doi:10.1038/nature13377

Bacterial phylogeny structures soil resistomes across habitats

Kevin J. Forsberg[1]*, Sanket Patel[1,2]*, Molly K. Gibson[1], Christian L. Lauber[3], Rob Knight[4,5], Noah Fierer[3,6] & Gautam Dantas[1,2,7]

Ancient and diverse antibiotic resistance genes (ARGs) have previously been identified from soil[1–3], including genes identical to those in human pathogens[4]. Despite the apparent overlap between soil and clinical resistomes[4–6], factors influencing ARG composition in soil and their movement between genomes and habitats remain largely unknown[3]. General metagenome functions often correlate with the underlying structure of bacterial communities[7–12]. However, ARGs are proposed to be highly mobile[4,5,13], prompting speculation that resistomes may not correlate with phylogenetic signatures or ecological divisions[13,14]. To investigate these relationships, we performed functional metagenomic selections for resistance to 18 antibiotics from 18 agricultural and grassland soils. The 2,895 ARGs we discovered were mostly new, and represent all major resistance mechanisms[15]. We demonstrate that distinct soil types harbour distinct resistomes, and that the addition of nitrogen fertilizer strongly influenced soil ARG content. Resistome composition also correlated with microbial phylogenetic and taxonomic structure, both across and within soil types. Consistent with this strong correlation, mobility elements (genes responsible for horizontal gene transfer between bacteria such as transposases and integrases) syntenic with ARGs were rare in soil by comparison with sequenced pathogens, suggesting that ARGs may not transfer between soil bacteria as readily as is observed between human pathogens. Together, our results indicate that bacterial community composition is the primary determinant of soil ARG content, challenging previous hypotheses that horizontal gene transfer effectively decouples resistomes from phylogeny[13,14].

Functional metagenomic selections permit deep interrogation of resistomes and can identify full-length, functionally verified ARGs without requiring sequence similarity to previously identified genes[2–4,16]. We constructed metagenomic libraries averaging 13.8 ± 8.3 (mean ± s.d.) gigabases (Gb) by shotgun cloning DNA fragments 1–5 kilobases (kb) long from 18 soils (Supplementary Table 1) into *Escherichia coli*, and screened these libraries for resistance against 18 antibiotics representing 8 drug classes. Resistance was conferred against 15 of the 18 antibiotics tested (Extended Data Fig. 1, Supplementary Table 2), and DNA fragments conferring resistance were sequenced, assembled, and annotated with PARFuMS[4] (see Methods).

We assembled 4,655 contigs over 500 base pairs (bp) in length (Fig. 1a, N50 size = 2.25 kb, or average contig length >1.76 kb) containing 8,882 open reading frames (ORFs) larger than 350 bp (Supplementary Data 1). Using profile Hidden Markov Models (HMMs), we annotated 2,895 of these 8,882 ORFs as ARGs (see Methods). Underscoring the immense functional diversity of soil resistomes, the identified soil ARGs were largely dissimilar from ARGs in public repositories (Fig. 1b). Only 15 soil ARGs (0.5%) have perfect amino acid identity to entries in the NCBI protein database, with just three having >99% nucleotide identity to nucleotide sequences in NCBI. In contrast, the average amino acid identity of all ARGs to their closest homologue in NCBI is only 61.1 ± 15.3%. Although we recently described cultured soil bacteria harbouring ARGs with perfect nucleotide identity to those in human pathogens[4], this phenomenon appears to be the exception rather than the rule: only one soil ARG from our current data set shares perfect nucleotide identity with a pathogen (NCBI accession number AY664504).

Our sampling depth (Extended Data Fig. 1) surpasses previous functional interrogations of soil metagenomes[8–10,16,17], permitting an unparalleled comparison of ARGs across soil types. Emphasizing the diversity recovered by our functional selections, 29% of assembled contigs over 1,500 bp did not contain an ORF that could be confidently assigned a known resistance function, representing a large repertoire of potentially novel ARGs. The ORFs assigned to known ARG functions represent all classical mechanisms of antibiotic resistance (Fig. 1c): antibiotic efflux, antibiotic inactivation and target protection/redundancy[15].

Resistance to amphenicol and tetracycline antibiotics occurred predominantly via the action of drug transporters, of which the majority belonged to the major facilitator superfamily (Fig. 1c). In contrast, selections with aminoglycoside and β-lactam antibiotics most frequently uncovered ARGs with antibiotic-inactivating capabilities, via covalent modification of aminoglycosides and enzymatic degradation of β-lactams (Fig. 1c). Excluding selections with trimethoprim and D-cycloserine (for which the ARGs selected were predominantly overexpressed target alleles from diverse bacterial lineages[16], Extended Data Fig. 2), β-lactamases were the most frequently encountered soil ARGs, mirroring observations from hospital settings[18]. We observed metallo-β-lactamases most frequently, followed by Ambler class-A and class-D enzymes (Extended Data Fig. 3).

We predicted the source phylum of each functionally selected resistance contig greater than 500 bp using a composition-based, semi-supervised, taxonomic binning algorithm[19]. Proteobacteria and Actinobacteria were the most prevalent predicted phyla, and each contained ARG families of all major resistance mechanisms[15] (Fig. 2a). Major facilitator superfamily transporters and β-lactamases showed the strongest, and most orthogonal, relationships with predicted bacterial phyla (Fig. 2b, Extended Data Fig. 4, Supplementary Table 3). β-lactamases were enriched within Verrucomicrobia, Acidobacteria and Cyanobacteria contigs, while major facilitator superfamily transporters were largely absent from Acidobacteria and were enriched among Actinobacteria and Proteobacteria contigs (Supplementary Table 3).

To quantitatively compare soil resistomes at higher resolution, a count matrix of unique gene sequences per functional annotation (that is, ARG family) was generated by summing across all antibiotic selections per soil and normalizing these counts to metagenomic library size (see Methods). The number of unique ARGs was significantly higher ($P < 0.01$, Wilcoxon rank sum test) in Cedar Creek (CC) grassland soils compared to agricultural soils from Kellogg Biological Station (KBS). Our selections resolved functional differences between CC and KBS soils regardless of whether Bray–Curtis distances were calculated using only ARG families (Fig. 3a) or all captured gene functions (Extended Data Fig. 5). Major facilitator superfamily transporters and β-lactamases were higher

[1]Center for Genome Sciences and Systems Biology, Washington University School of Medicine, St Louis, Missouri 63108, USA. [2]Department of Pathology and Immunology, Washington University School of Medicine, St Louis, Missouri 63110, USA. [3]Cooperative Institute for Research in Environmental Sciences, University of Colorado, Boulder, Colorado 80309, USA. [4]Department of Chemistry and Biochemistry and BioFrontiers Institute, University of Colorado, Boulder, Colorado 80309, USA. [5]Howard Hughes Medical Institute, Boulder, Colorado 80309, USA. [6]Department of Ecology and Evolutionary Biology, University of Colorado, Boulder, Colorado 80309, USA. [7]Department of Biomedical Engineering, Washington University, St Louis, Missouri 63130, USA.

*These authors contributed equally to this work.

©2014 Macmillan Publishers Limited. All rights reserved

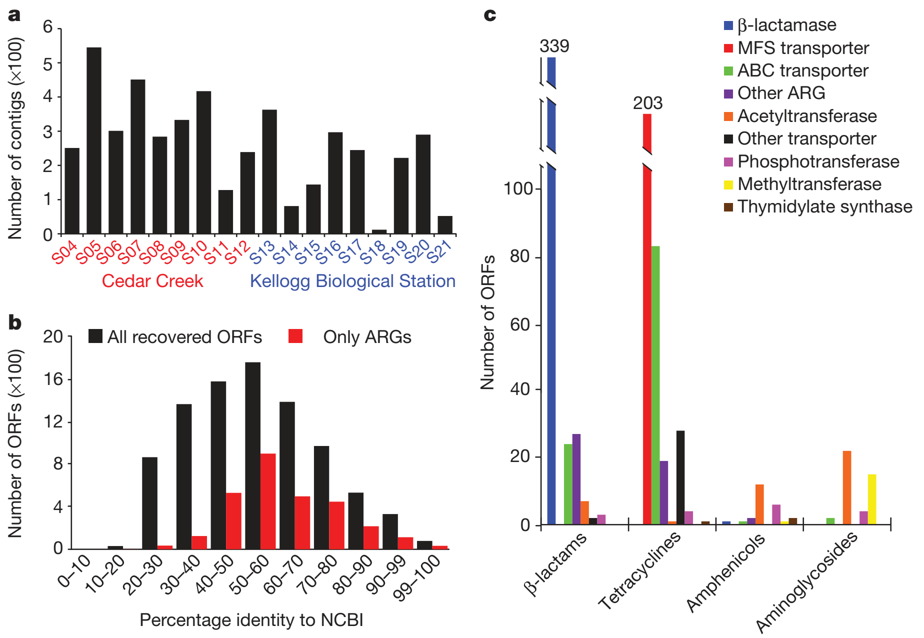


Figure 1 | Functional selections of 18 soil libraries yield diverse ARGs. **a**, Bar chart depicting contigs >500 bp across all antibiotic selections from CC (red) and KBS (blue) libraries. **b**, Amino acid identity between all ORFs (black, $n = 8{,}882$) or ARGs only (red, $n = 2{,}895$) and their top hit in the NCBI protein database. **c**, Total ARGs by antibiotic class (ARG types in the key); the y axis shows number of ORFs. MFS, major facilitator superfamily; ABC, ATP-binding cassette.

at CC compared to KBS, and ARG families of these resistance mechanisms most significantly differed between these soils (Supplementary Table 4). Only 2.6% of ARGs were shared across at least two soils at ≥99% nucleotide identity (Supplementary Table 5), with significantly more inter-soil sharing at CC versus KBS ($P < 0.05$, Fisher's exact test). No ARGs were shared between CC and KBS soils (≥99% nucleotide identity), and only two ARG mechanisms were observed in every soil (β-lactamase, major facilitator superfamily transporter).

We sampled CC soils across a gradient of added nitrogen (N) fertilizer. Similar to phylogenetic differences observed in community composition across the N gradient[20], we found that ARG composition of soils receiving higher N levels differed from the composition observed in other CC soils (Fig. 3b). These differences do not arise from a change in the number of unique ARGs between high-N and other soils ($P = 0.9$, Wilcoxon rank sum test), but rather were due to differences in relative proportions of ARG families in these soils (Supplementary Table 6). In high-N soils, β-lactamases were depleted whereas membrane transporters were enriched (Supplementary Table 6). High N levels favour particular organisms (for example, copiotrophs[8,10,20]), causing shifts in bacterial abundances, which in turn probably affect resistome composition.

We calculated differences in community structure of these CC and KBS soils using 16S ribosomal RNA gene sequences[8] (Extended Data Fig. 6). All bacterial phyla that were abundant (>3% relative abundance) in the samples, as determined by 16S rRNA gene sequencing, were also well-represented (>4% relative abundance) among phyla inferred from resistance-conferring contigs (Extended Data Fig. 7). Actinobacteria (which are characterized by GC-rich genomes and produce antibiotics in the soil[21]) were most enriched in resistance-conferring contigs relative

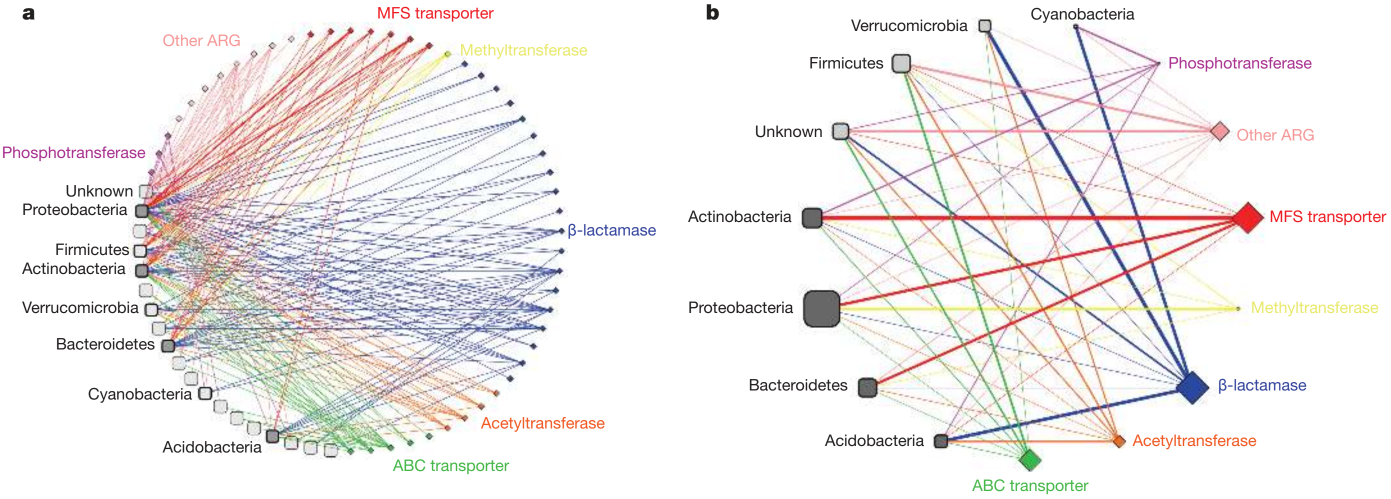


Figure 2 | Resistance is encoded by diverse soil phyla. **a**, Network of predicted bacterial phyla for each ARG used in cross-soil comparisons ($n = 880$). Edge thickness indicates number of ARGs within an ARG family (diamonds) from a predicted phylum (rounded squares). Phyla containing >15 ARGs are labelled, and are shaded dark grey at >3% 16S rRNA abundance. **b**, Simplified network of general ARG mechanisms; edge thickness represents significance of phylum and ARG mechanism co-occurrence (Fisher's exact test, line width increases with ranked significance). Node size indicates number of ARGs (diamonds) or contig count (rounded squares).

©2014 Macmillan Publishers Limited. All rights reserved

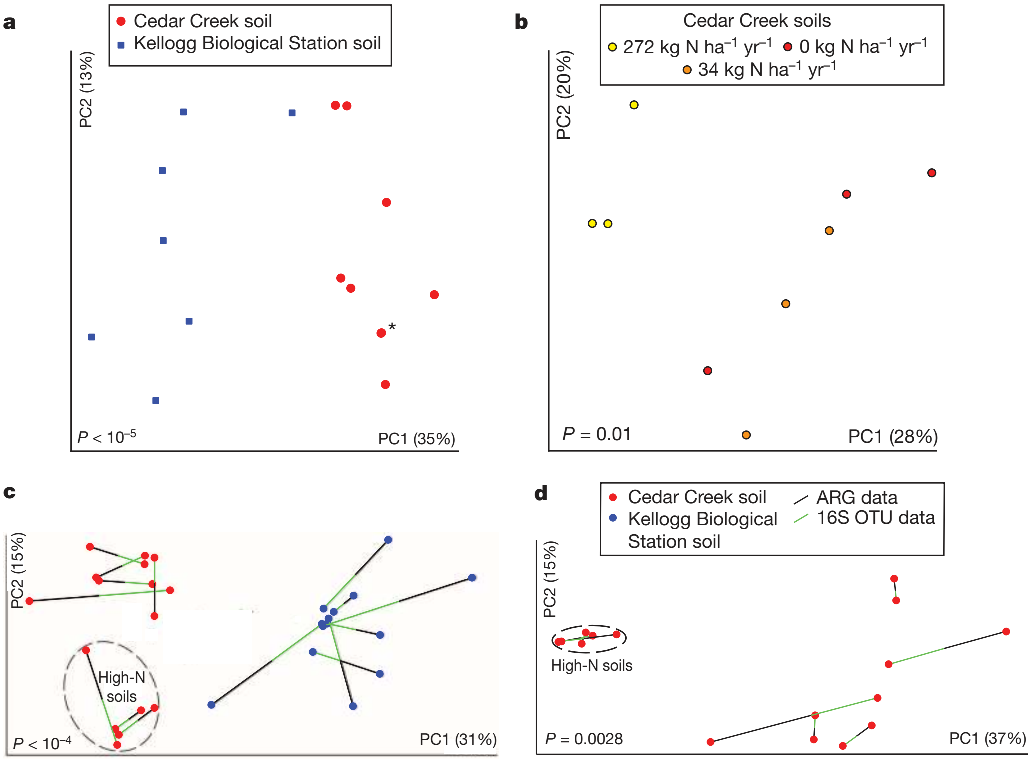


Figure 3 | Resistomes correlate with phylogeny across soil type and nitrogen amendment. **a**, **b**, Principal coordinate analysis (PCoA) plots depict Bray–Curtis distances between soils, using unique ARG counts. **a**, Resistomes from CC (red) and KBS (blue) soils cluster separately ($P < 10^{-5}$, analysis of similarity, ANOSIM). An asterisk denotes two soils with near-identical coordinates. **b**, CC soils amended with high N-levels cluster separately from other CC soils ($P = 0.01$, ANOSIM). **c**, **d**, Procrustes analyses depict significant correlation between ARG content (Bray–Curtis) and bacterial composition (Bray–Curtis) for CC (red) and KBS (blue) soils (**c**) and only CC soils (**d**). OTU, operational taxonomic unit.

to 16S rRNA gene abundance, whereas levels of Proteobacteria were similar in both data sets (Extended Data Fig. 7). Thus, any phylogenetic bias in functional selections due to heterologous expression in *E. coli* (a member of Proteobacteria) is minimal compared to the effect of the ARG content of source bacteria.

We next tested for correlations between soil resistomes and community composition. When both CC and KBS soils were considered, Bray–Curtis distances calculated from normalized ARG counts significantly correlated with bacterial operational taxonomic units inferred from 16S rRNA sequence data, whether taxonomic (Bray–Curtis) or phylogenetic (weighted and unweighted) dissimilarity metrics were used (Mantel tests, $P < 0.05$, Supplementary Table 7). Visualized by Procrustes analyses, both the ARG content and bacterial composition of CC and KBS soils clustered by sampling site, consistently displaying highly significant goodness-of-fit measures (Fig. 3c, Extended Data Fig. 8, Supplementary Table 8). Within sampling sites, the variability in phylogenetic community composition differed (Supplementary Table 9): more diversity was observed across CC soils than in KBS soils (Extended Data Fig. 6). Because of this disparity, we observed a significant within-site correlation between ARG content and community composition in CC soils (Supplementary Tables 7 and 8; Fig. 3d, Extended Data Fig. 8), but not in KBS soils (Extended Data Fig. 9).

The strong correlation between soil ARG content and bacterial composition suggests that the horizontal gene transfer (HGT) of ARGs is not sufficiently frequent to obscure their association with bacterial genomes. Corroborating this notion, soil ARGs show limited genetic potential for horizontal exchange. Only 0.42% of ORFs from our functional selections were predicted mobility elements (such as transposases, integrases and recombinases; Extended Data Fig. 10), and none of these genes were co-localized with an ARG containing >72% amino acid identity to a protein in NCBI. The limited mobility of the soil resistome may explain why ARGs are rarely shared between soil and human pathogens[4,22]. In contrast to soils, ARGs in pathogens often share near-perfect sequence identity[18], with origins that can be traced to the emergence of a single genotype disseminated broadly via HGT[23,24].

To test the hypothesis that ARGs in the soil have less potential for HGT than those in human pathogens, we compared ARGs from our functional selections to ARGs in fully sequenced genomes from 433 common human pathogens and 153 non-pathogenic soil organisms[13] (Supplementary Data 2). We modelled functional selections from each genome collection based on the fragment-size distribution observed in our soil selections (see Methods), and calculated the proportion of DNA fragments from each simulation that contained a predicted mobility element. Signatures of HGT were significantly more frequent in pathogen genomes than in soil genomes or soil selections (Fig. 4a). Importantly, we detected no difference in the HGT potential of ARGs between the two soil data sets (Fig. 4a), supporting the generality of the conclusions drawn from our soil functional selections. As the genetic distance from an ARG increased, the incidence of mobility elements in pathogen genomes was always higher than in soil genomes or functionally selected metagenomes (Fig. 4b), indicating that the higher potential for HGT seen in pathogens is independent of DNA fragment size or the method by which soil resistance is interrogated. Interestingly, enriching for multidrug-resistant Proteobacteria in the soil[4,25] (which are frequently encountered as opportunistic pathogens in hospital settings[26]) increases the detection of shared resistance between soil and clinic[4], suggesting that they may represent a major conduit through which ARGs move between these environments.

Unlike most hospital settings, soils contain a huge diversity of ARGs[1–3,16,22], and therefore increasing antibiotic exposure (as has occurred over the past 70 years[27]) may favour pre-existing genotypes[1] rather than the acquisition of new ARGs[4]. This key distinction explains our observation that, despite extensive sampling, very little evidence exists for HGT of ARGs across soil communities. Indeed, our evidence points to phylogeny, rather

©2014 Macmillan Publishers Limited. All rights reserved

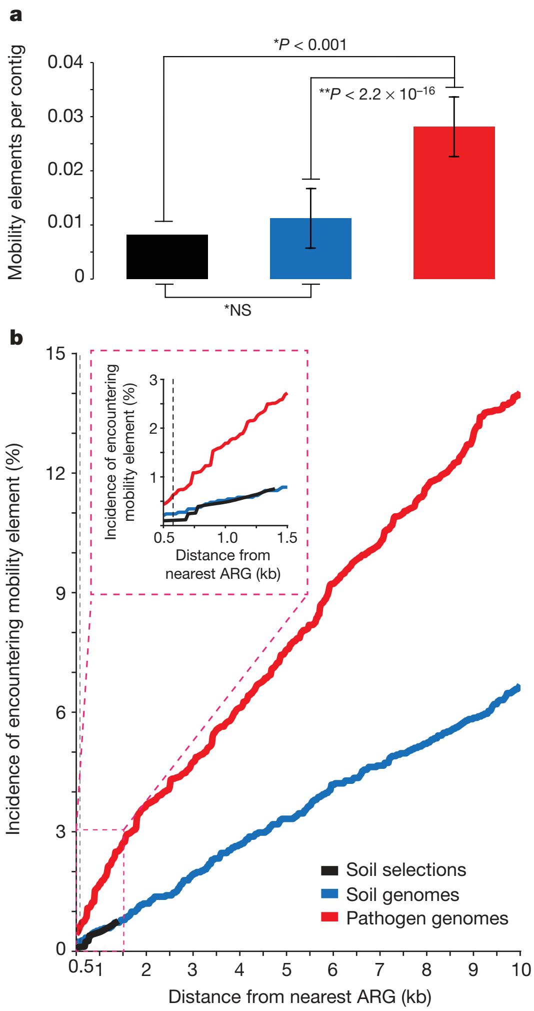


Figure 4 | Pathogen ARGs show higher HGT potential than soil ARGs. **a**, Mobility elements syntenic with ARGs are proportionally higher in pathogens ($n = 433$) than soil genomes ($n = 153$) or soil functional selections. *Significance determined from 1,000 Monte Carlo simulations; **significance determined by Student's t-test. NS, not significant. Error bars depict two standard deviations from the mean. **b**, Pathogens show significantly increased HGT potential relative to soil genomes and soil selections at all distances (20-bp intervals) greater than 580 bp (dashed line, $P < 0.05$, Fisher's exact test). The inset depicts mobility elements encountered within 1.5 kb of ARGs, demonstrating that data from soil selections resemble soil genomes.

than HGT[13,14], as the primary determinant of soil resistome content. Therefore, as bacterial type and diversity change across soils[8–10,20], so too do their associated ARGs, resulting in resistomes that may respond to anthropogenic modulations (for example, nitrogen fertilizer) that do not possess obvious antibiotic-related properties.

METHODS SUMMARY

Metagenomic DNA was extracted from all soils using the PowerMax soil DNA isolation kit (MoBio Laboratories). Small-insert metagenomic libraries were created by shearing this DNA into 1–5-kb fragments before ligation into the pZE21 expression vector[28] and electroporation into *E. coli* MegaX cells (Invitrogen). These libraries were plated on Mueller–Hinton agar containing one of 18 antibiotics, grown overnight at 37 °C, and resistant colonies collected in a liquid cell slurry. Metagenomic fragments conferring resistance were amplified via polymerase chain reaction (PCR) using this slurry as template, and barcoded libraries prepared from these amplicons for 101-bp paired-end sequencing using the Illumina HiSeq2000. Paired-end reads were assembled into resistance-conferring fragments using PARFuMs[4] and annotated using profile HMM databases. Cross-soil resistome comparisons were performed using Bray–Curtis distances calculated from a count matrix of unique gene sequences per ARG family, generated by summing across all antibiotic selections for a given soil and normalizing these counts to metagenomic library size. Taxonomies of assembled fragments were assigned using RAIphy[19], while the 16S rRNA gene sequence data generated previously were processed as described[8]. Genomes of human pathogens and non-pathogenic soil bacteria were collected based on published environmental metadata[13]. Functional selections were modelled from each genome collection based on the fragment-size distribution observed in our soil selections.

Online Content Any additional Methods, Extended Data display items and Source Data are available in the online version of the paper; references unique to these sections appear only in the online paper.

Received 19 February; accepted 14 April 2014.
Published online 21 May 2014.

1. D'Costa, V. M. *et al.* Antibiotic resistance is ancient. *Nature* **477,** 457–461 (2011).
2. Allen, H. K., Moe, L. A., Rodbumrer, J., Gaarder, A. & Handelsman, J. Functional metagenomics reveals diverse beta-lactamases in a remote Alaskan soil. *ISME J.* **3,** 243–251 (2009).
3. Allen, H. K. *et al.* Call of the wild: antibiotic resistance genes in natural environments. *Nature Rev. Microbiol.* **8,** 251–259 (2010).
4. Forsberg, K. J. *et al.* The shared antibiotic resistome of soil bacteria and human pathogens. *Science* **337,** 1107–1111 (2012).
5. Wright, G. D. Antibiotic resistance in the environment: a link to the clinic? *Curr. Opin. Microbiol.* **13,** 589–594 (2010).
6. Benveniste, R. & Davies, J. Aminoglycoside antibiotic-inactivating enzymes in actinomycetes similar to those present in clinical isolates of antibiotic-resistant bacteria. *Proc. Natl Acad. Sci. USA* **70,** 2276–2280 (1973).
7. Langille, M. G. *et al.* Predictive functional profiling of microbial communities using 16S rRNA marker gene sequences. *Nature Biotechnol.* **31,** 814–821 (2013).
8. Fierer, N. *et al.* Comparative metagenomic, phylogenetic and physiological analyses of soil microbial communities across nitrogen gradients. *ISME J.* **6,** 1007–1017 (2012).
9. Fierer, N. *et al.* Cross-biome metagenomic analyses of soil microbial communities and their functional attributes. *Proc. Natl Acad. Sci. USA* **109,** 21390–21395 (2012).
10. Fierer, N. *et al.* Reconstructing the microbial diversity and function of pre-agricultural tallgrass prairie soils in the United States. *Science* **342,** 621–624 (2013).
11. Muegge, B. D. *et al.* Diet drives convergence in gut microbiome functions across mammalian phylogeny and within humans. *Science* **332,** 970–974 (2011).
12. Zaneveld, J. R., Lozupone, C., Gordon, J. I. & Knight, R. Ribosomal RNA diversity predicts genome diversity in gut bacteria and their relatives. *Nucleic Acids Res.* **38,** 3869–3879 (2010).
13. Smillie, C. S. *et al.* Ecology drives a global network of gene exchange connecting the human microbiome. *Nature* **480,** 241–244 (2011).
14. Stokes, H. W. & Gillings, M. R. Gene flow, mobile genetic elements and the recruitment of antibiotic resistance genes into Gram-negative pathogens. *FEMS Microbiol. Rev.* **35,** 790–819 (2011).
15. Walsh, C. Molecular mechanisms that confer antibacterial drug resistance. *Nature* **406,** 775–781 (2000).
16. Pehrsson, E. C., Forsberg, K. J., Gibson, M. K., Ahmadi, S. & Dantas, G. Novel resistance functions uncovered using functional metagenomic investigations of resistance reservoirs. *Front. Microbiol.* **4,** 145 (2013).
17. Delmont, T. O. *et al.* Structure, fluctuation and magnitude of a natural grassland soil metagenome. *ISME J.* **6,** 1677–1687 (2012).
18. Jacoby, G. A. & Munoz-Price, L. S. The new beta-lactamases. *N. Engl. J. Med.* **352,** 380–391 (2005).
19. Nalbantoglu, O. U., Way, S. F., Hinrichs, S. H. & Sayood, K. RAIphy: phylogenetic classification of metagenomics samples using iterative refinement of relative abundance index profiles. *BMC Bioinform.* **12,** 41 (2011).
20. Ramirez, K. S., Lauber, C. L., Knight, R., Bradford, M. A. & Fierer, N. Consistent effects of nitrogen fertilization on soil bacterial communities in contrasting systems. *Ecology* **91,** 3463–3470; discussion 3503–3414 (2010).
21. Ventura, M. *et al.* Genomics of Actinobacteria: tracing the evolutionary history of an ancient phylum. *Microbiol. Mol. Biol. Rev.* **71,** 495–548 (2007).
22. Aminov, R. I. & Mackie, R. I. Evolution and ecology of antibiotic resistance genes. *FEMS Microbiol. Lett.* **271,** 147–161 (2007).
23. Davies, J. & Davies, D. Origins and evolution of antibiotic resistance. *Microbiol. Mol. Biol. Rev.* **74,** 417–433 (2010).
24. Medeiros, A. A. Evolution and dissemination of beta-lactamases accelerated by generations of beta-lactam antibiotics. *Clin. Inf. Diseases* **24** (Suppl. 1), 19–45 (1997).
25. Dantas, G., Sommer, M. O., Oluwasegun, R. D. & Church, G. M. Bacteria subsisting on antibiotics. *Science* **320,** 100–103 (2008).
26. Boucher, H. W. *et al.* Bad bugs, no drugs: no ESKAPE! *Clin. Inf. Diseases* **48,** 1–12 (2009).
27. Knapp, C. W., Dolfing, J., Ehlert, P. A. & Graham, D. W. Evidence of increasing antibiotic resistance gene abundances in archived soils since 1940. *Environ. Sci. Technol.* **44,** 580–587 (2010).
28. Lutz, R. & Bujard, H. Independent and tight regulation of transcriptional units in *Escherichia coli* via the LacR/O, the TetR/O and AraC/I1–I2 regulatory elements. *Nucleic Acids Res.* **25,** 1203–1210 (1997).

©2014 Macmillan Publishers Limited. All rights reserved

Supplementary Information is available in the online version of the paper.

Acknowledgements We thank M. Pesesky for access to and assistance with the data set used to benchmark RAlphy's performance, B. Wang for suggested improvements to Illumina library preparation, the Genome Technology Access Center at Washington University in St Louis for generating Illumina sequence data, M. Sherman for discussions on modelling pathogen HGT potential, and members of the Dantas laboratory for discussions on the results and analyses presented here. This work was supported by awards to G.D. through the Children's Discovery Institute (MD-II-2011-117), the International Center for Advanced Renewable Energy and Sustainability at Washington University, the National Academies Keck Futures Initiatives (Synthetic Biology, SB2), and the NIH Director's New Innovator Award (DP2-DK-098089). The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health. M.K.G. is supported by a Mr and Mrs Spencer T. Olin Fellowship for Women in Graduate Study at Washington University. K.J.F. received support from the NIGMS Cell and Molecular Biology Training Grant (GM 007067) and from the NHGRI Genome Analysis Training Program (T32 HG000045). K.J.F. and M.K.G. are NSF graduate research fellows (award number DGE-1143954).

Author Contributions N.F., C.L.L. and R.K. provided soils and 16S rRNA gene sequencing data; G.D. conceived the functional selections; S.P. created metagenomic libraries, performed functional selections, and prepared sequencing libraries; K.J.F. assembled sequence data from functional selections and annotated ARGs with assistance from M.K.G.; M.K.G. built the custom ARG profile HMM database; K.J.F. performed genomic and ecological analyses and wrote the manuscript with contributions from M.K.G., R.K., N.F. and G.D.

Author Information All assembled sequences have been deposited to Genbank with accession numbers KJ691878–KJ696532 and raw reads to SRA under the accession number SRP041174. Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to G.D. (dantas@wustl.edu).

©2014 Macmillan Publishers Limited. All rights reserved

METHODS

Construction of soil metagenomic libraries. For construction of soil metagenomic libraries, bulk community DNA was extracted from 10 g of each soil using the PowerMax Soil DNA Isolation Kit (MoBio Laboratories), following suggested protocols (http://www.mobio.com/images/custom/file/protocol/12988-10.pdf). Subsequently, DNA was sheared to a size range of approximately 500–5,000 bp using the Covaris E210 sonicator with the manufacturer's recommended settings (http://covarisinc.com/wp-content/uploads/pn_400069.pdf). Sheared DNA was size-selected by electrophoresis through a 1% low-melting point agarose gel in 0.5X Tris-Borate-EDTA (TBE) buffer stained with GelGreen dye (Biotium). A gel slice corresponding to 1,000–5,000 bp was excised from the gel and DNA was extracted using a QIAquick Gel Extraction Kit, eluting in 30 μl of warm nuclease-free H_2O (Qiagen). We chose this fragment size range because small fragment libraries, although they sacrifice the ability to capture large gene clusters or very large genes, typically contain many more unique clones and therefore provide significantly more sampling depth than do large-insert libraries. Given the tremendous genetic diversity in soil, and the fact that resistance is often (though not always) encoded by single genes, we favoured fragment sizes that typically encode one to three bacterial genes. Purified DNA was then end-repaired using the End-It DNA End Repair kit (Epicentre) with the following protocol.

(1) For each volume of 30 μl QIAquick eluate, add the following: 5 μl dNTP mix (2.5 mM), 5 μl ATP (10 mM), 5 μl 10X End-Repair Buffer, 1 μl End-Repair Enzyme Mix and 4 μl nuclease-free H_2O to a final volume of 50 μl.
(2) Mix gently and incubate at room temperature for 45 min.
(3) Heat-inactivate the reaction at 70 °C for 15 min.

End-repaired DNA was then purified using the QIAquick PCR purification kit (Qiagen) and quantified using the Qubit fluorometer BR assay kit (http://tools.invitrogen.com/content/sfs/manuals/mp32850.pdf) and ligated into the pZE21 MCS 1 vector[28] at the HincII site. The pZE21 vector was linearized at the HincII site using inverse PCR with the blunt-end PFX DNA polymerase (Life Technologies) per the following reaction conditions:

(1) Mix the following in a 50 μl reaction volume: 10 μl of 10X PFX reaction buffer, 1.5 μl of 10 mM dNTP mix (New England Biolabs), 1 μl of 50 mM $MgSO_4$, 5 μl of PFX enhancer solution, 1 μl of 100 pg μl^{-1} circular pZE21, 0.4 μl of PFX DNA polymerase, 0.75 μl forward primer (5′ GACGGTATCGATAAGCTTGAT 3′), 0.75 μl reverse primer (5′ GACCTCGAGGGGGGG 3′) and 29.6 μl of nuclease-free H_2O to a final volume of 50 μl.
(2) Cycle temperature as follows: 95 °C for 5 min, then 35 cycles of [95 °C for 45 s, 55 °C for 45 s, 72 °C for 2.5 min], then 72 °C for 5 min.

Linearized pZE21 was then size-selected (~2,200 bp) on a 1% low-melting-point agarose gel (0.5X TBE) stained with GelGreen dye (Biotium) and purified as described above. Pure vector was dephosphorylated using calf intestinal phosphatase (New England Biolabs) by adding 1/10th reaction volume of calf intestinal phosphatase, 1/10th reaction volume of New England Biolabs buffer 3, and nuclease-free H_2O to the vector eluate (exact volumes depend on reaction scale) and incubating at 37 °C overnight before heat inactivation for 15 min at 70 °C. End-repaired metagenomic DNA and linearized vector were then ligated together using the Fast Link Ligation Kit (Epicentre) at a 5:1 mass ratio of insert:vector using the following protocol (because the insert and vector were similarly sized, the mass ratio approximates a molar ratio):

(1) Mix the following: 1.5 μl 10X Fast-Link buffer, 0.75 μl ATP (10 mM), 1 μl Fast-Link DNA ligase (2 U μl^{-1}), 5:1 mass ratio of metagenomic DNA to vector, and nuclease-free H_2O to a final reaction volume of 15 μl.
(2) Incubate at room temperature overnight.
(3) Heat inactivate for 15 min at 70 °C.

After heat inactivation, ligation reactions were dialysed for 30 min using a 0.025 μm cellulose membrane (Millipore catalogue number VSWP09025) and the full reaction volume used for transformation by electroporation into 50 μl *E. coli* MegaX (Invitrogen). Electroporation was conducted using the manufacturer's recommendations (http://tools.invitrogen.com/content/sfs/manuals/megax_man.pdf), and cells were recovered in 1 ml Recovery Medium (Invitrogen) at 37 °C. Libraries were titred by plating out 0.1 μl and 0.01 μl of recovered cells onto Luria–Bertani (LB) agar (5 g yeast extract, 5 g NaCl, 10 g of tryptone, 12 g agar in 1 litre of water) plates containing 50 μg ml^{-1} kanamycin. For each library, insert size distribution was estimated by gel electrophoresis of PCR products obtained by amplifying the insert from 12 randomly picked clones using primers flanking the HincII site of the multiple cloning site of the pZE21 MCS1 vector (which contains a selectable marker for kanamycin resistance). The average insert size across all libraries was determined to be 2,000 bp, and library size estimates were calculated by multiplying the average PCR-based insert size by the number of titred colony forming units (CFUs) after transformation recovery. The rest of the recovered cells were inoculated into 10 ml of LB containing 50 μg ml^{-1} kanamycin and grown overnight. The overnight culture was frozen down with 15% glycerol and stored at −80 °C for subsequent screening.

Functional selections for antibiotic resistance. For each soil metagenomic library, selections for resistance to each of 18 antibiotics (at concentrations indicated in Supplementary Table 2) was performed using Mueller–Hinton (MH) agar (2 g beef infusion solids, 1.5 g starch, 17 g agar, 17.5 g casein hydrolysate, pH 7.4, in a final volume of 1 litre). For each metagenomic library, the number of cells plated on each antibiotic selection represented 10× the number of unique CFUs in the library, as determined by titres during library creation. Depending on the titre of live cells following library amplification and storage, the appropriate volume of freezer stocks were either diluted to 100 μl using LB broth or centrifuged and reconstituted in this volume for plating. After plating (using sterile glass beads), antibiotic selections were incubated at 37 °C for 18 h to allow the growth of clones containing an antibiotic-resistant DNA insert. After overnight growth, all colonies from a single antibiotic plate (soil by antibiotic selection) were collected by adding 750 μl of 15% LB-glycerol to the plate and scraping with an L-shaped cell scraper (Fisher Scientific catalogue number 03-392-151) to gently remove colonies from the agar. The liquid 'plate scrape culture' was then collected and this process was repeated a second time to ensure that all colonies were removed from the plate. The bacterial cells were then stored at −80 °C before PCR amplification of antibiotic-resistant metagenomic fragments and Illumina library creation.

Amplification of antibiotic resistant metagenomic DNA fragments. Freezer stocks of antibiotic-resistant transformants were thawed and 300 μl of cells pelleted by centrifugation at 13,000 revolutions per minute (r.p.m.) for two minutes and gently washed with 1 ml of nuclease-free H_2O. Cells were subsequently pelleted a second time and re-suspended in 30 μl nuclease-free H_2O. Re-suspensions were then frozen at −20 °C for one hour and thawed to promote cell lysis. The thawed re-suspension was then pelleted by centrifugation at 13,000 r.p.m. for two minutes and the resulting supernatant used as template for amplification of resistance-conferring DNA fragments by PCR with Taq DNA polymerase (New England Biolabs). A sample PCR reaction consisted of 2.5 μl of template, 2.5 μl of ThermoPol reaction buffer (New England Biolabs), 0.5 μl of 10 mM deoxynucleotide triphosphates (dNTPs, New England Biolabs), 0.5 μl of Taq polymerase (5 U μl^{-1}), 3 μl of a custom primer mix, and 16 μl of nuclease-free H_2O to bring the final reaction volume to 25 μl. The custom primer mix consisted of three forward and three reverse primers, each targeting the sequence immediately flanking the HincII site in the pZE21 MCS1 vector, and staggered by one base pair. The staggered primer mix ensured diverse nucleotide composition during early Illumina sequencing cycles and contained the following primer volumes (from a 10 μM stock) in a single PCR reaction: (primer F1, 5′-CCGAATTCATTAAAGAGGAGAAAG, 0.5 μl); (primer F2, 5′-CGAATTCATTAAAGAGGAGAAAGG, 0.5 μl); (primer F3, 5′-GAATTCATTAAAGAGGAGAAAGGTAC, 0.5 μl); (primer R1, 5′-GATATCAAGCTTATCGATACCGTC, 0.21 μl); (primer R2, 5′-CGATATCAAGCTTATCGATACCG, 0.43 μl); (primer R3, 5′-TCGATATCAAGCTTATCGATACC, 0.86 μl). PCR reactions were then amplified using the following thermocycler conditions: 94 °C for 10 min, 25 cycles of 94 °C for 5 min + 55 °C for 45 s + 72 °C for 5.5 min and 72 °C for 10 min. The amplified metagenomic inserts were then cleaned using the Qiagen QIAquick PCR purification kit and quantified using the Qubit fluorometer HS assay kit (http://tools.invitrogen.com/content/sfs/manuals/mp32851.pdf).

Illumina sample preparation and sequencing. For amplified metagenomic inserts from each antibiotic selection, 0.5 μg of PCR product was diluted to a total volume of 100 μl in Qiagen EB buffer and then sheared to 150–200 bp fragments using the BioRuptor XL (http://www.sibcb.ac.cn/cfmb/download/BioruptorManual.pdf). Sonication consisted of nine 10-min cycles of 30 s ON (high power setting), 30 s OFF. Between each 10-min cycle, ice was added to the water bath to prevent overheating. Following sonication, sheared DNA was purified and concentrated using the QIAGEN MinElute PCR Purification Kit and eluted in 20 μl pre-warmed nuclease-free H_2O. This eluate was then used as input for Illumina library preparation. In the first step of library preparation, sheared DNA was end-repaired by mixing the 20 μl of eluate with 2.5 μl T4 DNA ligase buffer with 10 mM ATP (10X, New England Biolabs), 1 μl dNTPs (10 mM, New England Biolabs), 0.5 μl T4 polymerase (3 U μl^{-1}, New England Biolabs), 0.5 μl T4 PNK (10 U μl^{-1}, New England Biolabs), and 0.5 μl Taq Polymerase (5 U μl^{-1}, New England Biolabs) for a total reaction volume of 25 μl. The reaction was incubated at 25 °C for 30 min followed by 20 min at 75 °C.

Next, to each end-repaired sample, 5 μl of 1 μM pre-annealed, barcoded sequencing adapters were added (adapters were thawed on ice). Barcoded adapters consisted of a unique 7-bp oligonucleotide sequence specific to each antibiotic selection, facilitating the de-multiplexing of mixed-sample sequencing runs. Forward and reverse sequencing adapters were stored in TES buffer (10 mM Tris, 1 mM EDTA, 50 mM NaCl, pH 8.0) and annealed by heating the 1 μM mixture to 95 °C followed by a slow cool (0.1 °C per second) to a final holding temperature of 4 °C. After the addition of barcoded adapters, samples were incubated at 16 °C for 40 min and then for 10 min at 65 °C. Before size-selection, 10 μl each of adapted-ligated samples were combined into pools of 12 and concentrated by elution through a Qiagen MinElute PCR Purification Kit, eluting in 14 μl of Qiagen elution buffer.

©2014 Macmillan Publishers Limited. All rights reserved

The pooled, adaptor-ligated, sheared DNA was then size-selected on a 2% agarose gel in 0.5X TBE, stained with GelGreen dye (Biotium). DNA fragments were combined with 2.5 μl 6X Fermentas Orange loading dye before loading on to the gel. Adaptor-ligated DNA was extracted from gel slices corresponding to DNA of 300–400 bp using a QIAGEN MinElute Gel Extraction kit. The purified DNA was enriched by PCR using 12.5 μl 2X Phusion HF Master Mix and 1 μl of 10 μM Illumina PCR Primer Mix in a 25 μl reaction using 2 μl of purified DNA as template. DNA was amplified at 98 °C for 30 s followed by 18 cycles of 98 °C for 10 s, 65 °C for 30 s, 72 °C for 30 s with a final extension of 5 min at 72 °C. Afterwards, the DNA concentration was measured using the Qubit fluorometer (HS assay) and 10 nM of each sample were pooled for sequencing. Subsequently, samples were submitted for Illumina Hi-Seq paired-end 101-bp sequencing using the HiSeq 2000 platform at GTAC (Genome Technology Access Center, Washington University in St Louis, USA). In total, four sequence runs were performed at concentrations ranging from 7 to 9 pM per lane.

Assembly and annotation of functional metagenomic selections. Illumina paired-end sequence reads were binned by barcode (exact match required), such that independent selections were assembled and annotated in parallel. Assembly of the resistance-conferring DNA fragments from each selection was achieved using PARFuMS (Parallel Annotation and Re-assembly of Functional Metagenomic Selections); a tool developed specifically for the high-throughput assembly and annotation of functional metagenomic selections[4]. Assembly with PARFuMS consists of: (1) three iterations of variable job size with the short-read assembler Velvet[29], (2) two iterations of assembly with Phrap[30], and (3) custom scripts to clean sequence reads, remove chimaeric assemblies, and link contigs by coverage and common annotation, as described[4]. In addition to outperforming traditional, Sanger-based methods for characterizing functional selections, PARFuMS has also been successfully applied to the interrogation of both soil[4] and faecal[31] resistomes. Of the 324 selections performed, 222 yielded antibiotic-resistant *E. coli* transformants (Extended Data Fig. 1), of which 219 were successfully sequenced and assembled into contigs larger than 500 bp. To annotate these assembled contigs, we opted to upgrade the previous implementation of PARFuMS, replacing annotation by BlastX homology to COG[32] with a profile HMM-based approach. ORFs were predicted using the gene-finding algorithm MetaGeneMark[33] and annotation was performed by searching the amino acid sequence against multiple profile HMM databases with HMMER3[34], including TIGRFAMS[35], PFams[36], and a collection of custom-built, resistance-gene-specific profile HMMs (http://dantaslab.wustl.edu/resfams). MetaGeneMark was run using default gene-finding parameters while hmmscan (HMMER3) was run with the option "–cut_ga", requiring that genes meet profile-specific gathering thresholds (rather than a global, more permissive, default log odds cutoff) before receiving annotation. An ORF from a resistance-conferring DNA fragment was labelled an ARG if it met one of the following criteria: (1) it surpassed strict, profile-specific gathering thresholds from the custom-built set of profile HMMs, (2) it matched obvious antibiotic resistance functions from the TIGRFAMS or PFams databases (for example, metallo β-lactamase, chloramphenicol phosphotransferase, major facilitator superfamily transporter), or (3) the ORF was sub-cloned from its original context and confirmed to confer antibiotic resistance when expressed in *E. coli*. In total, 2,895 of the 8,882 assembled ORFs (32.6%) could be confidently assigned an ARG label through one of these routes, representing 2,730 unique sequences. To generate more encompassing counts of general resistance functions (for example, β-lactamases), gene counts were summed across all annotations that clearly belonged to the parent function (for example, class A β-lactamases, metallo β-lactamases, TEM β-lactamases), informed by established ARG ontology[37]. Annotations were categorized as mobility elements based on string matches to one of the following keywords: transposase, transposon, conjugative, integrase, integron, recombinase, conjugal, mobilization, recombination, plasmid.

Percentage identity comparisons of recovered ORFS against NCBI. Percentage identity comparisons using either all ORFs or all ARGs were conducted via a BlastX query against the NCBI protein Non-Redundant (NR) database (retrieved 20 August 2013). For each ORF, the NCBI entry that generated the best local alignment was used to create global alignments with estwise (http://dendrome.ucdavis.edu/resources/tooldocs/wise2/doc_wise2.html). The following options were used in global alignment: "-init global" and "-alg 333". From this alignment, global percentage identities were calculated as the number of matched amino acids divided by the full length of the shorter of the two sequences compared.

Comparison of ARG content between soils. To compare the ARG composition of various soils, a count matrix was generated where each row represented a given soil metagenomic library and each column was represented by a specific annotation (that is, a profile HMM). Before populating each cell of the matrix, genes duplicated as a result of redundant assembly were collapsed into a single sequence with CD-HIT[38]. For each selection, all genes perfectly identical over the length of the shorter sequence were collapsed into a single sequence using CD-HIT with the parameters: -c 1.0 -aS 1.0 -g 1 -d 0 (the longest gene in the 100% identical cluster was retained for downstream analyses). Subsequently, fasta files of all genes sequences from all antibiotic selections for a given soil were concatenated and perfectly identical genes were again collapsed to a single sequence, using CD-HIT with the same parameters. Thus, the same gene captured on multiple selections from a given soil would be counted only once for that soil. These unique counts ("raw counts") only considered genes over 350 bp and were used in the creation of all figures summarizing the total resistance functions recovered (for example, Fig. 1c). Selections containing trimethoprim and D-cycloserine predominantly recovered dihydrofolate reductases, D-alanine–D-alanine ligases, and thymidylate synthases (these annotations accounted for 92.5% of ARGs from these selections). Because these genes represent target alleles of their respective antibiotics, and are present in nearly all bacterial genomes, their overexpression can provide resistance but the functions themselves do not represent an evolutionary response to overcome toxicity. Thus, we omitted these selections from cross-soil resistome comparisons (Extended Data Fig. 2).

As the size of metagenomic libraries varied stochastically by soil sample (Extended Data Fig. 1), raw ARG counts were normalized to metagenomic library size to account for inconsistent sampling depth, facilitating comparison between soils. Metagenomic libraries from soils S18 and S21 were both under 2 Gb in size, over fourfold smaller than next smallest library (S06), resulting in distinctly fewer selections yielding antibiotic resistance (Extended Data Fig. 1). Thus, these libraries were omitted from cross-soil comparisons and ARG counts were normalized to reflect the sampling depth achieved for the smallest remaining library, from soil S06 (library size was 6.9 Gb, roughly 3.5 million 2 kb DNA fragments). Both raw and normalized count matrices were created for the following gene sets: (1) all recovered ORFs (including ARGs and co-selected passenger genes), (2) only unique ARGs, (3) ARGs from only CC soils, and (4) ARGs from only KBS soils. When counts were normalized across only KBS soils, the smallest library from KBS (S14; 10.9 Gb) was used to normalize raw counts. Normalized count matrices were then used to calculate Bray–Curtis distances between soil samples (using the vegan package in R), which in turn were used in cross-soil analyses (for example, principal coordinate analyses, Procrustes analyses, Mantel tests, and so on).

The percentage of ARGs shared across any two soils was determined using sequences collected from selections without trimethoprim or D-cycloserine. Unique gene sequences from each soil were then clustered using CD-HIT with the following parameters: -c 0.99 (99% sequence identity) -aS 1.0 (over the full length of the shorter fragment) -g 1 (find the optimal cluster). Any sequences from more than one soil in a given 99% identity sequence bin were counted towards the fraction of shared sequences.

16S rRNA analysis. Sequencing of the 16S rRNA gene was performed on a Roche 454 GS FLX DNA sequencer using titanium chemistry and as described previously[8]. Briefly, the 16S primers 515F and 906R were used for their ability to generate accurate phylogenetic information with few taxonomic biases[8]. Primers also included a 454 sequencing adaptor and the reverse primer contained a 12-bp error-correcting barcode, generating approximately 300-bp sequencing reads. All downstream processing was performed with the QIIME (v1.4) software suite[39] and was reproduced from previous work[8]. After sequencing, 16S reads were split by sample, quality-filtered using the QIIME script split_libraries.py with the options: -w 50 -g -r -l 150 -L 350, and then de-noised using denoise_wrapper.py with default parameters. All sequence analyses were performed using the QIIME analysis pipeline[39], following the usage information found at http://qiime.org/tutorials/tutorial.html. Briefly, operational taxonomic units were picked at 97% sequence identity and taxonomic identity assigned by classification using the Ribosome Database Project. For comparisons across soils, samples were rarefied to 1,974 reads and both phylogenetic (unweighted Unifrac and weighted Unifrac distances[40,41]) and taxonomic (Bray–Curtis distance) were calculated. For one sample (soil S08), no 16S rRNA gene sequence data was available. Generally, all three measures of community similarity were used to examine relationships between the ARG content and phylogenetic composition of soil microbial communities, and supported the same conclusion: resistomes and bacterial community composition are correlated.

Assigning taxonomy to assembled sequences. To predict the taxonomic origin of functionally selected DNA fragments, we used RAIphy, a composition-based classifier that achieves accurate taxonomic prediction without a strict reliance on phylogenetically close sequences in public databases, as compared to similarity-based methods[19]. Specifically, RAIphy compares the relative abundance of all unique 7-mers within a query sequence to profiles of 7-mer abundance from RefSeq genomes, generating a score for each profile using the log-odd ratios between observed and expected frequency of each 7-mer. Prediction accuracy is then improved through an iterative refinement of genome models based on the 7-mer profiles of clusters of fragments of unknown origin in the query set. RAIphy reportedly performs well for classifying DNA fragments assembled from metagenomic sources[19], especially for lower-resolution taxonomic predictions. Thus, we reasoned it may be well-suited for phylum-level predictions of the originating taxa for our assembled contigs.

©2014 Macmillan Publishers Limited. All rights reserved

To convince ourselves of RAIphy's accuracy, we asked it to predict the source phylum of metagenomic DNA fragments originating from pools of genome-sequenced commensals of the human gut, selected via functional metagenomics for antibiotic resistance and assembled with PARFuMS (in exactly the same manner as soil resistomes were interrogated). Because full genomes existed for these organisms, we could determine with high confidence the true origin for functionally selected fragments: RAIphy's predictions of the source bacterial phylum correctly classified the assembled fragment with 95% accuracy ($n = 2747$), indicating that the software is well suited for the phylum-level classification of assembled metagenomic sequences. To predict the source phylum of resistance-conferring soil DNA fragments, we used all assembled fragments longer than 500 bp ($n = 4,655$), seeded predictions using the RAIphy's 2012 RefSeq database, and binned DNA fragments with the 'iterative refinement' option. For all downstream analyses, only the phylum-level predictions from RAIphy were used because we had higher confidence in these predictions, and also conclusions from these data may be more broadly applicable to different soils and other environments.

Assessing HGT potential for ARGs in soil and pathogens. To test the hypothesis that ARGs in the soil have less potential for HGT than those in human pathogens, we compared ARGs from our functional selections to ARGs in fully sequenced bacterial genomes from 433 common human pathogens and 153 non-pathogenic soil organisms. Genomes were stratified by pathogenicity and habitat according to the metadata presented in a recent paper examining general trends in horizontal gene transfer among bacteria across ecology[13]. A list of NCBI taxonomy IDs for human pathogens was obtained for all organisms from this publication with an 'Environment' label of "Human" and a 'Pathogenicity' label of "Pathogen". Taxonomy IDs for non-pathogenic soil bacteria were collected for all organisms with an 'Environment' labelled "non-human" that also contained the term "soil"; bacteria deemed pathogens were also omitted from the soil set. For each set of NCBI taxonomy IDs, all NCBI RefSeq genomes and plasmids were then downloaded. In total, 983 sequences from 433 human pathogens (downloaded 18 January 2014) and 296 sequences from 153 non-pathogenic soil bacteria (downloaded 3 February 2014) were obtained, and are enumerated in Supplementary Data 2.

We next re-annotated each bacterial genome using the same methods use to annotate assembly data from our functional selections. Briefly, ORFs were predicted using the gene-finding algorithm MetaGeneMark[33] and annotation performed by searching the amino acid sequence against multiple profile HMM databases with HMMER3[34], including TIGRFAMS[35], PFams[36], and a collection of custom-built, resistance gene-specific profile HMMs (http://dantaslab.wustl.edu/resfams). MetaGeneMark was run using default gene-finding parameters while hmmscan (HMMER3) was run with the option "–cut_ga", requiring that genes meet profile-specific gathering thresholds (rather than a global, more permissive, default log odds cutoff) before receiving annotation. Because our functional selections captured only those DNA fragments that confer antibiotic resistance, we modelled our functional metagenomic data set using genome collections by seeding mock 'metagenomic' DNA fragments at each predicted ARG across our genomes. For our Monte Carlo simulations of each genome set, we mimicked functional metagenomic DNA fragments by moving upstream and downstream from each seed ARG by a chosen genetic distance. These distances were selected from an empirical distribution of distances observed in our functional selections. For each ARG from our functional selections, the distance between the ARG boundary and the end of the assembled DNA fragment was recorded. These distance pairs were catalogued for all ARGs, and randomly selected to create a DNA fragment centred on each ARG in all bacterial genomes. In this fashion, we modelled a functional selection from each genome set based on the fragment-size distribution observed in our soil selections. Then, for each simulated DNA fragment, the number of mobile DNA elements over 350 bp contained within its boundaries were counted and ultimately displayed as a proportion of total DNA fragments queried. This sampling procedure was repeated 1,000 times for each genome set, each time with randomly selected upstream/downstream distance pairs, and the proportion of mobility elements compared to that observed from our soil functional selections was presented (Fig. 4a). If a single fragment contained multiple ARGs, the additional ARGs were not used to seed future fragments in that simulation. Since ARGs were drawn randomly from simulation to simulation, each member of a co-localized ARG group was equally likely to seed a DNA fragment.

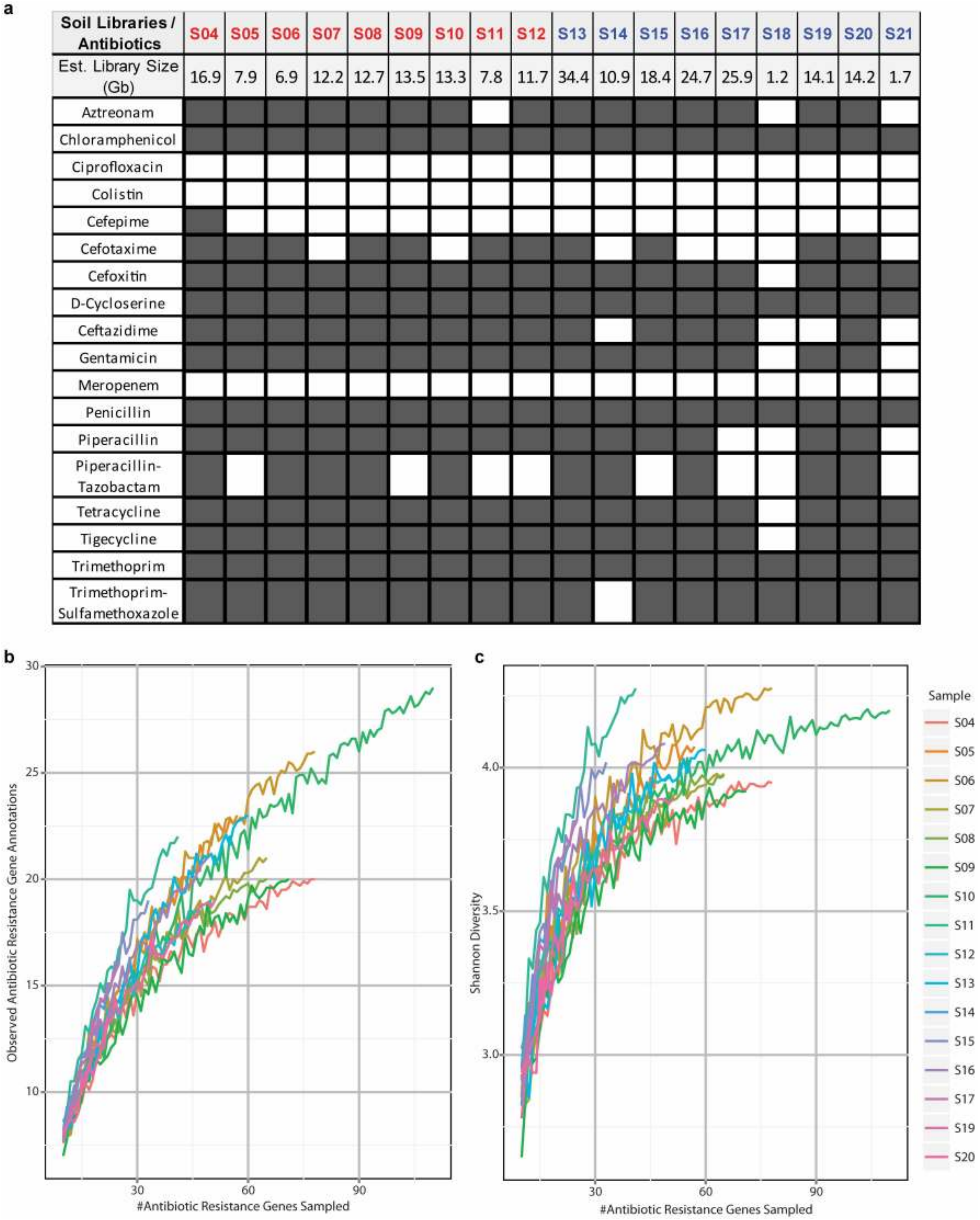
In Fig. 4b, the same data are plotted as a function of the distance from each ARG across pathogen genomes, non-pathogenic soil genomes, and assembled data from soil functional selections. Because the size of DNA fragments in our functional selections is constrained by the shearing conditions employed, their data could only be evaluated to a genetic distance of approximately 1.5 kb from each ARG. Nonetheless, the incidence of co-occurring ARGs and mobility elements is higher in pathogens than in soil genomes or functionally selected soil metagenomes, at all genetic distances tested that were greater than 580 bp. Annotations were categorized as mobility elements based on string matches to one of the following keywords: transposase, transposon, conjugative, integrase, integron, recombinase, conjugal, mobilization, recombination, plasmid.

Statistical analyses. QIIME was used to perform both principal coordinate analyses (PCoA, using principal_coordinates.py) and Procrustes transformations (using the script transform_coordinate_matrices.py, with two PCoA plots as input; one built from 16S rRNA gene sequence data and the other from resistome data; see ref. 11 for detailed description). The significance of any Procrustes transformation was determined by comparing the measure of fit, M^2, between matched-sample PCoA plots to a distribution of M^2 values empirically determined from 10,000 label permutations. In each of the 10,000 permutations, the M^2 value (the sum of squared distances between matched sample pairs) was recalculated and the original M^2 value compared to the simulated distribution in order to compute a P value. Because the M^2 value is dependent on the sample size and data structure, it is generally not comparable across Procrustes transformations. Rather, P values were used to compare different Procrustes plots. Regardless of whether transformations were performed using phylogenetic (unweighted or weighted Unifrac distances) or taxonomic (Bray–Curtis distances) measures of bacterial community composition, or considered only two or all dimensions of the relevant PCoA plots, we observed significant agreement between ARG content and bacterial composition when considering either all soils or just CC soils ($P < 0.05$, Supplementary Table 8).

Alpha-diversity plots (Extended Data Fig. 1) were generated by sampling (without replacement) an increasing subset of ARGs from each soil sample, and tabulating the observed number of unique annotations or Shannon diversity index at each rarefaction depth. Reported values (Extended Data Fig. 1) are the result of averaging ten independent samplings at each rarefaction depth, and were generated using the QIIME scripts multiple_rarefactions.py and alpha_diversity.py. The summary figure was generated in R (v2.15.2). Mantel tests were performed using the PRIMER-E software package[42] while all other statistics (ANOSIM, Fisher's exact test, Wilcoxon rank sum test, Student's t-test) were performed using the 'vegan' package in R.

29. Zerbino, D. R. & Birney, E. Velvet: algorithms for de novo short read assembly using de Bruijn graphs. *Genome Res.* **18,** 821–829 (2008).
30. de la Bastide, M. & McCombie, W. R. *Assembling Genomic DNA sequences with PHRAP* 2008/04/23 edn, Vol. 11 (John Wiley, 2007).
31. Moore, A. M. *et al.* Pediatric fecal microbiota harbor diverse and novel antibiotic resistance genes. *PLoS ONE* **8,** e78822 (2013).
32. Tatusov, R. L., Galperin, M. Y., Natale, D. A. & Koonin, E. V. The COG database: a tool for genome-scale analysis of protein functions and evolution. *Nucleic Acids Res.* **28,** 33–36 (2000).
33. Zhu, W., Lomsadze, A. & Borodovsky, M. Ab initio gene identification in metagenomic sequences. *Nucleic Acids Res.* **38,** e132 (2010).
34. Finn, R. D., Clements, J. & Eddy, S. R. HMMER web server: interactive sequence similarity searching. *Nucleic Acids Res.* **39,** W29–37 (2011).
35. Haft, D. H. *et al.* TIGRFAMs: a protein family resource for the functional identification of proteins. *Nucleic Acids Res.* **29,** 41–43 (2001).
36. Bateman, A. *et al.* The Pfam protein families database. *Nucleic Acids Res.* **28,** 263–266 (2000).
37. McArthur, A. G. *et al.* The comprehensive antibiotic resistance database. *Antimicrob. Agents Chemother.* **57,** 3348–3357 (2013).
38. Li, W. & Godzik, A. Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics* **22,** 1658–1659 (2006).
39. Caporaso, J. G. *et al.* QIIME allows analysis of high-throughput community sequencing data. *Nature Methods* **7,** 335–336 (2010).
40. Lozupone, C., Hamady, M. & Knight, R. UniFrac–an online tool for comparing microbial community diversity in a phylogenetic context. *BMC Bioinform.* **7,** 371 (2006).
41. Lozupone, C., Lladser, M. E., Knights, D., Stombaugh, J. & Knight, R. UniFrac: an effective distance metric for microbial community comparison. *ISME J.* **5,** 169–172 (2011).
42. Clarke, K. R. & Gorley, R. N. *PRIMER v6: User Manual/Tutorial* 6th edn, Ch. 13 (PRIMER-E, 2006).

©2014 Macmillan Publishers Limited. All rights reserved

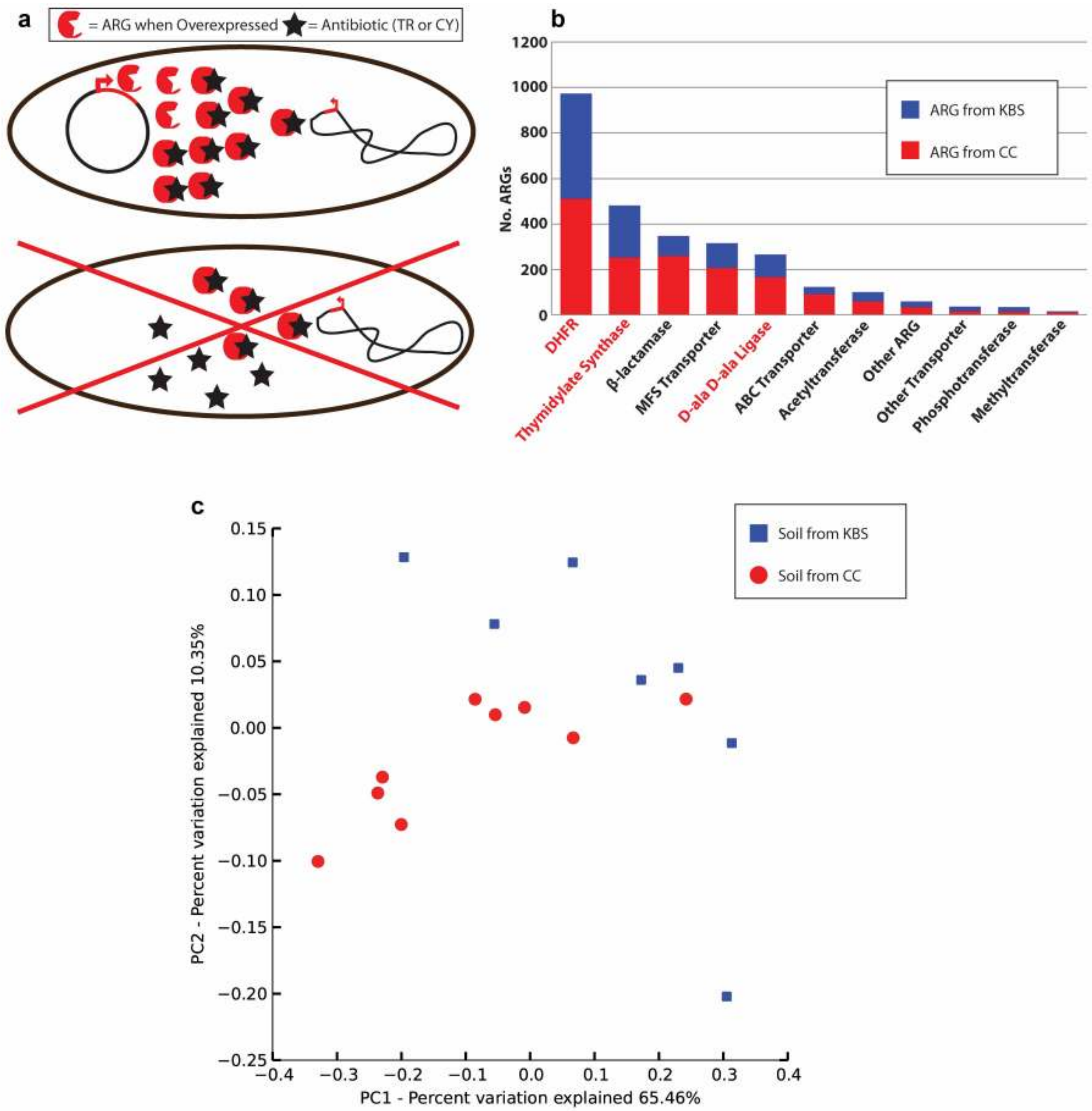


Soil Libraries / Antibiotics	S04	S05	S06	S07	S08	S09	S10	S11	S12	S13	S14	S15	S16	S17	S18	S19	S20	S21
Est. Library Size (Gb)	16.9	7.9	6.9	12.2	12.7	13.5	13.3	7.8	11.7	34.4	10.9	18.4	24.7	25.9	1.2	14.1	14.2	1.7
Aztreonam	■	■	■	■	■	■	■	□	■	■	■	■	■	■	□	■	■	□
Chloramphenicol	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■
Ciprofloxacin	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□
Colistin	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□
Cefepime	■	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□
Cefotaxime	■	■	■	□	■	■	□	■	■	■	□	■	□	□	□	■	■	□
Cefoxitin	■	■	■	■	■	■	■	■	■	■	■	■	■	■	□	■	■	■
D-Cycloserine	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■
Ceftazidime	■	■	■	■	■	■	■	■	■	■	□	■	■	■	□	□	■	□
Gentamicin	■	■	■	■	■	■	■	■	■	■	■	■	■	■	□	■	■	□
Meropenem	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□
Penicillin	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■
Piperacillin	■	■	■	■	■	■	■	■	■	■	■	■	■	□	□	■	■	□
Piperacillin-Tazobactam	■	□	■	■	■	□	■	□	□	■	■	□	■	□	□	■	■	□
Tetracycline	■	■	■	■	■	■	■	■	■	■	■	■	■	■	□	■	■	■
Tigecycline	■	■	■	■	■	■	■	■	■	■	■	■	■	■	□	■	■	■
Trimethoprim	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■
Trimethoprim-Sulfamethoxazole	■	■	■	■	■	■	■	■	■	■	□	■	■	■	■	■	■	■

Observed Antibiotic Resistance Gene Annotations
30
25
20
15
10
30
60
90
#Antibiotic Resistance Genes Sampled
Shannon Diversity
4.0
3.5
3.0
30
60
90
#Antibiotic Resistance Genes Sampled
Sample
S04
S05
S06
S07
S08
S09
S10
S11
S12
S13
S14
S15
S16
S17
S19
S20

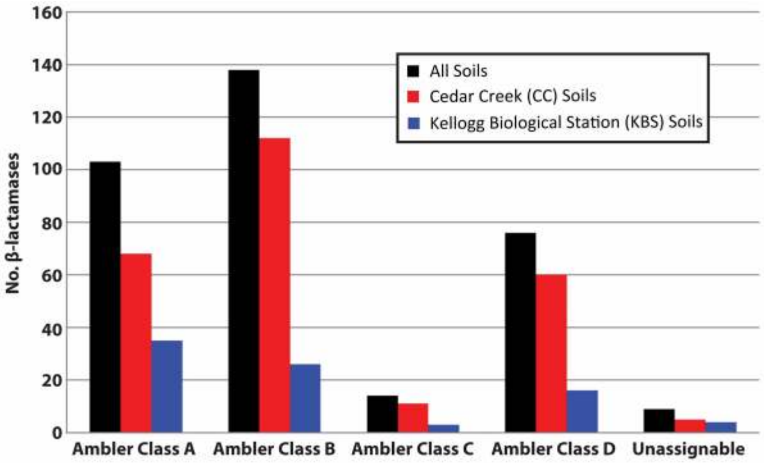
Extended Data Figure 1 | Functional selections of 18 soil metagenomes for resistance against 18 antibiotics. **a**, Phenotypic results of selections. A dark grey cell means that a resistance phenotype was observed whereas white cells indicate the absence of any drug-tolerant transformants. Grassland soils from CC are labelled in red and agricultural soils from KBS are labelled in blue. **b**, **c**, Alpha diversity representations. On the left is depicted the number of distinct ARG annotations observed as increasing numbers of ARGs are sampled from each soil. On the right, Shannon diversity scores (an ecological metric that quantifies within-sample diversity) are shown at each rarefaction step.

©2014 Macmillan Publishers Limited. All rights reserved



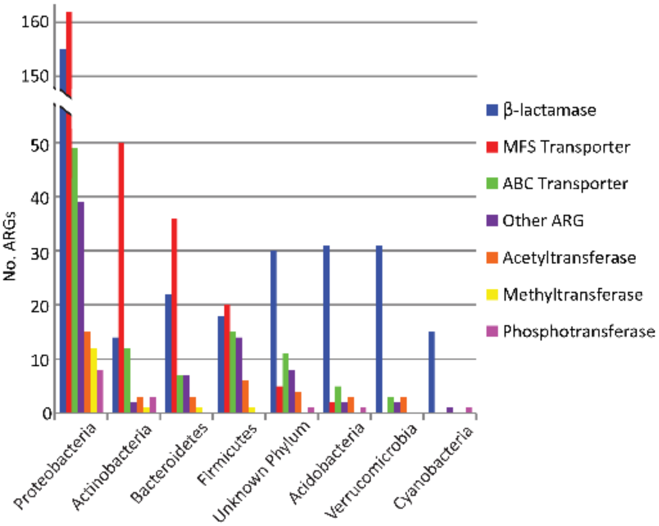
Extended Data Figure 2 | Three prominent ARG classes are present in nearly all bacterial genomes and can provide antibiotic resistance when overexpressed. **a**, Generalized as red circles are dihydrofolate reductases, D-alanine—D-alanine (D-ala D-ala) ligases, which are the molecular targets of the drugs trimethoprim (TR) and D-cycloserine (CY) respectively (black stars), and thymidylate synthases, which can provide trimethoprim resistance by circumventing the need for an active dihydrofolate reductase. When overexpressed in functional selections, these genes can provide antibiotic resistance. We found substantial diversity in these genes (average pairwise amino acid identity 39.3 ± 12.2%), suggesting that variants were captured from many bacterial lineages. **b**, Relative to other ARG mechanisms, large numbers of dihydrofolate reductases, thymidylate synthases, and D-ala D-ala ligases were found in all soils, with these ARGs representing 92.5% of resistance genes identified from selections containing trimethoprim or D-cycloserine antibiotics. Therefore, these selections encompass large genetic diversity, but constrained functional diversity, with a broad range of genes encoding limited functional traits. **c**, When considered in isolation, these functions were not different between the KBS and CC soils ($P > 0.05$, ANOSIM), indicating that trimethoprim and D-cycloserine resistance function is similarly distributed across the surveyed soil types.

©2014 Macmillan Publishers Limited. All rights reserved



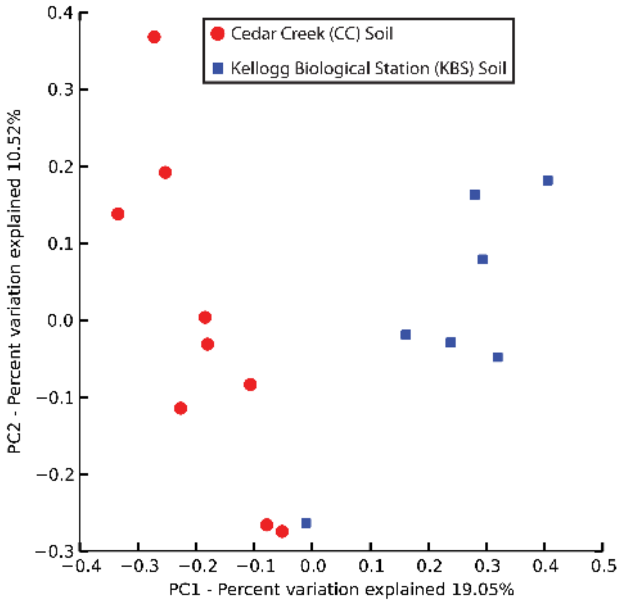
Extended Data Figure 3 | Total counts of β-lactamases recovered from antibiotic selections. All soils (black), CC soils (red), and KBS soils (blue).

©2014 Macmillan Publishers Limited. All rights reserved



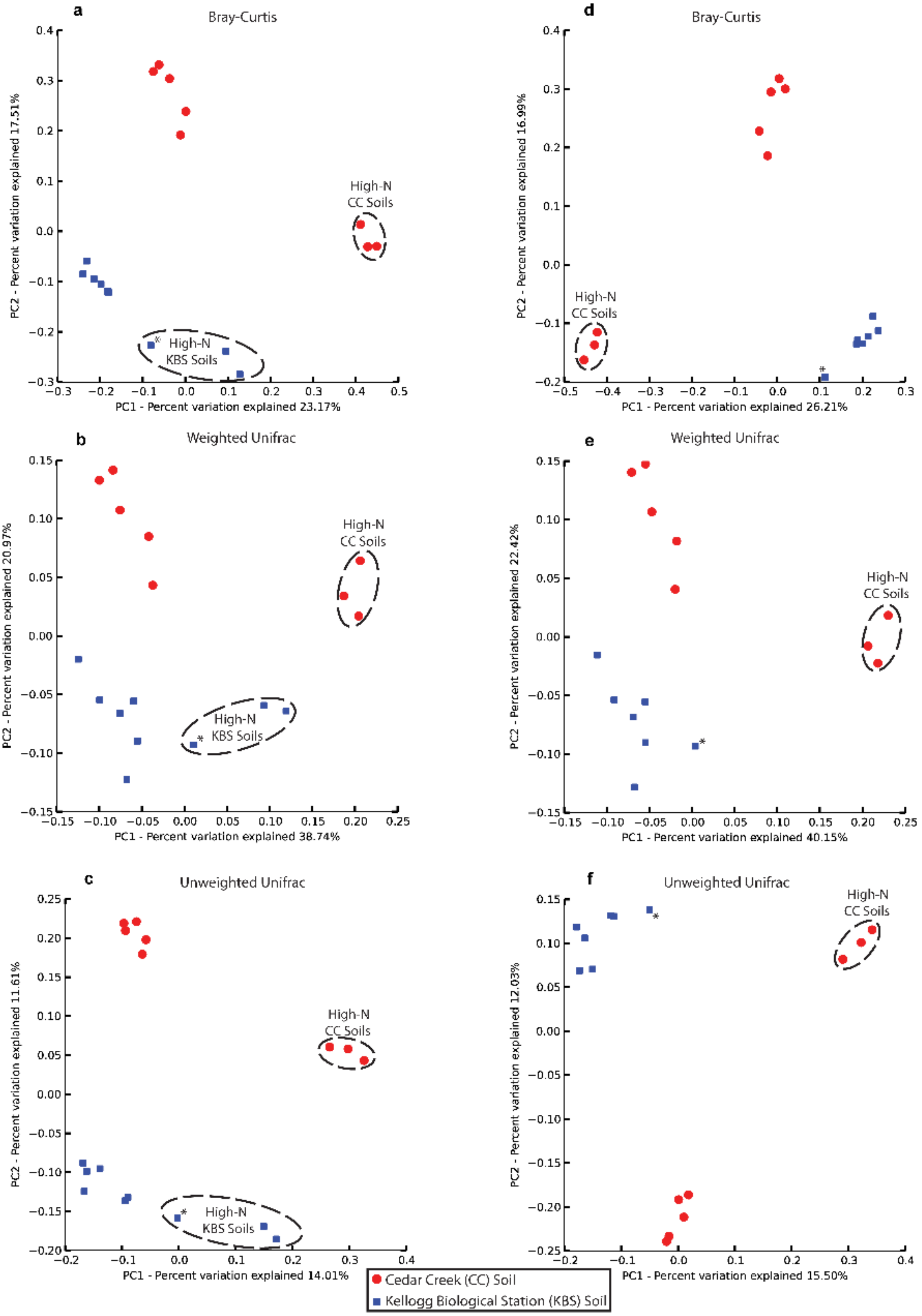
Extended Data Figure 4 | Total counts of ARGs categorized by their predicted phylogenetic origin. The number of ARGs is indicated on the *y* axis and the ARG types are colour-coded in the key.

©2014 Macmillan Publishers Limited. All rights reserved



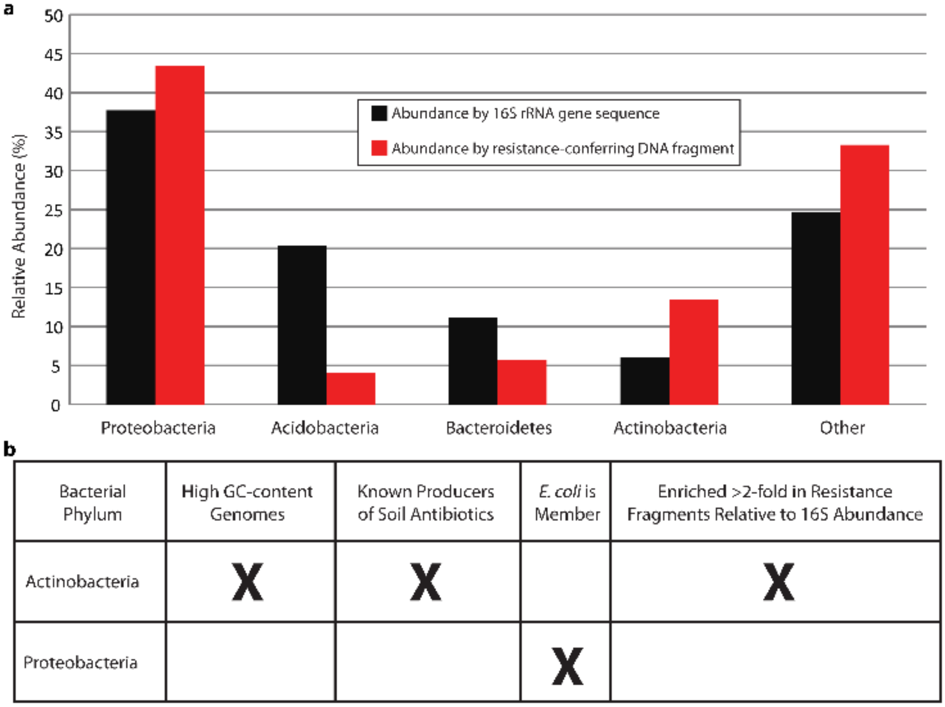
Extended Data Figure 5 | PCoA analysis plots of Bray–Curtis distances between soil resistomes. The PCoA was calculated using all ORFs captured from functional selections without trimethoprim and D-cycloserine, and shows significant separation between CC (red) and KBS (blue) resistomes ($P < 10^{-5}$, ANOSIM).

©2014 Macmillan Publishers Limited. All rights reserved



Extended Data Figure 6 | PCoA across CC (red, grassland) and KBS (blue, agricultural) soils. a–c, PCoA generated from all 16S data available from ref. 8, using Bray–Curtis (**a**), weighted Unifrac (**b**) and unweighted Unifrac (**c**) dissimilarity metrics. Samples cluster by soil location and N level, as previously demonstrated. **d–f**, The same PCoA plots generated using only samples with sufficient 16S and resistome data (that is, those used in Procrustes and Mantel analyses). Excluding the two high-N KBS soils with insufficient resistome data eliminates the clustering pattern observed for KBS soils in **a–c**. The asterisk denotes the high-N KBS soil common to both sets of analyses.

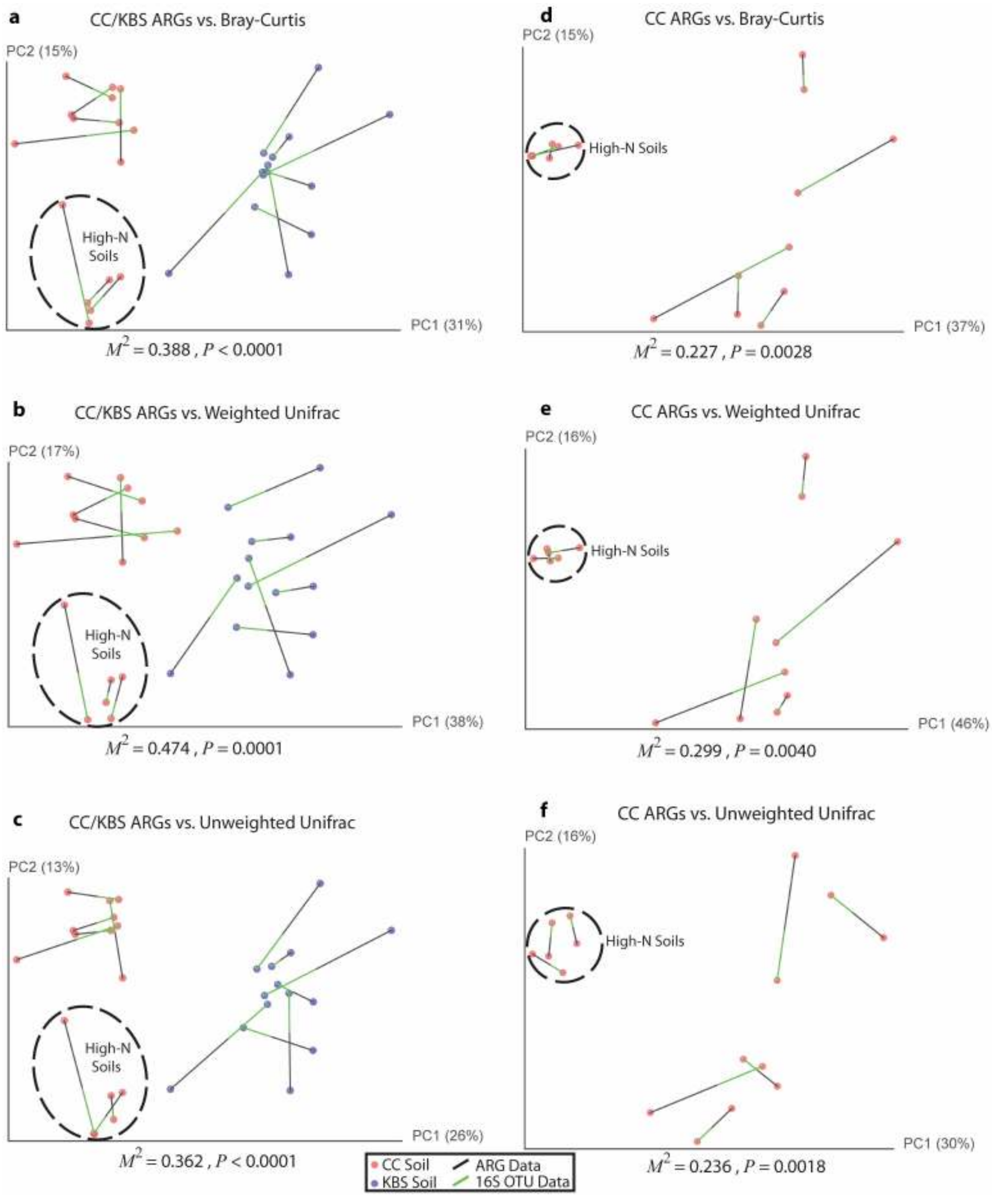
©2014 Macmillan Publishers Limited. All rights reserved



Bacterial Phylum	High GC-content Genomes	Known Producers of Soil Antibiotics	*E. coli* is Member	Enriched >2-fold in Resistance Fragments Relative to 16S Abundance
Actinobacteria	X	X		X
Proteobacteria			X	

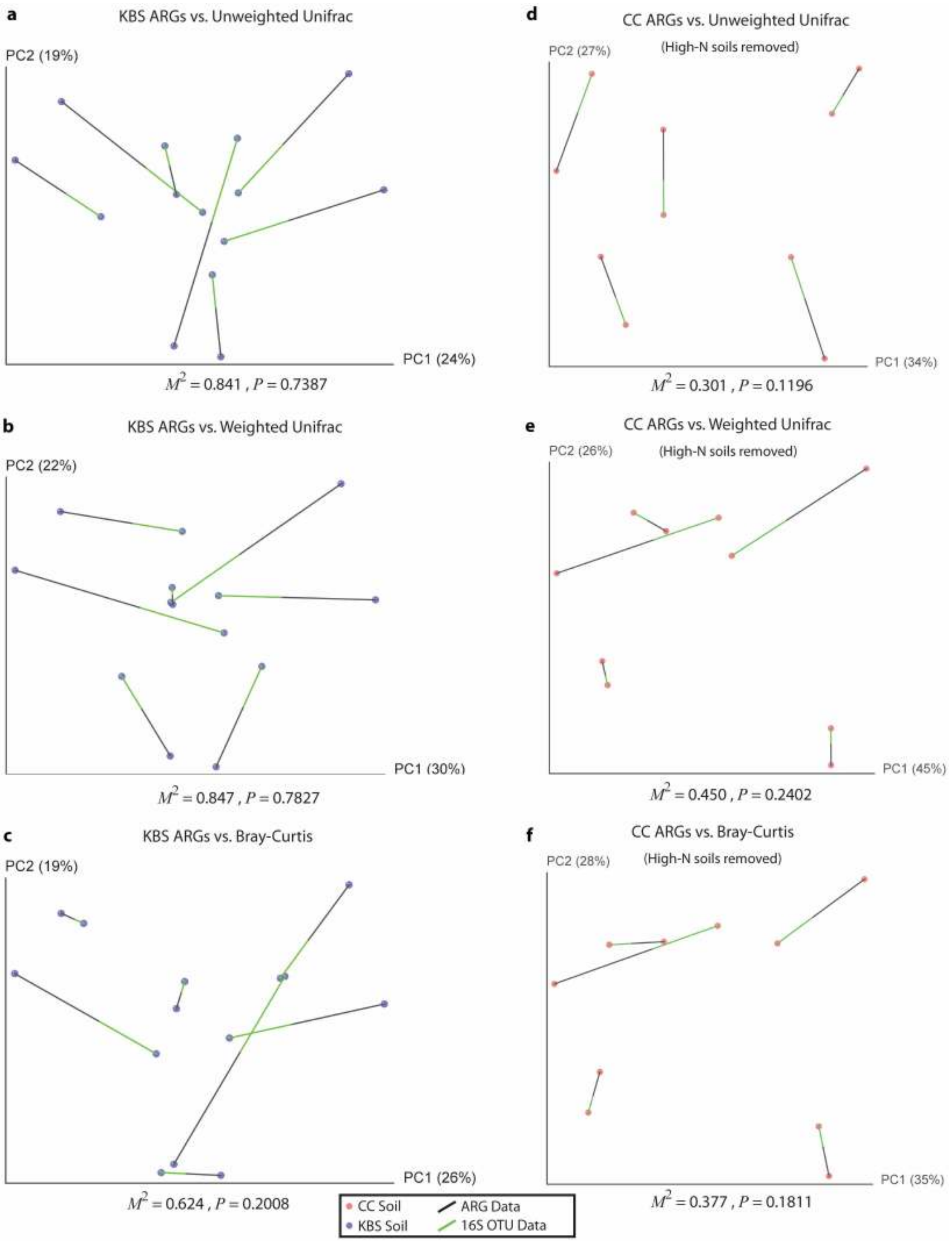
Extended Data Figure 7 | Phylum level relative abundance of combined CC and KBS data sets for major soil bacteria. **a**, 16s rRNA data are depicted in black. Phylogenetic inferences based on the sequence composition of the assembled, resistance-conferring DNA fragments are depicted in red. The relative abundances of Actinobacteria and Acidobacteria represent the largest discrepancies between data sets. **b**, Actinobacteria are most dramatically enriched in resistance-conferring DNA fragments, in accord with their role in producing antibiotics, but despite their high GC-content and predicted transcriptional incompatibilities with *E. coli*. Levels of Proteobacteria, the phylum to which *E. coli* belongs, are largely unchanged following functional selection, suggesting that any potential bias introduced to the selections by heterologous expression in *E. coli* is minimal compared to the effect of ARG-content of the source organisms.

©2014 Macmillan Publishers Limited. All rights reserved



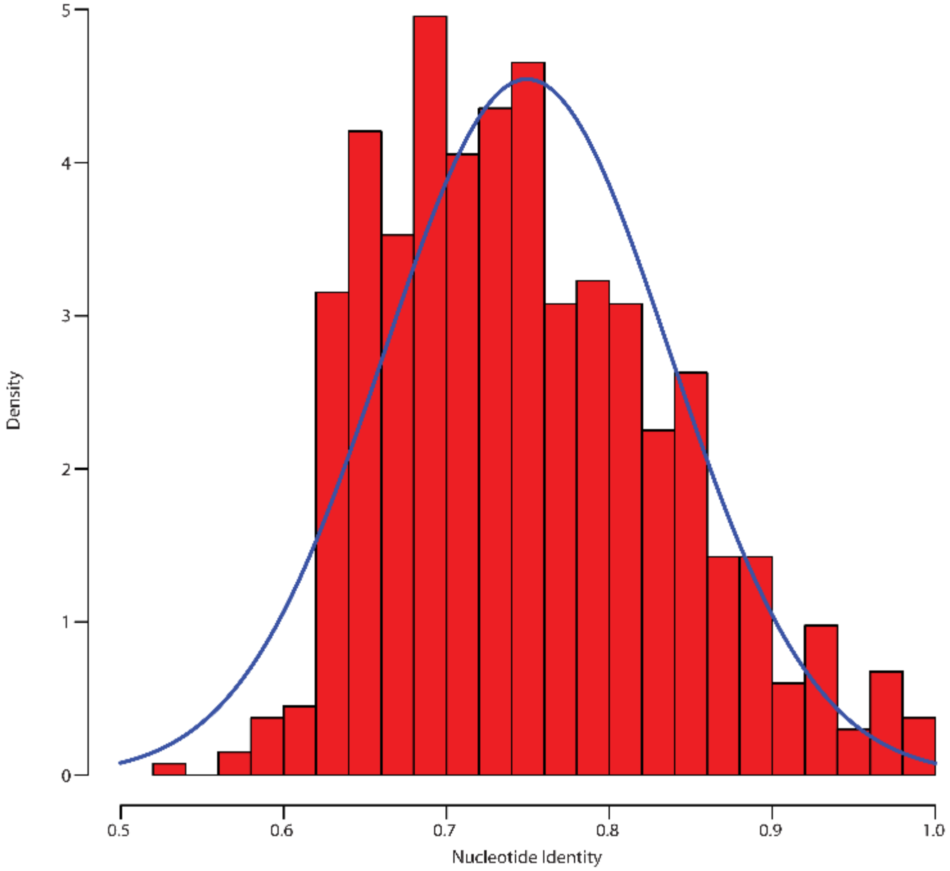
Extended Data Figure 8 | Procrustes analysis demonstrates that when soils cluster by bacterial composition, resistomes aggregate with phylogenetic groupings. **a**–**c**, Procrustes analysis of the ARG content (Bray–Curtis) of CC (red) and KBS (blue) soils compared to community composition calculated by Bray–Curtis (**a**), weighted Unifrac (**b**) and unweighted Unifrac (**c**) dissimilarity metrics. **d**–**f**, The same Procrustes transformations for CC soils only. For a given soil, black lines connect to functional resistome data while the green lines connect to points generated from 16S gene sequence data. The M^2 fit reported is from a Procrustes transformation over the first two principal coordinates while the P-value is calculated from a distribution of empirically determined M^2 values over 10,000 Monte Carlo label permutations. For M^2/P values calculated using all principal coordinates, refer to Supplementary Table 8.

©2014 Macmillan Publishers Limited. All rights reserved



Extended Data Figure 9 | Procrustes analysis demonstrates that when soils do not form distinct phylogenetic clusters, we are unable to detect significant correlation between ARG content and phylogenetic architecture. See Extended Data Fig. 6 for the phylogenetic relationships between these soils. **a**–**c**, Procrustes analysis of the ARG content (Bray–Curtis) of KBS (agricultural, blue) soils compared to 16S rRNA gene sequence using unweighted Unifrac (**a**), weighted Unifrac (**b**) and Bray–Curtis (**c**) similarity metrics. **d**–**f**, The same Procrustes transformations for the CC soils (grassland, red) without high-N amendment, showing that soil groupings must be distinguishable by bacterial composition to detect correlations with resistome content, regardless of soil type. For a given soil, black lines connect to functional resistome data while the green lines connect to points generated from 16S rRNA gene sequence data. The M^2 fit reported is from a Procrustes transformation over the first two principal coordinates while the P value is calculated from a distribution of empirically determined M^2 values over 10,000 Monte Carlo label permutations.

©2014 Macmillan Publishers Limited. All rights reserved



Extended Data Figure 10 | Histogram of nucleotide per cent identity from pairwise alignments of all predicted mobility elements, suggesting that assembly does not inappropriately condense mobile DNA elements into too few sequences. The blue trace depicts a normal distribution with the same mean and standard deviation empirically observed across all pairwise comparisons (*n* = 666).

©2014 Macmillan Publishers Limited. All rights reserved