Supplementary Materials for

Functions predict horizontal gene transfer and the emergence of antibiotic resistance

Hao Zhou, Juan Felipe Beltrán, Ilana Lauren Brito*

*Corresponding author. Email: ibrito@cornell.edu

Published 22 October 2021, Sci. Adv. 7, eabj5056 (2021)
DOI: 10.1126/sciadv.abj5056

The PDF file includes:

Supplementary Text
Figs S1 to S18
Tables S1 to S8
Supplementary Text

Details of the models used in specific figures

**Fig. 1B:** Each curve represents the average of 5 independent experiments ± one standard deviation (gray), each consisting of 500 randomly selected nodes with 7,849.2 ±714.06 positive edges and an equal number of negative HGT edges for the test set, and training data consisting of 8,433.6 ±125.2 genomes with 68,323.6 ±8,368.95 positive edges, such that there was no overlapping taxa between test and training datasets.

**Fig. 1C:** Each curve represents the average of 5 independent experiments ± one standard deviation (gray), using the same training and test nodes as in Fig. 1B. To ensure fair comparison, each dataset set was subsampled to the lowest number of edges, i.e. saline, with 5,069 edges. The “All” dataset for the RF and LR model consists of the same 2,000 subsampled from each environment. All test sets consisted of a balanced set of 250 HGT-positive and an equal number of HGT-negative edges, such that there was no overlapping taxa between test and training datasets. AUC values are shown. The number of training and test nodes per environment was: All: 3084.6 ±38.44, and 404.2 ±5.46; human: 1,846.6 ± 53.60 and 277.2 ±19.67; animal: 2,284.8 ±42.41 and 303.8 ±7.68; non-saline: 2,918.2 ±63.63 and 412.4 ±10.78; saline: 1,893.2 ±49.39 and 269.0 ±10.86, and plant: 2,055.8 ±44.80 and 288.8 ±9.87, respectively.

**Fig 1D:** These iterations were performed on the same training and test sets as in Fig. 1B, although the least censored model includes only 3,140 ±283.61 positive and an equal number of negative edges in each test set.

**Fig 3A:** Each boxplot represents the average of 5 independent experiments, each consisting of 500 randomly selected nodes with 3,395.7 ± 125.22 ARG-HGT-positive edges and an equal number of edges either lacking HGT or edges with HGT that do not involve ARGs for the test set. The training data consisted of 8433.6 ± 125.23 genomes with 29874.6 ± 4917.06 ARG-HGT positive edges, such that there was no overlapping taxa between test and training datasets. The middle of the boxplot is the median and edges are quartiles.

**Fig. 3D:** Each boxplot represents the average of 5 independent experiments, each consisting of 500 randomly selected genomes with HGT-positive edges involving class-specific ARGs and an equal number of HGT-positive edges excluding edges involving ARGs to that specific class, such that there was no overlapping taxa between test and training datasets. The middle of the boxplot is the median and edges are quartiles.
Fig. S1. Basic information about the HGT network.

(A) All genomes of the HGT network (12,518 genomes) according to their phyla.
(B) The number of genomes per taxonomic species name in our network after filtering.
(C) The 16S rRNA sequence similarity between all pairs in our genome (not including those over 97%).
(D) For 16S rRNA distances (sequence dissimilarity) above 3%, the rate of HGT per 100 genomes is plotted.
Fig. S2. Precision-Recall (PR) curves for models in Fig. 1B.

PR curves for the Lasso and Random Forests models shown in Fig. 1C. AUCs for the PR curves are shown.
Fig. S3. 16S rRNA similarity and KO composition correlate. A dot plot showing each pair of organisms according to their 16S rRNA similarity and the normalized number of shared KOs. The Spearman rank correlation is provided.
Fig. S4. KO assignments by organism.
(A) The distribution of the number of open reading frames (ORFs) identified in each genome (n=12,518) is plotted.
(B) The distribution of the number of identified ORFs with KO annotation per genome is plotted.
(C) The distribution of the percentage of ORFs with KO assignments per genome is plotted.
(D) The distribution of the number of KOs per genome is plotted.
(E) The distribution of the number of genomes per KO is plotted.
Fig. S5. The distribution of organisms across environments.

(A) The percent identity between the full-length 16S rRNA sequences from the isolates and the V4 region of the 16S rRNA in the EMP (n=9,439).

(B) The length of alignment between the full-length 16S rRNA sequences from the isolates and the V4 region of the 16S rRNA in the EMP.

(C) A Venn diagram showing the number of organisms that were identified, according to matching 16S rRNA V4 sequences, to human, animal, plant, saline and non-saline environments.

(D) The distribution of the percentages of ORFs per genome that could be annotated with KEGG is plotted according to environment.
**Fig. S6. Ecological co-occurrence correlates with HGT.**

(A) A boxplot showing the ecological co-occurrence values, calculated as SparCC correlations, between pairs of genomes (n=9,439) with and without observed HGT. Numbers of amplicon sequencing samples in different environments are provided. The middle of the boxplot is the median and edges are quartiles. **** represents Welch t-test p-value of less than 10^{-4}.

(B) Full-length 16S rRNA distance calculated between pairs of genomes, for which ecological co-occurrence data is available, is plotted against their ecological co-occurrence, as measured using SPARCC. The Spearman’s rho and p-value are provided.

(C) Normalized shared KO content for pairs of genomes, for which ecological co-occurrence data is available, is plotted against their ecological co-occurrence, as measured using SPARCC. The Spearman’s rho and p-value are provided.
Fig. S7. Neither ecological co-occurrence and phylogenetic distance, alone or in combination, predict HGT as accurately as functional gene content alone.

(A) PR curves for the models shown in Fig. 1C. AUCs for the PR curves are shown.

(B) ROC curves for logistic regression models using SparCC correlations between organisms’ distribution, inferred using near-identical (>99% sequence similarity) 16S V4 rRNA sequences across environmental samples available through the Earth Microbiome Project (EMP). AUCs for the ROC curves are shown. As in Fig. 1C, datasets were subsampled to the lowest number of edges, i.e. saline, with 5,069 edges. The “All” dataset for the RF and LR model consist of the same 2,000 subsampled from each environment.

(C) PR curves for the models in (B). AUCs for the PR curves are shown.
Fig. S8. Depiction of the GCN models

(A) An overview of the GCN model is provided. First, a test set of 500 genomes is chosen. Genomes with the same species name and/or ≥ 97% rRNA similarity (Overlap) are removed from the training set. The GCN model is shown, taking in an adjacency matrix (A) and a node attribute matrix (X(0)). The model consists of two hidden layers in which a weighted parameter matrix and edge and node attributes are learned. Predicted probabilities are computed using a sigmoid activation function. GraphLIME is used to examine the importance of edge attributes (shared or discordant KO features) for subgraphs of a randomly chosen set of HGT-positive edges. Consistently important KOs are identified by examining important KOs across edges across 5 GCN experiments.

(B) Uncensored networks of the test set are included are not used in the model and are only used upon calculating predictions. Uncensored edges may include those between test set nodes or those between test and training set nodes.
Fig. S9. Precision-Recall (PR) curves for models in Fig. 1D.
PR curves for the LR and RF models shown in Fig. 1D. AUCs for the PR curves are shown.
Fig. S10. Gini importance distribution for the RF model predicting HGT.

(A) The distribution of average Gini importances for features in 5 RF models predicting HGT using solely KOs, as shown in Fig. 1C. The inset shows the distribution of importance values over 0.001.

(B) ROC curves considering 26, 82, 153 important KOs, defined by having Gini importances over 0.004, 0.002, and 0.001, respectively. Each curve represents the average of 5 independent experiments ± one standard deviation (gray), each consisting of 500 randomly selected nodes for the test set, each, and training data consisting of 8433.6 ± 125.2 genomes. AUC values are shown.

(C) Precision-recall curves for the RF models shown in (B). AUC values are shown.
Fig. S11. Model performance on predicting inter-phylum transfers.

(A) Predicted probability of HGT from five independent experiments combined is plotted for HGT-positive and HGT-negative inter-phylum transfers using a logistic regression model, with 16S rRNA similarity as the sole input. A total of 6,735 HGT-positive and 6,735 HGT-negative edges were tested. The AUROC is provided.

(B) PR curves for the GCN models shown in Fig. 2A. AUCs for the PR curves are shown.
Fig. S12. Network characteristics and HGT predictions for HGT-negative edges.

(A) The number of common neighboring genomes for each pair of nodes with observed HGT (positive edges) is plotted against their predicted probability of HGT, as determined by the GCN with 60% uncensored edges shown in Fig. 2C. Spearman’s rho and p-value are provided.

(B) The number of common neighboring genomes for each pair of nodes without any observed HGT (HGT-negative edges) is plotted against their predicted probability of HGT, as determined by the GCN with 60% uncensored edges shown in Fig. 2C. Spearman’s rho and p-value are provided.

(C) The minimum degree (i.e. the minimum number of HGT partners) of each pair of genomes with observed HGT (HGT-positive edges) is plotted against their predicted probability of HGT, as determined by the GCN with 60% uncensored edges shown in Fig. 2C. Spearman’s rho and p-value are provided.

(D) The minimum degree (i.e. the minimum number of HGT partners) of each pair of genomes without observed HGT (HGT-negative edges) is plotted against their predicted probability of HGT, as determined by the GCN with 60% uncensored edges shown in Fig. 2C. Spearman’s rho and p-value are provided.
Fig. S13. PR values and important KOs from predictions of HGT involving ARGs.

(A) Area under the PR curves for models predicting HGT involving ARGs is plotted for the LR model using only 16S rRNA sequence similarity, a Lasso model using the presence/absence of KOs for each genome, a RF model using KOs, and GCN models using KOs with decreasing censorship of the network (0% to 60% edges), corresponding to Fig. 3A. Each boxplot represents the average of 5 independent experiments. The middle of the boxplot is the median and edges are quartiles. Mean AUPRC values (μ) are provided.

(B) Important KOs (Average Gini importance values > 0.006) for the RF model used to predict HGT involving any ARGs. For each feature, the percent of 10,000 randomly sampled ARG-HGT-positive and HGT-positive edges without ARG or HGT-negative edges where the KO is present in both, one, or neither genome is shown.

(C) Genomes linked by 46 ARG-specific HGT-negative randomly chosen edges from the RF ARG-HGT model depicted in Fig. 3B are plotted according to their genus, phylum and origin of isolation, where available. Organisms obtained from the gut microbiomes of individuals with inflammatory bowel disease, diabetes, or obesity were grouped with human pathogens.
**Fig. S14. PR and metrics of ARG-specific HGT models.**

(A) Distribution of HGT edges involving ARGs to one or more ARG classes. This includes ARG classes in addition to the 8 classes for which we performed multiclass predictions.

(B) Area under the PR curves for the experiments described in Fig. 3D. Boxplots show median and quartile values. Mean AUPRC values (μ) are provided. Below, for each gene, according to ARG class, the log(number of edges), and the number of genomes harboring that ARG is plotted. For each ARG class, the degree per genome is plotted (bottom) and colored according to phylum. Only ARGs involved in at least 5 HGT events are included in the plot.
Fig. S15. Area under the PR curves for APEC and *A. baumanii* isolates HGT-ARG class-specific predictions.

(A) Area under the PR curves for the class-specific ARG-HGT RF models for APEC isolates shown in Fig. 4A. Boxplots show median and quartile values over 5 experiments. Average AUROC values (μ) are provided.

(B) Area under the PR curves for the class-specific ARG-HGT RF models for *A. baumanii* isolates shown in Fig. 4F. Boxplots show median and quartile values over 5 experiments. Average AUROC values (μ) are provided.
Fig. S16. Predictions of tetracycline and beta-lactam ARG transfer in *Neisseria gonorrhoeae* isolates.

(A) A heatmap showing normalized shared KOs for 71 *N. gonorrhoeae* isolates (top left) and their ARG-specific HGT network Jaccard similarity. Isolates are clustered according to their network similarities. The phylogenetic clade for each isolate is plotted adjacent to the heatmap. Two genomes with divergent ARG-specific HGT networks are noted.

(B) Area under the ROC curves and PR curves are plotted for RF classification of tetracycline and beta-lactam ARG transfers in the *N. gonorrhoeae* isolates. Mean AUROC and AUPRC were calculated based on predicted probabilities of a balanced test set with equal number of positive and negative edges across 5 experiments.
Fig. S17. HGT is predictable in small environment-specific datasets.

(A) The area under the precision-recall curve is plotted for the models in Fig. 3C for the ocean (O), soil (S), plant (P), and human microbiome (H) datasets.

(B) For 16S rRNA distances between 3-20% (sequence dissimilarity), the rate of HGT per 100 genomes is plotted for the ocean, soil and human microbiome datasets. The plant dataset was not included here due to the low overall rate of HGT events (only 26 observed). The curves were fitted to a one phase decay model with a adjusted coefficient of determination ($R^2$) of 0.8081, 0.9113 and 0.2342 for ocean, soil and human gut, respectively.
Fig. S18. Receiver-operator curves (ROC) and Precision-Recall (PR) curves for the models using isolates from individual hosts.

(A) ROC curves for LR and RF models using the same data as in Fig. 5A (human gut microbiome isolates), each model trained on our original dataset (11,386 genomes, 84,363 edges, such that genomes that overlapped at the species-level were removed from the training set) and tested on isolates from individual human hosts. AUC values are shown.

(B) PR curves for the LR and RF models using the same data as in Fig. 5A (human gut microbiome isolates). AUCs for the PR curves are shown.
Supplemental Tables

**Table S1. 12,518 genomes used in this study.**
Genomes are listed according to their strain name, NCBI ID and taxonomy.

**Table S2. Consistent KOs (GraphLime coefficients > 0 in 30/500 edges in over 3 of 5 experiments) identified using GraphLIME.**
For each function, the KO number, gene name(s), gene definition, KEGG pathways, Percentage of HGT-positive edges and HGT-negative edges, respectively, where either both genomes in the pair share, are discordant, or both lack the function, and the number of genomes with that specific function.

**Table S3. Important KOs (Average Gini importance > 0.001) in the RF models.**
KO number, gene name(s), gene definition, KEGG pathways, Percentage of HGT-positive edges and HGT-negative edges, respectively, where either both genomes in the pair share, are discordant, or both lack the function, and the number of genomes with the specific function.

**Table S4. Important KOs associated with ARG-specific HGT.**
Important KOs are listed for the ARG-specific HGT RF classifier (Average Gini importance > 0.001).

**Table S5. HGT-ARG-negative edges with high predictions (over 0.9) or randomly chosen.**
Edges with ARG-HGT predictions over 0.9 (“high prediction), but with no observed HGT, are listed according to the genome names, genome identifiers (NCBI or PATRIC), 16S rRNA percent identity, predicted probability of RF predictor and source of isolation. Additionally, 46 ARG-HGT negative edges were randomly chosen (“random selection”) and listed according to the same features.

**Table S6. Top 1% important KOs associated with each antibiotic resistance class RF classifier.**
The top 1% important KOs are provided for the multi-class models in Fig. 3B. For each feature, the KO number, antibiotic resistance gene class, gene name(s), description, KEGG pathway and BRITE hierarchies, percentage of 10,000 randomly sampled ARG-class specific HGT versus HGT involving ARGs to other classes where the KO is present in both, one, or neither genome is shown.

**Table S7. APEC, A. baumanii and N. gonorrhoeae genomes used.**
Genomes used in Fig. 4, and Fig. S15 and S16.

**Table S8. Genomes used in Fig. 5.**
High quality genomes from environment-specific datasets are listed with their accession numbers according to study.