UCSF
UC San Francisco Electronic Theses and Dissertations

Title
Illuminating the intra-species diversity of bacterial populations from shotgun metagenomes

Permalink
https://escholarship.org/uc/item/0rv376sw

Author
Nayfach, Stephen

Publication Date
2017

Peer reviewed|Thesis/dissertation

eScholarship.org Powered by the California Digital Library
University of California

Illuminating the intra-species diversity of bacterial populations from shotgun metagenomes

by

Stephen Nayfach

DISSERTATION

Submitted in partial satisfaction of the requirements for the degree of

DOCTOR OF PHILOSOPHY

in

Biological and Medical Informatics

in the

GRADUATE DIVISION

of the

UNIVERSITY OF CALIFORNIA, SAN FRANCISCO

Copyright 2017

by

Stephen Nayfach

Illuminating the intra-species diversity of bacterial populations from shotgun metagenomes

Stephen Nayfach

Abstract

Deep metagenomic sequencing has the potential to illuminate the intra-species genomic variation of abundant microbial species. In this thesis, I develop a new tool MIDAS (Metagenomic *I*ntra-species Diversity Analysis System) for rapidly and automatically quantifying species abundance, single nucleotide polymorphisms (SNPs), and gene copy number variants (CNVs) from metagenomes. To illustrate the utility of this approach, I re-analyze three public datasets with MIDAS. First, I re-analyze stool metagenomes from 98 mother-infant pairs and used rare SNPs to track strain transmission. I find that early colonizers are likely transmitted from the mother whereas late colonizers are likely transmitted from the environment. Second, I re-analyze >300 stool metagenomes from healthy adults and use SNPs to identify examples of both strain co-existence and strain co-exclusion. Third, I re-analyze 198 globally distributed marine metagenomes and used gene copy number variants to show that many species have population structure that correlates with geographic location. Strain level genetic variants clearly reveal extensive structure and dynamics that are obscured when metagenomes are analyzed at coarser taxonomic resolution.

Table of Contents

List of Tables

List of Figures

Chapter 1

Defining bacterial species in the genomics era

Quantifying the intra-species variation of bacterial populations requires a clear and consistent definition of what is a bacterial species. Because taxonomic annotations of reference genomes derive from many different sources (e.g., individual labs, sequencing centers), Latin names of species are often erroneous and inconsistent. I address this problem in this chapter by applying a consistent, sequence-based definition of bacterial species to >30,000 currently sequenced bacterial reference genomes. I identify 30 universally distributed gene families that are optimal for defining species based on a comparison to genome-wide average nucleotide identity (ANI), which is considered a gold standard for microbial species delineation. Next, I systematically compare species derived from this approach to a reference taxonomy and identify many differences. Finally, I use metagenomes to address whether current reference genomes capture the diversity of bacterial species found in different environments and find that many types of communities are dominated by novel organisms.

1.1 Background

Quantifying the intra-species variation of bacterial populations requires a clear and consistent definition of what is a bacterial species. Current microbial genome taxonomies (e.g. NCBI) rely on Latin names provided by users [1]. This results in taxonomic annotations that are inconsistent, erroneous, and incomplete [2]. For example, over 20% of current microbial genomes are not annotated at the species level.

Over the past 10 years, there have been a number of efforts to systematically catalog bacterial species based on genomic information [2-5]. For example, Varghese et al. [4] used genome-wide average nucleotide identity, considered to be a gold standard [6], for delineating microbial species among 13,151 reference genomes (Table 1.1). Several years prior, Mende et al. [2] used 40 genes to found in nearly all prokaryotes to delineate species among 3,496 genomes (Table 1.1).

However, these high quality taxonomic annotations have not kept-up with the flood of new genome sequences. At start of my project in 2015 there were over 30,000 bacterial genomes available in the Pathosystems Resource Integration Center (PATRIC) [7] and two years later there are now over 90,000. Single cell genomics [8], discovery of genomes from metagenomes [9], and large-scale genome sequencing projects [10] will only increase the pace of new genome discovery.

To address this issue, I developed a procedure to hierarchically cluster bacterial genomes into species groups based on the pairwise percent identity across a set of 30 universal gene families.

These gene families were systematically identified from a set of 112 candidates to achieve equivalent results to whole genome comparisons. Thus, this approach is similar in speed to that described by Mende et al. because it involves comparison of only 30 genes between genomes. Furthermore, it achieves a resolution similar to that of Varghese et al. Finally, because this approach uses a small set of highly informative marker genes that comprise only ~1% of a typical genome it will be efficient to update these annotations as additional genomes are sequenced.

1.2 Methods

First I downloaded and quality controlled all currently available bacterial reference genomes. Reference genomes (N=33,252) were downloaded from the Pathosystems Resource Integration Center (PATRIC) [7] on March of 2015. I searched these genomes against 112 universally distributed gene families [11] using HMMER3 [12] and identified homologs with E-values $<1\times10^{-5}$. When there were multiple homologs of a gene family identified in a genome, I took the homolog with the lowest E-value. Low quality genomes – defined as having fewer than 100 universal genes (N=1,837) or greater than 1,000 contigs (N=618) – were removed. This left us with 31,007 high-quality genomes.

Next, I aligned all genomes to each other at the 112 universal genes using BLASTN [13]. I filtered out local alignments where either the query or target was covered by <70% of its length. I converted percent identities to distances using the formula: $D_{ab} = (100 - P_{ab})/100$, where P_{ab} was the percent identity of a gene between genomes a and b. This resulted in an undirected graph for each marker gene family where nodes were genomes and edges were distances.

To identify clusters of genomes that represented bacterial species, I performed average-linkage hierarchical clustering using the program MC-UPGMA [14]. The input to MC-UPGMA was the set of pairwise distances between genomes and the output of MC-UPGMA was a dendrogram. The dendrogram was cut at different distance thresholds (0.01 to 0.10, representing 90-99% identity) to identify genome-clusters (i.e. connected components). This procedure was performed separately for each of the 112 gene families.

1.3 Validation

For validation, I compared each set of genome-clusters to average nucleotide identity (ANI), which is considered to be a gold standard for delineating prokaryotic species [6, 15] but was too computationally intensive to compute for all genome-pairs. Specifically, I used the procedure described by Richter and Rossello-Mora [6] to compute ANI for >18,000 genome-pairs. Genome pairs with ANI ≥ 95% were labeled as members of the same species and genomes pairs with ANI < 95% were labeled as members of different species.

Next, I compared the true species labels to those predicted from genome clustering. Each genome pair was classified into one of the following categories: *true positive*: a clustered genome-pair with ANI ≥ 95%; *false positive*: a clustered genome-pair with ANI < 95%; *false negative*: a split genome-pair with ANI ≥ 95%; *true negative*: a split genome-pair with ANI < 95%. Using these classifications I calculated the true positive rate (TPR), precision (PPV), and F1-score for each set of genome-clusters. I used this procedure to evaluate performance for

genome-clusters defined using each of the 112 universal gene families and at each of the 10 distance cutoffs (Figure 1.1).

Based on this evaluation, I identified a subset of 30 universal gene families that produced genome-clusters that were in agreement with ANI (F1-scores > 0.93). Interestingly, I found that the best gene families for identifying bacterial species were less conserved and more widely distributed across the tree of life relative to other genes I tested (Figure 1.2). For example, many ribosomal gene families were too conserved to differentiate closely related species.

Finally, to increase clustering performance, I combined results across the 30 universal gene families. Specifically, for each genome-pair I averaged distances across these 30 gene-families to obtain a new genome distance graph. This graph was used to cluster genomes again using MC-UPGMA. Performance was evaluated again at different distance thresholds. I found that a distance cutoff of 0.035 (96.5% nucleotide identity) maximized the F1-score at 0.98 and resulted in 5,952 genome-clusters (Figure 1.3 and Table 1.2). In conclusion, I developed a fast procedure that produced bacterial species groups that were highly concordant with a gold standard definition of species based on 95% ANI.

1.4 Results

1.4.1 Comparison to the PATRIC taxonomy

I annotated each of the 5,952 genome-clusters according to the most common Latin name of genomes within the cluster. Interestingly, these names often differed from those specified by the PATRIC taxonomy (Figure 1.3). In particular this procedure clustered 2,666 genomes (8.6% of

total) that had not been previously annotated at the species level. About half of these genomes were assigned to cluster with at least one other reference genome, providing additional support that these represent new species of bacteria. Other genomes were previously annotated at the species level were assigned to a genome cluster with a different Latin name (4.3% of total). These represent potentially erroneous labels. Finally, I found other genomes that were split from a larger cluster with the same Latin name (5.22%). These represent diverse clades of bacteria that have been given a common Latin name. For example, genomes annotated as H. pylori often differ by as much as 80% ANI [1].

1.4.2 Quantifying the % of organisms from the environment with a sequenced reference genome

My approach for quantifying intra-species genomic variation relies on mapping reads to reference genomes. Therefore this cannot quantify genomic diversity for "novel" bacterial species without a sequenced representative in the reference database. Despite that microbial genome sequences are doubling in number every 18 months, the vast majority microorganisms in the environment are novel [10, 16]. To a lesser extent, even well studied environments like the human microbiome contain novel species [17, 18]. It was therefore important to quantify the percent of novel organisms with respect to the database of >30,000 genomes and identity the types of environments where the approach of mapping reads to reference genomes to uncover strain-level variants could be successful.

I developed a method to estimate the % of cellular organisms (i.e. Archaea, Bacteria and Fungi) in a metagenome that have a sequenced representative in the reference database (Figure 1.4).

First, the method aligns reads to a panel of 15 universal gene families and applied species-level mapping thresholds (94 to 98 % DNA identity). These genes and mapping thresholds were selected to optimize classification accuracy. DNA to DNA alignment was performed using HS-BLASTN [19]. Based on reads mapped to these genes, I estimated the total sequencing depth of all organisms with a sequenced genome in the database. Next, the method estimates the total sequencing depth across *all* organisms at the domain level, including those absent from the reference database. This was done using a tool I developed, called MicrobeCensus [20], which aligns reads to a panel of universal genes with parameters designed to recruit reads from distantly related organisms. Finally, to estimate the percent of organisms with a sequenced genome, I take the ratio of these two quantities, multiplied by 100.

I first applied this method to stool metagenomes from the Human Microbiome Project [21] and four other studies of the human gut [22-25] (Figure 1.5). I found that the majority of organisms from the human body had a sequenced reference genome at the species level. This included communities from skin (83%), nasal cavity (63%), urogenital tract (62%), mouth (55%), and gastrointestinal tract (49%). The best characterized human gut communities (i.e. greatest % of known organisms) came from individuals in the United States (52%), Europe (45%), and China (54%) that live urban lifestyles. In contrast, gut microbiomes of individuals from Tanzania and Peru that live hunter-gatherer and agricultural lifestyles had a much lower percentage of known organisms (9% and 13% respectively). This finding extends the previous discoveries of elevated levels of novel genera [26] and functions [23] in the gut microbiome of African hunter-gatherers.

Next, I was interested in identifying taxonomic groups that contained novel species in the human gut. Towards this goal, I performed Spearman correlations between the % of known organisms and the relative abundance of genera across HMP stool samples. Genus-level relative abundances were estimated using mOTU [17]. Gut communities with a higher % of novel organisms tended to have higher levels of several genera including *Coprococcus*, *Subdoligranulum*, *Dorea*, and *Blautia*, whereas communities with a higher % of known organisms tended to have higher levels of the genus *Bacteroides* (Figure 1.6). This analysis points to specific phylogenetic gaps in the set of currently sequenced bacterial genomes from the human gut.

Next, I quantified the % of known organisms across metagenomes from non-human environments. These included metagenomes from mouse stool [27], baboon stool [28], seawater [29], and soil [30]. Strikingly, I found that novel organisms consistently dominated these environments. For example, the best-characterized non-human environment I surveyed was marine surface water, where only 8.2% of organisms were estimated to have a sequenced genome at the species level. Even in the mouse gut microbiome, which is commonly used as a model system for the human gut microbiome, only 4.3% of genomes had a sequenced representative at the species level. This agrees with a previous report, which found that only 4% of microbial genes from the mouse gut overlapped with genes from the human gut [27]. In conclusion, there remains a massive gap between the microbial diversity found in non-human environments and that represented by sequenced bacterial reference genomes. Strain-level analyses can still be performed for these environments, but only for those species with sequenced representatives.

1.5 Conclusions & Discussion

Quantifying the intra-species variation of bacterial populations requires a clear and consistent definition of what is a bacterial species. In this chapter, I address this problem by applying a consistent, sequence-based definition of bacterial species to >30,000 currently sequenced bacterial reference genomes. This approach was fast and produced bacterial species groups that were highly concordant with a gold standard definition of species based on 95% ANI.

While my species groups were concordant with 95% ANI, I found 20% of genomes assigned to a species group disagreed with their annotated Latin name. About half of these discrepancies were because the genome was unannotated at the species level (e.g. *Streptococcus sp.)*. In other cases, I found diverse well-studied species groups, like H. pylori, that were split into 10's to 100's of clusters. Future work may be needed to refine species boundaries for these groups of organisms. Different species groups may also require different definitions. One promising solution is to use genome sequences from type strains to guide species definitions [1]. Another interesting idea would be to define bacterial species based on rates of recombination and lateral gene transfer, analogous to the current species definition in multicellular organisms. However, quantifying recombination and lateral gene transfer between genomes remains a major challenge in it of itself. Finally, it might be possible to create an ecological definition of species that relied on gene content, rather than sequence identity of core genes.

Scanning through >300 public metagenomes from 5 different biomes, I found that there remains a massive gap between the microbial diversity found in non-human environments and that represented by sequenced bacterial reference genomes. For example, I estimated that only ~1%

of genomes from soil have a sequenced representative. Luckily, genome sequences continue to rapidly increase in number and diversity, particularly with the advent of single cell genomics [8] and new computational [9] approaches to uncover genome sequences of uncultured microbes. For this reason, it will be important to update my species definitions as the number [31] and diversity [10] of microbial reference genomes continues to rapidly grow. It will also be useful to extend these sequenced-based annotations to Viruses, Eukaryotes, and Archaea.

1.6 Tables

Reference	Genomes	Species	Genomes/ Species	Naming conflicts	Distance estimation	Clustering algorithm	Identity cutoff
Schloissnig et al. Nature. 2012.	1,497	929	1.6	n/a	40 universal marker genes	Complete linkage	95.00%
Mende et al. Nature Methods. 2013.	3,496	1,753	2.0	19.80%	40 universal marker genes	Average linkage	96.50%
Varghese et al. Nucleic Acids Research. 2015.	13,151	3,032	4.3	18.00%	Bi-directional best hits	Complete linkage	96.50%
Nayfach et al. Genome Research. 2016.	31,007	5,952	5.2	18.40%	30 bacterial marker genes	Average linkage	96.50%

Table 1.1 A comparison of efforts to systematically cluster all known prokaryotic genomes into species groups.

% Identity Cutoff	TPR	PPV	F1-score
99.0	0.667	1.000	0.800
98.5	0.787	0.999	0.881
98.0	0.872	0.999	0.931
97.5	0.922	0.998	0.958
97.0	0.953	0.989	0.971
96.5*	0.974	0.984	0.979
96.0	0.984	0.973	0.979
95.5	0.989	0.952	0.970
95.0	0.993	0.923	0.957
94.5	0.994	0.901	0.945
94.0	0.994	0.858	0.921
93.5	0.994	0.832	0.906

Table 1.2 Genome clustering performance for 30 marker genes at different percent identity cutoffs. Genes were selected from a panel of 114 "PhyEco" marker gene families [11]. Performance was determined by comparison to genome-wide average nucleotide identity (ANI). True positive: genome pair with ANI ≥ 95% clustered together; false positive: genome pair with ANI < 95% clustered together; false negative: genome pair with ANI ≥ 95% assigned to different clusters; true negative: genome pair with ANI < 95% assigned to different clusters. Asterisk indicates the selected cutoff.

1.8 Figures

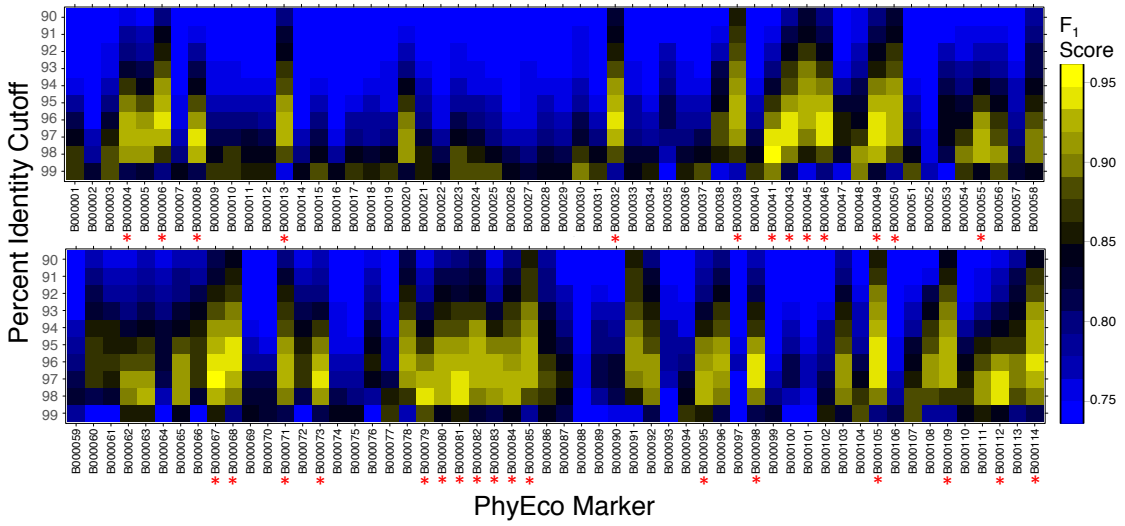

Figure 1.1. Genome clustering performance for 112 bacterial marker gene families. Marker-gene family names are listed on the horizontal-axis. The clustering percent identity cutoff is listed on the vertical-axis. Asterisks indicate selected gene families. Cell color indicates the F1-score,

which is a measure of clustering performance that balances the true positive rate with precision. True positive: genome pair with ANI ≥ 95% clustered together; false positive: genome pair with ANI < 95% clustered together; false negative: genome pair with ANI ≥ 95% assigned to different clusters; true negative: genome pair with ANI < 95% assigned to different clusters.

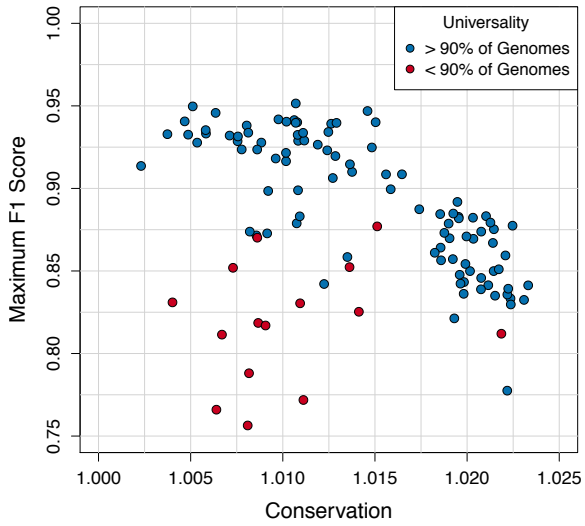

Figure 1.2. Features of gene families that explain genome-clustering performance. Clustering performance measured using the maximum F1-score across percent identity cutoffs. Universality is defined as the proportion of genomes where a gene family is found. Conservation is defined as the average ratio between the marker-gene percent identity ($PID_{i,j}$) and genome wide percent identity (ANI_i) across n genome pairs for each marker-gene j: $C_j = \frac{1}{n}\sum_i^n \frac{PID_{i,j}}{ANI_i}$. High conservation for a marker-gene indicates low sequence divergence relative to the genomic background.

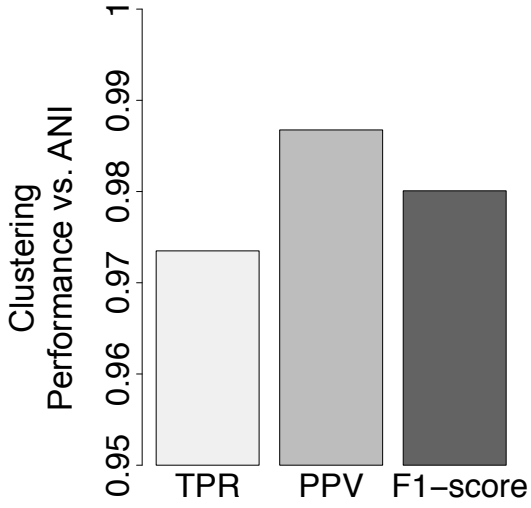

Figure 1.3. Clustering performance of top 30 marker genes versus ANI. TPR: true positive rate, PPV: precision, F1-score: harmonic mean of TPR and PPV. True positives and false positives are defined in Figure 1.1.

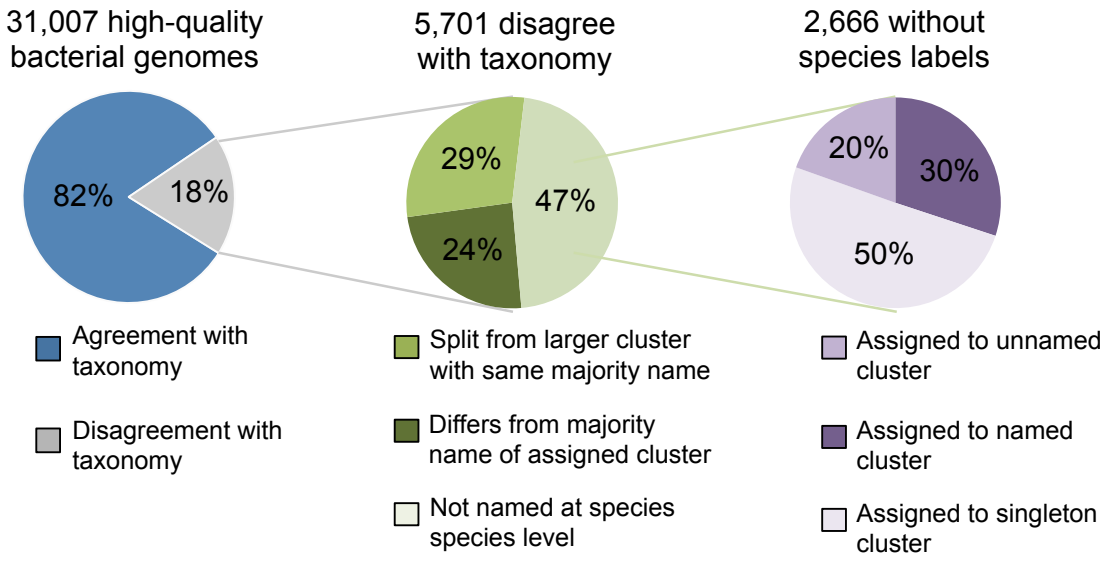

Figure 1.4. Concordance of genome-cluster names and annotated species names. Each genome-cluster was annotated according to the most common Latin name of genomes within the cluster. Of the 31,007 genomes assigned to a genome-cluster, 5,701 (18%) disagreed with the consensus Latin of the genome-cluster. Most disagreements are due to genomes lacking annotation at the species level (47%). Other disagreements are because a genome was split from a larger cluster with the same name (29%) or assigned to a genome-cluster with a different name (24%).

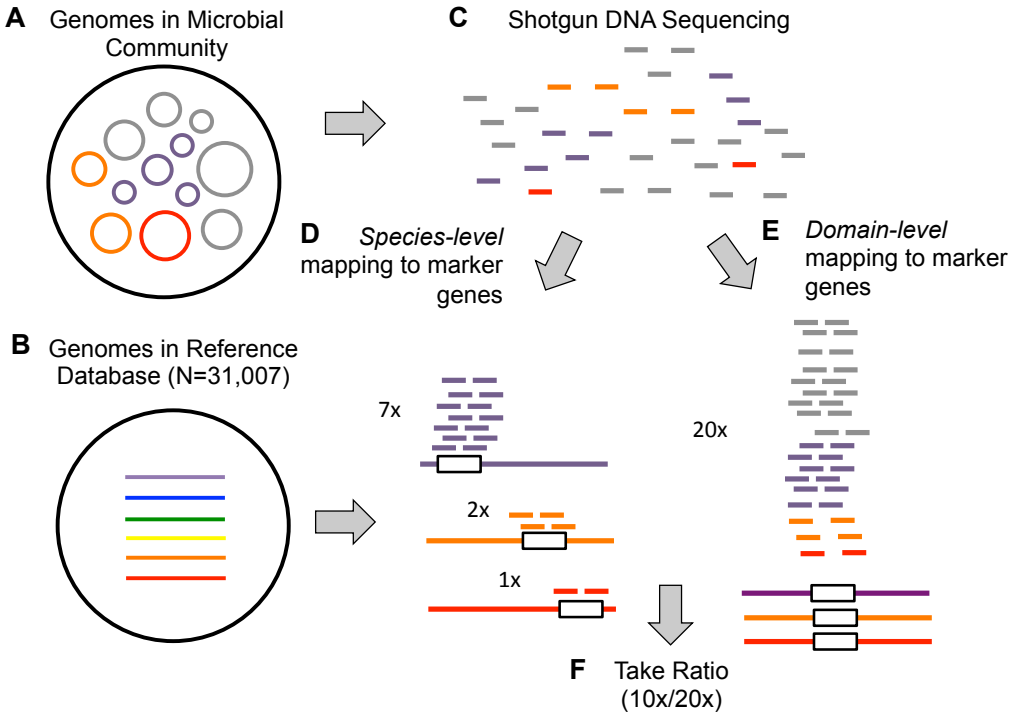

Figure 1.5. An approach for estimating the percent of genomes in a metagenome with a sequenced representative. Metagenomic reads are aligned to marker genes present in a collection of reference genomes. Reads are classified if their alignment satisfied either species or domain level mapping cutoffs. Species level mapping cutoffs recruit reads from strains of the same species as exist in the reference database. Domain level mapping cutoffs are much more lenient and recruit reads from any microbial genome (excluding viruses). The ratio of read-depth at marker-genes between the species and domain levels results in an estimate of the percent of genomes in a metagenome with a match to genome in the reference database at the species level.

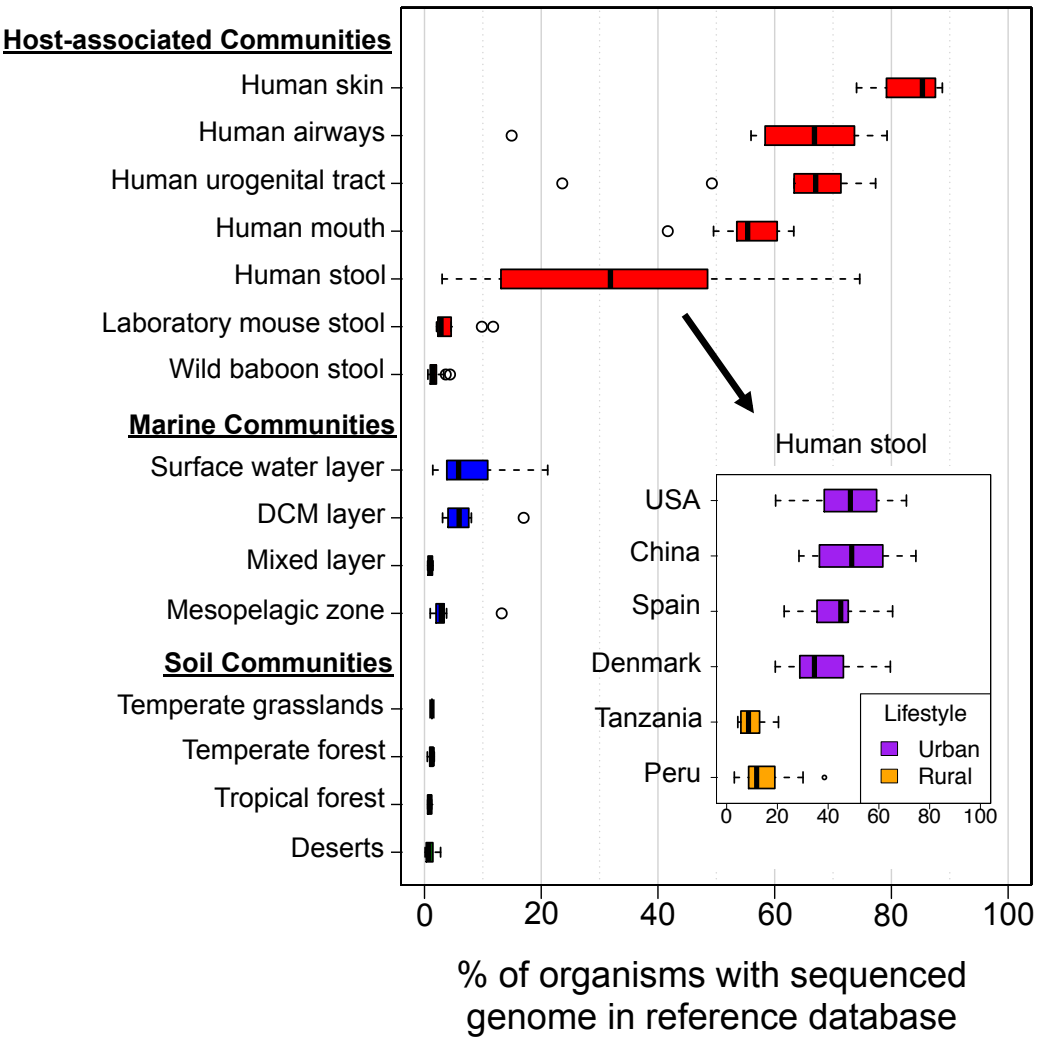

Figure 1.6. The percent of genomes with a sequenced representative across metagenomes from host-associated, marine, and terrestrial environments. Inset panel shows the distribution of database coverage across human stool metagenomes from six countries and two host lifestyles.

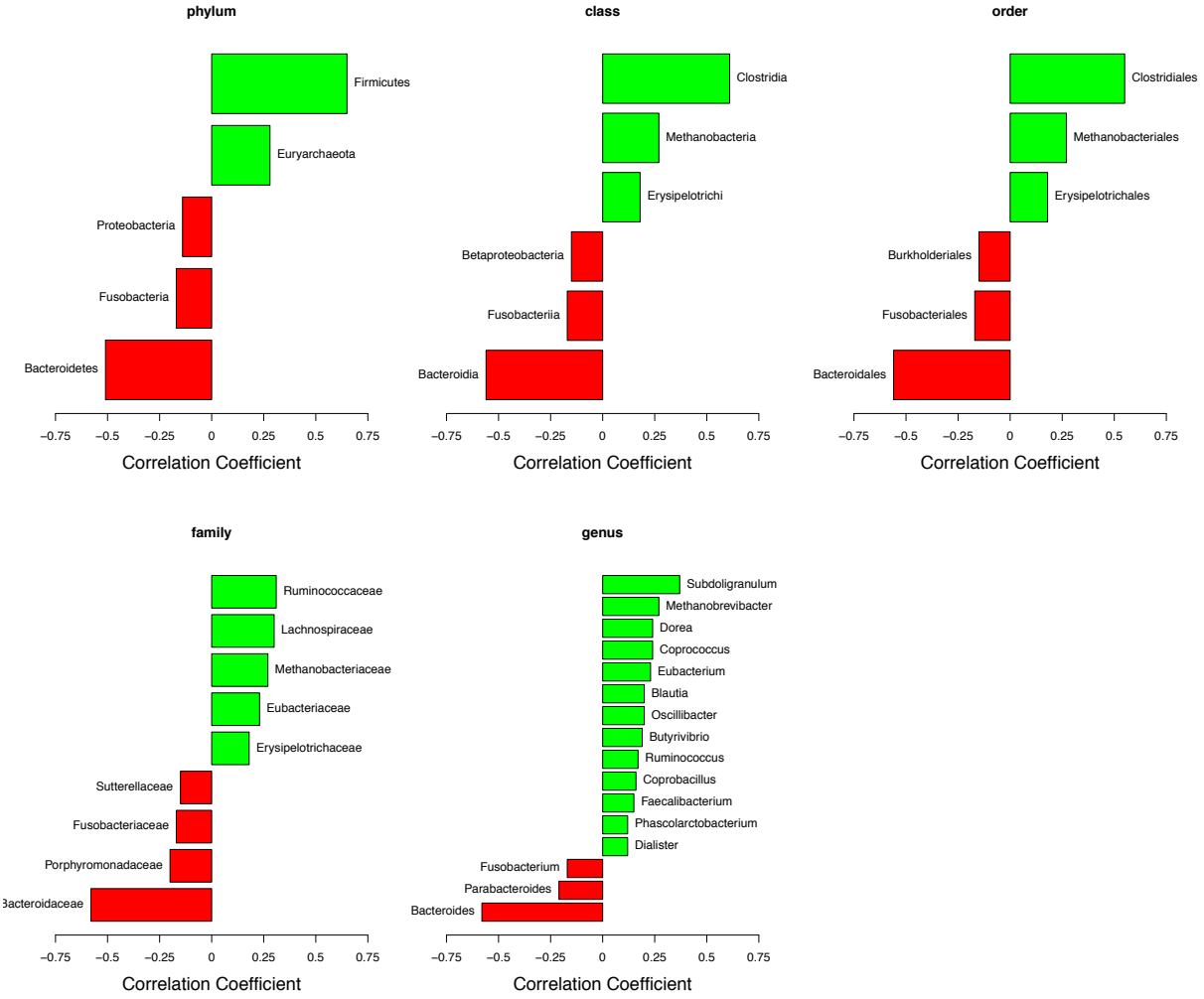

Figure 1.7 Correlations between taxon relative abundance and the percent of novel organisms in the human gut.

Chapter 2

An integrated pipeline for quantifying species abundance and strain-level genomic variation from metagenomes

In this chapter, I discuss construction of an integrated pipeline for automatically and accurately quantifying bacterial species abundance and strain-level genomic variation from shotgun metagenomes. The method is called the Metagenomic Intra-species Diversity Analysis System, or MIDAS (Figure 2.1). Open source software is freely available at http://github.com/snayfach/MIDAS. MIDAS leverages the genomic database of >30,000 bacterial reference genomes clustered into bacterial species (Chapter 1). Using realistic benchmark datasets, I show that MIDAS accurately quantifies species abundance, nucleotide variants and gene copy number variants, but requires at least 1 to 10x sequencing coverage.

2.1 Background

Deep metagenomic sequencing has the potential to illuminate the strain-level genomic diversity of microbial communities, yielding a genomic resolution not achievable by sequencing the 16S ribosomal RNA gene alone [17] and circumventing the need for culture-based approaches. However, limitations of existing computational methods and reference databases have hampered quantifying strain-level genomic variation from metagenomic data (Table 2.1).

One class of methods estimates the relative abundance of known strains from sequenced reference genomes (Table 2.1). Sigma [32] and Pathoscope [33] focus on accurately assigning reads to sequenced reference genomes. MetaPhlan2 [34] and GSMer [35] utilize strain-specific marker sequences identified from genomes. These methods are effective for well-characterized pathogens like *E. coli* that have thousands of sequenced genomes, but work poorly for the majority of species that have only a single sequenced genome.

A second class of tools uses SNP patterns to identify strains (Table 2.1). WG-FAST [36] uses consensus alleles at SNPs to reconstruct the core-genome sequence of the dominant strain in a metagenome. MetaMLST [37] does the same thing, but only for genes present in a multi-locus sequence typing (MLST) database. Neither of these approaches was designed to handle mixtures of two or more strains. To address this, binStrain [38] uses SNP allele frequencies to estimate mixtures of strains based on a SNP reference panel. However, this approach can only quantity the abundance of strains with sequenced reference genomes, which are utilized to build the SNP reference panel. ConStrains [39] overcomes this limitation by identifying strain haplotypes *de novo*. It does this by clustering together single nucleotide variants that co-vary in frequency

across metagenomic samples. However, ConStrains requires multiple samples where the same strains are found (e.g. time series samples) and may not be able to resolve strains in communities with high population heterogeneity.

A third class of tools uses sequence assembly to reconstruct microbial strains *de novo* from metagenomes (Table 2.1). Many tools exist for assembling metagenomic data [40-43], however, these nearly never produce complete microbial genomes. A number of clever strategies have been developed to overcome this limitation. MaxBin [44] bins contigs based on their coverage, nucleotide composition, and marker genes. Alneberg et al. [45] improved this algorithm by incorporating coverage *co-variation* of contigs across metagenomic samples; if two contigs have coverage levels that co-vary over tens to hundreds of metagenomes, they are likely linked together on the same chromosome. However, these approaches still require performing assembly on the entire metagenome, which can be time and memory intensive. To overcome this limitation, Clearly et al. designed a pre-assembly algorithm called LSA, which bins *sequencing reads* with k-mers that co-vary in abundance across metagenomic samples. Read bins of interest – for example corresponding to a specific strain – can then be assembled into genomes using significantly less memory. However, these latter methods require many metagenomic samples and may not work for rare strains that occur in only one metagenome.

A fourth class of methods identifies strain-level genomic variants based on read mapping to reference sequences, but does not attempt to identify discrete strains or haplotypes. Schloissnig et al. [3] described a procedure for identifying SNPs in human gut metagenomes and used SNPs to quantify population genetic parameters like pN/pS and nucleotide diversity. Greenblum et al.

[46] and Zhu et al. [47] described approaches for quantifying gene copy number variants and genes that are highly variable between strains of the same species. A similar approach was later implemented by Scholz et al [48] in the tool PanPhlAn. While these approaches are very promising, they are limited by a lack of existing software and good reference databases.

To address these issues, I developed the Metagenomic Intra-species Diversity Analysis System (MIDAS). MIDAS falls into the fourth class of methods described above. It automatically and accurately quantifies bacterial species abundance and strain-level genomic variation from shotgun metagenomes. MIDAS leverages the genomic database of >30,000 bacterial reference genomes clustered into bacterial species and identifies SNPs and CNVs in bacterial populations present with at least 1-10x sequencing depth. MIDAS processes ~5,000 reads per second and requires <1 gigabyte of RAM for a typical metagenome.

2.2 Methods

2.2.1 Estimating the relative abundance of bacterial species

In the first step in the pipeline, MIDAS estimates the sequencing depth and relative abundance of the 5,952 bacterial species in the reference database (Figure 2.1). This enables the automatic identification of species with sufficient sequencing depth for strain-level genomic variation analysis. Also, this dramatically increases pipeline speed and decreases RAM usage by limiting the size of the reference database: reads are only aligned to genes and genomes from species that are actually present (e.g. >1x sequencing depth). Finally, this enables users to quantify the taxonomic composition of a metagenome, which can be useful on its own.

Species abundance is estimated by mapping reads to a database of 15 gene families, which each occur in nearly all bacterial genomes at one copy per genome. Previous work has shown that these types of gene families are phylogenetically informative and can be used for metagenomic species profiling [17, 49]. In total, the database contains 87,895 genes from the 5,952 reference species. MIDAS aligns metagenomic reads to this database with HS-BLASTN [19] at an alignment rate of ~5,000 reads/second. To reduce false positives, local alignments that cover <70% of the read and alignments with low % identities are discarded. Each read is mapped one time according to its best hit in the reference database. Some reads align equally well to genes from two or more species. MIDAS probabilistically assigns these reads to species based on the number of uniquely mapped reads to each species. Finally, the mapped reads are used to estimate the sequencing depth and relative abundance of each species.

I selected these 15 gene families from a set of 112 candidates [11] on the basis of their ability to accurately recruit metagenomic reads to the correct species. To evaluate recruitment performance of each gene, I conducted an *in silico* metagenomic experiment in which the true species and gene family of origin for each read was known *a priori*. Specifically, I generated a dataset of 100-bp genomic fragments that were randomly sampled from bacterial reference genomes in the database. Each read was labeled based on its true species and gene family of origin. HS-BLASTN [19] was used to map these reads back to the database that contained gene sequences from all 112 gene families. All reference sequences were labeled with their true species and gene family. To simulate the presence of novel organisms, I discarded alignments between reads and reference sequences from the same genome. Each read was assigned to a species based on its top hit in the reference database. Recruitment performance was measured using the F1-score.

Based on this experiment, I identified 15 gene families that were able to accurately recruit metagenome reads (Table 2.2). Additionally, I made sure that these gene families were universally distributed and single copy. I also identified the optimal percent identity cutoffs for mapping reads to the database, which ranged from 94.5% to 98.0% identity depending on the gene family.

2.2.2 Estimating the pan-genome gene content of abundant species

To quantify the gene content of a bacterial population, MIDAS first uses the species abundance profile to identify bacterial species with sufficient coverage (Figure 2.1). Next, MIDAS dynamically builds a pan-genome database, which contains the set of non-redundant genes of all genomes from the abundant species. Bowtie 2 [50] is then used to map reads from the metagenome against the pan-genome database. Because the database typically contains genes from only a handful of abundant species, and redundant genes are removed, the mapping step is extremely fast. Each read is mapped a single time according to its best hit, and reads with an insufficient mapping percent identity (default=94%), alignment coverage (default=70%), or sequence quality (default=20) are discarded. Mapped reads are used to quantify the sequencing depth of each gene in the database. These values are normalized by the sequencing depth of single-copy genes, yielding an estimated copy number of each gene per cell in the bacterial population. Copy numbers are also thresholded to predict the presence or absence of the gene in the community.

MIDAS relies a reference database of pre-computed pan-genomes. A pan-genome is defined as the set of non-redundant genes across all genomes of a given species. I used the tool USEARCH [51] to cluster all genes from each species at 99% identity. This procedure clustered 116,978,184 genes from the 31,007 genomes into 31,840,245 gene families. I further clustered these genes at different levels of sequence identity (75-95% DNA identity) in order to identity gene families of varying size and diversity for downstream analyses. Functional annotations for all genes were obtained from PATRIC and include FIGfams [52], Gene Ontology [53], and KEGG Pathways [54].

2.2.3 Identifying core-genome SNPs in abundant species

To identify SNPs of individual species, MIDAS maps reads to a genome database (Figure 2.1). This database contains one representative genome sequence per species, and it only includes species with high sequencing coverage at universal single-copy genes in the metagenome being analyzed. Representative genomes are selected in order to maximize their sequence identity to all other genomes within the species. The core genome of each species is identified directly from the data using nucleotide positions in the representative genome that are at high coverage across multiple metagenomic samples. SNPs are quantified along the entire core genome, including at sites that are variable between samples, but fixed within individual samples. Core genome SNPs are useful because they occur in all strains of a species and facilitate comparative analyses.

To estimate core genome SNPs, MIDAS first uses the species abundance profile to identify species with sufficient coverage (e.g. >10x). A representative genome database is dynamically built, which contains a single genome per species that meets the coverage requirement. The

representative genome is a single genome chosen that has the greatest nucleotide identity, on average, to other members of the species. Only a single genome is needed for identifying the core genome, because this region should be present in all strains of a species. Bowtie 2 is used to globally map reads to the representative genome database. Each read is mapped a single time according to its best hit, and reads with an insufficient mapping percent identity (default=94%), alignment coverage (default=70%), mapping quality (default=20), or sequence quality (default=20) are discarded. Additionally, bases with low sequence quality scores are discarded (default=30). Samtools [55] is used to generate a pileup of nucleotides at each genomic position. Pileups are parsed to generate output files that report nucleotide variation at all genomic sites. To identify the core genome of a species, MIDAS uses the output from multiple metagenomic samples to identify regions at consistently high coverage (e.g. >10x coverage in 95% of samples). MIDAS then produces core genome SNP matrices for all species, which facilitate comparative analyses of nucleotide variation across genomic sites and metagenomic samples. MIDAS also gives the option of outputting all SNPs, including those that are not in the core genome.

2.3 Validation

I designed 20 realistic mock metagenomic datasets to validate each step in the MIDAS pipeline. I pooled short 100-bp Illumina reads from completed genome sequencing projects to create each mock metagenome. These datasets are expected to contain sequencing errors and other experimental artifacts found in real short-read sequencing data that might prevent accurate estimation of species abundance and strain-level genomic variation.

First, I identified short-read libraries to include in the mock metagenomes. I used the SRAdb MySQL database [56] to systematically scan through metadata from the NCBI Sequence Read Archive [57] and identify sequencing run accessions with the following criteria: 1) The read length was between 100 and 101 base-pairs, 2) The technology was Illumina GAIIx, Illumina HiSeq2000, or Illumina HiSeq2500, 3) Reads were paired-end, 4) There was a corresponding assembled genome in the MIDAS database, and 5) I selected a maximum of one short-read library per bacterial species. I identified 237 sequencing run accessions with these criteria. Short-reads were then pooled together to create the 20 mock metagenomes. Each metagenome contained reads from 20 randomly selected genome projects and contained 100x total genome coverage. The relative abundances of the 20 genomes were exponentially distributed in each simulation (50%, 25%, 12%, 6.5% etc.).

I ran each dataset through MIDAS and compared the output of MIDAS to the known species abundance, gene content, and SNPs in the simulated communities. To evaluate the accuracy of species abundance estimation I compared the expected relative abundance and coverage to the simulated relative abundance and coverage. I found that MIDAS accurately estimated species relative abundance but slightly underestimated the true sequencing depth of the species in the metagenome (Figure 2.2, left).

To evaluate the accuracy of gene content estimation, I ran MIDAS to estimate the copy-number of genes in the pan-genome of each species in each simulation (Figure 2.2, middle). I applied a cutoff to these values to predict gene presence-absence. True positives (TP) were present genes predicted as present, false positives (FP) were absent genes predicted as present, true negatives

(TN) were absent genes predicted as absent, and false negatives (FN) were present genes predicted as absent. Performance was measured across a range of copy-number cutoffs using balanced accuracy: (TPR+TNR)/2, where TPR=TP/(TP+ FN) and TNR=TN/(TN+FP). MIDAS accurately predicted the presence or absence of genes in species present with at least 1 to 3x sequencing coverage (Figure 2.2, middle). Prediction accuracy was maximized at 0.96 for strains with >3x coverage when using a threshold equal to 0.35x the coverage of universal single-copy genes – lower thresholds resulted in lower specificity and higher thresholds resulted in lower sensitivity.

To evaluate the accuracy of core genome SNPs, I ran MIDAS to estimate the frequency of nucleotide variants in the representative genome of each species in each simulation (Figure 2.2, right). I predicted SNPs using the consensus allele at each genomic position. True SNPs were identified by comparing genomes in the simulations to the representative genomes used for read mapping with the program MUMmer [58], which identified 3,971,528 total true SNPs. Because there is one genome per species in the mock community, all SNPs are differences from the reference genome. True positives were correctly called SNPs, false positives were incorrectly called SNPs, and false negatives were SNPs that were not called due to insufficient coverage. I compared predicted SNPs to true SNPs and measured performance using the true positive rate (TP/TP+FN) and precision (TP/TP+FP). I found that MIDAS called SNPs at a low false-discovery rate, but required between 5 to 10x coverage to identify the majority of SNPs present (Figure).

2.4 Conclusions & Discussion

Recent work has shown extensive strain-level genomic variation of bacteria at the level of gene copy number variants [46, 47] and single nucleotide variants [3], yet there is currently no method to automatically, efficiently, and accurately extract this information from shotgun metagenomes. Existing methods either estimate the relative abundance of known strains [32, 33, 35], or use SNPs to phylogenetically type strains [36, 39]. These methods not capture the functions of these organisms and therefore cannot shed light onto the ecological forces shaping their genomes.

To address these issues, I developed MIDAS, which is an integrated computational pipeline that quantifies bacterial species abundance and strain-level genomic variation from shotgun metagenomes. By coupling fast taxonomic profiling via a panel of universal-single-copy genes with sensitive pan-genome and whole-genome alignment, MIDAS can efficiently and automatically compare hundreds of metagenomes to >30,000 reference genomes to identify genetic variants present in the strains of each sample. The publicly available software and data resources will enable researchers to conduct large-scale population genetic analysis of metagenomes.

This first version of MIDAS has several limitations. Since it currently relies on bacterial reference genomes, MIDAS cannot quantify variation for novel species, novel genes, or known species from other groups of microbes (e.g. archaea, eukaryotes, and viruses). To accurately quantify strain-level gene content and SNPs, *MIDAS* requires greater than 1x and 10x sequencing coverage, respectively. This biases analyses towards the most abundant and prevalent species in an environment. MIDAS was nonetheless able to capture the majority of microbial species

abundance across human body sites, making it well suited for uncovering strain-level variation of human-associated bacteria.

2.5 Tables

Method	Reference	Description	Reference based	Species Abundance	Strain Abundance	CNVs	SNVs	Phylogenetic Reconstruction
MIDAS	Nayfach et al. (2016) Genome Research	Species abundance, intra-species genomic variants, & phylogenetic reconstruction	Yes	Yes	No	Yes	Yes	Dominant strain only
PanPhlAn	Scholz et al. (2016) Nature Methods	Intra-species genomic variants	Yes	No	No	Yes	No	No
ConStrains	Luo et al. (2015) Nature Biotechnology	Phylogentic reconstruction of strains	Yes	Yes	Yes	No	Only at marker genes	Yes
MetaMLST	Zolfo et al. (2016) Nucleic Acids Research	Phylogentic reconstruction of strains	Yes	No	No	No	Only at marker genes	Dominant strain only
WG-FAST	Sahl et al. (2015) Genome Medicine	Phylogentic reconstruction of strains	Yes	No	No	No	Yes	Dominant strain only
BinStrain	Joseph et al. (2015) bioArxiv	Strain relative abundance from SNP allele frequencies	Yes	No	Yes	No	Yes	Yes
Pathoscope	Francis et al. (2013) Genome Research	Strain relative abundance from accurate read assignment	Yes	Yes	Yes	No	No	No
Sigma	Ahn et al. (2014) Bioinformatics	Strain relative abundance from accurate read assignment	Yes	Yes	Yes	No	No	No
GSMer	Tu et al. (2014) Nucleic Acids Research	Strain relative abundance from strain-specific sequences	Yes	Yes	Yes	No	No	No
MetaPhlan2	Truong et al. (2015) Nature Methods	Strain relative abundance from strain-specific sequences	Yes	Yes	Yes	No	No	No
CONCOCT	Alenberg et a. (2014) Nature Methods	Post-assembly tool that bins contigs by coverage and composition	No	No	No	No	No	No
LSA	Cleary et al. (2015) Nature Biotechnology	Pre-assembly tool that bins reads by co-variation of k-mer frequencies	No	No	No	No	No	No

Table 2.1 A comparison of bioinformatics tools for high-resolution characterization of microbial communities from shotgun metagenomes.

PhyEco Marker	Universality	Copy #	% ID Cutoff	TPR	PPV	F1-score
B000032	0.98	0.99	95.50	0.88	0.83	0.86
B000039	0.99	1.02	94.75	0.89	0.82	0.86
B000041	1.00	1.01	98.00	0.85	0.79	0.82
B000062	0.99	0.99	97.25	0.85	0.79	0.82
B000063	1.00	1.01	96.00	0.88	0.76	0.82
B000065	1.00	1.01	98.00	0.81	0.81	0.81
B000071	0.99	1.01	95.25	0.87	0.80	0.84
B000079	1.00	1.01	98.00	0.84	0.81	0.83
B000080	0.99	1.00	95.25	0.88	0.78	0.83
B000081	1.00	1.01	97.00	0.86	0.81	0.84
B000082	1.00	1.03	95.25	0.86	0.81	0.84
B000086	0.99	1.00	96.75	0.84	0.77	0.81
B000096	0.99	1.03	96.75	0.86	0.82	0.84
B000103	1.00	1.02	95.25	0.88	0.78	0.83
B000114	1.00	1.04	94.50	0.89	0.81	0.85

Table 2.2 Selected marker gene families for metagenomic species profiling. Performance was based on a metagenomic simulation in which the true marker gene and species of each fragment was known. Mapping cutoffs maximize the F-score for each gene family, which balance classification sensitivity and precision.

2.6 Figures

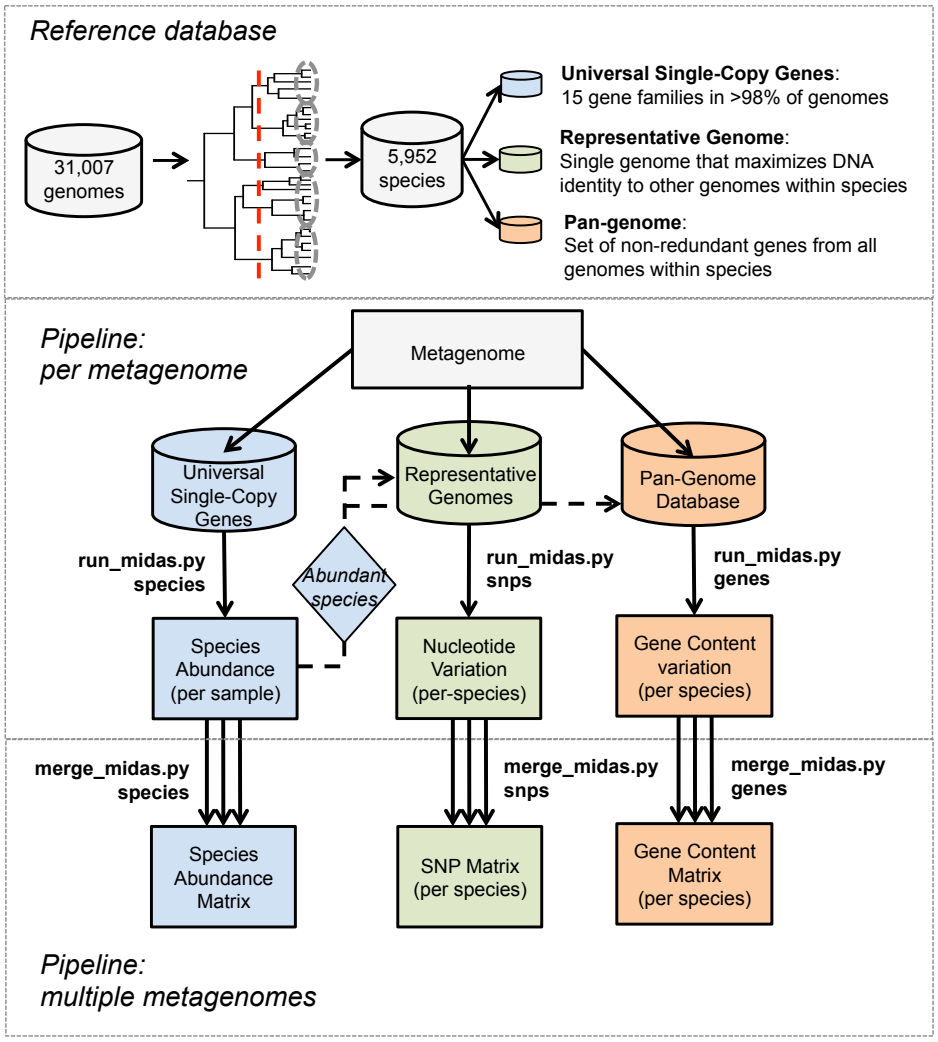

Figure 2.1 The MIDAS analysis pipeline. Reads are first aligned to a database of universal-single-copy genes to estimate species coverage and relative abundance per sample. For species with sufficient coverage, reads are next aligned to a pan-genome database of genes to estimate gene coverage, copy-number, and presence-absence. Finally, reads are aligned to a representative genome database to detect SNPs in the core genome. The core genome is defined directly from the data by identifying high coverage regions across multiple metagenomic samples.

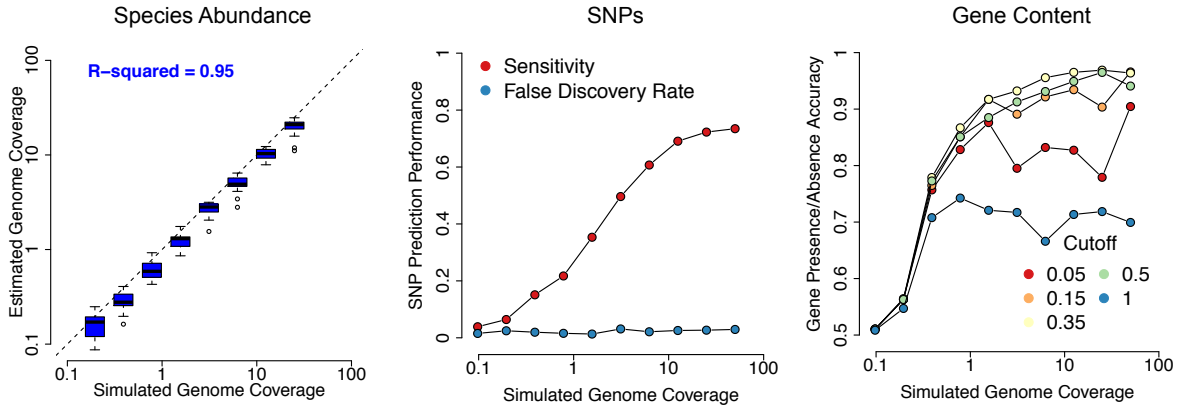

Figure 2.2 Shotgun simulations validate pipeline. To evaluate performance for each component of MIDAS, we analyzed 20 mock metagenomes composed of 100-bp Illumina reads from microbial genome-sequencing projects. Each community contained 20 organisms with exponentially decreasing relative abundance. We tested the ability of MIDAS to estimate species coverage and to predict genes and SNPs present in the strains of the mock communities compared to the reference gene and genome databases. Left) Species coverage is accurately estimated. Each boxplot indicates the distribution of estimated genome coverages across 20 mock communities for the top 8 most abundant species out of 20 analyzed. Middle) SNPs are detected with a low false discovery rate and good sensitivity when genome coverage is above 10x. Sensitivity = (# correctly called SNPs)/(# total SNPs); False Discovery Rate = (# incorrectly called SNPs)/(# called SNPs). Right) Gene presence-absence is accurately predicted when genome coverage is above 1x and a gene copy number cutoff of 0.35 is used. Accuracy = (Sensitivity + Specificity)/2; Sensitivity = (# genes correctly predicted as present)/(# total genes present); Specificity = (# genes correctly predicted as absent)/(# total genes absent).

Chapter 3

Mother-to-infant transmission of gut microbiome strains

In this chapter I use MIDAS to re-analyze stool metagenomes from 98 mothers and their infants [59] over the first year of life to assess whether gut microbiome strains are transmitted vertically (i.e. mother to infant) or not. Specifically, I develop a novel approach that utilizes rare SNPs to track gut bacteria from mother to infant. Based on this approach, I find that there is extensive vertical transmission of strains by 4 days after birth. However, by 1 year after birth, there is an emergence of late-colonizing bacterial strains in infants that are clearly distinct from those present in the mother. Interestingly, late colonizing bacteria contain spore-forming genes, whereas those that colonize early do not. Based on these results, I hypothesize that early colonizing bacteria are transmitted via direct physical contact between mother and infant, whereas late colonizing bacteria are transmitted via other sources in the environment.

3.1 Background

The gut microbiome is critical for many important processes. While infants are not sterile at birth [60], the gut microbiome is largely acquired after birth. Despite many infant gut microbiome studies to date [59, 61-65], it is not clear from what sources gut bacteria are acquired and the extent to which they are transmitted vertically (i.e. from mother to infant) or from other sources in the environment, including unrelated individuals. An understanding of vertical transmission is critical for determining the extent to which the microbiome – and by extension microbiome-mediated phenotypes – are inherited. Furthermore, disruption of vertical transmission by various factors – including birth mode, antibiotic use, and diet – may lead to abnormal development of the gut microbiome.

Several culture-based studies have found evidence of vertical transmission [66-69]. In general, these studies have focused on specific bacterial taxa (e.g. *Bifidobacterium spp*.) that were culturable from both the mother and infant stool. It is not clear whether other species are vertically transmitted and whether transmission rates vary over time. More recently, several culture-independent studies have been conducted on the mother and infant gut microbiome [59, 61]. These studies found significant overlap in species between mothers and their infants over the first year of life and concluded that this was a result of vertical transmission. However, many species of gut bacteria commonly occur in the human population, while individuals harbor distinct strains [3, 46]. Therefore, species sharing may not necessary indicate a transmission route. Other studies have examined the development of the infant gut microbiome [65], including at the strain level [39, 62], but did not assess vertical transmission. Thus, the extent and timescale of vertical transmission and the stability of transmitted stains are currently not well established.

I hypothesized that I could apply MIDAS to a recently published dataset of 98 Swedish mother and their infants [59] to quantify mother to infant transmission of gut bacterial strains. Specifically, I planned to use discriminative SNPs that are not shared between unrelated individuals to track transmission of gut bacteria to infants (Figure 3.1). Because these bacterial SNPs are extremely rare in the human population, their co-occurrence in mothers and infants would be strong evidence of vertical transmission.

3.2 Methods

I downloaded 391 stool metagenomes from the NCBI sequence read archive from a recently published dataset of Swedish mother and their infants [59]. Each metagenome contained an average of 40 million 100-bp sequencing reads. Mothers were sampled 1x at 4 days after birth, while infants were sampled at 4 days, 4 months, and 12 months after birth. Of the 98 infants, 83 were delivered vaginally and 15 via cesarean section.

Each metagenome was run through the MIDAS pipeline to estimate the relative abundance of the 5,952 reference species. I found an average of 56.6 species detected per metagenome. The number of shared species was computed between all mother-infant pairs, where a shared species is defined as a species with >1 mapped read to >=1 marker gene in both samples. Additionally Bray-Curtis dissimilarity was used to estimate the species-level compositional similarity of mother and infant microbiomes.

Next, I used the MIDAS pipeline to align reads to genomes for species with >10x sequencing depth in each sample. I found an average of 5.3 species with >10x sequencing per metagenome. Next, I identified core-genome sites by comparing the read depth of each genomic site across metagenomic samples. I defined a core-genomic site as having >1x depth in >95% of samples.

I used rare SNPs referred to as *marker alleles* to detect transmission of gut microbiota from mother to infant with high specificity and sensitivity (Figure 3.1). A marker allele was defined as an allele of a SNP that was present in one individual (or group of related individuals) and absent from all unrelated individuals. In my analysis, unrelated individuals included other Swedish mothers (N=97 at 1 time point), infants from other mothers (N=97 at 3 time points), and American individuals from the Human Microbiome Project (N=123 at up to 3 time points) [70]. Only bi-allelic SNPs were considered. To reduce noise from sequencing and mapping errors, I called a SNP if it was supported by ≥3 reads and ≥10% allele frequency. To sensitively call SNPs, I only considered bacterial species with ≥10x sequencing depth in a metagenome. To accurately identify rare alleles, I only considered bacterial species that occurred in ≥10 individuals. Marker alleles were identified separately for each species of gut bacteria and compared between all pairs of samples.

3.3 Results

3.3.1 Maturation and diversification of the infant gut microbiome

I first used MIDAS to quantify the relative abundance of bacterial species in mother and infant stool metagenomes. As described more fully in section 2.2.1, species relative abundances were estimated by mapping reads to a panel of 15 universal single copy genes. In agreement with

Backhed et al. and previous work [59, 61, 64, 65], I found that bacterial species alpha diversity was lowest in newborns and increased over time, species beta diversity was highest in newborns and decreased over time, and samples clustered by host age based on Bray-Curtis dissimilarity between species relative abundance profiles (Figure 3.2 and 3.3). These results illustrate the normal development of the infant microbiome to an adult-like state.

I found a large number of shared species between infants and their mothers that increased over time as the infant microbiome became more diverse (Figure 3.3). However, I found nearly as many shared species between mothers and *unrelated* infants (Figure 3.3) where there was no direct transmission. This indicates shared species are a result of the infant microbiome maturing to an adult-like state rather than an indication of mother to infant transmission.

3.3.2 Mother-infant strain similarity is high 4 days after birth but rapidly decreases over time

I used rare SNPs (i.e.) in order to detect transmission of gut microbiota from mother to infant with high specificity and sensitivity (Figure 3.1). I reasoned that a bacterial population from a mother's microbiome could harbor one or more SNPs would uniquely discriminate it from populations of the same species found in other unrelated individuals. Co-occurrence of these rare SNPs between mothers and their own infants would be strong evidence of vertical transmission. For the remainder of this chapter, I refer to these SNPs as *marker alleles*, since they serve as markers of individual strain populations.

Applying this procedure to the Backhed et al. dataset yielded a total of 278,924 marker alleles found in gut microbiomes the 98 mothers (217 ± 520 marker alleles per mother-species pair) (Figure 3.4). Together, this indicates that there are many SNPs in the microbiome that can be used for unique identification of a host in a cohort of ~300 individuals.

Next, I asked whether the large number of marker alleles found in mother microbiomes were shared with their infants, which would be evidence of vertical transmission. Towards this goal, I computed the fraction of shared alleles between mother-infants pairs for each species using the Jaccard Index. A value of 100% indicates that all marker alleles are shared for a species and a value of 0% indicates that no marker alleles are shared.

At 4 days after birth, allele sharing was remarkably high between mothers and their infants (mean=72%), indicating extensive mother-to-infant transmission of gut bacteria shortly after birth (Figure 3.5). However, over time there was a precipitous decrease in marker allele sharing (Figure 3.5). Across all species, allele sharing decreased from 72% at 4 days, to 58% at 4 months, to 35% at 12 months (Pearson's P=6e-17, r=-0.33). Thus, while the species level composition of mothers and infants converged over time, the strain level composition actually diverged.

To contextualize these results I performed two experiments. First, I identified and tracked marker alleles in healthy individuals from the HMP over a time period of 300-400 days. Across all individuals, I found high marker allele sharing (mean=77.0%) and no significant decrease in allele sharing over time (Figures 4.5 and 4.6), which indicates that strains are quite stable over

time in healthy adults and agrees with previous work [3, 71]. Therefore, the decrease in strain similarity between mothers and infants over time is not due to normal turnover of strains that occurs in healthy adults.

Next, I identified and tracked marker alleles in unrelated healthy individuals from the HMP. Because it is unlikely that unrelated individuals harbor the same strains, marker allele sharing between these individuals is expected to be close to 0%. For this experiment I separated individuals into training and testing groups (9:1 ratio); the training group was used to discover marker alleles and these alleles were then identified and tracked between individuals in the testing group. As expected, I found low allele sharing (mean=1.01%) between metagenomes from unrelated individuals (Figure 3.5), indicating that the high allele sharing between mothers and infants at 4 months is unlikely due to chance alone.

3.3.3 Early and late colonizers have distinct transmission patterns

I hypothesized that decreasing strain similarity between infants and mothers was due to late colonization of the infant gut by new species derived from the environment. Alternatively, early colonizing strains could lose marker alleles over time due to mutation and/or selection. Another possibility is that early colonizing strains are later lost and replaced by other strains not found in the mother.

To test these hypotheses, I compared vertical transmission patterns between microbiome species at the three different time points (Figure 3.7). Here I quantified a vertical transmission event as marker allele sharing >5% based on the distribution of allele sharing between related and

unrelated individuals (Figure 3.5), although the results were largely consistent at different cutoffs (Figure 3.8). Based on this analysis, I found that strains of *Bacteroides*, *Parabacteroides*, and *Bifidobacteria* were commonly vertically transmitted by 4 days. The vast majority of these persisted in the infants at 4 months (49/54 mother-infant pairs with >5% marker allele sharing) and at 12 months (47/51) indicating that the decrease in strain similarity over time is not due to loss of early colonizing strains. In contrast, at 4 months and 12 months after birth, many new species of bacteria appeared in the infants that were distinct from the mothers at the strain level. These included many Firmicutes like *Blautia*, *Faecalibacterium*, *Clostridium*, and *Ruminococcus*.

To further test my hypothesis, I compared the colonization timing of all species with the probability that they were vertically transmitted (Figure 3.9). In support of my hypothesis, I found a strong positive correlation between the relative abundance of a species in the infant at 4 days and vertical transmission ($P=9x10^{-14}$). In conclusion, early colonizing bacteria are often vertically transmitted whereas late colonizing bacteria are rarely vertically transmitted and likely derive from the environment.

3.3.4 Late colonizers are enriched for spore-forming bacteria

If late colonizing bacteria are derived from the environment, then they should be spore formers. Sporulation is a process in which a bacterium forms an endospore to protect it from environmental factors. Sporulation is particularly critical for anaerobic gut bacteria to survive outside of the host due to their sensitivity to ambient oxygen [18].

To address this question, I used data from a recent study, Browne et al. 2016 [18], to classify gut bacteria from this study as either spore-formers or non-spore formers. In their study, Browne et al. computed a sporulation score for various gut bacteria based on the presence/absence of 66 genes (scores > 0.4 indicate likely spore-formers). Strikingly, I observed high sporulation scores for species with low vertical transmission rates, supporting the hypothesis that these organisms colonize the infant from the environment (Figure 3.10). For example, 6/7 of species with low vertical transmission rates (<25% strain sharing) were predicted spore-formers. One exception to this pattern was the facultative anaerobe *E. coli*, which can survive in the environment without undergoing sporulation. In contrast, I observed low sporulation scores for species with high vertical transmission rates. For example, 10/11 of species with high vertical transmission rates (>65% strain sharing) were predicted non-spore-formers. These results suggest that direct physical contact between hosts (i.e. mother and infant) is required for transmission of these species.

3.3.5 Vertical transmission rates differ by birth mode

Interestingly, I found no species present with ≥10x sequencing depth in 15 C-section born infants and their mothers, and therefore had no way to assess transmission in these individuals. The lack of abundant shared species likely reflects lower vertical transmission of the mother's gut microbes, but I cannot directly test that hypothesis with the available data. However, C-section born infants did share species at 4 months and 12 months after birth and had fewer vertically strains transmitted strains compared to vaginally born infants at four months (chi-square $P=5\times10^{-8}$, 3/14 versus 128/149 shared species with >5% marker allele sharing) and to a lesser extent at 12

months (chi-square P=0.06, 13/34 versus 159/279). These results indicate that birth mode may affect where our gut microbiota is derived from.

3.4 Conclusions & Discussion

To illustrate the utility of MIDAS, I analyzed stool metagenomes from a recently published study of 98 mothers and their infants over one year [59] and used rare SNPs to track transmission of strains between hosts. Based on this analysis, I found extensive vertical transmission of early colonizing bacteria, which largely persisted in the infant for one year. While significant attention has been paid to transmission of *Bifidobacterium* spp. [66-68], I found high transmission rates for many *Bacteroides* spp. I also found that late colonizing bacteria, including *Blautia*, *Ruminococcus*, *Eubacterium*, and *Facelibacerium,* were rarely transmitted from the mother. Instead the mother was colonized by a different strain of these species. Comparing these species to a recent study of sporulation in the human gut [18], I found that late colonizers tended to be spore-formers capable of surviving in the environment, whereas early colonizers were non-spore-formers. Together, these results suggest that only certain taxonomic groups of bacteria may be vertically inherited, while others are acquired from the environment. My results build upon previous infant microbiome studies [59, 61, 62, 65] by showing that early and late colonizing species likely derive from different sources, which may be linked with their ability to form spores and survive in the environment. When the same metagenomes were analyzed at the species level, these patterns of transmission were missed, and a false signal of increasing transmission over time was detected due to convergence of the infant microbiome towards a more diverse and adult-like species profile.

My analysis of mother-infant strain sharing leaves a few questions unanswered. One intriguing issue is the source of the strains that colonize the infant but are not present in the mother's stool microbiome at 4 days after birth. It is possible that some strains colonize the mother's gut later in the year and are then passed along to the infant, though this is unlikely based on the temporal stability of strains in the adult microbiome. The new strains could also derive from other sites on the mother's body, such as skin and breast milk, other people, food, or the environment. One caveat of this analysis is that I did not distinguish which strains were transmitted to the infant from the mother in cases where mothers harbored multiple strains. Instead, I treated the transmission events as binary, whereby a transmission was defined as at least one strain being transmitted. It would be interesting to explore transmission as a quantitative variable in future work, including elucidating how the strain composition and genetic diversity of bacterial populations change as they are passed from mother to offspring and potentially undergo bottlenecks and selection.

3.5 Figures

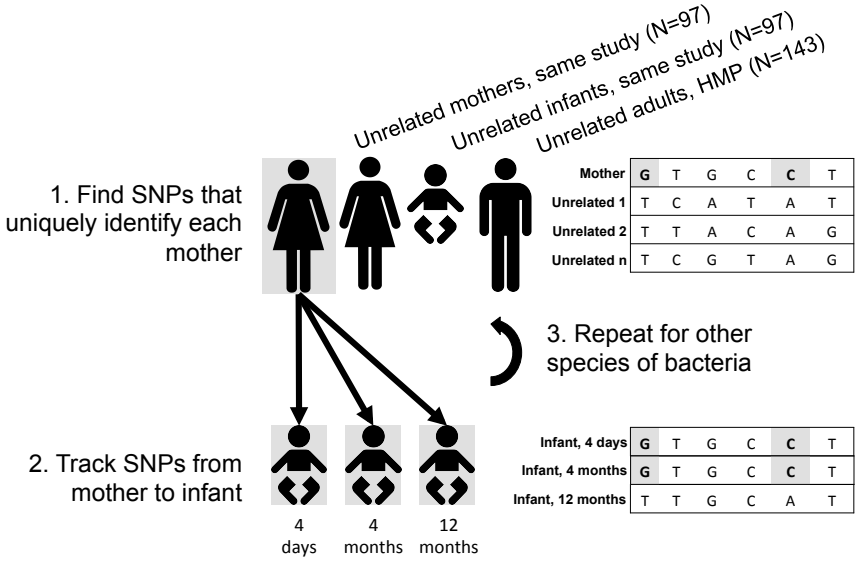

Figure 3.1 A SNP-based strategy for tracking strains between mothers and their infants. In the example, the G and C alleles at positions 1 and 5 in the genome of one microbiome species uniquely discriminate the mother from other unrelated individuals. These alleles are shared with the same species found in the mother's infant, indicated a shared strain and likely transmission route.

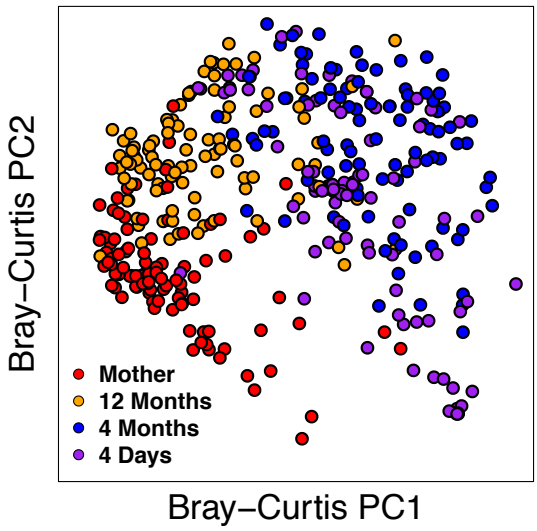

Figure 3.2 Principal coordinate analysis of Bray-Curtis dissimilarity between species relative abundance profiles of stool samples from mothers and infants at 4 days, 4 months, and 12 months following birth. Species composition of infant microbiomes is most similar to mothers at 12 months.

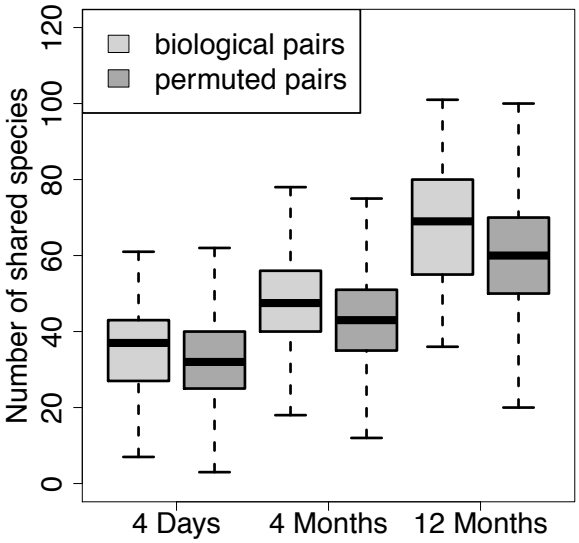

Figure 3.3 The number of shared species increases over time between mothers and their own infants. This pattern for biological mother-infant pairs is similar to that of unrelated mothers and infants (permuted pairs).

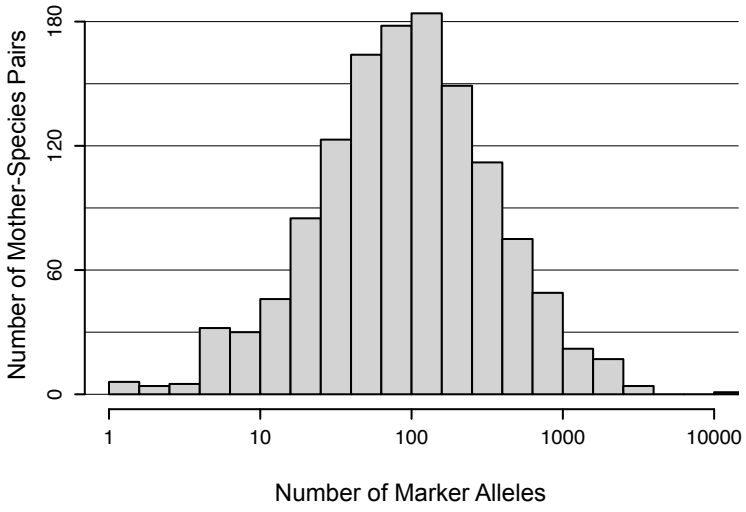

Figure 3.4 Distribution of the number of marker alleles found in mothers per species. On average there were 217 ± 520 marker alleles per mother-species pair.

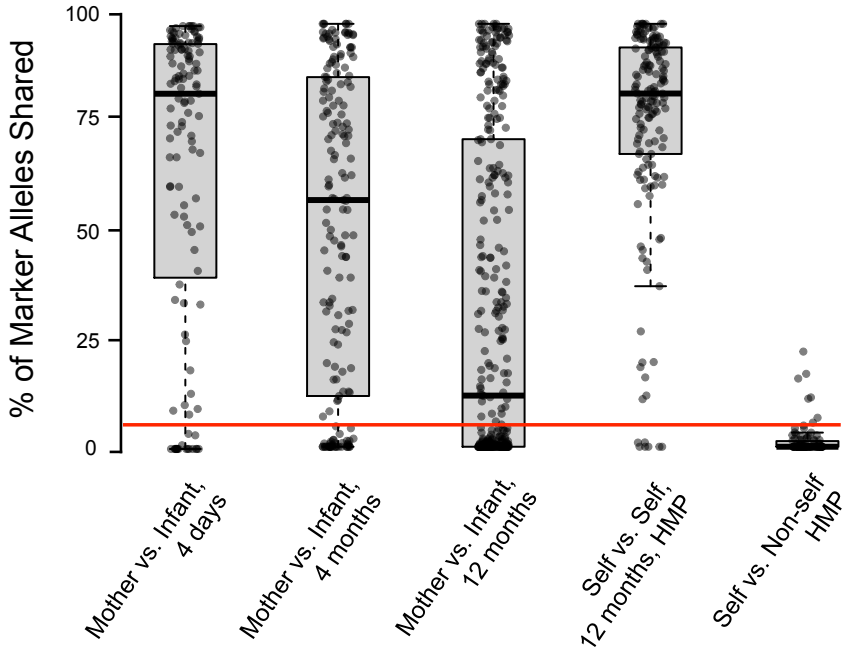

Figure 3.5 Percent of marker alleles shared between mothers and infants at 4 days, 4 months, and 12 months after birth. Each point indicates one species found in a mother and infant. Additionally plotted is allele sharing between the same healthy adults over 300-400 days (self vs. self) and healthy unrelated individuals (self vs. non-self) to provide additional context. The red horizontal line at 5% marker allele sharing defines the cutoff for determining a vertical transmission event.

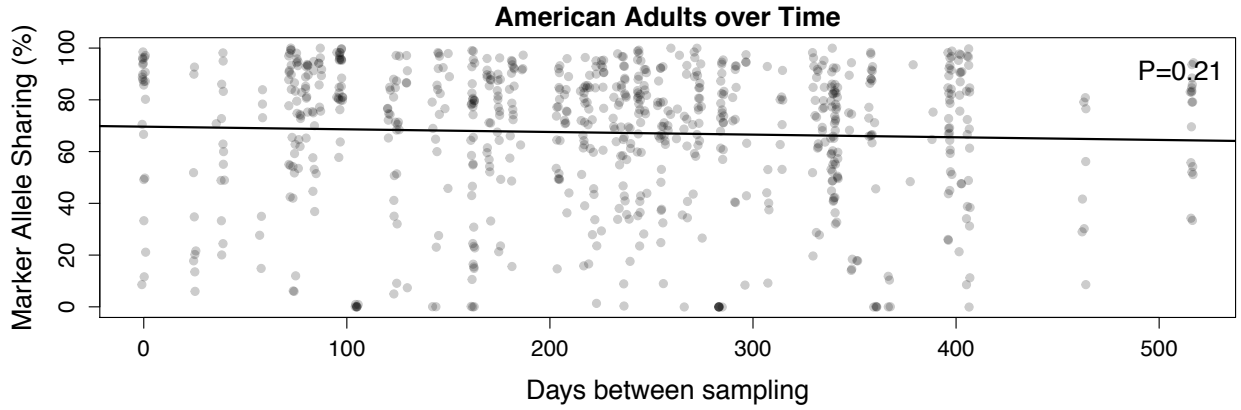

Figure 3.6 Allele sharing for gut microbiome species between different samples from the same healthy adults over time.

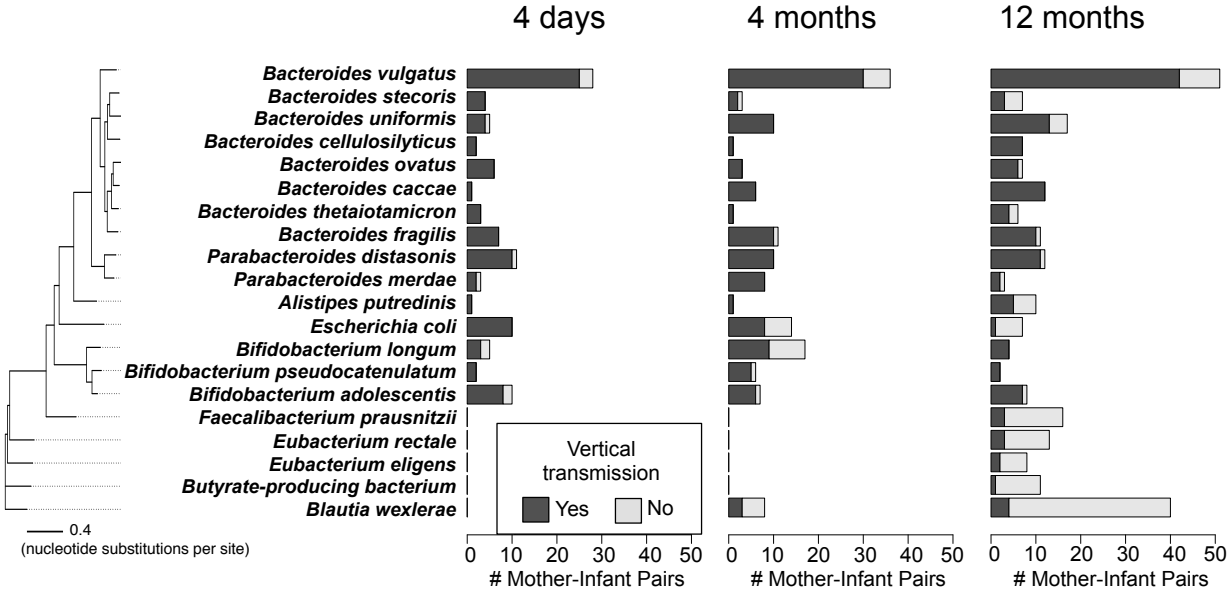

Figure 3.7 Vertical transmissions for bacterial species across mother-infant pairs at three time points. The 20 species with the greatest number of high-coverage mother-infant pairs are shown. A vertical transmission is defined as >5% marker allele sharing between mother and infant. The phylogenetic tree is constructed based on a concatenated DNA alignment of 30 universal genes and shows that phylogenetically related species have similar transmission patterns.

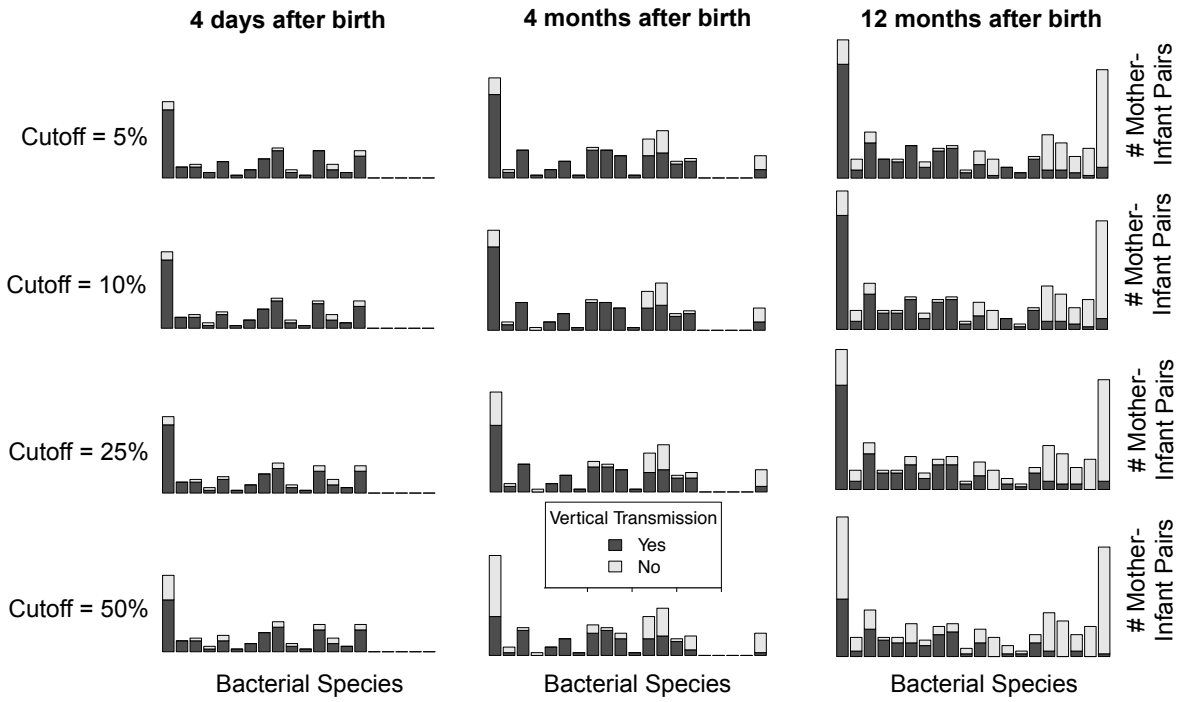

Figure 3.8 Vertical transmission patterns are robust to the marker-allele sharing cutoff used for defining transmission events.

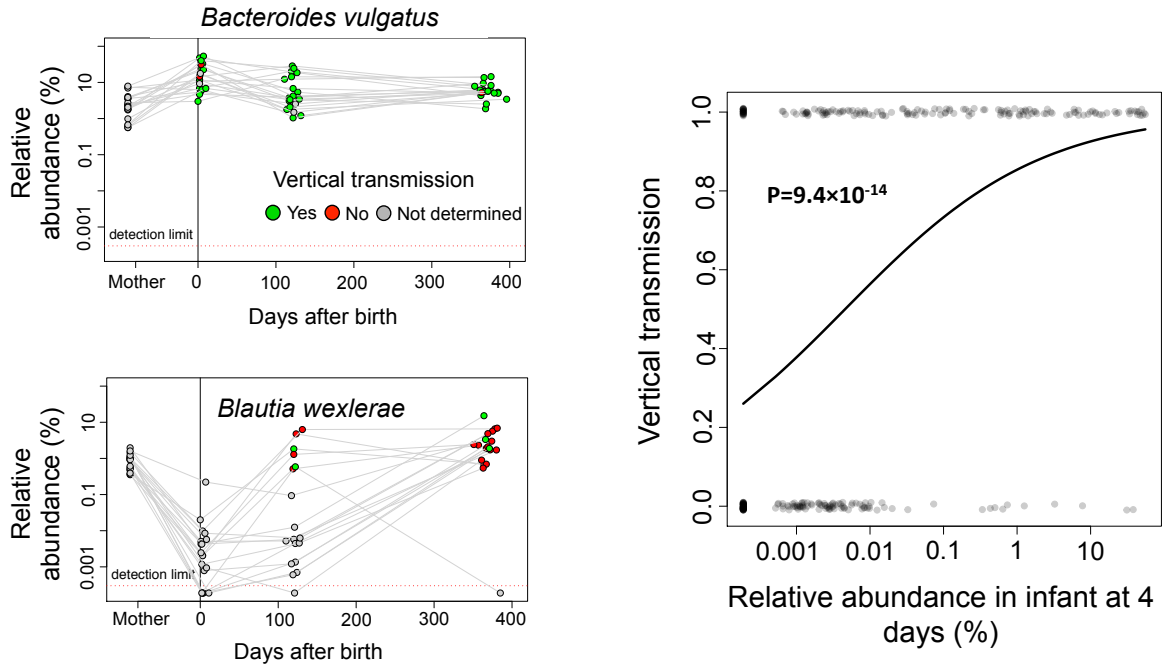

Figure 3.9 Colonization timing is correlated with vertical transmission. The dynamics of two species is shown on the left. *Bacteroides vulgatus* is an early colonizer that is vertically

transmitted and maintained over time. *Blautia wexlerae* is a late colonizer that is not vertically transmitted. On the right, colonization timing is plotted against vertical transmission. The horizontal axis indicates the relative abundance of bacterial species at 4 days. The vertical axis indicates whether a strain of the species was transmitted from the mother (y=1) or not (y=0) at 12 months. The curve is a logistic regression fitted to data points.

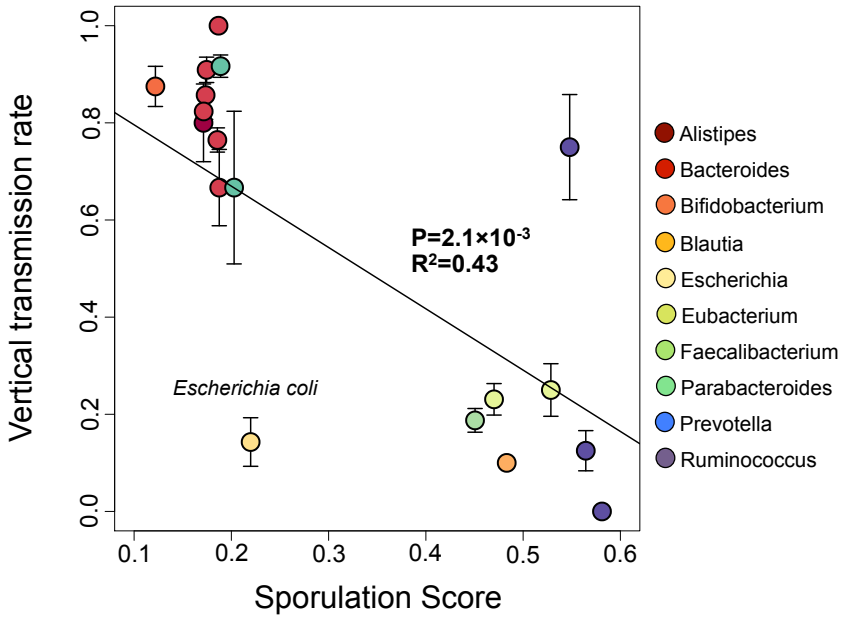

Figure 3.10 Species with low vertical transmission rates are predicted to be spore-formers with the ability to survive in the environment. Sporulation scores are genomic signatures of sporulation based on 66 genes. Error bars indicate one standard error in each direction. Only species with sporulation scores computed by Browne et al. and with ≥3 mother-infant pairs at 12 months are shown.

Chapter 4

Clonality of bacterial populations within and between host microbiomes

In this chapter I use MIDAS to reanalyze stool metagenomes from >500 stool metagenomes from the United States, Europe, China, Peru, and Tanzania and present the first global analysis of strain-level variation and biogeography in the human gut microbiome. On average, strain populations harbor 10x more nucleotide diversity between different individuals than within individuals. Many species are nearly clonal within individual hosts, but extremely diverse between hosts. I find that diversity is elevated in hosts from Peru and Tanzania that live rural lifestyles. For many, but not all common gut species, a significant proportion of between host genetic diversity is explained by geography. *Eubacterium rectale*, for example, has a highly structured population that tracks with host country, while strains of *Bacteroides uniformis* and other species are structured independently of their hosts. Finally, I discovered that the gene content of some bacterial strains diverges at short evolutionary timescales during which few nucleotide variants accumulate. These findings shed light onto the recent evolutionary history of microbes in the human gut and highlight the extensive differences in the gene content of closely related bacterial strains.

4.1 Background

Over the past several years, a number of studies have begun to shed light onto the extensive genomic variation of populations in the human gut microbiome. Some of these studies have focused on genomic variation *between* hosts. Greenblum et al. [46] and Zhu et al. [47] found that the gene content of bacterial populations varies significantly between individuals, while Schoissnig et al. [3] found extensive nucleotide diversity of microbiome species after pooling sequencing reads across metagenomic samples. Other studies have focused on genomic variation *within* hosts. Luo et al. [39] identified strains of the same species co-existing within the same host, and Kuleshov [72] used synthetic long reads to uncover mixtures of discrete haplotypes in one human stool sample.

While these studies have advanced our understanding of population genomics in the gut, they leave a number of questions unanswered. What is the relative magnitude of within and between host genomic variation of human gut bacteria? Are bacterial populations more clonal within individuals or is it common for individuals to harbor co-existing stains of the same species? What are the biological mechanisms that prevent or promote strain coexistence? Do bacterial populations differ significantly between individuals, and if so, do these differences track with host geography, disease, or other covariates? In this chapter, I use MIDAS to perform a large-scale population genomic meta-analysis of publicly metagenomes from human stool samples.

4.2 Methods

To understand global patterns of sequence variation within human gut species, I downloaded 372 publicly available metagenomes from the human gut. All metagenomes were downloaded from the NCBI Sequence Read Archive [57] and were identified with the aid of the SRAdb MySQL database [56]. These data included samples from healthy unrelated individuals from Denmark & Spain [25], the United States [21], China [24], Tanzania [23], and Peru [22] (Figure 4.1). I excluded all individuals with known diseases, including: diabetes, colorectal cancer, impaired glucose control, and inflammatory bowel disease.

Next, I developed an approach to use SNPs to estimate the intra-species nucleotide diversity, π, of bacterial populations within and between the metagenomic samples (Figure 4.2). Nucleotide diversity is defined as the expected number of variants per base pair between two randomly sampled genomes from a population. Conceptually, within-diversity is an indication of how clonal a bacterial population is within an individual, and between-diversity is an indication of how many differences there are between individuals.

In order to accurately estimate within and between host diversity, I focused on nucleotide variation within the core genome of each species. The presence or absence of SNPs in variable regions is a poor indicator of population heterogeneity if these regions are absent or duplicated in a subpopulation. For example, a region that occurs in only one of two subpopulations may appear to have abnormally low diversity whereas a gene that occurs at multiple copies may appear to have abnormally high diversity. To identify the core-genome of a species, I used the per-site distribution of read depth across metagenomic samples where the species was found at high depth. Specifically, I identified core-genomic sites that were covered by at least 15 reads in

> 95% of samples with >20x sequencing depth. Additionally, genomic sites with abnormally high read depth (>2x the per-sample mean) in >5% of samples were removed. This yielded an average of 1.2 million (range=0.14 to 2.74 million) core-genome sites per species.

I estimated genomic diversity at these sites using the following equation: $\pi = \frac{1}{n}\sum_i^n 2p_i q_i$, where p_i is the reference allele frequency at core-genomic site i, q_i=$1 - p_i$, and n is the number of core-genomic sites. This equation is adapted from [73] where it was used to estimate nucleotide diversity from human polymorphism data. In principle, π range from 0.0 to 0.5 when only considering bi-allelic sites. However, because two strains of the same species differ by a maximum of 5% by definition, the actual range is from 0.0 to 0.025 (i.e. 0 to 2.5%). To estimate π within hosts, I used reference allele frequencies estimated from individual metagenomic samples. To estimate π between hosts I used average reference allele frequencies across samples from distinct individuals (i.e. excluding replicates), which is akin to estimation of allele frequencies based on pooled metagenomes [3] but weights each population equally.

4.3 Validation

4.3.1 Population diversity estimates are robust to technical factors

Because MIDAS relies on read-mapping to reference genomes, I assessed the impact of different reference genomes on downstream estimates of species abundance and population genomics. Therefore, I picked four species with 1) multiple reference genomes in the MIDAS database and 2) at least 50 samples with >10x coverage in the stool metagenomes. For each species I picked two reference genomes: one "representative genome" as determined previously and one "alternate genome" which was picked at random. The selected species and genomes include: *F.*

prausnitzii (M21/2 and SL3/3), *A. muciniphila* (ATCC BAA-835 and Urmite), *E. rectale* (DSM 17629 and ATCC 33656), and *B. wexlerae* (DSM 19850 and DSM 17629). MIDAS was run using default parameters for each species using the representative and alternate reference genomes with a maximum of 20M reads per metagenome.

First, I compared alignment summary statistics obtained using the two different reference genomes and obtained nearly identical estimates of sequencing depth (R^2>0.99), (Figure 4.3). Next I quantified the SNP density and nucleotide diversity within the core-genome of each species using both reference genomes and obtained nearly identical estimates of SNP density (R^2=0.97) and nucleotide diversity (R^2=0.96), regardless of the reference genome used for read mapping (Figure 4.4). I also obtained similar estimated of pooled SNP density (mean difference of 1.3%) and nucleotide diversity (mean difference of 3.1%) across the four species. Together, these results indicate that estimates of population heterogeneity are not strongly influenced by the reference genome chosen for read mapping.

4.3.2 Current genomes adequately represent strains in the human gut

Next, I assessed whether the representative genomes used by MIDAS were representative of the genomes found in human gut communities. Reference genomes were covered by ≥40% of their length in 95% of sample-species pairs, by ≥60% in 92%, and by ≥80% in 52% (Figure 4.5). Further, >70% of low coverage reference genomes (<40% covered) corresponded to just four species: *Bacteroides rodentium* (ID:59708), *Escherichia fergusonii* (ID:56914), *Collinsella aerofaciens* (ID:61484), and *Collinsella sp* (ID:62205). For example, on average, *B. rodentium* recruited reads to only 9% of genomic sites but had >37x depth at these sites. These four species

were excluded from further analysis. In conclusion, the genomes used by MIDAS were generally good references for strains found in the stool metagenomes.

4.3.3 Minimum amount of data for unbiased estimates of population diversity

Species found in different metagenomes varied significantly in sequencing depth. I performed an experiment to explore the effect of sequencing depth on estimated within host nucleotide diversity. I down sampled the number of reads mapped to representative genomes of four difference species for 50 metagenomes. The number of reads sampled, n, ranged from 2 to 50. Reads were sampled without replacement. All core-genomic positions were down sampled to n reads. Genomic positions with less than n reads were discarded.

From this experiment, I found that nucleotide diversity strongly depends on sequencing depth (Figure 4.6). With fewer than 10 reads per site, diversity is consistently underestimated. This bias is most significant at 2 reads per site, where nucleotide diversity is underestimated by 50%. Thus, with at least 10-20 mapped reads per site, an unbiased estimate of nucleotide diversity is obtained and the diversity of populations found in different metagenomic samples can be directly compared.

4.4 Results

Next I applied MIDAS to the 372 human gut metagenomes with the goal of estimating the population diversity of the different species. To quantify population diversity and call SNPs, abundant species were first identified with ≥10x sequencing depth in any sample, which is the minimum depth required for sensitively calling nucleotide variants (Figure 2.2). Of these

abundant species, 273 were found in ≥1 sample, 114 in ≥1% of samples, 55 in ≥5%, 37 in ≥10%, and 13 in ≥20% (Figure 4.7). Many of these species occurred in multiple human populations.

By comparing diversity patterns within and between human hosts, I found many species with extremely low within-host population diversity but high between-host diversity (Figure 4.8). Across all species, there was 8x more nucleotide diversity for bacterial populations between different individuals compared to within individuals. However this ratio was as high as 30x for many species. Likewise, I found on average 12x more SNPs with a minor allele frequency ≥5% between individuals (mean SNPs Mb^{-1}=29,180) compared to within individuals (mean SNPs Mb^{-1}=3,652). Within and between-host diversity levels of microbiome species were generally consistent across different metagenomic studies and host continents, which suggests that they reflect consistent evolutionary, ecological, or demographic processes that occur in the gastrointestinal tract, such as growth rates [74], population bottlenecks [75] or priority effects during host colonization [76].

Many species, like *Ruminococcus bromii*, *Akkermansia muciniphila*, and *Bacteroides fragilis*, had consistently low within-host diversity levels (mean $\pi \leq$ 5e-4), whereas there were many more nucleotide differences when comparing these populations between hosts (Figure 4.9). For example, the core-genome of *Ruminococcus bromii* populations differed by an average of 1.2% between hosts, which was 24x greater than the average diversity levels within individual hosts. Together, these results indicate that populations of bacteria are commonly clonal within hosts and distinct between individuals.

Despite this overall trend, I found that intra-sample strain-level heterogeneity varied significantly across gut species (range=$4x10^{-4}$ to $1.1x10^{-2}$). While species like *Ruminococcus bromii* had consistently low intra-sample diversity (median π=$6.4x10^{-4}$), other species like *Prevotella copri* and *Faecalibacterium prauznizii* had very high levels of diversity (median π>$4.5x10^{-3}$) (Figure 4.9), which likely indicates the presence of multiple strains per host.

Next, I asked whether intra-sample strain-level heterogeneity varied systematically between host countries. Previous work has shown elevated species-level diversity in individuals living rural lifestyles in South America and Africa [22, 26, 64], but it is not clear whether these findings extend to the strain-level. To address this, I compared the distributions of intra-sample π for each species between host countries and found significant differences for 39% of species tested (Kruskal-Wallis q-value < 0.01). In particular, there was a trend towards increased intra-sample π in individuals from Peru and Tanzania, including *Ruminococcus bromii* (generally low intra-sample diversity) and *Facecalbacterium prauznizii* (relatively high intra-sample diversity)(Figure 4.10), indicating that host geography and/or lifestyle may influence strain-level diversity in individuals.

4.5 Conclusions & Discussion

I applied MIDAS to >300 faecal metagenomes from diverse human groups – including the United States, Europe, China, Tanzania, and Peru – and provide the first detailed characterization of global strain-level genomic variation of bacteria in the gut microbiome. By leveraging patterns of SNPs found in the core-genome of prevalent species, I find that bacterial strain-level diversity tends to be much greater between individuals than within individuals.

In particular, I found a number of species with high levels of diversity in the human population, but low diversity within individual hosts. Together this suggests that individual strains of these species colonize their hosts. A number of mechanisms could explain this result, including competitive exclusion, interaction with the host immune system, or direct competition or targeted killing between strains. While it is unclear which mechanism is the most common, previous work found that certain species of *Bacteroides* are resistant to colonization by members of the same species via competitive exclusion [76]. It's possible that this process may actually be quite common across gut species and not specific to the *Bacteroides* genus. In the future, this hypothesis could be experimentally tested in a gnobiotic mouse model.

In contrast, I found a number of other bacterial species with high diversity levels within individuals, suggesting that strains of these species are able to co-exist within individual hosts. One possible explanation is that strains of these species have functionFurther more is needed to gain a better understanding if these coexisting strains have differ

Furthermore, despite low levels of intra-sample diversity across many species, I still observed a large number of low-frequency segregating alleles in these species. It is unclear whether these variants are acquired during the lifespan of the host, as has been shown for certain pathogens [77], or whether they are inherited during colonization. Tracking genomic variation of microbiota between parent and offspring or other interacting individuals will shed light onto selection and demographic processes affecting bacterial populations during colonization.

Using a standard measure of population differentiation, F_{ST}, I found strong evidence for genomic differentiation – based on CNVs and SNVs – between host countries for most of the species examined in this study. In particular, I found that strains from China, Peru, and Africa were more differentiated from strains found in hosts from the Europe and the United States. It is currently unclear the extent to which these patterns are driven by adaptation to differences in the host environment, patterns of host migration, or a combination of factors. Additionally, only a minority of the genetic variation of bacterial species was explained by host geography. It is possible that increased travel between countries has led to mixing of bacterial populations that were at one point highly differentiated. Alternatively other environmental factors (e.g. diet, hygiene, genetics) that vary within human populations might explain the residual intra-species genetic variation. Lastly, these segregating genes and alleles could represent ancient genetic diversity that was present in ancestral bacterial strains and has been maintained across various human populations. As additional metagenomes are sequenced from African hosts, it may be possible to determine the proportion of current diversity that has arisen since migration out of Africa.

Finally, I find that gut communities from Tanzania and Peru are different from other gut communities in several important ways. They have elevated levels of novel species, greater species and strain-level level diversity, and more functionally and phylogenetically diverged strains. Because the hosts in Tanzania and Peru live rural hunter-gather and traditional agricultural lifestyles, their novel species and strains may represent ancestral taxa that have been lost from the microbiomes of humans living industrialized lifestyles. Furthermore, the increased species and strain level diversity might be explained by less hygiene and reduced antibiotics

exposure in these rural populations. As additional metagenomes are sequenced from diverse global populations, it will be possible to disentangle the degree to which these patterns are specific to particular lifestyles and/or geography. If hosts living more traditional lifestyles and with more limited access to health care indeed have more diverse microbiomes regardless of host migratory patterns, further exploration of strain-level variation with MIDAS may shed light on the "hygiene hypothesis" and the role that loss of ancestral microbiome diversity may play in the rise of autoimmune disease in industrialized countries.

While humans have been living with microbes throughout our evolution, we are just beginning to understand global patterns of variation. An understanding of how bacterial strains vary within and between hosts provides a necessary foundation for future studies and for linking the microbiome to human health and disease.

4.6 Figures

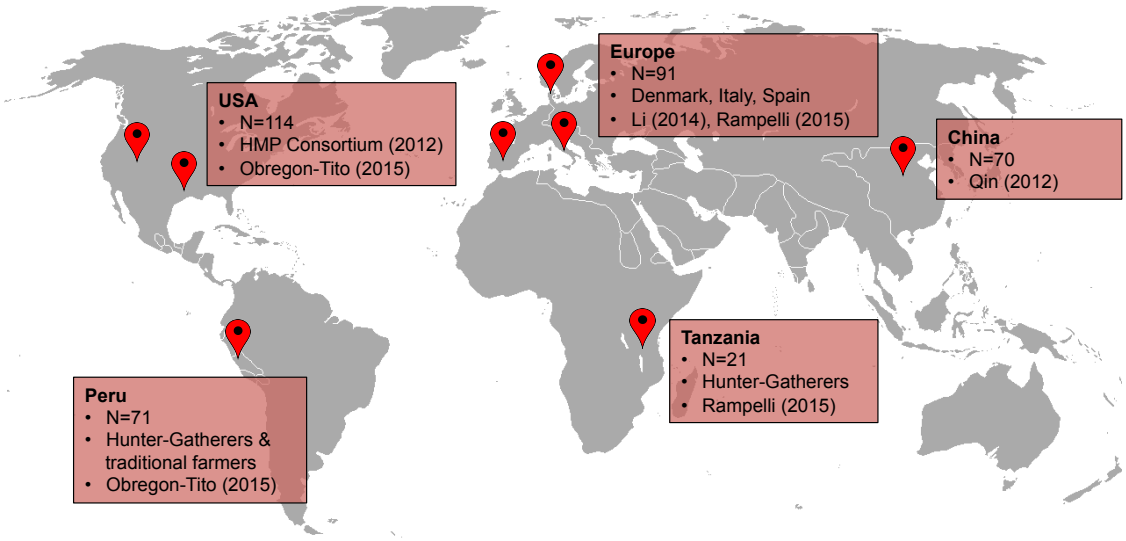

Figure 4.1 Information and sampling location of human gut metagenomes

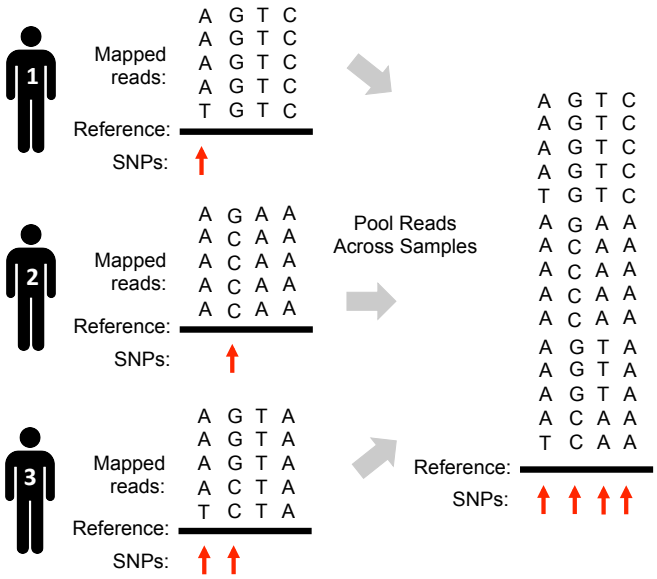

Figure 4.2 The general approach for estimating within and between host genomic diversity of individual populations in the gut microbiome. Within-host diversity is estimated from individual metagenomes, whereas between-host diversity is estimated by pooling reads across metagenomic samples. All estimates of diversity and SNP density are based on core-genomic regions that are consistently present across metagenomic samples.

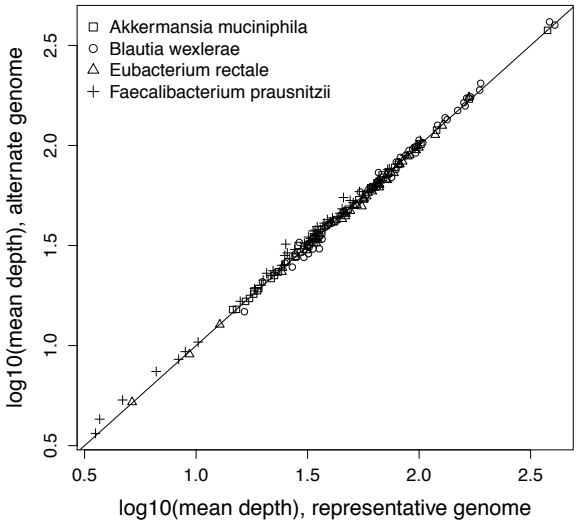

Figure 4.3 Average read depth is consistently estimated regardless of the reference genome used for read mapping. Each point represents one species-sample pair. Different species are indicated by point shape. Only genomic positions with non-zero depth were included in the average.

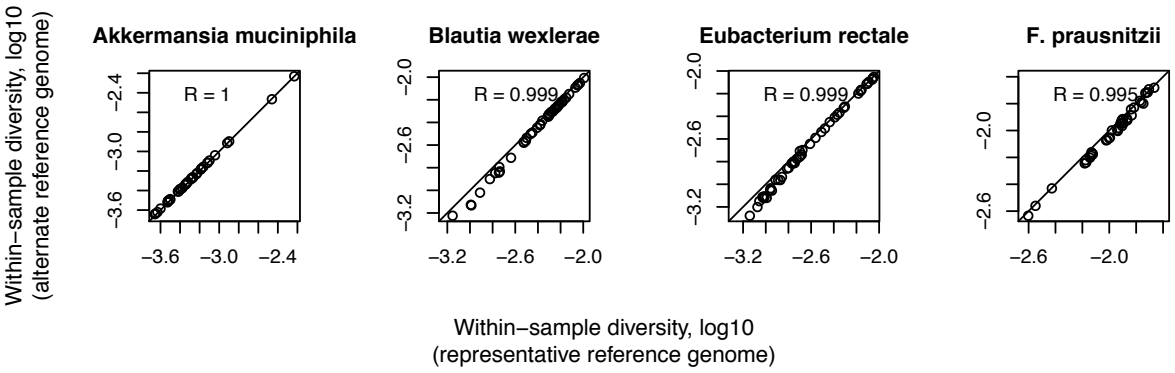

Figure 4.4 Within host diversity is consistently estimated regardless of the reference genome used for read mapping. Each point represents one species-sample pair. Different species are indicated the plot title.

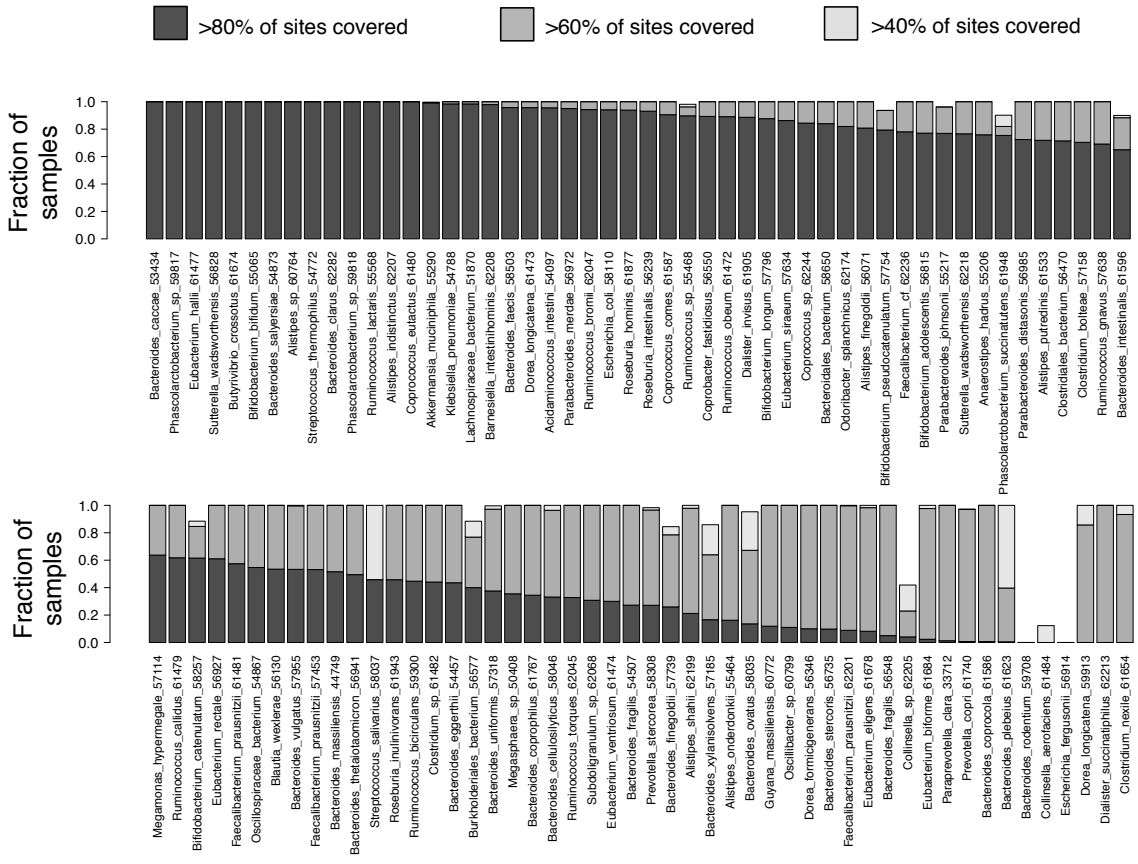

Figure 4.5 Representative genomes of most species are covered by at least 40% in nearly all metagenomic samples where the species is present. The horizontal axis indicates different species. The vertical axis indicates the fraction of metagenomes where the species is present. The bar color indicates the percent of the representative genome that is covered by at least 1 read. This result indicates that representative genomes are suitable for studying the population genomics of human gut bacteria.

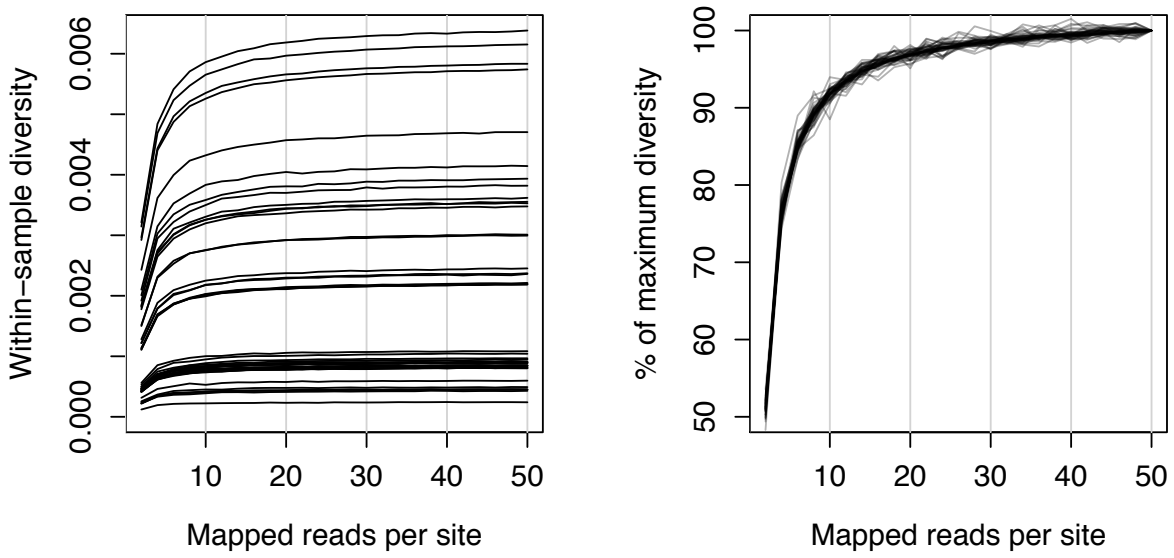

Figure 4.6 Minimum read depth for unbiased estimates of within-host nucleotide diversity.

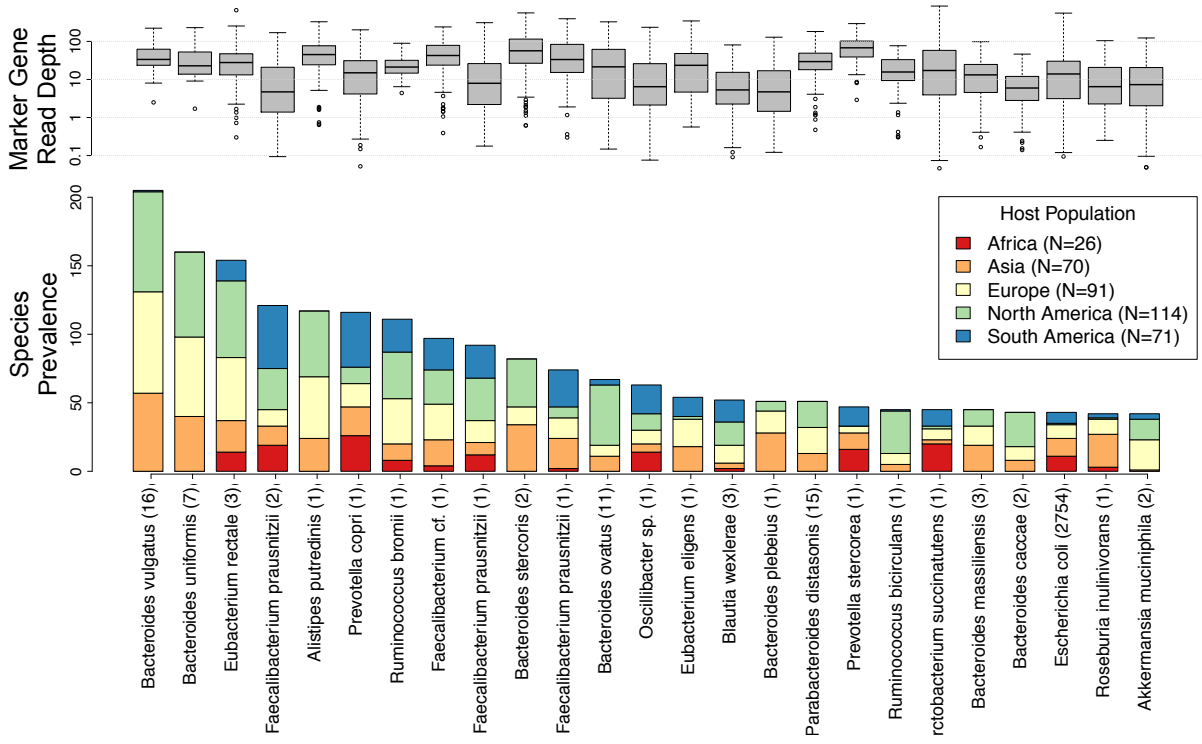

Figure 4.7 Prevalence and sequencing depth of 25 human gut species across 372 human gut metagenomes from 5 continents.

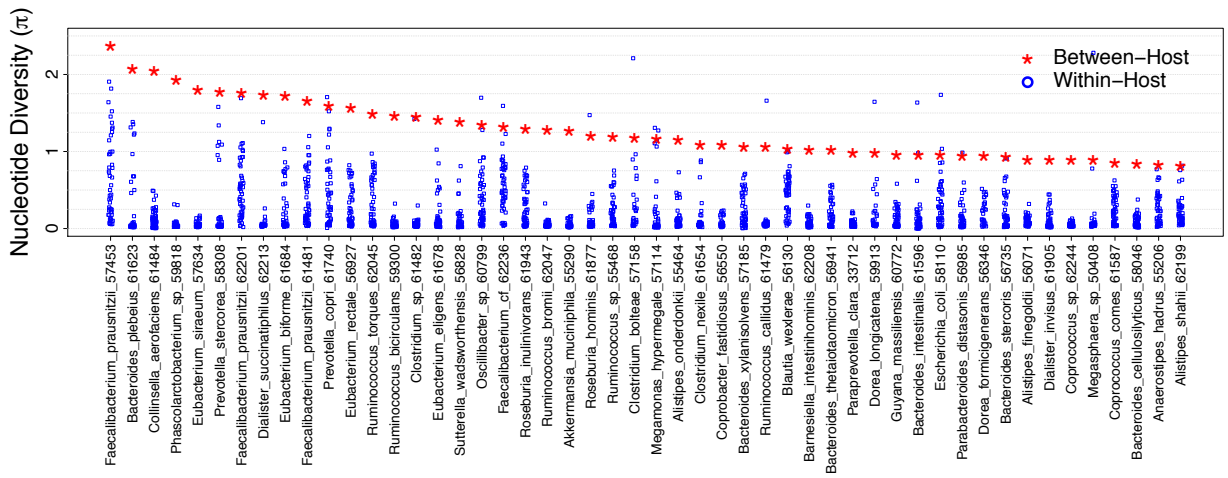

Figure 4.8 Within and between host nucleotide diversity for 50 species.

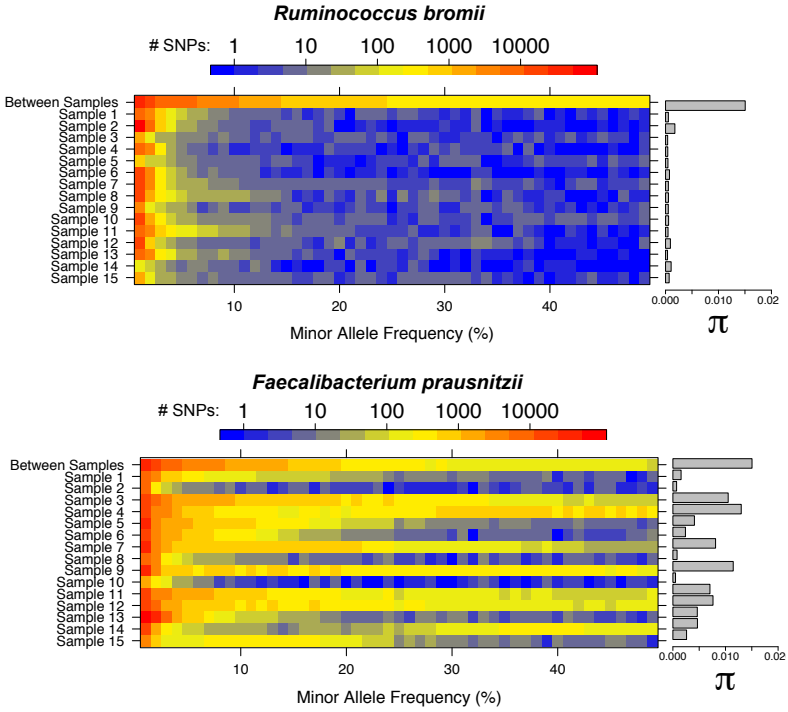

Figure 4.9 Allele frequency spectra for two species. Each row indicates the number of SNPs at different minor allele frequencies (0.0 to 0.5) for 15 randomly selected individuals. The top row indicates the number of SNPs at different frequencies after pooling reads across all samples. On the right nucleotide diversity is indicated for the same samples.

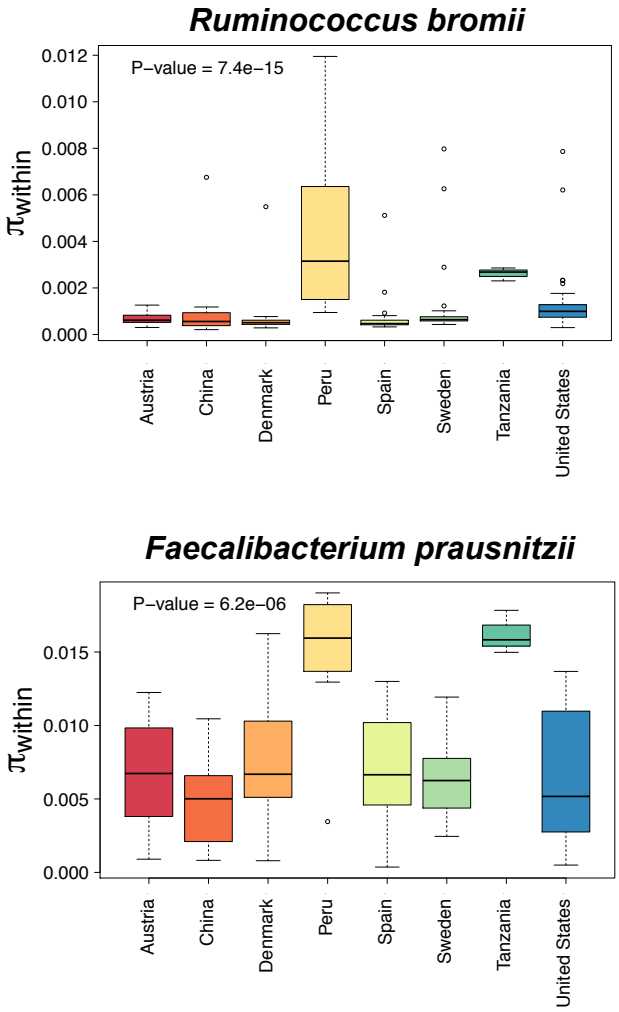

Figure 4.10 Within host nucleotide diversity is higher hunter-gatherers from Peru and Tanzania.

Chapter 5

Global population structure of prevalent marine bacteria

In this chapter I use MIDAS to quantify strain level gene copy number variation for 30 prevalent marine bacterial species across ~200 globally distributed ocean metagenomes [29]. I quantify the population structure of each species based on the presence or absence of genes between sampling stations. Using principal component analysis of gene presence absence, I show that many bacterial species have population structure that correlates with geographic location. In contrast other species have populations that are structured instead by depth.

5.1 Background

Marine microorganisms are ubiquitous and play key roles in biogeochemical processes. To explore marine microbial diversity, the *Tara* Oceans expedition collected and performed metagenomic shotgun sequencing on seawater samples from over 200 sampling locations in all of the world's oceans. In analysis of this data, Sunagawa et al. found that overall community composition mostly driven by temperature rather than other environmental factors or geographic location [29].

Within the marine microbiome, there are many widely distributed clades including *Procholococcus [78]*, *Pelagibacter* [79], and SAR86 [80]. However, recent work has shown that co-existing *Procholococcus* genomes can differ significantly in their gene content [78]. It is currently unclear how marine microbial populations differ in gene content across the global ocean, and whether these differences are structured by geography, depth, or other environmental variables. For example, while overall community structure could be driven by temperature, individual populations could be responding to different environmental stimuli.

Previously, Pritchard et al. conducting a principal component analysis (PCA) of human genetic variation data to uncover the geographic structure of human populations [81]. In this chapter, I follow a similar approach, in which I conduct a PCA of gene content variation for 30 widely distributed and abundant marine species.

5.2 Methods

To assess the global population structure of marine bacteria, I analyzed 198 shotgun metagenomes collected from the *Tara* Oceans expeditions that corresponded to prokaryotic size fractions. I utilized up to 100 million reads per metagenome and analyzed only one sequencing replicate per sample. I used MIDAS to quantify the relative abundance of the 5,952 reference species, and based on these results identified 30 species that occurred at >3x sequencing depth in the greatest number of metagenomes (Figure 5.1). The least prevalent species was found in 23% of metagenomes. Next, I used MIDAS to quantify the gene content of these species across metagenomic samples. Reads were mapped to the pan-genome database, and reads with <94% alignment identity were discarded. Mapped reads were used to compute the coverage of genes clustered at 95% identity. Gene coverages were normalized by the coverage of 15 universal single copy genes to estimate gene copy numbers. I estimated gene presence-absence by thresholding the gene copy numbers, whereby any gene with a copy number <0.35 was considered to be absent.

To uncover population structure, I performed a principle component analysis of the gene presence-absence matrix for each species. To assess the relationship between gene content and geography, I first quantified the PCA distance and geographic distance between metagenomic samples for each species. PCA distances were computed using the Euclidian distance between samples based on the first two principle components. Geographic distances were computed using the Great circle distance with the R package geosphere [82]. Mantel tests were computed using the R package vegan [83] to correlate the PCA distances to the geographic distances. Up to one million permutations were performed to assess significance.

5.3 Results

To explore the extent of population structure across different marine bacterial species on a global scale, I used MIDAS to quantify population genomics in 198 marine metagenomes from 66 stations along the *Tara* Oceans expedition [29].

First, I used MIDAS to estimate the sequencing depth of bacterial species in the metagenomes to identify marine species where gene content could be reliably estimated. Previously, I found that these metagenomes were dominated by novel organisms with no sequenced reference genome (Figure 1.5). Despite this, I found a number of species that occurred in >20% of metagenomes with at >1-10x sequencing depth (Figure 5.1). Among these species were several members of the genera *Pelagibacter*, *Alteromonas*, *Synechococcus*, and *Marinobacter,* a large group of closely related *Prochlorococcus* species, and several unnamed *Alphproteobacteria* species. Reference pan-genome sizes for these species ranged from 1,047 and 1,311 genes in the streamlined genomes of SAR406 and SAR86 (each with 1 genome) to 6,427 genes in the largest *Prochlorococcus* genome cluster (N=26 genomes) and 7,819 genes for *Alteromonas macleodii* (N=4 genomes).

Using MIDAS, I discovered extensive variability of gene content for these prevalent species of marine bacteria across the ocean metagenomes. Across all species, I found an average of 318 genes that differed between samples, ranging from 144 genes in SAR86 to 700 in *Alteromonas marina*. I next quantified the percent of genes that were different between samples using the Jaccard index and found that on average 19% of genes differed between samples. This level of genomic variability was higher than the 13% reported for human gut communities [47], although

this may be due to methodological differences. Regardless, the estimate of 19% is almost certainly an underestimate of the true level of gene content variation between populations, because MIDAS cannot measure the variation of genes that are present in strains but absent from sequenced reference genomes.

To explore how this variation correlated with geography and sampling depth, I conducted a principal component analysis (PCA) of gene content for each bacterial species, as has been done to study the geographic structure of human populations using polymorphism data [81]. Strikingly, I found that the populations of many species clustered together by ocean region based on the first two principal components of gene content, regardless of sampling depth (Figure 5.2). For example, populations of one *Pelagibacter* species formed three discrete clusters corresponding to the Mediterranean Sea, South Atlantic Ocean, and South Pacific Ocean, and each cluster contained samples from different water layers. Similar results were obtained for many other species. Furthermore, I found that the population structure of the marine bacteria examined was highly consistent, regardless of the percent identity threshold used for defining pan-genome gene families (75-99% identity).

To evaluate the extent of gene content biogeography across species, I computed the correlation between PCA distances and geographic distances and found significant distance-decay in gene content for the majority of species tested (Figure 5.3). Furthermore, this pattern was observed both in samples from the surface water layer and the deep chlorophyll maximum layer – the majority of species I examined were not found in the mesopelagic water layer. A previous study found season to be a major driver of biodiversity patterns in the global ocean [84]. To explore

whether season or other environmental variables were associated with strain-level population structure, I compared correlations of the first principal component of gene content (PC1) with geographic location and environmental variables. For 20/30 species tested, longitude (17/30) or latitude (3/30) was the strongest predictor of gene content, and each explained a significant proportion of gene content variation (22% and 8% on average). In contrast, day length (an indicator of season) explained less variation (4% on average) and was the most predictive covariate for only one species.

A few species showed relatively little geographic structure. Instead they had gene content variation that correlated with depth or marine layer (Figure 5.2). The most striking example of this was an unnamed *Alphproteobacteria* species which contained two genomes in the database obtained via single-cell sequencing [8]. This species was predominantly found in the mesopelagic layer (below 200m) and increased in relative abundance with decreasing depth. Looking only at mesopelagic samples, I found that the first principal component of gene content (PC1) was strongly correlated with depth (R^2=0.59), suggesting little mixing of strains across depth. When I included samples from all marine layers, I found that samples from the mesopelagic and epipelagic zone formed separate clusters based on gene content and there was still a strong correlation between PC1 and depth (R^2=0.57). These results could indicate that the populations at different depths contain genes for adaptation to the range of temperatures and nutrients across which this species is found. Supporting this hypothesis, I found hundreds of functions and pathways with gene-copy-numbers that were significantly correlated with depth.

Together, these results expand upon and even contradict patterns of marine bacterial biogeography observed at the species level. In particular, gene content analysis reveals that abundant and prevalent species are not ubiquitous at the strain level. Instead they show significant structure across geographic regions.

5.4 Conclusions & Discussion

To explore bacterial population structure using gene content, I applied MIDAS to metagenomes from the *Tara* Oceans expedition. I found a number of prevalent and abundant bacterial species, which shows that the method can be applied to different environments. Based on these results, I found that the gene content of many species in the epipelagic water layer (0-200m) was structured geographically. This contrasts with previous work at the species level, which found that depth and temperature were the strongest predictors of community structure [29]. However, the gene content of other species found in the mesopelagic layer (200-1000m) were structured by depth. As more genomes are sequenced from marine ecosystems, it should be possible to determine how generalizable these patterns are. Additionally, future work is needed to understand the extent to which these gene-level patterns are driven by adaptation to different environments in the ocean, or due to neutral processes, like genetic drift and/or migration.

5.5 Figures

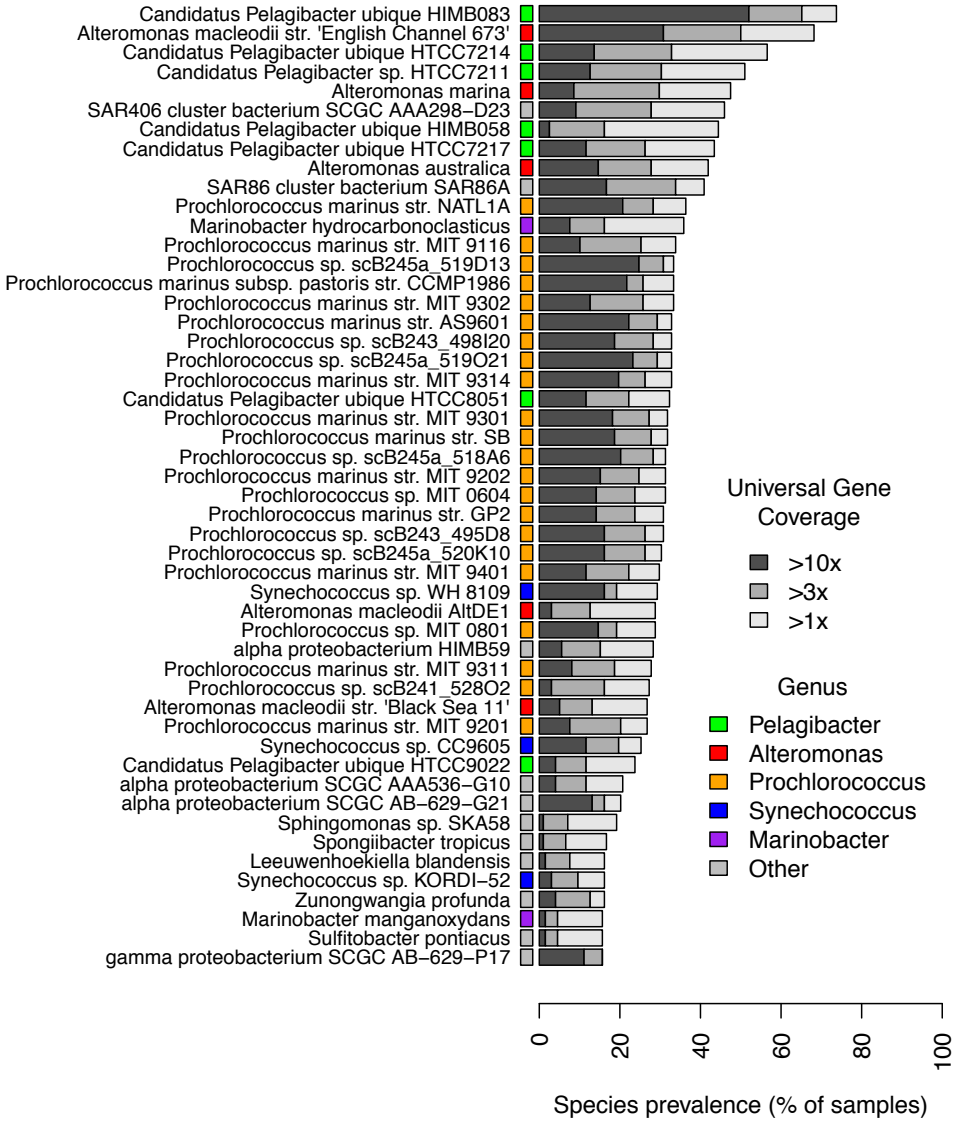

Figure 5.1 Prevalent bacterial species surveyed by the *Tara* Oceans expedition across 198 ocean metagenomes. Latin names of species are indicated on the vertical axis. In cases where multiple species had the same Latin name, the full name of the representative genome is shown. Many marine species have sufficient sequencing depth and prevalence for population genetic analyses.

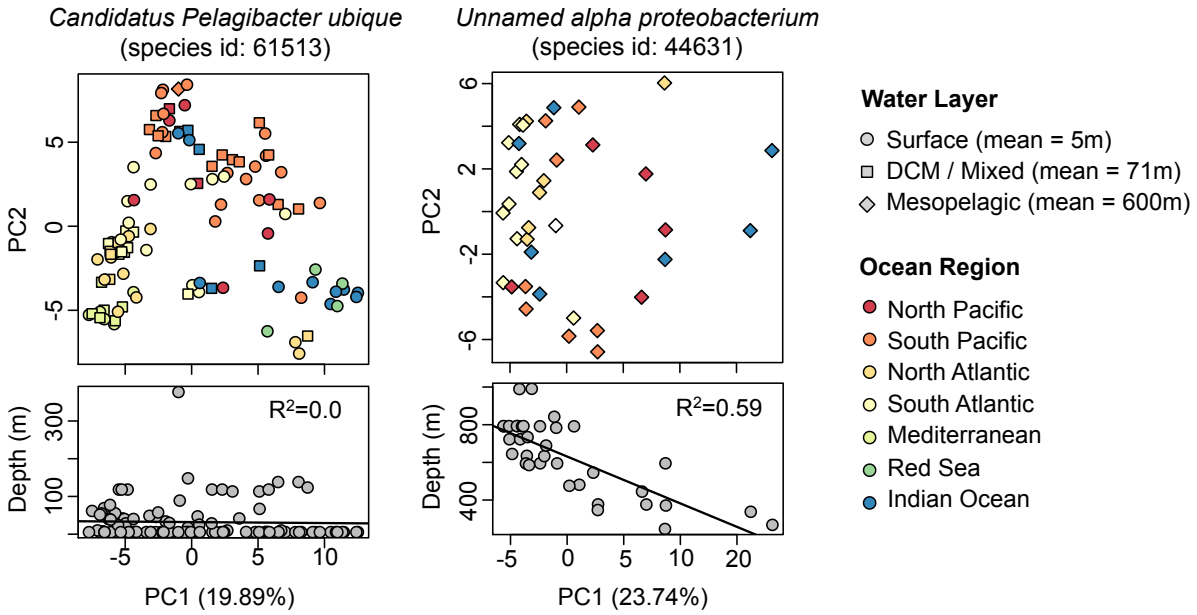

Figure 5.2 Gene content is correlated with geography and depth. Scatterplots show principal component analysis (PCA) of gene content for two bacterial species. Each point indicates a bacterial population from a different seawater sample. Point color and shape indicate the marine region and water layer respectively. *Candidatus Pelagibacter* populations tend to cluster together based on ocean region, not ocean depth. In contrast, *Alpha proteobacterium* populations tend to cluster together based on ocean depth, not ocean region.

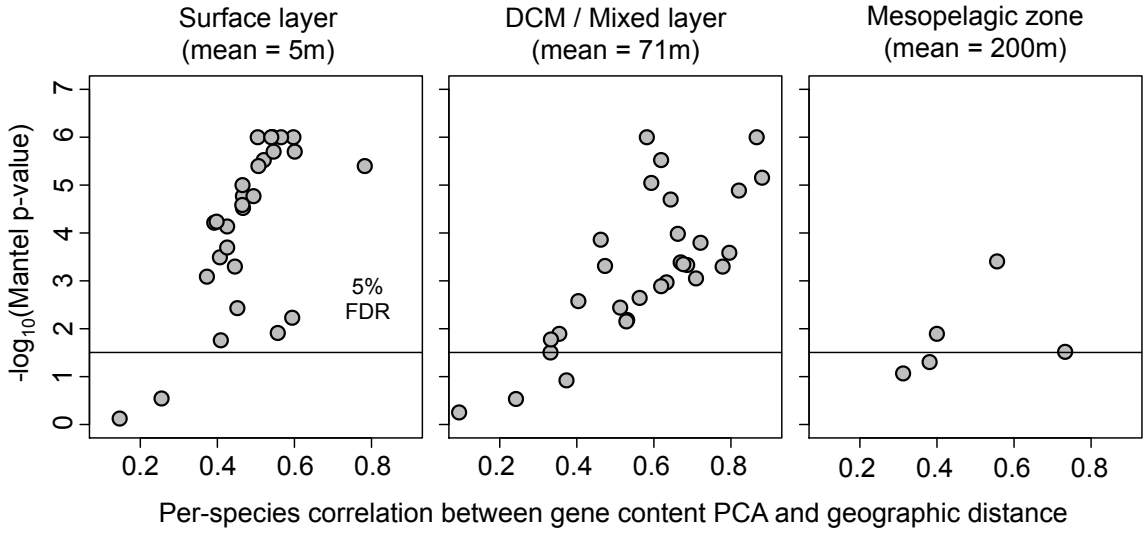

Figure 5.3 Gene content PCA and geographic distance are significantly correlated for most prevalent marine species. PCA distance was calculated using the Euclidian distance between PC1 and PC2 of the gene presence-absence matrix. Geographic distance was calculated using the

great-circle distance between sampling locations. For each species, the correlation of these two distances (horizontal axis) and associated p-value (vertical axis) were computed using the Mantel test with 1 million permutations. Only one metagenome per location was included in the tests. The population structure of marine bacteria, based on the first two principal components of gene content, is correlated with geography for many species of bacteria.

References

1. Federhen S, Rossello-Mora R, Klenk H-P, Tindall BJ, Konstantinidis KT, Whitman WB, Brown D, Labeda D, Ussery D, Garrity GM, et al: **Meeting report: GenBank microbial genomic taxonomy workshop (12–13 May, 2015).** *Standards in Genomic Sciences* 2016, **11**.
2. Mende DR, Sunagawa S, Zeller G, Bork P: **Accurate and universal delineation of prokaryotic species.** *Nat Methods* 2013, **10:**881-884.
3. Schloissnig S, Arumugam M, Sunagawa S, Mitreva M, Tap J, Zhu A, Waller A, Mende DR, Kultima JR, Martin J, et al: **Genomic variation landscape of the human gut microbiome.** *Nature* 2013, **493:**45-50.
4. Varghese NJ, Mukherjee S, Ivanova N, Konstantinidis KT, Mavrommatis K, Kyrpides NC, Pati A: **Microbial species delineation using whole genome sequences.** *Nucleic Acids Res* 2015, **43:**6761-6771.
5. Pei AY, Oberdorf WE, Nossa CW, Agarwal A, Chokshi P, Gerz EA, Jin Z, Lee P, Yang L, Poles M, et al: **Diversity of 16S rRNA genes within individual prokaryotic genomes.** *Appl Environ Microbiol* 2010, **76:**3886-3897.
6. Richter M, Rossello-Mora R: **Shifting the genomic gold standard for the prokaryotic species definition.** *Proc Natl Acad Sci U S A* 2009, **106:**19126-19131.
7. Wattam AR, Abraham D, Dalay O, Disz TL, Driscoll T, Gabbard JL, Gillespie JJ, Gough R, Hix D, Kenyon R, et al: **PATRIC, the bacterial bioinformatics database and analysis resource.** *Nucleic Acids Res* 2014, **42:**D581-591.
8. Stepanauskas R: **Single cell genomics: an individual look at microbes.** *Curr Opin Microbiol* 2012, **15:**613-620.
9. Nielsen HB, Almeida M, Juncker AS, Rasmussen S, Li J, Sunagawa S, Plichta DR, Gautier L, Pedersen AG, Le Chatelier E, et al: **Identification and assembly of genomes and genetic elements in complex metagenomic samples without using reference genomes.** *Nat Biotechnol* 2014, **32:**822-828.
10. Wu D, Hugenholtz P, Mavromatis K, Pukall R, Dalin E, Ivanova NN, Kunin V, Goodwin L, Wu M, Tindall BJ, et al: **A phylogeny-driven genomic encyclopaedia of Bacteria and Archaea.** *Nature* 2009, **462:**1056-1060.
11. Wu D, Jospin G, Eisen JA: **Systematic identification of gene families for use as "markers" for phylogenetic and phylogeny-driven ecological studies of bacteria and archaea and their major subgroups.** *PLoS One* 2013, **8:**e77033.
12. Eddy SR: **Accelerated Profile HMM Searches.** *PLoS Comput Biol* 2011, **7:**e1002195.
13. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ: **Basic local alignment search tool.** *Journal of Molecular Biology* 1990, **215:**403-410.
14. Loewenstein Y, Portugaly E, Fromer M, Linial M: **Efficient algorithms for accurate hierarchical clustering of huge datasets: tackling the entire protein space.** *Bioinformatics* 2008, **24:**i41-49.
15. Konstantinidis KT, Ramette A, Tiedje JM: **The bacterial species definition in the genomic era.** *Philos Trans R Soc Lond B Biol Sci* 2006, **361:**1929-1940.
16. Rinke C, Schwientek P, Sczyrba A, Ivanova NN, Anderson IJ, Cheng JF, Darling A, Malfatti S, Swan BK, Gies EA, et al: **Insights into the phylogeny and coding potential of microbial dark matter.** *Nature* 2013, **499:**431-437.

17. Sunagawa S, Mende DR, Zeller G, Izquierdo-Carrasco F, Berger SA, Kultima JR, Coelho LP, Arumugam M, Tap J, Nielsen HB, et al: **Metagenomic species profiling using universal phylogenetic marker genes.** *Nat Methods* 2013, **10:**1196-1199.
18. Browne HP, Forster SC, Anonye BO, Kumar N, Neville BA, Stares MD, Goulding D, Lawley TD: **Culturing of 'unculturable' human microbiota reveals novel taxa and extensive sporulation.** *Nature* 2016, **533:**543-546.
19. Chen Y, Ye W, Zhang Y, Xu Y: **High speed BLASTN: an accelerated MegaBLAST search tool.** *Nucleic Acids Res* 2015, **43:**7762-7768.
20. Nayfach S, Pollard KS: **Average genome size estimation improves comparative metagenomics and sheds light on the functional ecology of the human microbiome.** *Genome Biol* 2015, **16**.
21. The Human Microbiome Project Consortium: **A framework for human microbiome research.** *Nature* 2012, **486:**215-221.
22. Obregon-Tito AJ, Tito RY, Metcalf J, Sankaranarayanan K, Clemente JC, Ursell LK, Zech Xu Z, Van Treuren W, Knight R, Gaffney PM, et al: **Subsistence strategies in traditional societies distinguish gut microbiomes.** *Nat Commun* 2015, **6:**6505.
23. Rampelli S, Schnorr SL, Consolandi C, Turroni S, Severgnini M, Peano C, Brigidi P, Crittenden AN, Henry AG, Candela M: **Metagenome Sequencing of the Hadza Hunter-Gatherer Gut Microbiota.** *Curr Biol* 2015, **25:**1682-1693.
24. Qin J, Li Y, Cai Z, Li S, Zhu J, Zhang F, Liang S, Zhang W, Guan Y, Shen D, et al: **A metagenome-wide association study of gut microbiota in type 2 diabetes.** *Nature* 2012, **490:**55-60.
25. Li J, Jia H, Cai X, Zhong H, Feng Q, Sunagawa S, Arumugam M, Kultima JR, Prifti E, Nielsen T, et al: **An integrated catalog of reference genes in the human gut microbiome.** *Nat Biotechnol* 2014, **32:**834-841.
26. Schnorr SL, Candela M, Rampelli S, Centanni M, Consolandi C, Basaglia G, Turroni S, Biagi E, Peano C, Severgnini M, et al: **Gut microbiome of the Hadza hunter-gatherers.** *Nat Commun* 2014, **5:**3654.
27. Xiao L, Feng Q, Liang S, Sonne SB, Xia Z, Qiu X, Li X, Long H, Zhang J, Zhang D, et al: **A catalog of the mouse gut metagenome.** *Nat Biotechnol* 2015, **33:**1103-1108.
28. Tung J, Barreiro LB, Burns MB, Grenier JC, Lynch J, Grieneisen LE, Altmann J, Alberts SC, Blekhman R, Archie EA: **Social networks predict gut microbiome composition in wild baboons.** *Elife* 2015, **4**.
29. Sunagawa S, Coelho LP, Chaffron S, Kultima JR, Labadie K, Salazar G, Djahanschiri B, Zeller G, Mende DR, Alberti A, et al: **Ocean plankton. Structure and function of the global ocean microbiome.** *Science* 2015, **348:**1261359.
30. Fierer N, Leff JW, Adams BJ, Nielsen UN, Bates ST, Lauber CL, Owens S, Gilbert JA, Wall DH, Caporaso JG: **Cross-biome metagenomic analyses of soil microbial communities and their functional attributes.** *Proc Natl Acad Sci U S A* 2012, **109:**21390–21395.
31. Land M, Hauser L, Jun SR, Nookaew I, Leuze MR, Ahn TH, Karpinets T, Lund O, Kora G, Wassenaar T, et al: **Insights from 20 years of bacterial genome sequencing.** *Funct Integr Genomics* 2015, **15:**141-161.
32. Ahn TH, Chai J, Pan C: **Sigma: strain-level inference of genomes from metagenomic analysis for biosurveillance.** *Bioinformatics* 2015, **31:**170-177.

33. Francis OE, Bendall M, Manimaran S, Hong C, Clement NL, Castro-Nallar E, Snell Q, Schaalje GB, Clement MJ, Crandall KA, Johnson WE: **Pathoscope: species identification and strain attribution with unassembled sequencing data.** *Genome Res* 2013, **23:**1721-1729.
34. Truong DT, Franzosa EA, Tickle TL, Scholz M, Weingart G, Pasolli E, Tett A, Huttenhower C, Segata N: **MetaPhlAn2 for enhanced metagenomic taxonomic profiling.** *Nat Methods* 2015, **12:**902-903.
35. Tu Q, He Z, Zhou J: **Strain/species identification in metagenomes using genome-specific markers.** *Nucleic Acids Res* 2014, **42:**e67.
36. Sahl JW, Schupp JM, Rasko DA, Colman RE, Foster JT, Keim P: **Phylogenetically typing bacterial strains from partial SNP genotypes observed from direct sequencing of clinical specimen metagenomic data.** *Genome Med* 2015, **7:**52.
37. Zolfo M, Tett A, Jousson O, Donati C, Segata N: **MetaMLST: multi-locus strain-level bacterial typing from metagenomic samples.** *Nucleic Acids Res* 2016.
38. Joseph SJ, Li B, Petit RA, Qin ZS, Darrow LA, Read TD: **The single species metagenome: subtyping Staphylococcus aureus core genome sequences from shotgun metagenomic data.** *bioRxiv* 2015.
39. Luo C, Knight R, Siljander H, Knip M, Xavier RJ, Gevers D: **ConStrains identifies microbial strains in metagenomic datasets.** *Nat Biotechnol* 2015, **33:**1045-1052.
40. Ruby JG, Bellare P, Derisi JL: **PRICE: software for the targeted assembly of components of (Meta) genomic sequence data.** *G3 (Bethesda)* 2013, **3:**865-880.
41. Li D, Liu CM, Luo R, Sadakane K, Lam TW: **MEGAHIT: an ultra-fast single-node solution for large and complex metagenomics assembly via succinct de Bruijn graph.** *Bioinformatics* 2015, **31:**1674-1676.
42. Bankevich A, Nurk S, Antipov D, Gurevich AA, Dvorkin M, Kulikov AS, Lesin VM, Nikolenko SI, Pham S, Prjibelski AD, et al: **SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing.** *J Comput Biol* 2012, **19:**455-477.
43. Namiki T, Hachiya T, Tanaka H, Sakakibara Y: **MetaVelvet: an extension of Velvet assembler to de novo metagenome assembly from short sequence reads.** *Nucleic Acids Res* 2012, **40:**e155.
44. Wu YW, Tang YH, Tringe SG, Simmons BA, Singer SW: **MaxBin: an automated binning method to recover individual genomes from metagenomes using an expectation-maximization algorithm.** *Microbiome* 2014, **2:**26.
45. Alneberg J, Bjarnason BS, de Bruijn I, Schirmer M, Quick J, Ijaz UZ, Lahti L, Loman NJ, Andersson AF, Quince C: **Binning metagenomic contigs by coverage and composition.** *Nat Methods* 2014, **11:**1144-1146.
46. Greenblum S, Carr R, Borenstein E: **Extensive strain-level copy-number variation across human gut microbiome species.** *Cell* 2015, **160:**583-594.
47. Zhu A, Sunagawa S, Mende DR, Bork P: **Inter-individual differences in the gene content of human gut bacterial species.** *Genome Biol* 2015, **16:**82.
48. Scholz M, Ward DV, Pasolli E, Tolio T, Zolfo M, Asnicar F, Truong DT, Tett A, Morrow AL, Segata N: **Strain-level microbial epidemiology and population genomics from shotgun metagenomics.** *Nat Methods* 2016, **13:**435-438.

49. Liu B, Gibbons T, Ghodsi M, Treangen T, Pop M: **Accurate and fast estimation of taxonomic profiles from metagenomic shotgun sequences.** *BMC Genomics* 2011, **12 Suppl 2:**S4.
50. Langmead B, Salzberg SL: **Fast gapped-read alignment with Bowtie 2.** *Nat Methods* 2012, **9:**357-359.
51. Edgar RC: **Search and clustering orders of magnitude faster than BLAST.** *Bioinformatics* 2010, **26:**2460-2461.
52. Meyer F, Overbeek R, Rodriguez A: **FIGfams: yet another set of protein families.** *Nucleic Acids Res* 2009, **37:**6643-6654.
53. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT, et al: **Gene Ontology: tool for the unification of biology.** *Nature Genetics* 2000, **25:**25-29.
54. Kanehisa M, Goto S: **KEGG: Kyoto Encyclopedia of Genes and Genomes.** *Nucleic Acids Research* 2000, **28:**27-30.
55. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R, Genome Project Data Processing S: **The Sequence Alignment/Map format and SAMtools.** *Bioinformatics* 2009, **25:**2078-2079.
56. Zhu Y, Stephens RM, Meltzer PS, Davis SR: **SRAdb: query and use public next-generation sequencing data from within R.** *BMC bioinformatics* 2013, **14**.
57. Leinonen R, Sugawara H, Shumway M, International Nucleotide Sequence Database C: **The sequence read archive.** *Nucleic Acids Res* 2011, **39:**D19-21.
58. Kurtz S, Phillippy A, Delcher AL, Smoot M, Shumway M, Antonescu C, L. SS: **Versatile and open software for comparing large genomes.** *Genome Biology* 2004, **5:**R12.
59. Backhed F, Roswall J, Peng Y, Feng Q, Jia H, Kovatcheva-Datchary P, Li Y, Xia Y, Xie H, Zhong H, et al: **Dynamics and Stabilization of the Human Gut Microbiome during the First Year of Life.** *Cell Host Microbe* 2015, **17:**690-703.
60. Kjersti Aagaard, Jun Ma, Kathleen M. Antony, Radhika Ganu J, Joseph Petrosino, Versalovic J: **The Placenta Harbors a Unique Microbiome.** *Science Translational Medicine* 2014, **6:**1-11.
61. Bokulich NA, Chung J, Battaglia T, Henderson N, Jay M, Li H, A DL, Wu F, Perez-Perez GI, Chen Y, et al: **Antibiotics, birth mode, and diet shape microbiome maturation during early life.** *Sci Transl Med* 2016, **8:**343ra382.
62. Yassour M, Vatanen T, Siljander H, Hamalainen AM, Harkonen T, Ryhanen SJ, Franzosa EA, Vlamakis H, Huttenhower C, Gevers D, et al: **Natural history of the infant gut microbiome and impact of antibiotic treatment on bacterial strain diversity and stability.** *Sci Transl Med* 2016, **8:**343ra381.
63. Kostic AD, Gevers D, Siljander H, Vatanen T, Hyotylainen T, Hamalainen AM, Peet A, Tillmann V, Poho P, Mattila I, et al: **The dynamics of the human infant gut microbiome in development and in progression toward type 1 diabetes.** *Cell Host Microbe* 2015, **17:**260-273.
64. Yatsunenko T, Rey FE, Manary MJ, Trehan I, Dominguez-Bello MG, Contreras M, Magris M, Hidalgo G, Baldassano RN, Anokhin AP, et al: **Human gut microbiome viewed across age and geography.** *Nature* 2012, **486:**222-227.
65. Koenig JE, Spora A, Scalfonea N, Frickera AD, Stombaugh J, Knight R, Angenent LT, Ley RE: **Succession of microbial consortia in the developing infant gut microbiome.** *Proc Natl Acad Sci U S A* 2011, **108:**4578–4585.

66. Martin V, Maldonado-Barragan A, Moles L, Rodriguez-Banos M, Campo RD, Fernandez L, Rodriguez JM, Jimenez E: **Sharing of bacterial strains between breast milk and infant feces.** *J Hum Lact* 2012, **28:**36-44.
67. Makino H, Kushiro A, Ishikawa E, Kubota H, Gawad A, Sakai T, Oishi K, Martin R, Ben-Amor K, Knol J, Tanaka R: **Mother-to-infant transmission of intestinal bifidobacterial strains has an impact on the early development of vaginally delivered infant's microbiota.** *PLoS One* 2013, **8:**e78331.
68. Milani C, Mancabelli L, Lugli GA, Duranti S, Turroni F, Ferrario C, Mangifesta M, Viappiani A, Ferretti P, Gorfer V, et al: **Exploring Vertical Transmission of Bifidobacteria from Mother to Child.** *Appl Environ Microbiol* 2015, **81:**7078-7087.
69. Tannock GW, Fuller R, Smith SL, Hall MA: **Plasmid Profiling of Members of the Family *Enterobacteriaceae*, *Lactobacilli*, and *Bifidobacteria* To Study the Transmission of Bacteria from Mother to Infant.** *Journal of Clinical Microbiology* 1990, **28:**1225-1228.
70. Group NHW, Peterson J, Garges S, Giovanni M, McInnes P, Wang L, Schloss JA, Bonazzi V, McEwen JE, Wetterstrand KA, et al: **The NIH Human Microbiome Project.** *Genome Res* 2009, **19:**2317-2323.
71. Faith JJ, Guruge JL, Charbonneau M, Subramanian S, Seedorf H, Goodman AL, Clemente JC, Knight R, Heath AC, Leibel RL, et al: **The long-term stability of the human gut microbiota.** *Science* 2013, **341:**1237439.
72. Kuleshov V, Jiang C, Zhou W, Jahanbani F, Batzoglou S, Snyder M: **Synthetic long-read sequencing reveals intraspecies diversity in the human microbiome.** *Nat Biotechnol* 2016, **34:**64-69.
73. Mu XJ, Lu ZJ, Kong Y, Lam HY, Gerstein MB: **Analysis of genomic variation in non-coding elements using population-scale sequencing data from the 1000 Genomes Project.** *Nucleic Acids Res* 2011, **39:**7058-7076.
74. Tal Korem DZ, Jotham Suez, Adina Weinberger, Tali Avnit-Sagi, Maya Pompan-Lotan, Elad Matot, Ghil Jona, Alon Harmelin, Nadav Cohen, Alexandra Sirota-Madi, Christoph A. Thaiss, Meirav Pevsner-Fischer, Rotem Sorek, Ramnik J. Xavier, Eran Elinav, Eran Segal: **Growth dynamics of gut microbiota in health and disease inferred from single metagenomic samples.** *Science* 2015, **349**.
75. Lam LH, Monack DM: **Intraspecies competition for niches in the distal gut dictate transmission during persistent Salmonella infection.** *PLoS Pathog* 2014, **10:**e1004527.
76. Lee SM, Donaldson GP, Mikulski Z, Boyajian S, Ley K, Mazmanian SK: **Bacterial colonization factors control specificity and stability of the gut microbiota.** *Nature* 2013, **501:**426-429.
77. Lieberman TD, Flett KB, Yelin I, Martin TR, McAdam AJ, Priebe GP, Kishony R: **Genetic variation of a bacterial pathogen within individuals with cystic fibrosis provides a record of selective pressures.** *Nat Genet* 2014, **46:**82-87.
78. Kashtan N, Roggensack SE, Rodrigue S, Thompson JW, Biller SJ, Coe A, Ding H, Marttinen P, Malmstrom RR, Stocker R, et al: **Single-Cell Genomics Reveals Hundreds of Coexisting Subpopulations in Wild Prochlorococcus.** *Science* 2014, **344:**416-420.
79. Rappe MS, Connon SA, Vergin KL, Giovannoni SJ: **Cultivation of the ubiquitous SAR11 marine bacterioplankton clade.** *Nature* 2002, **418:**630-633.

80. Dupont CL, Rusch DB, Yooseph S, Lombardo MJ, Richter RA, Valas R, Novotny M, Yee-Greenbaum J, Selengut JD, Haft DH, et al: **Genomic insights to SAR86, an abundant and uncultivated marine bacterial lineage.** *ISME J* 2012, **6:**1186-1199.
81. Novembre J, Johnson T, Bryc K, Kutalik Z, Boyko AR, Auton A, Indap A, King KS, Bergmann S, Nelson MR, et al: **Genes mirror geography within Europe.** *Nature* 2008, **456:**98-101.
82. Hijmans RJ: **geosphere: Spherical Trigonometry.** 2016.
83. Oksanen J, Blanchet F, Friendly M, Kindt R, Legendre P, McGlinn D, Minchin P, O'Hara RB, Simpson GL, Solymos P, et al: **vegan: Community Ecology Package.** 2016.
84. Ladau J, Sharpton TJ, Finucane MM, Jospin G, Kembel SW, O'Dwyer J, Koeppel AF, Green JL, Pollard KS: **Global marine bacterial diversity peaks at high latitudes in winter.** *ISME J* 2013, **7:**1669-1677.

Publishing Agreement

It is the policy of the University to encourage the distribution of all theses, dissertations, and manuscripts. Copies of all UCSF theses, dissertations, and manuscripts will be routed to the library via the Graduate Division. The library will make all theses, dissertations, and manuscripts accessible to the public and will preserve these to the best of their abilities, in perpetuity.

I hereby grant permission to the Graduate Division of the University of California, San Francisco to release copies of my thesis, dissertation, or manuscript to the Campus Library to provide access and preservation, in whole or in part, in perpetuity.

Author Signature Stephen Vaughan Date 03/13/17