Metagenomic targets of balancing selection in the human gut

Andrew H. Moeller[1]

[1]Cornell University

November 17, 2021

Abstract

Bacteria in the human gut contend with numerous fluctuating environmental variables, including bouts of extreme selective agents like antibiotics. Theory predicts that oscillations in the adaptive landscape can impose balancing selection on bacterial populations, leaving characteristic signatures in the sequence variation of functionally significant genomic loci. Despite their potential importance for gut bacterial adaptation, the metagenomic targets of balancing selection have not been identified. Here, I present population genetic evidence that balancing selection maintains allelic diversity in multidrug efflux pumps of multiple predominant bacterial species in the human gut metagenome. Metagenome wide scans of 566,958 core open reading frames (CORFs) from 287 bacterial species represented by 118,617 metagenome assembled genomes (MAGs) indicated that most CORFs have been conserved by purifying selection. However, dozens of CORFs displayed positive Tajima's D values that deviated significantly from their species' genomic backgrounds, indicating the action of balancing selection. The AcrB subunit of a multidrug efflux pump (MEP) in Bacteroides dorei displayed the highest Tajima's D of any CORF, and AcrB and other MEPs from a diversity of bacterial species were significantly enriched among the CORFs with the highest Tajima's D values. Crystal structures indicated that the regions under balancing selection bind tetracycline and macrolide antibiotics. Other proteins identified as targets of balancing selection included synthases, hydrolases, and ion transporters. Intriguingly, bacterial species experiencing balancing selection were the most abundant in the human gut based on metagenomic data, further suggesting fitness benefits of the allelic variation identified.

Metagenomic targets of balancing selection in the human gut

Andrew H. Moeller[1,2]

[1]Department of Ecology and Evolutionary Biology, Cornell University, Ithaca, NY 14805

[2]corresponding author: andrew.moeller@cornell.edu

Bacteria in the human gut contend with numerous fluctuating environmental variables, including bouts of extreme selective agents like antibiotics. Theory predicts that oscillations in the adaptive landscape can impose balancing selection on bacterial populations, leaving characteristic signatures in the sequence variation of functionally significant genomic loci. Despite their potential importance for gut bacterial adaptation, the metagenomic targets of balancing selection have not been identified. Here, I present population genetic evidence that balancing selection maintains allelic diversity in multidrug efflux pumps of multiple predominant bacterial species in the human gut metagenome. Metagenome wide scans of 566,958 core open reading frames (CORFs) from 287 bacterial species represented by 118,617 metagenome assembled genomes (MAGs) indicated that most CORFs have been conserved by purifying selection. However, dozens of CORFs displayed positive Tajima's D values that deviated significantly from their species' genomic backgrounds, indicating the action of balancing selection. The AcrB subunit of a multidrug efflux pump (MEP) in *Bacteroides dorei* displayed the highest Tajima's D of any CORF, and AcrB and other MEPs from a diversity of bacterial species were significantly enriched among the CORFs with the highest Tajima's D values. Crystal structures indicated that the regions under balancing selection bind tetracycline and macrolide antibiotics. Other proteins identified as targets of balancing selection included synthases, hydrolases, and ion transporters.

Intriguingly, bacterial species experiencing balancing selection were the most abundant in the human gut based on metagenomic data, further suggesting fitness benefits of the allelic variation identified.

Introduction

The populations of bacteria that reside within the gastrointestinal tracts of humans experience a diversity of oscillating environmental variables. Within and among hosts, gut bacterial populations face fluctuating selection pressures imposed by variation in host diet (David et al., 2014), drug use (e.g., antibiotics) (Modi et al., 2014), and immunity (Schluter et al., 2020) as well as by variation in biotic interactions with other constituents of the microbiota (Coyte and Rakoff-Nahoum, 2019). These cyclic changes in the adaptive landscapes on which gut bacterial populations evolve may promote the maintenance of genetic diversity within gut bacterial species. However, the degree to which balancing/diversifying selection operates on gut bacterial genomes has not been widely investigated, and the genomic loci that represent the targets of such selective forces have not been identified.

Theory predicts that genes under balancing selection will display a greater number of pairwise sequence differences between copies in a population than expected under neutral evolution based on the number of polymorphic sites in the population. The difference between these values—the observed average number of pairwise differences (π_o) and the expected number of pairwise differences based on the number of segregating sites under neutrality (π_e)—provides a test statistic for balancing selection termed Tajima's D (Tajima, 1989). Recent advances in assembling bacterial genomes directly from metagenomic sequence data have generated unprecedented opportunities for interrogating the strength and genomic targets of balancing selection in the human gut microbiota (Pasolli, 2019). Metagenome assembled genomes are now available for nearly all of the bacterial species detected at appreciable abundances in the human gut microbiota and, for many species, multiple genomes from a diversity of strains have been assembled from metagenomes of numerous host populations and individuals.

Here, we analyzed 118,617 metagenome assembled genomes (MAGs) to identify the targets of balancing selection in 288 species of human gut bacteria. We find that gut bacterial genomes evolve primarily under purifying selection. However, a subset of loci displayed significant population genetic evidence of balancing selection. In multiple prominent gut bacterial species, these loci included coding regions for components of multidrug efflux pumps, which were overrepresented among the gene functions displaying the most significant evidence of balancing selection. Integrating comparative genomic analyses with metagenomic measurements of microbiota composition revealed that bacterial species whose genomes contain targets of balancing selection tend to be more abundant in the human gut than do other bacterial species, implying a relationship between the loci under selection and fitness. Cumulatively, these findings reveal adaptive genomic diversity maintained by balancing selection within gut bacterial species.

Materials and Methods

Identifying core genomes of gut bacterial species

To enable scans for balancing selection across the gut bacterial species of humans, all 154,723 metagenome assembled genomes from Pasoli *et al.* (2019) were downloaded from http://segatalab.cibio.unitn.it/data/Pasolli_et_al.html. Of these SGBs, bacterial SGBs that were represented by >100 genomes and retained for downstream analyses of intraspecific patterns of DNA polymorphism. To identify the core genomes of the well-represented SGBs, Open Reading Frame (ORF) Finder from the National Center for Biotechnological Information was used with default settings to identify all ORFs in each of the genomes. Next, CoreCruncher (Harris et al., 2021) was used with default settings to identify the core set of ORFS for each SGB. In this analysis, the largest genome from each SGB was chosen as the pivot genome from which to identify the core ORFs (CORFs). All CORFs were then annotated against the updated Clusters of Orthologous Groups database (Galperin et al., 2021).

Calculating Tajima's D

The CORFS for each SGB were aligned with MAFFT (Katoh et al., 2002) using default settings. Tajima's

D (Tajima, 1989) for each aligned set of CORFs was calculated with the 'tajimasD' function in the R package 'strataG' (Archer et al., 2017). P-values output by the 'tajimasD' function were corrected for multiple comparisons with the 'p.adjust' function in stats package in R based on the total number of ORFs per genome using the Benjamini and Hochberg method with the 'method = "fdr" ' option (Benjamin and Hochberg, 1995).

Phylogenetic analyses

Representative genomes identified by Pasoli et al. (2019) from each SGB were used to construct a phylogeny of SGBs. Alignments for phylogenetic analyses were generated using the Genome Taxonomy Database Toolkit (GTDB-Tk) (Chaumeil et al., 2020). Marker genes in each SGB representative genome were identified with 'gtdbtk identify' using default settings and aligned against the BAC120 reference gene set with 'gtdbtk align' using default settings. Alignments were then used for phylogenetic inference with IQTree2 (Minh et al., 2020). For these analyses, model search was constrained to only LG and WAG models of protein substitution, and 1000 ultrafast bootstrap replicates were performed. Phylogenetic tree off SGBs was visualized in the Interactive Tree of Life (iTOL) web interface (Letunic and Bork, 2021).

Statistical analyses

Significant differences in means between per-species Tajima's D estimates were tested in R with 'pairwise.t.test' in 'stats' using the flag 'p.adjust.method = "fdr" '. Phylogenetic Generalized Least Sqaures (PGLS) was conducted within the 'phytools' package (Revell, 2012) with the 'gls' and 'corBrownian' functions. Phylomorphospace plot was constructed within 'phytools' with 'phylomorphospace' using default settings. Tests for phylogenetic signal of genome-wide Tajima's D were conducted with 'phylosig' using the flag 'method = "k" '. Histograms were generated with the function 'hist' in the 'graphics' package in R. Skewness of distributions was calculated with 'skewness' in the 'e1071' package (Meyer et al., 2019) in R.

Analyses of protein structure and chemistry

To visualize the putative tertiary structure of proteins coded for by CORFs under balancing selection, the Protein Data Bank was searched for homologs with resolved structures at www.rcsb.org (Burley et al., 2021). Top hits from *Escherichia coli* was used for downstream analysis. To identify the putative location of CORFs under balancing selection within the *E. coli* protein complex, CORF sequences from the representative MAGs for SGBs were aligned against the sequence of the*E. coli* protein and visualized with the Mol* 3D Viewer. Disorder was estimated along the protein sequence with IUPred2 (Mészáros, et al., 2018) and hydrophobicity was calculated using the Kyte and Doolittle method (Kyte and Doolittle, 1982) with a window size of 21.

Relative abundance estimates and associations with Tajima's D

Relative abundances of all SGBs in the metagenomes of the Human Microbiome Project healthy human subjects cohort were estimated with CoverM using default settings. Significance of associations between these relative abundance estimates and Tajima's D values were tested with the function 'lm' in the 'stats' package in R.

Results

Intraspecific genomic comparisons of 288 bacterial species

Filtering the MAGs generated by Pasoli et al., (2019) for MAGs belonging to SGBs represented by >100 genomes yielded 118,617 metagenome assembled genomes (MAGs) with >50% completeness and <5% contamination as estimated by CheckM (Parks et al. 2015) belonging to 287 bacterial species-level genome bins (SGBs). From these genomes, CoreCruncher (Harris et al., 2021) identified 566,958 core open reading frames (CORFs). Tajima's D values and representative amino acid sequences for each of these CORFs are presented in Table S1.

Genome wide Tajima's D is associated with bacterial evolutionary history

Phylogenetic analyses of all SGBs in combination with intraspecific scans for balancing selection revealed that genome-wide estimates of Tajima's D were associated with bacterial phylogenetic history (Figure 1) (phylosig p -value = 1.63e-06). In addition, we observed significant differences among the mean genome-wide Tajima's D among bacterial genera. Figure 1B presents these genome-wide Tajima's D estimates for all bacterial genera represented by >6 SGBs.*Bifidobacterium* and *Faecalibacterium* displayed the most negative genome-wide Tajima's D, whereas *Bacteroides* displayed the highest. After correction for multiple testing, significant differences in genome-wide Tajima's D at the $p < 0.001$ significance threshold were observed between *Bifidobacterium* and*Bacteroides* and between *Faecalibacterium* and*Bacteroides* . A table of false-discovery rate corrected p-values generated from all pairwise comparisons of genome-wide Tajima's D is presented in Table S2.

A long tail of CORFs display significant evidence of balancing selection

The majority of CORFs displayed negative Tajima's D values (Figure 2). However, the distribution of CORFs was skewed right, and a long tail of CORFs displayed significantly positive Tajima's D values. The CORF with the most positive Tajima's D was the Multidrug efflux pump subunit AcrB from *Bacteroides dorei* (test for non-zero Tajima's D, adjusted$p$ -value = 4.46e-07). In addition, the Multidrug efflux pump subunit AcrA (membrane-fusion protein) from *B. dorei* was the fourth most positive CORF of all 566,958 CORFs (test for non-zero Tajima's D, adjusted p -value = 9.84e-06). Similarly, several other multidrug transporters, including multiple homologs of AcrB and AcrA in other bacterial genera, displayed Tajima's D values greater than 3, indicative of histories of balancing selection (Table S1, Figure 2).

Multidrug efflux pumps are overrepresented among the CORFs with the highest Tajima's D values

The CORFs displaying the highest Tajima's D values also included multidrug efflux pump subunits from bacterial species other than*Bacteroides dorei* , including *Bacteroides eggerthii* ,*Bacteroides fragilis* , *Coprobacter fastidiosus* ,*Bifidobacterium breve* , and others (Figure 2, Table S1). Post-hoc Fischer's exact tests were employed to test whether ORFs coding for peptides that belong to multidrug efflux pump complexes were significantly overrepresented among the ORFs with Tajima's D values >3. A total of 329 CORFs displayed Tajima's values >3, and these included 9 CORFs coding for components of multidrug efflux pumps. In contrast, only 3,692 of the total 566,985 CORFs were annotated as coding for components of multidrug efflux pumps. These analyses indicated that CORFs coding for components of multidrug efflux pumps were significantly overrepresented among the ORFs with Tajima's D values >3 (Fisher's exact test, p -value = 4e-4).

Balancing selection near the active site of multidrug efflux pump of multiple bacterial genera

To visualize the location of the AcrB CORF inferred to be under balancing selection in *Bacteroides dorei* , this sequence was searched against the Protein Data Bank. This analysis yielded a top hit of a multidrug efflux pump from *Escherichia coli.* Alignment of the *B. dorei* AcrB CORF against the *E. coli* reference enabled visualization of the putative location of the *B. dorei*AcrB CORF's peptide in the multidrug efflux pump. The crystal structure of this *E. coli* protein with the homologous region from *B. dorei* indicated is presented in Figure 2B. Theses analyses revealed that the peptide lays within a binding site of the protein complex. Ligands that interact with AcrB in *E. coli* include several notable antibiotics such as Erythromycin A, a macrolide, and Minocycline, a tetracycline. The binding site of Erythromycin A, which includes residues in AcrB, is shown in Figure 2C. To further investigate the potential functional significance of the *B. dorei* AcrB CORF, we conducted analyses of disorder and hydrophobicity, both of which are expected to differ near active sites of protein complexes from the background levels. These analyses indicated that this AcrB CORF codes for a peptide with relatively low levels of disorder and high hydrophobicity (Figure S1).

Degree of balancing selection predicts relative abundances of bacterial species in the human gut.

The maintenance by balancing selection of genetic variation in functionally important genes, such as multidrug efflux pumps, in bacterial species may provide a competitive advantage to these species in the human gut microbiome. To test the ecological consequences of balancing selection for gut bacterial species, we asked

whether the relative abundance of bacterial species in the human gut microbiome was associated with the degree of balancing selection evident in the genomes of that species. For this analysis, the relative abundances of all SGBs in the Human Microbiome Project (HMP) metagenomes sequenced from the healthy human subjects cohort (Huttenhower et al., 2012) were estimated using CoverM (https://github.com/wwood/CoverM). We then tested for an association of species' relative abundance estimates with the species' genome-wide Tajima's D values as well as with the species' maximum CORF Tajima's D value. Species' relative abundances were not associated with genome-wide Tajima's D (p -value > 0.05), but they were significantly positively associated with the species' maximum CORF Tajima's D value (Figure 3A). Similarly, PGLS indicated that the association between species' relative abundance and the species' maximum CORF Tajima's D value was also evident after controlling for bacterial phylogenetic history (Figure 3B). These analyses showed that the bacterial species containing CORFs with the highest Tajima's D values also displayed the highest relative abundances in the human gut microbiome.

Discussion

Comparisons of 118,617 metagenome assembled genomes from 287 gut bacterial species enabled the identification of genes targeted by balancing selection in the human gut. Results revealed that multidrug efflux pumps (MEPs) display the strongest signatures of balancing selection of any gut bacterial core open reading frames (CORFs). MEPs from a diversity of prominent gut bacterial species, including*Bacteroides* and *Bifidobacterium* , displayed evidence of balancing selection (Table S1, Figure 2), suggesting that adaptive allelic variation within these loci has been maintained in parallel in multiple bacterial lineages. MEPs were also overrepresented among the CORFs displaying the highest Tajima's D values, further supporting that balancing selection shapes allelic variation within this functional category of loci.

MEPs serve myriad functions for bacteria, including the extrusion of antibiotics that are commonly used as medicines. Previous work has shown that antibiotic therapies can act as harsh selective agents in the gut, reshaping the community composition of the gut microbiota (Modi et al., 2014) as well as the adaptive trajectories of individual gut bacterial lineages (Banerjee et al., 2021; Card et al., 2021). The findings reported here are consistent with the possibility that medical antibiotic use also contributes to the maintenance of allelic variation within multiple prominent gut bacterial species. In particular, the observation that the CORF displaying the highest Tajima's D value was a homolog of the AcrB subunit of the RND superfamily of multidrug efflux pumps (Figure 2) suggests medical antibiotic usage as an agent of balancing selection. In *Escherichia coli* , the periplasmic distal binding pocket of the AcrB subunit binds minocycline, a tetracycline antibiotic, and erythromycin A, a macrolide antibiotic (Du et al., 2018). Moreover, allelic variation at this locus in *E. coli* has been shown to contribute to antibiotic resistance (Okusu et al., 1996, Blair et al., 2015). However, MEPs are widely distributed among bacterial genomes and serve ancient functions that predate the usage of antibiotics in medical contexts (Blanco et al., 2016), including the extrusion of heavy metals, organic pollutants, plant-produced compounds, and bacterial metabolites. Therefore, it is possible that selective agents other than medical antibiotics may contribute to the maintenance of allelic variation in MEP loci displaying positive Tajima's D values. If medical antibiotics are in fact driving balancing selection in the MEP loci identified, results presented her imply that these treatments may be among the most influential selective agents maintaining allelic variation in the human gut.

Positive Tajima's D values are consistent with a history of balancing selection, but they can also be caused by fluctuations in population size. In particular, recent population contractions can generate positive Tajima's D values in the absence of balancing selection. Here, Tajima's D was estimated genome-wide for each CORF in each bacterial species analyzed, allowing identification of loci that deviated substantially from the genomic background. This approach provided tests for balancing selection that accounted for genome-wide patterns of nucleotide variation caused by demographic processes. For example, genome-wide Tajima's values differed significantly among bacterial clades, with species within the *Bifidobacterium* displaying the most negative values and species within the *Bacteroides*displaying the most positive values. These differences among genome-wide Tajima's D values likely reflect difference among the demographic histories of these clades, whereas loci with Tajima's D values that deviate from the genomic background are more likely to represent

targets of balancing selection.

Multidrug efflux pumps were significantly enriched among the CORFs displaying the highest Tajima's D values included, but this set of CORFs also included a diversity of other functional categories of proteins. Magnesium transporters, helicases, and various synthases and hydrolases were all represented among the ORFs with the highest Tajima's D values (Table S1). Allelic variation in these enzymes may be maintained by cyclic fluctuations in the availabilities of different substrates. The CORFs with Tajima's D values greater than three (Table S1) represent excellent candidates for experimental study of the functional consequences of allelic variation within these loci.

Interestingly, the bacterial species that contained CORFs with the highest Tajima's D values also tended to be the most abundant bacterial species in the human gut based on metagenomic data (Figure 3). This positive association suggests a relationship between balancing selection and fitness in human gut bacteria. One possible explanation for this pattern is that balancing selection is more effective in more abundant bacterial species, given that selection is in general expected to be more efficient in larger populations than smaller populations (Lanfear et al., 2014). Alternatively, the CORFs identified as targets of balancing selection may confer fitness benefits to bacterial species, increasing their competitive advantage over other species in the gut. This hypothesis is supported by the observation that the relationship between balancing selection and relative abundance in the gut remained evident after controlling for bacterial phylogenetic history (Figure 3B). Under this scenario, the allelic variation in MEPs displaying evidence of balancing selection may underlie the success in the human gut of lineages like *Bacteroides* spp., which are overrepresented in industrialized human populations relative to non-industrialized populations and non-human primates (Yatsunenko et al., 2011; Moeller et al., 2014; Sonnenburg and Sonnenburg, 2019).

Acknowledgements

This work was funded by a Maximizing Investigator's Research Award to AHM (R35 GM138284).

Data Availability Statement

All metagenome assembled genomes analyzed in this study are available at http://segatalab.cibio.unitn.it/data/Pasolli_et_al.html and http://opendata.lifebit.ai/table/?project=SGB. The complete list of CORF sequences and their corresponding SGBs and Tajima's D values is available at http://moellerlab.com/publications.

References

David, L. A., Maurice, C. F., Carmody, R. N., Gootenberg, D. B., Button, J. E., Wolfe, B. E., ... & Turnbaugh, P. J. (2014). Diet rapidly and reproducibly alters the human gut microbiome. *Nature* , *505*(7484), 559-563.

Modi, S. R., Collins, J. J., & Relman, D. A. (2014). Antibiotics and the gut microbiota. *The Journal of Clinical Investigation* , *124* (10), 4212-4218.

Schluter, J., Peled, J. U., Taylor, B. P., Markey, K. A., Smith, M., Taur, Y., ... & Xavier, J. B. (2020). The gut microbiota is associated with immune cell dynamics in humans. *Nature* , *588* (7837), 303-307.

Coyte, K. Z., & Rakoff-Nahoum, S. (2019). Understanding competition and cooperation within the mammalian gut microbiome. *Current Biology* , *29* (11), R538-R544.

Tajima, F. (1989). Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. *Genetics* , *123* (3), 585-595.

Pasolli, E., Asnicar, F., Manara, S., Zolfo, M., Karcher, N., Armanini, F., ... & Segata, N. (2019). Extensive unexplored human microbiome diversity revealed by over 150,000 genomes from metagenomes spanning age, geography, and lifestyle. *Cell* , *176* (3), 649-662.

Harris, C. D., Torrance, E. L., Raymann, K., & Bobay, L. M. (2021). CoreCruncher: Fast and Robust Construction of Core Genomes in Large Prokaryotic Data Sets. *Molecular Biology and Evolution* , *38* (2), 727-734.

Galperin, M. Y., Wolf, Y. I., Makarova, K. S., Vera Alvarez, R., Landsman, D., & Koonin, E. V. (2021). COG database update: focus on microbial diversity, model organisms, and widespread pathogens. *Nucleic Acids Research* , *49* (D1), D274-D281.

Katoh, K., Misawa, K., Kuma, K. I., & Miyata, T. (2002). MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucleic Acids Research* , *30* (14), 3059-3066.

Archer, F. I., Adams, P. E., & Schneiders, B. B. (2017). stratag: An r package for manipulating, summarizing and analysing population genetic data. *Molecular Ecology Resources* , *17* (1), 5-11.

Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society: Series B (Methodological)* , *57* (1), 289-300.

Chaumeil, P. A., Mussig, A. J., Hugenholtz, P., & Parks, D. H. (2020). GTDB-Tk: a toolkit to classify genomes with the Genome Taxonomy Database. *Molecular Biology and Evolution* , *36* (6), 1925-1927.

Minh, B. Q., Schmidt, H. A., Chernomor, O., Schrempf, D., Woodhams, M. D., Von Haeseler, A., & Lanfear, R. (2020). IQ-TREE 2: new models and efficient methods for phylogenetic inference in the genomic era. *Molecular Bology and Evolution* , *37* (5), 1530-1534.

Letunic, I., & Bork, P. (2021). Interactive Tree Of Life (iTOL) v5: an online tool for phylogenetic tree display and annotation. *Nucleic Acids Research* , *49* (W1), W293-W296.

Revell, L. J. (2012). phytools: an R package for phylogenetic comparative biology (and other things). *Methods in Ecology and Evolution* , *3* (2), 217-223.

Meyer, D., Dimitriadou, E., Hornik, K., Weingessel, A., Leisch, F., Chang, C. C., ... & Meyer, M. D. (2019). Package 'e1071'. *The R Journal* .

Burley, S. K., Bhikadiya, C., Bi, C., Bittrich, S., Chen, L., Crichlow, G. V., ... & Zhuravleva, M. (2021). RCSB Protein Data Bank: powerful new tools for exploring 3D structures of biological macromolecules for basic and applied research and education in fundamental biology, biomedicine, biotechnology, bioengineering and energy sciences. *Nucleic Acids Research* , *49* (D1), D437-D451.

Mészáros, B., Erdős, G., & Dosztányi, Z. (2018). IUPred2A: context-dependent prediction of protein disorder as a function of redox state and protein binding. *Nucleic Acids Research* , *46* (W1), W329-W337.

Kyte, J., & Doolittle, R. F. (1982). A simple method for displaying the hydropathic character of a protein. *Journal of Molecular Biology* , *157* (1), 105-132.

Parks, D. H., Imelfort, M., Skennerton, C. T., Hugenholtz, P., & Tyson, G. W. (2015). CheckM: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. *Genome Research* , *25* (7), 1043-1055.

Huttenhower, C.J., Human Microbiome Project Consortium. (2012). Structure, function, and diversity of the healthy human microbiome. *Nature* , *486* (7402), 207.

Banerjee, S., Lo, K., Ojkic, N., Stephens, R., Scherer, N. F., & Dinner, A. R. (2021). Mechanical feedback promotes bacterial adaptation to antibiotics. *Nature Physics* , *17* (3), 403-409.

Card, K. J., Thomas, M. D., Graves, J. L., Barrick, J. E., & Lenski, R. E. (2021). Genomic evolution of antibiotic resistance is contingent on genetic background following a long-term experiment with Escherichia coli. *Proceedings of the National Academy of Sciences USA* , *118* (5).

Du, D., Wang-Kan, X., Neuberger, A., van Veen, H. W., Pos, K. M., Piddock, L. J., & Luisi, B. F. (2018). Multidrug efflux pumps: structure, function and regulation. *Nature Reviews Microbiology* , *16* (9), 523-539.

Okusu, H., Ma, D., & Nikaido, H. (1996). AcrAB efflux pump plays a major role in the antibiotic resistance phenotype of Escherichia coli multiple-antibiotic-resistance (Mar) mutants. *Journal of Bacteriology* , *178* (1), 306-308.

Blair, J. M., Bavro, V. N., Ricci, V., Modi, N., Cacciotto, P., Kleinekathöfer, U., ... & Piddock, L. J. (2015). AcrB drug-binding pocket substitution confers clinically relevant resistance and altered substrate specificity. *Proceedings of the National Academy of Sciences USA* , *112* (11), 3511-3516.

Blanco, P., Hernando-Amado, S., Reales-Calderon, J. A., Corona, F., Lira, F., Alcalde-Rico, M., ... & Martinez, J. L. (2016). Bacterial multidrug efflux pumps: much more than antibiotic resistance determinants. *Microorganisms* , *4* (1), 14.

Lanfear, R., Kokko, H., & Eyre-Walker, A. (2014). Population size and the rate of evolution. *Trends in Ecology & Evolution* , *29* (1), 33-41.

Sonnenburg, E. D., & Sonnenburg, J. L. (2019). The ancestral and industrialized gut microbiota and implications for human health. *Nature Reviews Microbiology* , *17* (6), 383-390.

Yatsunenko, T., Rey, F. E., Manary, M. J., Trehan, I., Dominguez-Bello, M. G., Contreras, M., ... & Gordon, J. I. (2012). Human gut microbiome viewed across age and geography. *Nature* , *486* (7402), 222-227.

Moeller, A. H., Li, Y., Ngole, E. M., Ahuka-Mundeke, S., Lonsdorf, E. V., Pusey, A. E., ... & Ochman, H. (2014). Rapid changes in the gut microbiome during human evolution. *Proceedings of the National Academy of Sciences USA* , *111* (46), 16431-16435.

Figure 1. Patterns of intraspecific polymorphism are associated with bacterial phylogenetic history. (A) Phylogenetic tree shows the evolutionary relationships among the Species-level Genome Bins (SGBs) in the human gut for which >100 metagenome assembled genomes were recovered by Pasoli et al., (2019). Colors correspond to bacterial phyla as indicated by the inset. Phylogeny was constructed with IQTree2, and all internal nodes were supported by >70% of 1000 ultrafast bootstrap replicates. Barplots encircling phylogeny show genome-wide Tajima's D for SGBs, with larger bars indicating more negative values. Bars are colored based on whether the value is below (dark grey) or above (light grey) the median value across all SGBs. Arcs on the outside of the phylogeny indicate clades corresponding to bacterial genera that displayed significantly different genome-wide Tajima's D estimates a the $p < 0.001$ significance threshold. (B) Boxplots show the median and inner-quartile range of per-species genome-wide Tajima's D estimates for bacterial genera represented by >6 SGBs. Horizontal bars indicate significant differences between comparisons of genera after false discovery rate correction for multiple pairwise tests; $p < 0.05$ *; $p < 0.001$ ***.

Figure 2. Balancing selection targets the active sites of multidrug efflux pumps in multiple prominent gut bacterial species. A) Histogram shows the distribution of Tajima's D values across all 566,958 CORFs analyzed from 287 gut bacterial species. Highlighted are CORFs from multidrug efflux pumps, including AcrB and AcrA subunits, from multiple prominent gut bacterial species. B) Crystal structure of a multidrug efflux pump from *Escherichia coli* containing homologous AcrB subunits to the top hit identified in *Bacteroides dorei* (PDB ID: 4DX5) C) Binding pocket of AcrB homolog in E. coli interacting with ligand Erythromycin A (PDB ID: 4ZJO).

Figure 3. Signatures of balancing selection predict relative abundances of gut bacterial species. A) Scatter plot shows the relationship between the maximum Tajima's D value for a CORF in a SGB and the relative abundance of the SGB in the human gut as estimated by CoverM. The SGBs with ORFs displaying the top five Tajima's D values (Table S1) are labelled. Polynomial regression was fitted in R, and asterisks show significance of linear, quadratic, and cubic coefficients; $p < 0.05$ *, $p < 0.01$ **, $p < 0.001$ ***. B) Phylomorphospace plot shows the projection of the phylogenetic tree of SGBs onto the space of relative abundances of the SGBs and maximum Tajima's D values for a CORF in the SGBs. Large circles represent tips, and smaller circles represented states at ancestral nodes inferred under a Brownian motion model of evolution. Asterisks indicates significance of PGLS; * $p< 0.05$.

Figure 1

Hosted file

image1.emf available at https://authorea.com/users/446586/articles/545761-metagenomic-targets-of-balancing-selection-in-the-human-gut

Figure 2

Hosted file

image2.emf available at https://authorea.com/users/446586/articles/545761-metagenomic-targets-of-balancing-selection-in-the-human-gut

Figure 3

Hosted file

image3.emf available at https://authorea.com/users/446586/articles/545761-metagenomic-targets-of-balancing-selection-in-the-human-gut