Nucleic Acids Research, 2025, **53**, D163–D168
https://doi.org/10.1093/nar/gkae869
Advance access publication date: 7 October 2024
Database issue

oriTDB: a database of the origin-of-transfer regions of bacterial mobile genetic elements

Guitian Liu[1,2,†], Xiaobin Li[3,†], Jiahao Guan[2,†], Cui Tai[2], Yuqing Weng[4], Xiaohua Chen[1,*] and Hong- Yu Ou[1,2,*]

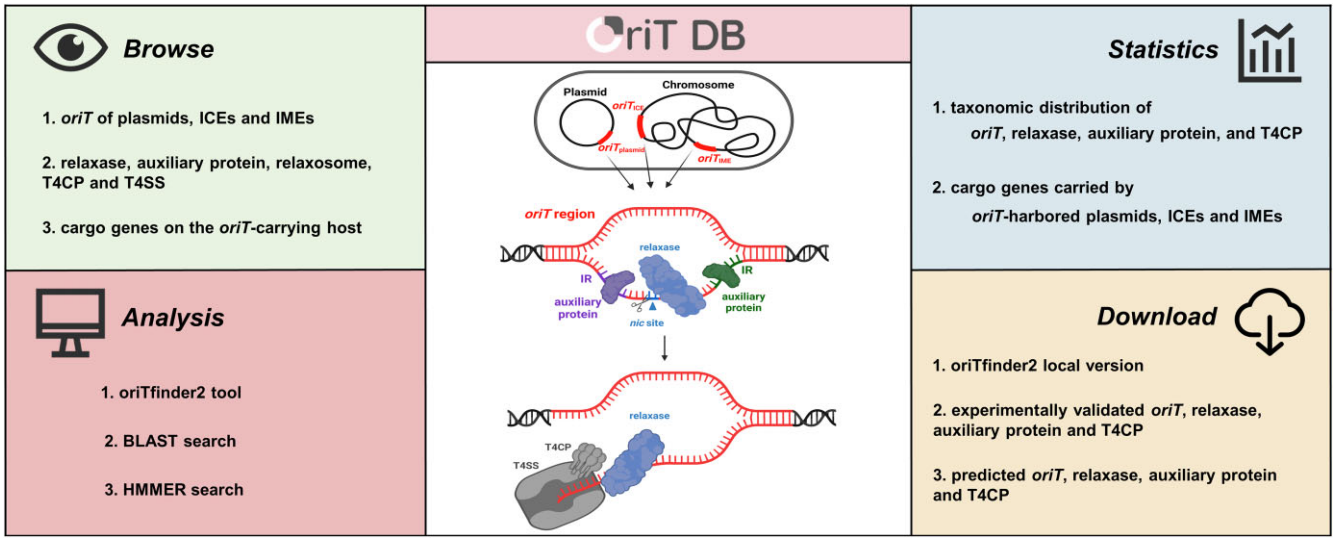
[1]Department of Infectious Diseases, Shanghai Sixth People's Hospital Affiliated to Shanghai Jiao Tong University School of Medicine, Shanghai 200233, China
[2]State Key Laboratory of Microbial Metabolism, Joint International Laboratory on Metabolic & Developmental Sciences, School of Life Sciences & Biotechnology, Shanghai Jiao Tong University, Shanghai 200240, China
[3]Guangdong Provincial Key Laboratory of Tumor Interventional Diagnosis and Treatment, Zhuhai People's Hospital (Zhuhai Clinical Medical College of Jinan University), Zhuhai 519000, China
[4]Department of Pulmonary and Critical Care Medicine, Zhuhai People's Hospital (Zhuhai Clinical Medical College of Jinan University), Zhuhai 519000, China

[*]To whom correspondence should be addressed: Tel: +86 2134204710; Fax: +86 2134204710; Email: hyou@sjtu.edu.cn
Correspondence may also be addressed to Xiaohua Chen. Tel: +86 2124058673; Fax: +86 2124058673; Email: chenxiaohua2000@163.com
[†]The first three authors should be regarded as Joint First Authors.

Abstract

Conjugation and mobilization are two important pathways of horizontal transfer of bacterial mobile genetic elements (MGEs). The origin-of-transfer (*oriT*) region is crucial for this process, serving as a recognition site for relaxase and containing the DNA nicking site (*nic* site), which initiates the conjugation or mobilization. Here, we present a database of the origin-of-transfer regions of bacterial MGEs, oriTDB (https://bioinfo-mml.sjtu.edu.cn/oriTDB2/). Incorporating data from text mining and genome analysis, oriTDB comprises 122 experimentally validated and 22 927 predicted *oriT*s within bacterial plasmids, Integrative and Conjugative Elements, and Integrative and Mobilizable Elements. Additionally, oriTDB includes details about associated relaxases, auxiliary proteins, type IV coupling proteins, and a gene cluster encoding the type IV secretion system. The database also provides predicted secondary structures of *oriT* sequences, dissects *oriT* regions into pairs of inverted repeats, *nic* sites, and their flanking conserved sequences, and offers an interactive visual representation. Furthermore, oriTDB includes an enhanced *oriT* prediction pipeline, oriTfinder2, which integrates a functional annotation module for cargo genes in bacterial MGEs. This resource is intended to support research on bacterial conjugative or mobilizable elements and promote an understanding of their cargo gene functions.

Graphical abstract



Received: August 15, 2024. **Revised:** September 12, 2024. **Editorial Decision:** September 19, 2024. **Accepted:** September 23, 2024

Introduction

Bacterial mobile genetic elements (MGEs) are crucial in disseminating virulence factor genes and antimicrobial resistance genes(1). These elements, including conjugative or mobilizable plasmids, integrative and conjugative elements (ICEs), and Integrative and Mobilizable Elements (IMEs), typically consist of four modules in their conjugative regions: the origin-of-transfer (*oriT*) region, relaxase gene, type IV coupling protein (T4CP) gene, and a gene cluster encoding type IV secretion system (T4SS). The *oriT* region, usually ranging in length from tens to hundreds of base pairs, contains a *nic* site and pairs of inverted repeats (IRs) (2). It is vital for the transfer process as it is recognized by the relaxase and undergoes nicking at a conserved site (*nic*), resulting in the formation of single-stranded DNA (ssDNA) (3). This ssDNA, along with the relaxase and auxiliary proteins, forms a relaxosome and is transferred with the assistance of T4SS and its T4CPs. Therefore, it is essential to curate data on the *oriT* regions and their corresponding relaxase genes, auxiliary protein genes, T4CP genes, and the T4SS gene cluster, which help investigate the self-transfer or mobilization of bacterial MGEs.

Many studies have been conducted to investigate the potential mobility of plasmids with a focus on relaxases (4,5), by using experimental methods (6,7) and *in silico* approaches (8). Computational tools have also been developed to predict the *oriT* regions, which helps examine the transfer of MGEs. The prediction tool oriTfinder, developed by our group, employs similarity searches for *oriT* sequences and the co-localization of flanking relaxase homologous genes (9). Furthermore, a novel approach has utilized intergenic positioning, relaxase distance and MOB-type association to detect *oriT*s in plasmids of specific bacterial species (10), such as *Escherichia coli*, *Klebsiella pneumoniae* and *Acinetobacter baumannii*. Recently, it was discovered that plasmids lacking a relaxase retain the capability to transfer (11–13), indicating the necessity to predict plasmid mobility independent of relaxase information. Certain structural conservation has been observed in *oriT* sequences (14), allowing non-conjugative plasmids to become mobilizable by carrying *oriT*-mimics (15,16), which suggests a need for further *in silico* dissection of the *oriT* regions. The IRs and some specific regions in the *oriT* sequence serve as the binding site for the relaxase and auxiliary proteins, such as the IR2 within $oriT_{\mathrm{ICE}st3}$ for the relaxase RelSt3 (17) and the *ossA* and *ossB* region within $oriT_{\mathrm{pWBG749}}$ for the auxiliary protein SmpO (16). However, existing bioinformatics resources lack comprehensive data on the *oriT* regions, including their essential *nic* sites, IRs, and the corresponding relaxases, auxiliary proteins, relaxosomes and T4CPs. Moreover, analytical tools connecting the mobility of different MGEs with the diverse cargo genes they carry are insufficient, hindering the understanding of the relationship between MGEs and their cargo genes.

Here, we present the release of oriTDB, a comprehensive database containing an expanded collection of curated *oriT* regions. Within this database are details about the IRs and *nic* sites located within the *oriT* regions, along with information about their corresponding relaxases, auxiliary proteins, and T4CPs. Additionally, a tool capable of identifying the *oriT* region in a MGE lacking a relaxase through sequence comparison against oriTDB. Moreover, oriTfinder2 integrates a novel functional annotation module for cargo genes in *oriT*-carrying MGEs. We anticipate that oriTDB will streamline the precise localization of *oriT* regions and the functional assessment of diverse cargo genes within *oriT*-carrying MGEs.

Materials and methods

Data updated by text mining and manual curations

Through a meticulous manual curation of PubMed literature using '*oriT*' and 'origin of transfer' keywords, we have amassed a collection of 741 papers (on 25 March 2024). After text mining and manual curations, oriTDB collected 122 experimentally validated *oriT*s from various MGEs of 45 species, including those in plasmids ($n = 91$), ICEs ($n = 18$) and IMEs ($n = 13$). Additionally, oriTDB now incorporates information on *oriT*-related and experimentally validated relaxases ($n = 50$), auxiliary proteins ($n = 44$) and T4CPs ($n = 16$) (Supplementary Table S1). We have also compiled 21 entries for relaxosomes, providing details on the *oriT* regions, associated relaxases and auxiliary proteins.

Prediction of *oriT* region and cargo genes in MGE sequences using oriTfinder2

The updated tool, oriTfinder2, aims to predict potential *oriT*s, relaxases and T4CPs in plasmids, ICEs, and IMEs (Supplementary Figure S1). It has been enhanced to improve the prediction accuracy for relaxases by utilizing hidden Markov model (HMM) profiles for the MOB protein family (Supplementary Table S2). Additionally, oriTfinder2 can identify *oriT* regions in a relaxase-lacking mobilizable plasmid by integrating BLASTn searches against the *oriT* sequences archived in the oriTDB (Supplementary Method and Supplementary Figure S1). The plasmid carrying the *oriT* but lacking a relaxase gene was found to be mobilized with the help of a conjugative plasmid(13). By utilizing oriTfinder2, we have identified 22 927 putative *oriT* regions (22 390 in plasmids, 482 in ICEs, and 55 in IMEs) across 972 species from MGE sequences sourced from the NCBI RefSeq (18) plasmid dataset ($n = 86\ 009$) (on 21 March 2024) and our ICE database ICEberg 3.0 (1323 ICEs and 324 IMEs) (Supplementary Table S1). The number and ratio of *oriT*-carrying plasmid, ICEs and IMEs predicted by oriTfinder 2.0 are 22 199 (25.8%), 457 (34.5%) and 68 (20.9%), respectively. These pre-computing results were stored by oriTDB.

Furthermore, oriTfinder2 also incorporates a functional annotation pipeline for cargo genes in *oriT*-carrying MGEs, encompassing acquired antibiotic resistance genes sourced from Resfinder (19), virulence factors obtained from VFDB (20), metal resistance determinants from BacMet2 (21), anti-CRISPR proteins from Anti-CRISPRdb v2.2 (22), microbial degradation proteins from mibPOPdb (23) and symbiotic proteins compiled by ICEberg 3.0 (24). More details are available in the Supplementary Methods and Supplementary Table S3. This advancement may contribute to a comprehensive understanding of various biological functionalities associated with bacterial MGEs.

Implementation of the web-based database oriTDB

The oriTDB utilizes a PostgreSQL relational database, a PHP data pipeline, and HTML web interfaces for data management. It also incorporates the Bootstrap library (https://getbootstrap.com/) and JavaScript-powered data visualization libraries, such as ECharts (25) and SVGene (https://github.com/kblin/svgene), to improve user interaction.

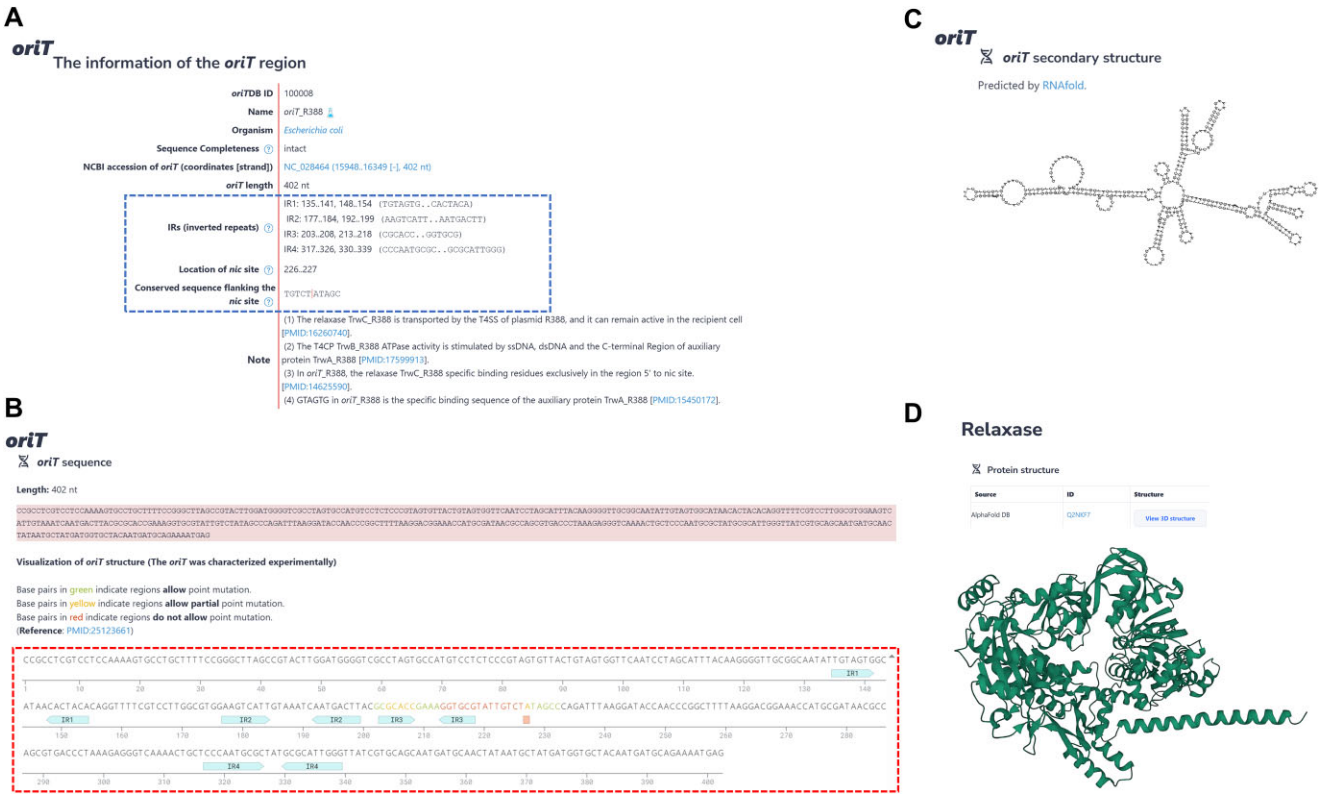


Figure 1. Overview of the dissection and visualization of the *oriT* region in oriTDB. (**A**) The detailed information of the *oriT*_R388 region in the plasmid R388 of *Escherichia coli*. The *oriT* region was dissected into pairs of IRs, *nic* sites, and their flanking conserved sequences (blue dashed box). (**B**) Visual representation of the *oriT*_R388 region. The sequence of *oriT* is presented in the interactive visualization form (red dashed box), with the IR pairs in the *oriT* region as aqua arrows, and the *nic* sites as orange rectangles. (**C**) The RNAfold-predicted secondary structure of the single strand *oriT*_R388 region. (**D**) The AlphaFold2-predicted structure of the relaxase protein TrwC_R388 related to *oriT*_R388.

Sequence feature annotations of *oriT*, including IRs and conserved nick regions, are visualized using seqviz (https://github.com/Lattice-Automation/seqviz). Additionally, DNA secondary structures of *oriT* regions were predicted by RNAfold with the DNA parameters (26). The circular genome visualization tool CGView (27) is integrated into the oriTfinder2 result page to depict the distribution of the putative *oriT* and other transfer modules in the MGE sequence. Moreover, the Pfam (28) domains of the proteins related to *oriT* are predicted using InterProScan(29). Experimentally determined and predicted 3D structures for each protein are obtained from the RCSB Protein Data Bank (PDB) (https://www.rcsb.org/)(30) and the AlphaFold Protein Structure Database (31), respectively, and are interactively visualized using PDBe Mol* (32). Finally, DNA and protein homologs are identified using NCBI BLASTp (33) and HMMER3 (34) in the BLASTp and HMMER search tools.

Results and discussion

Compilation and visualization of *oriT* regions

Through a rigorous process of text mining and manual curation, we have introduced 23 527 *oriT* entries, 13 116 relaxase entries, 7658 auxiliary protein entries, 21 relaxosome entries and 16 985 T4CP entries into oriTDB, making it a robust and comprehensive database for *oriT* and its related elements. Notably, a substantial subset of 122 *oriT*s, 50 relaxases, 44 auxiliary proteins, 6 relaxosomes and 16 T4CPs has been experimentally validated, emphasizing the reliability of the database content (Supplementary Table S1). Additionally, our analysis has documented six categories of cargo genes found on *oriT*-carrying MGEs, encompassing acquired antimicrobial resistance genes, virulence factor genes, metal resistance genes, degradation genes, symbiosis genes, and anti-CRISPR genes. This information holds potential for enriching the understanding of the network involving *oriT*-MGEs-cargo genes across diverse bacterial species.

To further study in *oriT* region, we have dissected and provided interactive visualization of the compiled *oriT* in oriTDB (Figure 1). In the *oriT* entries, comprehensive information about *oriT*s, including the IRs, *nic* sites, conserved flanking sequences, regular sequences, and other specialized regions was provided (Figure 1A). Detailed information about *oriT* regions is visually presented in oriTDB, especially for the experimentally validated ones (Figure 1B). Furthermore, for the recorded *oriT* sequences, oriTDB employed RNAfold to predict and display their secondary structures, which display the complex relationships between the dissected elements within the *oriT* region through user-friendly web pages (Figure 1C) (21). Additionally, oriTDB provides information on the *oriT*-related relaxase (Figure 1D), auxiliary protein, T4CP and T4SS. Together, based on the dissection and interactive visualization of the *oriT* region, oriTDB provides a deeper understanding of the dissected elements within the *oriT* region.

Categorization for cargo genes in *oriT*-carrying MGEs

Bacterial MGEs carry a variety of cargo genes that can provide a competitive advantage to host bacteria. However, there

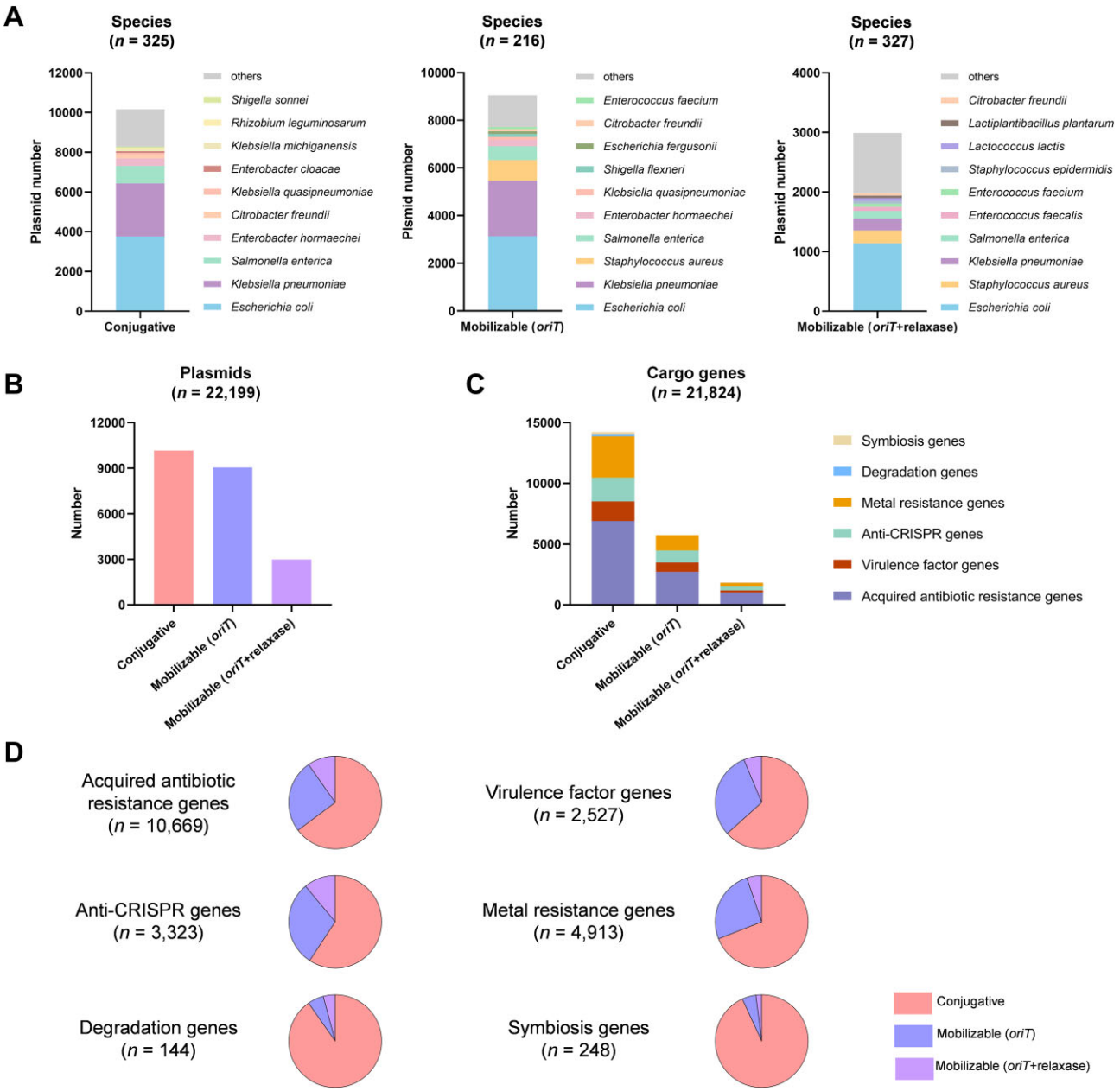


Figure 2. Species distribution and cargo genes categorization of the *oriT*-carrying plasmids predicted by oriTfinder2. (**A**) Species distribution of the *oriT*-carrying plasmids. (**B**) The number of the conjugative, mobilizable (*oriT*) and mobilizable (*oriT* + relaxase) plasmids. (**C**) The number of the six categorized cargo genes in the *oriT*-carrying plasmids. (**D**) The proportion of the three categorized *oriT*-carrying plasmids in the six categorized cargo genes.

is limited research on the relationship between the mobility of MGEs and their cargo genes (15,35). Previous studies have primarily focused on the mobility of specific types of MGEs and their associated cargo genes (11,36–38). To address this gap, we employed the oriTDB databases to investigate the potential mobility of plasmids and their correlation with carried cargo genes. Utilizing the plasmid classification based on transfer potential described in our previous work (13), we categorized the *oriT*-carrying plasmids (n = 22 199) predicted by oriTfinder2 into 'Conjugative' (carrying the *oriT*, relaxase gene, T4CP gene, and T4SS gene cluster) (n = 10 164), 'Mobilizable (*oriT*)' (carrying the *oriT* but lacking a relaxase gene) (n = 9 046) and 'Mobilizable (*oriT* + relaxase)' (carrying the *oriT* and a relaxase gene) (n = 2 989) (Figure 2B). Then we analyzed the species distribution (Figure 2A) and cargo genes associated with each category of the *oriT*-carrying plasmids (Figures 2C and D).

The 22 199 *oriT*-carrying plasmids are distributed across 583 species, such as *Escherichia coli*, *Klebsiella pneumoniae*, *Salmonella enterica*, *Enterobacter hormaechei*, *Enterococcus faecalis*, *Staphylococcus aureus* and *Citrobacter freundii* (Figure 2A). Of these, conjugative, mobilizable (*oriT*), and mobilizable (*oriT* + relaxase) plasmids are found in 325, 216 and 327 species, respectively (Supplementary Table S4). Notably, about 40.7% of *oriT*-carrying plasmids do not contain the relaxase gene (Figure 2B). Recent research has shown that more

than half of the plasmids in *S. aureus* are classified as 'Mobilizable (*oriT*)'(11,15). Our previous conjugation experiments have observed that mobilizable plasmids lacking the relaxase gene were transferable in *K. pneumoniae* through interaction with a helper conjugative plasmid (Supplementary Figure S2) (12,13). These findings imply that a considerable proportion of 'Mobilizable (*oriT*)' plasmids possess transferability. Additionally, the species of *oriT*-carrying ICEs and IMEs are detailed in Supplementary Figure S3 and Supplementary Table S5.

The primary focus of oriTDB has been to systematically classify and integrate cargo gene functions associated with *oriT*-carrying MGEs. The 'Browse' web page of oriTDB presents categorized cargo gene function lists for individual MGEs, allowing users to access detailed information about the cargo genes conveniently. Through *in silico* analysis using oriTfinder2, prevalent cargo genes within plasmids, ICEs, and IMEs have been categorized into six distinct groups (Figure 2C). Among 22 199 *oriT*-carrying plasmids, 21 824 cargo genes were annotated (Figure 2C). Furthermore, 14 233, 5761 and 1830 cargo genes were found in conjugative plasmids, mobilizable (*oriT*) plasmids, and mobilizable (*oriT* + relaxase) plasmids, respectively (Supplementary Table S4). Additionally, the cargo genes of *oriT*-carrying ICEs and IMEs are presented in Supplementary Figure S4 and Supplementary Table S5. Utilizing the oriTDB to analyze the cargo genes carried by MGEs may assist researchers in exploring the correlation between the potential mobility of the MGEs and their cargo genes, thereby providing further insights into the realm of MGE biology.

Conclusion

In this report, we introduced a user-friendly *oriT* database, accompanied by an enhanced *oriT* prediction tool. The oriTDB provides a comprehensive compilation of both experimentally validated and predicted *oriT* regions, along with their associated relaxases, auxiliary proteins, and T4CPs in bacterial plasmids, ICEs and IMEs. Additionally, oriTDB offers comprehensive information on the IRs and *nic* sites in the *oriT* region to support experimental research. Furthermore, the upgraded oriTfinder2 has not only improved the accuracy of its *oriT* region predictions but also incorporates annotation features for cargo genes in MGEs. The oriTDB database will be continuously maintained and updated to align with the rapidly expanding microbial genome database and ensure its ongoing relevance.

Data availability

oriTDB is freely available at https://bioinfo-mml.sjtu.edu.cn/oriTDB2/ .

Supplementary data

Supplementary Data are available at NAR Online.

Acknowledgements

The graphical abstract was created with BioRender (BioRender.com).

Funding

National Key Research and Development Program of China [2023YFC2307103]; National Natural Science Foundation of China [32070572, 82270630]; Shanghai Hospital Development Center Foundation [SHDC2022CRD039]; Talent Program on Public Health System Construction of Shanghai [GWV1-11.2-XD01]; Talent Program on academic clinicians of Shanghai Jiao Tong University School of Medicine [20240818]. Funding for open access charge: National Key Research and Development Program of China.

Conflict of interest statement

None declared.

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. Micro.*, **3**, 722–732.
2. 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.
3. Ilangovan,A., Kay,C.W.M., Roier,S., El Mkami,H., Salvadori,E., Zechner,E.L., Zanetti,G. and Waksman,G. (2017) Cryo-EM structure of a relaxase reveals the molecular basis of DNA unwinding during bacterial conjugation. *Cell*, **169**, 708–721.
4. Garcillán-Barcia,M.P., Redondo-Salvo,S. and de la Cruz,F. (2023) Plasmid classifications. *Plasmid*, **126**, 102684.
5. Francia,M.V., Varsaki,A., Garcillán-Barcia,M.P., Latorre,A., Drainas,C. and de la Cruz,F. (2004) A classification scheme for mobilization regions of bacterial plasmids. *FEMS Microbiol. Rev.*, **28**, 79–100.
6. Cuartas,R., Coque,T.M., de la Cruz,F. and Garcillán-Barcia,M.P. (2022) PLASmid TAXonomic PCR (PlasTax-PCR), a multiplex relaxase MOB typing to assort plasmids into taxonomic units. *Methods Mol. Biol.*, **2392**, 127–142.
7. Alvarado,A., Garcillán-Barcia,M.P. and de la Cruz,F. (2012) A degenerate primer MOB typing (DPMT) method to classify gamma-proteobacterial plasmids in clinical and environmental settings. *PLoS One*, **7**, e40438.
8. Garcillán-Barcia,M.P., Redondo-Salvo,S., Vielva,L. and de la Cruz,F. (2020) MOBscan: automated annotation of MOB relaxases. *Methods Mol. Biol.*, **2075**, 295–308.
9. Li,X., Xie,Y., Liu,M., Tai,C., Sun,J., Deng,Z. and Ou,H.-Y. (2018) oriTfinder: a web-based tool for the identification of origin of transfers in DNA sequences of bacterial mobile genetic elements. *Nucleic Acids Res.*, **46**, W229–W234.
10. Ares-Arroyo,M., Nucci,A. and Rocha,E.P.C. (2024) Identification of novel origins of transfer across bacterial plasmids. bioRxiv doi: https://doi.org/10.1101/2024.01.30.577996, 30 January 2024, preprint: not peer reviewed.
11. Ares-Arroyo,M., Coluzzi,C. and Rocha,E.P.C. (2023) Origins of transfer establish networks of functional dependencies for plasmid transfer by conjugation. *Nucleic Acids Res.*, **51**, 3001–3016.
12. Xu,Y., Zhang,J., Wang,M., Liu,M., Liu,G., Qu,H., Liu,J., Deng,Z., Sun,J., Ou,H.-Y., *et al.* (2021) Mobilization of the nonconjugative virulence plasmid from hypervirulent *Klebsiella pneumoniae*. *Genome Med.*, **13**, 119.
13. Zhang,J., Xu,Y., Wang,M., Li,X., Liu,Z., Kuang,D., Deng,Z., Ou,H.-Y. and Qu,J. (2023) Mobilizable plasmids drive the spread of antimicrobial resistance genes and virulence genes in *Klebsiella pneumoniae*. *Genome Medicine*, **15**, 106.
14. Zrimec,J. and Lapanje,A. (2018) DNA structure at the plasmid origin-of-transfer indicates its potential transfer range. *Sci. Rep.*, **8**, 1820.

15. O'Brien,F.G., Yui Eto,K., Murphy,R.J.T., Fairhurst,H.M., Coombs,G.W., Grubb,W.B. and Ramsay,J.P. (2015) Origin-of-transfer sequences facilitate mobilisation of non-conjugative antimicrobial-resistance plasmids in *Staphylococcus aureus*. *Nucleic Acids Res.*, **43**, 7971–7983.
16. Yui Eto,K., Kwong,S.M., LaBreck,P.T., Crow,J.E., Traore,D.A.K., Parahitiyawa,N., Fairhurst,H.M., Merrell,D.S., Firth,N., Bond,C.S., *et al.* (2021) Evolving origin-of-transfer sequences on staphylococcal conjugative and mobilizable plasmids-who's mimicking whom?*Nucleic Acids Res.*, **49**, 5177–5188.
17. Laroussi,H., Aoudache,Y., Robert,E., Libante,V., Thiriet,L., Mias-Lucquin,D., Douzi,B., Roussel,Y., Chauvot de Beauchêne,I., Soler,N., *et al.* (2022) Exploration of DNA processing features unravels novel properties of ICE conjugation in gram-positive bacteria. *Nucleic Acids Res.*, **50**, 8127–8142.
18. O'Leary,N.A., Wright,M.W., Brister,J.R., Ciufo,S., Haddad,D., McVeigh,R., Rajput,B., Robbertse,B., Smith-White,B., Ako-Adjei,D., *et al.* (2016) Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. *Nucleic Acids Res.*, **44**, D733–D745.
19. Bortolaia,V., Kaas,R.S., Ruppe,E., Roberts,M.C., Schwarz,S., Cattoir,V., Philippon,A., Allesoe,R.L., Rebelo,A.R., Florensa,A.F., *et al.* (2020) ResFinder 4.0 for predictions of phenotypes from genotypes. *J. Antimicrob. Chemother.*, **75**, 3491–3500.
20. Liu,B., Zheng,D., Zhou,S., Chen,L. and Yang,J. (2022) VFDB 2022: a general classification scheme for bacterial virulence factors. *Nucleic Acids Res.*, **50**, D912–D917.
21. Pal,C., Bengtsson-Palme,J., Rensing,C., Kristiansson,E. and Larsson,D.G. (2014) BacMet: antibacterial biocide and metal resistance genes database. *Nucleic Acids Res.*, **42**, D737–D743.
22. Dong,C., Wang,X., Ma,C., Zeng,Z., Pu,D.-K., Liu,S., Wu,C.-S., Chen,S., Deng,Z. and Guo,F.-B. (2022) Anti-CRISPRdb v2.2: an online repository of anti-CRISPR proteins including information on inhibitory mechanisms, activities and neighbors of curated anti-CRISPR proteins. *Database (Oxford)*, **2022**, baac010.
23. Ngara,T.R., Zeng,P. and Zhang,H. (2022) mibPOPdb: an online database for microbial biodegradation of persistent organic pollutants. *Imeta*, **1**, e45.
24. Wang,M., Liu,G., Liu,M., Tai,C., Deng,Z., Song,J. and Ou,H.-Y. (2024) ICEberg 3.0: functional categorization and analysis of the integrative and conjugative elements in bacteria. *Nucleic Acids Res.*, **52**, D732–D737.
25. Li,D., Mei,H., Shen,Y., Su,S., Zhang,W., Wang,J., Zu,M. and Chen,W. (2018) ECharts: a declarative framework for rapid construction of web-based visualization. *Visual Informatics*, **2**, 136–146.
26. Gruber,A.R., Lorenz,R., Bernhart,S.H., Neuböck,R. and Hofacker,I.L. (2008) The Vienna RNA websuite. *Nucleic Acids Res.*, **36**, W70–W74.
27. Grant,J.R. and Stothard,P. (2008) The CGView Server: a comparative genomics tool for circular genomes. *Nucleic Acids Res.*, **36**, W181–W184.
28. Mistry,J., Chuguransky,S., Williams,L., Qureshi,M., Salazar,G.A., Sonnhammer,E.L.L., Tosatto,S.C.E., Paladin,L., Raj,S., Richardson,L.J., *et al.* (2021) Pfam: the protein families database in 2021. *Nucleic Acids Res.*, **49**, D412–D419.
29. Jones,P., Binns,D., Chang,H.-Y., Fraser,M., Li,W., McAnulla,C., McWilliam,H., Maslen,J., Mitchell,A., Nuka,G., *et al.* (2014) InterProScan 5: genome-scale protein function classification. *Bioinformatics*, **30**, 1236–1240.
30. Berman,H.M., Westbrook,J., Feng,Z., Gilliland,G., Bhat,T.N., Weissig,H., Shindyalov,I.N. and Bourne,P.E. (2000) The Protein Data Bank. *Nucleic Acids Res.*, **28**, 235–242.
31. Varadi,M., Anyango,S., Deshpande,M., Nair,S., Natassia,C., Yordanova,G., Yuan,D., Stroe,O., Wood,G., Laydon,A., *et al.* (2022) AlphaFold Protein Structure Database: massively expanding the structural coverage of protein-sequence space with high-accuracy models. *Nucleic Acids Res.*, **50**, D439–D444.
32. Sehnal,D., Bittrich,S., Deshpande,M., Svobodová,R., Berka,K., Bazgier,V., Velankar,S., Burley,S.K., Koča,J. and Rose,A.S. (2021) Mol* viewer: modern web app for 3D visualization and analysis of large biomolecular structures. *Nucleic Acids Res.*, **49**, W431–W437.
33. Camacho,C., Coulouris,G., Avagyan,V., Ma,N., Papadopoulos,J., Bealer,K. and Madden,T.L. (2009) BLAST+: architecture and applications. *BMC Bioinf.*, **10**, 421.
34. Potter,S.C., Luciani,A., Eddy,S.R., Park,Y., Lopez,R. and Finn,R.D. (2018) HMMER web server: 2018 update. *Nucleic Acids Res.*, **46**, W200–W204.
35. Nayar,G., Terrizzano,I., Seabolt,E., Agarwal,A., Boucher,C., Ruiz,J., Slizovskiy,I.B., Kaufman,J.H. and Noyes,N.R. (2022) ggMOB: elucidation of genomic conjugative features and associated cargo genes across bacterial genera using genus-genus mobilization networks. *Front. Genet.*, **13**, 1024577.
36. Garcillán-Barcia,M.P., Alvarado,A. and de la Cruz,F. (2011) Identification of bacterial plasmids based on mobility and plasmid population biology. *FEMS Microbiol. Rev.*, **35**, 936–956.
37. Nguyen,Q., Nguyen,Y.T.P., Ha,T.T., Tran,D.T.N., Voong,P.V., Chau,V., Nguyen,P.L.N., Le,N.T.Q., Nguyen,L.P.H., Nguyen,T.T.N., *et al.* (2024) Genomic insights unveil the plasmid transfer mechanism and epidemiology of hypervirulent *Klebsiella pneumoniae* in Vietnam. *Nat. Commun.*, **15**, 4187.
38. Mahendra,C., Christie,K.A., Osuna,B.A., Pinilla-Redondo,R., Kleinstiver,B.P. and Bondy-Denomy,J. (2020) Broad-spectrum anti-CRISPR proteins facilitate horizontal gene transfer. *Nat. Microbiol.*, **5**, 620–629.

Received: August 15, 2024. **Revised:** September 12, 2024. **Editorial Decision:** September 19, 2024. **Accepted:** September 23, 2024

oriTDB: a database of the origin-of-transfer regions of bacterial mobile genetic elements

SUPPLEMENTARY METHODS

oriTfinder2 predicting *oriT*s and cargo genes in bacterial MGE sequences

SUPPLEMENTARY DATA

Table S1. Statistical summary of *oriT*s, relaxases, auxiliary proteins, T4CPs and relaxosomes archived in oriTDB.

Table S2. Profile Hidden Markov Models (HMM) for relaxases and T4CPs used by oriTfinder2.

Table S3. The prediction strategy of detecting cargo genes in a bacterial plasmid, ICE, or IME used by oriTfinder2.

Table S4. Species distribution and categorized cargo genes of the conjugative, mobilizable (*oriT*) and mobilizable (*oriT*+relaxase) plasmids.

Table S5. Species distribution and categorized cargo genes of the *oriT*-carrying ICEs and IMEs.

Figure S1. The prediction strategy used by oriTfinder2 to identify the *oriT region*, relaxase gene, T4CP gene, T4SS gene cluster, and cargo genes in a DNA sequence of bacterial plasmid, ICE, or IME.

Figure S2. Overview of oriTfinder2 outputs with two examples of the *oriT*-carrying plasmid and the mobilizable (*oriT*) plasmid.

Figure S3. Species distribution of the *oriT*-carrying ICEs and IMEs.

Figure S4. The number of the categorized cargo genes in the *oriT*-carrying ICEs and IMEs.

SUPPLEMENTARY METHODS

oriTfinder2 predicting *oriT*s and cargo genes in bacterial MGE sequences

The oriTfinder2 has been updated to improve *oriT* detection capabilities by expanding the experimental dataset from 52 to 122 *oriT* sequences in the background dataset oriTDB. It can now identify nine relaxase families and three T4CP families through an increased number of HMM profiles (Supplementary Table S2). Additionally, the tool can detect T4SS by integrating the CONJScan model from the macsyfinder software(1).

In addition to *oriT* detection, oriTfinder2 includes a functional annotation module for identifying cargo genes in bacterial MGEs. We established a comprehensive dataset containing acquired antibiotic resistance genes sourced from Resfinder(2), virulence factors obtained from VFDB(3), anti-CRISPR proteins from Anti-CRISPRdb v2.2(4), metal resistance determinants from BacMet2(5), microbial degradation proteins from mibPOPdb(6), and symbiotic proteins compiled by ICEberg3.0(7). Specifically, the symbiotic proteins are focused on the nitrogen-fixing symbiosis of rhizobia. The identification of cargo genes is carried out through BLAST searches using specified parameters (Supplementary Table S3).

Table S1. Statistical summary of *oriT*s, relaxases, auxiliary proteins, T4CPs and relaxosomes archived in oriTDB.

		Number of elements derived from experimental data	Number of elements derived from predicted data	Total
*oriT*DB (released in 2018)				
oriT	Plasmid	42	954	996
	ICE	6	67	73
	IME	4	2	6
relaxase		27	956	983
auxiliary protein		29	73	102
T4CP		12	452	464
*oriT*DB (updated in 2024)				
oriT	Plasmid	91	22,390	22,481
	ICE	18	482	500
	IME	13	55	68
relaxase		50	13,066	13,116
auxiliary protein		44	7,614	7,658
T4CP		16	16,969	16,985
relaxosome		6	15	21

Table S2. Profile Hidden Markov Models (HMM) for relaxases and T4CPs used by oriTfinder2

Conserved domain	Pfam ID [a]	Family
Relaxase		
Replic_Relax	PF13814	MOB_C
TrwC	PF08751	MOB_F
TraI	PF07057	MOB_F
TraI_2	PF07514	MOB_H
Relaxase	PF03432	MOB_P
MobA_MobL	PF03389	MOB_Q
Mob_Pre	PF01076	MOB_V
Rep_trans	PF02486	MOB_T
DUF5712 [b]	PF18976	MOB_B
MobL [b]	PF18555	MOB_L
T4CP		
T4SS-DNA_transf	PF02534	t4cp1
TrwB_AAD_bind	PF10412	t4cp2
FtsK_SpoIIIE [b]	PF01580	tcpA

[a] Pfam ID, protein family database accession number (http://pfam-legacy.xfam.org/).
[b] The HMM profile newly added by oriTfinder2.

Table S3. The prediction strategy of detecting cargo genes in a bacterial plasmid, ICE, or IME used by oriTfinder2.

Cargo genes	Method	Dataset	Ref.
Antibiotic resistance genes (ARGs)	Abricate (identities≥80%, coverage≥80%)	Resfinder database	(2)
Virulence factors	BLASTp (E-value ≤0.0001; Ha-value ≥ 0.64)	VFDB database	(3,7)
Metal resistance determinants	BLASTp (E-value ≤0.0001; Ha-value ≥ 0.64)	metal resistance proteins from BacMet2	(7)
Degradation proteins	BLASTp (E-value ≤0.0001; Ha-value ≥ 0.64)	degradation proteins from mibPOPdb	(7)
Symbiosis proteins	BLASTp (E-value ≤0.0001; Ha-value ≥ 0.64)	symbiotic proteins compiled by ICEberg3	(7)
Anti-CRISPR proteins	BLASTp (E-value ≤0.001)	Anti-CRISPRdbv2.2-Verified+Pliterature	(4,8)

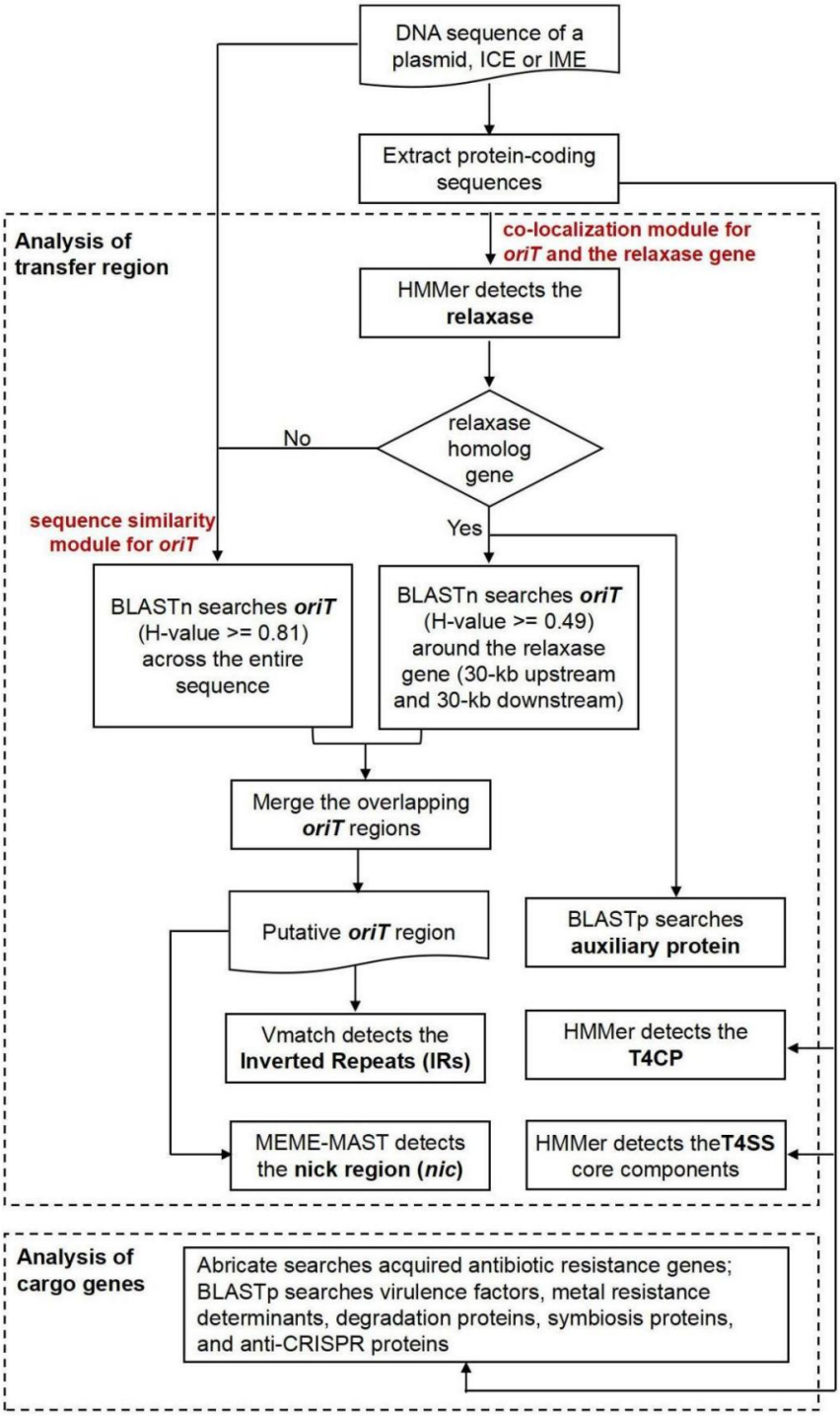


Figure S1. The prediction strategy used by oriTfinder2 to identify the *oriT region*, relaxase gene, T4CP gene, T4SS gene cluster, and cargo genes in the DNA sequence of a bacterial plasmid, ICE, or IME.

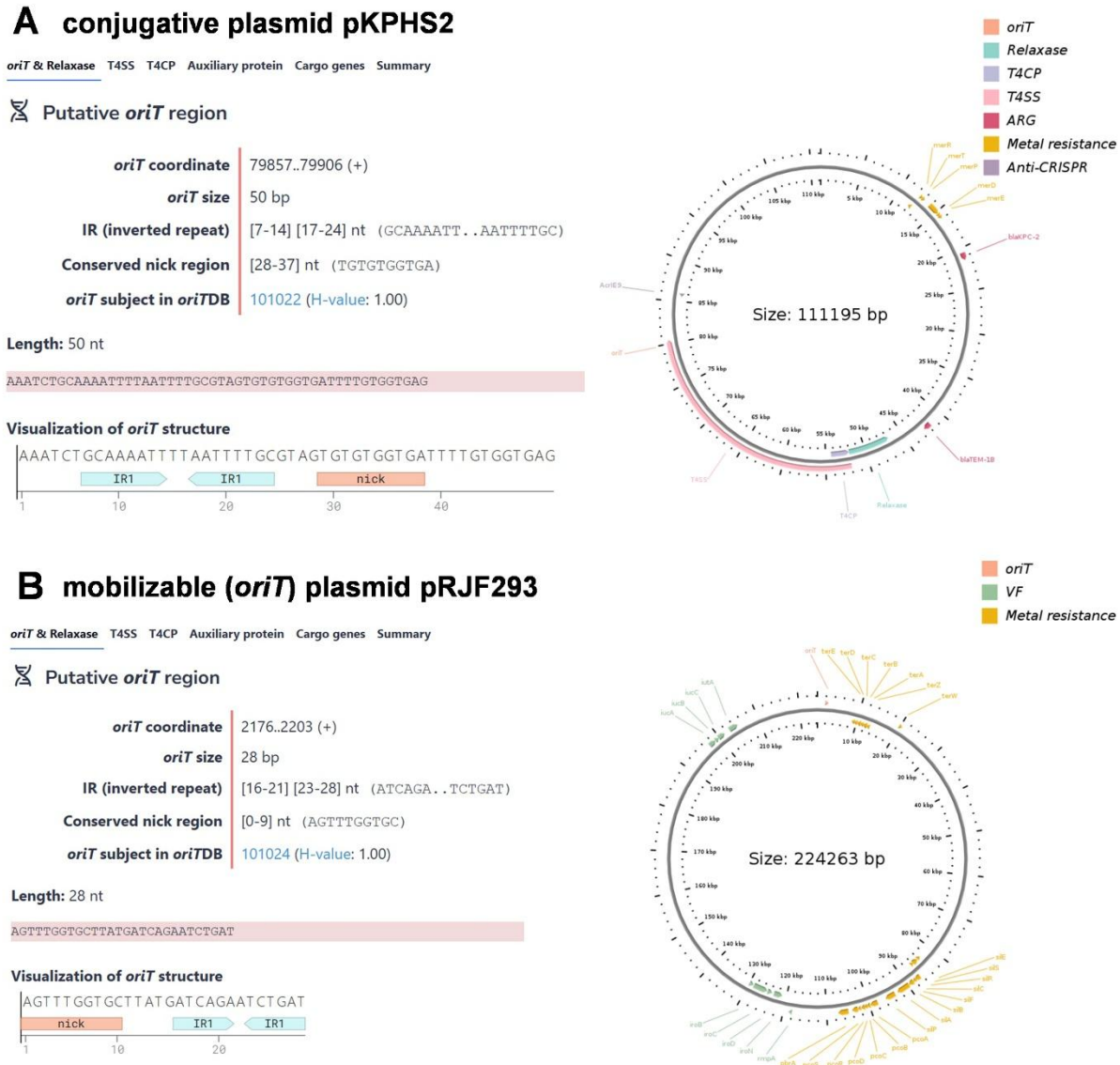


Figure S2. Overview of oriTfinder2 outputs with two examples of the *oriT*-carrying plasmid and the mobilizable (*oriT*) plasmid. (A) The conjugative plasmid pKPHS2 of *Klebsiella pneumoniae* HS11286 (NCBI accession no. NC_016846). (B) The mobilizable (*oriT*) plasmid pRJF293 of *Klebsiella pneumoniae* RJF293 (NCBI accession no. NZ_CP014009). The *oriT* region, relaxase gene, T4CP gene, gene cluster coding for T4SS, and cargo genes are marked in the circular visualization of the plasmid in different colors.

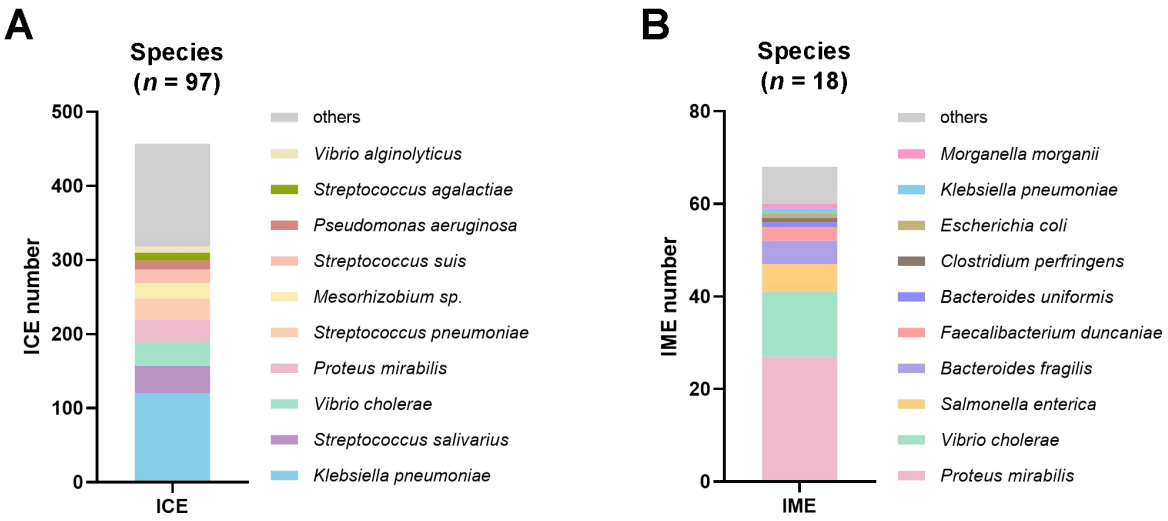


Figure S3. Species distribution of the *oriT*-carrying ICEs (A) and IMEs (B).

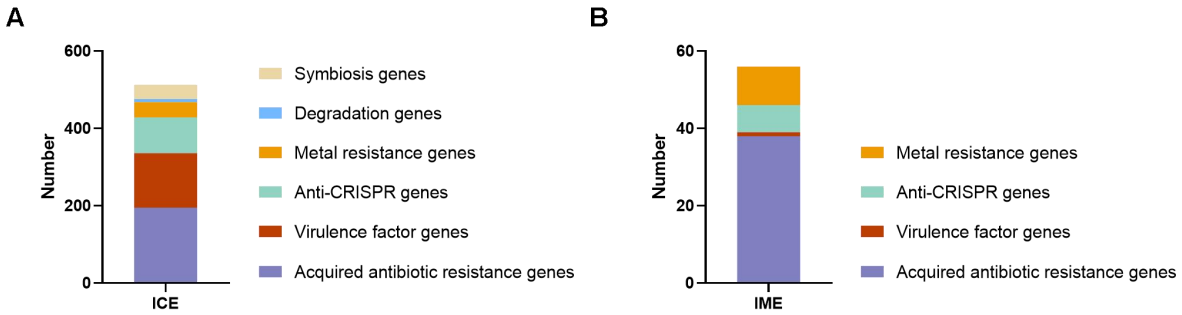


Figure S4. The number of the categorized cargo genes in the *oriT*-carrying ICEs (A) and IMEs (B).

REFERENCES

1. Abby, S.S., Néron, B., Ménager, H., Touchon, M. and Rocha, E.P.C. (2014) MacSyFinder: a program to mine genomes for molecular systems with an application to CRISPR-Cas systems. *PLoS One*, **9**, e110726.
2. Wang, M., Goh, Y.X., Tai, C., Wang, H., Deng, Z. and Ou, H.Y. (2022) VRprofile2: detection of antibiotic resistance-associated mobilome in bacterial pathogens. *Nucleic Acids Res*, **50**, W768-w773.
3. Liu, B., Zheng, D., Zhou, S., Chen, L. and Yang, J. (2022) VFDB 2022: a general classification scheme for bacterial virulence factors. *Nucleic Acids Research*, **50**, D912-D917.
4. Dong, C., Wang, X., Ma, C., Zeng, Z., Pu, D.K., Liu, S., Wu, C.S., Chen, S., Deng, Z. and Guo, F.B. (2022) Anti-CRISPRdb v2.2: an online repository of anti-CRISPR proteins including information on inhibitory mechanisms, activities and neighbors of curated anti-CRISPR proteins. *Database (Oxford)*, **2022**.
5. Pal, C., Bengtsson-Palme, J., Rensing, C., Kristiansson, E. and Larsson, D.G. (2014) BacMet: antibacterial biocide and metal resistance genes database. *Nucleic Acids Res*, **42**, D737-743.
6. Ngara, T.R., Zeng, P. and Zhang, H. (2022) mibPOPdb: An online database for microbial biodegradation of persistent organic pollutants. *Imeta*, **1**, e45.
7. Wang, M., Liu, G., Liu, M., Tai, C., Deng, Z., Song, J. and Ou, H.-Y. (2024) ICEberg 3.0: functional categorization and analysis of the integrative and conjugative elements in bacteria. *Nucleic Acids Research*, **52**, D732-D737.
8. Liu, M., Li, X., Xie, Y., Bi, D., Sun, J., Li, J., Tai, C., Deng, Z. and Ou, H.Y. (2019) ICEberg 2.0: an updated database of bacterial integrative and conjugative elements. *Nucleic acids research*, **47**, D660-d665.