Training and Benchmarking of Resfams profile HMMs

A curated list of antibiotic resistance (AR) proteins used to generate Resfams profile HMMs, along with associated reference gene sequences, is available at **http://dantaslab.wustl.edu/resfams**. These proteins were compiled using the Comprehensive Antibiotic Resistance Database (CARD) (McArthur *et al.*, 2013), the Antibiotic Resistance Database (ARDB) (Liu and Pop, 2009), the Lactamase Engineering Database (LacED) (Thai *et al.*, 2009) and Jacoby and BushÕs collection of curated beta-lactamase proteins (http://www.lahey.org/Studies/). All proteins were hand-curated to ensure functional homology using phylogenetic analysis and literature searches, eliminating all incorrectly annotated and truncated protein sequences.

Resfams profile HMMs were trained using CARD (McArthur *et al.*, 2013), LacED (Thai *et al.*, 2009), and Jacoby & BushÕs collection of beta-lactamases while retaining ARDB (Liu and Pop, 2009) as an independent test set. Gathering thresholds were optimized such that precision and recall metrics on this independent set of known genes were as close to 1.0 as possible. Precision was calculated as (the total number of correct annotations)/(total number of annotations). A precision of 1.0 for a profile HMM indicates that all annotations of proteins recruited by that HMM in the ARDB test set matched the annotation of the profile HMM. Recall was calculated as (the total number of correct annotations)/(the total number of proteins in ARDB with annotations matching the annotation of the profile HMM). A recall of 1.0 for a profile HMM indicates that all proteins contained in the ARDB test set that matched the annotation of the profile HMM were correctly recruited by the profile HMM.

Resfams profile HMM Precision and Recall Optimization Overview

The full set of precision and recall metrics of Resfams profile HMMs can be found in Supplementary Figure S1. We first evaluated the ability of each profile HMM to recruit the sequences used to initially train the profile HMMs. Protein sequences were recruited into an AR family or sub-family if they covered greater than 80% of the profile HMM and had an e-value score of less than $1x10^{-50}$. These global coverage and e-value scores were optimized to achieve maximum precision and recall. The distribution of precision and recall across all Resfams using these universal significance thresholds shows low precision for individual protein families (Supplementary Fig. S1A). These results indicate that a large number of AR families are structurally similar, highlighting an important challenge in predicting AR functions from sequence.

To improve Resfams prediction accuracy, we optimized profile specific gathering thresholds, which set an inclusion bit score cut-off for a protein sequence alignment on a profile-by-profile basis (Supplementary Fig. S1B). Finally, we tested the prediction accuracy of these optimized Resfams families on AR proteins from ARDB (Liu and Pop, 2009) not used in training of the original profile HMMs (Supplementary Fig. S1C). These recruited protein sequences were subsequently incorporated into the corresponding Resfams protein families, resulting in the final database of AR profile HMMs used for all further analysis in this study.

Resfams profile HMM Annotation from Microbial Sequence Alone

To test the ability of Resfams to accurately distinguish between AR and all other genomic functions, we used the well-curated UniProtKB/SwissProt database (Supplementary Fig. S7). The UniProtKB/Swiss-Prot protein database was downloaded on September 15, 2013, containing 540,732 reviewed proteins. The full set of reviewed proteins was aligned to the core Resfams database of profile HMMs (*Resfams.hmm*) using the hmmscan function of the HMMER3 (Finn *et al.*, 2011) software package using the following parameters: `--cut_ga, --tblout`. All proteins in the UniProtKB/Swiss-Prot database that were recruited to at least one Resfams AR protein family are represented on Supplementary Figure S7. Hits were designated as 'true positive' if the annotation in the UniProtKB/Swiss-Prot database matched the Resfams AR family annotation for the top hit. In addition, for all efflux/transporter and quinolone resistance AR mechanisms, 'true positive' hits required the protein to be designated as an 'Antibiotic Resistance' protein in the UniProtKB/Swiss-Prot database as these proteins are often also associated with other functions beyond resistance. All other hits were designated as 'false positive' hits.

Resfams AR protein families had less than 5% false discovery rate with the exception of ABC transporters, a class of transmembrane proteins with extremely diverse functions that are categorized together due to a common ATP binding domain. Because ABC transporters are difficult to predict *a priori* as resistance proteins, these Resfams annotations were excluded from analysis in the absence of functional confirmation. Excluding ABC transporters, Resfams very accurately predicted AR function from microbial genomes. Combined with its ability to annotate sequence-divergent AR proteins, this indicates that Resfams is adept at predicting AR without the need for a functional verification assay (e.g. functional metagenomics). The accuracy of Resfams profile HMMs for predicting AR function from microbial sequence alone is supported by the results obtained from functional metagenomic selections of the human gut and soil microbiotas. For example, we found that AR to tetracycline is mediated almost exclusively by tetracycline MFS efflux pumps in the soil microbiota and by ribosomal protection (TetM/TetO/TetW/TetS) in the human gut microbiota (Fig. 3). Resfams profile HMMs predict the same profile of tetracycline resistance mechanisms across habitats obtained from functional selections in bacterial genomes. Conversely, pairwise sequence alignment to AR specific databases incorrectly predicts enrichment of all tetracycline resistance mechanisms in the human gut versus soil (Fig. 5b).

Comparison of Resfams HMMs to BLAST to AR-specific databases

For comparison of Resfams HMMs to BLAST to AR-specific databases in functional metagenomic selections (Fig. 1), we used assembled contigs from functional metagenomics studies number 1 (MDR soil isolates) and 2 (pediatric gut resistome) described in Supplementary Table S2. A total of 161 assembled contigs from the MDR soil isolate resistome study and 3,692 assembled contigs from the pediatric gut resistome study were used for method comparison. Open reading frames were predicted in the assembled contigs using the stand-alone version of MetaGeneMark (Zhu *et al.*, 2010) with default parameters. Within each functional selection investigated, all proteins 100% identical over the length of the shorter sequence were collapsed into a single sequence using CD-HIT with the following parameters: `-c 1.0 -aS 1.0 -g 1-d 0`. The longest protein in the identified cluster was retained for downstream analysis. A total of 281 and 9,795 proteins were identified in the MDR soil isolates and pediatric gut resistome study, respectively. The same predicted protein sequences for each study were then annotated using either the full Resfams database of profile HMMs (*Resfams-full.hmm*) or BLAST to AR-specific databases as described above. The hand curated annotations for the MDR soil isolates study previously reported in Forsberg *et al.*, 2012 were used as a gold standard for that study for comparison of BLAST and Resfams HMMs. A recent report using pairwise sequence alignment to the ARDB (Liu and Pop, 2009) concluded that AR is highly enriched in the human gut as compared to natural environments, such as the soil (Hu *et al.*, 2013). In contrast, we find no statistical difference between the total AR using either functional selection data from metagenomes or sequenced isolate genomes. This emphasizes that studies of AR in microbial genomes and communities and comparisons across habitats requires functional or consensus-based annotation methods in order to provide a complete, unbiased representation of AR reservoirs.

Resistome analysis using functional metagenomic selections

Functional selections using antibiotics common to all three functional metagenomic selections described (see Methods) were used for comparative resistome analysis (Penicillin, Piperacillin, Tetracycline, Chloramphenicol, and Gentamicin). Contigs were assembled using PARFuMS (Forsberg *et al.*, 2012) and open reading frames were predicted using MetaGeneMark (Zhu *et al.*, 2010) as described above. Within each sample, all proteins 100% identical over the length of the shorter sequence were collapsed into a single sequence using CD-HIT with the following parameters: -c 1.0 ÐaS 1.0 Ðg 1-d 0. All proteins over 350bp and unique within sample were then used for downstream analysis. All proteins were then annotated using Resfams HMMs as described above, resulting in a total of 3,099 AR proteins used for comparative resistome analysis (64, MDR soil isolates; 1,082, pediatric gut resistome, 1,953; soil resistome). A count matrix of unique protein sequences per Resfams family annotation for each resistome sample was generated by summing unique annotation counts across all antibiotic selections for a sample and normalizing them by metagenomic library size (Supplementary Table S2). The normalized AR protein count table was then used to generate Bray-Curtis and binary Jaccard distance matrices and perform principal coordinate analysis (PCoA) using the `beta_diversity.py`

and `principal_coordinates.py` scripts (QIIME (Caporaso *et al.*, 2010)). The significance of clusters by was determined using ANOSIM and performed using the `compare_categories.py` script (QIIME (Caporaso *et al.*, 2010)). Random forests analysis was performed using the `supervised_learning.py` script (QIIME (Caporaso *et al.*, 2010) and randomForest R package (Liaw and Wiener, 2002)) to determine the Resfams families that most discriminate resistomes between habitats. PCoA plots were plotted along with the six most discriminating Resfams families as determined by random forests analysis as biplots for bray-curtis distance matrix. Biplot positions were calculated as the weighted average of the coordinate positions of all samples along the first two PCoA axes, where the weights are the relative abundances of the Resfams family. The size of the biplot points represents the aggregate abundance of the Resfams family across all samples. A bipartite network (Fig. 3) was generated from the normalize AR protein count table using the `make_bipartite_network.py` script (QIIME (Caporaso *et al.*, 2010)) and then visualized using Cytoscape (Shannon *et al.*, 2003) version 3.0.2 using the edge weighted spring embedded format.

Generation of binary heatmaps for genome comparisons

Binary heatmaps corresponding to genome annotations (e.g., Fig. 4) were created using the Heatplus package in R. Sections of the heatmap were colored and outlined if there was a significant enrichment of that AR mechanism (listed in Supplementary Table S1) as determined by Fisher's Exact Test ($P < 0.01$; Supplementary Table S5). If there was not a significant enrichment, the heatmap was colored black.

References

Caporaso, JG, Kuczynski, J, Stombaugh, J, Bittinger, K, Bushman, FD, Costello, EK, *et al.* (May 2010). QIIME allows analysis of high-throughput community sequencing data. *Nature methods* **7:** 335–336.

Finn, RD, Clements, J, Eddy, SR. (July 2011). HMMER web server: interactive sequence similarity searching. *Nucleic acids research* **39:** W29–37.

Forsberg, KJ, Reyes, A, Wang, B, Selleck, EM, Sommer, MOA, Dantas, G. (Aug. 2012). The shared antibiotic resistome of soil bacteria and human pathogens. *Science (New York, N.Y.)* **337:** 1107–1111.

Hu, Y, Yang, X, Qin, J, Lu, N, Cheng, G, Wu, N, *et al.* (2013). Metagenome-wide analysis of antibiotic resistance genes in a large cohort of human gut microbiota. *Nature communications* **4:** 2151.

Liaw, A, Wiener, M. (2002). Classification and regression by randomforest. *R News* **2:** 18–22.

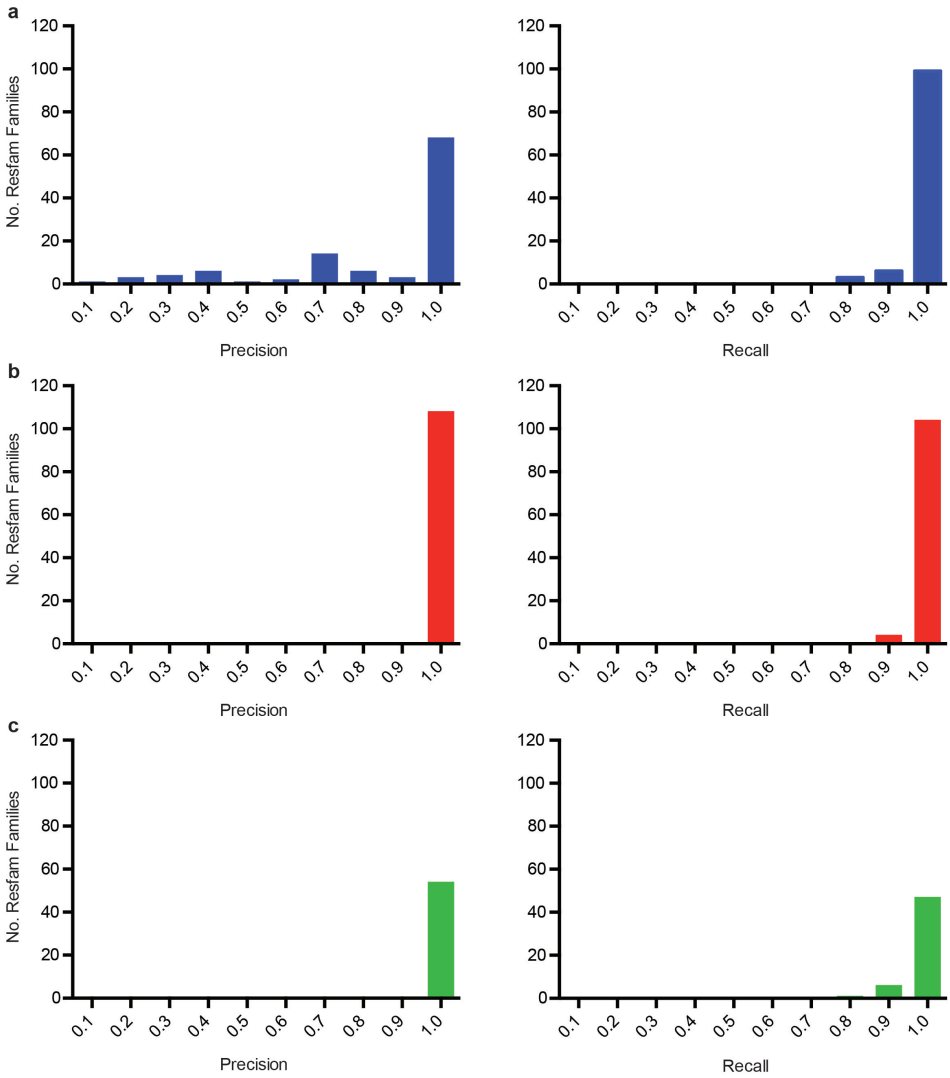
Liu, B, Pop, M. (Jan. 2009). ARDB–Antibiotic Resistance Genes Database. *Nucleic acids research* **37:** D443–7.

McArthur, AG, Waglechner, N, Nizam, F, Yan, A, Azad, MA, Baylay, AJ, *et al.* (July 2013). The comprehensive antibiotic resistance database. *Antimicrobial agents and chemotherapy* **57:** 3348–3357.

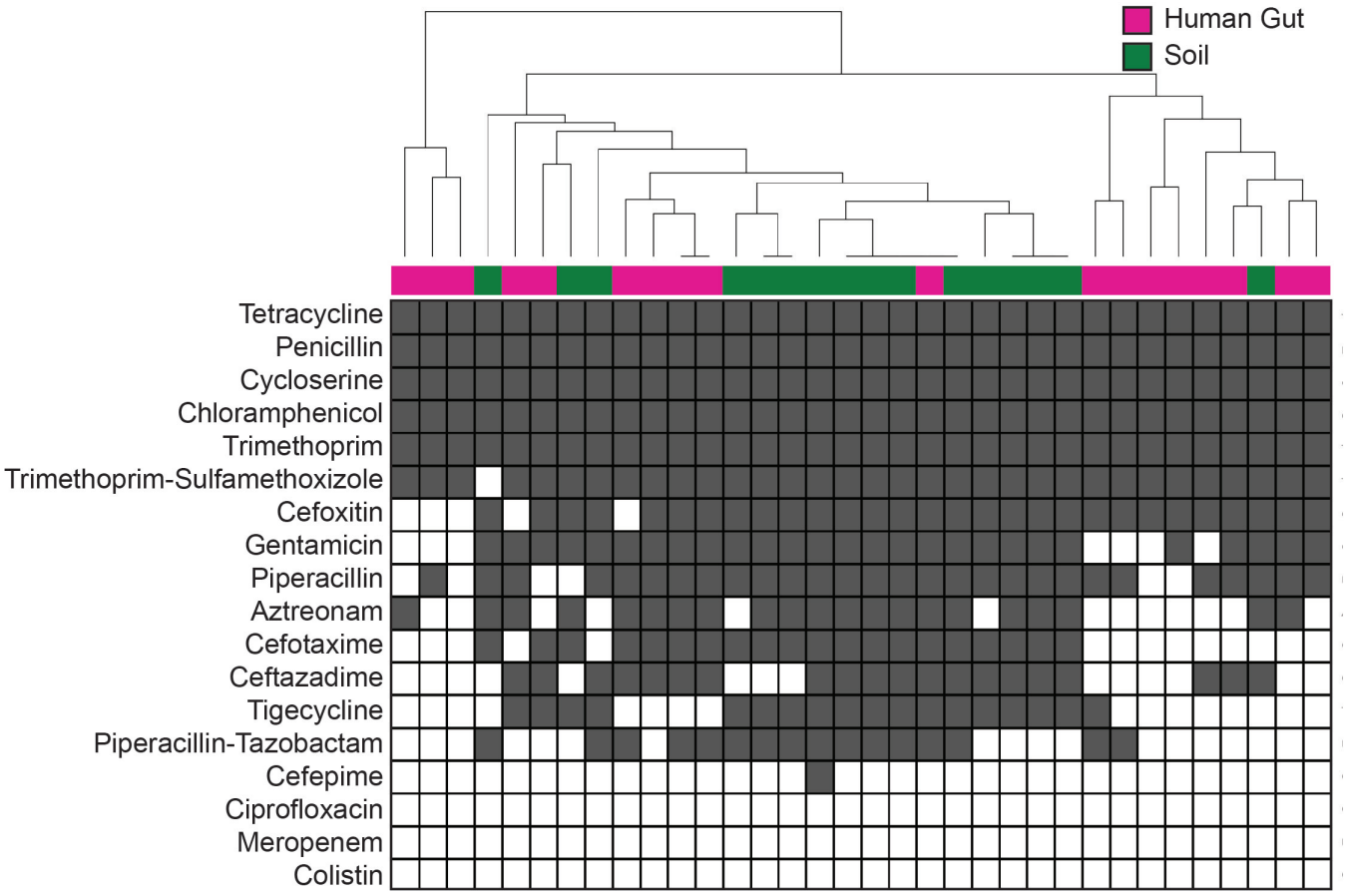
Shannon, P, Markiel, A, Ozier, O, Baliga, NS, Wang, JT, Ramage, D, *et al.* (Nov. 2003). Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome research* **13:** 2498–2504.

Thai, QK, Bös, F, Pleiss, J. (2009). The Lactamase Engineering Database: a critical survey of TEM sequences in public databases. *BMC genomics* **10:** 390.

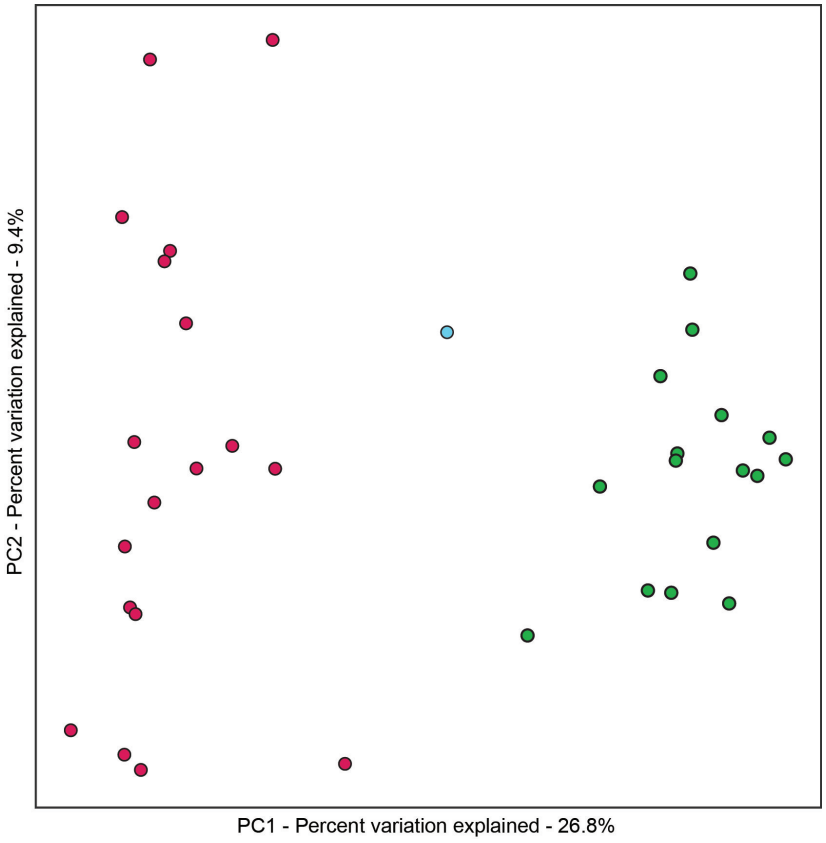
Zhu, W, Lomsadze, A, Borodovsky, M. (July 2010). Ab initio gene identification in metagenomic sequences. *Nucleic acids research* **38:** e132.



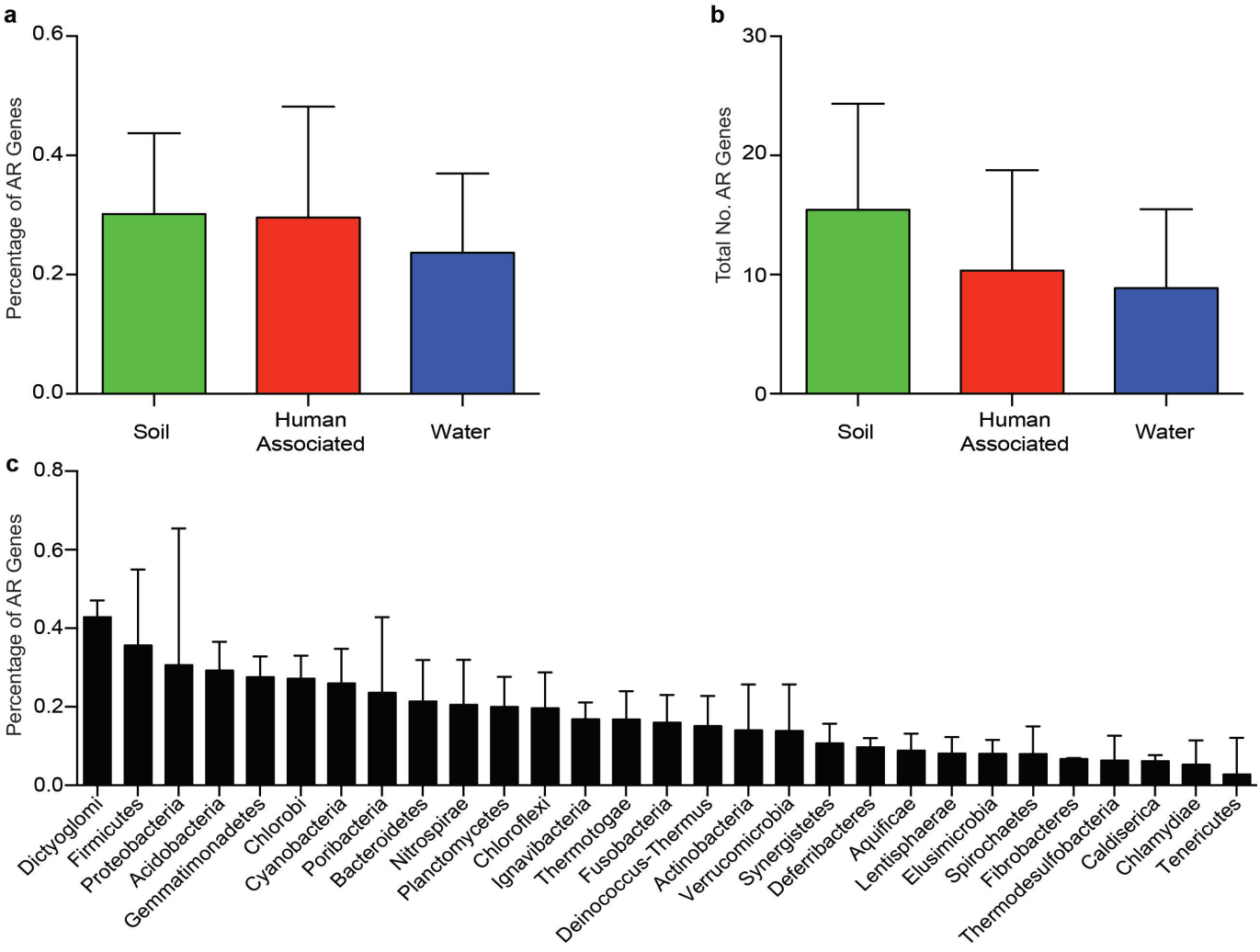
Supplementary Figure S1: Resfams family precision and recall distributions. The x-axis represents the precision or recall values while the y-axis represents the number of Resfams families of a given value. (A) Precision and recall distributions of Resfams families on the proteins used to train the original profile using universal thresholds. (B) Precision and recall distributions of Resfams families on the proteins used to train the original profile HMMs using defined gathering thresholds. (C) Precision and recall distributions of Resfams families on an independent validation set of antibiotic resistance proteins not used to train the Resfams profile HMMs using defined gathering thresholds.



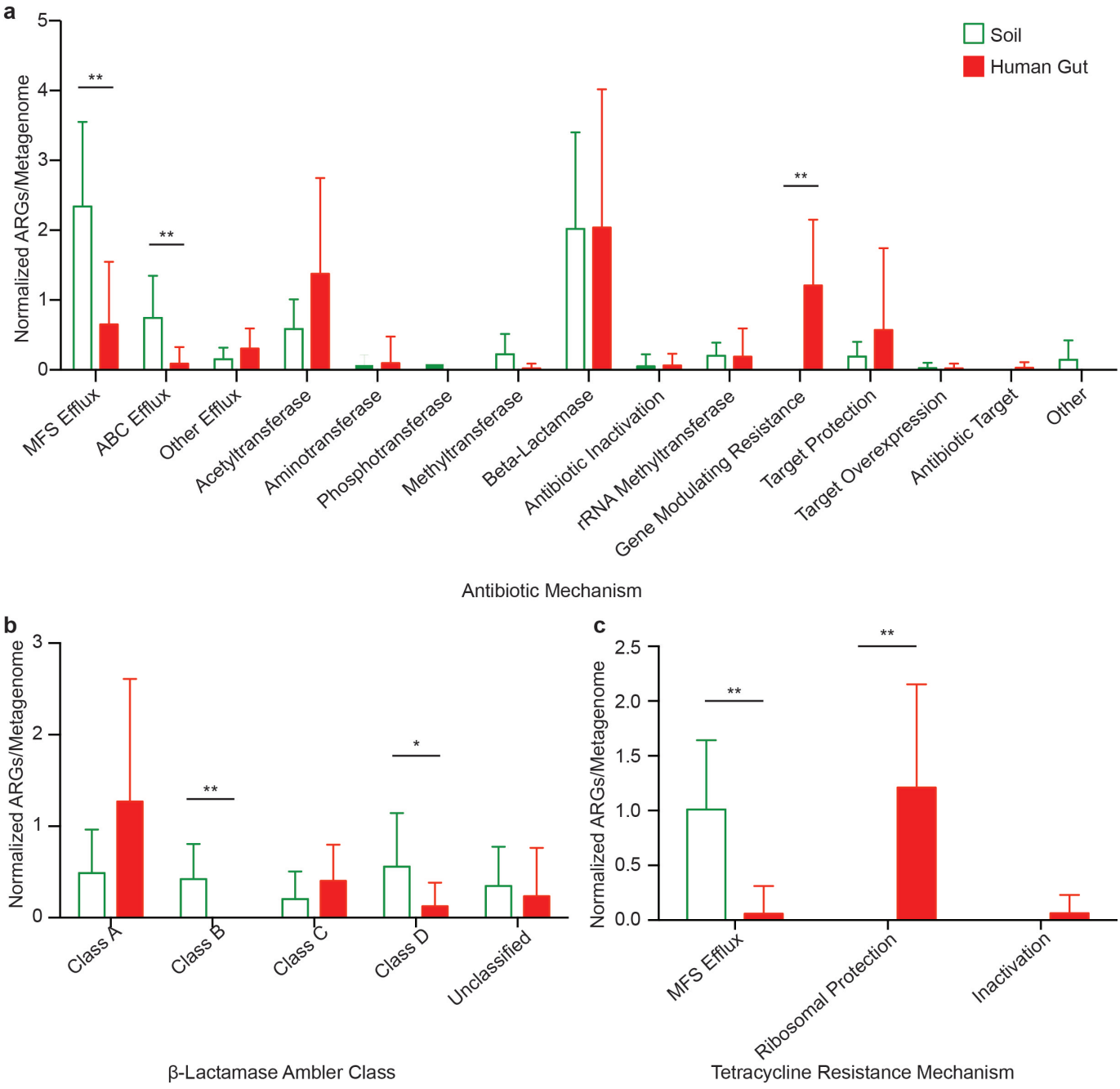
Supplementary Figure S2: Phenotypic resistance profiles across ecologies. Binary heatmap (gray, resistance observed; white, no resistance) of phenotypic profiles of 16 soil microbiota (green) and 18 human gut microbiota (magenta) to 18 antibiotics. Samples are clustered using the Jaccard Index and resistance profiles are herarcically clustered.



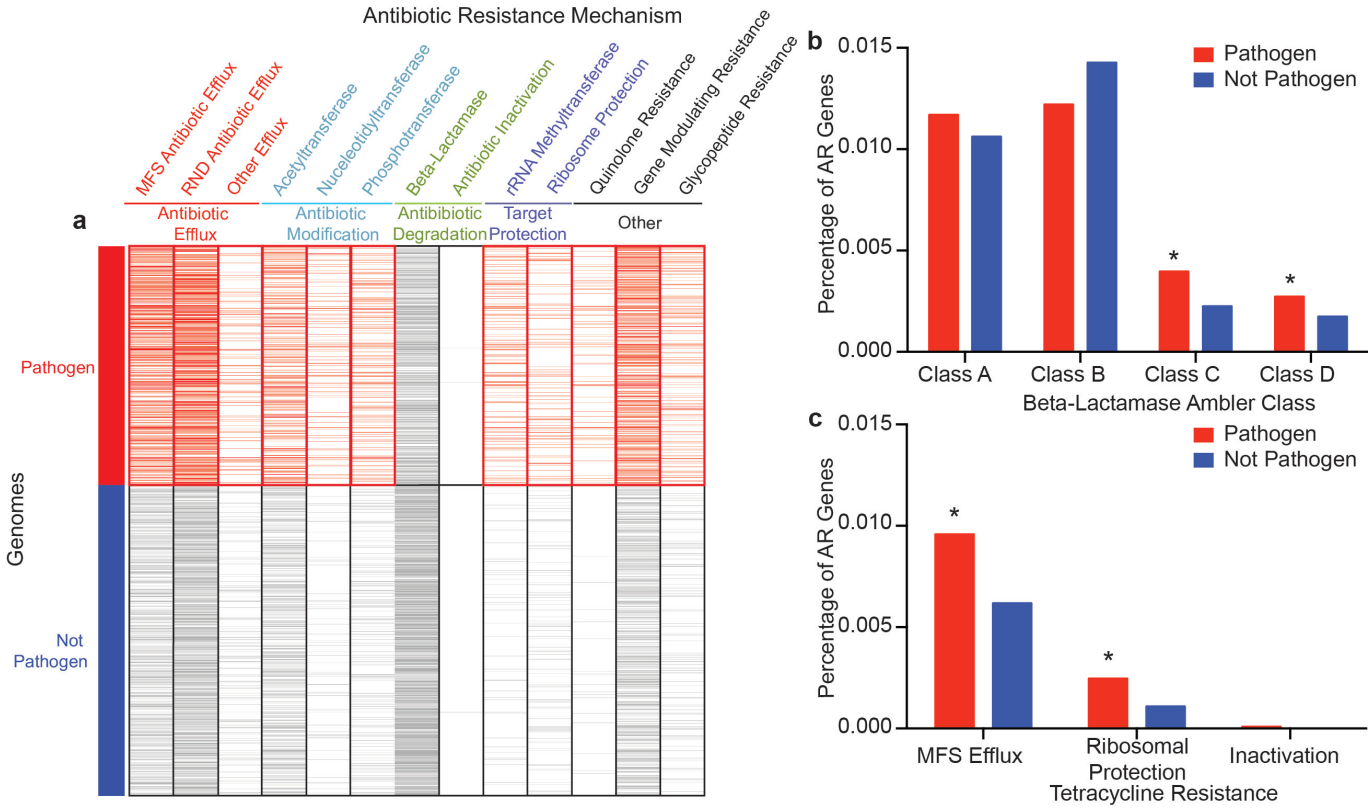
Supplementary Figure S3: Principal coordinate analysis (PCoA) plots of functional metagenomic selections. PCoA plot depicting binary Jaccard distances between resistomes of soil microbiota (green), human gut microbiota (magenta), and MDR soil isolates (blue), calculated using unique ARG counts generated by Resfams annotations. Resistomes of different ecologies cluster separately ($P < 0.001$, ANOSIM).



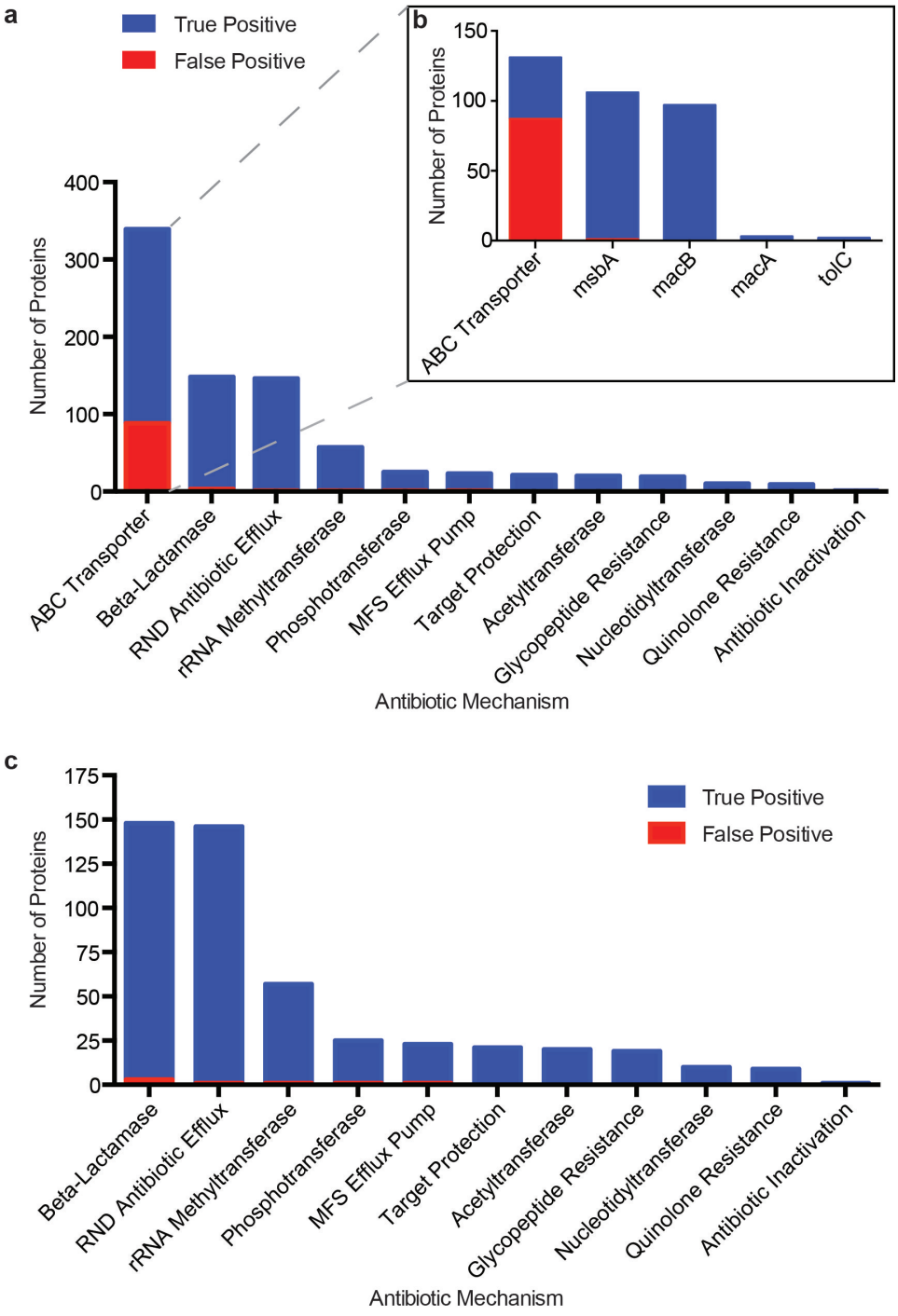
Supplementary Figure S4: Distribution of total antibiotic resistance (AR) genome composition across habitat and bacteria phyla. The total AR genome composition was calculated by taking the total number of genes annotated by Resfams profile HMMs as AR genes in the genome divided by the total number of genes in the genome. The total AR genome composition was averaged across (A) habitats and (C) bacterial phyla. Error bars represent one standard deviation. The raw number of total AR genes per genome was averaged across all genomes by (B) habitat. Error bars represent one standard deviation.



Supplementary Figure S5: Enrichment of antibiotic resistance (AR) mechanisms in functional metagenomic selections. Distribution of Resfams family AR genes identified in functional metagenomic selections of the soil (green) and human gut (red), including (A) all mechanisms, (B) β-lactamase Ambler classes, and (C) tetracycline resistance mechanisms. The normalized count of unique antibiotic resistance genes per metagenome investigated in each mechanistic category along the x-axis is depicted along the y-axis (*P $<$ 0.05, **P $<$ 0.001; Student's t-test).



Supplementary Figure S6: Enrichment of antibiotic resistance functions in pathogens. (A) Binary heatmap of resistomes organized by pathogen status of 2,966 genome sequenced bacterial isolates. The heatmap is colored by enrichment of a particular AR mechanism within pathogenic or non-pathogenic organisms ($P < 0.01$, Fisher's exact). Enrichment of (B) β-lactamase ambler class and (C) tetracycline resistance functions within pathogenic or non-pathogenic organisms (*$P < 0.01$, Fisher's exact).



Supplementary Figure S7: Resfams profile HMM annotation of 540,732 proteins in the UniProtKB/Swiss-Prot curated database. The y-axis represents the total number of proteins recruited to Resfams familes in each of the antibiotic resistance mechanism categories represented along the x-axis. True positive hits (blue) both match the annotation category and are flagged as antibiotic resistant protein forms in the UniProtKB/Swiss-Prot Database. (A) Recruitment distribution of all Resfams families. (B) ABC Transporter mechanistic class broken down into the individual Resfams families, showing that the majority of false positive arise from the general ABC Transporter profile HMM. (C) Recruitment distribution of all Resfams families used in annotation of genomes in absence of functional assay.