Microbiome source tracking using single nucleotide variants

Leah Briscoe[1]*, Eran Halperin[2,3,4,5,6], Nandita R. Garud[3,7]*

1. Bioinformatics Interdepartmental Program, University of California Los Angeles, Los Angeles, Los Angeles, CA, United States of America
2. Department of Computer Science, University of California Los Angeles, Los Angeles, CA, United States of America
3. Department of Human Genetics, David Geffen School of Medicine, University of California Los Angeles, Los Angeles, CA, United States of America
4. Department of Computational Medicine, David Geffen School of Medicine, University of California Los Angeles, Los Angeles, CA, United States of America
5. Department of Anesthesiology and Perioperative Medicine, David Geffen School of Medicine, University of California Los Angeles, Los Angeles, CA, United States of America
6. Institute of Precision Health, University of California Los Angeles, CA, United States of America
7. Department of Ecology and Evolutionary Biology, University of California Los Angeles, CA, United States of America

*Corresponding author

Abstract

Microbiomes are composed of hundreds to thousands of species of microorganisms living on and in our body and also in our environment. Elucidating the sources of these community members has been of great interest in the field to understand underlying ecological and colonization dynamics. Microbiomes are likely mixtures of several other microbiomes. The estimation of the contribution of various source microbiomes to a given community is known as source tracking. While emphasis has been placed on source tracking using species composition, single nucleotide variants (SNVs) within species may be more informative because rare variants can be highly specific to certain sources. However, to date, SNV frequencies have not been leveraged for source tracking despite their success with strain tracking, in which individual strains per species rather than contributions from whole communities are tracked. We assess the ability of SNVs versus species in a previously designed source tracking algorithm FEAST (Shenhav et al., 2019) and find that SNVs can more accurately identify sources and their contributions. With SNV source tracking, we recapitulate previous findings that transmissions from mothers to their infants decreases with the age of the infant and that the built environment of NICUs play an important role in seeding infant microbiomes. Additionally, with SNV source tracking, we track migration of microbes across oceanic regions, including across the Suez and Panama canals, and observe a distance-decay relationship in the source contribution, which we do not observe with species source tracking. In sum, source tracking with SNVs can offer new insights into microbiome transmission and colonization sources that species cannot.

Introduction

Understanding how microbiomes are formed has important implications for human and environmental health, such as determining the impact of a hospital environment on the early infant gut microbiome, and in quantifying the extent of exchange of microorganisms over large distances and long spans of time. Approaches for determining the proportion of a microbiome of interest (the “sink”) that is attributed to different microbiomes (the “sources”) is known as “source tracking” (Knights et al., 2011; Shenhav et al., 2019). Source tracking is useful for forensics, categorization of samples, and tracing transmissions between different hosts or environments.

Current approaches for quantifying the microbiome as a mixture of other source microbiomes include SourceTracker (Knights et al., 2011) and more recently FEAST (Shenhav et al., 2019). These source tracking methods are designed to use the species abundance profile of the sample of interest (the sink) and of putative sources, and compute percentages of the microbiome of interest that are traced to each putative source. Microbiome source tracking is analogous to estimation of human admixture, which seeks to quantify the proportion of a person’s genome that is attributable to different ancestries (Alexander et al., 2009; Chiu et al., 2022).

16S amplicon sequencing data, which is used to determine abundance of species, has been an appealing datatype to use with current source tracking methods (Knights et al., 2011; Shenhav et al., 2019) because of the low cost of sequencing and the availability of this data in public repositories. However, 16S data is often limited to providing abundance information at the species-level, but rarely at the sub-species level (Callahan et al., 2016). By contrast, species are comprised of thousands of single nucleotide variants (SNVs) and hosts are frequently colonized by their own genetically distinct set of strains (Schloissnig et al., 2013), making SNVs an appealing source of high-resolution information about transmission patterns. While whole metagenomic sequencing data has been used to quantify species counts for purposes of estimating source contributions (McGhee et al., 2020), single nucleotide variants (SNVs) have not been leveraged to date.

Previous studies have used SNVs to determine sharing of strains for tracking transmission between hosts. For example, (Nayfach et al., 2016) quantified vertical transmissions from mother to infant by tracking the sharing of SNVs private to mothers and their infants. They found that private SNV sharing decreases over the first year of life while species sharing increases. This suggests that while the infant microbiome increasingly resembles the adult microbiome ecologically, sources other than the mother also colonize the infant. Thus, species-level resolution may obscure true underlying ecological dynamics and true sources of microbes while SNVs more adequately represent actual transmission from sources to the infant. Other studies have also used private SNVs to detect transmission of strains between hosts (Korpela et al., 2018; Li et al., 2016; Schmidt et al., 2019). Additionally, many studies have inferred strain haplotypes to track transmission (Asnicar et al., 2017a; Brooks et al., 2017; D. W. Chen & Garud, 2021; Enav & Ley, 2021; Ferretti et al., 2018; Hildebrand et al., 2021; Mitchell et al., 2020; Olm et al., 2021; Yassour et al., 2018).

Current approaches to strain tracking are limited because they do not provide a quantity of source contributions and instead provide a binarization of whether a strain transmission occurred per species. Some of these studies quantify strain sharing as a percentage but only between the host of interest and one source of interest at a time (Asnicar et al., 2017a; Ferretti et al., 2018; Nayfach et al., 2016; Olm et al., 2021). By contrast, with source tracking, the proportions for multiple sources contributing to a given sink (e.g. 25% from mother, 10% from dog, 30% from unknown, etc), integrated over all community members in the sink, can be inferred simultaneously (Knights et al., 2011; Shenhav et al., 2019).

Additionally, most source tracking studies have focused on human systems where transmission of strains have occurred in more recent time scales. However, it is less clear how these methods perform in systems where strain migrations may have occurred in the more distant past, such as across different oceans. A study on travel times and mixing in the ocean using satellite-tracked surface drifting buoys found that drifters in the Southern Ocean could take up to 13 years to travel to the Mediterranean (Laso-Jadart et al., 2021). Another study found that travel times between different oceanic regions could be over two decades (O'Malley et al., 2021). A benefit of using SNVs in the ocean microbiome is that SNVs can track fragments of DNA that have moved due to horizontal gene transfer in the distant past rather than relying on inference of whole genomes or presence of private SNVs that may been transmitted in the recent past. This global-level source tracking is analogous to admixture estimation with human genotypes.

Here, we evaluate the ability of FEAST with SNVs (SNV-FEAST) to accurately estimate true sources in simulated mixtures and to detect trends in source estimates along time and distances, and compare this performance to FEAST with species abundance profiles (species-FEAST). We show that SNVs can be used directly for source tracking, allowing us to estimate the percentage of the sink microbiome explained by different sources without having to identify discrete taxonomic units. FEAST is faster and more accurate than previous source tracking tools (Knights et al., 2011; Shenhav et al., 2019), and therefore, is ideal for adaptation to SNV source tracking. Because there are potentially millions of polymorphic sites of interest across all present species, we introduce a method within SNV-FEAST to determine informative SNVs to use as input into the original FEAST algorithm. This both reduces memory requirements and computation time in the FEAST estimation, allowing us to optimally estimate the source contribution of a sink. We find that SNV-FEAST and species-FEAST yield different outcomes when applied to fecal samples from infants in the first year of life, fecal samples from infants in the neonatal intensive care unit (NICU), and water samples from world oceans. We show use of genetic variants to trace migration across oceanic regions, particularly across the Suez Canal and across the Panama Canal and find a distance decay-relationship between source and sink with SNVs but not species. In sum, we show that SNVs can be used to estimate transmission across hosts and across environments.

Results

FEAST algorithm

The goal of source tracking is to estimate the contribution of various sources to a sink. This requires defining a probabilistic model for inferring mixture proportions for both known and unknown sources. Current methods estimate source contributions of sinks using species abundance profiles from a set of potential source microbiomes. SourceTracker (Knights et al., 2011) estimates these contributions using a Bayesian approach with Gibbs sampling to identify sources and their proportions using species counts for the sources and sink. Shenhav et al. later introduced an expectation maximization algorithm, FEAST (Shenhav et al., 2019). FEAST models the species read counts in the sink as a mixture of multinomial distributions that represent the sources. The inferred mixture parameters are the relative contributions of sources to the sink. FEAST is both faster and more accurate than Source Tracker (Knights et al., 2011), enabling the use of larger feature sets and a larger number of input sources.

The intuition behind the estimation process is that sources with similar species distribution to the sink would have higher estimated contributions to the sink. For example, a source and a sink harboring a species private to these two samples would increase the estimated contribution of that source. However, in many cases, the same species are found in multiple sources simultaneously. The algorithm does not uniquely assign each species to a source, but rather, utilizes all the species together to infer the source contributions.

SNV-FEAST

FEAST was originally tested and evaluated with species data but not with SNVs. Yet, SNVs can potentially provide higher resolution information. Our objective here is to assess whether source contributions estimated with SNVs are more accurate than with species by utilizing two approaches we designate as SNV-FEAST and species-FEAST (Shenhav et al., 2019), respectively.

A computational challenge, though, is that the number of different species comprising a microbiome can range from a few hundred to a few thousand, while the number of possible SNVs for a given species alone can be in the thousands (Schloissnig et al., 2013). This can result in runtimes that last several hours instead of a few minutes. Additionally, more features do not necessarily increase accuracy as the same FEAST estimates could potentially be obtained with much less data. Thus, to reduce the number of SNVs for source tracking with FEAST, we used a likelihood approach to define a 'signature score' for each SNV (see **Methods**). A signature score quantifies the extent to which SNVs in the sink that are most likely derived from one of the sources. This is analogous to identifying SNVs private to sources and their sinks, but more generalized to include SNVs that may be found in multiple sources, albeit at high frequency in one of the sources (see **Methods**).

Simulations

To compare the accuracy of species-FEAST and SNV-FEAST, we performed simulations mimicking mother-infant transmissions with the goal of estimating contributions of different sources to an infant sink. Concretely, we tested the ability of SNVs and species to recapitulate the true source composition in synthetic samples comprised of a mixture of reads drawn from multiple real fecal adult samples. To construct these synthetic infant microbiomes, we leveraged the metagenomic data from mothers sampled in a mother-infant dataset (Bäckhed et al., 2015).

The difficulty of source tracking increases with the number of contributing sources (Shenhav et al., 2019). Thus, we simulate infants that have a low (<=4) versus high (5 – 10) number of contributing sources with different proportions ranging from 5 - 90% (**Supplementary Table 1**). A single unknown source (e.g. a randomly selected unrelated mother) was also selected to contribute to the simulated infant. However, this unknown source was not presented to FEAST as a potential known source.

Additionally, not all species in a mother are transmitted to the infant (Asnicar et al., 2017b; Ferretti et al., 2018; Korpela et al., 2018; Sprockett et al., 2020; Yassour et al., 2018). Thus, in our simulations, species transmission rates were determined using a beta distribution, which is a natural model for values between (0,1) and often proposed for microbial abundance data (E. Z. Chen & Li, 2016; Martin et al., 2020; Sloan et al., 2006, 2007) (see **Methods**). We thus consider four simulated scenarios: a combination of low versus high number of sources and low versus high transmission rates (see **Methods**).

In **Figure 1**, our results show the performance of SNV-FEAST and species-FEAST in estimating the true contribution of sources. Generally, SNVs outperform species in most scenarios, especially when transmission rates are low. SNVs have a lower root mean squared error (RMSE) compared to species in three of the four scenarios and higher Spearman correlation between true and estimated contributions in all four scenarios. The difference in these correlations for SNVs versus species is significant in an unpaired Wilcoxon rank sum test (p = 0.01429, but p = 0.06 when paired test is used). These results suggest that SNVs may offer useful signatures of transmission.

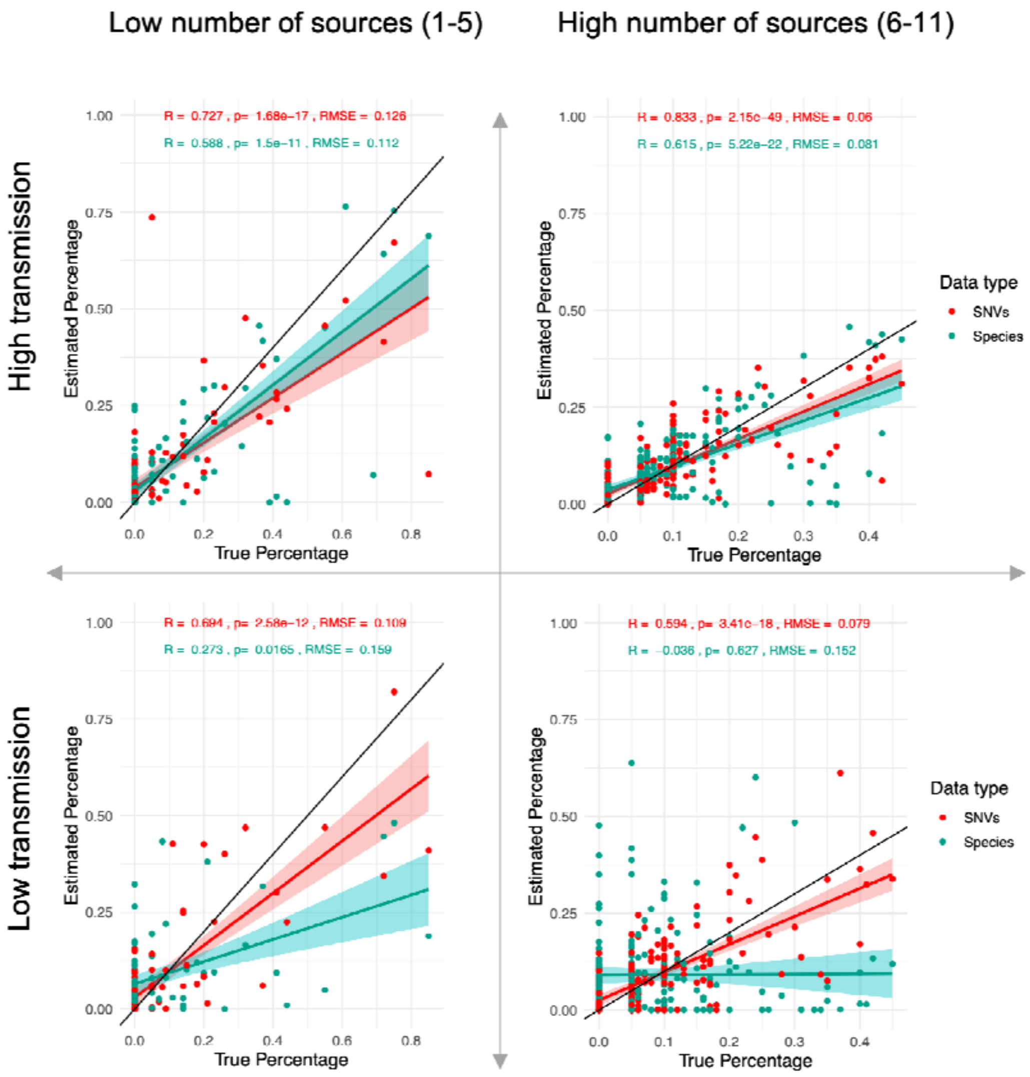

Figure 1: Ability of SNV and species-FEAST to recapitulate true contributions in simulations. Estimated known and unknown source proportions for infant microbiomes simulated with in silico mixtures of real maternal fecal microbiomes under different scenarios: either low number of contributing sources (1-4) or high number of sources (5-10), and high transmission rate of species or low transmission rate. Transmission rate is the probability of an infant being colonized by a given species, simulated using a beta distribution centered on the relative abundance of species in sources (**Methods**). Ten infants were simulated with low number of sources and 18 infants were simulated with high number of sources (**Table S1**). The black line indicates the ground truth for proportions. For each simulated infant, there are 11

points plotted, whereby 10 correspond to known sources (some of which have zero contribution), and one corresponds to an unknown source.

Source tracking in infants over the first year of life

We estimated the contribution from the true mother over time to the infant with SNV and species-FEAST. We once again analyzed the Backhed et al. 2015 dataset, composed of metagenomic samples from infants collected at four days, four months, and 12 months after birth, as well as their mothers at the time of delivery. Previous analyses on this data have shown that infants and their mothers share fewer proportions of strains over time, even while species similarity increases (Nayfach et al., 2016). Thus, SNVs belonging to strains shared only by the infant and their mother may be more revealing of the true source compared to species.

When applying FEAST with species, the input sources included samples from the mother at birth, three randomly selected unrelated mothers, as well as samples from previous time points when applicable (sample at birth when sink is four months, samples at birth and four months when sink is 12 months). We utilized all species present in the infant whereas SNV-FEAST used signature SNVs from only a subset of species (mean for 4 days, 4 months, 1 year). Shown in **Figure S1** are the distribution of species included in species and SNV FEAST.

We estimated the contribution of the mother to the infant over the first year of life with species and SNV-FEAST (**Figure 2).** Consistent with previous findings made with species and SNV (Nayfach et al., 2016), species-FEAST estimates an increasing contribution of the mother over time ($p = 5.1 \times 10^{-4}$), but SNV-FEAST estimates a decrease over time (p = 0.063).

We also assessed the ability of species and SNV-FEAST to distinguish the true mother from three randomly selected unrelated mothers. We find that species-FEAST estimates an increasing contribution of unrelated mothers over time (p= 0.014) while SNV-FEAST estimates no significant change over time (p = 0.59) (**Figure 2**).

We also estimated contributions from unknown sources, i.e. the portion of the infant microbiome not explainable by the true mother or the three randomly selected unrelated mothers. Interestingly, species-FEAST estimates a sharp decline in contribution of unknown sources ($p=7.1 \times 10^{-12}$) (**Figure 2**), whereas SNV-FEAST estimates little change in the contribution of unknown (p = 0.49) (**Figure 2**).

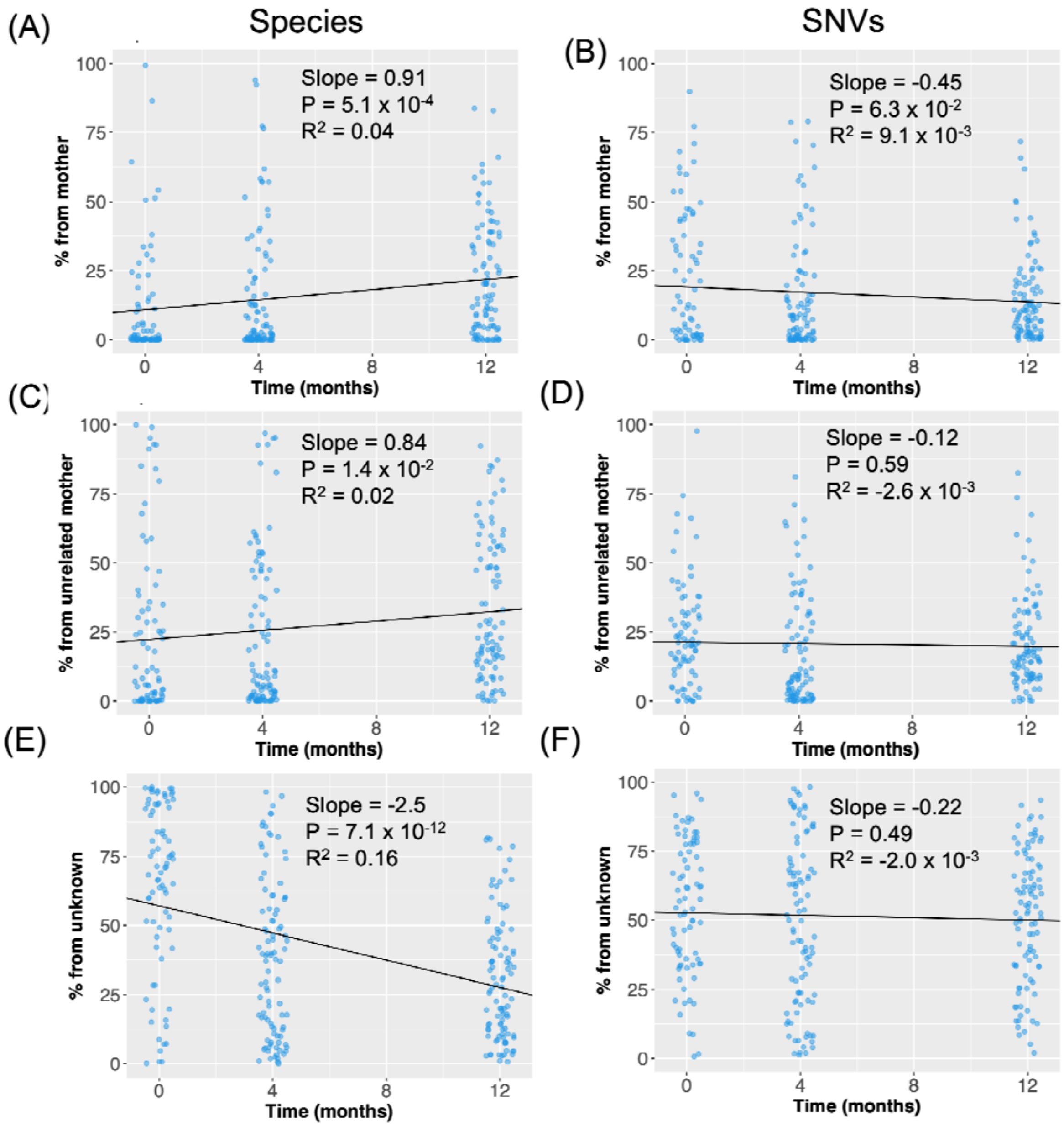

Figure 2. Source tracking in the infant gut microbiome over the first year of life. For infants at birth, four months, and twelve months, we utilize species-FEAST and SNV-FEAST to estimate the contribution of (A, B) mother, (C, D) unrelated mothers and (E, F) unknown sources. The black line and inset statistics pertain to the linear regression fit for the source estimates as a function of age of the infant. **Figure S1A** shows the species that were included in species-FEAST and **Figure S1B** shows the species for which SNVs were included in SNV-FEAST.

Contribution of the NICU built environment to infant microbiomes

Brooks et al. 2017 studied the contribution of the hospital environment to the infant gut microbiome in the neonatal intensive care unit (NICU). They sampled the microbiomes of infant stool, as well as their rooms at frequently touched surfaces, sink basins, the floor, and isolette-top (Brooks et al., 2017) over an 11-month period. We re-analyzed this data with SNV and species-FEAST to assess the contribution of the infant's own NICU room as well as a different NICU room (**see Methods**) in the vicinity, to estimate possible transmissions across rooms.

Concordant with the findings of Brooks et al., both SNV and species-FEAST detected that the most common source contributing to the infant microbiome was the floor and isolette-top from the infant's own room. SNV-FEAST found Infant 18 also had large contributions from their own room's touched surfaces at multiple time points (**Figure 3**), which may be explained by a finding by Brooks et al. that three strains found in Infant 18 perfectly matched (> 99.999% average nucleotide identity) strains found in the touched surfaces samples of Infant 18's own room. Lastly, we found Infant 6's microbiome was explained almost entirely by samples from a different room including a sizeable contribution from the sink basin in this different room. This is concordant with Brooks et al. finding of multiple cases of strain sharing across rooms of Infant 6 and 12 for the different surfaces. SNV-FEAST was able to quantify the extent to which Infant 6's microbiome was influenced by strains present in the built environment.

Through application of SNV and species-FEAST, we are able to quantify any trends over time in the influence of the built environment on the infant microbiome. For example, both SNV and species-FEAST estimated a large unknown component for all four infants, with Infant 18 showing the largest mean unknown component across the NICU stay based on SNVs. This unknown component is important because it signifies the extent to which other sources such as the mother and diet are impacting infant gut colonization. In assessing the known sources, we found that SNV-FEAST shows more consistency in the contribution from different sources compared to species-FEAST over multiple time points.

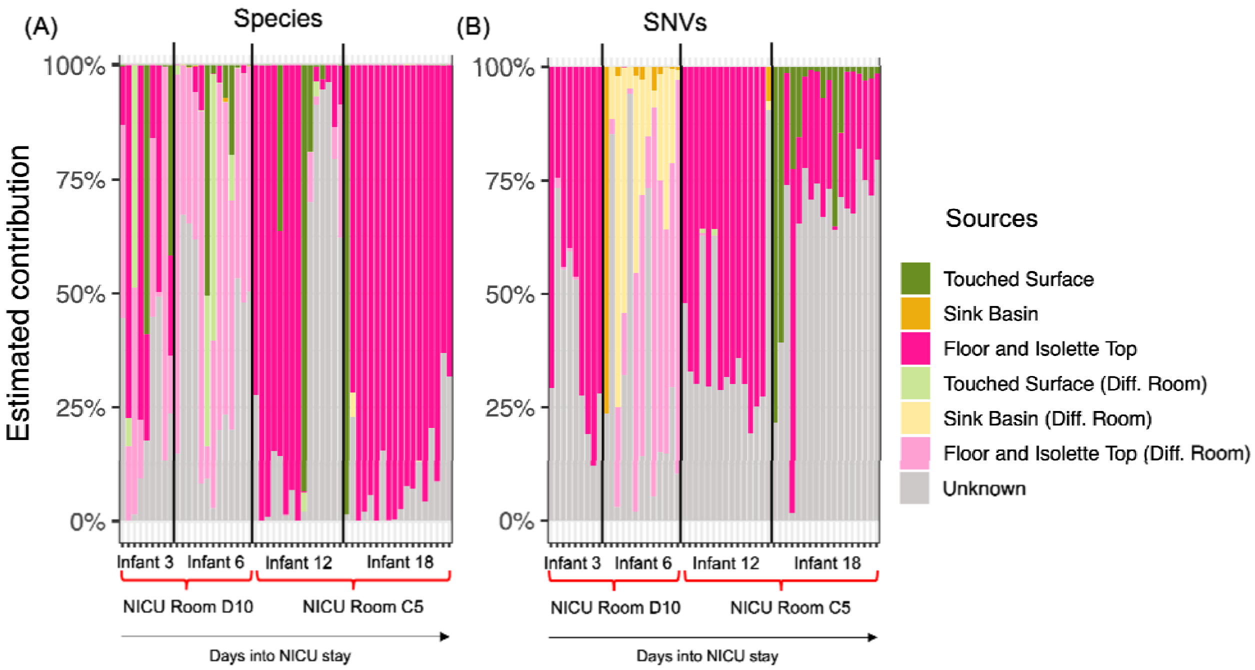

Figure 3: SNVs estimate more diverse sources of infant microbiomes in the NICU. Each bar represents one sampling day in the NICU stay of an infant. Infants 3 and 6 as well as Infants 12 and 18 stayed in the same NICU room at different times, respectively. The contribution of a different room are determined by using samples from Infant 12's room for Infants 3 and 6, and samples from Infants 6's room for Infants 12 and 18 for each of the categories of surfaces per infant: touched surface, sink basin, or floor and isolette top surface.

Global source tracking of ocean microbiomes

The ocean microbiome is a complex community that displays biogeography at the species and functional levels (Nayfach et al., 2016; Sunagawa et al., 2015). To further understand global migration patterns of ocean microbiomes, we applied SNV-FEAST and species-FEAST to the Tara Oceans microbiome dataset (Sunagawa et al., 2015). Tara Oceans is composed of 182 whole metagenomic sequencing samples derived from 64 stations at multiple depths. Previous research indicates that temperature is one of the highest drivers of variability in microbial composition in the ocean ((Ladau et al., 2013; Sunagawa et al., 2015). For this reason, we restricted the source tracking analysis to sinks and sources from the same temperature range:

above 20 degrees Celsius. We additionally focused on samples from the surface water layer at an average of five meters below the surface.

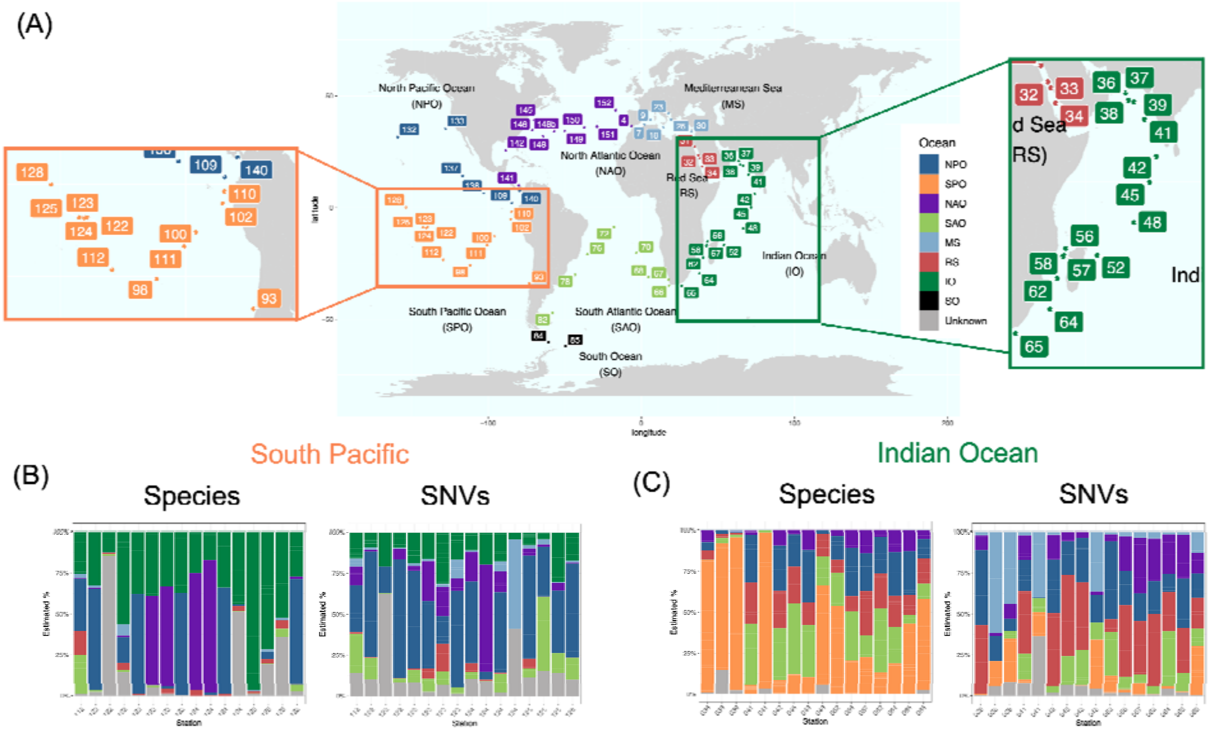

Figure 4. Microbial source tracking in the Tara Oceans dataset with SNV and species-FEAST. World map indicating the location of sampling sites (A). Source tracking estimates for the contribution of different oceans to the South Pacific (n=16) (B) and Indian Oceans (n=16) (C) are depicted with vertical bars. In each experiment, all stations around the world excluding those from the “sink” ocean are considered potential sources. Light blue, for example, represents the total contribution of four stations from the Mediterranean Sea.

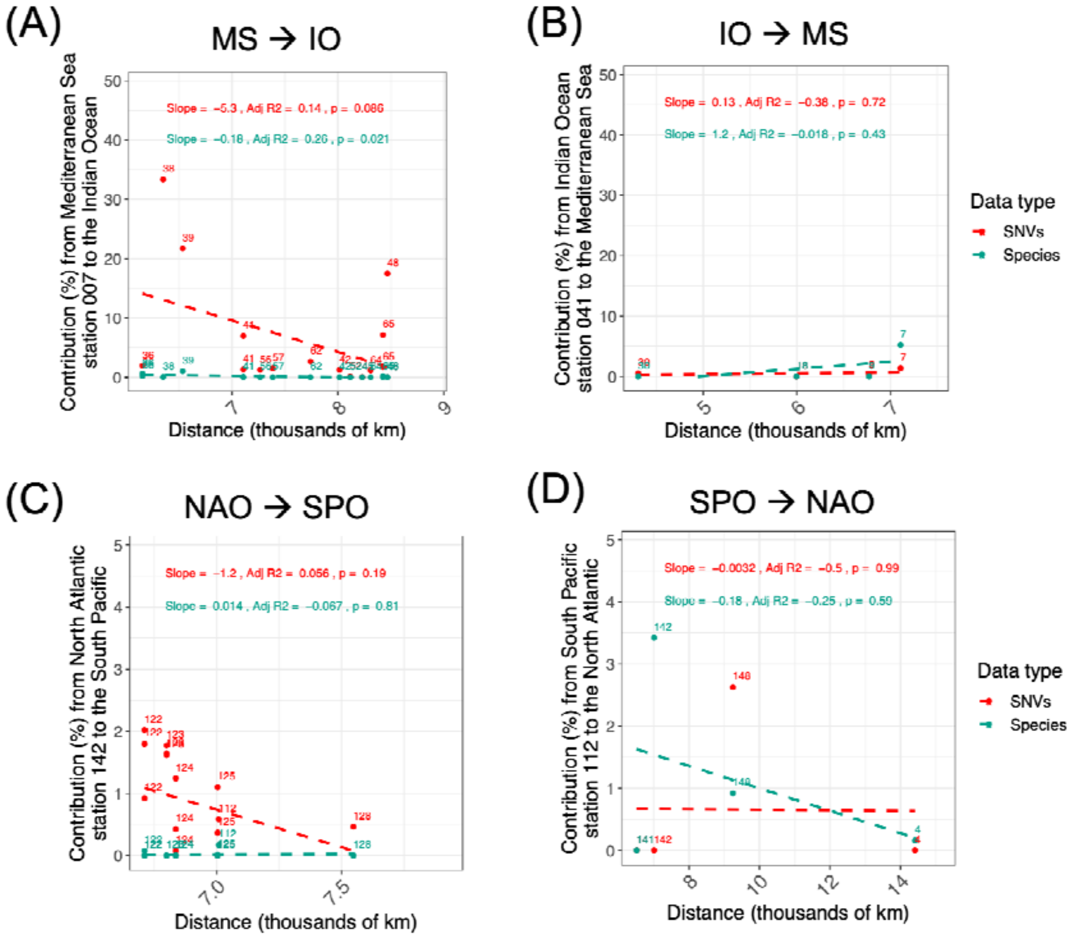

Figure 5. Distance decay in contribution of a "source" ocean to a "sink" ocean. (A) Estimated contribution of Mediterranean station 7 to various Indian Ocean samples as a function of geographic distance. (B) Estimated contribution of Indian Ocean station 41 to various Mediterranean samples as a function of geographic distance. (C) Estimated contribution of North Atlantic station 142 to various South Pacific samples as a function of geographic distance. (D) Estimated contribution of South Pacific station 112 to various North Atlantic samples as a function of geographic distance. Inlaid text shows the linear fit for contribution as a function of distance (slope, adjusted R-squared, and p-value for slope).

First, we performed source tracking between oceans using SNV and species-FEAST. We treated each station around the world as a sink and estimated the contribution of different oceans around the world to that sink (Methods). Unknown represents any portion of the microbiome in these sink samples that cannot be explained by any of the provided source samples. We found that species and SNV-FEAST estimate different amounts of sharing between oceans, where SNVs estimate a higher unknown on average. The finding that SNV-FEAST estimates a higher unknown contribution on average is most evident in the North Pacific, North Atlantic, South Atlantic, and Mediterranean oceans (**Figure S1**). Additionally, in some oceans, SNVs identify contributions from oceans that species-FEAST does not detect (**Figure 5, Figure S1**). For example, in applying FEAST to Indian Ocean samples we find that there is measurable sharing of microbes with the Mediterranean Sea, but this is not detected with species (**Figure 4C**). Such differences were found in samples from other oceans as well (**Figure S1**).

Second, we assessed whether source tracking estimates display a distance-decay relationship. Previous studies found that genetic distance, such as that represented by fixation index F_{ST}, increases with geographic distance between populations (Cavalli-Sforza & Feldman, 2003; DeGiorgio & Rosenberg, 2013). Based on these findings, our expectation was that samples that are further away from a given station will have a lower contribution from that station.

To assess this distance-decay relationship, we focused on sharing between the Indian Ocean and Mediterranean Sea, which are connected by the man-made Suez Canal. Migration from the Red Sea to the Mediterranean Sea, known as Lessepsian migration, is well-documented for not only microorganisms but also macroorganisms like fish (Bentur et al., 2008; Bianchi & Morri, 2003; Golani, 2009). Additionally, recent studies suggest that anti-Lessepsian migration of bacteria (Mediterranean to Red Sea) may be more common than Lessepsian migration (Elsaeed et al., 2021). We hypothesized that the source estimates for the Indian Ocean will show a large contribution from Mediterranean-derived microbes detected by SNV source tracking, consistent with anti-Lessepsian migration through the Red Sea.

We assessed the changing contribution of Mediterranean-derived microbes in the SNV and species profiles to the Indian Ocean with distance (**Figure 5A**). We also assessed contributions in the opposite direction from the Indian Ocean to the Mediterranean Sea (**Figure 5B**). Only in the Mediterranean to Indian Ocean direction (**Figure 5A)** do we observe a distance decay relationship (p= 0.72 for SNV-FEAST and p=0.43 for species-FEAST) for the estimated contribution of a sample from Mediterranean station 7 to Indian Ocean samples. By contrast, the estimated contribution of a sample from Indian Ocean station 41 to Mediterranean samples shows little change over distance for both SNV and species-FEAST. Despite there being a distance decay relationship detected with both species and SNVs from the Mediterranean to IO, the contribution from species is overall very low (0.04% mean contribution across all Mediterranean stations, 1% max) compared to that of SNVs (3.4% mean across all Mediterranean stations, 33.4% max).

Additionally, we examined distance-decay relationships between the North Atlantic and South Pacific, which are connected by the man-made Panama Canal. Migrations are more commonly recorded along the Suez compared to the Panama Canal. It has been previously suggested that the low salinity of the Panama Canal waters could be a barrier to migration of organisms across the canal (Menzies, 1968). However, migrations of certain microogranisms have been detected in either direction (Carlton et al., 2011).

Figure 5C shows the estimated contribution of a sample from North Atlantic station 142 to South Pacific samples. SNV-FEAST finds a distance decay relationship between a North Atlantic sample and South Pacific stations (p = -0.19) while species-FEAST does not (p = 0.81). In **Figure 4D**, we show the estimated contribution of a sample from South Pacific station 112 to samples in the North Atlantic. For both SNV and species-FEAST, the mean contribution of North Atlantic to South Pacific is higher than South Pacific to North Atlantic with a mean estimated contribution of North Atlantic samples to the South Pacific of 3.2 % (SNV-FEAST) and 5.6 % (species-FEAST) and mean contribution of South Pacific to North Atlantic of 0.26 % (SNV-FEAST) and 2.3 % (species-FEAST).

Discussion

Source tracking provides insight into source contributions to a metagenomic sample as well as similarities between metagenomic samples. While species abundances have been informative in source tracking in several studies (Flores et al., 2011; Knights et al., 2011; McGhee et al., 2020; Shenhav et al., 2019), they may be limited in their resolution. SNVs provide a potential alternative because of their ability to distinguish sources of strain transmissions. Here we compared the ability of a previously published source tracking algorithm FEAST using species versus SNVs as input data. In application of species and SNV-FEAST to three case studies, we confirm that SNVs indeed can provide insight into the ecological processes shaping microbial communities that species information alone cannot.

In the first case study, we confirm previous findings that SNV sharing between mothers and infants decreases over the first year of life while species sharing increases (Nayfach et al., 2016), suggesting that while the infant microbiome matures to resemble adults, sources other than the mother may seed the infant over time. In the second case study, we confirmed source contributions from the NICU built environment to the infant microbiome (Brooks et al., 2017), and found that SNVs detect a more consistent estimate in source contributions overtime compared to species as well as detecting contribution from sources not detect by species-FEAST. Finally, in the third case study in the TARA oceans dataset, we found SNVs but not species display a distance decay relationship, paralleling recent results made with gene content (Dlugosch et al., 2022). This last case study provides novel insight into sharing of microbiomes across oceans. While previous studies have examined the biogeography of the ocean using species profiles, genes (Dlugosch et al., 2022; Nayfach et al., 2016) or amino acid variants from a single species (SAR11) (Delmont et al., 2019), for the first time, we leverage the use of SNVs across all detected prevalent species in the ocean microbiome to identify proportions of sharing across oceans.

Several previous studies have relied on tracking transmissions of strains with private SNVs shared only between the sink and putative source (Bäckhed et al., 2015; Korpela et al., 2018; Nayfach et al., 2016; Schmidt et al., 2019). While this has been an effective way to track transmissions, such analyses are restricted to a binary quantification of sharing or not sharing for each species. We instead used any SNV with an informative distribution across sources as determined by our signature scoring method (see **Methods**) and are able to quantify the relative contribution of all the sampled environments. Additionally, with source tracking, we can quantify the contribution of unknown sources. For example, with SNV FEAST applied to ocean samples, we found an overall higher proportion attributable to "unknown" sources compared to findings made with species FEAST. This unknown component suggests that a significant fraction of marine biodiversity may be endemic, as previously noted in the Mediterranean (Katsanevakis et al., 2014).

Another popular approach used to track strain transmissions is to resolve haplotypes and then identify matches with high average nucleotide identity (ANI). However, this approach may miss strain sharing of lower-abundance strains whose haplotypes cannot be confidently resolved. For example Brooks et al.'s study of strain transmission was restricted to only strains whose presence in a sample was defined with an ANI > 99.999% and genome breadth > 90%. As a result, many potentially informative strains may not have been considered. The benefit conferred by using SNVs however, is that we bypass computationally intensive phasing of haplotypes and maximize use of the data. Moreover, with strict requirements of high ANI between samples, transmissions of fragments of DNA arising from recombination may not be detected. However, SNV tracking would potentially reveal such transmission events. This feature of SNV tracking resembles work in human genetics estimating admixture proportions by tracking transmission of genetic fragments across generations (Alexander et al., 2009; Chiu et al., 2022).Thus, tracking SNV frequencies may be important in detecting genetic heterogeneity over long spans of time, especially when gene-specific sweeps or movement of mobile genetic elements could be important contributors to gene flow (Bendall et al., 2016; Reveillaud et al., 2019).

A drawback, however, with using SNVs over species is deeper, whole genome sequencing is required to accurately call SNVs. Moreover, even when there is sufficient coverage, there is still the challenge of a large number of SNVs. We demonstrate one way to subset SNVs that uses a scoring method for informativeness, but there may yet be other methods for filtering SNVs to the most informative set. Another potential caveat of SNV filtering is that not all species present will be represented in the final informative SNV set (**Figure S1**). However, we show that not all species need to contribute informative SNVs in order to make accurate inferences.

Ascertainment of SNVs from metagenomic data in a high-throughput manner, especially common SNVs with microbiome genotyping technology (Shi et al., 2021), is becoming an increasing priority for the field as metagenomic datasets become more abundant. A genotyper for prokaryotes has already been developed and tested on a catalog of over 100 million SNVs in order to characterize population structure (Shi et al., 2021). Such a catalog of informative SNVs could be invaluable for source tracking. With source tracking enabling us to characterize samples by their relationship to known samples, we have a powerful tool to explore samples in new contexts we have yet to discover.

Methods

Data

For simulations and analyses of infant microbiomes in the first year of life (Bäckhed et al., 2015), we downloaded the raw shotgun metagenomic sequencing reads from public read archives under accession number PRJEB6456. We downloaded the raw sequence reads for the NICU analysis (Brooks et al., 2017) from accession number PRJEB323631, and the equivalent for the Tara Oceans analyses (Sunagawa et al., 2015) were downloaded from accession number PRJEB402.

Estimation of species 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 species abundance and SNV content per species in each metagenomic shotgun sequencing sample. The database we used to apply MIDAS consisted of 31,007 bacterial genomes that are clustered into 5,952 species. The parameters we used to estimate species abundances and SNVs were described in (Garud et al., 2019). A species was considered present if there are at least 3 reads mapping to a set of single copy marker genes on average. To call SNVs, we used the default MIDAS settings in order to map reads to a single representative reference genome. The mapping was done with Bowtie 2 (Langmead & Salzberg, 2012): global alignment, MAPID≥94.0%, READQ≥20, ALN_COV≥0.75, and MAPQ≥20, where species with reads mapped to less than 40% of the genome were excluded from the SNV calls.

Application of FEAST algorithm

FEAST, originally introduced by Shenhav et al., is an R-based method that models the mixture proportions for various “source” microbial samples for a given “sink” (Shenhav et al., 2019). This method utilizes Expectation Maximization to estimate the proportions when given any sort of count-based feature matrix representing the sources and sinks. The intuition behind the estimation process is that a source with a similar species distribution to the sink would have a higher contribution estimate to the sink. A species with non-zero counts only in source *j* and the sink would increase the estimated contribution of source *j*. However, in many cases, the same species are found in multiple sources simultaneous. The algorithm does not uniquely assign a species to a source but rather simultaneously utilizes all species information to infer the source contributions. The method was originally tested and evaluated on species and not previously tested on more fine scale genetic data such as SNVs. The number of different species, on average, range in number from a few hundred to a few thousand, while the number of possible nucleotide sites that vary across different sources can number in millions. For this reason, a SNV-filtering process is necessary so that the algorithm can run within a reasonable time and with reasonable memory requirements.

For both species and SNV-FEAST, the same set of sources and sinks were fed into the FEAST algorithm. In the case study of infants in the first year of life (Bäckhed et al., 2015), the sink consisted of the infant fecal sample at either four days, four months, or 12 months and the sources consisted of fecal samples from the true mother, three randomly selected mothers from the same dataset, and also any previous time points for the infant. For the case study of infants in the NICU (Brooks et al., 2017), the sink consisted of the fecal sample of the infant at a given time point and the sources consisted of pooled reads from the touched surfaces, the sink basin and the floor and isolette top from both the infant's own room as well as a different room. The different room was Infant 12's room for Infants 3 and 6, Infants 6's room for Infants 12 and 18. For the Tara Ocean (Sunagawa et al., 2015) samples, the sink consisted of the surface water sample from the ocean station of interest while the sources consisted of surface water samples from every other station from every other ocean in the world.

To obtain single nucleotide variation, we applied MIDAS (Nayfach et al., 2016) at the "species" and "snps" step to the publicly available fastq files provided by each publication. The merge_midas.py script was applied to process the final "snps" output with the following parameters: sample depth 5, site depth 3, min samples 1, site prevalence 0 and threads 7.

Determining the signature SNV set

To find the signature SNVs, the following steps are followed:

(1) Minimum coverage filtering: only sites of the genome with at least the required number of reads mapping to the site are considered. In the case study of infants in the first year of life (Bäckhed et al., 2015) and infants in the NICU (Brooks et al., 2017), the minimum coverage requirement is 10 across the sink and *J* sources. For the Tara Ocean (Sunagawa et al., 2015) samples, the minimum coverage is five reads (Sunagawa et al., 2015). Additionally, sites that are biallelic must have more than one read mapped to each allele to be considered.

(2) Signature score calculation: Two hypotheses are compared for each SNV: (1) only source *i* out of *J* possible sources explains the observed sink's distribution of reference and alternative alleles in the reads and (2) all other sources except *i* (sources $j \neq i$) explain the observed distribution of reads in the sink. Given *n* reads with the reference allele and *m* reads with the alternative reads in the sink and learned parameter θ for the reference allele frequency, the binomial log-likelihood for the observed read count distribution is calculated as follows:

$$l(\theta) = n\log p + m\log(1-p)$$

Allele frequency θ for the sink is learned using one of two hypotheses:

Hypothesis 1: Source *i* explains the allele counts in the sink

$$\theta = \theta_i$$

where θ_i is the allele frequency in source *i*.

Hypothesis 2: The combination of all sources except source *i* explain the allele counts in the sink.

$$\theta = \sum_{j \neq i}^{J} \alpha_j \theta_j$$

where θ_j is the allele frequency in source *j* and α_j is the mixing parameter representing how much of source *j* explains the allele frequency in the sink. The *j* length vector α is learned by applying a solver implemented in scipy.minimize() with Sequential Least Squares Programming and subject to the constraint of summing to 1 with bounds of 0 to 1 inclusive.

A likelihood ratio is calculated as $l_1(\theta) - l_2(\theta)$ representing the ratio of the likelihood of hypothesis 1 to hypothesis 2 for each source *j* of interest such that there are *J* likelihood ratios. The maximum of the likelihood ratios calculated for all *J* sources is saved as the one signature score for that SNV, representing how favorably one source explains the sink better than all other sources. All the scores are ranked across SNVs and SNVs with scores that are greater than two standard deviations over the mean signature score within each 200 kbp window of the genome are retained as signature SNVs. This window size was chosen for to optimize run time and memory requirements.

Note, if only one source passes minimum coverage filtering, $l_2(\theta) = 0$ resulting in a very high likelihood ratio as represented by $l_1(\theta)$ for the one source . These SNVs are more likely to pass the signature score filtering. One exception for SNVs that are included in the signature SNV set without passing signature score filtering are SNVs with an allele that is completely unique to the infant, as these represent SNVs that are potentially derived from an unknown source. Signature SNVs are obtained from the SNV profile of every species for which we have MIDAS output for.

Simulating mother to infant transmission

The mixture proportions for 28 simulated infants is shown in **Table S1**. Four possible scenarios are simulated using a combination of either low or high number of sources and low or high transmission probabilities of species. High transmission of species was simulated by drawing separate transmission probabilities for each species in each contributing source based on a beta distribution with a mean equal to the species relative abundance and variance equal to 0.1, a value selected to emulate Backhed et al.'s mean relative abundance and variance. For the low transmission scenario, transmission probabilities were drawn from a beta distribution with mean 0.1 times the relative abundance and variance at 0.1. To determine if a species from each source was transmitted to a given infant, a binomial draw was performed J times, where J = number of sources, and the probability of a mother transmitting the species is p_j based on the beta-drawn transmission probability. If any of the draws yields a one, that species is transmitted to the infant from all sources. The same simulated data under these scenarios is used for both SNV and species source tracking.

The source tracking estimates are compared to the true mixing proportions using Spearman correlation. The significance of correlation is calculated using the stat_cor function in the 'ggpubr' package (*CRAN - Package Ggpubr*, n.d.).

Distance Decay Analysis

To study the relationship between source tracking estimates and geographic distance, we selected a single station from the Red Sea and North Atlantic by which to compute the distance to stations in the ocean of interest, namely, the Indian Ocean and South Pacific. To compute geographic distance between stations, we applied the Haversine distance to the longitude and latitude of the sampling sites provided by (Sunagawa et al., 2015) using the package "geosphere" (Hijmans et al., 2021). Source tracking estimates were computed as described above using either SNV-FEAST or Species FEAST.

Acknowledgements

We sincerely thank Nicole Zeltser for processing the TARA Oceans data through the MIDAS pipeline. We sincerely thank the Garud lab for their feedback on the manuscript.
LB was supported by the NSF Graduate Research Fellowship Program grant numbers DGE-1650604 and DGE-2034835. N.R.G. received support from the Paul Allen Frontiers Group, and the Research Corporation for Science Advancement.

Code Availability

The scripts used to implement the analyses including signature scoring are available at https://github.com/garudlab/SNV-FEAST.

Data Availability

All metagenomic data was obtained from public repositories. The applicable accessions numbers are PRJEB6456 for Backhed et al. 2015 (mother-infant), PRJEB323631 for Brooks et al. 2017 (NICU), and PRJEB402 for Sunagawa et al. 2015 (Tara Oceans).

References

Alexander, D. H., Novembre, J., & Lange, K. (2009). Fast model-based estimation of ancestry in unrelated individuals. *Genome Research*, *19*(9), 1655–1664. https://doi.org/10.1101/GR.094052.109

Asnicar, F., Manara, S., Zolfo, M., Truong, D. T., Scholz, M., Armanini, F., Ferretti, P., Gorfer, V., Pedrotti, A., Tett, A., & Segata, N. (2017a). Studying Vertical Microbiome Transmission from Mothers to Infants by Strain-Level Metagenomic Profiling. *MSystems*, *2*(1). https://doi.org/10.1128/msystems.00164-16

Asnicar, F., Manara, S., Zolfo, M., Truong, D. T., Scholz, M., Armanini, F., Ferretti, P., Gorfer, V., Pedrotti, A., Tett, A., & Segata, N. (2017b). Studying Vertical Microbiome

Transmission from Mothers to Infants by Strain-Level Metagenomic Profiling. *MSystems*, *2*(1). https://doi.org/10.1128/MSYSTEMS.00164-16/ASSET/54C4C531-C6DB-421B-8C8A-10C0ECFE3BE9/ASSETS/GRAPHIC/SYS0011720800004.JPEG

Bäckhed, F., Roswall, J., Peng, Y., Feng, Q., Jia, H., Kovatcheva-Datchary, P., Li, Y., Xia, Y., Xie, H., Zhong, H., Khan, M. T., Zhang, J., Li, J., Xiao, L., Al-Aama, J., Zhang, D., Lee, Y. S., Kotowska, D., Colding, C., … Jun, W. (2015). Dynamics and stabilization of the human gut microbiome during the first year of life. *Cell Host and Microbe*, *17*(5), 690–703. https://doi.org/10.1016/j.chom.2015.04.004

Bendall, M. L., Stevens, S. L. R., Chan, L. K., Malfatti, S., Schwientek, P., Tremblay, J., Schackwitz, W., Martin, J., Pati, A., Bushnell, B., Froula, J., Kang, D., Tringe, S. G., Bertilsson, S., Moran, M. A., Shade, A., Newton, R. J., McMahon, K. D., & Malmstrom, R. R. (2016). Genome-wide selective sweeps and gene-specific sweeps in natural bacterial populations. *The ISME Journal 2016 10:7*, *10*(7), 1589–1601. https://doi.org/10.1038/ismej.2015.241

Bentur, Y., Ashkar, J., Lurie, Y., Levy, Y., Azzam, Z. S., Litmanovich, M., Golik, M., Gurevych, B., Golani, D., & Eisenman, A. (2008). Lessepsian migration and tetrodotoxin poisoning due to Lagocephalus sceleratus in the eastern Mediterranean. *Toxicon*, *52*(8), 964–968. https://doi.org/10.1016/J.TOXICON.2008.10.001

Bianchi, C. N., & Morri, C. (2003). Global sea warming and “tropicalization” of the Mediterranean Sea: biogeographic and ecological aspects. *Biogeographia – The Journal of Integrative Biogeography*, *24*(1). https://doi.org/10.21426/B6110129

Brooks, B., Olm, M. R., Firek, B. A., Baker, R., Thomas, B. C., Morowitz, M. J., & Banfield, J. F. (2017). Strain-resolved analysis of hospital rooms and infants reveals overlap between the human and room microbiome. *Nature Communications*, *8*(1), 1–7. https://doi.org/10.1038/s41467-017-02018-w

Callahan, B. J., McMurdie, P. J., Rosen, M. J., Han, A. W., Johnson, A. J. A., & Holmes, S. P. (2016). DADA2: High-resolution sample inference from Illumina amplicon data. *Nature Methods 2016 13:7*, *13*(7), 581–583. https://doi.org/10.1038/nmeth.3869

Carlton, J. T., Newman, W. A., Pitombo, F. B., Carlton, J. T., Newman, W. A., & Pitombo, F. B. (2011). Barnacle Invasions: Introduced, Cryptogenic, and Range Expanding Cirripedia of North and South America. *In the Wrong Place - Alien Marine Crustaceans: Distribution,* *Biology and Impacts*, *6*, 159–213. https://doi.org/10.1007/978-94-007-0591-3_5

Cavalli-Sforza, L. L., & Feldman, M. W. (2003). The application of molecular genetic approaches to the study of human evolution. *Nature Genetics 2003 33:3*, *33*(3), 266–275. https://doi.org/10.1038/ng1113

Chen, D. W., & Garud, N. R. (2021). Rapid evolution and strain turnover in the infant gut microbiome. *BioRxiv*, 2021.09.26.461856. https://doi.org/10.1101/2021.09.26.461856

Chen, E. Z., & Li, H. (2016). A two-part mixed-effects model for analyzing longitudinal microbiome compositional data. *Bioinformatics*, *32*(17), 2611–2617. https://doi.org/10.1093/BIOINFORMATICS/BTW308

Chiu, A. M., Molloy, E. K., Tan, Z., Talwalkar, A., & Sankararaman, S. (2022). Inferring population structure in biobank-scale genomic data. *The American Journal of Human Genetics*. https://doi.org/10.1016/J.AJHG.2022.02.015

CRAN - Package ggpubr. (n.d.). Retrieved March 6, 2022, from https://cran.r-project.org/web/packages/ggpubr/index.html

DeGiorgio, M., & Rosenberg, N. A. (2013). Geographic Sampling Scheme as a Determinant of the Major Axis of Genetic Variation in Principal Components Analysis. *Molecular Biology and Evolution*, *30*(2), 480–488. https://doi.org/10.1093/MOLBEV/MSS233

Delmont, T. O., Kiefl, E., Kilinc, O., Esen, O. C., Uysal, I., Rappé, M. S., Giovannoni, S., & Eren, A. M. (2019). Single-amino acid variants reveal evolutionary processes that shape the biogeography of a global SAR11 subclade. *ELife*, *8*. https://doi.org/10.7554/ELIFE.46497

Dlugosch, L., Poehlein, A., Wemheuer, B., Pfeiffer, B., Badewien, T. H., Daniel, R., & Simon, M. (2022). Significance of gene variants for the functional biogeography of the near-surface Atlantic Ocean microbiome. *Nature Communications 2022 13:1*, *13*(1), 1–11. https://doi.org/10.1038/s41467-022-28128-8

Elsaeed, E., Fahmy, N., Hanora, A., & Enany, S. (2021). Bacterial Taxa Migrating from the Mediterranean Sea into the Red Sea Revealed a Higher Prevalence of Anti-Lessepsian Migrations. *OMICS A Journal of Integrative Biology*, *25*(1), 60–71. https://doi.org/10.1089/omi.2020.0140

Enav, H., & Ley, R. E. (2021). SynTracker: a synteny based tool for tracking microbial strains. *BioRxiv*, 2021.10.06.463341. https://doi.org/10.1101/2021.10.06.463341

Ferretti, P., Pasolli, E., Tett, A., Asnicar, F., Gorfer, V., Fedi, S., Armanini, F., Truong, D. T., Manara, S., Zolfo, M., Beghini, F., Bertorelli, R., De Sanctis, V., Bariletti, I., Canto, R., Clementi, R., Cologna, M., Crifò, T., Cusumano, G., … Segata, N. (2018). Mother-to-Infant Microbial Transmission from Different Body Sites Shapes the Developing Infant Gut Microbiome. *Cell Host and Microbe*, *24*(1), 133-145.e5. https://doi.org/10.1016/j.chom.2018.06.005

Flores, G. E., Bates, S. T., Knights, D., Lauber, C. L., Stombaugh, J., Knight, R., & Fierer, N. (2011). Microbial Biogeography of Public Restroom Surfaces. *PLoS ONE*, *6*(11), e28132. https://doi.org/10.1371/journal.pone.0028132

Garud, N. R., Good, B. H., Hallatschek, O., & Pollard, K. S. (2019). Evolutionary dynamics of bacteria in the gut microbiome within and across hosts. *PLoS Biology*, *17*(1), e3000102. https://doi.org/10.1371/JOURNAL.PBIO.3000102

Golani, D. (2009). Distribution of Lessepsian migrant fish in the Mediterranean. *Http://Dx.Doi.Org/10.1080/11250009809386801*, *65*(S1), 95–99. https://doi.org/10.1080/11250009809386801

Hijmans, R. J., Karney, C., Geographiclib,] (, Williams, E., Vennes, C., & Maintainer,]. (2021). *Package "geosphere."* https://doi.org/10.1007/s00190012

Hildebrand, F., Gossmann, T. I., Frioux, C., Özkurt, E., Myers, P. N., Ferretti, P., Kuhn, M., Bahram, M., Nielsen, H. B., & Bork, P. (2021). Dispersal strategies shape persistence and evolution of human gut bacteria. *Cell Host and Microbe*, *29*(7), 1167-1176.e9. https://doi.org/10.1016/J.CHOM.2021.05.008/ATTACHMENT/B6A2A326-7A64-425F-A15E-F46BE993903C/MMC5.XLSX

Katsanevakis, S., Coll, M., Piroddi, C., Steenbeek, J., Lasram, F. B. R., Zenetos, A., & Cardoso, A. C. (2014). Invading the Mediterranean Sea: Biodiversity patterns shaped by human activities. *Frontiers in Marine Science*, *1*(SEP), 32. https://doi.org/10.3389/FMARS.2014.00032/ABSTRACT

Knights, D., Kuczynski, J., Charlson, E. S., Zaneveld, J., Mozer, M. C., Collman, R. G., Bushman, F. D., Knight, R., & Kelley, S. T. (2011). Bayesian community-wide culture-independent microbial source tracking. *Nature Methods*, *8*(9), 761–765. https://doi.org/10.1038/nmeth.1650

Korpela, K., Costea, P., Coelho, L. P., Kandels-Lewis, S., Willemsen, G., Boomsma, D. I., Segata, N., & Bork, P. (2018). Selective maternal seeding and environment shape the human gut microbiome. *Genome Research*, *28*(4), 561–568. https://doi.org/10.1101/GR.233940.117/-/DC1

Ladau, J., Sharpton, T. J., Finucane, M. M., Jospin, G., Kembel, S. W., O'Dwyer, J., Koeppel, A. F., Green, J. L., & Pollard, K. S. (2013). Global marine bacterial diversity peaks at high latitudes in winter. *The ISME Journal 2013 7:9*, *7*(9), 1669–1677. https://doi.org/10.1038/ismej.2013.37

Langmead, B., & Salzberg, S. L. (2012). Fast gapped-read alignment with Bowtie 2. *Nature Methods 2012 9:4*, *9*(4), 357–359. https://doi.org/10.1038/nmeth.1923

Laso-Jadart, R., Sykulski, A. M., Ambroise, C., & Madoui, A. (2021). How marine currents and environment shape plankton genomic differentiation: a mosaic view from Tara Oceans metagenomic data. *BioRxiv*, 2021.04.29.441957. https://doi.org/10.1101/2021.04.29.441957

Li, S. S., Zhu, A., Benes, V., Costea, P. I., Hercog, R., Hildebrand, F., Huerta-Cepas, J., Nieuwdorp, M., Salojärvi, J., Voigt, A. Y., Zeller, G., Sunagawa, S., De Vos, W. M., & Bork, P. (2016). Durable coexistence of donor and recipient strains after fecal microbiota transplantation. *Science*, *352*(6285), 586–589. https://doi.org/10.1126/SCIENCE.AAD8852/SUPPL_FILE/LI-SM.PDF

Martin, B. D., Witten, D., & Willis, A. D. (2020). MODELING MICROBIAL ABUNDANCES AND DYSBIOSIS WITH BETA-BINOMIAL REGRESSION. *The Annals of Applied Statistics*, *14*(1), 94. https://doi.org/10.1214/19-AOAS1283

McGhee, J. J., Rawson, N., Bailey, B. A., Fernandez-Guerra, A., Sisk-Hackworth, L., & Kelley, S. T. (2020). Meta-SourceTracker: application of Bayesian source tracking to shotgun metagenomics. *PeerJ*, *8*, e8783. https://doi.org/10.7717/peerj.8783

Menzies, R. J. (1968). Transport of Marine Life between Oceans through the Panama Canal. *Nature 1968 220:5169*, *220*(5169), 802–803. https://doi.org/10.1038/220802a0

Mitchell, C. M., Mazzoni, C., Hogstrom, L., Bryant, A., Bergerat, A., Cher, A., Pochan, S., Herman, P., Carrigan, M., Sharp, K., Huttenhower, C., Lander, E. S., Vlamakis, H., Xavier, R. J., & Yassour, M. (2020). Delivery Mode Affects Stability of Early Infant Gut Microbiota. *Cell Reports Medicine*, *1*(9). https://doi.org/10.1016/J.XCRM.2020.100156/ATTACHMENT/C379FFAF-51B2-4926-B735-FE6251D4BFB2/MMC2.XLSX

Nayfach, S., Rodriguez-Mueller, B., Garud, N., & 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. https://doi.org/10.1101/gr.201863.115

O'Malley, M., Sykulski, A. M., Laso-Jadart, R., & Madoui, M. A. (2021). Estimating the Travel Time and the Most Likely Path from Lagrangian Drifters. *Journal of Atmospheric and Oceanic Technology*, *38*(5), 1059–1073. https://doi.org/10.1175/JTECH-D-20-0134.1

Olm, M. R., Crits-Christoph, A., Bouma-Gregson, K., Firek, B. A., Morowitz, M. J., & Banfield, J. F. (2021). inStrain profiles population microdiversity from metagenomic data and sensitively detects shared microbial strains. *Nature Biotechnology 2021 39:6*, *39*(6), 727–736. https://doi.org/10.1038/s41587-020-00797-0

Reveillaud, J., Bordenstein, S. R., Cruaud, C., Shaiber, A., Esen, Ö. C., Weill, M., Makoundou, P., Lolans, K., Watson, A. R., Rakotoarivony, I., Bordenstein, S. R., & Eren, A. M. (2019). The Wolbachia mobilome in Culex pipiens includes a putative plasmid. *Nature Communications 2019 10:1*, *10*(1), 1–11. https://doi.org/10.1038/s41467-019-08973-w

Schloissnig, S., Arumugam, M., Sunagawa, S., Mitreva, M., Tap, J., Zhu, A., Waller, A., Mende, D. R., Kultima, J. R., Martin, J., Kota, K., Sunyaev, S. R., & Weinstock, G. M. (2013). Genomic variation landscape of the human gut microbiome. *Nature*. https://doi.org/10.1038/nature11711

Schmidt, T. S. B., Hayward, M. R., Coelho, L. P., Li, S. S., Costea, P. I., Voigt, A. Y., Wirbel, J., Maistrenko, O. M., Alves, R. J. C., Bergsten, E., de Beaufort, C., Sobhani, I., Heintz-Buschart, A., Sunagawa, S., Zeller, G., Wilmes, P., & Bork, P. (2019). Extensive transmission of microbes along the gastrointestinal tract. *ELife*, *8*. https://doi.org/10.7554/ELIFE.42693

Shenhav, L., Thompson, M., Joseph, T. A., Briscoe, L., Furman, O., Bogumil, D., Mizrahi, I., Pe'er, I., & Halperin, E. (2019). FEAST: fast expectation-maximization for microbial source tracking. *Nature Methods*, *16*(7). https://doi.org/10.1038/s41592-019-0431-x

Shi, Z. J., Dimitrov, B., Zhao, C., Nayfach, S., & Pollard, K. S. (2021). Fast and accurate metagenotyping of the human gut microbiome with GT-Pro. *Nature Biotechnology 2021 40:4*, *40*(4), 507–516. https://doi.org/10.1038/s41587-021-01102-3

Sloan, W. T., Lunn, M., Woodcock, S., Head, I. M., Nee, S., & Curtis, T. P. (2006). Quantifying the roles of immigration and chance in shaping prokaryote community structure.

Environmental Microbiology, *8*(4), 732–740. https://doi.org/10.1111/J.1462-2920.2005.00956.X

Sloan, W. T., Woodcock, S., Lunn, M., Head, I. M., & Curtis, T. P. (2007). Modeling taxa-abundance distributions in microbial communities using environmental sequence data. *Microbial Ecology*, *53*(3), 443–455. https://doi.org/10.1007/S00248-006-9141-X/FIGURES/4

Sprockett, D. D., Martin, M., Costello, E. K., Burns, A. R., Holmes, S. P., Gurven, M. D., & Relman, D. A. (2020). Microbiota assembly, structure, and dynamics among Tsimane horticulturalists of the Bolivian Amazon. *Nature Communications 2020 11:1*, *11*(1), 1–14. https://doi.org/10.1038/s41467-020-17541-6

Sunagawa, S., Coelho, L. P., Chaffron, S., Kultima, J. R., Labadie, K., Salazar, G., Djahanschiri, B., Zeller, G., Mende, D. R., Alberti, A., Cornejo-Castillo, F. M., Costea, P. I., Cruaud, C., d'Ovidio, F., Engelen, S., Ferrera, I., Gasol, J. M., Guidi, L., Hildebrand, F., … Bork, P. (2015). Structure and function of the global ocean microbiome. *Science*, *348*(6237). https://doi.org/10.1126/SCIENCE.1261359

Yassour, M., Jason, E., Hogstrom, L. J., Arthur, T. D., Tripathi, S., Siljander, H., Selvenius, J., Oikarinen, S., Hyöty, H., Virtanen, S. M., Ilonen, J., Ferretti, P., Pasolli, E., Tett, A., Asnicar, F., Segata, N., Vlamakis, H., Lander, E. S., Huttenhower, C., … Xavier, R. J. (2018). Strain-Level Analysis of Mother-to-Child Bacterial Transmission during the First Few Months of Life. *Cell Host & Microbe*, *24*(1), 146-154.e4. https://doi.org/10.1016/J.CHOM.2018.06.007