



RefSoil+: a Reference Database for Genes and Traits of Soil Plasmids

Taylor K. Dunivin,^{a,b}  Jinlyung Choi,^c Adina Howe,^c  Ashley Shade^{a,d,e,f}

^aDepartment of Microbiology and Molecular Genetics, Michigan State University, East Lansing, Michigan, USA

^bEnvironmental and Integrative Toxicological Sciences, Michigan State University, East Lansing, Michigan, USA

^cDepartment of Agricultural and Biosystems Engineering, Iowa State University, Ames, Iowa, USA

^dDepartment of Plant, Soil and Microbial Sciences, Michigan State University, East Lansing, Michigan, USA

^eProgram in Ecology, Evolutionary Biology and Behavior, Michigan State University, East Lansing, Michigan, USA

^fPlant Resilience Institute, Michigan State University, East Lansing, Michigan, USA

ABSTRACT Plasmids harbor transferable genes that contribute to the functional repertoire of microbial communities, yet their contributions to metagenomes are often overlooked. Environmental plasmids have the potential to spread antibiotic resistance to clinical microbial strains. In soils, high microbiome diversity and high variability in plasmid characteristics present a challenge for studying plasmids. To improve the understanding of soil plasmids, we present RefSoil+, a database containing plasmid sequences from 922 soil microorganisms. Soil plasmids were larger than other described plasmids, which is a trait associated with plasmid mobility. There was a weak relationship between chromosome size and plasmid size and no relationship between chromosome size and plasmid number, suggesting that these genomic traits are independent in soil. We used RefSoil+ to inform the distributions of antibiotic resistance genes among soil microorganisms compared to those among nonsoil microorganisms. Soil-associated plasmids, but not chromosomes, had fewer antibiotic resistance genes than other microorganisms. These data suggest that soils may offer limited opportunity for plasmid-mediated transfer of described antibiotic resistance genes. RefSoil+ can serve as a reference for the diversity, composition, and host associations of plasmid-borne functional genes in soil, a utility that will be enhanced as the database expands. Our study improves the understanding of soil plasmids and provides a resource for assessing the dynamics of the genes that they carry, especially genes conferring antibiotic resistances.

IMPORTANCE Soil-associated plasmids have the potential to transfer antibiotic resistance genes from environmental to clinical microbial strains, which is a public health concern. A specific resource is needed to aggregate the knowledge of soil plasmid characteristics so that the content, host associations, and dynamics of antibiotic resistance genes can be assessed and then tracked between the environment and the clinic. Here, we present RefSoil+, a database of soil-associated plasmids. RefSoil+ presents a contemporary snapshot of antibiotic resistance genes in soil that can serve as a reference as novel plasmids and transferred antibiotic resistances are discovered. Our study broadens our understanding of plasmids in soil and provides a community resource of important plasmid-associated genes, including antibiotic resistance genes.


KEYWORDS antibiotic resistance genes, database, isolates, metagenomics, microbial ecology, microbiome dynamics, one health, plasmid, soil microbiology, targeted gene assembly

Citation Dunivin TK, Choi J, Howe A, Shade A. 2019. RefSoil+: a reference database for genes and traits of soil plasmids. *mSystems* 4:e00349-18. <https://doi.org/10.1128/mSystems.00349-18>.

Editor Thomas J. Sharpton, Oregon State University

Copyright © 2019 Dunivin et al. This is an open-access article distributed under the terms of the [Creative Commons Attribution 4.0 International license](https://creativecommons.org/licenses/by/4.0/).

Address correspondence to Ashley Shade, shadeash@msu.edu.

 Diversity and genomic locations of antibiotic resistance genes for soil organisms provide an important baseline to understand their potential for transfer.

Received 3 January 2019

Accepted 29 January 2019

Published 26 February 2019

Soil is a unique and ancient environment that harbors immense microbial biodiversity. The soil microbiome has functional consequences for ecosystems, such as supporting plant growth (1, 2) and mediating key biogeochemical transformations (3). It also serves as a reservoir of microbial functional genes of interest to human and animal welfare. Within microbial genomes, important functions can be encoded on both chromosomes and extrachromosomal mobile genetic elements such as plasmids. Plasmids can be laterally transferred among community members, both among and between phyla (4–6). This causes a propagation of plasmid functional genes and allows them to spread among divergent host strains. Within microbial communities, plasmids influence microbial diversification (7) and contribute to functional gene pools (4). Plasmids can alter the fitness of individuals in a community as they can be gained or lost in the environment, which alters their functional gene content and can have consequences for their local competitiveness.

Antibiotic resistance genes (ARGs) provide a prime example of the importance that functional genes encoded on plasmids can have. ARGs can undergo plasmid-mediated horizontal gene transfer (HGT) (8, 9). There is particular concern about the potential for spread of ARGs between environmental and clinically relevant bacterial strains. Studies of ARGs in soil have shown overlap between environmental and clinical strains that suggests HGT (10–12). For example, plasmid-encoded quinolone resistance (*qnrA*) in clinical *Enterobacteriaceae* strains likely originated from the environmental strain *Shewanella algae* (11). The extent of the impact of environmental reservoirs of ARGs is unknown (13), but studies have shown evidence for predominantly vertical, rather than horizontal, transfer of these genes (14). Additionally, it is speculated that rates of transfer in bulk soil are low compared to that in environments with higher population densities, such as the rhizosphere, phyllosphere, and gut microbiomes of soil microorganisms (15). In the case of antibiotic resistance, mobilization is a public health risk. Broadly, the ability of plasmids to rapidly move genes both between and among memberships is linked to diversification in complex systems, especially soils (7).

Despite their ecological and functional relevance, plasmids are not well characterized in soil. Plasmids vary in copy number, host range, transfer potential, and genetic makeup (4, 16), making them difficult to assemble and characterize from complex soil metagenomes that contain tens of thousands of bacteria and archaea (17). Plasmid extraction from soil is biased toward smaller plasmids and excludes linear plasmids (4). Additionally, mosaic gene content on plasmids makes their assembly from metagenomes difficult (4). Though new methods for plasmid assembly from metagenomes are being developed (18, 19), the resulting contigs represent a population average of plasmid gene content and size because they are very likely not derived from an individual cell. Thus, the size ranges of plasmids in soils are largely unknown but of consequence, because size is one factor reported to contribute to plasmid potential for transferability (5). Furthermore, “plasmidome” analysis and plasmid assembly from metagenomes do not provide host information. New methods, such as single-cell analysis and proximity ligation of chromosomes to plasmids prior to sequencing (20), are still expected to assemble plasmids with some degree of mosaicism. However, whole genomes sequenced from soil-associated microorganisms, inclusive of both chromosomes and plasmids, could provide plasmid host and size information. A database including this information could also provide information as to the extent functional genes encoded on plasmids overlap with the host cell chromosome(s).

To aid in the study of plasmids and their associated functional genes in soil, we established a resource to compare genetic locations of functional genes in soil microorganisms. We extended the RefSoil database (21) of 922 soil microorganisms to include their plasmids. We used this database to test whether soil-associated plasmids are distinct from plasmids from a broad general database of microorganisms, RefSeq (22). We focused our comparisons on plasmid size and the content, diversity, and location of ARGs on plasmids and chromosomes. We used hidden Markov models from the ResFams database (23) to search for ARGs in the extended soil database, RefSoil+, and RefSeq. RefSoil+ provides insights into the range of plasmid sizes and their

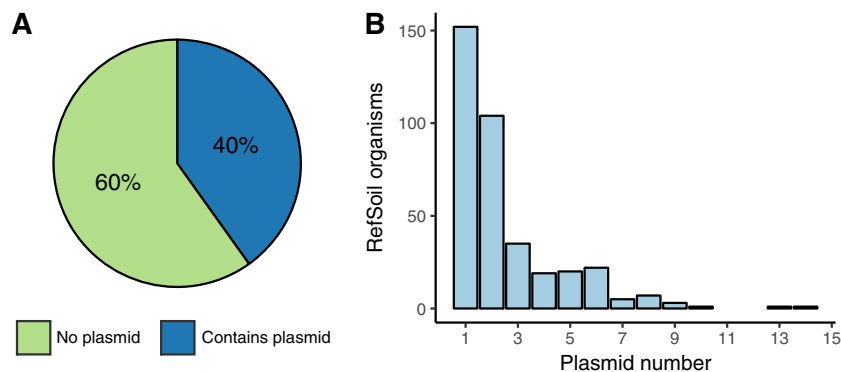


FIG 1 Summary of RefSoil plasmids. (A) Percentages of RefSoil microorganisms with (blue) and without (green) detected plasmids. (B) Distribution of the number of plasmids per RefSoil microorganism.

functional potential within soil microorganisms. RefSoil+ can be used to inform and test hypotheses about the traits, functional gene content, and spread of soil-associated plasmids and can serve as a reference for plasmid assembly from metagenomes.

RESULTS AND DISCUSSION

Plasmid characterization. RefSoil+ is an extension of the RefSoil database inclusive of soil-associated plasmids. RefSoil+ includes taxonomic information, amino acid sequences, coding nucleotide sequences, and GenBank files for a curated set of 922 soil-associated microorganisms. A total of 928 plasmids were associated with RefSoil microorganisms, and 370 RefSoil microorganisms (40.1%) had at least one plasmid (Fig. 1A). This is high compared to the proportion of noneukaryotic plasmids in the general RefSeq database (34%; Mann-Whitney U, $P < 0.01$). The mean number of plasmids per RefSoil organism was 1.01, but the number of plasmids per organism varied greatly (variance, 3.2) (Fig. 1B). For example, strain *Bacillus thuringiensis* serovar *thuringiensis* (RefSoil 738) had 14 plasmids, ranging from 6,880 to 328,151 bp. The mean number of plasmids per RefSoil organism was also greater than for RefSeq (Mann-Whitney U, $P < 0.01$). The abundance of plasmids found in RefSoil genomes highlights plasmids as an important component of soil microbiomes (7, 24).

Soil-associated plasmids tended to be larger than plasmids from other environments (Mann-Whitney U, $P < 0.01$). Plasmid size in RefSoil microorganisms ranged from 1,286 bp to 2.58 Mbp (Fig. 2A), which rivals the range of all known plasmids from various environments (744 bp to 2.58 Mbp) (16). In the distribution of plasmid size, both upper and lower extremes had representatives from soil. Plasmids from all habitats were previously shown to have a characteristic bimodal size distribution with peaks at 5 kb and 35 kb (15–17). In this analysis, the subset RefSeq plasmids had a multimodal distribution (Hartigan's dip test, $P < 0.01$; bimodality coefficient, 0.745) and modes at 3 kb and 59 kb (Fig. 2). Soil-associated plasmids in RefSoil+ also had a multimodal size distribution (Hartigan's dip test, $P < 0.05$; bimodality coefficient, 0.800) but had modes at 1 kb, 3 kb, 49 kb, and 183 kb. Additionally, RefSoil+ plasmids were larger than RefSeq plasmids (Mann-Whitney U, $P < 0.01$) (Fig. 2). Specifically, RefSoil+ proportionally contained more plasmids of >100 kb (Fig. 2B). Thus, while soil-associated plasmids vary in size, they are, on average, large. This is of particular importance because of the established differences in mobility of plasmids in different size ranges (5). Smillie and colleagues showed that mobilizable plasmids, which have relaxases, tend to be larger than nontransmissible plasmids, with median values of 35 and 11 kbp, respectively (5). The majority of soil-associated plasmids (68.2%) were >35 kbp (Fig. 2), suggesting they are more likely to be mobile. Additionally, conjugative plasmids, which encode type IV coupling proteins, have a larger median size (181 kbp) (5). Similarly, RefSoil+ plasmids had a mode of 183 kb (Fig. 2), suggesting that these soil-associated plasmids are more likely to be conjugative. Future works should examine the genetic potential for the transfer of plasmids associated with different ecosystems to test this hypothesis.

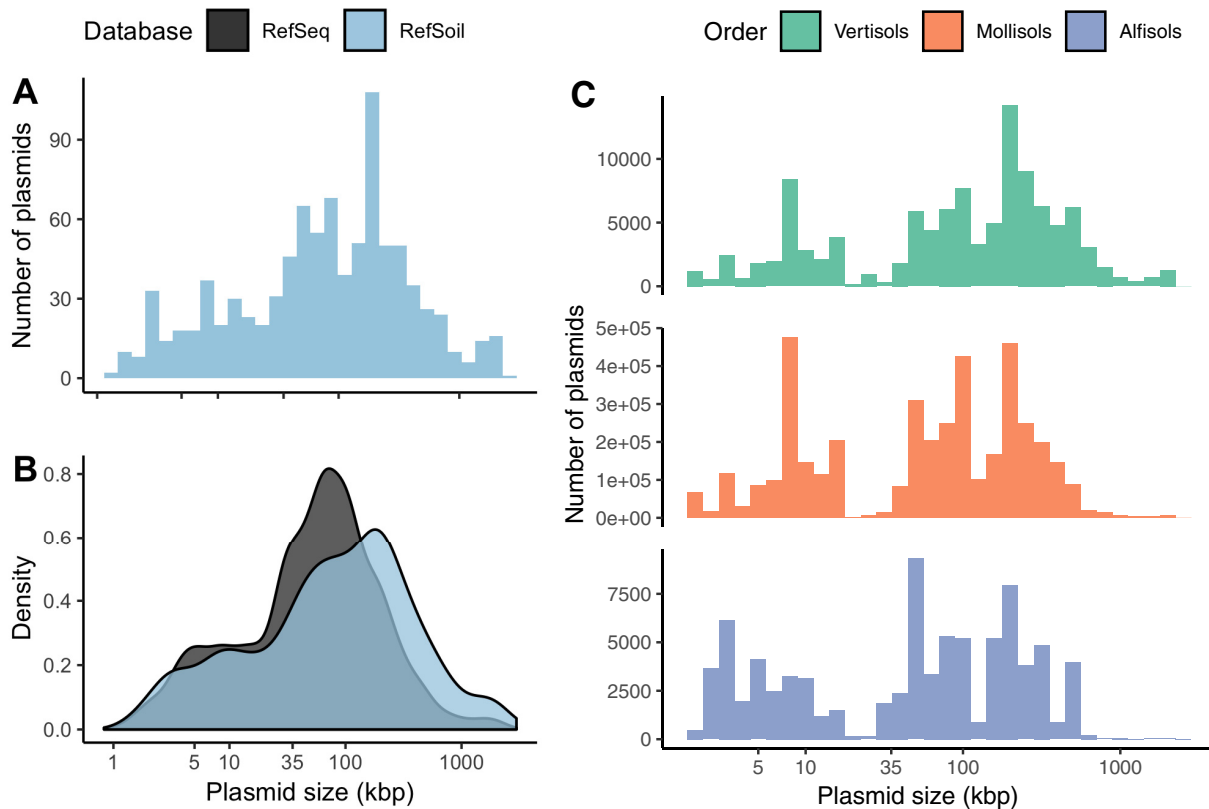


FIG 2 Plasmid size distributions. (A) Histogram of plasmid size (kbp) from RefSoil plasmids. (B) RefSoil (blue) and RefSeq (gray) plasmid size distributions. (C) Estimation of plasmid size distribution in three soil orders. Color indicates soil order and the number indicates the community size.

Plasmid size may vary in the environment. To estimate the environmental size distributions of plasmids, we used estimates of the environmental abundance of RefSoil microorganisms (21). We focused on soil orders previously shown to include the most RefSoil representatives (alfisols, mollisols, and vertisols) (21). We found that plasmid size distributions varied based on soil order (Kruskal-Wallis, $P < 0.01$) (Fig. 2C). True environmental abundance may vary based on plasmid copy number within individuals and plasmids from uncultivated microorganisms, but this estimation gives a rough idea of plasmid size distributions in the environment and provides some baseline information because there are methodological challenges to accurately measuring plasmid size *in situ* (4, 18, 19).

Genome size, inclusive of chromosomes and plasmids, is an important ecological trait that is difficult to estimate from metagenomes (25). Due to incomplete assemblies, genome size must be approximated based on the estimated number of individuals through single-copy gene abundance (26). Extrachromosomal elements, however, inflate these estimated genome sizes, because they contribute to the sequence information of the metagenome often without contributing single-copy genes (27). While our methodologies do not account for plasmid copy number (28), we examined the relationship between genome size and plasmid size in soil-associated microorganisms and found a weak but significant correlation (Spearman's $\rho = 0.12$; $P < 0.001$) (Fig. 3). Additionally, chromosome size was not predictive of the number of plasmids (Fig. 3; see also Fig. S1 in the supplemental material). For example, *Bacillus thuringiensis* serovar *thuringiensis* strain IS5056 had the most plasmids in RefSoil+, but these plasmids spanned the size range of 6.8 to 328 kbp. This strain's plasmids make up 19% of its coding sequences (29), but its chromosome (5.4 Mbp) is average for soils (27). Despite the weak relationship between genome size and plasmid characteristics within these

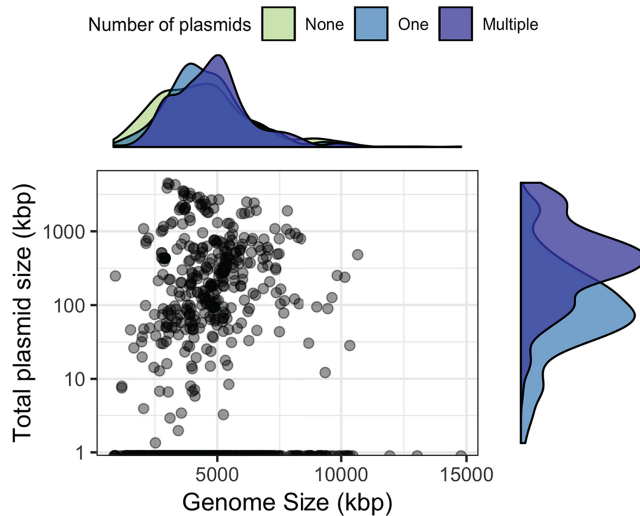


FIG 3 Relationship between plasmid size and genome size. Total plasmid size (sum of all plasmids in a microorganism; kbp) is plotted on a log scale against total genome size for each RefSoil microorganism. Density plots are included for each axis to represent the distribution of RefSoil microorganisms with different numbers of plasmids (none, green; one, blue; or multiple, purple).

data, the plasmid database can be used to inform estimates of average genome sizes from close relatives detected within metagenomes.

ARGs on soil plasmids. It is unclear whether soil ARGs are predominantly on chromosomes or mobile genetic elements. While mobile gene pools are not static, there is evidence to suggest low transfer of ARGs in soil (14, 15, 30). For example, bulk soils are not a “hot spot” for HGT because they are often resource-limited (31), and surveys of ARGs in soil metagenomes have suggested a predominance of vertical transfer, rather than horizontal transfer, of ARGs (14, 30). Using RefSoil+ sequences and ResFams hidden Markov models (HMMs) (23), we examined 174 genes encoding resistance to beta-lactams, tetracyclines, aminoglycosides, chloramphenicol, glycopeptides, macrolides, quinolones, and trimethoprim. After quality filtering, we detected 154,392 ARG sequences in RefSoil chromosomes and plasmids (Fig. 4; see also Table S1).

Adding plasmids to the RefSoil database increased the number of functional gene types, or genes that have functional potential (32), represented in the database, as 7 ARGs (16S rRNA methyltransferase, AAC6-Ib, ANT6, CTXM, ErmC, KPC, and TetD) were only detected on plasmids. Notably, these functional genes would be missed if only chromosomes were considered. However, the majority of ARGs were chromosomally encoded in RefSoil+ microorganisms (Fig. 4A and B) (chromosome versus plasmid; Mann Whitney U, $P < 0.01$). We next examined the genomic distributions of ARGs in RefSoil+ based on taxonomy (Fig. 4C and D). Proteobacteria had the most plasmid-associated ARGs, which has been reported previously (33).

We were curious whether ARGs were more commonly detected on chromosomes than plasmids in general or if this trend was specific to soil microorganisms. We found that the number of ARGs per genome was comparable for RefSoil and RefSeq (Mann Whitney U, $P > 0.05$), but RefSoil plasmids had fewer ARGs than RefSeq plasmids (Mann Whitney U, $P < 0.05$) (Fig. 5). Normalizing to individual microorganisms is biased toward chromosomes, however, because chromosomes typically have more base pairs than plasmids. To account for this, we also normalized ARGs to base pairs, and there were more ARGs in plasmids from both databases than in chromosomes (Mann Whitney U, $P < 0.05$). Notably, RefSoil+ had fewer ARGs than RefSeq (Mann Whitney U, $P < 0.01$) (Fig. S3). This suggests that plasmid-mediated HGT rates of ARGs may be relatively low in these soil microorganisms. We note that the RefSoil database is limited in representatives of *Verrucomicrobia* and *Acidobacteria*, which may change these estimates (21); however, this will improve as the database grows.

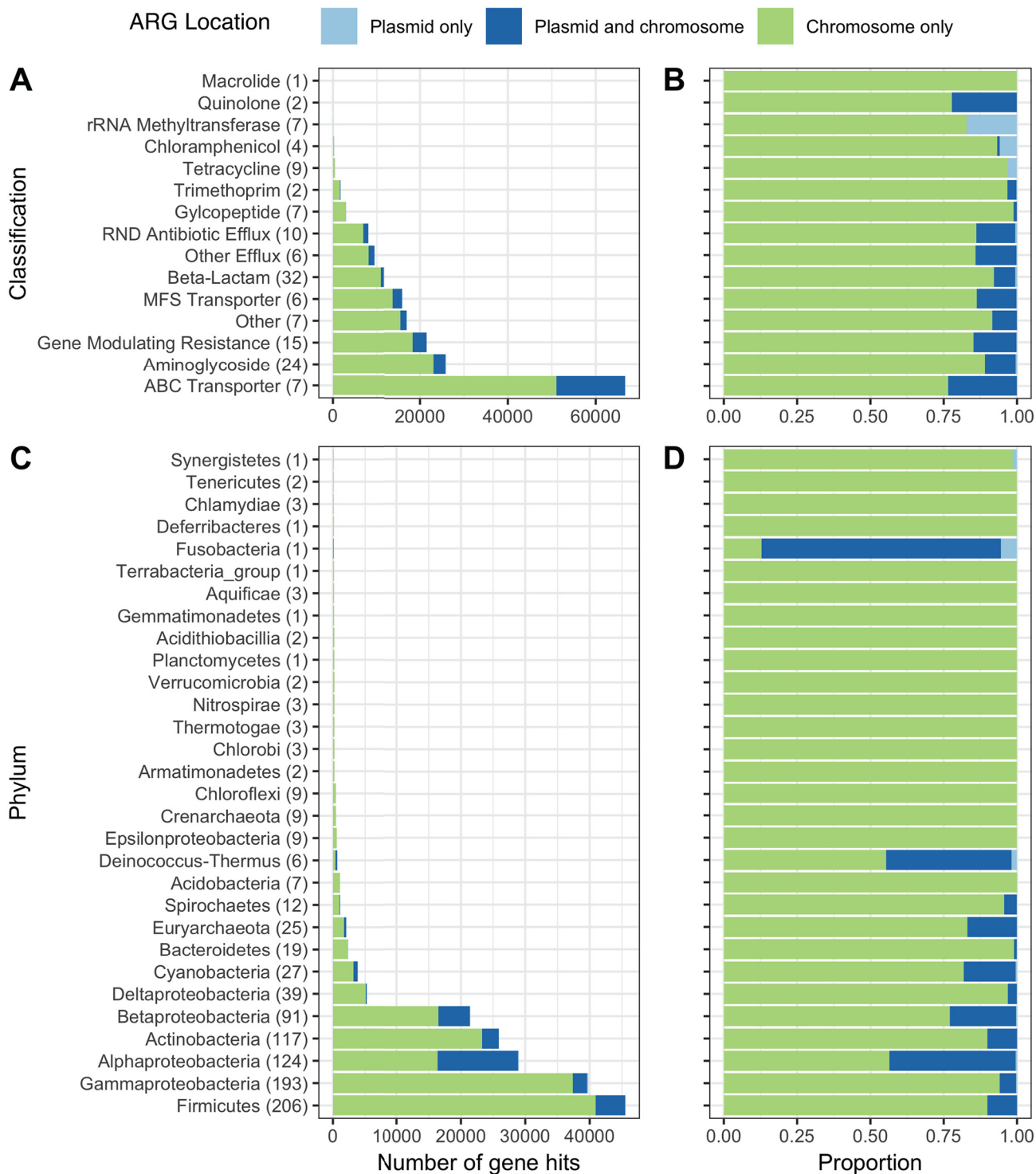


FIG 4 Distribution of ARGs in RefSoil genomes and plasmids. The raw numbers (A) and proportions (B) of ARGs on plasmids (light blue), genomes (green), or both (dark blue) in RefSoil+ microorganisms by antibiotic resistance gene group. The numbers of genes included in each group are shown in parentheses. The raw numbers (C) and proportions (D) of detected ARGs on plasmids (light blue), genomes (green), or both (dark blue) in RefSoil+ microorganisms by phylum-level taxonomy. The numbers of taxa included in each phylum are shown in parentheses.

We examined this trend for each antibiotic class and observed a greater proportion of ARG sequences on plasmids in RefSeq than in RefSoil+ for genes encoding glycopeptide and tetracycline resistance (see Fig. S2). Gibson and colleagues also found a lack of tetracycline resistance genes in soil-associated isolates compared to that in

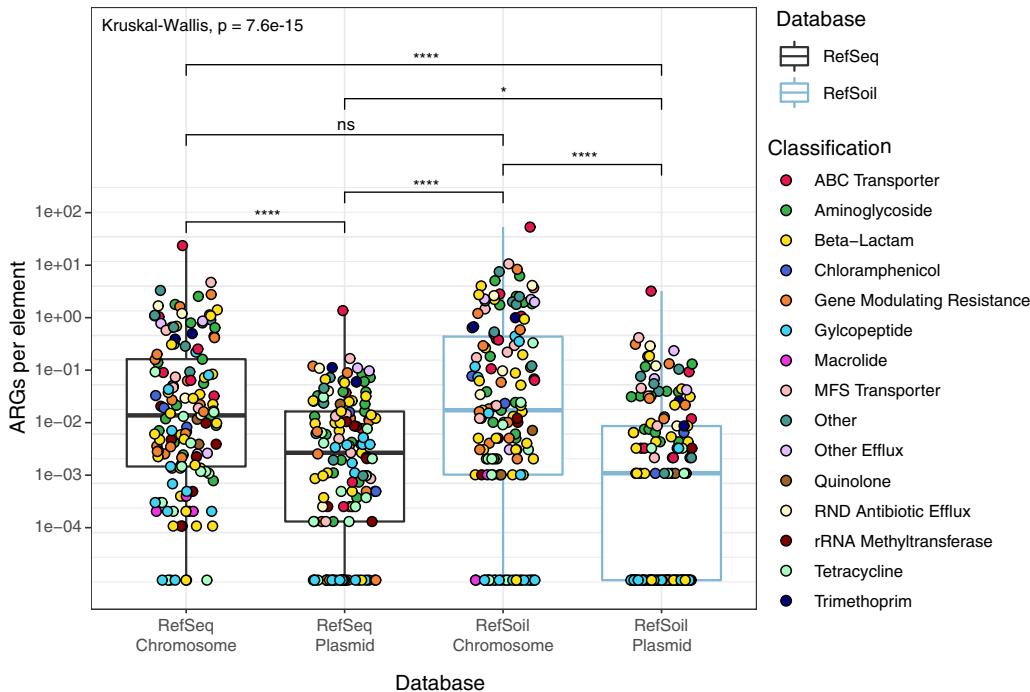


FIG 5 Proportion of ARGs on genomes and plasmids in RefSeq+ and RefSeq databases. Numbers of ARGs were normalized to numbers of genetic elements. Boxplots are colored by database. Points represent individual ARGs and are colored based on classification. Kruskal-Wallis test results are shown in addition to significant results from pairwise Mann-Whitney U tests (****, $P \leq 0.0001$; *, $P \leq 0.05$; ns, $P > 0.05$).

water- and human-associated strains (23). By determining whether ARGs were encoded on plasmids or chromosomes, our analysis suggests that these patterns were due to chromosomal genes and more likely vertically transferred (Fig. 5). Thus, these soil bacteria harbor relatively few ARGs on plasmids, suggesting that RefSeq+ microorganisms have limited capacity for plasmid-mediated transfer of these genes. Future assessments of functional gene content on chromosomes and plasmids together will help to delineate changes in transfer potential and reveal selective or environmental factors that impact transfer potential.

While genome data from isolates cannot inform on the environmental abundance of ARGs, our data support observations of ARGs in mobile genetic elements in soil from cultivation-independent studies as well. Luo and colleagues observed a low abundance of chloramphenicol, quinolone, and tetracycline resistance genes in soil mobile genetic elements (24), and Xiong and colleagues (34) also observed low abundance of *qnr* genes. Similarly, we observed fewer plasmid-encoded tetracycline resistance genes in soil-associated microorganisms than in RefSeq microorganisms (Fig. S2). We did not observe significant differences for genes encoding quinolone or chloramphenicol resistance; however, these had small sample sizes ($n = 2$ and 3 , respectively). Mobile genetic elements in soil have also been shown to have an abundance of genes encoding multidrug efflux pumps and resistance to beta-lactams, aminoglycosides, and glycopeptides (24). Genes encoding beta-lactam and aminoglycoside resistance were comparable between RefSeq+ and RefSeq (Kruskal-Wallis, $P > 0.05$) (Fig. S2). However, plasmid-borne glycopeptide resistance genes were less common in RefSeq+ plasmids (Mann-Whitney U, $P < 0.05$).

RefSeq+ applications. RefSeq+ is publicly available on GitHub (https://github.com/ShadeLab/RefSeq_plasmids). It includes an excel file linking RefSeq+ organism taxonomy with accession numbers for corresponding chromosomes and plasmids. It also contains several fasta files with coding DNA sequence (CDS) and amino acid sequences. These files can be downloaded directly from GitHub. RefSeq+ has been

used to better estimate genome sizes in soil (27) and to estimate the distribution of arsenic resistance genes in soil-associated chromosomes and plasmids (35).

Our results show that soil-associated plasmids have distinctive traits and can harbor functional genes that are not encoded on host chromosomes. RefSoil+ expands the knowledge of functional genes with potential for transfer among soil microorganisms and offers insights into plasmid size and host ranges in soil (and improves the accuracy of estimates of their genome sizes).

Because it is populated by the chromosomes and plasmids of isolates, RefSoil+ links host taxonomy to plasmid content. This linkage is important especially for heterogeneous ecosystems with high microbial richness, such as soils, which rely heavily on cultivation-independent methods for observing microbially diverse populations. RefSoil+ can guide the assembly and support the annotation of plasmids from soil metagenomes and also direct hypotheses of host identity (18, 36). Notably, plasmid gene content is not static (37), and individuals can gain or lose plasmids (38, 39). Despite this, historical data of the genetic makeup and host range of plasmids can be used to better understand plasmid ecology, and to serve as an important reference to understand by how much host plasmid numbers and contents change in the future. This information contributes to information needed to understand patterns of plasmid dissemination, both across environments and among hosts.

RefSoil+ can be used as a reference database or as a database for primer design to target plasmids in the environment. Advances microbiome sequencing methods such as presequencing proximity linkage (e.g., Hi-C [20]), long-read technology (40), or single cell sequencing (41) could add to and leverage RefSoil+ to improve the characterization of plasmid-host relationships in soil. As movements of ARGs are observed in the clinic and the environment, RefSoil+ can also serve as a reference for comparison with legacy plasmid and chromosome contents and distributions. Novel genomes and plasmids could be added in future RefSoil+ versions, and plasmid-host relationships as well as encoded functions could be compared between cultivation-dependent and -independent methodologies. RefSoil+ provides a rich community resource for research frontiers in plasmid ecology and evolution within wild microbiomes.

MATERIALS AND METHODS

RefSoil plasmid database generation. Accession numbers from RefSoil genomes were used to collect assembly accession numbers for all 922 strains. Assembly accession numbers were then used to obtain a list of all genetic elements from the assembly of one strain. Because all RefSoil microorganisms have completed genomes, all plasmids present at the time of sequencing are included in the assembly. Plasmid accession numbers were compiled for each strain and added to the RefSoil database to make RefSoil+ (see Table S1 in the supplemental material). Plasmid accession numbers were used to download amino acid sequences, coding nucleotide sequences, and GenBank files. To ease comparisons between genome and plasmid sequence information, sequence descriptors for plasmid protein sequences were adjusted to mirror the format used for bacterial and archaeal RefSoil files.

Accessing RefSeq genomes and plasmids. Complete RefSeq genomes and plasmids were downloaded from NCBI to compare with RefSoil. All RefSeq bacteria and archaea protein sequences were downloaded from release 89 (<ftp://ftp.ncbi.nlm.nih.gov/refseq/release>). All GenBank files for complete RefSeq assemblies were downloaded from NCBI. A total of 10,270 bacterial and 259 archaeal assemblies were downloaded. GenBank files were used to extract plasmid size and to compile a list of chromosomal and plasmid accession numbers. GenBank information was read into R, and accession numbers for plasmids and chromosomes were separated. Additionally, all RefSoil accession numbers were removed from the RefSeq accession numbers. Ultimately, 10,335 chromosome and 8,271 plasmids were collected to represent non-RefSoil microorganisms. Protein files were downloaded and tidied using the protocol for RefSoil plasmids as described above.

Plasmid characterization. We summarized the RefSoil+ and RefSeq plasmids in several ways. Plasmid size was extracted from GenBank files for each RefSoil genome and plasmid. For comparison, size was also extracted from RefSeq plasmids. These data were compiled and analyzed in the R statistical environment for computing (42). The RefSoil metadata (Table S1), which contains host information for each plasmid, was used to calculate proportions of RefSoil microorganisms with plasmids. Both the number of plasmids per organism and the number of RefSoil microorganisms with one plasmid were examined. Plasmid size distributions were compared using Mann Whitney U tests, Hartigan's dip test (43), and bimodality coefficients (44). The environmental abundances of RefSoil plasmids were calculated using estimations of RefSoil organism environmental abundance (21). Only soil orders with the most RefSoil+ representatives (alfisols, mollisols, and vertisols [21]) were included in the analysis.

Antibiotic resistance gene detection. We examined ARGs from the ResFams database (174 total [23] in RefSoil+) (see Table S3). We then used HMMs from the ResFams database (23) to search amino acid sequence data from RefSoil genomes and plasmids with a publicly available custom script and HMMER (45). To perform the search, `hmmsearch` (45) was used with `-cut_ga` and `-tblout` parameters. These steps were repeated for protein sequence data from the complete RefSeq database (accessed 24 July 2018). Tabular outputs from both data sets were analyzed in R. Quality scores and percent alignments were plotted to determine quality cutoff values for each gene (Fig. S1). All final hits were required to be within 10% of the model length and to have a score of at least 30% of the maximum score for that gene. When one amino acid sequence was annotated twice (i.e., for similar genes), the hit with the lower score was discarded. The final quality filtered hits were used to plot the distribution of ARGs in RefSoil genomes and plasmids.

Data availability. All data and workflows are publicly available on GitHub (https://github.com/ShadeLab/RefSoil_plasmids). A table of all RefSoil microorganisms with genome and plasmid accession numbers is available in Table S2 and GitHub in the DATABASE_plasmids repository. This repository also hosts amino acid and nucleotide sequences for RefSoil+ genomes and plasmids. Plasmid retrieval workflows are included in the BIN_retrieve_plasmids directory.

All workflows are included on GitHub as well in the ANALYSIS_antibiotic_resistance repository.

SUPPLEMENTAL MATERIAL

Supplemental material for this article may be found at <https://doi.org/10.1128/mSystems.00349-18>.

FIG S1, EPS file, 0.4 MB.

FIG S2, EPS file, 1.3 MB.

FIG S3, EPS file, 1 MB.

TABLE S1, CSV file, 14.1 MB.

TABLE S2, CSV file, 0.2 MB.

TABLE S3, XLSX file, 0.1 MB.

ACKNOWLEDGMENTS

This work was supported in part by the National Science Foundation under grants DEB number 1655425 and DEB number 1749544, by the USDA National Institute of Food and Agriculture and Michigan State AgBioResearch, and by the Great Lakes Bioenergy Research Center U.S. Department of Energy, Office of Science, Office of Biological and Environmental Research under award number DE-SC0018409 (to A.S.). Support was also provided by the Michigan State University Department of Microbiology and Molecular Genetics Russell B. DuVall Fellowship (to T.K.D.). This work was funded in part by the DOE Center for Advanced Bioenergy and Bioproducts Innovation (U.S. Department of Energy, Office of Science, Office of Biological and Environmental Research, under award number DE-SC0018420, to J.C. and A.H.). Any opinions, findings, and conclusions or recommendations expressed in this publication are those of the author(s) and do not necessarily reflect the views of the U.S. Department of Energy.

We thank the Jim Cole and the Ribosomal Database Project for helpful feedback on the work.

REFERENCES

- Glick BR. 1995. The enhancement of plant growth by free-living bacteria. *Can J Microbiol* 41:109–117. <https://doi.org/10.1139/m95-015>.
- Hu J, Wei Z, Friman VP, Gu SH, Wang XF, Eisenhauer N, Yang TJ, Ma J, Shen QR, Xu YC, Jousset A. 2016. Probiotic diversity enhances rhizosphere microbiome function and plant disease suppression. *mBio* 7:e01790-16. <https://doi.org/10.1128/mBio.01790-16>.
- Falkowski PG, Fenchel T, Delong EF. 2008. The microbial engines that drive Earth's biogeochemical cycles. *Science* 320:1034–1039. <https://doi.org/10.1126/science.1153213>.
- Smalla K, Jechalke S, Top EM. 2015. Plasmid detection, characterization and ecology. *Microbiol Spectr* 3:PLAS-0038-2014. <https://doi.org/10.1128/microbiolspec.PLAS-0038-2014>.
- Smillie C, Garcillan-Barcia MP, Francia MV, Rocha EPC, de la Cruz F. 2010. Mobility of plasmids. *Microbiol Mol Biol Rev* 74:434–452. <https://doi.org/10.1128/MMBR.00020-10>.
- Aminov RI. 2011. Horizontal gene exchange in environmental microbiota. *Front Microbiol* 2:158. <https://doi.org/10.3389/fmicb.2011.00158>.
- Heuer H, Smalla K. 2012. Plasmids foster diversification and adaptation of bacterial populations in soil. *FEMS Microbiol Rev* 36:1083–1104. <https://doi.org/10.1111/j.1574-6976.2012.00337.x>.
- van Hoek AHAM, Mevius D, Guerra B, Mullany P, Roberts AP, Aarts HJM. 2011. Acquired antibiotic resistance genes: an overview. *Front Microbiol* 2:203. <https://doi.org/10.3389/fmicb.2011.00203>.
- Sentchilo V, Mayer AP, Guy L, Miyazaki R, Green Tringe S, Barry K, Malfatti S, Goessmann A, Robinson-Rechavi M, van der Meer JR. 2013. Community-wide plasmid gene mobilization and selection. *ISME J* 7:1173–1186. <https://doi.org/10.1038/ismej.2013.13>.
- Forsberg KJ, Reyes A, Wang B, Selleck EM, Sommer MO, Dantas G. 2012. The shared antibiotic resistome of soil bacteria and human pathogens. *Science* 337:1107–1111. <https://doi.org/10.1126/science.1220761>.
- Poirel L, Rodriguez-Martinez J-M, Mammari H, Liard A, Nordmann P. 2005. Origin of plasmid-mediated quinolone resistance determinant QnrA. *Antimicrob Agents Chemother* 49:3523–3525. <https://doi.org/10.1128/AAC.49.8.3523-3525.2005>.
- Patel R, Piper K, Cockerill FR, Steckelberg JM, Yousten AA. 2000. The biopesticide *Paenibacillus popilliae* has a vancomycin resistance gene

- cluster homologous to the enterococcal VanA vancomycin resistance gene cluster. *Antimicrob Agents Chemother* 44:705–709. <https://doi.org/10.1128/AAC.44.3.705-709.2000>.
13. Finley RL, Collignon P, Larsson DGJ, McEwen SA, Li X-Z, Gaze WH, Reid-Smith R, Timinouni M, Graham DW, Topp E. 2013. The scourge of antibiotic resistance: the important role of the environment. *Clin Infect Dis* 57:704–710. <https://doi.org/10.1093/cid/cit355>.
 14. Forsberg KJ, Patel S, Gibson MK, Lauber CL, Knight R, Fierer N, Dantas G. 2014. Bacterial phylogeny structures soil resistomes across habitats. *Nature* 509:612–616. <https://doi.org/10.1038/nature13377>.
 15. van Elsas JD, Bailey MJ. 2002. The ecology of transfer of mobile genetic elements. *FEMS Microbiol Ecol* 42:187–197. <https://doi.org/10.1111/j.1574-6941.2002.tb01008.x>.
 16. Thomas CM, Nielsen KM. 2005. Mechanisms of, and barriers to, horizontal gene transfer between bacteria. *Nat Rev Microbiol* 3:711–721. <https://doi.org/10.1038/nrmicro1234>.
 17. Schloss PD, Girard RA, Martin T, Edwards J, Thrash JC. 2016. Status of the archaeal and bacterial census: an update. *mBio* 7:e00201-16. <https://doi.org/10.1128/mBio.00201-16>.
 18. Krawczyk PS, Lipinski L, Dziembowski A. 2018. PlasFlow: predicting plasmid sequences in metagenomic data using genome signatures. *Nucleic Acids Res* 46:e35. <https://doi.org/10.1093/nar/gkx1321>.
 19. Rozov R, Brown Kav A, Bogumil D, Shterzer N, Halperin E, Mizrahi I, Shamir R. 2017. Recycler: an algorithm for detecting plasmids from *de novo* assembly graphs. *Bioinformatics* 33:475–482. <https://doi.org/10.1093/bioinformatics/btw651>.
 20. Burton JN, Liachko I, Dunham MJ, Shendure J. 2014. Species-level deconvolution of metagenome assemblies with Hi-C–based contact probability maps. *G3 (Bethesda)* 4:1339–1346. <https://doi.org/10.1534/g3.114.011825>.
 21. Choi J, Yang F, Stepanauskas R, Cardenas E, Garoutte A, Williams R, Flater J, Tiedje JM, Hofmockel KS, Gelder B, Howe A. 2017. Strategies to improve reference databases for soil microbiomes. *ISME J* 11:829–834. <https://doi.org/10.1038/ismej.2016.168>.
 22. O’Leary NA, Wright MW, Brister JR, Ciufu S, Haddad D, McVeigh R, Rajput B, Robbertse B, Smith-White B, Ako-Adjei D, Astashyn A, Badretdin A, Bao Y, Blinkova O, Brover V, Chetvernin V, Choi J, Cox E, Ermolaeva O, Farrell CM, Goldfarb T, Gupta T, Haft D, Hatcher E, Hlavina W, Joardar VS, Kodali VK, Li W, Maglott D, Masterson P, McGarvey KM, Murphy MR, O’Neill K, Pujar S, Rangwala SH, Rausch D, Riddick LD, Schoch C, Shkeda A, Storz SS, Sun H, Thibaud-Nissen F, Tolstoy I, Tully RE, Vatsan AR, Wallin C, Webb D, Wu W, Landrum MJ, Kimchi A, et al. 2016. Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. *Nucleic Acids Res* 44:D733–D745. <https://doi.org/10.1093/nar/gkv1189>.
 23. Gibson MK, Forsberg KJ, Dantas G. 2015. Improved annotation of antibiotic resistance determinants reveals microbial resistomes cluster by ecology. *ISME J* 9:207–216. <https://doi.org/10.1038/ismej.2014.106>.
 24. Luo W, Xu Z, Riber L, Hansen LH, Sørensen SJ. 2016. Diverse gene functions in a soil mobilome. *Soil Biol Biochem* 101:175–183. <https://doi.org/10.1016/j.soilbio.2016.07.018>.
 25. Beszteri B, Temperton B, Frickenhaus S, Giovannoni SJ. 2010. Average genome size: a potential source of bias in comparative metagenomics. *ISME J* 4:1075–1077. <https://doi.org/10.1038/ismej.2010.29>.
 26. Nayfach S, Pollard KS. 2015. Average genome size estimation improves comparative metagenomics and sheds light on the functional ecology of the human microbiome. *Genome Biol* 16:51. <https://doi.org/10.1186/s13059-015-0611-7>.
 27. Sorensen JW, Dunivin TK, Tobin TC, Shade A. 2019. Ecological selection for small microbial genomes along a temperate-to-thermal soil gradient. *Nat Microbiol* 4:55–61. <https://doi.org/10.1038/s41564-018-0276-6>.
 28. Lee C, Kim J, Shin SG, Hwang S. 2006. Absolute and relative QPCR quantification of plasmid copy number in *Escherichia coli*. *J Biotechnol* 123:273–280. <https://doi.org/10.1016/j.jbiotec.2005.11.014>.
 29. Murawska E, Fiedoruk K, Bideshi DK, Swiecicka I. 2013. Complete genome sequence of *Bacillus thuringiensis* subsp. *thuringiensis* strain I55056, an isolate highly toxic to *Trichoplusia ni*. *Genome Announc* 1:e0010813. <https://doi.org/10.1128/genomeA.00108-13>.
 30. Dunivin TK, Shade A. 2018. Community structure explains antibiotic resistance gene dynamics over a temperature gradient in soil. *FEMS Microbiol Ecol* 94:fy016. <https://doi.org/10.1093/femsec/fiy016>.
 31. Sørensen SJ, Bailey M, Hansen LH, Kroer N, Wuertz S, Sorensen SJ, Bailey M, Hansen LH, Kroer N, Wuertz S. 2005. Studying plasmid horizontal transfer in situ: a critical review. *Nat Rev Microbiol* 3:700–710. <https://doi.org/10.1038/nrmicro1232>.
 32. Fish JA, Chai B, Wang Q, Sun Y, Brown CT, Tiedje JM, Cole JR. 2013. FunGene: the functional gene pipeline and repository. *Front Microbiol* 4:291. <https://doi.org/10.3389/fmicb.2013.00291>.
 33. Pal C, Bengtsson-Palme J, Kristiansson E, Larsson DGJ. 2015. Co-occurrence of resistance genes to antibiotics, biocides and metals reveals novel insights into their co-selection potential. *BMC Genomics* 16:964. <https://doi.org/10.1186/s12864-015-2153-5>.
 34. Xiong W, Sun Y, Ding X, Zhang Y, Zhong X, Liang W, Zeng Z. 2015. Responses of plasmid-mediated quinolone resistance genes and bacterial taxa to (fluoro)quinolones-containing manure in arable soil. *Chemosphere* 119:473–478. <https://doi.org/10.1016/j.chemosphere.2014.07.040>.
 35. Dunivin TK, Yeh SS, Shade A. 2018. Targeting microbial arsenic resistance genes: a new bioinformatic toolkit informs arsenic ecology and evolution in soil genomes and metagenomes. *bioRxiv* <https://doi.org/10.1101/445502>.
 36. Beaulaurier J, Zhu S, Deikus G, Mogno I, Zhang XS, Davis-Richardson A, Canepa R, Triplett EW, Faith JJ, Sebra R, Schadt EE, Fang G. 2018. Metagenomic binning and association of plasmids with bacterial host genomes using DNA methylation. *Nat Biotechnol* 36:61–69. <https://doi.org/10.1038/nbt.4037>.
 37. Jechalke S, Broszat M, Lang F, Siebe C, Smalla K, Grohmann E. 2015. Effects of 100 years wastewater irrigation on resistance genes, class 1 integrons and IncP-1 plasmids in Mexican soil. *Front Microbiol* 6:163. <https://doi.org/10.3389/fmicb.2015.00163>.
 38. Smalla K, Haines AS, Jones K, Krögerrecklenfort E, Heuer H, Schlöter M, Thomas CM. 2006. Increased abundance of IncP-1β plasmids and mercury resistance genes in mercury-polluted river sediments: first discovery of IncP-1β plasmids with a complex mer transposon as the sole accessory element. *Appl Environ Microbiol* 72:7253–7259. <https://doi.org/10.1128/AEM.00922-06>.
 39. Riber L, Burmolle M, Alm M, Milani SM, Thomsen P, Hansen LH, Sorensen SJ. 2016. Enhanced plasmid loss in bacterial populations exposed to the antimicrobial compound irgasan delivered from interpenetrating polymer network silicone hydrogels. *Plasmid* 87–88:72–78. <https://doi.org/10.1016/j.plasmid.2016.10.001>.
 40. White RA, III, Callister SJ, Moore RJ, Baker ES, Jansson JK. 2016. The past, present and future of microbiome analyses. *Nat Protoc* 11:2049–2053. <https://doi.org/10.1038/nprot.2016.148>.
 41. Stepanauskas R. 2015. Wiretapping into microbial interactions by single cell genomics. *Front Microbiol* 6:258. <https://doi.org/10.3389/fmicb.2015.00258>.
 42. R Core Team. 2017. R: a language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria.
 43. Hartigan JA, Hartigan PM. 1985. The dip test of unimodality. *Ann Statist* 13:70–84. <https://doi.org/10.1214/aos/1176346577>.
 44. Ellison AM. 1987. Effect of seed dimorphism on the density-dependent dynamics of experimental populations of *Atriplex triangularis* (Chenopodiaceae). *Am J Bot* 74:1280–1288. <https://doi.org/10.1002/j.1537-2197.1987.tb08741.x>.
 45. Johnson L, Eddy S, Portugaly E. 2010. Hidden Markov model speed heuristic and iterative HMM search procedure. *BMC Bioinformatics* 11:431. <https://doi.org/10.1186/1471-2105-11-431>.