Title

RefSoil+: A reference for antimicrobial resistance genes on soil plasmids

Authors

TK Dunivin[1,2], J Choi[3], AC Howe[3] and A Shade[1,4]

1. Department of Microbiology and Molecular Genetics, Michigan State University, East Lansing MI 48840 USA
2. Environmental and Integrative Toxicological Sciences, Michigan State University, East Lansing MI 48840
3. Department of Agricultural and Biosystems Engineering, Iowa State University Ames, IA 50011
4. Department of Plant, Soil and Microbial Sciences; Program in Ecology, Evolutionary Biology and Behavior; and the Plant Resilience Institute, Michigan State University, East Lansing, MI 48840

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 understanding of soil plasmids, we present RefSoil+, a database containing plasmid sequences from 922 soil microorganisms. Soil plasmids were relatively larger than other described plasmids, which is a trait associated with plasmid mobility. There was no relationship between chromosome size and plasmid size or number,

suggesting that these genomic traits are independent in soil. 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 baseline 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 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 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 for investigating clinic-environment dynamics of important plasmid-associated genes, including antibiotic resistance genes.

Introduction

Soil is a unique and ancient environment that harbors immense microbial biodiversity. The soil microbiome has functional consequences for ecosystems, like 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 propagation of plasmid functional genes and allows for 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 organisms in a community as they can be gained or lost by environmental organisms, 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 (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 environments with higher population densities such as the rhizosphere, phyllosphere, and gut microbiomes of soil

74 organisms (15). In the case of antibiotic resistance, mobilization is a public health risk. Broadly,
75 the ability of plasmids to rapidly move genes both between and among membership is linked to
76 diversification in complex systems, especially soils (7).

77 Despite their ecological and functional relevance, plasmids are not well characterized in
78 soil. Plasmids vary in copy number, host range, transfer potential, and genetic makeup (4, 16),
79 making them difficult to assemble and characterize from complex soil metagenomes that contain
80 tens of thousands of bacteria and archaea (17). To aid in the study of plasmid-mediated transfer
81 of functional genes in soil, we establish a resource to compare genetic locations of functional
82 genes in soil organisms. We extended the RefSoil database (18) of 922 soil microorganisms to
83 include their plasmids. We used this database to test whether soil-associated plasmids are distinct
84 from plasmids from a broad, general database of microorganisms, RefSeq (19). We focused our
85 comparisons on the content, diversity, and location of ARGs on plasmids and chromosomes. We
86 used hidden markov models to search for clinically and agriculturally relevant ARGs in the
87 extended soil database, RefSoil+, and RefSeq. RefSoil+ provides insights into the range of
88 plasmid sizes and their functional potential within soil microorganisms. RefSoil+ can be used to
89 inform and test hypotheses about the traits, functional gene content, and spread of soil-associated
90 plasmids and can serve as a reference for plasmid assembly from metagenomes.

91

92 **Results and discussion**

93 *Plasmid characterization*

94 RefSoil+ is a database of soil-associated plasmids as an extension of RefSoil, which
95 includes taxonomic information, amino acid sequences, coding nucleotide sequences, and
96 GenBank files for a curated set of 922 soil-associated organisms. A total of 927 plasmids were

97 associated with RefSoil organisms, and 370 RefSoil organisms (40.1%) had at least one plasmid
98 (**Figure 1A**). This is high compared to the proportion of non-eukaryotic plasmids in the general
99 RefSeq database (20%). The mean number of plasmids per RefSoil organism was 1.01, but the
100 number of plasmids per organism varied greatly (**Figure 1B**). For example, strain *Bacillus*
101 *thuringiensis* serovar thuringiensis (RefSoil 738) had 14 plasmids, ranging from 6,880 to
102 328,151 bp. The abundance of plasmids found in RefSoil genomes highlights plasmids as an
103 important component of soil microbiomes (7, 20).

104 Soil-associated plasmids tended to be larger than plasmids from other environments.
105 RefSoil plasmids contained > 195,000 kbp and increased the number of base pairs included in
106 RefSoil by 4.4%. Plasmid size in RefSoil organisms ranged from 1,286 bp to 2.58 Mbp (**Figure**
107 **2A**), which rivals the range of all known plasmids from various environments (744 bp – 2.58
108 Mbp) (16). In the distribution of plasmid size, both upper and lower extremes had representatives
109 from soil. Plasmids from all habitats had a characteristic bimodal size distribution with peaks at 5
110 kb and 35 kb (15–17). Soil-associated plasmids in RefSoil+, however, trended larger and did not
111 have many representatives in the lower size range (**Figure 2**). Specifically, RefSoil+
112 proportionally contained more plasmids > 100 kb (**Figure 2B,** Mann-Whitney U test $p < 0.001$).
113 Thus, while soil-associated plasmids vary in size, they are, on average, large. This is of particular
114 importance because of the established differences in mobility of plasmids in different size ranges
115 (5). Mobilizable plasmids, which have relaxases, tend to be larger than non-transmissible
116 plasmids, with median values of 35 and 11 kbp respectively (5). The majority of soil-associated
117 plasmids were > 35 kbp (**Figure 2**), suggesting they are more likely to be mobile. Additionally,
118 conjugative plasmids, which encode type IV coupling proteins, have a larger median size (181
119 kbp) (5). The median size of soil-associated plasmids was 91 kbp (**Figure 2**), suggesting that

120 these soil-associated plasmids are more likely to be conjugative. Future works should examine
121 genetic potential for transfer of plasmids associated with different ecosystems to test this
122 hypothesis.

123 Genome size, inclusive of chromosomes and plasmids, is an important ecological trait
124 that is difficult to estimate from metagenomes (24). Due to incomplete assemblies, genome size
125 must be approximated based on the estimated number of organisms through single-copy gene
126 abundance (25). Extrachromosomal elements, however, inflate these estimated genome sizes
127 because they contribute to the sequence information of the metagenome often without
128 contributing single-copy genes (26). While our methodologies do not account for plasmid copy
129 number (27), we examined the relationship between genome size and plasmid size in soil-
130 associated organisms, and found none (**Figure 3**). Additionally, chromosome size was not
131 predictive of the number of plasmids (**Figure 3**; **Figure S1**). For example, *Bacillus thuringiensis*
132 subsp. thuringiensis Strain IS5056 had the most plasmids in RefSoil+, but these plasmids
133 spanned the size range of 6.8 - 328 kbp. This strain's plasmids make up 19% of its coding
134 sequences (28), but its chromosome (5.4 Mbp) is average for soils (26). Despite that there is no
135 clear relationship between genome size and plasmid characteristics within these data, the plasmid
136 database can be used to inform estimates of average genome sizes from close relatives detected
137 within metagenomes.
138
139 *ARGs in soil plasmids*
140 It is unclear whether soil ARGs are predominantly on chromosomes or mobile genetic
141 elements. While mobile gene pools are not static, there is evidence to suggest low transfer of
142 ARGs in soil (14, 15, 29). For example, bulk soils are not a "hot spot" for HGT because they are

143 often resource-limited (30), and surveys of ARGs in soil metagenomes have suggested a
144 predominance of vertical transfer, rather than horizontal transfer, of ARGs (14, 29). Using
145 RefSoil+, we examined 36 genes encoding resistance to beta-lactams, tetracyclines,
146 aminoglycosides, chloramphenicol, vancomycin, sulfonamides, macrolides, and trimethoprim
147 (29). After quality filtering, we detected 3,217 ARGs in RefSoil chromosomes and plasmids
148 (**Figure 4; Table S1**).

149 Adding plasmids to the RefSoil database increased functional genes in the database, as
150 128 ARG sequences were only detected on plasmids (**Figure 4C**). These functional genes would
151 be missed if only chromosomes were considered. With the exception of sulfonamides, the
152 majority of ARGs were chromosomally encoded in soils (**Figure 4AB**). We examined the
153 genomic distributions of ARGs in RefSoil+ based on taxonomy (**Figure S3**). Proteobacteria had
154 the most plasmid-associated ARGs, which has been reported previously (31). ARGs were found
155 on chromosomes more often than plasmids, but we were curious whether this phenomenon was
156 specific to soil. Therefore, we compared ARG content in RefSoil to all other known plasmids
157 (RefSeq database; n = 9,132, (19)) and found that the number of ARGs per genome was
158 comparable for RefSoil and RefSeq, but RefSoil plasmids had proportionally fewer ARGs than
159 RefSeq plasmids (**Figure S4;** Mann-Whitney U test p-value = 0.002). This suggests that
160 plasmid-mediated HGT rates of ARGs may be relatively low in these soil organisms. We note
161 that the RefSoil database is limited in representatives of Verrucomicrobia and Acidobacteria
162 which may change these estimates (18); however, this will improve as the database grows. We
163 examined this trend for each gene individually and still observed a greater proportion of ARG
164 sequences on plasmids in RefSeq compared with RefSoil+ with one exception, ANT9 which
165 encodes a Streptomycin 3"-adenylyltransferase (**Figure 5**). Additionally, 12 genes (ANT3, CEP,

166 *dfra1*, *ermB*, *intI*, *qnr*, *repA*, *strA*, *strB*, *sul2*, *tetD*, *vanZ*) were more common on plasmids in
167 RefSeq compared to only 3 genes (CEP, *dfra1*, *repA*) in RefSoil+ (**Figure 5**). Thus, these soil
168 bacteria harbor relatively fewer ARGs on plasmids, suggesting that RefSoil+ organisms have
169 limited capacity for plasmid-mediated transfer of these genes. These data represent a baseline of
170 ARGs present on chromosomes and plasmids in soil microorganisms. This is important because
171 some data suggest that soil ARGs are increasing over time due to increased antibiotic exposure
172 (32). Future assessments of functional gene content on chromosomes and plasmids together will
173 help to delineate changes in transfer potential and reveal selective or environmental factors that
174 impact transfer potential.

175 We examined the abundance of ARGs in RefSoil+ and RefSeq strains and asked whether
176 these ARGs were more commonly detected on chromosomes or plasmids. Gibson and colleagues
177 (2015) compared soil-associated isolates with water and human-associated strains and found an
178 abundance of genes encoding multidrug efflux pumps and beta lactam resistance but not
179 tetracycline resistance in soil (33). This was also observed in our analysis (**Figure 5**). By
180 determining whether ARGs were encoded on plasmids or chromosomes, we were also able to
181 show that these patterns were due to chromosomal genes and more likely vertically transferred
182 (**Figure 5**). While genome data from isolates cannot speak to environmental abundance of
183 ARGs, our data support observations of ARGs in mobile genetic elements in soil from
184 cultivation-independent studies as well. Luo and colleagues (2016) observed a low abundance of
185 chloramphenicol, quinolone, and tetracycline resistance genes in soil mobile genetic elements
186 (20), and Xiong and colleagues (2015) also observed low abundance of *qnr* genes in a soil
187 mobile genetic elements (34). While plasmids are not the sole mobile genetic element, we
188 observed fewer plasmid-encoded chloramphenicol, quinolone, and tetracycline resistance genes

189 in soil-associated microorganisms than RefSeq microorganisms (**Figure 5**). Mobile genetic
190 elements in soil have also been shown to have an abundance of genes encoding multidrug efflux
191 pumps and resistance to beta-lactams, aminoglycosides, and glycopeptides (20). While we
192 detected genes encoding aminoglycoside and beta-lactam resistance and multi drug efflux pumps
193 in RefSoil+, we observed lower counts on plasmids as compared with chromosomes (**Figure 4**;
194 **Figure 5**). Additionally, we did not detect plasmid-borne vancomycin resistance genes, despite
195 that environmental samples have shown vancomycin resistance genes on mobile genetic
196 elements (20). Though all isolate databases are biased by common cultivation conditions, these
197 data point to gaps in our soil collections with a specific eye towards representation of plasmid
198 content.
199
200 *RefSoil+ applications*
201 Plasmid assembly tools rely on existing databases to assemble plasmids from metagenomes
202 (35, 36), but this work shows that soil-associated plasmids are distinct. While this RefSoil+ is
203 biased towards cultured strains, characterization of known plasmids is essential to improve
204 detection of novel plasmids (21). This database of soil-associated plasmids expands knowledge
205 of functional genes with potential for transfer in soil microbiomes, highlights the contribution of
206 plasmids to metagenome-estimated genome size, offers insights into plasmid host ranges in soil,
207 and serves as a reference for future works.
208 Host taxonomy can be observed in RefSoil+ because it is populated by the chromosomes and
209 plasmids of isolates. While RefSoil+ does not predict plasmid presence or gene content in the
210 environment, annotation of cultivable organisms with plasmids is important for soil systems
211 because traditional methods of assembly and annotation from metagenomes allows only for

coarse estimation of host identity (35, 37). Plasmid gene content is not static (38), and organisms can gain or lose plasmids (39, 40). 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 changes in the future.

RefSoil+ can be used to better target plasmids in the environment, whether it is used as a reference database or as a database for primer design. New microbiome sequencing techniques such as Hi-C sequencing (41), long-read technology (42), or single cell sequencing (43) could add to and leverage RefSoil+ to improve characterization of plasmid-host relationships in soil. As movement of ARGs are observed in the clinic and the environment, RefSoil+ can also serve as a reference for comparison with legacy plasmid and chromosome content 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 resource for research frontiers in plasmid ecology and evolution within wild microbiomes.

Materials and methods

Data availability

All data and workflows are publicly available on GitHub (github.com/ShadeLab/RefSoil_plasmids). A table of all RefSoil organisms 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.

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. Plasmid accession numbers were compiled for each strain and added to the RefSoil database to make RefSoil+ (**Table S1**). 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, 10,359 chromosome and 9,132 plasmid accession numbers were collected to represent non-RefSoil

257 plasmids. Protein files were downloaded and tidied using the protocol for RefSoil plasmids as
258 described above.
259
260 *Plasmid characterization*
261 We summarized the RefSoil+ and RefSeq plasmids in several ways. Plasmid size was
262 extracted from GenBank files for each RefSoil genome and plasmid. For comparison, size was
263 also extracted from RefSeq plasmids. These data were compiled and analyzed in the R statistical
264 environment for computing (44). The RefSoil metadata (**Table S1**), which contains host
265 information for each plasmid, was used to calculate proportions of RefSoil organisms with
266 plasmids. Both the number of plasmids per organism and the number of RefSoil organisms with
267 one plasmid were examined.
268
269 *Antibiotic resistance gene detection*
270 We examined 36 clinically-relevant ARGs in RefSoil+, including *AAC6-Ia, adeB, ANT3,*
271 *ANT6, ANT9, blaA, blaB, blaC, CAT, cmlA, dfra1, dfra12, ermB, ermC, intI, mexC, mexE, qnr,*
272 *repA, strA, strB, sul2, tetA, tetD, tetM, tetQ, tetW, tetX, tolC, vanA, vanC, vanH, vanT, vanW,*
273 *vanX,* and *vanZ*. For each gene of interest, hidden makrov models were downloaded from the
274 FunGene database (45), which includes some models from the Resfams database (33). We then
275 used these models to search amino acid sequence data from RefSoil genomes and plasmids with
276 a publicly available, custom script and HMMER (46). To perform the search, hmmsearch (46)
277 was used with an e-value cutoff of 10^{-10}. These steps were repeated for protein sequence data
278 from the complete RefSeq database (accessed 24 July 2018). Tabular outputs from both datasets
279 were analyzed in R. Quality scores and percent alignments were plotted to determine quality

cutoff values for each gene (**Figure S2**). All final hits were required to be within 10% of the model length and to have a score of at least 40% of the maximum score for that gene. Based on quality distributions and GenBank function assignments, additional quality filtering by score was applied to genes *adeB*, *CEP*, *vanA*, *vanC*, *vanH*, *vanX*, and *vanW*. 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.

Acknowledgements. AS acknowledges support in part from the National Science Foundation under Grants DEB #1655425 and DEB#1749544, from the USDA National Institute of Food and Agriculture and Michigan State AgBioResearch, and from 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. TKD acknowledges support from the Michigan State University Department of Microbiology and Molecular Genetics Russell B. DuVall Fellowship. We thank the Jim Cole and the Ribosomal Database Project for helpful feedback on the work.

References

1. Glick BR. 1995. The enhancement of plant growth by free-living bacteria. Can J Microbiol 41:109–117.
2. 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 7:1–8.
3. Falkowski PG, Fenchel T, Delong EF. 2008. The Microbial Engines That Drive Earth's Biogeochemical Cycles. Science (80-).
4. Smalla K, Jechalke S, Top EM. 2015. Plasmid detection, characterization and ecology. Cancer 121:1265–1272.
5. Smillie C, Garcillan-Barcia MP, Francia M V., Rocha EPC, de la Cruz F. 2010. Mobility of Plasmids. Microbiol Mol Biol Rev 74:434–452.
6. Aminov RI. 2011. Horizontal gene exchange in environmental microbiota. Front Microbiol 2:1–19.
7. Heuer H, Smalla K. 2012. Plasmids foster diversification and adaptation of bacterial populations in soil. FEMS Microbiol Rev 36:1083–1104.

8. Van Hoek AHAM, Mevius D, Guerra B, Mullany P, Roberts AP, Aarts HJM. 2011. Acquired antibiotic resistance genes: An overview. Front Microbiol 2:1–27.
9. 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 7:1173–86.
10. Forsberg KJ, Reyes A, Wang B, Selleck EM, Sommer MO, Dantas G. 2012. The shared antibiotic resistome of soil bacteria and human pathogens. Science (80-) 337:1107–1111.
11. Poirel L, Liard A, Nordmann P, Mammeri H. 2005. Origin of Plasmid-Mediated Quinolone Resistance Determinant QnrA. Antimicrob Agents Chemother 49:3523–3525.
12. 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.
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.
14. Forsberg KJ, Patel S, Gibson MK, Lauber CL, Fierer N, Dantas G. 2014. Bacterial phylogeny structures soil resistomes across habitats. Nature 509:612–616.
15. van Elsas JD, Bailey MJ. 2002. The ecolgy of transfer of mobile genetic elements. FEMS Microb Ecol 42:187–197.
16. Thomas CM, Nielsen KM. 2005. Mechanisms of, and barriers to, horizontal gene transfer between bacteria. NatRevMicrobiol 3:711–721.
17. Schloss PD, Girard RA, Martin T, Edwards J, Thrash JC. 2016. Status of the archaeal and bacterial census: An update. MBio 7:1–10.
18. 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.
19. O'Leary NA, Wright MW, Brister JR, Ciufo 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, Tatusova T, DiCuccio M, Kitts P, Murphy TD, Pruitt KD. 2016. Reference sequence (RefSeq) database at NCBI: Current status, taxonomic expansion, and functional annotation. Nucleic Acids Res 44:D733–D745.
20. Luo W, Xu Z, Riber L, Hansen LH, Sørensen SJ. 2016. Diverse gene functions in a soil mobilome. Soil Biol Biochem 101:175–183.
21. Shintani M, Sanchez ZK, Kimbara K. 2015. Genomics of microbial plasmids: Classification and identification based on replication and transfer systems and host taxonomy. Front Microbiol 6:1–16.
22. Garcillán-Barcia MP, Alvarado A, De la Cruz F. 2011. Identification of bacterial plasmids based on mobility and plasmid population biology. FEMS Microbiol Rev 35:936–956.
23. Li L-G, Xia Y, Zhang T. 2017. Co-occurrence of antibiotic and metal resistance genes revealed in complete genome collection. ISME J 11:651–662.

24. Beszteri B, Temperton B, Frickenhaus S, Giovannoni SJ. 2010. Average genome size: a potential source of bias in comparative metagenomics. ISME J.
25. 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.
26. Sorensen JW, Dunivin TK, Tobin TC, Shade A, (in review) . Ecological selection for small microbial genomes along a temperate-to-thermal soil gradient. bioRxiv.
27. 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.
28. Murawska E, Fiedoruk K, Bideshi DK, Swiecicka I. 2013. Complete Genome Sequence of Bacillus thuringiensis subsp. thuringiensis Strain IS5056, an Isolate Highly Toxic to Trichoplusia ni. Genome Announc 1:e00108-13-e00108-13.
29. Dunivin TK, Shade A. 2018. Community structure explains antibiotic resistance gene dynamics over a temperature gradient in soil. FEMS Microbiol Ecol 94.
30. 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.
31. 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.
32. Knapp CW, Dolfing J, Ehlert PA, Graham DW. 2010. Evidence of Increasing Antibiotic Resistance Gene Abundances in Archived Soils since 1940. Environ Sci Technol 44:580–587.
33. Gibson MK, Forsberg KJ, Dantas G. 2014. Improved annotation of antibiotic resistance determinants reveals microbial resistomes cluster by ecology. ISME J 9:1–10.
34. Musovic S, Klümper U, Dechesne A, Magid J, Smets BF. 2014. Long-term manure exposure increases soil bacterial community potential for plasmid uptake. Environ Microbiol Rep 6:125–130.
35. Krawczyk PS, Lipinski L, Dziembowski A. 2018. PlasFlow: predicting plasmid sequences in metagenomic data using genome signatures. Nucleic Acids Res 46.
36. Rozov R, Brown Kav A, Bogumil D, Shterzer N, Halperin E, Mizrahi I, Shamir R. 2016. Recycler: an algorithm for detecting plasmids from *de novo* assembly graphs. Bioinformatics.
37. 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.
38. 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 6:1–10.
39. Smalla K, Haines AS, Jones K, Krögerrecklenfort E, Heuer H, Schloter 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.
40. Riber L, Burmolle M, Alm M, Milani SM, Thomsen P, Hansen LH, Sï¿½rensen 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.

41. Burton JN, Liachko I, Dunham MJ, Shendure J. 2014. Species-Level Deconvolution of Metagenome Assemblies with Hi-C–Based Contact Probability Maps. G3: Genes|Genomes|Genetics 4:1339–1346.
42. White RA, Callister SJ, Moore RJ, Baker ES, Jansson JK. 2016. The past, present and future of microbiome analyses. Nat Protoc 11:2049–2053.
43. Stepanauskas R. 2015. Wiretapping into microbial interactions by single cell genomics. Front Microbiol 6:2014–2016.
44. R Core Team. 2017. R: A Language and Environment for Statistical Computing. Vienna, Austria.
45. Fish JA, Chai B, Wang Q, Sun Y, Brown CT, Tiedje JM, Cole JR. 2013. FunGene: The functional gene pipeline and repository. Front Microbiol.
46. Johnson L, Eddy S, Portugaly E. 2011. Hidden Markov Model Speed Heuristic and Iterative HMM Search Procedure. BMC Bioinformatics 39.

Table and Figure legends

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

Figure 2. Plasmid size distributions. A) Histogram of plasmid size (kbp) from RefSoil plasmids. **B)** RefSoil (blue) and RefSeq (gray) plasmid size distributions.

Figure 3. Relationship between plasmid size and genome size. Total plasmid size (sum of all plasmids in an 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)).

Figure 4. Distribution of ARGs in RefSoil genomes and plasmids. A) The proportion of

ARGs on plasmids (light blue), genomes (green) or both (dark blue) in RefSoil+ microorganisms. **B)** The raw numbers of detected ARGs. Bars are colored by location of genomic element (as in panel A) and categorized by antibiotic resistance gene group. The number of genes included in each group is shown in parentheses. **C)** A table with the number different ARGs that were only found on plasmids. Genes are ordered by ranked abundance.

Figure 5. Proportion of genes on genomes and plasmids in RefSoil+ and RefSeq databases. Number of ARGs was normalized to number of genetic elements. Bars are colored by genetic element

Figure S1. Relationship between plasmid number and genome size. Boxplots showing the distribution of genome sizes based on the number of plasmids. Numbers above boxplots show the number of organisms in that category. P-value from an ANOVA is also shown.

Figure S2. Quality of RefSoil+ ARG hits. Percent alignment was plotted against the score for each ARG hit for quality filtering purposes.

Figure S3. Distribution of ARGs in RefSoil chromosomes and plasmids by taxonomy. The number of detected ARGs were normalized to the number of RefSoil organisms in each phylum and Proteobacteria class. ARG hits are colored by genetic location. The number of taxa included in each phylum is shown in parentheses.

463 **Figure S4. Proportion of ARGs in RefSoil and RefSeq databases.** Boxplots of the proportion
464 of ARGs per genetic element. Each ARG was normalized to the number of genetic
465 elements in the database. Points are colored by ARG category, and P-values for Mann-
466 Whitney U test are 0.55 (n.s. is not significant) and 0.007 (**) for chromosomes and
467 plasmids respectively.
468
469 **Table S1.** Quality filtered ARG hits in RefSoil genomes and plasmids. Information on quality
470 scores and accession numbers for each ARG hit.
471
472 **Table S2.** RefSoil taxonomy table with plasmid and genome accession numbers.
473

Figure 1

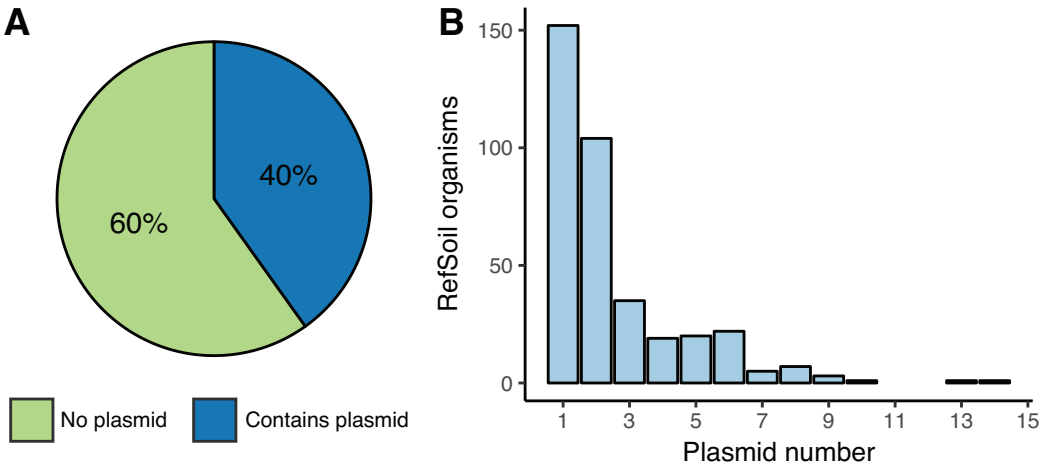

Figure 2

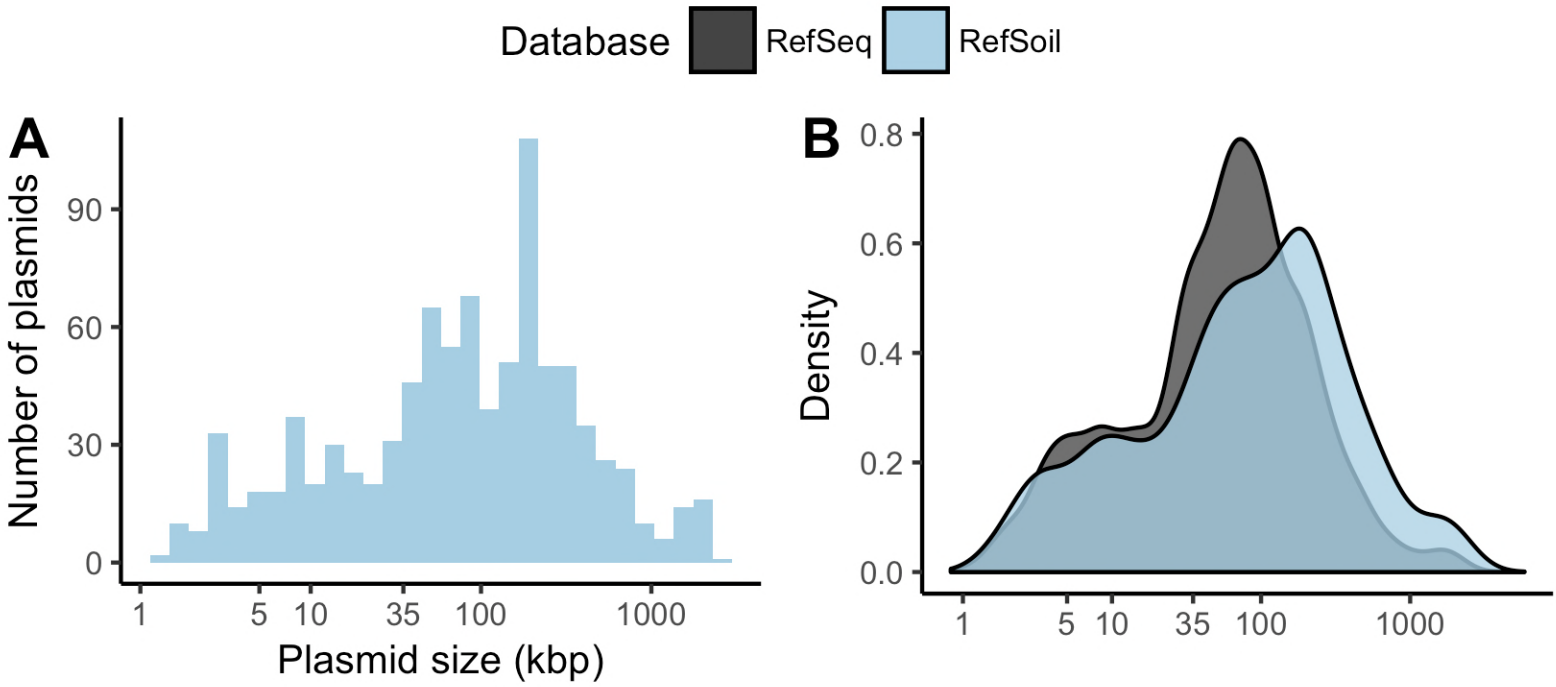

Figure 3

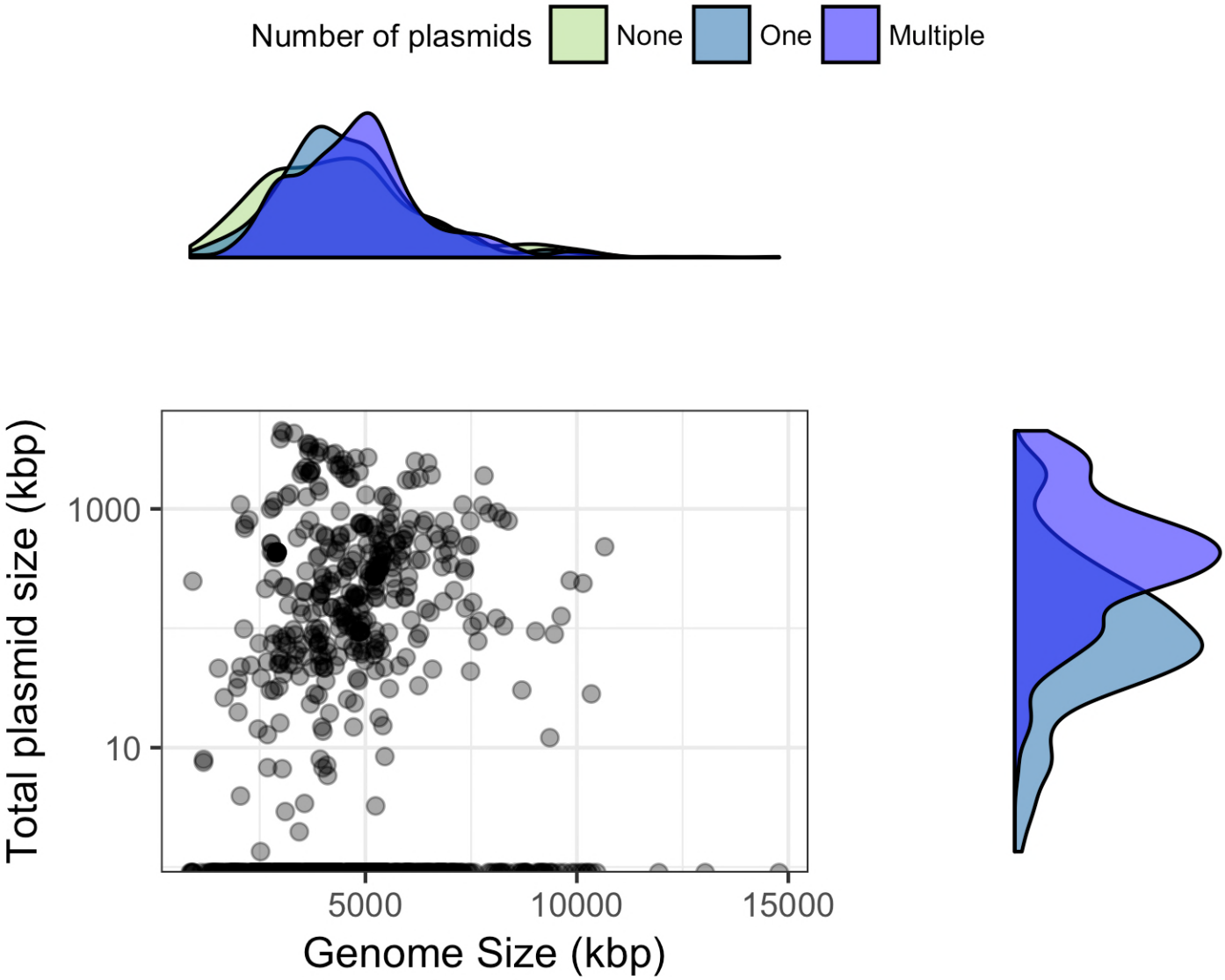

Figure 4

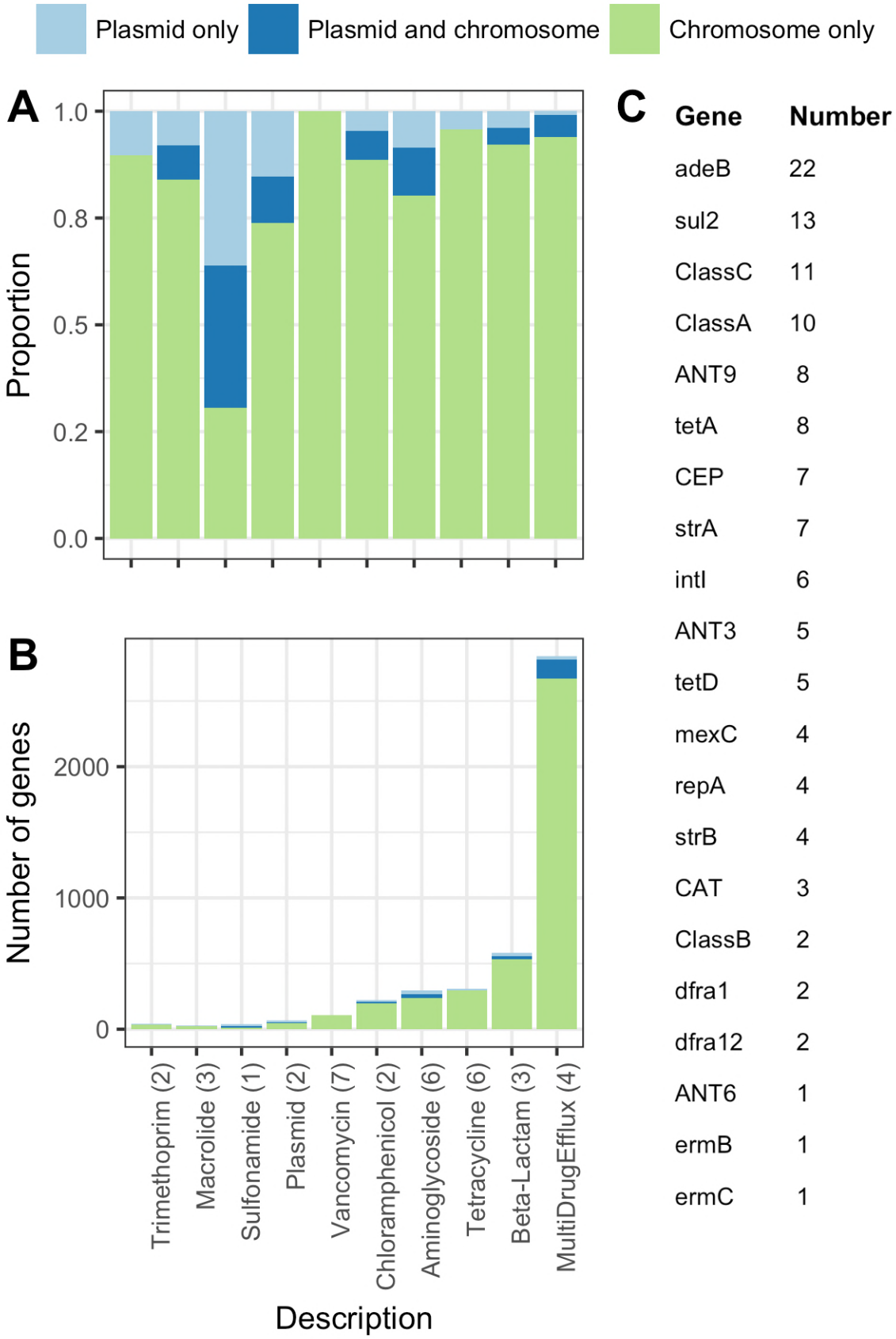

Figure 5