bioRxiv preprint doi: https://doi.org/10.1101/2022.03.08.483496; this version posted March 8, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

Community diversity is associated with intra-species genetic diversity and gene loss in the human gut microbiome

Naïma Madi[1], Daisy Chen[2,3,^], Richard Wolff[4,^], B. Jesse Shapiro[1,5,6,7,*], and Nandita Garud[4,8,*]

1. Département de sciences biologiques, Université de Montréal, Canada;
2. Computational and Systems Biology, University of California, Los Angeles
3. Bioinformatics and Systems Biology Program, University of California, San Diego
4. Department of Ecology and Evolutionary Biology, University of California, Los Angeles
5. Department of Microbiology and Immunology, McGill University, Canada;
6. McGill Genome Centre, McGill University, Canada
7. Quebec Centre for Biodiversity Science, Canada
8. Department of Human Genetics, University of California, Los Angeles

* Correspondence to jesse.shapiro@mcgill.ca and ngarud@ucla.edu. These authors contributed equally.

^ These authors contributed equally.

bioRxiv preprint doi: https://doi.org/10.1101/2022.03.08.483496; this version posted March 8, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

Abstract

The human gut microbiome contains a diversity of microbial species that varies in composition over time and across individuals. These species are comprised of diverse strains, which are known to evolve by mutation and recombination within hosts. How the ecological process of community assembly interacts with sub-species diversity and evolutionary change is a longstanding question. Two hypotheses have been proposed based on ecological observations and theory: Diversity Begets Diversity (DBD), where taxa tend to become more diverse in already diverse communities, and Ecological Controls (EC), where higher community diversity impedes diversification within taxa. Recently we showed with 16S rRNA gene amplicon data from the Earth Microbiome Project that DBD is detectable in natural bacterial communities from a range of environments at high taxonomic levels (ranging from phylum to species-level), but that this positive relationship between community diversity and within-taxon diversity plateaus at high levels of community diversity. Whether increasing community diversity is associated with sub-species genetic diversity within microbiomes, however, is not yet known. To test the DBD and EC hypotheses at a finer genetic resolution, we analyzed sub-species strain and nucleotide variation in static and temporally sampled shotgun sequenced fecal metagenomes from a panel of healthy human hosts. We find that both sub-species single nucleotide variation and strain number are positively correlated with community diversity, supporting DBD. We also show that higher community diversity predicts gene loss in a focal species at a future time point and that community metabolic pathway richness is inversely correlated with the pathway richness of a focal species. These observations are consistent with the Black Queen Hypothesis, which posits that genes with functions provided by the community are less likely to be retained in a focal species' genome. Together, our results show that DBD and Black Queen may operate simultaneously in the human gut microbiome, adding to a growing body of evidence that these eco-evolutionary processes are key drivers of biodiversity and ecosystem function.

bioRxiv preprint doi: https://doi.org/10.1101/2022.03.08.483496; this version posted March 8, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

Introduction

Our understanding of the evolution and diversification has been enriched by experimental studies of bacterial isolates in the laboratory, but it remains a challenge to study evolution in the context of more complex communities (Lenski, 2017). Ongoing advances in culture-independent technologies have allowed us to study bacteria in the complex and dense communities in which they naturally occur (Garud and Pollard, 2020). Within a community, individual players engage in many negative and positive ecological interactions. Negative interactions can originate from competition for resources and biomolecular warfare (Hibbing et al., 2010), while positive interactions can stem from secreted metabolites that are used by other members of the community (cross feeding) (Venturelli et al., 2018). These ecological interactions can create new niches and selective pressures, leading to eco-evolutionary feedbacks whose nature are yet to be fully understood.

Ecological interactions can yield positive or negative effects on the diversification of a focal species. Under the "Diversity Begets Diversity" (DBD) hypothesis, higher levels of community diversity increase the rate of speciation (or diversification, more generally) due to positive feedback mechanisms such as niche construction (Calcagno et al., 2017; Schluter and Pennell, 2017). By contrast, the "Ecological Controls" (EC) hypothesis posits that competition for a limited number of niches at high levels of community diversity results in a negative effect on further diversification. Metabolic models predict that DBD may initially spur diversification due to cross feeding, but the diversification rate eventually slows and reaches a plateau as metabolic niches are filled (San Roman and Wagner, 2021). These theoretical predictions are largely supported by our previous study involving 16S rRNA gene amplicon sequencing data from the Earth Microbiome Project, in which we observed a generally positive relationship between community diversity and focal-taxon diversity at most taxonomic levels, reaching a plateau at the highest levels of diversity (Madi et al., 2020). A recent experiment on soil bacteria also found evidence of DBD at the family level, most likely driven by niche construction and metabolic cross-feeding (Estrela et al., 2022). Both of these studies show that DBD shapes microbial communities at higher taxonomic levels – involving community assembly and species sorting – but lacked the genetic resolution to interrogate sub-species strain-level dynamics such as strain colonization dynamics, polymorphism levels, and gene gain and loss events. Moreover, these studies also lacked time series data to enable

bioRxiv preprint doi: https://doi.org/10.1101/2022.03.08.483496; this version posted March 8, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

directly tracking the dynamics of DBD, in which community diversity at one time point influences diversity of a focal species in a future time point.

Like DBD and EC, the Black Queen Hypothesis (BQH) also makes predictions about the effects of community diversity on the evolutionary dynamics of a focal species. BQH predicts that a focal species will be less likely to encode genes with functions provided by other members of the surrounding community, if such functions are "leaky" and available as diffusible public goods (Morris et al., 2014, 2012). Gene loss may even be adaptive, provided that there is a cost to encoding and expressing the relevant genes (Albalat and Cañestro, 2016; Koskiniemi et al., 2012; Simonsen, 2022). The BQH has been invoked to explain the distribution of genes involved in vitamin B metabolism (Sharma et al., 2019) and iron acquisition (Vatanen et al., 2019) in the gut microbiome, but we still lack a complete picture of how the BQH applies to natural microbial communities.

Here we examine evidence for DBD and BQH in human gut microbiome data at a sub-species level. We use static and temporal shotgun metagenomic data from a large panel of healthy hosts from the Human Microbiome Project as well as from the same individual sampled almost daily over the course of one year (Poyet et al., 2019). In our previous study, we found strong support for DBD in the animal gut compared to more diverse microbiomes such as soils and sediments which were closer to a plateau of diversity (Madi et al., 2020). As such, the human microbiome represents an ideal model for further studying DBD dynamics. Using metagenomic data to track gene gain and loss events within a focal species allows us to simultaneously test the predictions of DBD and BQH, which are not mutually exclusive (**Figure 1**).

bioRxiv preprint doi: https://doi.org/10.1101/2022.03.08.483496; this version posted March 8, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

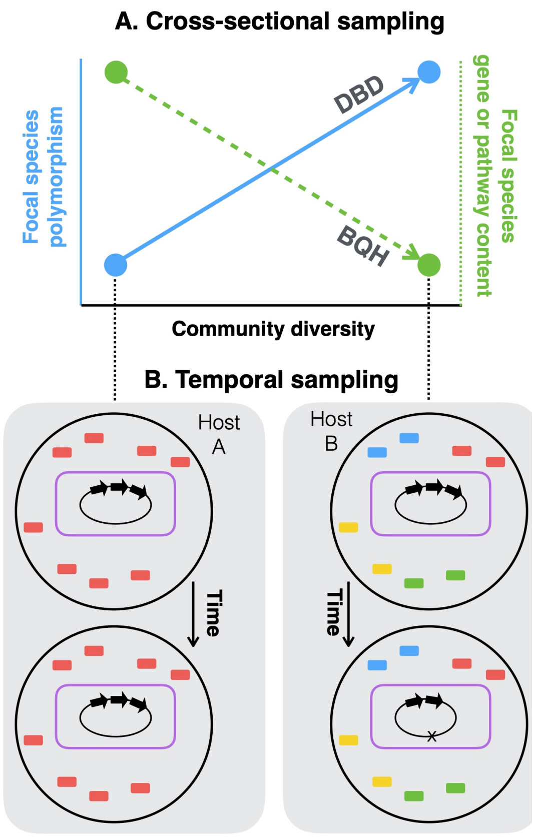
A. Cross-sectional sampling
Focal species polymorphism
Focal species gene or pathway content
DBD
BQH
Community diversity
B. Temporal sampling
Host A
Host B
Time
Time

Figure. 1. Diversity begets diversity (DBD) and the Black Queen Hypothesis (BQH) illustrated. (A) Each point represents a different individual microbiome sampled from a cross-sectional cohort, ranging from low to high community diversity along the x-axis. Under DBD, high community diversity is associated with high focal species polymorphism (blue line) yielding a positive diversity slope. Alternatively (not shown) a niche filling model as described by the EC hypothesis would yield a flat or negative diversity slope. Under the BQH, high community diversity is associated with fewer genes or metabolic pathways encoded by the focal species (dashed green line). **(B)** Two hypothetical gut microbiomes (hosts) sampled over time, illustrating the predictions of DBD and BQH. The bacterial community is shown as small rectangles with different bacterial species in different colors. The focal species is shown as an enlarged rectangle outlined in violet, containing a circular genome encoding genes (bold black arrows). In Host A, the microbiome is not diverse (only one 'red' species) and does not affect the diversity of the focal species. In Host B, a more diverse community at the first time point selects for increased polymorphism (point mutation shown as an 'X') and the loss of genes or pathways within the focal species. For simplicity these are depicted as *de novo* mutation or gene loss events of resident lineages, but could equally come about due to invasions of new strains or changes in the relative abundance of preexisting strain diversity.

bioRxiv preprint doi: https://doi.org/10.1101/2022.03.08.483496; this version posted March 8, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

Results

We assess evidence for the DBD hypothesis within species using two shotgun metagenomic datasets. First, we analyze data from a panel of 249 healthy hosts (Human Microbiome Project Consortium, 2012; Lloyd-Price et al., 2017), in which stool samples were collected 1-3 times at approximately 6-month intervals. Second, we analyze data from a single individual sampled more densely (206 samples) over the course of ~18 months (Poyet et al., 2019). We analyze both cross-sectional and temporal data to understand the relationship between community diversity and genetic diversity at the sub-species level.

To test the DBD hypothesis, we examined several metrics of community diversity and intra-species diversity and calculated the diversity slope (**Figure 1**). To quantify community diversity, we calculated Shannon diversity and richness at the species level. To control for variation in sequencing depth across samples, richness was computed on rarefied data. We also used unrarefied data and included the number of reads per sample as a covariate in our models, yielding similar results (described below). To quantify intra-species diversity, we used a reference genome-based approach to call single nucleotide variants (SNVs) and gene copy number variants (CNVs) within each focal species and computed polymorphism rates, measured as the fraction of synonymous nucleotide sites in the core genome with intermediate allele frequencies (between 0.2 and 0.8) within a host (Methods). As an additional metric of intra-species diversity, we inferred the number of strains within each species using StrainFinder (Smillie et al., 2018).

Community diversity is associated with sub-species polymorphism in the human gut microbiome

We began by plotting the relationship between community diversity and intra-species polymorphism rate in cross-sectional HMP metagenomes (**Figure 2A and B**). The slope of this relationship (which we call the diversity slope; **Figure 1**) provides an indicator of the extent of DBD (positive slope) or EC (flat or negative slope). As a descriptive analysis, we first computed Spearman correlations between Shannon diversity and intra-species polymorphism rate. Out of the 68 bacterial species with sufficient prevalence (present in at least four samples), we found 15 significant correlations (*P* < 0.05, uncorrected for multiple tests), of which 14 were positive (**Fig S1**). Similarly, we found 18 significant correlations between richness and intra-species

bioRxiv preprint doi: https://doi.org/10.1101/2022.03.08.483496; this version posted March 8, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

polymorphism rate (**Fig S2**), all of which were positive. These positive associations are broadly consistent with the DBD hypothesis, although we cannot establish the direction of causality in this cross-sectional data.

The relationship between polymorphism rate and community diversity was found to be non-linear (**Figures 2A, B, S1 and S2**). Polymorphism rates across HMP hosts span several orders of magnitude (10^{-5}/bp to 10^{-2}/bp), largely due to the fact that strain content is variable across hosts. Polymorphism rates of ~10^{-2}/bp or more are inconsistent with within-host diversification of a single colonizing lineage, and instead represent mixtures of multiple strains that diverged before colonizing a host. By contrast, rates <10^{-4}/bp are more consistent with a single strain colonizing the host (Garud et al., 2019).

To more formally test the predictions of DBD and to account for non-linear relationships between polymorphism and community diversity, we used generalized additive models (GAMs). Using GAMs, we are able to model non-linear relationships and account for random variation in the strength of the diversity slope across samples, human hosts, and bacterial species (Methods). We find that GAMs support the overall positive association between within-species polymorphism and Shannon diversity (GAM, P=0.031, Chi-square test) as well as between within-species polymorphism and community richness after controlling for coverage as a covariate (P=0.017) or rarefying samples to an equal number of reads (P=1.93e-05) (**Fig. S3**). While the polymorphism-community diversity relationships were generally positive, it appears that polymorphism reaches a plateau at high levels of community richness (**Fig S2**), as supported by GAMs **(Fig S3 B,C)**; see **Table S1** and **Supplementary File 1** for additional model details.

These generally positive correlations between focal species polymorphism and species-level measures of community diversity also hold when community diversity is measured at higher taxonomic levels; specifically, polymorphism rate was significantly positively associated with Shannon diversity calculated at the genus and family levels (GAMs, P<0.05, Chi-square test) (**Fig S4**). However, polymorphism rate was not significantly associated with Shannon diversity calculated at the highest taxonomic levels (order, class and phylum, GAMs, P>0.05, Chi-square test). The positive correlation between polymorphism rate and richness held at all taxonomic levels

bioRxiv preprint doi: https://doi.org/10.1101/2022.03.08.483496; this version posted March 8, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

(GAMs, *P*<0.05, Chi-square test) (**Fig S4**, **Table S2**, **Supplementary File 1**). Overall, these results are consistent with DBD acting within the human gut microbiome at most taxonomic levels, as previously observed in environmental samples (Madi et al., 2020) and experimental soil communities (Estrela et al., 2022).

Community diversity is associated with sub-species strain diversity

To more explicitly account for the strain structure within hosts, we next inferred the number of strains per focal species with StrainFinder (Smillie et al., 2018) (Methods) and used strain number as another quantifier of intra-species diversity. Strain-level variation has important functional and ecological consequences; among other things, strains are known to engage in interactions that cannot be predicted from their species identity alone (Goyal et al. 2021). How ecological processes at the strain level affect and are affected by community composition and dynamics, however, remains poorly characterized.

We found that the number of strains per focal species follows an approximately linear relationship with community Shannon diversity (**Figure 1C** and **S5)**. We therefore calculated Pearson correlations between community diversity and the number of strains per focal species. Out of the 134 species for which strains were inferred (Methods), we found a total of 23 significant correlations between Shannon diversity and strain number (*P*<0.05), of which 21 were positive (**Fig S5**). By contrast, out of the 16 significant correlations between richness and strain number, 13 were negative (**Figures 1D** and **S6**). We note that the 16 species with a significantly positive Shannon-strain number correlation were completely non-overlapping with the 13 with a significant negative richness-strain number correlation, suggesting species-specific effects.

We next used generalized linear mixed models (GLMMs) to investigate the relationship between the number of strains per focal species and community diversity, while taking into account coverage per sample as a covariate and variation between species, hosts and samples as random effects. The number of strains per focal species was positively correlated with community Shannon diversity (GLMM, *P*= 1.549e-04, likelihood ratio test (LRT)) (Table S3, supplementary file 1). This is consistent with the positive correlation between polymorphism rates and Shannon diversity and is also generally concordant with DBD.

bioRxiv preprint doi: https://doi.org/10.1101/2022.03.08.483496; this version posted March 8, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

While Shannon diversity was positively correlated with strain number, species richness was negatively correlated with strain number (GLMM, *P*=1.5e-05, LRT) (**Table S3**, **Supplementary File 1**). The negative relationship with richness was unlikely to be confounded by sequencing depth, since the same result was obtained using rarefied data, albeit with a weaker negative relationship (GLMM, *P*=0.037, LRT) (**Table S3**, **Supplementary File 1**). The negative strain number-richness relationship also held at all other taxonomic ranks (GLMM, *P*<0.05, LRT) (**Table S4**, **Supplementary File 1**), while the strain number-Shannon diversity relationship was generally positive (**Fig S7**). Together, these results show that richness and Shannon diversity are both positively correlated with polymorphism rates, consistent with DBD, whereas richness and Shannon diversity have contrasting correlations with strain number.

bioRxiv preprint doi: https://doi.org/10.1101/2022.03.08.483496; this version posted March 8, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

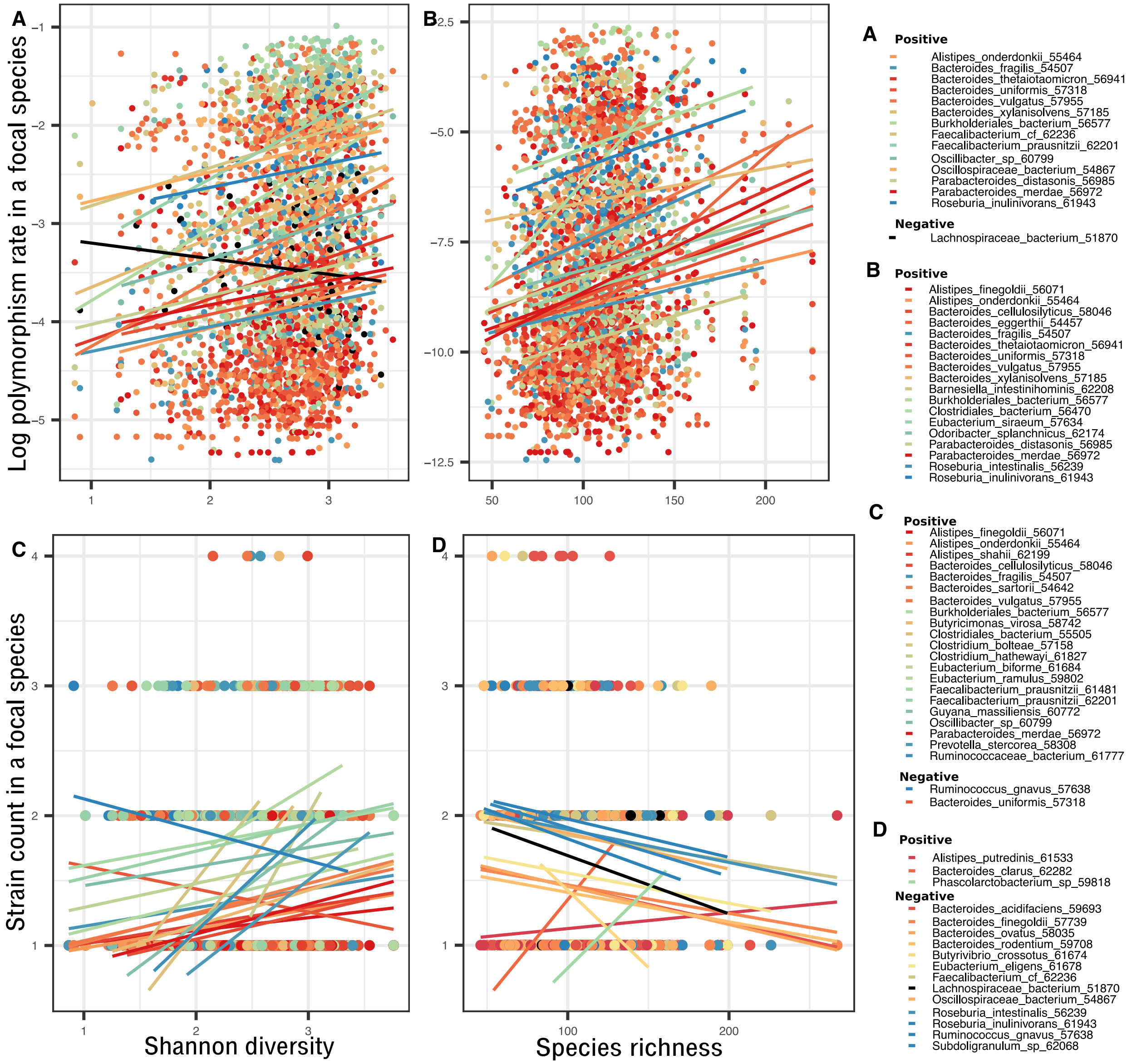

Fig 2. Relationship between intra-species diversity and community diversity in human gut microbiomes. (A) Intra-species polymorphism rate versus Shannon diversity (15 out of 68 species have a significant Spearman correlation with Shannon alpha diversity; 14 are positive and 1 is negative, shown with a black trendline). **(B)** Intra-species polymorphism rate versus rarefied richness (18 species out of 68 are significantly correlated; all are positive). **(C)** Intra-species strain number versus Shannon diversity (23 out of 134 species are significantly correlated with richness; 21 are positive). **(D)** Intra-species strain number versus rarefied richness (16 out of 134 species are significantly correlated with richness; 13 are negative). Each species is represented by a different color. Only significant correlations (points and lines) are plotted ($P<0.05$, Spearman in A and B and Pearson in C and D). Species present in fewer than 4 samples were not considered. The Y-axis in A and B is on a log10 scale.

bioRxiv preprint doi: https://doi.org/10.1101/2022.03.08.483496; this version posted March 8, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

Testing DBD over time in the gut

Our analyses thus far have considered only individual time points, which represent static snapshots of the dynamic processes of community assembly and evolution in the microbiome. To test the effects of DBD over time, we analyzed 160 HMP hosts with multiple time points, in which the same person was sampled 2-3 times ~6 months apart. Under a DBD model, we expect community diversity at an earlier time point to result in higher within-species polymorphism at a future time-point. To test this expectation, we defined 'polymorphism change' as the difference between polymorphism rates at the two time points (Methods). We also investigated the effects of community diversity on gene loss and gain events within a focal species, as such changes in gene content are known to occur frequently within host gut microbiomes (Garud et al., 2019; Groussin et al., 2021; Zhao et al., 2019). Here a gene was considered absent if its coverage (*c*) was <0.05 and present if $0.6 \leq c \leq 1.2$ (Methods). As in the cross-sectional analyses above, we also controlled for sequencing depth of the sample and excluded genes with aberrant coverage or that were present in multiple species.

In HMP samples, polymorphism change showed no significant relationships with community diversity at the earlier time point (**Fig S8**, GAM, P=0.4, P=0.64 and P=0.497, for Shannon, richness and rarefied richness respectively), nor did gene gains show any relationships (**Fig S9**, GLMM, *P*= 0.733, *P*= 0.617 and *P*= 0.508, LRT for Shannon, richness and rarefied richness respectively) (**Supplementary File 1**). These results suggest that DBD is negligible or undetectable over ~6-month time lags in the human gut. By contrast, we found that gene loss in a focal species between two consecutive time points was positively correlated with community diversity at the earlier time point (**Figure 3**, **S10**, GLMM, *P*= 0.028, *P*= 0.036 and *P*= 0.01, LRT for Shannon, richness and rarefied richness respectively) (**Table S5**, supplementary file 1). Elevation of gene loss in more diverse communities is consistent with the BQH, which we investigate in further detail below. Most species in the HMP hosts lost fewer than ten genes over ~6 months, but occasionally hundreds of genes were lost (**Figure 3**), suggesting a mixture of *de novo* deletion of a few genes as well as selection of strains encoding fewer genes in more diverse communities.

bioRxiv preprint doi: https://doi.org/10.1101/2022.03.08.483496; this version posted March 8, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

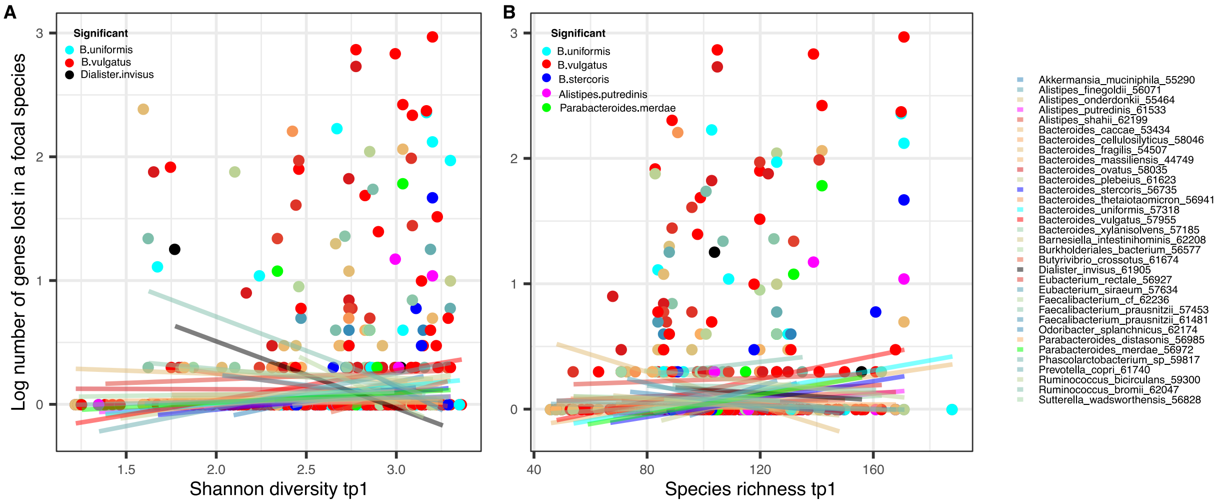

Figure 3. Community diversity predicts focal species gene loss in the human microbiome over time. Plots show the number of genes lost between two time points in a focal species as a function of **(A)** Shannon diversity, or **(B)** richness measured at an earlier time point in the HMP. The species listed in the top left of each panel are those with individually significant correlation (Pearson correlation, *P*<0.05) out of 33 listed on the right. Only the 33 species present in more than three samples with at least one gene loss event were tested. All significant correlations were positive except for the correlation between gene loss and Shannon diversity in *Dialister invisus* in **A** (black). The Y-axis in A and B is on a log10 scale.

To assess the evidence for DBD at higher temporal resolution, we analyzed shotgun metagenomic data from the most frequently sampled healthy individual (host *am*) from a previous study (Poyet et al., 2019). This individual donated stool samples that were sequenced over 18 months with a median of one day (mean of 2.6 days) between time points. In this data, we tracked both polymorphism change and gene gains and losses between two successive time points in *Bacteroides vulgatus*, the most abundant species across samples (mean coverage=58.5; median=54.2; Methods). Polymorphism and strain-level diversity within *B. vulgatus* were positively associated with community diversity in the HMP cross-sectional data (**Figures S1**, **S2**, **S5**). We would therefore expect similar associations in time series data.

We asked whether community diversity in the gut microbiome at one time point could predict increases in *B. vulgatus* polymorphism at the next time point, typically a few days later. Consistent with DBD, Shannon diversity at the earlier time point was positively correlated with changes in polymorphism (**Figure 4A**, Pearson correlation *P*=0.002). Notably, this positive correlation was

bioRxiv preprint doi: https://doi.org/10.1101/2022.03.08.483496; this version posted March 8, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

not evident in the HMP time series, perhaps due to insufficient density of sampling to capture rapid dynamics. Even when individual correlations were tested in HMP data, *B.vulgatus* did not show a significant relationship (Pearson, *P*>0.05).

Consistent with observations from HMP time series (**Figure 3**), we found a positive relationship between gene loss and Shannon diversity in *B. vulgatus* in individual *am* (**Figure 4B,** Pearson correlation, *P*=0.06). The positive association with gene loss was mirrored by a negative association with gene gain, although both with borderline statistical significance due to relatively few observed gain or loss events over these short time intervals (**Figure 4C,** Pearson correlation, *P*=0.09). All genes gained and lost in *B. vulgatus* in *am* were annotated as hypothetical proteins. Neither polymorphism change nor gene gains or losses in *B. vulgatus* were correlated with species richness in individual *am* (**Figure S11A, B, and C,** Pearson correlation, *P*>0.2).

Overall, these results are consistent with community diversity promoting changes within the *B. vulgatus* genome over timescales of a few days. We note that this is an example of one abundant species in one well-sampled individual and may not generalize to other species and hosts. However, it does suggest that changes in polymorphism captured over daily time scales could be obscured over the timescales on the order of months, as reflected in the HMP samples. The association between Shannon diversity and gene loss, in both HMP and Poyet time series, is suggestive of adaptive gene loss as posited by the Black Queen Hypothesis (BQH). Under BQH, genes are lost from an individual genome when their functions are provided by other members of a (diverse) community.

bioRxiv preprint doi: https://doi.org/10.1101/2022.03.08.483496; this version posted March 8, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

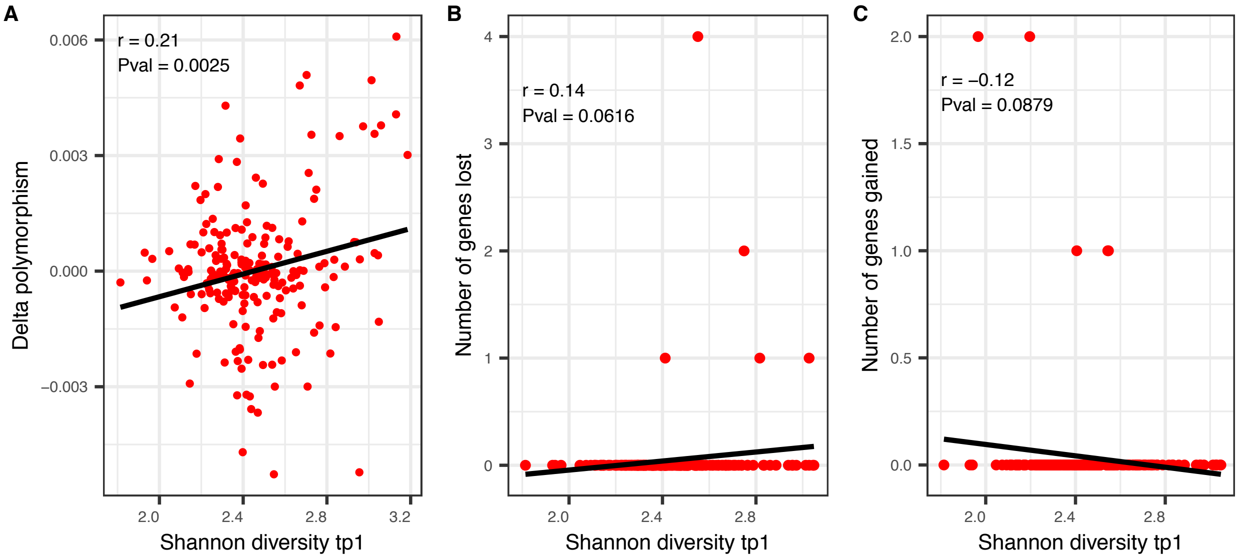

Figure 4. Correlations between polymorphism and gene content changes in *B. vulgatus* and Shannon diversity in one human gut microbiome over 18 months. A) Correlation between polymorphism change (time point 2 – time point 1) and Shannon diversity at time point 1 (tp1; the earlier time point). B) Correlation between the number of genes lost between tp1 and tp2 and the Shannon diversity at tp1. C) Correlation between the number of genes gained between tp1 and tp2 and the Shannon diversity at tp1. Each point represents the change measured between a pair of consecutive time points. Pearson correlations and *P*-values are reported in each panel.

Testing the Black Queen hypothesis in the human gut microbiome

To further assess evidence for the BQH in the HMP data, we tested the hypothesis that a focal species encodes fewer genes in a community that collectively harbors more genes. This would be expected under adaptive gene loss, provided that the genes encoded by the community provide 'leaky' functions to the focal species. Contrary to this simple expectation, we observed a significant positive relationship between community gene richness and focal species gene richness (see Methods for computation of gene richness) (**Figures 5A and S12A**; GAM, *P*=2.92e-06, Chi-square test) (Table S6, supplementary file 1). By estimating Spearman correlation between gene richness per focal species and community gene richness, we found that out of 134 species, 42 had significant correlations, of which 39 were positive (**Fig S13**). This result is inconsistent with a simple version of the BQH acting on individual gene families assuming that all gene functions are equally 'leaky'. It is, however, broadly consistent with DBD, provided that gene content is

bioRxiv preprint doi: https://doi.org/10.1101/2022.03.08.483496; this version posted March 8, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

correlated with polymorphism rate, which we already showed to be correlated with community diversity (**Figure 2**). In other words, DBD is supported both in terms of within-species single nucleotide polymorphism and gene content variation.

Next, we tested the hypothesis that the BQH acts at the level of metabolic pathways rather than individual gene families. Specifically, cellular pathways that are encoded by the community need not be encoded by a focal species provided that the pathway product or function is leaky. Consistent with the BQH acting at the pathway level, we found that community pathway richness, measured as the number of pathways present with non-zero abundance inferred with HUMAnN2 (Franzosa et al., 2018) (Methods) was negatively correlated with focal species pathway richness (**Figures 5B, S12B;** GAM, P<2e-16, Chi-square test) (Table S6, supplementary file 1). When testing 239 prevalent species, we found 107 significant Spearman correlations ($P < 0.05$), of which 95 (89%) were positive (**Fig S14**). Note that three species (*Escherichia coli, Enterobacter cloacae* and *Klebsiella pneumoniae*, shown respectively with green, orange, and red points and trendlines in **Figure 5B**) with particularly high pathway richness had much steeper negative slopes, but they are not solely responsible for the overall negative trend **(Fig 5B**).

bioRxiv preprint doi: https://doi.org/10.1101/2022.03.08.483496; this version posted March 8, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

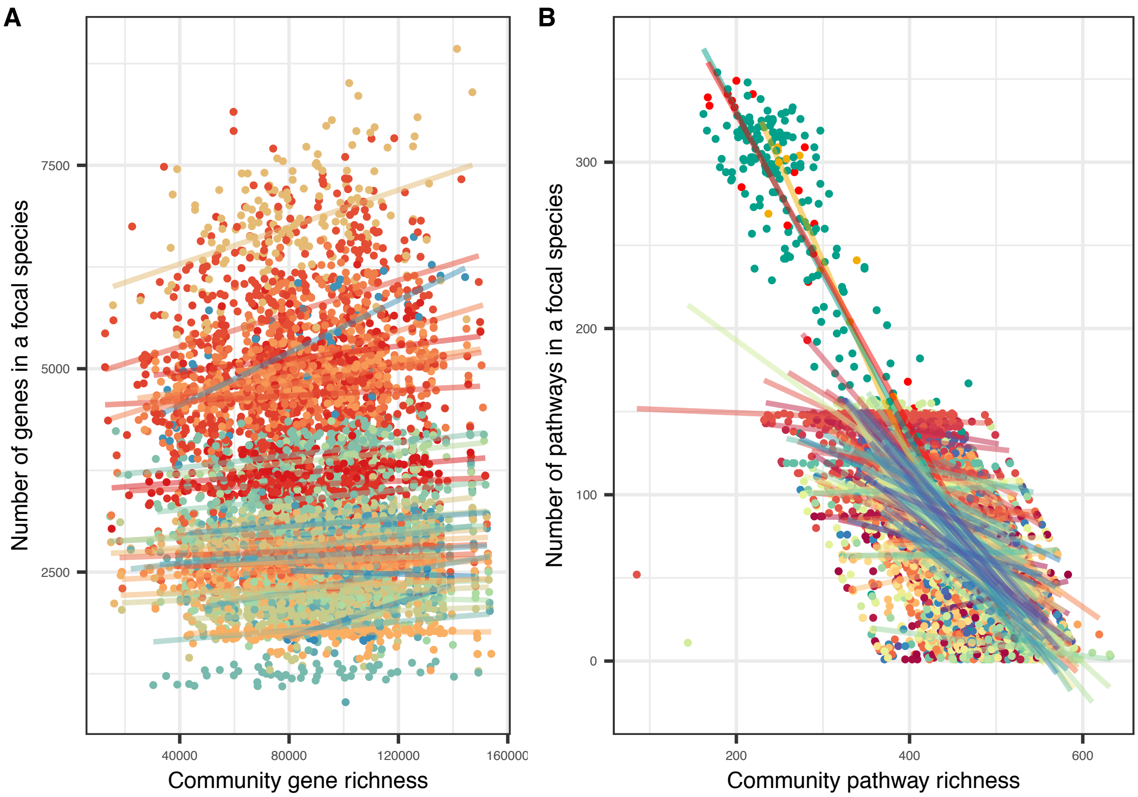

Figure 5. Testing predictions of the Black Queen Hypothesis in the human microbiome. Plots show correlations between (A) the number of genes present in a focal species and genes present in other members of the community (out of 124 species tested, 42 had significant correlations, of which 39 positive and 3 negative), and B) the number of pathways present in focal species and community pathway richness (out of 239 species tested, 107 had significant correlations, of which 95 negative and 12 positive). Only species present in at least 4 samples with a significant Spearman correlation (P<0.05) are plotted.

Discussion

In this paper we investigated whether community diversity begets genetic diversity within species in gut microbiota using static and temporally resolved fecal shotgun metagenomic data from a panel of healthy hosts. In support of the DBD hypothesis, we found that focal species often have higher polymorphism rates and strain counts in more diverse communities, whether community diversity was estimated with Shannon index or species richness. The same pattern held when community diversity was estimated at higher taxonomic ranks, consistent with our previous analysis of amplicon sequence data across environments (Madi et al., 2020) and a recent experimental study of soil bacteria community assembly (Estrela et al., 2022). Together, these

bioRxiv preprint doi: https://doi.org/10.1101/2022.03.08.483496; this version posted March 8, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

results indicate that the DBD hypothesis is relevant at multiple taxonomic levels, and extends past the species level to the sub-species genetic level.

While sub-species strain diversity is generally positively correlated with Shannon diversity, it is inversely correlated with species richness, suggesting that the ability of strains to colonize a host may be associated with higher community evenness rather than their total count. Although Shannon diversity is considered to be more robust and informative than richness in estimating bacterial diversity (He et al., 2013; Reese and Dunn, 2018), we observe the same contrasting results between Shannon and richness when community diversity is calculated at higher taxonomic levels, suggesting that this pattern is not due to artifacts such as sequencing effort.

Another study also recently found evidence for eco-evolutionary feedbacks in the HMP, in the form of a positive relationship between evolutionary modifications or strain replacements in a focal species and community diversity (Good and Rosenfeld, 2022). Using a model, they further showed that these eco-evolutionary dynamics could be explained by resource competition and did not require the cross-feeding interactions previously invoked to explain DBD at higher taxonomic levels (Estrela et al., 2022; San Roman and Wagner, 2021, 2018). This could be because cross-feeding operates at the family- or genus- level, and is less relevant as a finer-scale evolutionary process.

Perhaps compatible with the recent work, we found that community diversity predicts gene loss in a future time point and that community pathway richness is negatively correlated with pathway richness of a focal species. This suggests that both DBD and BQH might be at play in the gut microbiome, in which high community diversity may simultaneously select for diversification (at the SNV and strain level) while also selecting for adaptive gene loss as predicted by BQH (that is, relaxed selective pressure to maintain pathways already provided by the community). While it is possible that gene deletion events could explain the loss of functional metabolic pathways, it is also possible that there is a propensity for strains with fewer pathways to colonize hosts with more complex communities. Higher resolution time series data can help to disentangle these possibilities as well as to more deeply quantify the effect of BQH on microbiome diversity.

The tendency for reductive genome evolution in bacteria has already been reported by comparing hundreds of genomes (Albalat and Cañestro, 2016; Koskiniemi et al., 2012; Puigbò et al., 2014). Genome reduction is also a hallmark of endosymbiotic bacteria, which receive many

bioRxiv preprint doi: https://doi.org/10.1101/2022.03.08.483496; this version posted March 8, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

metabolites from their hosts (McCutcheon and Moran, 2012; Nikoh et al., 2011). It has been shown that uncultivated bacteria from the gut have undergone considerable genome reduction, which may be an adaptive process that results from use of public goods (Nayfach et al., 2019). Our findings suggest that genome reduction in the gut is higher in more diverse gut communities, and future work could establish whether this effect is indeed due to metabolic cross-feeding as posited by some models (Estrela et al., 2022; San Roman and Wagner, 2021, 2018), but not others (Good and Rosenfeld, 2022).

The BQH may help explain why the majority of gut microbial species remain recalcitrant to cultivating under laboratory conditions (Nayfach et al., 2019; Walker et al., 2014). Specifically, gut microbes may lack the necessary pathways to survive in culture in absence of their natural counterparts that may otherwise provide essential goods. For instance, menaquinone and fatty acids have been shown to promote the growth of uncultured bacteria, and both pathways were missing from many uncultured bacteria identified in (Nayfach et al., 2019). Additionally, more than 70% of the recent created Unified Human Gastrointestinal Genome (UHGG) collection lack cultured representatives (Almeida et al., 2019).

As noted in our previous study (Madi et al., 2020), we cannot establish causal relationships between community diversity and focal species diversity using cross-sectional survey data; doing so requires controlled experiments. In the case of DBD, the correlations observed in naturally occurring microbiomes are generally concordant with experimental (Estrela et al., 2022; Jousset et al., 2016) and metabolic modeling studies (San Roman and Wagner, 2021), strengthening the plausibility of the hypothesis. Although they also note that causality is difficult to establish, Good and Rosenfeld (2022) suggest the importance of focal species evolution as a driver of changes in community structure, as shown in an experimental study of *Pseudomonas* in compost communities (Padfield et al., 2020). Further work is therefore needed to establish the extent and relative rates of eco-evolutionary feedbacks in both directions. How these feedbacks among bacteria are influenced by abiotic factors and by interactions with fungi, archaea, and phages also deserve further study.

In summary, our results show support for both DBD and the BQH within the human gut microbiome. Using metagenomic time series data, we find a positive association between community diversity and sub-species strain-level diversity. Higher community diversity is also associated with losses of genes and metabolic pathways in a focal species. Whether these reductive

bioRxiv preprint doi: https://doi.org/10.1101/2022.03.08.483496; this version posted March 8, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

genome evolution events are adaptive, as predicted by BQH, and if they can be explained by metabolic cross-feeding, remains to be seen.

Data and materials availability

The raw sequencing reads for the metagenomic samples used in this study were downloaded from Human Microbiome Project Consortium 2012 and Lloyd-Price et al. (2017) (URL: https://aws.amazon.com/datasets/human-microbiome-project/); and Poyet et al. 2019 (NCBI accession number PRJNA544527). All computer code for this paper is available at https://github.com/Naima16/DBD_in_gut_microbiome.

Methods

Metagenomic analyses

Estimation of species, gene, and SNV content of metagenomic samples

We used MIDAS (Metagenomic Intra-Species Diversity Analysis System, version 1.2, downloaded on November 21, 2016) (Nayfach et al., 2016) to estimate within-species nucleotide and gene content of raw metagenomic whole genome shotgun sequencing data for HMP1-2 and Poyet et al. 2019 data. MIDAS relies on a reference database comprised of 31,007 bacterial genomes that are clustered into 5,952 species, covering roughly 50% of species found in human stool metagenomes from “urban” individuals. Described below are the parameters used to estimate species abundances, SNVs, and gene copy numbers variants (CNVs) with MIDAS:

Estimation of species content:

To assess evidence for community diversity begetting genetic diversity, we estimated species diversity and SNVs and CNVs by mapping reads to reference genomes. Since a component of this work relies on quantifying polymorphism and CNV changes over time, we constructed a

bioRxiv preprint doi: https://doi.org/10.1101/2022.03.08.483496; this version posted March 8, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

"personal" reference database to avoid spurious inferences of allele frequency and CNV changes due to errors in mapping of reads to regions of the genome shared by multiple species. This per-host reference database was comprised of the union of all species present at one or more timepoints so as to be as inclusive as possible to prevent reads from being "donated" to reference genome, while also being selective to prevent a reference genome from "stealing" reads from a species truly present.

To estimate the species relative abundances for each host x timepoint sample, we mapped reads to 15 universal single-copy marker genes that are a part of the MIDAS pipeline (Nayfach et al., 2016; Wu et al., 2013) and belong to the 5,952 species. A species with an average marker gene coverage ≥ 3 was considered present for the purposes of inferring SNVs and CNVs below. The per-host database was constructed by including all species present at one or more timepoints with coverage ≥3.

Estimation of copy number variation:

To estimate gene copy number variation (CNV) we mapped reads to the pangenomes of species present in a host's personal database using Bowtie2 (Langmead and Salzberg, 2012) with default MIDAS settings (local alignment, MAPID≥94.0%, READQ≥20, and ALN_COV≥0.75). Each gene's coverage was estimated by dividing the total number of reads mapped to a given gene by the gene length. These genes included the aforementioned 15 universal single-copy marker genes. A given gene's copy number (*c*) was estimated by taking the ratio of its coverage and the median coverage of the single-copy marker genes.

With these copy number values, we estimated the prevalence of genes in the broader population, defined as the fraction of samples with copy number $c \leq 3$ and $c \geq 0.3$ (conditional on the mean single gene marker coverage being ≥ 5x). For each species, we computed "core genes", defined as genes in the MIDAS reference database that are present in at least 90% of samples within a given cohort. Within-host polymorphism rates were computed in core genes.

However, orthologous genes present in multiple species can result in read stealing and read donating. Thus, we excluded a set of genes belonging to a 'blacklist' comprised of genes present in multiple species. This blacklist was constructed in Garud et al. 2019 using USEARCH (Edgar, 2010) to cluster all genes in human-associated reference genomes with a 95% identity threshold. Since some genes may be absent from the MIDAS database that may also be shared across species,

bioRxiv preprint doi: https://doi.org/10.1101/2022.03.08.483496; this version posted March 8, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

we implemented another filter in Garud et al. 2019 in which genes with $c \geq 3$ in at least one sample in our cohort was excluded from analysis of polymorphism rate or gene changes over time.

Inferring single nucleotide variants (SNVs) within bacterial species

To call SNVs, we mapped reads to a single representative reference genome as per the default MIDAS software. Reads were mapped with Bowtie2, with default MIDAS mapping thresholds: global alignment, MAPID≥94.0%, READQ≥20, ALN_COV≥0.75, and MAPQ≥20. Species were excluded from further analysis if reads mapped to ≤ 40% of their genome. We further excluded samples from further analysis if they had low median read coverage ($\overline{D}$) at protein coding sites. Specifically, samples with $\overline{D} < 5$ of across all protein coding sites with nonzero coverage were excluded. This MIDAS SNV output was then used subsequently for computing within-species polymorphism rates and inferring the number of strains present for each species in each sample (see below).

To compute polymorphic rates, additional bioinformatic filters were imposed to avoid read stealing and donating across different species. First, we did not call SNVs in blacklisted genes present in multiple species. Additionally, we excluded sites in a given sample if $D < 0.3\overline{D}$ or $D > 3\overline{D}$ as these sites harbor coverage anomalously low or high compared to the genome-wide average $\overline{D}$. An additional coverage threshold requirement of 20 reads/site was imposed for inclusion of SNVs in the polymorphism rate computation.

Shannon diversity, species richness and polymorphism rate calculations

Shannon diversity and richness were computed within each sample by including any species with abundance greater than zero. Rarefied species richness estimates are based on HMP1-2 samples rarefied to 20 million reads and Poyet samples rarefied to 5 million reads.

The polymorphism rate of a species in a sample was computed as the proportion of synonymous sites in core genes with intermediate allele frequencies ($0.2 \leq f \leq 0.8$). This is quantitatively similar to the more traditional population genetic measure of heterozygosity, H=E[$2f(1\text{-}f)$], in which intermediate frequency alleles contribute the most weight. By computing polymorphism with the

bioRxiv preprint doi: https://doi.org/10.1101/2022.03.08.483496; this version posted March 8, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

criteria $0.2 \leq f \leq 0.8$, we avoid inclusion of low frequency sequencing errors, which can otherwise greatly influence the mean heterozygosity.

Temporal changes in polymorphism rates and gene content

Delta polymorphism (or changes in polymorphism) was computed as the difference in polymorphism rates between time points. Gene gains and losses between time points were computed by identifying genes with copy number c <=0.05 (indicating gene absence) in one sample and 0.6 <= c <= 1.2 in another (indicating single copy gene presence). These thresholds were used in Garud et al. 2019 when inferring gene changes in temporal data and reflect a range of copy numbers expected in either the absence of a gene or presence of a single copy of a gene. Higher copy numbers were not considered to avoid confounding our analysis with read stealing or donating among different species. Filters for coverage and blacklisted genes were applied as described above.

Strain number inference

We used StrainFinder (Smillie et al., 2018) to infer the number of strains present for each species in each HMP1-2 metagenomic sample. To do so, we used allele frequencies from MIDAS SNV output, generated as described above. For each species in each host, all multi-allelic sites with coverage of 20x or greater were passed as input to StrainFinder. Species in which no sites passed the 20x threshold were assumed to have only a single strain. StrainFinder was then run on each sample separately for strain number 1, 2, 3, and 4, and the optimal strain number was chosen based on BIC. This range of strain number was chosen for biological reasons. Based on multiple analyses of the densely longitudinally sampled metagenomic data from four healthy hosts in Poyet et al, a maximum of three strains were shown to be present at any one time within a host for the ~30 most prevalent species (Poyet et al. 2019, Wolff et al. 2021, Zheng et al. 2020). Thus, four strains were chosen as the maximum to accommodate the range of observed possibilities, as well as possible rare cases outside of this, without overfitting.

Gene and pathway richness

To determine gene richness of each sample, we used the default MIDAS threshold of 0.35 copy number to define gene presence and absence. All genes from the species' pangenome with

bioRxiv preprint doi: https://doi.org/10.1101/2022.03.08.483496; this version posted March 8, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

minimum read-depth of 1, including core and accessory genes, were considered for this analysis. Finally, we define “community gene richness” of a sample, with respect to a focal species, as the number of gene clusters present in any of the species in the sample, excluding the focal species. Gene clusters are defined as any set of genes with 95% nucleotide identity.

In addition to examining gene sets, we utilized previously generated functional profiling output from HUMAnN 2.0 (Franzosa et al., 2018) (downloaded from https://www.hmpdacc.org/hmmrc2/) to estimate pathway richness in each species present in a sample. HUMAnN 2.0 takes in whole genome metagenomes and reports gene family (UniRef) and metabolic pathway (MetaCyc) abundances in reads per kilobase (RPK); here, we count all pathways with nonzero RPK as present in a sample.

Statistical analyses

Model construction and evaluation

Using data from the HMP and Poyet et al. 2019, we examined the relationship between intra-species diversity and gut microbiome community diversity. Intra-species diversity was estimated with polymorphism rate and strain count within each species at individual time points. When two or more time points were available from the same person, delta polymorphism and gene content variation (gain and loss) between time points were used to track DBD over time. Community diversity was estimated with the Shannon index, species richness and rarefied richness (to 20 million reads per sample). When the relationship between the response variable (intra species genetic diversity) and the predictor (community diversity) was approximately linear by visual inspection, we fit generalized linear mixed models (GLMMs) (glmmTMB function from the glmmTMB R package - RStudio version 1.2.5042) with community diversity as the predictor of within-species genetic diversity, otherwise we fit Generalized additive mixed models (GAMs) (mgcv function from the mgcv R package - RStudio version 1.2.5042) to account for the non-linearity of the relationships.

To account for variation in sequencing depth, we added read count per sample (coverage) as a covariate to all generalized mixed models except when richness was calculated on the rarefied

bioRxiv preprint doi: https://doi.org/10.1101/2022.03.08.483496; this version posted March 8, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

data. Species name and sample identifier nested within subject identifier were added as random effects to account for variation between different species, subjects, and samples.

In generalized mixed models, the predictors were standardized to zero mean and unit variance before analyses. We first assessed random effects significance by comparing nested models where each random effect was dropped one at a time using the likelihood-ratio test (LRT, anova function from the R stats package). We then assessed the fixed effects significance with LRTs implemented in drop1 function in the stats package (this function drops individual terms from the full model and report the AIC and the LRT *p*-value). We again used LRTs to compare the full significant models to null models including all random effects but no fixed effects other than the intercept. The difference in Akaike information criterion (ΔAIC) between full and null model and their associated *p*-value are reported in Supplementary Tables S3, S4 and S5. As an additional evaluation of the goodness of fits, we estimated the coefficient of determination (R^2) using the r2 function from the performance R package. Two values are reported: the marginal R^2, a measure of the variance explained only by fixed effects, and the conditional R^2, a measure of the variance explained by the entire model (Supplementary Table S5). We evaluated the GLMM fits by inspecting the residuals using the DHARMa library in R (simulateResiduals and plot functions). In generalized additive mixed models (GAMs), we evaluated the fits by inspecting residual distributions and fitted-observed values plots using the gam.check function from the mgcv R package. Adjusted R^2 (from summary function from the mgcv R package) values are reported as a goodness of fits. All model outputs (summary function from mgcv and glmmTMB R packages) are reported in the **Supplementary File 1**.

Correlation analyses and scatter plots between community diversity and within-species genetic diversity

Only species present in at least four samples were retained to produce the scatter plots (ggplot function in the ggplot2 R package) and to test the relationship between community diversity and within-species genetic diversity with correlation analyses (Pearson when the relationship is linear and Spearman otherwise; cor.test function from the stats R package).

Community diversity is correlated with strain-level diversity

bioRxiv preprint doi: https://doi.org/10.1101/2022.03.08.483496; this version posted March 8, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

To assess evidence for DBD in the gut microbiome, we first tested the relationship between community diversity and within-species polymorphism rate. Because scatter plots (**Figs 2A,B, S1,S2**) showed non-linear trends, we fitedt a separate generalized additive mixed model (GAM) with polymorphism rate in a focal species as a function of each of the community diversity metrics (Shannon index, species richness and rarefied species richness).

We then sought to test this relationship with community diversity calculated at higher taxonomic ranks (from genus to phylum). We used GTDBK and the Genome Taxonomy Database (GTDB) (Chaumeil et al., 2020) to annotate MIDAS reference genomes. Richness at each level was estimated with the total number of distinct units in the sample. Shannon index was calculated based on the relative abundances table from MIDAS (469 samples*5952 species). At each level and for every distinct unit from the sample, we used the sum of the abundances of all species belonging to the focal unit to calculate the Shannon index (using the diversity function from R vegan library). We then fit two GAMs for each taxonomic rank (from genus to phylum) with Shannon diversity and richness as the predictors of polymorphism rate in a focal species (with the coverage per sample as a covariate and species name, sample and subject identifiers as random effects). All the GAMs in this section were fitted with a beta error distribution with logit-link function because polymorphism rate is a continuous value strictly bounded by 1, and all the terms were smooth terms (See **Table S2** and **Supplementary File 1** for additional model details).

As a second test of DBD in the HMP data, we looked at the relationship between strain count in a focal species and community diversity. Because scatter plots (**Figs 2C**,**D**, **S5**,**S6**) showed a linear trend, we fit separate generalized linear mixed models (GLMMs) with strain count in a focal species as a function of community diversity estimated with Shannon diversity, species richness, or rarefied species richness. As strain number is positive count data, we compared many zero-truncated count models based on the Akaike information criterion (AIC) score (AICtab function from bbmle R library) (Brooks et al., 2017). We fit the model with the truncated negative binomial distribution (truncated_nbinom2 in glmmTMB; the second best fit) in order to resolve the overdispersion detected in the best fit (the truncated Poisson model) using the check_overdispersion function from the performance R package as described here: https://bbolker.github.io/mixedmodels-misc/glmmFAQ.html.

bioRxiv preprint doi: https://doi.org/10.1101/2022.03.08.483496; this version posted March 8, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

As in the previous section, we tested the relationship between strain count and community diversity at higher taxonomic levels from genus to phylum, fitting a separate GLMM with strain diversity in a focal species as a function of each metric of diversity (Shannon and richness) at higher taxonomic levels. All GLMMs details are reported in **Table S4** and **Supplementary File 1**.

Genetic diversity as a function of community diversity over time

To test DBD over time, we used HMP samples with multiple time points from the same person to look at the relationship between polymorphism change (delta polymorphism) between two time points and community diversity at the earlier time point. We fit Generalized additive mixed models with delta polymorphism as a function of community diversity at the earlier time point, and added the coverage per sample at the earlier time point as a covariate when diversity was not estimated on rarefied data, as well as species name, sample and subject identifiers as random effects. We used a Gaussian GAM since delta polymorphism is a continuous number that can take on negative values (**Supplementary File 1**).

In addition, we investigated the effect of community diversity at one time point on gene variation at the subsequent time point. We used separate negative binomial generalized linear mixed models with gene gain as the response and each of the metrics of community diversity as the predictor with the same covariates and random effects used in the previous models (**Supplementary File 1**). The same method was used to test how gene loss was related to community diversity (**Table S5**, **Supplementary File 1**).

HMP longitudinal data were sampled at a time lag of ~6 months. To analyze time series at higher resolution, we used longitudinal metagenomic data from a highly sampled healthy donor (host *am,* sampled 206 times spanning 539 days between 2014-12-03 and 2016-05-25) (Poyet et al., 2019). We tested the relationship between community diversity and genetic variation (polymorphism change and gene content variation) in *B. vulgatus*. *B. vulgatus* is the most abundant species in all *am* samples (mean coverage=58.46 and median=54.22). Community diversity was estimated with richness and Shannon index calculated on rarefied data to 5 million reads per sample. We used a Spearman correlation test (cor.test function from the stats R package) for the diversity-delta

bioRxiv preprint doi: https://doi.org/10.1101/2022.03.08.483496; this version posted March 8, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

polymorphism relationship (a nonlinear relationship) and Pearson correlations for both diversity-gene loss and diversity-gene gain relationships (linear relationships) (**Figures 4** and **S11**).

Testing the Black Queen Hypothesis in HMP

The negative relationship between gene loss in focal species and community diversity observed in HMP and Poyet et al. (2019) data suggested the Black Queen Hypothesis (BQH) in the gut microbiome. We sought to further test the BQH by comparing the content in genes and pathways in a focal species to those present in the surrounding community. We used generalized additive models (GAMs) to account for the non-linearity of the relationships (**Figures 5**, **S13**, **S14**). As in all our models, we added the coverage per sample as a covariate as well as species name, sample, and subject identifiers as random effects. Because both responses were count data, we compared Poisson and negative binomial GAMs in both cases by looking at residual distribution and fitted-observed values plots (gam.check function from the mgcv R package). We used a negative binomial GAM for gene richness and a Poisson GAM for pathway richness, both with log-link function. All the terms were specified as smooth terms, see **Table S6** and **Supplementary File 1** for additional model details.

Acknowledgements

We sincerely thank members of the Garud and Shapiro labs for their feedback during the development of this paper. NRG received support from the Paul Allen Frontiers Group, a University of California Hellman fellowship, a UCLA Faculty Career Development award, and the Research Corporation for Science Advancement. DWC received funding support from NIH R25 MH 109172. BJS was supported by a Natural Sciences and Engineering Research Council of Canada (NSERC) Discovery Grant and a Canada Research Chair.

References

Albalat R, Cañestro C. 2016. Evolution by gene loss. *Nat Rev Genet*. doi:10.1038/nrg.2016.39

Almeida A, Mitchell AL, Boland M, Forster SC, Gloor GB, Tarkowska A, Lawley TD, Finn RD. 2019. A new genomic blueprint of the human gut microbiota. *Nature* **568**:499–504. doi:10.1038/s41586-019-0965-1

bioRxiv preprint doi: https://doi.org/10.1101/2022.03.08.483496; this version posted March 8, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

Brooks ME, Kristensen K, Benthem KJ van, Magnusson A, Berg CW, Nielsen A, Skaug HJ, Mächler M, Bolker BM. 2017. Modeling zero-inflated count data with glmmTMB. *BioRxiv*.

Calcagno V, Jarne P, Loreau M, Mouquet N, David P. 2017. Diversity spurs diversification in ecological communities. *Nat Commun* **8**. doi:10.1038/ncomms15810

Chaumeil PA, Mussig AJ, Hugenholtz P, Parks DH. 2020. GTDB-Tk: A toolkit to classify genomes with the genome taxonomy database. *Bioinformatics* **36**. doi:10.1093/bioinformatics/btz848

Edgar RC. 2010. Search and clustering orders of magnitude faster than BLAST. *Bioinformatics* **26**. doi:10.1093/bioinformatics/btq461

Estrela S, Diaz-Colunga J, Vila JCC, Sanchez-Gorostiaga A, Sanchez A. 2022. Diversity begets diversity under microbial niche construction. *BioRxiv*.

Franzosa EA, McIver LJ, Rahnavard G, Thompson LR, Schirmer M, Weingart G, Lipson KS, Knight R, Caporaso JG, Segata N, Huttenhower C. 2018. Species-level functional profiling of metagenomes and metatranscriptomes. *Nat Methods* **15**:962–968. doi:10.1038/s41592-018-0176-y

Garud NR, Good BH, Hallatschek O, Pollard KS. 2019. Evolutionary dynamics of bacteria in the gut microbiome within and across hosts. *PLoS Biol* **17**:e3000102. doi:10.1371/JOURNAL.PBIO.3000102

Garud NR, Pollard KS. 2020. Population Genetics in the Human Microbiome. *Trends Genet* **36**:53–67. doi:10.1016/j.tig.2019.10.010

Good BH, Rosenfeld LB. 2022. Eco-evolutionary feedbacks in the human gut microbiome. *bioRxiv*.

Groussin M, Poyet M, Sistiaga A, Kearney SM, Moniz K, Noel M, Hooker J, Gibbons SM, Segurel L, Froment A, Mohamed RS, Fezeu A, Juimo VA, Lafosse S, Tabe FE, Girard C, Iqaluk D, Nguyen LTT, Shapiro BJ, Lehtimäki J, Ruokolainen L, Kettunen PP, Vatanen T, Sigwazi S, Mabulla A, Domínguez-Rodrigo M, Nartey YA, Agyei-Nkansah A, Duah A, Awuku YA, Valles KA, Asibey SO, Afihene MY, Roberts LR, Plymoth A, Onyekwere CA, Summons RE, Xavier RJ, Alm EJ. 2021. Elevated rates of horizontal gene transfer in the industrialized human microbiome. *Cell* **184**. doi:10.1016/j.cell.2021.02.052

He Y, Zhou BJ, Deng GH, Jiang XT, Zhang H, Zhou HW. 2013. Comparison of microbial diversity determined with the same variable tag sequence extracted from two different PCR

bioRxiv preprint doi: https://doi.org/10.1101/2022.03.08.483496; this version posted March 8, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

amplicons. *BMC Microbiol* **13**. doi:10.1186/1471-2180-13-208

Hibbing ME, Fuqua C, Parsek MR, Peterson SB. 2010. Bacterial competition: Surviving and thriving in the microbial jungle. *Nat Rev Microbiol*. doi:10.1038/nrmicro2259

Human Microbiome Project Consortium T. 2012. A framework for human microbiome research The Human Microbiome Project Consortium*. *Nature* **486**.

Jousset A, Eisenhauer N, Merker M, Mouquet N, Scheu S. 2016. High functional diversity stimulates diversification in experimental microbial communities. *Sci Adv* **2**. doi:10.1126/sciadv.1600124

Koskiniemi S, Sun S, Berg OG, Andersson DI. 2012. Selection-driven gene loss in bacteria. *PLoS Genet* **8**. doi:10.1371/journal.pgen.1002787

Langmead B, Salzberg SL. 2012. Fast gapped-read alignment with Bowtie 2. *Nat Methods* **9**. doi:10.1038/nmeth.1923

Lenski RE. 2017. Experimental evolution and the dynamics of adaptation and genome evolution in microbial populations. *ISME J*. doi:10.1038/ismej.2017.69

Lloyd-Price J, Mahurkar A, Rahnavard G, Crabtree J, Orvis J, Hall AB, Brady A, Creasy HH, McCracken C, Giglio MG, McDonald D, Franzosa EA, Knight R, White O, Huttenhower C. 2017. Strains, functions and dynamics in the expanded Human Microbiome Project. *Nature* **550**:61–66. doi:10.1038/nature23889

Madi N, Vos M, Murall CL, Legendre P, Shapiro BJ. 2020. Does diversity beget diversity in microbiomes? *Elife* **9**. doi:10.7554/eLife.58999

McCutcheon JP, Moran NA. 2012. Extreme genome reduction in symbiotic bacteria. *Nat Rev Microbiol*. doi:10.1038/nrmicro2670

Morris JJ, Lenski RE, Zinser ER. 2012. The black queen hypothesis: Evolution of dependencies through adaptive gene loss. *MBio* **3**. doi:10.1128/mBio.00036-12

Morris JJ, Papoulis SE, Lenski RE. 2014. Coexistence of evolving bacteria stabilized by a shared Black Queen function. *Evolution (N Y)* **68**. doi:10.1111/evo.12485

Nayfach S, Rodriguez-Mueller B, Garud N, Pollard KS. 2016. An integrated metagenomics pipeline for strain profiling reveals novel patterns of bacterial transmission and biogeography. *Genome Res* **26**:1612–1625. doi:10.1101/gr.201863.115

Nayfach S, Shi ZJ, Seshadri R, Pollard KS, Kyrpides NC. 2019. New insights from uncultivated genomes of the global human gut microbiome. *Nature* **568**:505–510. doi:10.1038/s41586-

bioRxiv preprint doi: https://doi.org/10.1101/2022.03.08.483496; this version posted March 8, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

019-1058-x

Nikoh N, Hosokawa T, Oshima K, Hattori M, Fukatsu T. 2011. Reductive evolution of bacterial genome in insect gut environment. *Genome Biol Evol* **3**. doi:10.1093/gbe/evr064

Padfield D, Vujakovic A, Paterson S, Griffiths R, Buckling A, Hesse E. 2020. Evolution of diversity explains the impact of pre-adaptation of a focal species on the structure of a natural microbial community. *ISME J* **14**. doi:10.1038/s41396-020-00755-3

Poyet M, Groussin M, Gibbons SM, Avila-Pacheco J, Jiang X, Kearney SM, Perrotta AR, Berdy B, Zhao S, Lieberman TD, Swanson PK, Smith M, Roesemann S, Alexander JE, Rich SA, Livny J, Vlamakis H, Clish C, Bullock K, Deik A, Scott J, Pierce KA, Xavier RJ, Alm EJ. 2019. A library of human gut bacterial isolates paired with longitudinal multiomics data enables mechanistic microbiome research. *Nat Med* **25**:1442–1452. doi:10.1038/s41591-019-0559-3

Puigbò P, Lobkovsky AE, Kristensen DM, Wolf YI, Koonin E V. 2014. Genomes in turmoil: Quantification of genome dynamics in prokaryote supergenomes. *BMC Med* **12**. doi:10.1186/s12915-014-0066-4

Reese AT, Dunn RR. 2018. Drivers of microbiome biodiversity: A review of general rules, feces, and ignorance. *MBio* **9**. doi:10.1128/mBio.01294-18

San Roman M, Wagner A. 2021. Diversity begets diversity during community assembly until ecological limits impose a diversity ceiling. *Mol Ecol* **30**. doi:10.1111/mec.16161

San Roman M, Wagner A. 2018. An enormous potential for niche construction through bacterial cross-feeding in a homogeneous environment. *PLoS Comput Biol* **14**. doi:10.1371/journal.pcbi.1006340

Schluter D, Pennell MW. 2017. Speciation gradients and the distribution of biodiversity. *Nature*. doi:10.1038/nature22897

Sharma V, Rodionov DA, Leyn SA, Tran D, Iablokov SN, Ding H, Peterson DA, Osterman AL, Peterson SN. 2019. B-Vitamin sharing promotes stability of gut microbial communities. *Front Microbiol* **10**. doi:10.3389/fmicb.2019.01485

Simonsen AK. 2022. Environmental stress leads to genome streamlining in a widely distributed species of soil bacteria. *ISME J* **16**. doi:10.1038/s41396-021-01082-x

Smillie CS, Sauk J, Gevers D, Friedman J, Sung J, Youngster I, Hohmann EL, Staley C, Khoruts A, Sadowsky MJ, Allegretti JR, Smith MB, Xavier RJ, Alm EJ. 2018. Strain Tracking

bioRxiv preprint doi: https://doi.org/10.1101/2022.03.08.483496; this version posted March 8, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

Reveals the Determinants of Bacterial Engraftment in the Human Gut Following Fecal Microbiota Transplantation. *Cell Host Microbe*. doi:10.1016/j.chom.2018.01.003

Vatanen T, Plichta DR, Somani J, Münch PC, Arthur TD, Hall AB, Rudolf S, Oakeley EJ, Ke X, Young RA, Haiser HJ, Kolde R, Yassour M, Luopajärvi K, Siljander H, Virtanen SM, Ilonen J, Uibo R, Tillmann V, Mokurov S, Dorshakova N, Porter JA, McHardy AC, Lähdesmäki H, Vlamakis H, Huttenhower C, Knip M, Xavier RJ. 2019. Genomic variation and strain-specific functional adaptation in the human gut microbiome during early life. *Nat Microbiol* **4**. doi:10.1038/s41564-018-0321-5

Venturelli OS, Carr A V, Fisher G, Hsu RH, Lau R, Bowen BP, Hromada S, Northen T, Arkin AP. 2018. Deciphering microbial interactions in synthetic human gut microbiome communities. *Mol Syst Biol* **14**. doi:10.15252/msb.20178157

Walker AW, Duncan SH, Louis P, Flint HJ. 2014. Phylogeny, culturing, and metagenomics of the human gut microbiota. *Trends Microbiol*. doi:10.1016/j.tim.2014.03.001

Wu D, Jospin G, Eisen JA. 2013. Systematic Identification of Gene Families for Use as “Markers” for Phylogenetic and Phylogeny-Driven Ecological Studies of Bacteria and Archaea and Their Major Subgroups. *PLoS One* **8**:e77033. doi:10.1371/journal.pone.0077033

Zhao S, Lieberman TD, Poyet M, Kauffman KM, Gibbons SM, Groussin M, Xavier RJ, Alm EJ. 2019. Adaptive Evolution within Gut Microbiomes of Healthy People. *Cell Host Microbe* **25**:656-667.e8. doi:10.1016/j.chom.2019.03.007