**Across-host population genetics analysis of common commensal gut microbiota suggests varied trends in demographic history**

Jonathan C. Mah1, Kirk E. Lohmueller1, 2, 3, Nandita Garud1, 2, 3

1. Bioinformatic Interdepartmental Program, University of California, Los Angeles

2. Department of Ecology and Evolutionary Biology, University of California, Los Angeles

3. Department of Human Genetics, University of California, Los Angeles

**Abstract**

Human commensal gut microbes play a crucial role in host health, including aiding with the digestion of foods that humans cannot digest themselves. Despite the importance of such microbes to human health, there is little knowledge about the evolutionary history of commensal gut microbes, including their demographic histories and how selective forces shape their genetic variation. In this study, we infer the demographic histories of ~25 of the most highly prevalent commensal gut microbial species in North Americans. We find reductions in diversity observed at the species and genetic level in commensal gut microbes sampled from Western populations relative to non-Western rural populations whose diets consist of higher amounts of fiber. Additionally, we find that species experienced a variety of demographic trends, including contractions in effective population size and expansions in effective population size. These findings suggest that commensal gut microbes have varied demographic histories.

**Introduction**

Humans have coexisted for several millions of years with commensal gut microbes, which have been shown to be important for host health (Marchesi et al., 2016). Several of these microbial species exhibit similar phylogenetic patterns as their human hosts, suggesting that humans an microbiota have co-evolved (Moeller et al., 2016). Understanding the evolutionary history of the human microbiome in relation to human anthropological development is of particular importance for investigating how host-microbial interactions are impacted by external forces. Additionally, understanding the interplay between human behavior and gut microbial dynamics on long timescales, i.e., several thousands of years, is necessary to further investigate the interplay of human behavior and gut microbial dynamics on shorter timescales, i.e., spanning months, weeks, and days.

Previous work has shown that many aspects of urbanization, e.g., antibiotic usage (Sonnenburg and Sonnenburg, 2019), and changes in diet (Blaser, 2018), have had profound impacts on microbial diversity. For example, Western populations have reduced microbial species diversity (Sonnenburg et al., 2016), loss of gene content (Tett et al., 2019), reduced nucleotide diversity (Nayfach et al., 2016), and even loss of entire bacterial clades (Tett et al., 2019). Additionally, recent work has shown that the gut microbiomes of rural agrarian populations have depleted genetic diversity compared to hunter-gatherer populations, possibly due to a significant decrease in dietary fiber (Sonnenburg et al., 2016). Sonnenburg et al performed a study in mice in which the fiber content of diets was modulated to investigate impacts on microbiome health – this study provided further evidence that reductions in dietary fiber led to reductions in genetic diversity in the gut microbiome (Sonnenburg et al., 2016). The onset of agricultural expansion and the subsequent anthropological events of urbanization may have similarly caused a reduction in genetic diversity due to shifts in diet, which suggests that such forces may be important environmental stressors on human hosts and their microbiota.

Reductions in genetic diversity often result in large decreases of population size and even extinctions (Zalewski et al., 2016; Lande, 1998), whereas increases in genetic diversity can often lead to improved demographic stability or even expansions (Pennekamp et al., 2018). Previous work by Cornejo et al found that there was a demographic expansion in the cavity-causing oral bacteria, *Streptococcus mutans*, coincident with the onset of agricultural expansion ~ 20,000 years go (Cornejo et al., 2013). This suggests that *S. mutans* found an expanded niche via new available resources from the shift in diet. However, gut microbiota have instead experienced a reduction in diversity at the species and genetic levels (Sonnenburg et al., 2016; Tett et al., 2019; Nayfach et al., 2016). While these findings provide insight into the big-picture ecological effects of urbanization, we lack insight into detailed evolutionary histories, such as demography and adaptation of the human gut microbiome over long timescales spanning history.

The site-frequency-spectrum, or SFS, is a summary statistic which describes the distribution of derived allele frequencies from a given sample of DNA sequences. Although this statistic is simple, the SFS has a handful of useful key characteristics, e.g., it is highly sensitive to the underlying evolutionary impacts of demography and linked selection (Nielsen, 2000). This sensitivity allows for powerful maximum-likelihood inference methods to infer demography (Gutenkunst et al., 2009).

Here, we leverage an SFS-based approach to perform population genetics on 26 highly prevalent commensal microbial species in the gut microbiome.

**Results**

North American gut microbiomes have decreased nucleotide diversity relative to Madagascar gut microbiomes

We computed nucleotide diversity in gut microbial species for individuals sampled from a North American cohort (Turnbaugh et al., 2007) and from an African cohort (Pasolli et al., 2019). This African cohort consists of the gut microbiomes of two rural communities from Northeastern Madagascar, which contrasts with the predominantly urban cohort sampled from North America. Figure 1 shows the distribution of average genome-wide within-host nucleotide diversity for several highly prevalent gut microbial species.

We find a significant difference (as computed by a Wilcoxon rank sum test) between the North American and African cohorts in nine species: *Ruminococcus bromii* (p = 0.020), *Bifidobacterium longum* (p = 0.044), *Eubacterium eligens* (p = 0.008), *Faecalibacterium prausnitzii* 62201 (p = 2.4E-11), *Oscillibacter sp.* (p = 3.4E-15), *Faecalibacterium cf.* (p = 6.9E-15), *Prevotella copri* (p < 2E-16), and *Faecalibacterium prausnitzii* 57453 (p < 2E-16). In all of these cases, species found in the North American gut microbiome have decreased nucleotide diversity compared to the African gut microbiome.

Similarly, a paired Wilcoxon rank sum test of the across-host nucleotide diversity values of the African and North American cohort also reveals a significant difference (p = 0.02661), suggesting that African gut microbiomes have greater across-host nucleotide diversity compared to their North American counterparts.

The distribution of within-host nucleotide diversity typically ranges between 1E-3 and 3E-2 – previous work (Garud, Good et al., 2019) suggests that nucleotide diversity near the upper bound of this range, i.e., approximately 1E-2, is inconsistent with the clonal expansion of a single colonization event. In this study, we find that the across-host nucleotide diversity for both the African and North American cohorts is higher than the mean of the distribution of within-host nucleotide diversity for each species. Taken together, these two findings suggest that there exist multiple strains of these bacterial species circulating within hosts. Furthermore, these findings suggest that, in both cohorts, there is an even greater number of strains present across hosts.

These observed decreases in nucleotide diversity in North American gut microbiomes relative to African gut microbiomes match findings in literature (Sonnenburg et al., 2016; Blaser, 2018; Tett et al., 2019).

A picture containing scatter chart

Description automatically generated

**Figure 1: Nucleotide diversity within-hosts and aggregated across hosts in African and North American gut microbiomes.**  
For species present in at least 5 hosts for both the African and North American cohorts, we report the distribution of within-host nucleotide diversity and the average nucleotide diversity aggregated across hosts. Each dot represents the average within-host nucleotide diversity of a single host; larger diamonds represent the across-host nucleotide diversity as computed via a pairwise approach (see Methods). For most species, we find decreased within-host and across-host nucleotide diversity in the North American cohort relative to the African cohort. In nine of these species, we find a significant difference in the distribution of within-host nucleotide diversity, as computed via a Wilcoxon rank sum test.

Simulations give insight into the interpretation of the shape and distribution of the across-host site-frequency spectrum

Estimates of nucleotide diversity in the human microbiome show that North American gut microbiomes have decreased nucleotide diversity relative to African gut microbiomes. To better understand the underlying evolutionary forces which have contributed to this difference in nucleotide diversity, e.g., demography and linked selection, we analyzed a site-frequency spectrum (SFS) of synonymous variants. Analysis of the shape and distribution of the SFS is a key population genetics approach (Wakeley and Hey, 1997; Nielson, 2000). The SFS is known to be sensitive to several different underlying evolutionary effects, such as demographic changes, linked selection, and selective effects; however, the SFS of synonymous variants is thought to be under neutral selection, thereby allowing for greater inference of demography and linkage. We construct our SFS from synonymous variants found across hosts to better represent the effects of demography and linkage for common commensal gut microbial species.

To illustrate an example of interpreting the shape and distribution of the SFS, we used a forward-in-time population genetics simulation framework (Haller and Messer, 2019) to model a population under three different simplified demographic scenarios: a population undergoing demographic equilibrium, a population undergoing demographic expansion, and a population undergoing demographic contraction. We report the SFS’s of these simulated populations in Figure 2.

A picture containing text

Description automatically generated

**Figure 2: Across-host site-frequency spectra of a simulated population undergoing different demographic trends.**   
Caption TBD.

In the first scenario, we model a population undergoing demographic equilibrium, i.e., the case in which there is no recent change in population size and no gene flow. We observe an SFS which closely matches the expected allele frequency spectrum that can be calculated for this scenario via either a coalescent or diffusion approach (Wakeley 2009, Crow 2017). For a sample of size , the expected allele frequency spectrum is computed as

Where is the population scaled mutation rate, given by the produce of two multiplied by the population size and the mutation rate .

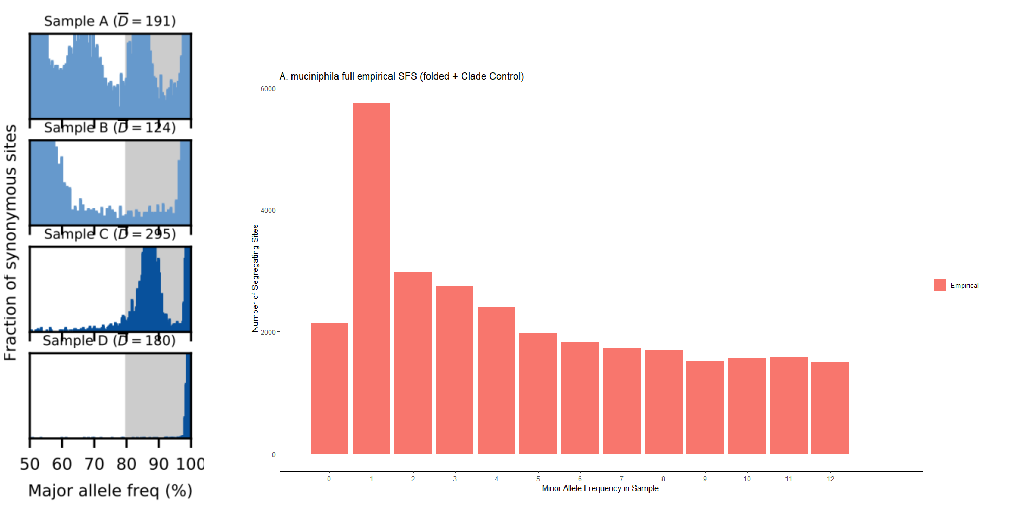
In the second scenario, we model a population undergoing demographic expansion, i.e., an increase in effective population size over time. Of note, relative to the SFS of a population undergoing demographic equilibrium, we observe a proportional increase in frequency for rare variants, i.e., SNPs which are present in a small number of individuals relative to the overall sample size.

In the third scenario, we model a population undergoing demographic contraction, i.e., a decrease in effective population size over time. In contrast with a demographic expansion, here instead observe a proportional increase in frequency for common variants relative to the SFS of a population undergoing demographic equilibrium.

One case of previous work leveraged an SFS-based approach to perform population genetic analysis of the whole genome of *Streptococcus mutans* (Cornejo et al., 2013). Here, we employed a similar SFS-based method to infer the demography of several different commensal human gut microbials species simultaneously. Cornejo et al., used isolate data; however, here we used shotgun metagenomic data, from which we inferred “quasi-phased” haplotypes essentially describing the most prominent strain of bacteria for each species and for each host (see Methods).

Within-host site-frequency spectrums differ from across-host site-frequency spectrums

In contrast to the across-host SFS, which gives insight into the population genetic processes which manifest as demographic changes and genome-wide linkage effects, the within-host SFS grants insight into host-specific patterns of genetic, species, and strain diversity. Typically, the within-host SFS contains multiple peaks, whereas the across-host SFS is more often characterized by a smooth and gradual monotonic decline in allele frequency from rare variants to common variants (Garud, Good et al., 2019; Truong et al., 2017).



**Figure 3A: Four within-host SFS’s of Akkermansia muciniphila sampled from a North American gut microbiomes. Figure 3B: The empirical across-host SFS of Akkermansia muciniphila sampled across North American hosts.**   
Caption TBD.

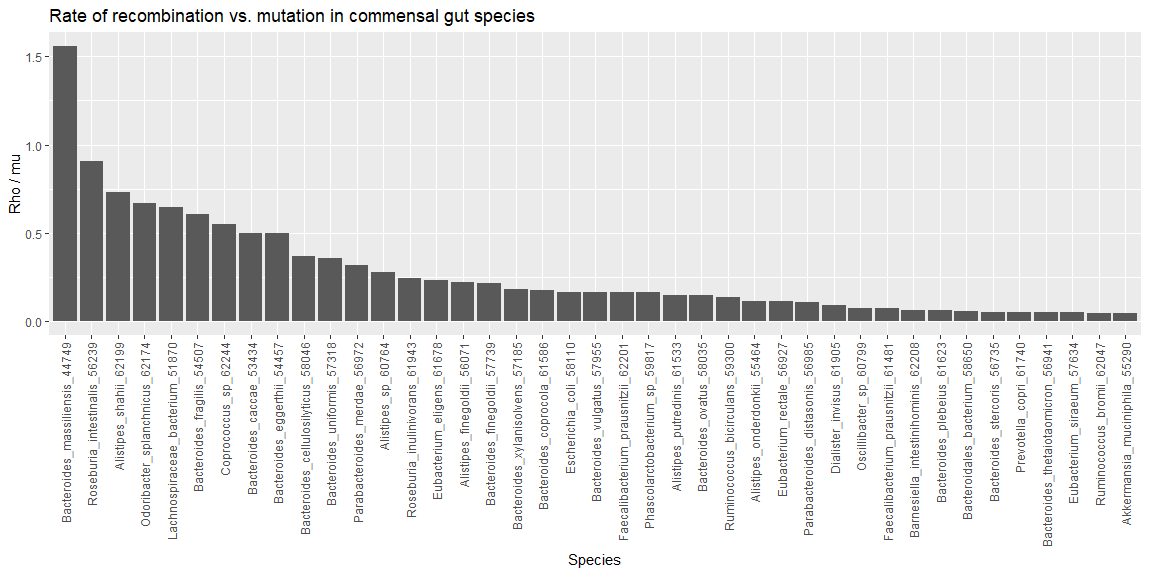
The presence of peaks in the within-host SFS may be indicative of the number of strains colonizing a specific host, which is the primary source of difference between the within-host and across-host SFS. More specifically, this difference between the within-host and across-host SFS arises due to the fact that a given host can only be colonized by a few strains, which represent a subset of all strains that can colonize host. In Figure 3A we report four within-host SFSs of *Akkermansia muciniphila*, while in Figure 3B we report the empirical SFS computed across hosts.

Estimates of recombination are qualitatively consistent with other bacterial species

An important assumption made when leveraging the SFS for demographic inference is that of site independence, i.e., the assumption of free recombination between sites (Bustamante et al., 2001).

Recombination in bacteria typically occurs via some combination of three methods: 1) natural transformation, i.e., uptake of genetic material from the environment, 2) transduction, i.e., sharing of genetic material between bacteria through bacteriophage infection, and 3) conjugation, i.e., direct transfer of genetic material between bacterial cells such as through a pilus or fimbria (Fraser et al., 2007). Of particular relevance to common commensal gut microbial species is homologous recombination, in which horizontal gene transfer between closely related species, strains, or individual cells can result in existing genomic regions being overwritten (Hanage, 2016). Previous studies have reported that there exists a great amount of heterogeneity in both the rate of recombination within a species, as well as the length of segments which are transferred between and within species (Liu and Good, 2022).

When considering evolutionary histories, it can be useful to organize descendants into monophyletic groups known as clades, i.e., a grouping of lineages consisting of one common ancestor and its descendants. Previous studies have shown that there is a much higher degree of recombination within a clade than between different clades (Garud, Good et al., 2019). In some species, such as *Bacteroides vulgatus*, studies have reported up to a five-fold decrease in recombination between clades as opposed to within clades (Liu and Good, 2022). Furthermore in the case of *B.vulgatus*, work by Costea and Bork has shown that different clades have sufficient evolutionary distance to be considered separate subspecies, capable of stable coexistence in the same host (Costea et al., 2017). Thus, by restricting samples to those which belong to the top clade, we can sample our SFS from a well-mixed population. In Figure 4, we report the ratio of recombination rate to mutation rate in common commensal gut bacterial species found in North American gut microbiomes. We observe recombination rates that are qualitatively consistent with observations in other bacterial species, i.e., approximately on the order of a ratio of 0.1 to 3.0 for a comparison of recombination rate to mutation rate (Vos and Didelot, 2009).



**Figure 4: Ratio of recombination rate over mutation rate for 26 highly prevalent common commensal human gut microbial species.** Caption TBD.

Demographic analysis of gut microbiota reveals a large range of changes in effective population size across species

Chart, bar chart

Description automatically generated

**Figure 5: Comparison of Empirical and model site-frequency spectrum for R. bromii.** Caption WIP.

Chart, scatter chart

Description automatically generated

**Figure 6: Likelihood surface of demographic inference for R. bromii.** Caption WIP.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Table 1: Maximum likelihood inferred demographic parameters for highly prevalent commensal human gut microbial species.** | | | | |
|  | **Two Epoch** | | | **One Epoch** |
| **Species** | Nu | Tau | Likelihood | Likelihood |
| A. finegoldii | 0.0536736 | 0.192409 | -51.2531 | -226.6443517359121 |
| A. muciniphila | 0.412214 | 0.184787 | -35.0968 | -394.4640223844726 |
| A. onderdonkii | 0.00077792 | 0.00578665 | -35.7025 | -367.26556546208485 |
| A. putredinis | 9.61159 | 3.63098 | -36.1261 | -863.2523605907211 |
| A. shahii | 18.7669 | 0.0499848 | -37.2099 | -92.12656956679211 |
| B. bacterium | 0.685901 | 0.116485 | -47.2415 | -83.95198106999351 |
| B. caccae | 1.66655 | 0.0717302 | -32.5175 | -57.22309009287301 |
| B. cellulosilyticus | 2.66512 | 1.39504 | -35.0243 | -605.3428693350315 |
| B. fragilis | 2.92751 | 0.288772 | -35.0196 | -621.4516522293925 |
| B. intestinihominis | 0.540065 | 0.050769 | -47.1256 | -159.78453733334754 |
| B. ovatus | 3.52144 | 0.0446575 | -30.6202 | -57.29119777634651 |
| B. thetaiotaomicron | 0.0427133 | 0.22371 | -47.1437 | -79.68659050254428 |
| B. uniformis | 1.74006 | 1.18761 | -30.3353 | -66.4434523929051 |
| B. vulgatus | 1.13481 | 0.0430459 | -33.5302 | -34.81032024811884 |
| B. xylanisolvens | 0.265088 | 0.876258 | -32.9757 | -37.87379244237218 |
| D. invisus | 18.4977 | 0.0217412 | -37.1637 | -95.18332533145986 |
| E. eligens | 55.5916 | 115.696 | -36.5358 | -61.91821099785466 |
| E. rectale | 2.16442 | 1.87647 | -35.8725 | -238.73120577496502 |
| F. prausnitzii | 1.06848 | 0.243666 | -35.9031 | -40.18847936906059 |
| Oscillibacter sp. | 5.09949 | 0.0852462 | -36.7551 | -245.37306153640384 |
| O. splanchnicus | 2.43161 | 7.38277 | -31.8308 | -32.427532235250965 |
| P. copri | 4.4606 | 0.315347 | -78.4052 | -1494.2754772140547 |
| P. distasonis | 0.807558 | 0.136155 | -41.1449 | -68.50672337859214 |
| P. merdae | 0.737274 | 0.391528 | -35.3025 | -73.68111715850773 |
| Phascolarcto. sp. | 2.57614 | 0.355933 | -34.4246 | -393.170544586656 |
| R. bicirculans | 377.278 | 307.728 | -38.5849 | -1650.5914319833837 |
| R. bromii | 2.96492 | 1.46003 | -34.8367 | -658.0589498106256 |

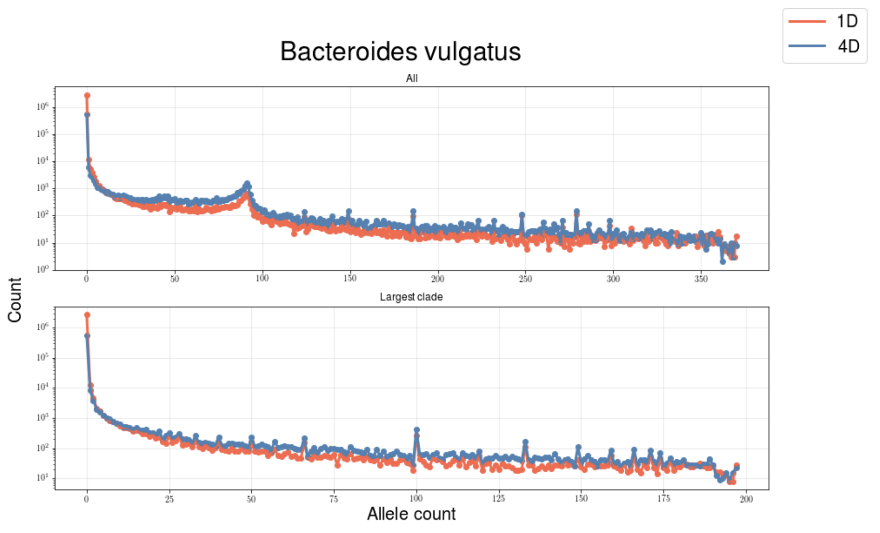
DFE

Comparison of nucleotide diversity and SFS

Clade control reduces confounding effects of population structure

Chart, box and whisker chart

Description automatically generated



**Figure 7: Example dendrogram.** Caption WIP.

* Toy example showing potential diagram vs. actual annotated dendrogram?
* If toy example maybe toy SFS, and then real dendrogram vs. real SFS?

By grouping sequences into clades based on their common evolutionary history, clade control can reduce the effects of certain demographic processes, such as population structure and migration, which can complicate interpretation of the SFS. Clade control can account for the shared ancestry of sequences, resulting in a more accurate representation of genetic diversity. As an example, in Figure 7A, which illustrate a simple phylogeny in which A and B form separate monophyletic groups; however, the entirety of clades A and B are themselves also a monophyletic group. In the case in which we take the SFS for all of the shown individuals, then mutations which arise in the evolutionary timeline represented by the blue slash mark in clade A will have an allele frequency of 8/14, resulting in a potential bump in intermediate frequencies. In Figure 7B, we show an empirical example of this phenomenon, by which when we don’t perform clade control, the resulting SFS may have a large bump in intermediate frequencies, such as with *Bacteroides vulgatus*.

**Discussion**

Several commensal gut microbial species show demographic patterns which parallel human history

Despite broad reductions in genetic diversity observed in North American gut microbiomes, several commensal gut microbial species demonstrate an expansion in effective population size

False evidence of contraction

* Clade control issues
* Balancing Selection

False evidence of expansion

* Purifying Selection
* Positive Selection

**Methods**

Data Availability

All necessary metadata, as well as the source code for the bioinformatics pipeline, downstream analysis, and figure generation are available at Github: <https://github.com/jon-mah/microbiome_demography_manuscript/>

Raw whole-genome shotgun sequencing reads for metagenomic samples from North American fecal microbiomes were downloaded from the Human Microbiome Project Consortium Portal (Turnbaugh et al., 2007). Raw whole-genome shotgun sequencing reads for Madagascar fecal microbiomes were downloaded from the NCBI Sequence Read Archive with the accession numbers PRJNA485056 (Pasolli et al., 2019). Accession numbers associated with this data are in Supplemental Tables S1, S2, and S3.

For North American gut microbiomes, we analyzed 471 fecal samples from 250 healthy North American individuals. For Madagascar gut microbiomes, we analyzed 112 fecal samples from 112 healthy Madagascar individuals.

Identification of bacterial species, genes, and SNVs

We used a standard reference-based approach to quantify bacterial species’ abundance and strain-level genomic variation (Nayfach et al, 2016). FASTQ files for technical and sample replicates from the same time point were merged to increase coverage.

We analyzed gene and SNV content of samples using the MIDAS pipeline, with several additional postprocessing steps similar to as performed in Garud, Good et al., 2019.

The first step of the MIDAS pipeline is to determine the species present in each set of sample(s) for each host. The goal of this “species” step is to include as many present species as possible while omitting species that are absent from the sample or could be mistakenly recruited from other species. MIDAS quantifies the relative abundance of species in a given sample by mapping sequencing reads to a reference database of single-copy “marker” genes unique to each species. We used database version 1.2, downloaded on November 21st, 2016 (Nayfach et al., 2016). In the case that a given species has an average marker gene coverage of at least 3x for a given sample, then it is assigned as truly present. In the case that longitudinal data is available, species are included if marker coverage is at least 3x in at least one time point. Species are identified using an identity threshold of 95% Average Nucleotide Identity (ANI).

Second, MIDAS quantifies SNV counts for each species for each sample. Similar to the “species” step, this “SNV” step leverages a standard reference-based approach. In summary, sequencing reads are aligned to a host-specific panel of reference genomes using Bowtie2 (Langmead and Salzberg, 2012) with the following default MIDAS mapping thresholds: global alignment, MAPID >= 94.0%, READQ >= 20, ALN\_COV >= 0.75, and MAP >= 20.

Third, MIDAS aligns reads to a “pangenome”, i.e., the set of genes shared by all strains in a clade for a given species. Additional postprocessing is applied to account for any genes which may have >= 95% ANI with more than one pangenome. Species are assigned if a marker gene has at least 1x coverage in at least one sample.

Quasi-phasing

We followed the approach as outlined in Garud, Good, et al., 2019 to identify “quasi-phaseable” (QP) samples. In summary, to infer evolutionary changes found in metagenomic samples, must detect changes in allele frequency over time within lineages for each strain. It is important to account for confounding factors which might also manifest as a change in allele frequency, such as strain fluctuation and sampling error, as these also result in allele frequency changes.

To do so, we “quasi-phase” samples in the case when we have sufficiently simple lineage structure, i.e., when we are able to, we identify pairs of alleles which can be confidently assigned to the genome of single lineages. This approach is similar to that of Truong et al., 2017; however, as expanded upon in Garud, Good et al., 2019, here there are bounds on the error for phasing. For further statistical details, please consult the supplement of Garud, Good et al., 2019.

Complete SFS Calculation

Clade Control

MIDAS uses a purely operational species definition, i.e., 95% ANI Thus, it may occur that geographically and genetically isolated populations may exceed this threshold and thus have sequence reads mapped to the same reference genome. These mixtures of genetically isolated populations may confound typically SFS-based methods due to inaccurate estimates of recombination as a result of increased between-population genetic diversity. To address this issue, we manually partitioned species into top-level clades based on sequence similarity clusters.

In Figure XYZ, we illustrate this method of clade control with a dendrogram. Based on these dendrograms, lineages are manually assigned to clades, and further downstream analysis is limited to only the top-most clade for each species. After “quasi-phasing”, there were no cases in which QP samples had < 95% ANI.

Estimating nucleotide diversity

The core-genome nucleotide diversity of each species, denoted here as , was estimated using SNP frequencies via the following estimation:

Where is the frequency of the reference allele at site *i* of the core-genome,

Representing the frequency of alternative alleles relative to the reference, and denotes the length in base pairs of the core-genome. This approach is adapted from Mu et al., 2011 to estimate nucleotide diversity from polymorphism data.

We estimate intra-sample nucleotide diversity, denoted as using reference allele frequencies from individual metagenomic samples by taking the mean nucleotide diversity of all sites for a given host.

We estimate inter-sample nucleotide diversity, denoted as by calling consensus alleles at each site for each pair of samples, estimating allele frequencies at the given site for the given pair to gain a pairwise estimate of allele frequencies. We then take the average of pairwise estimates across all sites for a given pair to gain a paired average nucleotide diversity estimate, and then take the average of all paired average nucleotide diversity estimates to estimate aggregated across all individuals.

Down Sampling

As part of our analysis, we projected the empirical SFS down from a larger sample size to a smaller sample size, i.e., 14 QP samples. This involves randomly averaging over all possible resamplings of the larger sample size data. Down sampled sites are randomly drawn, with replacement, from the hypergeometric distribution (Gutenkunst et al., 2009).

This down sampling is done to account for the cases of missing data, particularly for rare variants, as some sites are not accurately called for all individuals; thus, SNPs with more successful calls are projected down to lower frequencies. Species for which we had fewer than 20 QP samples were discarded.

Inference using the site-frequency spectrum

To infer demographic histories of common commensal human gut microbiota, we use data summarized from the site-frequency spectrum (SFS) under a maximum-likelihood framework (Gutenkunst et al., 2009). The SFS is a summary statistic which describes the distribution of derived allele frequencies across the genome. The SFS can be represented as a vector, in which each index of the vector describes the number of SNPs at frequency given chromosomes. In simpler terms, here, the SFS describes the number of QP samples a SNPs appear in.

Patterns of segregating sites are affected by a combination of linked selection, demography, and selective effects; thus, we remove non-synonymous mutations from our SFS. This is done so with the assumption that synonymous sites are less impacted by selective effects; thus allowing us to estimate a demographic model for putatively neutral, synonymous sites for which we only consider the effects of linked selection and demographic in our demographic models.

Inference of demographic models

To infer demographic models, we use data summarized from site-frequency spectra (SFSs) to fit a maximum-likelihood inference method to find the best-fit demographic parameters under a given set of model specifications. To fit demographic parameters, we use (Gutenkunst et al., 2009) to recover two main demographic parameters, being , which denotes the time in generations scaled by , where is the effective ancestral population size, and , which denotes the change in effective population size relative to the ancestral effective population size.

finds the maximum composite likelihood solution to the diffusion equation to infer the distribution of allele frequencies given some demographic model specifications, , where denotes the set of parameters specified by the model. Then, the expected SFS can be defined as:

The multinomial likelihood, computed from the folded SFS, is then maximized, which provides maximum-likelihood demographic parameters as such:

Where denotes the observed count of SNPs at frequency in the folded SFS. Note that the multinomial likelihood uses the proportion of SNPs at a given frequency, rather than the raw count of SNPs from the model. Thus, the likelihood computation does not require *a priori* estimates of either the mutation rate or the ancestral population size. The population-scaled mutation rate of synonymous sites, denoted as , is computed as the scaling factor between the optimized proportional SFS and the empirical data.

To interpret the inferred parameters, and , we convert population values to effective population sizes. To convert time in generations to years, we assumed a mutation rate of substitutions per site per generation with an estimated generation time of one division per day.

We employ a set of two different population demographic model specifications: a one-epoch model, which consists of no free parameters or inferred change in effective population size, and a two-epoch model, which consists of an instantaneous change in effective population size after an inferred number of generations. The one-epoch demographic model is a useful null expectation whose likelihood gives insight into whether or not there was a change in effective population size over time.

**Acknowledgements**

This work was funded in part through the NIGMS T32 5T32GM008185 (Systems in Integrative Biology) training grant. The authors would like to thank members of the Garud and Lohmueller labs, both past and present, for helpful discussions, shared code, and edits / revision of the manuscript.

**Competing interests**

The authors declare no competing interests.

**References**

Blaser, M. J. (2018). The past and future biology of the human microbiome in an age of

extinctions. Cell, 172(6):1173–1177.

Bustamante, C. D., Wakeley, J., Sawyer, S., and Hartl, D. L. (2001). Directional selection and

the site-frequency spectrum. Genetics, 159(4):1779–1788.

Cornejo, O. E., Lefebure, T., Pavinski Bitar, P. D., Lang, P., Richards, V. P., Eilertson, K., Do, T., ´Beighton, D., Zeng, L., Ahn, S.-J., et al. (2013). Evolutionary and population genomics of the

cavity causing bacteria streptococcus mutans. Molecular biology and evolution, 30(4):881–

893.

Costea, P. I., Coelho, L. P., Sunagawa, S., Munch, R., Huerta-Cepas, J., Forslund, K., Hildebrand, F., Kushugulova, A., Zeller, G., and Bork, P. (2017). Subspecies in the global human

gut microbiome. Molecular systems biology, 13(12):960.

Crow, James Franklin. An introduction to population genetics theory. Scientific Publishers, 2017.

Fraser, C., Hanage, W. P., and Spratt, B. G. (2007). Recombination and the nature of bacterial

speciation. Science, 315(5811):476–480.

Garud, N. R., Good, B. H., Hallatschek, O., and Pollard, K. S. (2019). Evolutionary dynamics

of bacteria in the gut microbiome within and across hosts. PLoS biology, 17(1):e3000102.

Gutenkunst, R. N., Hernandez, R. D., Williamson, S. H., and Bustamante, C. D. (2009). Inferring the joint demographic history of multiple populations from multidimensional snp frequency data. PLoS genetics, 5(10):e1000695.

Haller, B. C. and Messer, P. W. (2019). Slim 3: forward genetic simulations beyond the wright–

fisher model. Molecular biology and evolution, 36(3):632–637.

Hanage, W. P. (2016). Not so simple after all: bacteria, their population genetics, and recombination. Cold Spring Harbor perspectives in biology, 8(7):a018069.

Lande, R. (1998). Anthropogenic, ecological and genetic factors in extinction and conservation.

Population Ecology, 40(3):259–269.

Langmead, B. and Salzberg, S. L. (2012). Fast gapped-read alignment with bowtie 2. Nature

methods, 9(4):357–359.

Liu, Z. and Good, B. H. (2022). Dynamics of bacterial recombination in the human gut microbiome. bioRxiv.

Marchesi, J. R., Adams, D. H., Fava, F., Hermes, G. D., Hirschfield, G. M., Hold, G., Quraishi,

M. N., Kinross, J., Smidt, H., Tuohy, K. M., et al. (2016). The gut microbiota and host health:

a new clinical frontier. Gut, 65(2):330–339.

Moeller, A. H., Caro-Quintero, A., Mjungu, D., Georgiev, A. V., Lonsdorf, E. V., Muller, M. N.,

Pusey, A. E., Peeters, M., Hahn, B. H., and Ochman, H. (2016). Cospeciation of gut microbiota with hominids. Science, 353(6297):380–382.

Mu, X. J., Lu, Z. J., Kong, Y., Lam, H. Y., and Gerstein, M. B. (2011). Analysis of genomic variation in non-coding elements using population-scale sequencing data from the 1000 genomes

project. Nucleic acids research, 39(16):7058–7076.

Nayfach, S., Rodriguez-Mueller, B., Garud, N., and Pollard, K. S. (2016). An integrated

metagenomics pipeline for strain profiling reveals novel patterns of bacterial transmission

and biogeography. Genome research, 26(11):1612–1625.

Nielsen, R. (2000). Estimation of population parameters and recombination rates from single

nucleotide polymorphisms. Genetics, 154(2):931–942.

Pasolli, E., Asnicar, F., Manara, S., Zolfo, M., Karcher, N., Armanini, F., Beghini, F., Manghi,

P., Tett, A., Ghensi, P., et al. (2019). Extensive unexplored human microbiome diversity revealed by over 150,000 genomes from metagenomes spanning age, geography, and lifestyle.

Cell, 176(3):649–662.

Pennekamp, F., Pontarp, M., Tabi, A., Altermatt, F., Alther, R., Choffat, Y., Fronhofer, E. A.,

Ganesanandamoorthy, P., Garnier, A., Griffiths, J. I., et al. (2018). Biodiversity increases and

decreases ecosystem stability. Nature, 563(7729):109–112.

Sonnenburg, E. D., Smits, S. A., Tikhonov, M., Higginbottom, S. K., Wingreen, N. S., and

Sonnenburg, J. L. (2016). Diet-induced extinctions in the gut microbiota compound over

generations. Nature, 529(7585):212–215.

Sonnenburg, J. L. and Sonnenburg, E. D. (2019). Vulnerability of the industrialized microbiota.

Science, 366(6464).

Tett, A., Huang, K. D., Asnicar, F., Fehlner-Peach, H., Pasolli, E., Karcher, N., Armanini,

F., Manghi, P., Bonham, K., Zolfo, M., et al. (2019). The prevotella copri complex comprises four distinct clades underrepresented in westernized populations. Cell host & microbe,

26(5):666–679.

Truong, D. T., Tett, A., Pasolli, E., Huttenhower, C., and Segata, N. (2017). Microbial

strain-level population structure and genetic diversity from metagenomes. Genome research,

27(4):626–638.

Turnbaugh, P. J., Ley, R. E., Hamady, M., Fraser-Liggett, C. M., Knight, R., and Gordon, J. I.

(2007). The human microbiome project. Nature, 449(7164):804–810.

Vos, M. and Didelot, X. (2009). A comparison of homologous recombination rates in bacteria

and archaea. The ISME journal, 3(2):199–208.

Wakeley, J. and Hey, J. (1997). Estimating ancestral population parameters. Genetics,

145(3):847–855.

Wakeley, John. Coalescent theory: an introduction. No. 575: 519.2 WAK. 2009.

Zalewski, A., Zalewska, H., Lunneryd, S.-G., Andre, C., and Mikusi ´ nski, G. (2016). Reduced ´

genetic diversity and increased structure in american mink on the swedish coast following

invasive species control. PLoS One, 11(6):e0157972.

**Supplementary Information**

Table S1: North American fecal sample accession numbers

Table S2: Madagascar fecal sample accession numbers

Table S3: Maximum likelihood demographic parameters for different down sampling conditions

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **Two Epoch** | | | **One Epoch** |
| **Species** | Nu | Tau | Likelihood | Likelihood |
| A. finegoldii (10) | 0.0024135 | 0.0160841 | -28.0073 | -158.66738007844833 |
| A. finegoldii (12) | 0.00141166 | 0.0102196 | -38.0423 | -195.1654192197775 |
| A. finegoldii (14) | 0.0536736 | 0.192409 | -51.2531 | -226.6443517359121 |
| A. finegoldii (16) | 0.0216463 | 0.098936 | -66.807 | -254.92853145861318 |
| A. finegoldii (18) | 0.0545651 | 0.199532 | -84.1735 | -279.5723881256772 |
| A. muciniphila (10) | 0.40622 | 0.166818 | -25.3404 | -271.860384337564 |
| A. muciniphila (12) | 0.406192 | 0.221325 | -30.1067 | -340.2099727784189 |
| A. muciniphila (14) | 0.412214 | 0.184787 | -35.0968 | -394.4640223844726 |
| A. muciniphila (16) | 0.385545 | 0.263801 | -40.7719 | -441.06190927399075 |
| A. muciniphila (18) | 0.391512 | 0.240012 | -48.1178 | -482.1546733933401 |
| A. onderdonkii (10) | 0.227832 | 0.0881793 | -25.5034 | -257.2562129893595 |
| A. onderdonkii (12) | 0.000663046 | 0.00503453 | -30.3591 | -314.12664610120737 |
| A. onderdonkii (14) | 0.00077792 | 0.00578665 | -35.7025 | -367.26556546208485 |
| A. onderdonkii (16) | 0.00874967 | 0.0439118 | -41.3301 | -411.1944780140302 |
| A. onderdonkii (18) | 0.000125171 | 0.00116278 | -47.9952 | -452.4473745351561 |
| A. putredinis (10) | 8.43938 | 3.10627 | -22.5545 | -590.910821013696 |
| A. putredinis (12) | 9.19344 | 3.44869 | -28.8994 | -735.249998197641 |
| A. putredinis (14) | 9.61159 | 3.63098 | -36.1261 | -863.2523605907211 |
| A. putredinis (16) | 10.6528 | 4.16344 | -44.1492 | -974.1492077833045 |
| A. putredinis (18) | 10.7771 | 4.21105 | -53.1155 | -1070.3713299493795 |
| A. shahii (10) | 13.9005 | 0.035724 | -23.6239 | -49.8864536305 |
| A. shahii (12) | 17.2738 | 0.0394458 | -28.1309 | -69.2119480339311 |
| A. shahii (14) | 18.7669 | 0.0499848 | -37.2099 | -92.12656956679211 |
| A. shahii (16) | 23.2719 | 0.0338066 | -37.0198 | -117.66999818480463 |
| A. shahii (18) | 3.48215 | 0.0630991 | -42.7689 | -145.3302055638601 |
| B. bacterium (10) | 0.656385 | 0.0924681 | -25.9385 | -50.45437839261831 |
| B. bacterium (12) | 0.692618 | 0.135731 | -34.689 | -66.42410461333293 |
| B. bacterium (14) | 0.685901 | 0.116485 | -47.2415 | -83.95198106999351 |
| B. bacterium (16) | 0.72274 | 0.221349 | -61.9099 | -102.24641285809412 |
| B. bacterium (18) | 0.747965 | 0.26547 | -78.8852 | -121.5256301759091 |
| B. caccae (10) | 1.91278 | 0.0536306 | -23.7114 | -36.301684917431885 |
| B. caccae (12) | 1.40648 | 0.120183 | -28.261 | -46.27593526221608 |
| B. caccae (14) | 1.66655 | 0.0717302 | -32.5175 | -57.22309009287301 |
| B. caccae (16) | 1.60429 | 0.0791623 | -37.06 | -68.42155328457329 |
| B. caccae (18) | 3.27649 | 0.0316115 | -41.312 | -80.89728650631787 |
| B. cellulosilyticus (10) | 2.29372 | 0.828225 | -25.4978 | -416.57659483365387 |
| B. cellulosilyticus (12) | 2.75049 | 1.50239 | -30.3851 | -515.8026292276945 |
| B. cellulosilyticus (14) | 2.66512 | 1.39504 | -35.0243 | -605.3428693350315 |
| B. cellulosilyticus (16) | 2.37427 | 0.975554 | -39.3666 | -686.3523745472339 |
| B. cellulosilyticus (18) | 3.18192 | 1.99256 | -45.0077 | -760.0016756856094 |
| B. fragilis (10) | 3.22129 | 0.248444 | -24.7025 | -371.6239429929792 |
| B. fragilis (12) | 2.40281 | 0.406655 | -39.3029 | -495.2058080779716 |
| B. fragilis (14) | 2.92751 | 0.288772 | -35.0196 | -621.4516522293925 |
| B. fragilis (16) | 3.21259 | 0.237611 | -38.4512 | -747.0948649758061 |
| B. fragilis (18) | 3.0888 | 0.26575 | -43.9282 | -871.2663353748503 |
| B. intestinihominis (10) | 0.149372 | 0.00621968 | -27.798 | -92.51292865669711 |
| B. intestinihominis (12) | 0.437146 | 0.0315171 | -36.2919 | -126.12642765565943 |
| B. intestinihominis (14) | 0.540065 | 0.050769 | -47.1256 | -159.78453733334754 |
| B. intestinihominis (16) | 0.653196 | 0.118365 | -62.1081 | -193.57683995560183 |
| B. intestinihominis (18) | 0.643165 | 0.092557 | -77.7833 | -227.04279680796026 |
| B. ovatus (10) | 2.25642 | 0.0603274 | -22.3608 | -34.6241364497273 |
| B. ovatus (12) | 7.20843 | 0.0339838 | -26.4903 | -44.85214406663749 |
| B. ovatus (14) | 3.52144 | 0.0446575 | -30.6202 | -57.29119777634651 |
| B. ovatus (16) | 8.04801 | 0.0348558 | -34.6214 | -71.16037130275527 |
| B. ovatus (18) | 2.12642 | 0.108158 | -44.4466 | -85.96252530619563 |
| B. thetaiotaomicron (10) | 0.0163749 | 0.100595 | -31.0133 | -56.57709360220542 |
| B. thetaiotaomicron (12) | 0.0143413 | 0.0909163 | -39.1149 | -67.86495644336355 |
| B. thetaiotaomicron (14) | 0.0427133 | 0.22371 | -47.1437 | -79.68659050254428 |
| B. thetaiotaomicron (16) | 0.00659293 | 0.0473145 | -55.6264 | -90.53149830453367 |
| B. thetaiotaomicron (18) | 0.13555 | 0.546058 | -63.9759 | -101.09623839611413 |
| B. uniformis (10) | 32.233 | 49.2075 | -22.083 | -47.09826242811914 |
| B. uniformis (12) | 7.59117 | 10.6886 | -26.1642 | -57.68891387756207 |
| B. uniformis (14) | 1.74006 | 1.18761 | -30.3353 | -66.4434523929051 |
| B. uniformis (16) | 1.77798 | 1.27066 | -34.4863 | -74.54592033988752 |
| B. uniformis (18) | 1.76202 | 1.2601 | -38.5199 | -80.89423736027015 |
| B. vulgatus (10) | 4.9289 | 0.0037879 | -23.6192 | -23.829172435319833 |
| B. vulgatus (12) | 3.01055 | 0.00824215 | -28.3958 | -29.059697599464016 |
| B. vulgatus (14) | 1.13481 | 0.0430459 | -33.5302 | -34.81032024811884 |
| B. vulgatus (16) | 6.1956 | 0.0100282 | -36.182 | -40.15132639392118 |
| B. vulgatus (18) | 1.78105 | 0.0311021 | -39.7184 | -47.51509635860066 |
| B. xylanisolvens (10) | 0.0132507 | 0.0850143 | -22.7839 | -27.502007390339713 |
| B. xylanisolvens (12) | 0.0369925 | 0.202084 | -27.761 | -32.67997566774466 |
| B. xylanisolvens (14) | 0.265088 | 0.876258 | -32.9757 | -37.87379244237218 |
| B. xylanisolvens (16) | 0.0104182 | 0.0720721 | -38.4724 | -43.37195364729632 |
| B. xylanisolvens (18) | 0.125021 | 0.540209 | -44.0514 | -49.182891442799246 |
| D. invisus (10) | 1.57348 | 0.0827088 | -25.8386 | -50.10320348806272 |
| D. invisus (12) | 2.42975 | 0.0476867 | -30.8393 | -71.05069010726947 |
| D. invisus (14) | 18.4977 | 0.0217412 | -37.1637 | -95.18332533145986 |
| D. invisus (16) | 3.02789 | 0.0450853 | -42.962 | -122.28821161779706 |
| D. invisus (18) | 2.21215 | 0.0579241 | -50.6428 | -156.26389196593573 |
| E. eligens (10) | 44.8382 | 92.7077 | -25.7159 | -44.11193197504326 |
| E. eligens (12) | 106.953 | 222.988 | -31.0836 | -53.411666204312496 |
| E. eligens (14) | 55.5916 | 115.696 | -36.5358 | -61.91821099785466 |
| E. eligens (16) | 32.0696 | 65.638 | -42.6105 | -72.00705318163409 |
| E. eligens (18) | 83.0965 | 170.788 | -47.6559 | -80.02451076590751 |
| E. rectale (10) | 1.64542 | 0.769789 | -25.918 | -163.86495196229225 |
| E. rectale (12) | 1.58777 | 0.580629 | -30.7554 | -203.9217927887821 |
| E. rectale (14) | 2.16442 | 1.87647 | -35.8725 | -238.73120577496502 |
| E. rectale (16) | 1.65568 | 0.832631 | -40.2183 | -271.2183950435392 |
| E. rectale (18) | 3.53508 | 4.03943 | -45.6508 | -300.6688830067369 |
| F. prausnitzii (10) | 1.09867 | 1.299 | -26.2847 | -27.848481108148917 |
| F. prausnitzii (12) | 9.21845 | 30.6123 | -31.1614 | -33.83407165128665 |
| F. prausnitzii (14) | 1.06848 | 0.243666 | -35.9031 | -40.18847936906059 |
| F. prausnitzii (16) | 1.06737 | 0.337445 | -40.6422 | -46.20320308065311 |
| F. prausnitzii (18) | 1.09659 | 0.243501 | -45.4701 | -54.57869935754752 |
| Oscillibacter sp. (10) | 5.02523 | 0.0759358 | -24.9166 | -122.49240820956948 |
| Oscillibacter sp. (12) | 8.01039 | 0.0744874 | -29.6372 | -183.95614434688196 |
| Oscillibacter sp. (14) | 5.09949 | 0.0852462 | -36.7551 | -245.37306153640384 |
| Oscillibacter sp. (16) | 4.02196 | 0.0964394 | -45.7138 | -313.25361229792315 |
| Oscillibacter sp. (18) | 2.93863 | 0.133149 | -58.2456 | -386.60885749472527 |
| O. splanchnicus (10) | 9.55467 | 31.6094 | -23.0582 | -23.613390012760647 |
| O. splanchnicus (12) | 1.5589 | 3.85637 | -27.5014 | -28.079710261162745 |
| O. splanchnicus (14) | 2.43161 | 7.38277 | -31.8308 | -32.427532235250965 |
| O. splanchnicus (16) | 11.5442 | 40.7758 | -36.2858 | -36.8739292604796 |
| O. splanchnicus (18) | 1.31261 | 3.03154 | -40.6788 | -41.17584009768916 |
| P. copri (10) | 3.47521 | 0.358743 | -28.7823 | -839.2162791757946 |
| P. copri (12) | 4.11266 | 0.305928 | -37.0422 | -1143.2168014732433 |
| P. copri (14) | 4.4606 | 0.315347 | -78.4052 | -1494.2754772140547 |
| P. copri (16) | N/A | N/A | N/A | N/A |
| P. copri (18) | N/A | N/A | N/A | N/A |
| P. distasonis (10) | 0.776827 | 0.0916526 | -26.7736 | -44.38095023353162 |
| P. distasonis (12) | 0.791112 | 0.105926 | -33.356 | -56.45492961817399 |
| P. distasonis (14) | 0.807558 | 0.136155 | -41.1449 | -68.50672337859214 |
| P. distasonis (16) | 0.813611 | 0.161783 | -49.9537 | -79.98190123672157 |
| P. distasonis (18) | 0.836116 | 0.228785 | -58.9054 | -91.25185015802981 |
| P. merdae (10) | 0.742039 | 0.379442 | -24.8629 | -50.53570047635185 |
| P. merdae (12) | 0.755594 | 0.341855 | -29.8825 | -61.8852600976561 |
| P. merdae (14) | 0.737274 | 0.391528 | -35.3025 | -73.68111715850773 |
| P. merdae (16) | 0.687759 | 0.605537 | -40.7138 | -83.64791875639548 |
| P. merdae (18) | 0.0352227 | 0.176049 | -46.3478 | -93.21962114603184 |
| Phascolarcto. sp. (10) | 2.58953 | 0.328679 | -23.7862 | -244.39453441684418 |
| Phascolarcto. sp. (12) | 2.46391 | 0.322072 | -29.7279 | -319.0501304679924 |
| Phascolarcto. sp. (14) | 2.57614 | 0.355933 | -34.4246 | -393.170544586656 |
| Phascolarcto. sp. (16) | 2.51285 | 0.540036 | -46.7268 | -458.55350307162144 |
| Phascolarcto. sp. (18) | N/A | N/A | N/A | N/A |
| R. bicirculans (10) | 512.46 | 417.649 | -26.9991 | -1168.6064597172044 |
| R. bicirculans (12) | 310.657 | 252.905 | -32.6474 | -1430.8374296203147 |
| R. bicirculans (14) | 377.278 | 307.728 | -38.5849 | -1650.5914319833837 |
| R. bicirculans (16) | 286.055 | 233.838 | -44.6031 | -1841.5728263737565 |
| R. bicirculans (18) | 258.088 | 210.431 | -50.5602 | -2007.462031447595 |
| R. bromii (10) | 2.99009 | 1.50668 | -25.1369 | -455.1322437282979 |
| R. bromii (12) | 3.01357 | 1.50539 | -29.9893 | -563.2324294410473 |
| R. bromii (14) | 2.96492 | 1.46003 | -34.8367 | -658.0589498106256 |
| R. bromii (16) | 2.85747 | 1.30462 | -39.9897 | -746.0033986139315 |
| R. bromii (18) | 3.00012 | 1.49253 | -46.0828 | -825.2435880425423 |