Title

RefSoil+: A reference database for genes and traits of 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

*Correspondence: shadeash@msu.edu

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 a weak relationship between chromosome size and plasmid size or 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 as compared to non-soil 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 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 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 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 (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

74 population densities such as the rhizosphere, phyllosphere, and gut microbiomes of soil
75 microorganisms (15). In the case of antibiotic resistance, mobilization is a public health risk.
76 Broadly, the ability of plasmids to rapidly move genes both between and among membership is
77 linked to diversification in complex systems, especially soils (7).

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

To aid in the study of plasmids and their associated functional genes in soil, we establish 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 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 (**Figure 1A**). This is high compared to the proportion of non-eukaryotic 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; **Figure 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 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 (**Figure 2A**), which rivals the range of all known plasmids from various environments (744 bp – 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 (Hartigans' dip test $p < 0.01$; Bimodality coefficient = 0.745) and modes at 3 kb and 59 kb (**Figure 2**). Soil-associated plasmids in RefSoil+ also had a multimodal size distribution (Hartigans' 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 $p < 0.01$) (**Figure 2**). Specifically, RefSoil+ proportionally contained more plasmids > 100 kb (**Figure 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 (2010) showed that mobilizable plasmids, which have relaxases, tend to be larger than non-transmissible plasmids, with median values of 35 and 11 kbp respectively (5). The majority of soil-associated plasmids (68.2%) were > 35 kbp (**Figure 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 (**Figure 2**), suggesting that these soil-associated plasmids are more likely to be conjugative.

142 Future works should examine genetic potential for transfer of plasmids associated with different
143 ecosystems to test this hypothesis.
144 Plasmid size may vary in the environment. To estimate the environmental size
145 distributions of plasmids, we used estimates of the environmental abundance RefSoil
146 microorganisms (21). We focused on soil orders previously shown to include the most RefSoil
147 representatives (Alfisols, Mollisols, Vertisols) (21). We found that plasmid size distributions
148 varied based on soil order (Kruskal-Wallis $p < 0.01$; **Figure 2C**). True environmental abundance
149 may vary based on plasmid copy number within individuals and plasmids from uncultivated
150 microorganisms, but this estimation gives a rough idea of plasmid size distributions in the
151 environment, and provides some baseline information because there are methodological
152 challenges to accurately measuring plasmid size *in situ* (4, 18, 19).
153 Genome size, inclusive of chromosomes and plasmids, is an important ecological trait
154 that is difficult to estimate from metagenomes (28). Due to incomplete assemblies, genome size
155 must be approximated based on the estimated number of individuals through single-copy gene
156 abundance (29). Extrachromosomal elements, however, inflate these estimated genome sizes
157 because they contribute to the sequence information of the metagenome often without
158 contributing single-copy genes (30). While our methodologies do not account for plasmid copy
159 number (31), we examined the relationship between genome size and plasmid size in soil-
160 associated microorganisms and found a weak but significant correlation (Spearman's $\rho = 0.12$; p
161 < 0.001; **Figure 3**). Additionally, chromosome size was not predictive of the number of plasmids
162 (**Figure 3**; **Figure S1**). For example, *Bacillus thuringiensis* subsp. thuringiensis Strain IS5056
163 had the most plasmids in RefSoil+, but these plasmids spanned the size range of 6.8 - 328 kbp.
164 This strain's plasmids make up 19% of its coding sequences (32), but its chromosome (5.4 Mbp)

165 is average for soils (30). Despite that there is a weak relationship between genome size and
166 plasmid characteristics within these data, the plasmid database can be used to inform estimates of
167 average genome sizes from close relatives detected within metagenomes.
168
169
170 *ARGs on soil plasmids*
171 It is unclear whether soil ARGs are predominantly on chromosomes or mobile genetic
172 elements. While mobile gene pools are not static, there is evidence to suggest low transfer of
173 ARGs in soil (14, 15, 33). For example, bulk soils are not a "hot spot" for HGT because they are
174 often resource-limited (34), and surveys of ARGs in soil metagenomes have suggested a
175 predominance of vertical transfer, rather than horizontal transfer, of ARGs (14, 33). Using
176 RefSoil+ sequences and ResFams HMMs (23), we examined 174 genes encoding resistance to
177 beta-lactams, tetracyclines, aminoglycosides, chloramphenicol, glycopeptides, macrolides,
178 quinolones, and trimethoprim. After quality filtering, we detected 154,392 ARG sequences in
179 RefSoil chromosomes and plasmids (**Figure 4; Table S1**).
180 Adding plasmids to the RefSoil database increased the number of functional gene types,
181 or genes that have functional potential (35), represented in the database, as 7 ARGs (16S rRNA
182 methyltransferase, AAC6-Ib, ANT6, CTXM, ErmC, KPC, TetD) were only detected on
183 plasmids. Notably, these functional genes would be missed if only chromosomes were
184 considered. However, the majority of ARGs were chromosomally encoded in RefSoil+
185 microorganisms (**Figure 4AB**; chromosome v. plasmid Mann Whitney U $p < 0.01$). We next
186 examined the genomic distributions of ARGs in RefSoil+ based on taxonomy (**Figure 4CD**).
187 Proteobacteria had the most plasmid-associated ARGs, which has been reported previously (36).

188 We were curious whether ARGs were more commonly detected on chromosomes than
189 plasmids in general, or if this trend was specific to soil microorganisms. We found that the
190 number of ARGs per genome was comparable for RefSoil and RefSeq (Mann Whitney U $p >$
191 0.05), but RefSoil plasmids had fewer ARGs than RefSeq plasmids (Mann Whitney U $p < 0.05$;
192 **Figure 5**). Normalizing to individual microorganisms is biased towards chomosomes, however,
193 because chromsomes typically have more base pairs than plasmids. To account for this, we also
194 normalized ARGs to base pairs, and plasmids from both databases had more ARGs compared to
195 chromosomes (Mann Whitney U $p < 0.05$). Notably, RefSoil+ had less ARGs compared with
196 RefSeq (Mann Whitney U $p < 0.01$) (**Figure S3**). This suggests that plasmid-mediated HGT
197 rates of ARGs may be relatively low in these soil microorganisms. We note that the RefSoil
198 database is limited in representatives of Verrucomicrobia and Acidobacteria which may change
199 these estimates (21); however, this will improve as the database grows.
200 We examined this trend for each antibiotic class and observed a greater proportion of ARG
201 sequences on plasmids in RefSeq compared with RefSoil+ for genes encoding glycopeptide and
202 tetracycline resistance (**Figure S2**). Gibson and colleagues (2015) also found an lack of
203 tetracycline resistance genes in soil-associated isolates compared to with water and human-
204 associated strains (23). By determining whether ARGs were encoded on plasmids or
205 chromosomes, our analysis suggests that these patterns were due to chromosomal genes and
206 more likely vertically transferred (**Figure 5**). Thus, these soil bacteria harbor relatively fewer
207 ARGs on plasmids, suggesting that RefSoil+ microorganisms have limited capacity for plasmid-
208 mediated transfer of these genes. Future assessments of functional gene content on chromosomes
209 and plasmids together will help to delineate changes in transfer potential and reveal selective or
210 environmental factors that impact transfer potential.

While genome data from isolates cannot speak to 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 (2016) observed a low abundance of chloramphenicol, quinolone, and tetracycline resistance genes in soil mobile genetic elements (24), and Xiong and colleagues (2015) also observed low abundance of *qnr* genes. Similarly, we observed fewer plasmid-encoded tetracycline resistance genes in soil-associated microorganisms than RefSeq microorganisms (**Figure 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 RefSoil+ and RefSeq (Kruskal-Wallis $P > 0.05$; **Figure S2**. However, plasmid-borne glycopeptide resistance genes were less common in RefSoil+ plasmids (Mann Whitney U $P < 0.05$).

RefSoil+ applications

RefSoil+ is publicly available on GitHub (github.com/ShadeLab/RefSoil_plasmids). It includes an excel file linking RefSoil+ organism taxonomy with accession numbers for corresponding chromosomes and plasmids. It also contains several fasta files with CDS and amino acid sequences. These files can be downloaded directly from GitHub. RefSoil+ has been used to better estimate genome sizes in soil (37) and to estimate the distribution of arsenic resistance genes in soil-associated chromosomes and plasmids (38).

233 Our results show that soil-associated plasmids have distinctive traits and can harbor
234 functional genes that are not encoded on host chromosomes. RefSoil+ expands knowledge of
235 functional genes with potential for transfer among soil microorganisms and offers insights into
236 plasmid size and host ranges in soil (and improves accuracy of estimates of their genome sizes),.
237 Because it is populated by the chromosomes and plasmids of isolates, RefSoil+ links host
238 taxonomy to plasmid content. This linkage is important especially for heterogeneous ecosystems
239 with high microbial richness like soils, which rely heavily on cultivation-independent methods
240 for observing microbial diversity. RefSoil+ can guide assembly and support annotation of
241 plasmids from soil metagenomes, and also direct hypotheses of host identity (18, 39). Notably,
242 plasmid gene content is not static (40), and individuals can gain or lose plasmids (41, 42).
243 Despite this, historical data of the genetic makeup and host range of plasmids can be used to
244 better understand plasmid ecology, and to serve as an important reference to understand by how
245 much host plasmid numbers and contents changes in the future. This information contributes to
246 information needed to understand patterns of plasmid dissemination, both across environments
247 and among hosts.
248 RefSoil+ can be used as a reference database or as a database for primer design to target
249 plasmids in the environment. Advances microbiome sequencing methods such as pre-sequencing
250 proximity linkage (e.g. Hi-C; 20), long-read technology (43), or single cell sequencing (44)
251 could add to and leverage RefSoil+ to improve characterization of plasmid-host relationships in
252 soil. As movement of ARGs are observed in the clinic and the environment, RefSoil+ can also
253 serve as a reference for comparison with legacy plasmid and chromosome content and
254 distributions. Novel genomes and plasmids could be added in future RefSoil+ versions, and
255 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

Data availability

All data and workflows are publicly available on GitHub (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.

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+ (**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, 10,335 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 (45). 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 (46), and bimodality coefficients (47). 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, Vertisols; (21)) were included in the analysis.

Antibiotic resistance gene detection

We examined ARGs from the ResFams database (174 total (23) in RefSoil+ (**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 (48). To perform the search, hmmsearch (48) 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 datasets were analyzed in R. Quality scores and percent alignments were plotted to determine quality cutoff values for each gene (**Figure 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.

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 320:1034-1039.
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 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. Nat Rev Microbiol 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. Krawczyk PS, Lipinski L, Dziembowski A. 2018. PlasFlow: predicting plasmid sequences in metagenomic data using genome signatures. Nucleic Acids Res 46.
19. 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 33(4):475-482.
20. Burton JN, Liachko I, Dunham MJ, Shendure J. 2014. Species-Level Deconvolution of Metagenome Assemblies with Hi-C–Based Contact Probability Maps. G3Genes|Genomes|Genetics 4:1339–1346.
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.
22. 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.
23. Gibson MK, Forsberg KJ, Dantas G. 2014. Improved annotation of antibiotic resistance determinants reveals microbial resistomes cluster by ecology. ISME J 9:1–10.
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.
25. 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.
26. 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.
27. 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.
28. 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.
29. 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.
30. 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.
31. 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.
32. 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(2): e00108-13.
33. Dunivin TK, Shade A. 2018. Community structure explains antibiotic resistance gene dynamics over a temperature gradient in soil. FEMS Microbiol Ecol 94.

34. 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.
35. 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.
36. 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.
37. Sorensen JW, Dunivin TK, Tobin TC, Shade A. 2018. Ecological selection for small microbial genomes along a temperate-to-thermal soil gradient. Nat Microbiol 4.
38. Dunivin TK, Yeh SS, Shade A. 2018. Targeting microbial arsenic resistance genes: a new bioinformatic toolkit informs arsenic ecolog and evolution in soil genomes and metagenomes. bioRxiv.
39. 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.
40. 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.
41. 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.
42. 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.
43. White RA, Callister SJ, Moore RJ, Baker ES, Jansson JK. 2016. The past, present and future of microbiome analyses. Nat Protoc 11:2049–2053.
44. Stepanauskas R. 2015. Wiretapping into microbial interactions by single cell genomics. Front Microbiol 6:2014–2016.
45. R Core Team. 2017. R: A Language and Environment for Statistical Computing. Vienna, Austria.
46. Hartigan JA, Hartigan PM. 1985. The dip test of unimodality. Ann Stat 13:70–84.
47. 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.
48. 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)

462 and without (green) detected plasmids. **B)** Distribution of the number of plasmids per
463 RefSoil microorganism.
464
465 **Figure 2. Plasmid size distributions. A)** Histogram of plasmid size (kbp) from RefSoil
466 plasmids. **B)** RefSoil (blue) and RefSeq (gray) plasmid size distributions. **C**) Estimation of
467 plasmid size distribution in three soil orders. Color indicates soil order and n indicates the
468 community size.
469
470 **Figure 3. Relationship between plasmid size and genome size.** Total plasmid size (sum of all
471 plasmids in an microorganism, kbp) is plotted on a log scale against total genome size for
472 each RefSoil microorganism. Density plots are included for each axis to represent the
473 distribution of RefSoil microorganisms with different numbers of plasmids (none (green),
474 one (blue), or multiple (purple)).
475
476 **Figure 4. Distribution of ARGs in RefSoil genomes and plasmids. A)** The raw numbers and
477 **B**) proportions of ARGs on plasmids (light blue), genomes (green) or both (dark blue) in
478 RefSoil+ microorganisms by antibiotic resistance gene group. The number of genes
479 included in each group is shown in parentheses. **C**) The raw numbers and **D**) proportions
480 of detected ARGs on plasmids (light blue), genomes (green) or both (dark blue) in
481 RefSoil+ microorganisms by phylum-level taxonomy. The number of taxa included in
482 each phylum is shown in parentheses.
483
484 **Figure 5. Proportion of ARGs on genomes and plasmids in RefSoil+ and RefSeq databases.**

Number of ARGs was normalized to number 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.

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

Figure S2. Proportion of ARGs by classification in RefSoil and RefSeq databases. Boxplots of the proportion of ARGs per genetic element. Each ARG was normalized to the number of genetic elements in the database. Points are colored by ARG category. Kruskal-Wallis P values are shown and where applicable, significant Mann-Whitney U test results are shown.

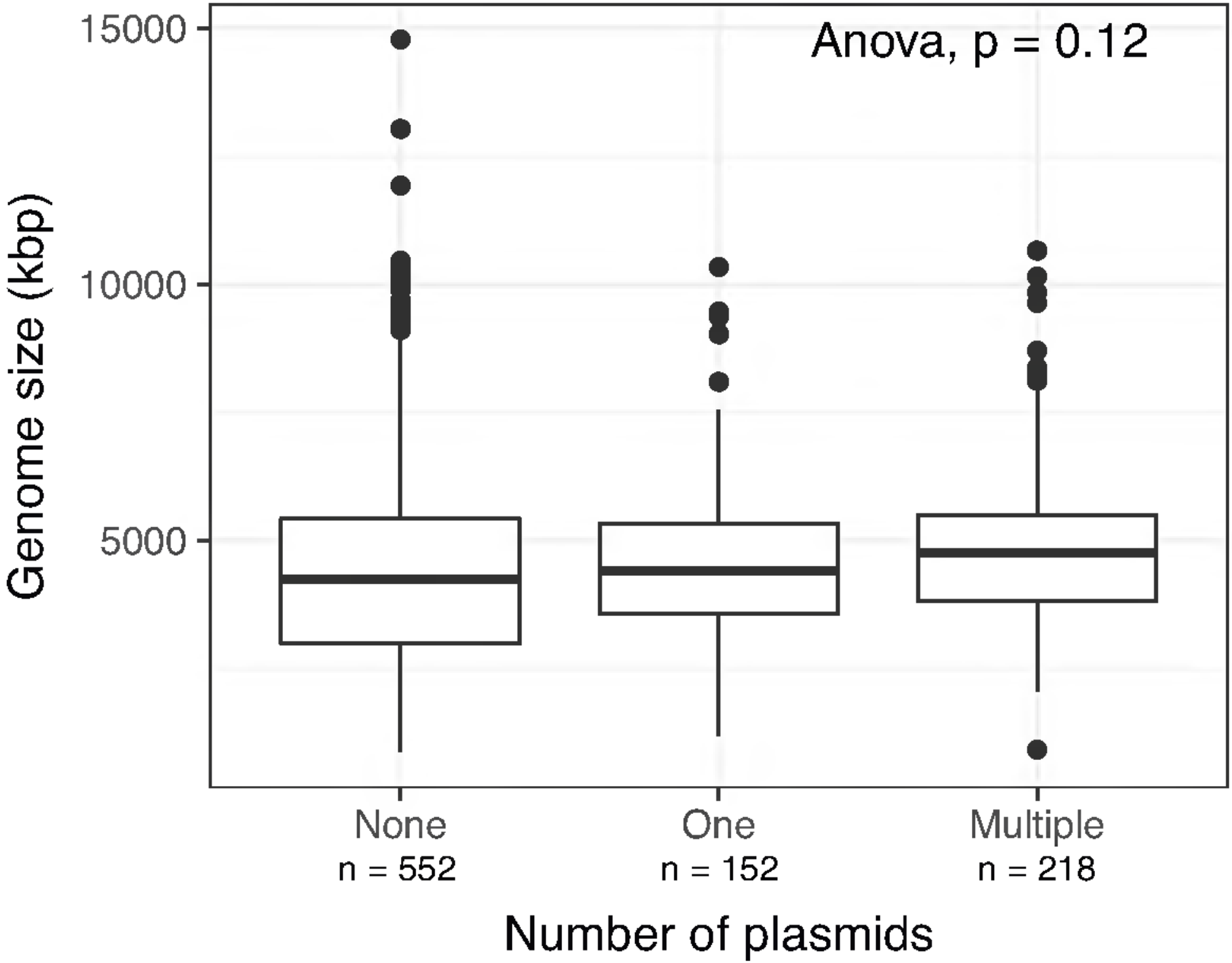
Figure S3. Proportion of ARGs on genomes and plasmids in RefSoil+ and RefSeq databases normalized to base pairs. Number of ARGs was normalized to total number of base pairs. Boxplots are colored by database. Points represent individual ARGs and are colored based on classification. Kruskal-Wallis test results are shown in addition to significant results from pairwise Mann Whitney U tests.

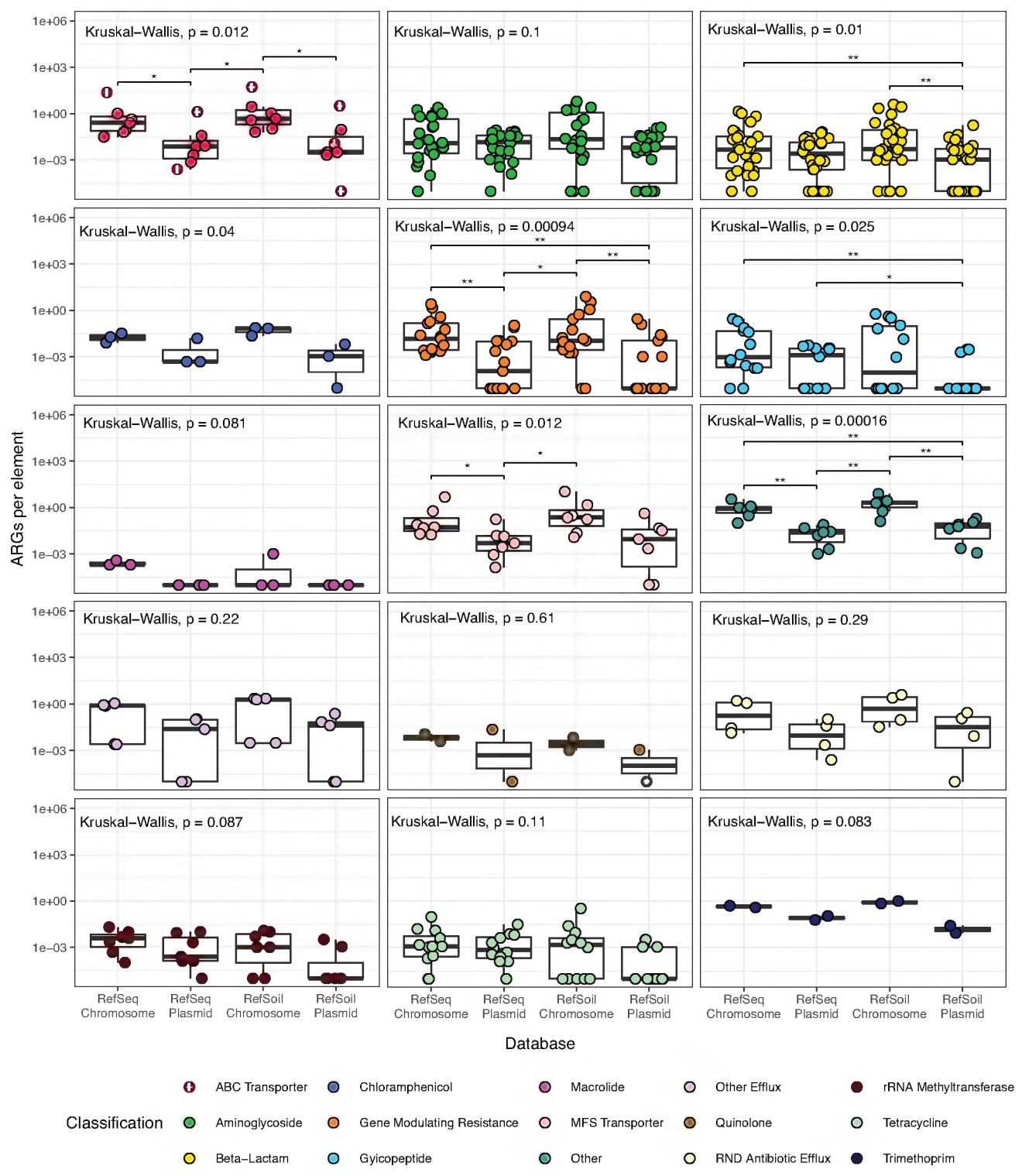
Table S1. Quality filtered antibiotic resistance gene (ARG) hits in RefSoil genomes and

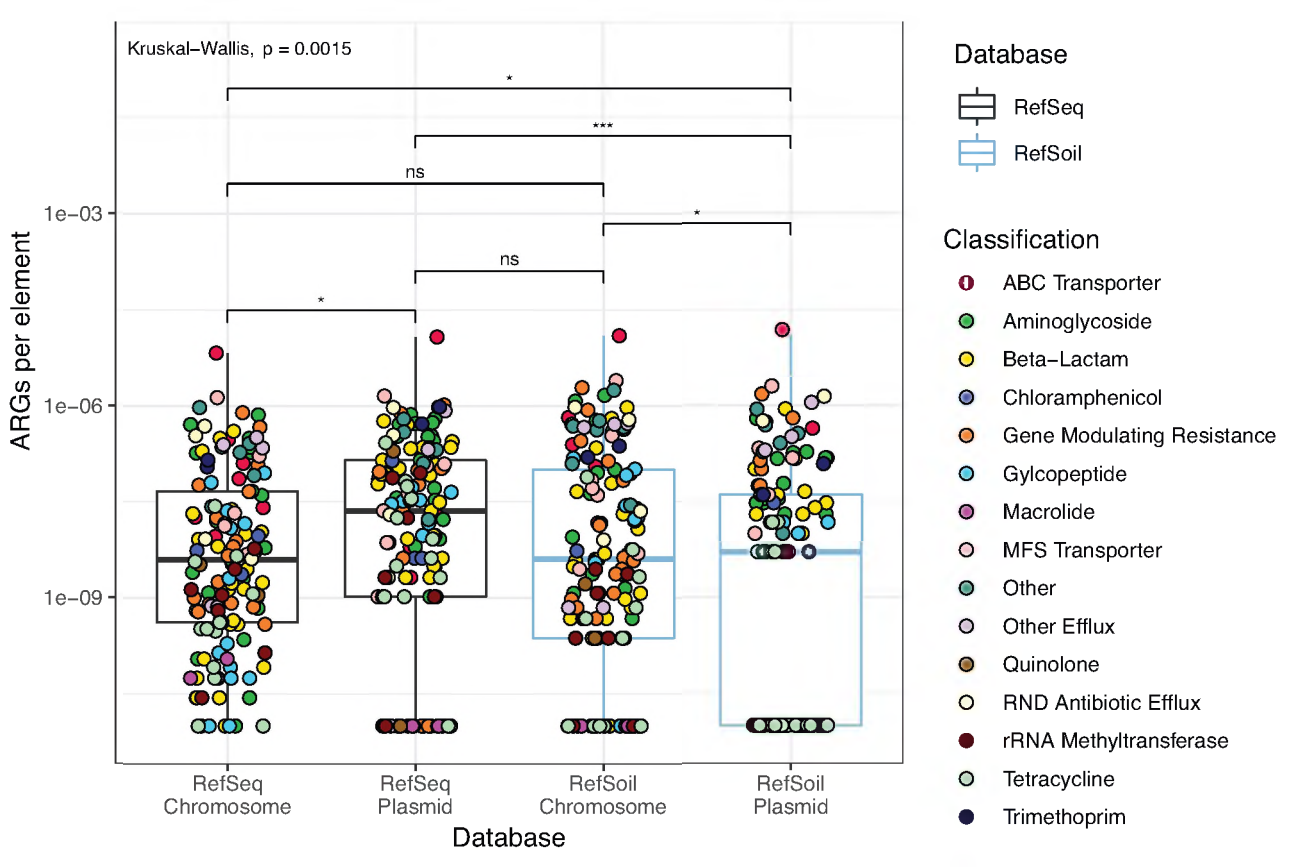
plasmids. Information on quality scores and accession numbers for each ARG hit.

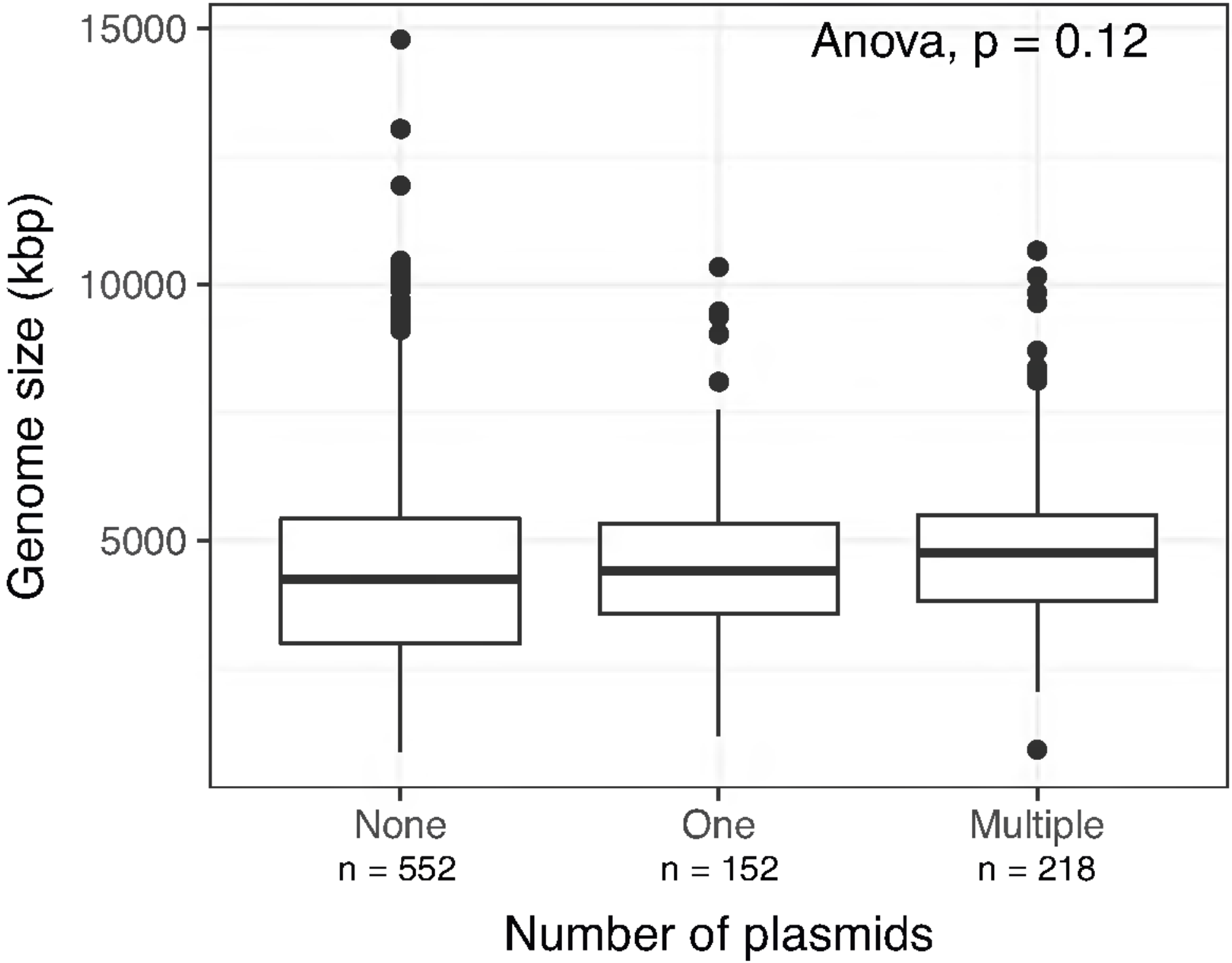
Table S2. RefSoil taxonomy table with plasmid and genome accession numbers.

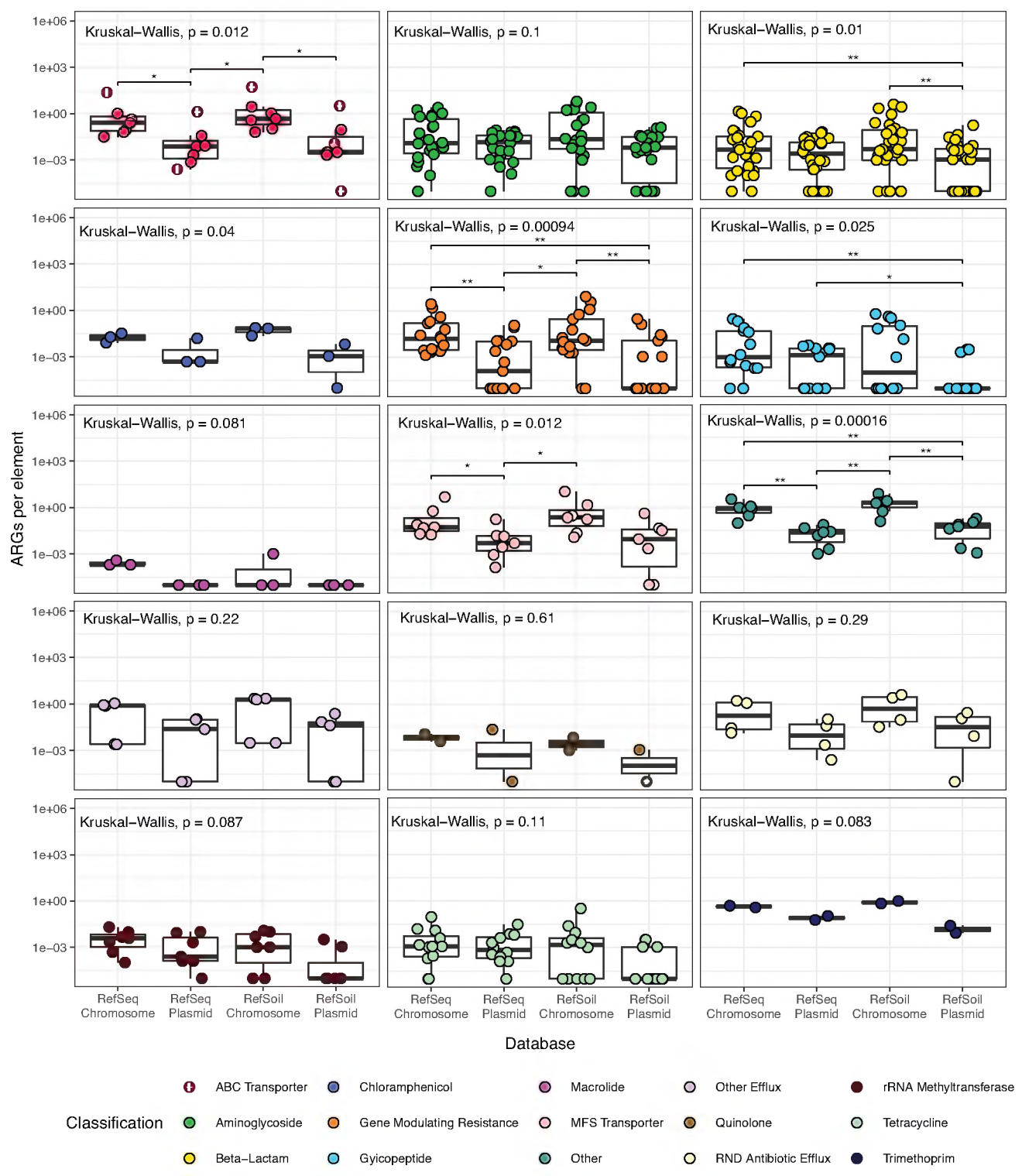
Table S3. ResFams HMMs and antibiotic classifications.

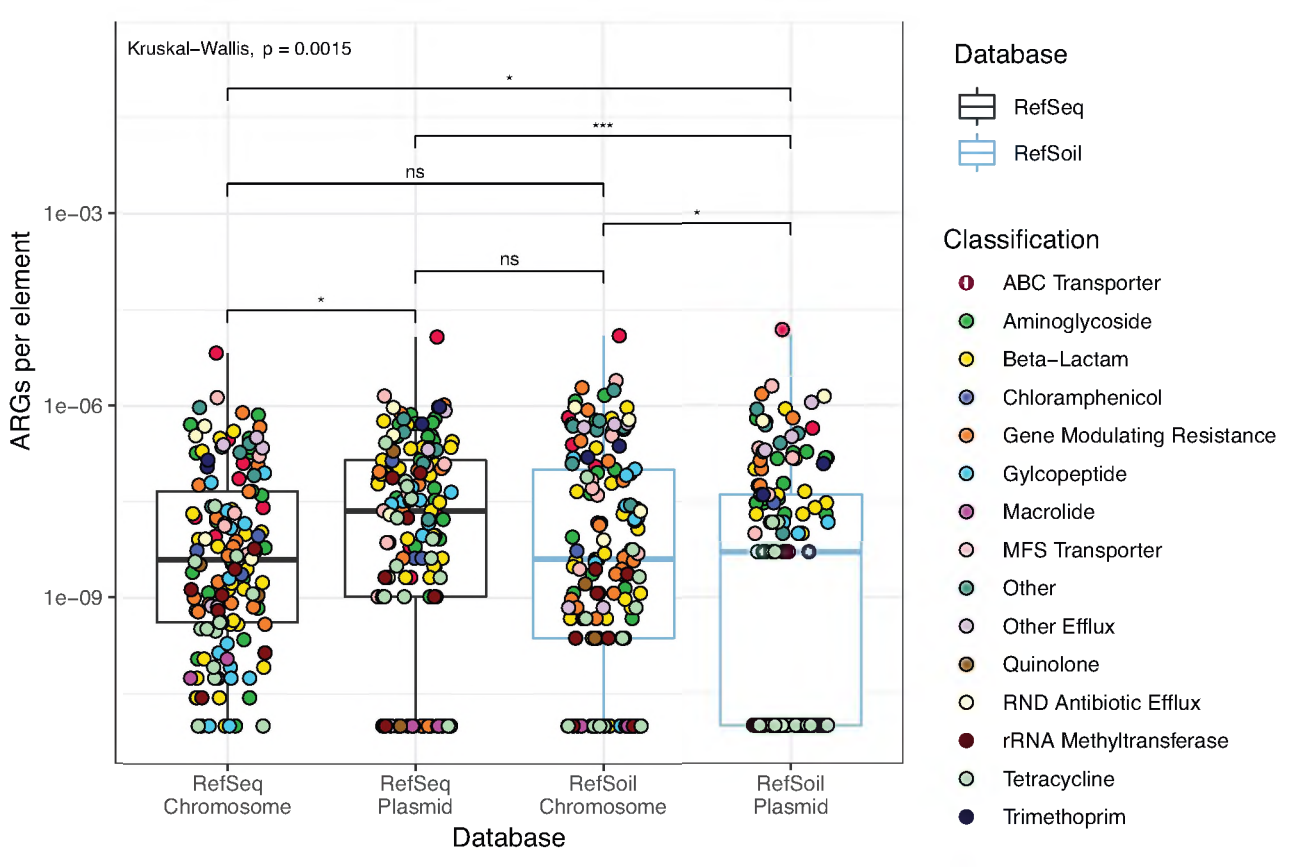

Anova, p = 0.12
15000
10000
5000
Genome size (kbp)
None
n = 552
One
n = 152
Multiple
n = 218
Number of plasmids


Kruskal–Wallis, p = 0.012
Kruskal–Wallis, p = 0.1
Kruskal–Wallis, p = 0.01
Kruskal–Wallis, p = 0.04
Kruskal–Wallis, p = 0.00094
Kruskal–Wallis, p = 0.025
Kruskal–Wallis, p = 0.081
Kruskal–Wallis, p = 0.012
Kruskal–Wallis, p = 0.00016
Kruskal–Wallis, p = 0.22
Kruskal–Wallis, p = 0.61
Kruskal–Wallis, p = 0.29
Kruskal–Wallis, p = 0.087
Kruskal–Wallis, p = 0.11
Kruskal–Wallis, p = 0.083
1e+06
1e+03
1e+00
1e−03
ARGs per element
RefSeq Chromosome
RefSeq Plasmid
RefSoil Chromosome
RefSoil Plasmid
Database
Classification
ABC Transporter
Aminoglycoside
Beta−Lactam
Chloramphenicol
Gene Modulating Resistance
Gyicopeptide
Macrolide
MFS Transporter
Other
Other Efflux
Quinolone
RND Antibiotic Efflux
rRNA Methyltransferase
Tetracycline
Trimethoprim


Kruskal–Wallis, p = 0.0015
*
***
ns
*
ns
*
1e−03
1e−06
1e−09
ARGs per element
RefSeq Chromosome
RefSeq Plasmid
RefSoil Chromosome
RefSoil Plasmid
Database
Database
RefSeq
RefSoil
Classification
ABC Transporter
Aminoglycoside
Beta−Lactam
Chloramphenicol
Gene Modulating Resistance
Gylcopeptide
Macrolide
MFS Transporter
Other
Other Efflux
Quinolone
RND Antibiotic Efflux
rRNA Methyltransferase
Tetracycline
Trimethoprim


Anova, p = 0.12
15000
10000
5000
Genome size (kbp)
None
n = 552
One
n = 152
Multiple
n = 218
Number of plasmids


Kruskal–Wallis, p = 0.012
Kruskal–Wallis, p = 0.1
Kruskal–Wallis, p = 0.01
Kruskal–Wallis, p = 0.04
Kruskal–Wallis, p = 0.00094
Kruskal–Wallis, p = 0.025
Kruskal–Wallis, p = 0.081
Kruskal–Wallis, p = 0.012
Kruskal–Wallis, p = 0.00016
Kruskal–Wallis, p = 0.22
Kruskal–Wallis, p = 0.61
Kruskal–Wallis, p = 0.29
Kruskal–Wallis, p = 0.087
Kruskal–Wallis, p = 0.11
Kruskal–Wallis, p = 0.083
1e+06
1e+03
1e+00
1e−03
ARGs per element
RefSeq Chromosome
RefSeq Plasmid
RefSoil Chromosome
RefSoil Plasmid
Database
Classification
ABC Transporter
Aminoglycoside
Beta−Lactam
Chloramphenicol
Gene Modulating Resistance
Gyicopeptide
Macrolide
MFS Transporter
Other
Other Efflux
Quinolone
RND Antibiotic Efflux
rRNA Methyltransferase
Tetracycline
Trimethoprim


Kruskal–Wallis, p = 0.0015
*
***
ns
*
ns
*
1e−03
1e−06
1e−09
ARGs per element
RefSeq Chromosome
RefSeq Plasmid
RefSoil Chromosome
RefSoil Plasmid
Database
Database
RefSeq
RefSoil
Classification
ABC Transporter
Aminoglycoside
Beta−Lactam
Chloramphenicol
Gene Modulating Resistance
Gylcopeptide
Macrolide
MFS Transporter
Other
Other Efflux
Quinolone
RND Antibiotic Efflux
rRNA Methyltransferase
Tetracycline
Trimethoprim