secretion system gene cluster, and the putative genes coding for virulence factors and acquired antibiotic resistance determinants. oriTfinder may contribute to meeting the increasing demands of re-annotations for bacterial conjugative, mobilizable or non-transferable elements and aid in the rapid risk assessment of disease-relevant trait dissemination in pathogenic bacteria of interest. oriTfinder is freely available to all users without any login requirement at http://bioinfo-mml.sjtu.edu.cn/oriTfinder.

INTRODUCTION

Bacterial mobile genetic elements (MGEs), such as conjugative plasmids and integrative and conjugative elements (ICEs), have been highlighted as important vehicles for the dissemination of pathogenesis and antimicrobial-resistance determinants (1). The conjugative transfer regions of the self-transmissible MGEs typically consist of four modules: an origin of transfer (oriT) region, relaxase gene, type IV coupling protein (T4CP) gene and gene cluster for the bacterial type IV secretion system (T4SS) apparatus (2). In the process of conjugation, the single-stranded DNA (ssDNA) conjugation process is initially recognized, bound and cleaved by relaxase at the oriT site (3). After rolling-circle replication, the ssDNA is recruited by T4CP and subsequently transferred from the donor cell into the recipient cell via T4SS (4). In addition, large numbers of non-conjugative MGEs, including mobilizable plasmids and integrative and mobilizable elements (IMEs), typically carry a limited number of mob genes for their own DNA processing in conjugation (5), which are transferable but not self-transmissible. Interestingly, non-conjugative MGEs carrying functional oriT sequences can be mobilized by conjugative elements (6). For example, the Vibrio cholerae genomic islands carrying the oriTs were mobilized by the SXT/R391 ICE (7). The SXT element also mobilized the plasmid RSF1010 in trans, which encoded resistance to sulphonamide and streptomycin (8). The oriT region, which is usually tens to hundreds of base pairs in length, contains a conserved nick region (flanking the nic site) and variable numbers of inverted repeats (IRs) (9). The nic site is recognized and cleaved by a relaxase, while the IRs are involved in the localization to a precise nic site as well as the termination of ssDNA transfer (10). Thus, the identification of the oriT region in the MGE sequence is important to investigate the self-transfer or mobilizing transfer capability of MGEs.

Several bioinformatic sources for predicting the T4SS modules of MGEs are available so far, such as the SecReT4 database (11), the AtlasT4SS database (12), the EffectiveDB database (13), the web-based tool T346Hunter (14) and the online tool VRprofile (15). However, all these bioinformat-
ics sources are deficient in the data involved in the initiation
of ssDNA transfer, including the oriT regions, relaxase and
T4CP. To preferably predict the transferability of putative
MGEs and to investigate the transmission of antibiotic re-
sistance or virulence factors (VF) carried by the bacterial
MGEs, it is necessary to obtain the whole picture of the con-
jugational transfer components, especially the oriT region, which
has a widespread distribution in both conjugative and mo-
mobilizable MGEs.

In this study, we report a web tool, named ‘oriTfinder’,
as a public resource for in silico detection of oriT in bac-
terial MGE sequences, especially in antibiotic resistance
plasmids. It can also recognize the putative relaxase genes,
T4CP genes, T4SS gene clusters, VF genes and acquired an-
tibiotic resistance genes within the genetic context of oriT.

We first developed a back-end database oriTDB using our
collections of known oriT loci, relaxases and T4CPs of bac-
terial MGEs. The oriTfinder then performs rapid homology
searches of a query genome sequence against oriTDB based
on both the oriT nucleotide sequence similarity and the re-
laxase protein similarity. It outputs a simple list and gener-
ates a graphic overview of not only the predicted transfer-
related functional site (or genes) but also the extended
putative virulence or acquired antibiotic resistance genes.

The oriTfinder might facilitate the rapid detection of vari-
ous conjugative regions in the dynamic MGEs of bacterial
pathogens.

MATERIALS AND METHODS

oriTDB collecting the sequences of known oriT regions and
relaxase and T4CP genes

The database oriTDB was developed to collect 1074 oriT re-
gions, including 996 in plasmids, 74 in ICEs and 4 in IMEs
(Supplementary Table S1). The oriT region was tagged as
‘experimentally validated’ in oriTDB only if the transfer-
associated function was clearly reported in a peer-reviewed
scientific publication. After manual curation of the PubMed
search results with the keyword ‘oriT’, 324 published pa-
pers were collected and added into oriTDB. The archived
oriT loci are linked to the corresponding literature with ex-
perimental data. In oriTDB, 50 sequences of experimentally
validated oriT regions were collected (Supplementary Table
S1). In addition, oriTDB contains 982 relaxase genes that
were located closely to the known oriT regions (Supple-
mentary Figure S1A), typically with a distance of 20 bp to 27 kb
(Supplementary Table S3). It also records 464 T4CP genes
encoded within the MGEs (Supplementary Figure S1B).

oriTfinder predicting oriTs in DNA sequences of bacterial
MGEs

By analyzing the 1074 oriT sequences from the conjugation-
related regions of the natural plasmids and chromosome-
borne ICEs, the oriTfinder was developed to catch two typi-
fical features of the known oriTDB-archived oriT regions:
(i) containing the conserved DNA sequence flanking the
nic sites (Supplementary Figure S2); and (ii) flanked by re-
laxase genes (Supplementary Table S3). The server com-
bines the similarity searches of the oriTDB-archived oriT
nucleotide sequence module and the co-localization mod-
ule of the flanking relaxase homologous gene (Figure 1), al-
lowing the enhanced prediction performance for the oriT
regions. Briefly, the oriTfinder starts by finding the protein-
coding regions in the query DNA sequence and detects the
relaxase homologs using the HMMer searches with nine
HMM-profiles (Supplementary Table S2). These obtained
homologs were subsequently filtered with the BLASTp
searches (16) against the oriTDB-collected relaxases with
the cut-off identities of 30%. Second, each of the oriT se-
quences recorded by oriTDB is searched against the 30-
kbp upstream and 30-kb downstream regions of the relax-
ase gene when a relaxase homolog was found. It employs
BLASTn (16) using an H-value cut-off ≤ 0.49 for sig-
nificant similarities. The BLASTn-based H-value (0 ≤ H-
value ≤ 1.0) reflects the degree of similarity in terms of
the length of the matching region and the degree of iden-
tity at a nucleotide level between the matching region in
the users’ sequence and the oriT examined (see the Sup-
lementary Methods). Third, each oriTDB-archived oriT
sequence is searched with BLASTn against the other re-
gions of the whole sequence with a cut-off H-value of 0.81.

Then, the putative directed repeats (IRs) within the oriT
obtained above are also detected by Vmatch (available at
http://www.vmatch.de/). The conserved nick region (Sup-
plementary Figure S2) is identified by MEME-MAST (16).
At last, the oriTfinder outputs and visualizes the identified
oriT region containing the information of the sequence co-
ordinates, region length, IRs, nick region and relaxase gene.

In addition, the T4SS gene cluster is predicted by the co-
localization of the homologs of at least five core compo-
ents, similar to the VRprofile (15). The T4CP homolog
is also detected by using HMMer with four HMM-profiles
(Supplementary Table S2). The putative VF and acquired
antibiotic resistance determinants (AR) that are frequently
encoded by the accessory regions of MGEs are also pre-
dected based on BLASTp searches for homologs with a cut-
off Ha-value of 0.64 (15).

To evaluate the oriTfinder algorithm, 43 transferable
plasmids with experimental supports were used as a bench-
mark dataset (Supplementary Table S4). Notably, these
transferable plasmids were not recorded by oriTDB due to
the absence of the characterized oriT regions. Meanwhile,
50 putative non-transferable plasmids (17), which do not
code for a known relaxase, T4CP nor T4SS, were also con-
tained by the benchmark dataset. Then, three frequently
used metrics (18), sensitivity (Sn), specificity (Sp) and posi-
tive predictive value (PPV), were employed (see the Sup-
plementary Methods) to assess the performance of oriTfinder.

Implementation of the oriTfinder server

The oriTfinder web server is applicable to a wide range of
bacterial plasmids and other MGEs. It allows users to up-
load a nucleotide sequence and its annotation (in GenBank
format) as a query. The server consists of two basic com-
ponents: the computational pipeline and the web interface.
The computational component is written in Perl/Bioperl
and uses NCBI BLAST (16) and HMMer3 (19) to predict
the oriT regions and other modules of MGEs involved in
Figure 1. The prediction strategy used by oriTfinder to identify the putative oriT region of a conjugative plasmid or a chromosome-borne ICE. It combines the similarity searches of the oriTDB-archived oriT nucleotide sequence with the co-localization of the flanking relaxase gene.
the conjugal transfer. Tools including MEME-MAST (20), Vmatch (http://vmatch.de/), EMBOSS (21) and Prodigal (22) are also integrated into oriTfinder to allow both enhancements of the prediction performance and extended downstream analyses. The oriTfinder server was developed using Perl/Bioperl and PHP on a Linux platform with an Apache web server. The web interface typically consists of an input page, a status page and a result page, which are generated with HTML, CSS and JavaScript. The CGView circular genome visualization tool (23) was integrated into the result page to display the distribution of the predicted oriT and other transfer modules in the MGE sequence. We also developed a plug-in into the web interface for the visualization of the features of the oriT regions, such as the conserved nick region and IRs. Google Chrome is the recommended Web browser to run oriTfinder. It runs on a high-performance cluster, which contains a computing node equipped with four eight-core processors and 512-gigabyte memory and a 20-terabyte storage node. In general, a job can be completed within 30 s for a 200-kb plasmid genome sequence or 3 min for a 5-Mb bacterial genome sequence.

RESULTS AND DISCUSSION

Validating the performance of oriTfinder prediction for oriT regions

The oriTfinder provides rapid computational identification for conjugative regions in the MGE sequence, including the oriT region, relaxase gene, T4CP gene and T4SS gene cluster. The benchmark dataset contains 43 transferable plasmids and 50 non-transferable plasmids (Supplementary Table S4). The oriTfinder performed well in identifying the oriT regions in these plasmids with the Sn of 88.4%, the Sp of 100% and the PPV of 100.0%. Notably, the oriTfinder had a higher performance than the BLASTn-based oriT sequence search module (Sn = 83.7%, Sp = 100% and PPV = 100% with an H-value ≥ 0.81). This result suggested that the oriT prediction accuracy could be enhanced by the combination of oriT sequence searches with the co-localization of the oriT and flanking relaxase genes.

The well-documented IncP-alpha plasmid RK2 (also called RP4) is shown as an example (Figure 2). It is a conjugative drug-resistance plasmid broadly distributed in gram-negative bacteria and a canonical model for plasmid transfer study (24). There have been 436 vectors derived from the plasmid RK2/RP4 with the same oriT regions (Collection in oriTDB; URL: http://bioinfo-mml.sjtu.edu.cn/oriTDB/browse_vector.php). With the input of the GenBank file of the plasmid RK2 (GenBank accession no. BN000925), oriTfinder identified the oriT regions and three other modules participating in the process of plasmid conjugal transfer. The oriTfinder result page contained five tabs: (i) ‘oriT&Relaxase’, containing the information of oriT co-ordinate, IRs, conserved nick region and relaxase gene, and the visual representation of the nick region and IRs (Figure 2A); (ii) ‘T4SS’, including the graphic presentation and tabulated view of the T4SS gene cluster; (iii) ‘T4CP’, displaying detailed information of the T4CP gene; (iv) ‘AR&VF’, tabulating the information of the AR and/or VF genes; and (v) ‘Summary’, tabulating the above information and displaying a CGview-generated circular genomic map (Figure 2B).

Case study: Prediction of potential transmission of antibiotic-resistant plasmids and virulence plasmids from Klebsiella pneumoniae

Klebsiella pneumoniae is an important drug-resistant bacterial pathogen and causative of nosocomial infections throughout the world. It is currently regarded as a major worldwide source and shuttle for antibiotic resistance (25). Here, we collected 311 plasmid sequences from 107 completely sequenced K. pneumoniae genomes available in GenBank up to 1 December 2017 (Supplementary Table S5). The incompatibility groups of these plasmids were determined by PlasmidFinder (26). With oriTfinder, the transfer regions of these plasmids were successfully identified (Supplementary Figure S3), including the oriTs, relaxases, T4SSs gene clusters and T4CPs. Among the 311 K. pneumoniae plasmids, 26.4% (82/311) were found to possess a whole set of oriTs, relaxases, T4CPs and T4SSs, indicating their high potential for self-transferability (17). Notably, 63 of these 82 conjugative plasmids were found to carry putative acquired AR genes, indicating that these plasmids could potentially disseminate antibiotic resistance genes. For example, the 111-kb carbapenemase-encoded plasmid pKPHS2 from the clinical isolate K. pneumoniae HS11286 (27) was found to contain an oriT region homologous to that of the oriTDB-archived plasmid R100 with an H-value of 0.51 (Supplementary Figure S4A). In addition, 12.2% (38/311) of all K. pneumoniae plasmids under study were found to have both oriTs and relaxases, but were lacking T4CPs and/or T4SSs, indicating that they are potentially mobilizable plasmids (17). Thirty-one of these 38 mobilizable plasmids were found to carry acquired AR genes, which might be mobilized by conjugal MGEs (6). At last, 61.4% (191/311) of all K. pneumoniae plasmids were found to contain no oriTDB-archived conjugal modules, and 20 out of the 191 predicted non-transferable plasmids were found to carry AR genes.

CONCLUSION

We have developed a user-friendly web server, oriTfinder, to perform quick detection of oriTs and three other transfer-associated modules (relaxase, T4CP and T4SS) in bacterial MGE sequences, especially in plasmids carrying antimicrobial resistance genes. To our knowledge, oriTfinder is the only tool providing a specific service for predicting oriT regions in MGE sequences so far. In the near future, we expect to collect more oriT information, update oriTDB regularly and identify a broader set of oriT regions to enhance the prediction performance of the oriTfinder. We propose that a tool such as oriTfinder will support the rapidly escalating demands of comparative genomics studies aimed at defining self-transferable or mobilizable plasmids and other MGEs with potential transfer across clinically relevant bacterial pathogens.
Figure 2. An overview of oriTfinder outputs using the oriT region of the plasmid RK2 as an example. (A) List of the features of the oriT region: location, sequence, subject in oriTDB and predicted relaxase. The detected nic site and IRs are displayed within the oriT sequence. Hyperlinks to oriTDB and NCBI are provided as appropriate. (B) A scaled representation of the circular RK2 plasmid generated by the oriTfinder-integrated CGView (23) utility showing the locations and sizes of oriT (saddle brown), the relaxase gene (green), T4CP (dodger blue), genes coding for components of both T4SSs (blue) and AR (pink) within this replicon.
**SUPPLEMENTARY DATA**

Supplementary Data are available at NAR Online.

**ACKNOWLEDGEMENTS**

The genome sequence analysis was supported by the Center for High Performance Computing (HPC), Shanghai Jiao Tong University.

**FUNDING**

National Key R&D Program of China [2017YFC1600105 to H.Y.O.]; National Natural Science Foundation of China [31670074 to H.Y.O., 21661140002 to Z.D.]; Medicine and Engineering Interdisciplinary Research Fund of Shanghai Jiao Tong University [YG2015MS59 to J.S.]. Funding for open access charge: National Key R&D Program of China [2017YFC1600105 to H.Y.O.]; National Natural Science Foundation of China [31670074 to J.S.].

**REFERENCES**

1. Frost, L.S., Leplae, R., Summers, A.O. and Toussaint, A. (2005) Mobile genetic elements: the agents of open source evolution. *Nat. Rev. Microbiol.*, **3**, 722–732.

2. Bürkus, V. (2017) Mechanisms of stabilization of integrative and conjugative elements. *Curr. Opin. Microbiol.*, **38**, 44–50.

3. Llosa, M., Gomis-Ruth, F.X., Coll, M. and de la Cruz Fd, F. (2002) Bacterial conjugation: a two-step mechanism for DNA transport. *Mol. Microbiol.*, **45**, 1–8.

4. Grohmann, E., Christie, P.J., Waksman, G. and Backert, S. (2018) Type IV secretion in Gram-negative and Gram-positive bacteria. *Mol. Microbiol.*, **107**, 455–471.

5. Lanka, E. and Wilkins, B.M. (1995) DNA processing reactions in bacterial conjugation. *Annu. Rev. Biochem.*, **64**, 141–169.

6. Ramsay, J.P. and Firth, N. (2017) Diverse mobilization strategies facilitate transfer of non-conjugative mobile genetic elements. *Curr. Opin. Microbiol.*, **38**, 1–9.

7. Daccord, A., Ceccarelli, D. and Burrus, V. (2010) Integrating conjugative elements of the SXT/R391 family trigger the excision and drive the mobilization of a new class of *Vibrio* genomic islands. *Mol. Microbiol.*, **78**, 576–585.

8. Hochhut, B., Marrero, J. and Waldor, M.K. (2000) Mobilization of plasmids and chromosomal DNA mediated by the SXT element, a consin found in *Vibrio cholerae* O139. *J. Bacteriol.*, **182**, 2043–2047.

9. de la Cruz, F., Frost, L.S., Meyer, R.J. and Zechner, E.L. (2010) Conjugative DNA metabolism in Gram-negative bacteria. *FEMS Microbiol. Rev.*, **34**, 18–40.

10. Furuya, N. and Komano, T. (2000) Initiation and termination of DNA transfer during conjugation of IncI plasmid R64: roles of two sets of inverted repeat sequences within oriT in termination of R64 transfer. *J. Bacteriol.*, **182**, 3191–3196.

11. Bi, D., Liu, L., Tai, C., Deng, Z., Rajakumar, K. and Ou, H.Y. (2013) SecReT4: a web-based bacterial type IV secretion system resource. *Nucleic Acids Res.*, **41**, D660–D665.

12. Souza, R.C., del Rosario Quispe Saji, G., Costa, M.O., Netto, D.S., Lima, N.C., Klein, C.C., Vasconcelos, A.T. and Nicolas, M.F. (2012) AtlasT4SS: a curated database for type IV secretion systems. *BMC Microbiol.*, **12**, 172.

13. Eichinger, V., Nussbauer, T., Platzner, A., Jehl, M.A., Arnold, R. and Rattie, T. (2016) EffectiveDB—updates and novel features for a better annotation of bacterial secreted proteins and Type III, IV, VI secretion systems. *Nucleic Acids Res.*, **44**, D669–D674.

14. Martinez-Garcia, P.M., Ramos, C. and Rodriguez-Palenzuela, P. (2015) T346Hunter: a novel web-based tool for the prediction of type III, type IV and type VI secretion systems in bacterial genomes. *PLoS One*, **10**, e0119317.

15. Li, J., Tai, C., Deng, Z., Zhong, W., He, Y. and Ou, H.Y. (2017) VRprofile: gene-cluster-detection-based profiling of virulence and antibiotic resistance traits encoded within genome sequences of pathogenic bacteria. *Brief. Bioinform.*, doi:10.1093/bib/bbw141.

16. Camacho, C., Coulouris, G., Avagyan, V., Ma, N., Papadopoulos, J., Bealer, K. and Madden, T.L. (2009) BLAST+: architecture and applications. *BMC Bioinformatics*, **10**, 421.

17. Smithies, C., Garcillan-Barcia, M.P., Francia, M.V., Rocha, E.P. and de la Cruz, F. (2010) Mobility of plasmids. *Microbiol. Mol. Biol. Rev.*, **74**, 434–452.

18. Fawcett, T. (2006) An introduction to ROC analysis. *Pattern Recognit. Lett.*, **27**, 861–874.

19. Finn, R.D., Clements, J., Arndt, W., Miller, B.L., Wheeler, T.J., Schreiber, F., Bateman, A. and Eddy, S.R. (2015) HHMmer web server: 2015 update. *Nucleic Acids Res.*, **43**, W30–W38.

20. Bailey, T.L., Johnson, J., Grant, C.E. and Noble, W.S. (2015) The MEME Suite. *Nucleic Acids Res.*, **43**, W39–W49.

21. Rice, P., Longden, I. and Bleasby, A. (2000) EMBOSS: the European molecular biology open software suite. *Trends Genet.*, **16**, 276–277.

22. Hyatt, D., Chen, G.L., Locascio, P.F., Land, M.L., Larimer, F.W. and Hauser, L.J. (2010) Prodigal: prokaryotic gene recognition and translation initiation site identification. *BMC Bioinformatics*, **11**, 119.

23. Grant, I.R. and Stothard, P. (2008) The CGView Server: a comparative genomics tool for circular genomes. *Nucleic Acids Res.*, **36**, W181–W184.

24. Guiney, D.G. and Yakobson, E. (1983) Location and nucleotide sequence of the transfer origin of the broad host range plasmid RK2. *Proc. Natl. Acad. Sci. U.S.A.*, **80**, 5359–5364.

25. Navon-Venezia, S., Kondratyeva, K. and Carattoli, A. (2017) Klebsiella pneumoniae: a major worldwide source and shuttle for antibiotic resistance. *FEMS Microbiol. Rev.*, **41**, 252–275.

26. Carattoli, A., Zankari, E., Garcia-Fernandez, A., Voldby Larsen, M., Lund, O., Villa, L., Moller Aurestrup, P. and Hasman, H. (2014) *Insilico* detection and typing of plasmids using PlasmidFinder and plasmid multilocus sequence typing. *Antimicrob. Agents Chemother.*, **58**, 3895–3903.

27. Bi, D., Jiang, X., Sheng, Z.K., Ngamenterebo, D., Tai, C., Wang, M., Deng, Z., Rajakumar, K. and Ou, H.Y. (2015) Mapping the resistance-associated mobilome of a carbapenem-resistant *Klebsiella pneumoniae* strain reveals insights into factors shaping these regions and facilitates generation of a ‘resistance-disarmed’ model organism. *J. Antimicrob. Chemother.*, **70**, 2770–2774.