The contribution of plasmids to trait diversity in a soil bacterium

Sarai S. Finks[#], Pranav Moudgalya[a], Claudia Weihe[b], Jennifer B.H. Martiny[b]

Department of Ecology and Evolutionary Biology, University of California Irvine, Irvine, California, USA

Running Header: Plasmid diversity in a soil bacterium

#Corresponding author. Current address: Department of Biology, The Pennsylvania State University, University Park, Pennsylvania, USA, ssf5197@psu.edu, W-239A Millennium Science Complex, Pollock Road, University Park, PA 16802 US

ABSTRACT

Plasmids are so closely associated with pathogens and antibiotic resistance that their potential for conferring other traits is often overlooked. Few studies consider how the full suite of traits encoded by plasmids is related to a host's environmental adaptation, particularly for gram-positive bacteria. To investigate the role that plasmid traits might play in microbial communities from natural ecosystems, we identified plasmids carried by isolates of *Curtobacterium* (phylum *Actinomycetota*) from a variety of soil environments. We found that plasmids were common, but not ubiquitous, in the genus and varied greatly in their size and genetic diversity. There was little evidence of phylogenetic conservation among *Curtobacterium* plasmids even for closely-related bacterial strains within the same ecotype, indicating that horizontal transmission of plasmids is common. The plasmids carried a wide diversity of traits that were not a random subset of the host chromosome. Further, the composition of these plasmid traits was associated with the environmental context of the host bacterium. Together, the results indicate that plasmids contribute substantially to the microdiversity of a soil bacterium and that this diversity may play a role in niche differentiation and a bacterium's adaptation to its local environment.

KEY WORDS: microdiversity, mobile genetic elements, plasmids, genetic traits, HGT, plant litter, *Curtobacterium*, *Actinobacteria*, *Actinomycetota*

A high degree of genetic variation is encompassed within traditional operational taxonomic units (OTUs) of bacteria [1]. This so-called microdiversity encompasses an enormous amount of variability in traits that influence a bacterium's ecological role and its contributions to community functioning [2–4]. Plasmids may contribute to this microdiversity as they can encode a diversity of traits [5] that may allow a bacterium to adapt rapidly to environmental changes [6]. The most striking examples of this are the transfer of metal and antibiotic resistance, particularly in the human gut microbiome and clinical environments [7–10]. Beyond toxin resistance, however, evidence of the importance of plasmids to broader niche-adaptation is sporadic [11, 12] Most of what we currently know is based on a handful of well represented genera (e.g., *Vibrio, Pseudomonas,* and *Burkholdaria*) within the phylum *Pseudomonadota* (e.g., reference [13]) and few studies consider gram-positive bacteria (e.g., reference [14]) but see, Finks and Martiny, 2023 [5].

A general understanding of plasmid evolution, the diversity of traits that they carry, and their importance for adaptation in most bacterial communities thus remains elusive [5, 15]. To investigate these unknowns in a soil bacterium, we focused on the widespread genus *Curtobacterium* [16] for which we have isolated a number of closely-related strains from the top layer of soil (plant litter) in different environments. *Curtobacterium* strains associated with plant disease can carry plasmids encoding for putative virulence encoded genes [17]. However, plasmid prevalence and diversity for this genus, as in other soil bacteria, is largely uncharacterized.

Plasmids can mobilize across broad bacterial host ranges [18], interact with other types of mobile genetic elements [19], and recombine with their hosts [20]. We thus expected that *Curtobacterium* plasmids would also be subject to a high degree of mobility and recombination.

However, plasmids are also vertically transmitted to daughter cells during host cell replication such that, at some level of genetic resolution, they will be phylogenetically conserved. Thus, plasmids might be conserved within *Curtobacterium* ecotypes, previously defined as genetic clades with similar phenotypes that are adapted to local environmental conditions including temperature and moisture [21]. Alternatively, selection might act on plasmids separately from that of an ecotype's chromosome such that plasmid traits vary by environment rather than host phylogeny. To test these alternatives, here we asked: (1) Are plasmids within the *Curtobacterium* genus phylogenetically conserved? (2) What traits do the plasmids encode and how do these compare to the chromosome? (3) Are plasmid traits correlated with the environment from which they were isolated?

Long-read sequencing of 23 strains and additional reference genomes resulted in analysis of 26 putative plasmids from 18 *Curtobacterium* strains (**Figure 1**; **Supplemental Methods**). Three lines of evidence suggest that these sequences are indeed plasmids. First, the average plasmid GC content was approximately 7% lower relative to the host chromosomes (**Figure 1D**). Second, the topology (usually circular) and replicon sizes (smaller than the chromosome) of the sequences are well-known signatures of plasmids [22, 23]. Third, all but one plasmid (pD03b) carried some kind of plasmid feature. Interestingly, two plasmids of strain P990 showed % GC contents that were half that of other plasmids (32.3 % and 35.3 % versus ~ 67 %; **Table S1**), suggesting more recent acquisition of these mobile genetic elements. Approximately half of the plasmid sequences encoded genes for known plasmid replicon types (RepA-type, n=4; **Table S6**) or MOB relaxases (MOBF or MOBP, n=12; **Table S7**). In addition, some plasmids carried genes necessary for conjugative, cell-to-cell DNA transfer (e.g., *trwC*) and for partitioning to daughter cells during host replication and division (e.g., *parA/B/G*; **Figure S1**). Based on sequencing

coverage, most plasmids appeared to be present in single copies, whereas some smaller ones were present in high-copy numbers (**Table S1**). None of the *Curtobacterium* plasmid sequences grouped into known plasmid taxonomic units (PTUs), although this is not surprising given the low representation of *Actinomycetota* in databases (**Supplementary Methods** [18, 24]).

Plasmids were common among *Curtobacterium* strains, but their distribution across the phylogeny was not random. Plasmids were notably absent from ecotype IV and very common in ecotype I (**Figure 1A**). That said, plasmid size varied greatly even within clades (1.5 – 607 kb, mean=136 kb), supporting the idea that plasmids are not phylogenetically conserved in this genus (**Figure 1C**). Indeed, genetic (mash) similarity of the plasmids was not correlated with the genetic similarity of the host chromosomes (**Figure 1B;** RELATE: r = 0.28; $P = 0.08$).

Curtobacterium plasmids encoded more than 4,000 gene calls that clustered into 2,396 distinct orthologous groups (**Figure S1**). Despite making up only 3% of the gene content of the entire dataset, this genetic diversity spanned 22 COG functional categories. Based on whole genome alignments, *Curtobacterium* plasmids did not appear to share a conserved backbone, such as is commonly observed for some IncF type plasmids found in *Enterobacteriaceae* [25]. Only one gene, *lsr2* (a putative histone-like protein), was shared by 38% of the 26 plasmids, whereas most other genes were shared by fewer than 3 plasmids (**Figure S1**). BlastP searches of consensus amino acid sequence alignments of Lsr2 against the NCBI Reference Proteins (refseq_protein) database reveals this small protein (~12 kDa) is ubiquitous throughout the genus. In *M. smegmatis,* this protein appears to be involved in the biosynthesis of mycolyl-diacylglycerols, an apolar lipid in the cell wall, as well as a DNA-binding function having a transcriptional regulatory role [26–28].

The *Curtobacterium* plasmids encoded a diversity of traits that were not a random subset of chromosomal traits (G (21) = 1203.2, $P < 0.001$; **Figure 2A**). Not surprisingly, genes associated with the mobilome, prophages and transposons (X) were relatively more prevalent on plasmids than the chromosome, but other functions including those associated with cell motility (N) were relatively more abundant on plasmids than on chromosomes (**Figure 2B**). Conversely, carbohydrate transport and metabolism functions (G) were more prevalent on *Curtobacterium* chromosomes than plasmids. Given their role in soil carbon cycling, it is notable that 11 plasmids carried 46 CAZyme (carbohydrate active enzyme) genes (**Table S9; Figure S2A**), and in more than half of these cases, the CAZyme family was not present on the associated host chromosome. We also identified two genes encoding nitrate assimilation (*narB*) on a plasmid (**Figure S2B**).

Finally, plasmid trait composition differed significantly by the environment from which the host was isolated, explaining ~14% of variation in COG functional categories (PERMANOVA: Pseudo-F (7): 1.424, $P = 0.042$). For instance, plasmids isolated from grassland and alpine environments encoded a higher prevalence of carbohydrate transport and metabolism (G) genes, whereas those isolated from two arid environments (Desert and Salton-Sea), encoded a relatively high number of genes associated with cell motility (N) and translation, ribosomal structure and biogenesis (J) (**Figure 2B**).

Our results indicate that plasmids contribute substantially to the microdiversity of *Curtobacterium* and that this diversity may play a role in its adaptation to the local environment. Horizontal transfer appeared to break up any signal of vertical transmission of plasmids, even within *Curtobacterium* ecotypes. However, only about half the plasmids encoded for genes known to facilitate mobility from one bacterium to another. This result is similar to that of

marine *Vibrio spp*., where plasmids also appear to spread rapidly by horizontal gene transfer, many by unknown mechanisms [29].

This work also highlights the paucity in knowledge about which plasmid traits will be favored in natural ecosystems. Models investigating the evolutionary mechanisms that sustain plasmid diversity suggest that they should encode traits, like antibiotic resistance, that are widely beneficial to many bacterial species and come under relatively strong selection [30]. Future investigations into how, when, and where plasmid traits such as cell motility provide soil bacteria with an advantage would provide a more in-depth understanding of the eco-evolutionary role of these mobile genetic elements in soil.

DATA AVAILABILTY

All scripts are accessible on GitHub (https://github.com/SaraiFinks). The ONT sequenced *Curtobacterium* and assemblies can be retrieved from NCBI SRA and Genome databases under BioProject ID PRJNA391502.

ACKNOWLEDGEMENTS

We thank Adam C. Martiny, Katrine L. Whiteson, Alex B. Chase, Kristin M. Barbour, Alberto Barron Sandoval, and Lauren Lui for helpful insights and comments on this work. This work was supported by a National Science Foundation Graduate Research Fellowship and UCI Chancellor's Club Fellowship to SSF, and by the U.S. Department of Energy, Office of Science, Office of Biological and Environmental Research grants DE-SC0016410 and DE-SC0020382.

ETHICS DECLARATIONS

Competing interests

The authors declare no conflicts of interest.

REFERENCES

1. Jaspers E, Overmann J. Ecological significance of microdiversity: Identical 16S rRNA gene sequences can be found in bacteria with highly divergent genomes and ecophysiologies. *Appl Environ Microbiol* 2004; **70**: 4831–4839.
2. Schloter M, Lebuhn M, Heulin T, Hartmann A. Ecology and evolution of bacterial microdiversity. *FEMS Microbiol Rev* 2000; **24**: 647–660.
3. Brazelton WJ, Ludwig KA, Sogin ML, Andreishcheva EN, Kelley DS, Shen CC, et al. Archaea and bacteria with surprising microdiversity show shifts in dominance over 1,000-year time scales in hydrothermal chimneys. *Proc Natl Acad Sci U S A* 2010; **107**: 1612–1617.
4. Acinas SG, Klepac-Ceraj V, Hunt DE, Pharino C, Ceraj I, Distel DL, et al. Fine-scale phylogenetic architecture of a complex bacterial community. *Nature 2004 430:6999* 2004; **430**: 551–554.
5. Finks SS, Martiny JBH. Plasmid-Encoded Traits Vary across Environments. *mBio* 2023; **14**.
6. Steinberg AP, Lin M, Kussell E. Core genes can have higher recombination rates than accessory genes within global microbial populations. *Elife* 2022; **11**.
7. Carattoli A. Plasmids and the spread of resistance. *Int J Med Microbiol* 2013; **303**: 298–304.
8. Peter S, Bosio M, Gross C, Bezdan D, Gutierrez J, Oberhettinger P, et al. Tracking of Antibiotic Resistance Transfer and Rapid Plasmid Evolution in a Hospital Setting by Nanopore Sequencing. *mSphere* 2020; **5**.
9. Gumpert H, Kubicek-Sutherland JZ, Porse A, Karami N, Munck C, Linkevicius M, et al. Transfer and persistence of a multi-drug resistance plasmid in situ of the infant gut microbiota in the absence of antibiotic treatment. *Front Microbiol* 2017; **8**: 298526.
10. Smillie CS, Smith MB, Friedman J, Cordero OX, David LA, Alm EJ. Ecology drives a global network of gene exchange connecting the human microbiome. *Nature* 2011; **480**: 241–244.
11. Estrada-De Los Santos P, Bustillos-Cristales R, Caballero-Mellado J. Burkholderia, a Genus Rich in Plant-Associated Nitrogen Fixers with Wide Environmental and Geographic Distribution. *Appl Environ Microbiol* 2001; **67**: 2790–2798.
12. Yang LL, Jiang Z, Li Y, Wang ET, Zhi XY. Plasmids Related to the Symbiotic Nitrogen Fixation Are Not Only Cooperated Functionally but Also May Have Evolved over a Time Span in Family Rhizobiaceae. *Genome Biol Evol* 2020; **12**: 2002–2014.
13. Chibani CM, Roth O, Liesegang H, Wendling CC. Genomic variation among closely related Vibrio alginolyticus strains is located on mobile genetic elements. *BMC Genomics* 2020; **21**: 1–14.

14. Gushgari-Doyle S, Lui LM, Nielsen TN, Wu X, Malana RG, Hendrickson AJ, et al. Genotype to ecotype in niche environments: adaptation of Arthrobacter to carbon availability and environmental conditions. *ISME Communications 2022 2:1* 2022; **2**: 1–10.
15. Martiny JBH, Martiny AC, Brodie E, Chase AB, Rodríguez-Verdugo A, Treseder KK, et al. Investigating the eco-evolutionary response of microbiomes to environmental change. *Ecol Lett* 2023.
16. Chase AB, Arevalo P, Polz MF, Berlemont R, Martiny JBH. Evidence for Ecological Flexibility in the Cosmopolitan Genus Curtobacterium . *Frontiers in Microbiology* . 2016. , **7**: 1874
17. Chen G, Khojasteh M, Taheri-Dehkordi A, Taghavi SM, Rahimi T, Osdaghi E. Complete Genome Sequencing Provides Novel Insight Into the Virulence Repertories and Phylogenetic Position of Dry Beans Pathogen Curtobacterium flaccumfaciens pv. flaccumfaciens. *Phytopathology* 2020; **111**: 268–280.
18. Redondo-Salvo S, Fernández-López R, Ruiz R, Vielva L, de Toro M, Rocha EPC, et al. Pathways for horizontal gene transfer in bacteria revealed by a global map of their plasmids. *Nat Commun* 2020; **11**: 3602.
19. Horne T, Orr VT, Hall JP. How do interactions between mobile genetic elements affect horizontal gene transfer? *Curr Opin Microbiol* 2023; **73**: 102282.
20. Heuer H, Smalla K. Plasmids foster diversification and adaptation of bacterial populations in soil. *FEMS Microbiol Rev* 2012; **36**: 1083–1104.
21. Chase AB, Weihe C, Martiny JBH. Adaptive differentiation and rapid evolution of a soil bacterium along a climate gradient. *Proceedings of the National Academy of Sciences* 2021; **118**: e2101254118.
22. Krawczyk PS, Lipinski L, Dziembowski A. PlasFlow: predicting plasmid sequences in metagenomic data using genome signatures. *Nucleic Acids Res* 2018; **46**: e35.
23. Smillie C, Garcillan-Barcia MP, Francia MV, Rocha EPC, de la Cruz F. Mobility of plasmids. *Microbiol Mol Biol Rev* 2010; **74**: 434–452.
24. Redondo-Salvo S, Bartomeus-Peñalver R, Vielva L, Tagg KA, Webb HE, Fernández-López R, et al. COPLA, a taxonomic classifier of plasmids. *BMC Bioinformatics* 2021; **22**: 1–9.
25. Fernandez-Lopez R, de Toro M, Moncalian G, Garcillan-Barcia MP, de la Cruz F. Comparative genomics of the conjugation region of F-like plasmids: Five shades of F. *Front Mol Biosci* 2016; **3**: 71.
26. Chen JM, Ren H, Shaw JE, Wang YJ, Li M, Leung AS, et al. Lsr2 of Mycobacterium tuberculosis is a DNA-bridging protein. *Nucleic Acids Res* 2008; **36**: 2123.
27. Colangeli R, Helb D, Vilchèze C, Hazbón MH, Lee CG, Safi H, et al. Transcriptional Regulation of Multi-Drug Tolerance and Antibiotic-Induced Responses by the Histone-Like Protein Lsr2 in M. tuberculosis. *PLoS Pathog* 2007; **3**: e87.
28. Chen JM, German GJ, Alexander DC, Ren H, Tan T, Liu J. Roles of Lsr2 in colony morphology and biofilm formation of Mycobacterium smegmatis. *J Bacteriol* 2006; **188**: 633–641.
29. Xue H, Cordero OX, Camas FM, Trimble W, Meyer F, Guglielmini J, et al. Eco-Evolutionary Dynamics of Episomes among Ecologically Cohesive Bacterial Populations. *mBio* 2015; **6**.
30. Lehtinen S, Huisman JS, Bonhoeffer S. Evolutionary mechanisms that determine which bacterial genes are carried on plasmids. *Evol Lett* 2021; **5**: 290–301.

232

233

FIGURE LEGENDS

Figure 1. *Curtobacterium* plasmids by host ecotype and their genetic similarity, size distribution, and GC content. (**A**) Cladogram of complete chromosomes of *Curtobacterium* constructed from a phylogenomic analysis of 916 single-copy core genes. All branches displayed represent bootstrap values of 95% confidence or greater. Bolded values next to strain identifiers are nucleotide lengths of plasmids in bp. The branches are colored by ecotype designation with adjacent color tiles indicating the environment from which the strain was isolated. Note: asterisks indicate the two plasmids in host P990 with relatively lower % GC content (see panel D) compared to the others (**B**) Heatmap of plasmids constructed from mash pairwise similarities. The strain identifiers are listed by row and the plasmid identifier (**Table S1**) as columns. The color tiles beside the row labels indicate the environment as in panel A. (**C**) The frequency of plasmid sizes in kilobases across all strains, where the x-axis is the lower bound of each 25kB bin. (**D**) Percent GC content for each plasmid and its corresponding chromosome.

Figure 2. *Curtobacterium* plasmid COG functions are distinct from chromosomal functions and vary by environment. (**A**) Percentages (Log_{10} scaled) of COG functional category counts of the *Curtobacterium* plasmids (top) and chromosome (bottom) sequences. The total number of COG functions identified on *Curtobacterium* plasmids and chromosomes are shown in parentheses. No COG functions for category Z were identified on the plasmids, and plasmids pCff2, pCff3, and pD03b are not included as no COG functions were identified. **(B)** Normalized frequencies of COG categories encoded by the plasmids by environment. The counts of COG functions were first converted into proportional abundances within an environment after removal of COG functions (n = < 6), and COG counts were then normalized across environments using

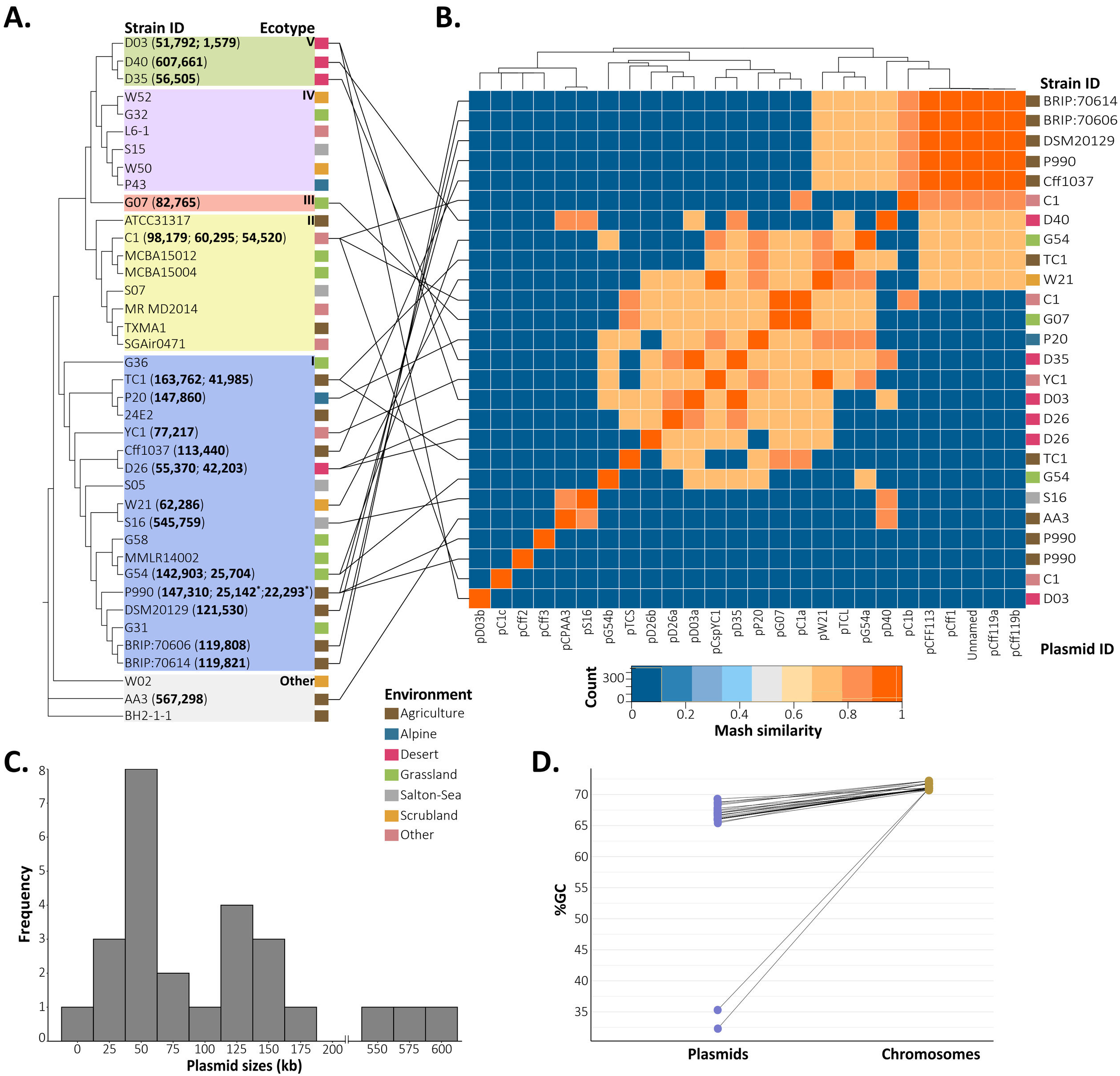
Z-scores to standardize for uneven representation of plasmids across environments. **C** - Energy production and conversion; **D** - Cell cycle control, cell division, chromosome partitioning; **E** - Amino acid transport and metabolism; **F** - Nucleotide transport and metabolism; **G** - Carbohydrate transport and metabolism; **H** - Coenzyme transport and metabolism; **I** - Lipid transport and metabolism; **J** - Translation, ribosomal structure and biogenesis; **K** – Transcription; **L** - Replication, recombination and repair; **M** - Cell wall/membrane/envelope biogenesis; **N** - Cell motility; **O** - Posttranslational modification, protein turnover, chaperones; **P** - Inorganic ion transport and metabolism; **Q** - Secondary metabolites biosynthesis, transport and catabolism; **R** - General function; **S** - Unknown function; **T** - Signal transduction mechanisms; **U** - Intracellular trafficking, secretion, and vesicular transport; **V** - Defense mechanisms; **X** - Mobilome: prophages, transposons.

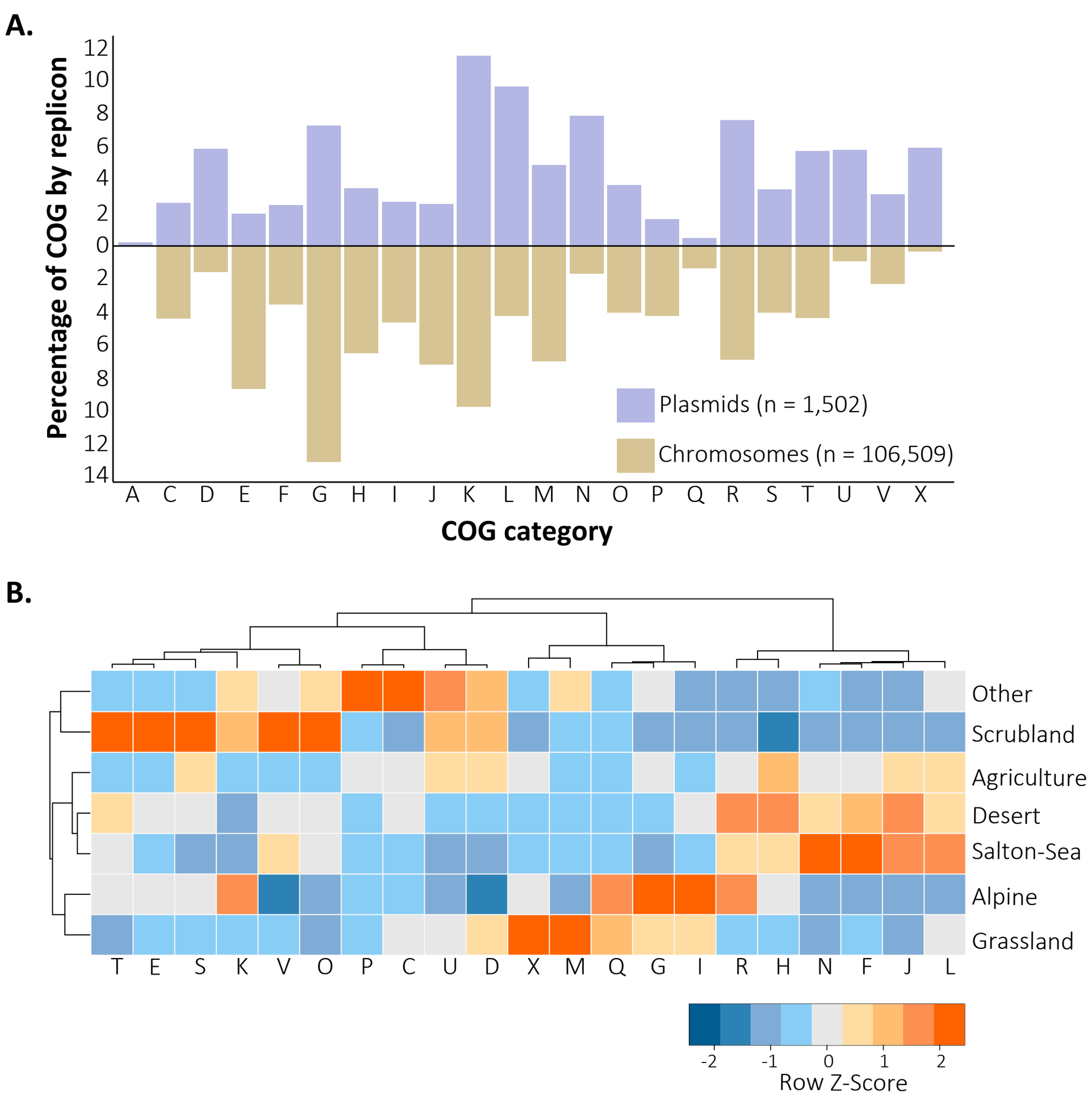
A.

B.

C.

D.


Strain ID
Ecotype
D03 (51,792; 1,579)
D40 (607,661)
D35 (56,505)
W52
G32
L6-1
S15
W50
P43
G07 (82,765)
ATCC31317
C1 (98,179; 60,295; 54,520)
MCBA15012
MCBA15004
S07
MR MD2014
TXMA1
SGAir0471
G36
TC1 (163,762; 41,985)
P20 (147,860)
24E2
YC1 (77,217)
Cff1037 (113,440)
D26 (55,370; 42,203)
S05
W21 (62,286)
S16 (545,759)
G58
MMLR14002
G54 (142,903; 25,704)
P990 (147,310; 25,142*;22,293*)
DSM20129 (121,530)
G31
BRIP:70606 (119,808)
BRIP:70614 (119,821)
W02
AA3 (567,298)
BH2-1-1
V
IV
III
II
I
Other
Environment
Agriculture
Alpine
Desert
Grassland
Salton-Sea
Scrubland
Other
Strain ID
BRIP:70614
BRIP:70606
DSM20129
P990
Cff1037
C1
D40
G54
TC1
W21
C1
G07
P20
D35
YC1
D03
D26
D26
TC1
G54
S16
AA3
P990
P990
C1
D03
pD03b
pC1c
pCff2
pCff3
pCPAA3
pS16
pG54b
pTCS
pD26b
pD26a
pD03a
pCspYC1
pD35
pP20
pG07
pC1a
pW21
pTCL
pG54a
pD40
pC1b
pCFF113
pCff1
Unnamed
pCff119a
pCff119b
Plasmid ID
Count
300
0
0
0.2
0.4
0.6
0.8
1
Mash similarity
Frequency
0
1
2
3
4
5
6
7
8
0
25
50
75
100
125
150
175
200
550
575
600
Plasmid sizes (kb)
%GC
35
40
45
50
55
60
65
70
Plasmids
Chromosomes


A.
Percentage of COG by replicon
12
10
8
6
4
2
0
2
4
6
8
10
12
14
A C D E F G H I J K L M N O P Q R S T U V X
COG category
Plasmids (n = 1,502)
Chromosomes (n = 106,509)
B.
Other
Scrubland
Agriculture
Desert
Salton-Sea
Alpine
Grassland
T E S K V O P C U D X M Q G I R H N F J L
-2 -1 0 1 2
Row Z-Score

Curtobacterium Plasmids

2,396 gene clusters

4,382 gene calls

Pfam

COG20 Functions

pG07 - 82,765
pD26a - 55,370
pTCS - 41,985
pC1a - 98,179
pD26b - 42,203
pD03a - 51,792
pD35 - 56,505
pD40 - 607,661
pG54a - 142,903
pCspYC1 - 77,217
pW21 - 62,286
pP20 - 147,860
pTCL - 163,762
pCff119a - 119,821
pCff119b - 119,808
pCFF113 - 113,440
Unnamed - 121,530
pCff1 - 147,310
pC1b - 60,295
pC1c - 54,520
pCPAA3 - 567,298
pS16 - 545,759
pG54b - 25,704
pD03b - 1,579
pCff2 - 25,142
pCff3 - 22,293

Ecotype

436

Singleton gene clusters

0

2.65

Num genes per kbp

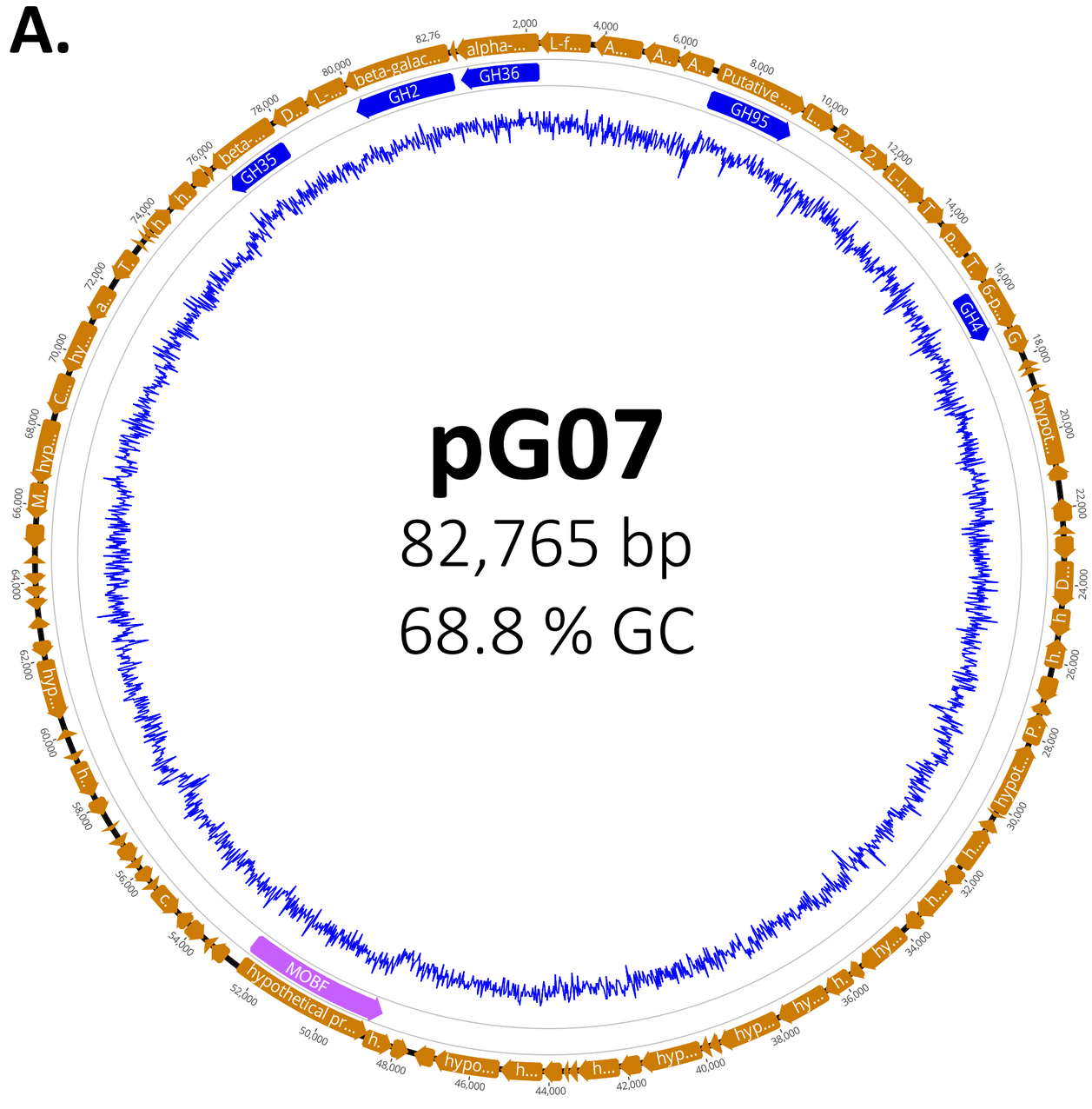
0

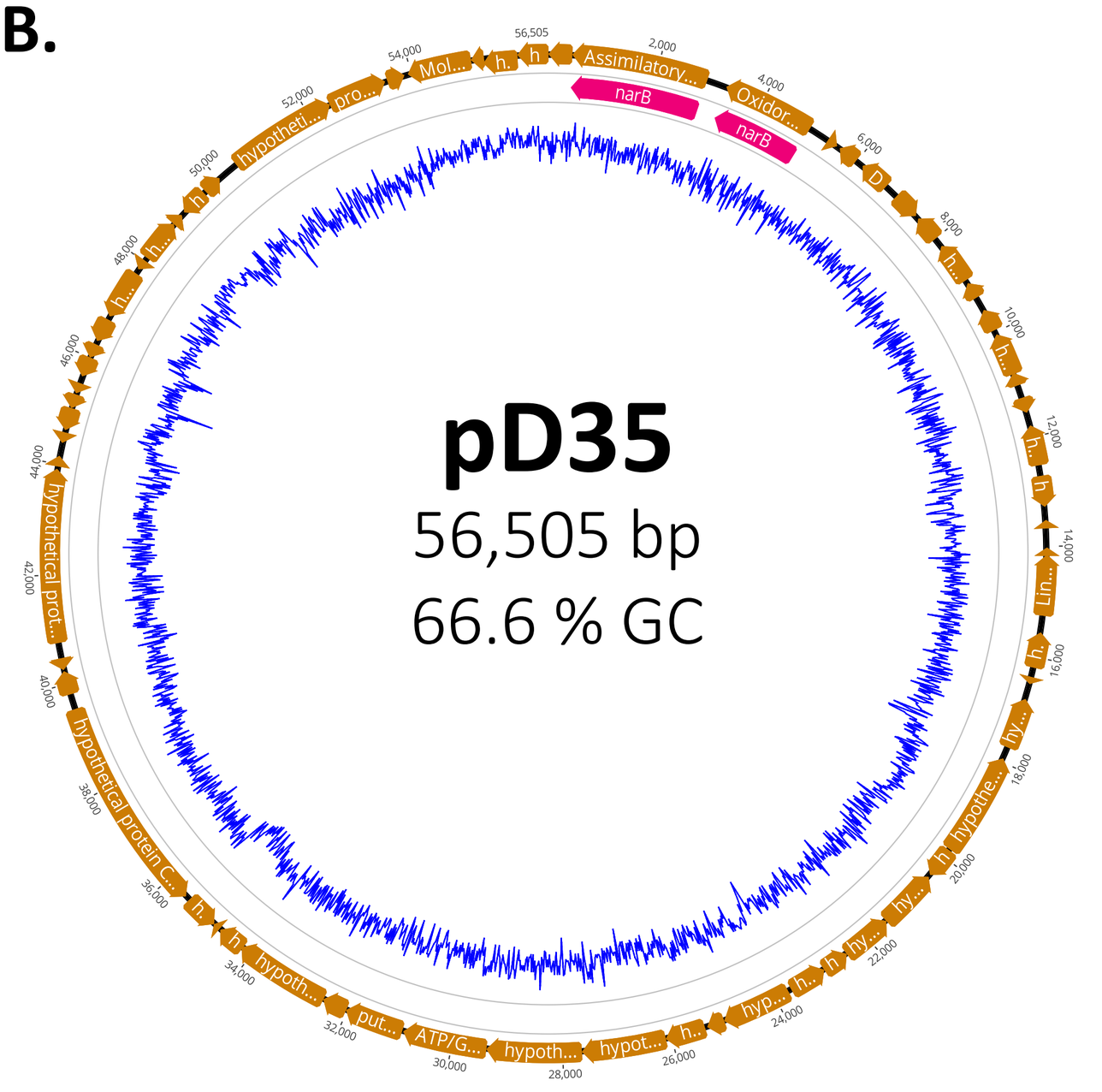
Ecotype

- I (17)
- II (3)
- III (1)
- V (4)
- Other (1)

Plasmid Features

- Conjugative, cell-to-cell DNA transfer proteins
- Chromosome/plasmid partitioning proteins
- Histone-like proteins
- CAZymes


A.
pG07
82,765 bp
68.8 % GC
GH2
GH36
GH35
GH95
GT4
MOBF
hypothetical pr...
beta-galac...
alpha-...
Putative ...
hypot...
hypo...


B.
pD35
56,505 bp
66.6 % GC
narB
narB
Assimilatory...
Oxidor...
hypothetical prot...
hypothetical protein C...
hypoth...
ATP/G...
hypot...
hypothe...
Lin...

SUPPLEMENTAL METHODS

Culture collection and reference genomes. We long-read sequenced 23 *Curtobacterium* strains from our culture collection that were obtained from senescent plant litter (the top 0-5 cm of soil) along an elevation gradient in Southern California. The strains were stored in 25% *v/v* glycerol at -80 ºC and had been previously sequenced on an Illumina platform [1–3]. In addition, we retrieved 14 complete plasmid sequences (and associated host chromosomes) representing diverse *Curtobacterium spp*. hosts that were deposited in NCBI GenBank and RefSeq databases on March 31, 2022. The search criteria we used included: '*Curtobacterium'* and 'Plasmid' or 'Chromosome'. In total, we include 39 *Curtobacterium* genomes in our analyses (**Table S1**). Notably, several attempts were made to isolate plasmids from several strains in our culture collection using Qiagen® Plasmid Maxi Kit (Qiagen, Hilden, Germany) and ZymoPURE™ II Plasmid Midiprep Kit (Zymo Research, Irvine, CA, USA), and custom protocols designed to extract plasmid DNA from gram-positive taxa. However, these approaches missed many of the plasmids that the long read sequencing revealed, presumably because larger plasmids and those with low copy numbers can evade detection with traditional approaches.

DNA preparation and sequencing. Cultures were revived from glycerol stocks, and DNA extractions performed using the Qiagen Blood and Cell Culture DNA Mini Kit (Qiagen, Hilden, Germany). The DNA extraction generated high molecular weight gDNA (> 60 ng), free of small DNA contamination, which was suited for sequencing on Oxford Nanopore Technologies (ONT) platform. DNA quality was assessed via Nanodrop (Thermo Fisher; Massachusetts, USA) and quantified by Qubit (BioTek; Vermont, USA). ONT sequencing libraries were prepared with Ligation Sequencing Kit V14 with Native barcoding (Oxford Nanopore Technologies; Oxford, UK), multiplexed, and run on three different MinION devices

with R9.4.1 flow cells by The SeqCenter Team, formerly known as The Microbial Genome Sequencing Center (Pennsylvania, USA), generating 300 Mbp per isolate. Basecalls of the raw nanopore reads were performed using Guppy v5.0.16.

Sequence assemblies. *De novo* 'hybrid' assemblies of ONT and Illumina sequenced *Curtobacterium* strains were performed with quality checked short and long reads using the default settings of Unicycler version 0.4.8 [4]. Prior to assembly, quality checks for both ONT and Illumina sequencing data were checked using FastQC version 0.11.9 and reports compiled using MultiQC version 1.9 [5]. For ONT generated reads, low quality (PRED < 8), adaptor, and chimeric sequences were removed using Porechop version 0.2.4 along with sequences < 2 kbp in length per previously described methods [6]. For Illumina generated reads, low quality (PHRED < 30), adapter, and PhiX sequences were removed using FastP version 0.20.0 [7]. The read quality for both ONT and Illumina quality filtered reads were reassessed with FastQC and MultiQC. A 'hybrid' assembly (combining long and short read sequencing data) approach was used to obtain complete replicon assemblies, as many long reads can exceed the length of repeats in bacterial genomes, which are also a characteristic of many types of MGE, and short reads can improve accuracy of detecting plasmids in WGS data [6]. Notably, for the Scrubland-52 (W52) and Pine-Oak-43 (P43) genomes, these hybrid assemblies failed, and long-read only assemblies using Trycycler v0.5.3 was performed along with a final polishing step using Medaka version 1.6.0 [8]. All assembly graphs were assessed using Bandage version 0.8.1 [9], and completeness of genome assemblies (e.g., contiguity, N50, and %GC) determined using the web interface of Quast [10].

Phylogenomic analysis. To determine the similarity of NCBI retrieved plasmid and chromosomes sequences to previously described ecotypes (genetic clades with similar

phenotypes that are adapted to local environmental conditions including temperature and moisture) of *Curtobacterium* [1] from our culture collection, reference sequences were imported into Anvi'o version 7.0 [11]. First, 916 single-copy core genes within chromosomes sequences were identified, concatenated, and nucleotide positions that were gap characters in more than 50% of the sequences removed using trimAl version 1.4.1. Next, IQ-TREE [12, 13] with the 'WAG' [14] general matrix model was used to construct a maximum likelihood tree, which was visualized using iTOL version 5 [15]. Except for three strains (AA3, BH2-1-1 and W02), the *Curtobacterium* strains in this study fell within five previously described ecotypes (based on clade designations).

Putative plasmids were identified as closed, circular sequences that were distinct from the chromosome (those having similar percent GC content to known *Curtobacterium* plasmids). No genes were conserved across all plasmid sequences, and the nucleotide lengths of putative plasmids varied significantly. Therefore, pairwise estimates of plasmid similarities were calculated using Mash version 2.3 [16, 17]. The parameters for calculating mash distances were as follows: K-mer = 21 and minimum-hashes per sketch = 1000 (**Table S2** and **S3**). This comparison method was chosen because it allows for the similarity of the original sequences to be rapidly estimated with a bounded error. It depends only on the size of the sketch (i.e., the mash similarities are independent of the genome sizes) and is strongly correlated with ANI [16]. Mash distances for chromosomal sequences were also calculated using the same approach as for plasmids (**Table S2**). To evaluate whether putative plasmids of *Curtobacterium* grouped into known plasmid taxonomic units (380 PTUs constructed from 9,894 plasmid sequences from a curated reference database - RefSeq84), the web version of COPLA was used [18]. To investigate whether any of the plasmids shared conserved backbone region as is common with

other types of plasmids [19], whole genome alignments were performed using Mauve v1.1.3 [20] with a seed weight set to 15 and minimum LCB score of 30,000.

Trait analyses. To determine the trait content of chromosomes and plasmids, gene calls were made in Anvi'o using Prodigal version 2.6.3 [21] and searched against the COG20 (Clusters of Orthologous Groups of genes/proteins) [22] and Pfam version 33.1 [23] databases via DIAMOND v0.9.14 [24] in sensitive mode (**Tables S4** and **S5**). Putative plasmid replicases (used in plasmid replicon typing/incompatibility grouping) were identified from hits to the Pfam databases (**Table S6**). Clustering analysis of plasmid and chromosome amino acid sequence similarities were performed in Anvi'o using the MCL algorithm [25], under the following parameters: exclude partial gene calls, minimum gene cluster occurrence = 1, and default settings for minbit heuristic and MCL inflation parameter. Gene clusters for plasmid and chromosome replicons, visualized via the anvio-display-pan feature of the interactive interface. All COG functions, Pfam hits, and corresponding gene calls were exported as tables from Anvi'o and merged into one data table before importing into R version 4.2.2 [26] for statistical analysis. To determine the potential for plasmids to be mobilizable, sequences were searched for MOB family relaxases, enzymes essential for conjugative DNA processing [27] using MobScan (**Table S7)** [28].

Additionally, chromosome and plasmid sequences were analyzed for genes involved in carbohydrate and nitrogen utilization. To identify carbohydrate active enzymes (CAZymes), we used run_dbcan 4.0.0 and dbCAN2 databases released in 2022 [29]. Query matches were included if two or more of the three search tools (HMMER, DIAMOND, Hotpep) identified the same CAZyme family annotation per the developer's recommendation [29]. Query results were included in analyses for HMMER searches of dbCAN and dbCAN-sub with E-values < 1e-15

and coverage > 0.35; and for DIAMOND searches of the CAZy database with E-value < 1e-102 (**Tables S8** and **S9**). To identify genes associated with nitrogen-cycling pathways, BLASTp searches of queries against a curated database of nitrogen (N) gene families, the NCycDB release 2019 [30] at 100% sequence identity were performed and gene calls having E-values 10^{-5} and > 50 % query coverages were included in the analyses (**Tables S10)**.

Statistical analysis. To determine whether the pairwise similarities for plasmid and chromosome sequences varied by ecotype and/or environment type, similarity matrices for each sequence type were tested separately via permutational multivariate analysis of variance (PERMANOVA; permutations $n = 999$ with unrestricted permutations of raw data using type III sums of squares) in PRIMER-e version 6 [31, 32] with ecotype and/or environment designated as fixed factors. Distance-based tests for homogeneity of multivariate dispersions were also performed using the PERMDISP function in PRIMER-e, grouping by either ecotype or environment. To account for sampling biases for rare ecotypes (i.e., *Curtobacterium* chromosomes outside ecotype/clade I or V; **Table S1**) and environments (i.e., *Curtobacterium* isolated from algae or unknown origins; **Table S1**), the number of plasmids/chromosomes by category were grouped together into an 'Other' category. The estimated variance explained was determined by dividing terms with significant p-values plus the residual variation by the sum of the estimates of components of variation given as output from PRIMER-e. To test whether plasmids and chromosome genetic similarities varied similarly by ecotype and environment, a RELATE test [32] using Spearman correlation was performed in PRIMER-e.

To determine whether the COG and CAZyme composition of plasmid and chromosomes varied by ecotype and/or environment type, euclidean distances were calculated from COG and CAZyme counts using the *vegdist* function of the 'vegan' package in R [33], and PERMANOVA

and RELATE tests performed as previously mentioned. Heatmaps for plasmid pairwise similarities and COG traits were passed to the *heatmap.2* function of the 'gplots' package (https://github.com/talgalili/gplots) in R for visualization. Plasmid sequences and alignments were visualized and annotated in Geneious Prime® 2023.2.1, Build 2023-07-20 (https://www.geneious.com). Additional, G-tests were performed on contingency tables of non-standardize trait counts with rare traits (traits counts < 6 across all environments) removed to confirm trends were not stochastic attributes of these sequences.

REFERENCES

1. Chase AB, Gomez-Lunar Z, Lopez AE, Li J, Allison SD, Martiny AC, et al. Emergence of soil bacterial ecotypes along a climate gradient. *Environ Microbiol* 2018; **20**: 4112–4126.
2. Glassman SI, Weihe C, Li J, Albright MBN, Looby CI, Martiny AC, et al. Decomposition responses to climate depend on microbial community composition. *Proceedings of the National Academy of Sciences* 2018; **115**: 11994 LP – 11999.
3. Glassman SI, Martiny JBH. Broadscale Ecological Patterns Are Robust to Use of Exact Sequence Variants versus Operational Taxonomic Units. *mSphere* 2018; **3**: e00148-18.
4. Wick RR, Judd LM, Gorrie CL, Holt KE. Unicycler: Resolving bacterial genome assemblies from short and long sequencing reads. *PLoS Comput Biol* 2017; **13**: e1005595.
5. Ewels P, Magnusson M, Lundin S, Käller M. MultiQC: summarize analysis results for multiple tools and samples in a single report. *Bioinformatics* 2016; **32**: 3047–3048.
6. Wick RR, Judd LM, Gorrie CL, Holt KE. Completing bacterial genome assemblies with multiplex MinION sequencing. *Microb Genom* 2017; **3**: e000132.
7. Chen S, Zhou Y, Chen Y, Gu J. fastp: an ultra-fast all-in-one FASTQ preprocessor. *Bioinformatics* 2018; **34**: i884–i890.
8. Wick RR, Judd LM, Cerdeira LT, Hawkey J, Méric G, Vezina B, et al. Trycycler: consensus long-read assemblies for bacterial genomes. *Genome Biol* 2021; **22**: 1–17.
9. Wick RR, Schultz MB, Zobel J, Holt KE. Bandage: interactive visualization of de novo genome assemblies. *Bioinformatics* 2015; **31**: 3350–3352.
10. Gurevich A, Saveliev V, Vyahhi N, Tesler G. QUAST: quality assessment tool for genome assemblies. *Bioinformatics* 2013; **29**: 1072–1075.
11. Eren AM, Kiefl E, Shaiber A, Veseli I, Miller SE, Schechter MS, et al. Community-led, integrated, reproducible multi-omics with anvi'o. *Nat Microbiol* 2021; **6**: 3–6.
12. Nguyen L-T, Schmidt HA, von Haeseler A, Minh BQ. IQ-TREE: A Fast and Effective Stochastic Algorithm for Estimating Maximum-Likelihood Phylogenies. *Mol Biol Evol* 2015; **32**: 268–274.
13. Minh BQ, Schmidt HA, Chernomor O, Schrempf D, Woodhams MD, von Haeseler A, et al. IQ-TREE 2: New Models and Efficient Methods for Phylogenetic Inference in the Genomic Era. *Mol Biol Evol* 2020; **37**: 1530–1534.
14. Whelan S, Goldman N. A General Empirical Model of Protein Evolution Derived from Multiple Protein Families Using a Maximum-Likelihood Approach. *Mol Biol Evol* 2001; **18**: 691–699.
15. Letunic I, Bork P. Interactive Tree Of Life (iTOL) v5: an online tool for phylogenetic tree display and annotation. *Nucleic Acids Res* 2021; **49**: W293–W296.
16. Ondov BD, Treangen TJ, Melsted P, Mallonee AB, Bergman NH, Koren S, et al. Mash: Fast genome and metagenome distance estimation using MinHash. *Genome Biol* 2016; **17**: 1–14.
17. Ondov BD, Starrett GJ, Sappington A, Kostic A, Koren S, Buck CB, et al. Mash Screen: High-throughput sequence containment estimation for genome discovery. *Genome Biol* 2019; **20**: 1–13.
18. Redondo-Salvo S, Bartomeus-Peñalver R, Vielva L, Tagg KA, Webb HE, Fernández-López R, et al. COPLA, a taxonomic classifier of plasmids. *BMC Bioinformatics* 2021; **22**: 1–9.

19. Fernandez-Lopez R, de Toro M, Moncalian G, Garcillan-Barcia MP, de la Cruz F. Comparative genomics of the conjugation region of F-like plasmids: Five shades of F. *Front Mol Biosci* 2016; **3**: 71.
20. Darling ACE, Mau B, Blattner FR, Perna NT. Mauve: Multiple Alignment of Conserved Genomic Sequence With Rearrangements. *Genome Res* 2004; **14**: 1394–1403.
21. Hyatt D, Chen G-L, LoCascio PF, Land ML, Larimer FW, Hauser LJ. Prodigal: prokaryotic gene recognition and translation initiation site identification. *BMC Bioinformatics* 2010; **11**: 119.
22. Galperin MY, Wolf YI, Makarova KS, Vera Alvarez R, Landsman D, Koonin E V. COG database update: focus on microbial diversity, model organisms, and widespread pathogens. *Nucleic Acids Res* 2021; **49**: D274–D281.
23. Mistry J, Chuguransky S, Williams L, Qureshi M, Salazar GA, Sonnhammer ELL, et al. Pfam: The protein families database in 2021. *Nucleic Acids Res* 2021; **49**: D412–D419.
24. Buchfink B, Xie C, Huson DH. Fast and sensitive protein alignment using DIAMOND. *Nat Methods* 2015; **12**: 59–60.
25. van Dongen S, Abreu-Goodger C. Using MCL to Extract Clusters from Networks BT - Bacterial Molecular Networks: Methods and Protocols. In: van Helden J, Toussaint A, Thieffry D (eds).2012. Springer New York, New York, NY, pp 281–295.
26. R Core Team (2020). R: A language and environment for statistical computing. *R: A language and environment for statistical computing R Foundation for Statistical Computing, Vienna, Austria* . 2020.
27. Garcillán-Barcia MP, Francia MV, de La Cruz F. The diversity of conjugative relaxases and its application in plasmid classification. *FEMS Microbiol Rev* 2009; **33**: 657–687.
28. Garcillán-Barcia MP, Redondo-Salvo S, Vielva L, de la Cruz F. MOBscan: Automated Annotation of MOB Relaxases. *Methods Mol Biol* 2020; **2075**: 295–308.
29. Zheng J, Ge Q, Yan Y, Zhang X, Huang L, Yin Y. dbCAN3: automated carbohydrate-active enzyme and substrate annotation. *Nucleic Acids Res* 2023; **51**: W115–W121.
30. Tu Q, Lin L, Cheng L, Deng Y, He Z. NCycDB: a curated integrative database for fast and accurate metagenomic profiling of nitrogen cycling genes. *Bioinformatics* 2019; **35**: 1040–1048.
31. Clarke KR, Gorley RN. PRIMER v6: *Primer V6: User Manual/Tutorial* . 2006.
32. Anderson MJ, Gorley RN, Clarke KR. PERMANOVA+ for PRIMER: Guide to Software and Statistical Methods. *Plymouth, UK*. 2008.
33. Oksanen J, Blanchet FG, Friendly M, Kindt R, Legendre P, McGlinn D, et al. vegan: Community Ecology Package. R package version 2.5-2. *Cran R* 2019.