An integrated metagenomics pipeline for strain profiling reveals novel patterns of transmission and global biogeography of bacteria

Stephen Nayfach[1,2], Beltran Rodriguez-Mueller[2], Nandita Garud[2], Katherine S. Pollard[1, 2, 3 §]

1 Integrative Program in Quantitative Biology, University of California San Francisco, CA 94158
2 Gladstone Institutes, San Francisco, CA 94158
3 Institute for Human Genetics, Institute for Computational Health Sciences, and Department of Epidemiology and Biostatistics, University of California San Francisco, CA 94158

§ Corresponding author

Email addresses:
SN: stephen.nayfach-battilana@ucsf.edu
KSP: kpollard@gladstone.ucsf.edu

Abstract

We present the *Metagenomic Intra-species Diversity Analysis System (MIDAS)*, which is an integrated computational pipeline for quantifying bacterial species abundance and strain-level genomic variation, including gene content and single nucleotide polymorphisms, from shotgun metagenomes. Our method leverages a database of >30,000 bacterial reference genomes which we clustered into species groups. These cover the majority of abundant species in the human microbiome but only a small proportion of microbes in other environments, including soil and seawater. We applied *MIDAS* to stool metagenomes from 98 Swedish mothers and their infants over one year and used rare single nucleotide variants to reveal extensive vertical transmission of strains at birth but colonization with strains unlikely to derive from the mother at later time points. This pattern was missed with species-level analysis, because the infant gut microbiome composition converges towards that of an adult over time. We also applied *MIDAS* to 198 globally distributed marine metagenomes and used gene content to show that many prevalent bacterial species have population structure that correlates with geographic location. Strain-level genetic variants present in metagenomes clearly reveal extensive structure and dynamics that are obscured when data is analyzed at a higher taxonomic resolution.

Introduction

Microbial communities play a myriad of important roles in the different environments that they inhabit. These communities are typically comprised of various distinct species that each exists as a complex and heterogeneous population of cells with differences in gene content (Greenblum et al. 2015; Zhu et al. 2015) and single nucleotide polymorphisms (SNPs) (Schloissnig et al. 2013; Kashtan et al. 2014; Lieberman et al. 2014). Several recent studies used within-species differences in gene content and SNPs as a window into the ongoing evolutionary history of microbes on earth. For example, this approach has revealed genomic events that lead to ecologically distinct species (Shapiro et al. 2012), uncovered the presence of ancient subpopulations of ecologically differentiated marine bacteria (Kashtan et al. 2014), and highlighted extensive intra-species recombination in pathogens (Snitkin et al. 2011) and free-living bacteria (Rosen et al. 2015). Additionally, an understanding of strain-level variation is critical for studying the interaction of microbes with humans and for understanding microbial pathogenicity. Differences at the nucleotide level can lead to within-host adaptation of pathogens (Lieberman et al. 2014), and differences in gene content can confer drug resistance, convert a commensal bacterium into a pathogen (Snitkin et al. 2011), or lead to outbreaks of highly virulent strains (Rasko et al. 2011).

Metagenomic shotgun sequencing has the potential to shed light onto strain-level heterogeneity among bacterial genomes within and between microbial communities, yielding a genomic resolution not achievable by sequencing the 16S ribosomal RNA gene alone (Sunagawa et al. 2013) and circumventing the need for culture-based approaches. However, limitations of existing computational methods and reference databases have prevented most researchers from obtaining this level of resolution from metagenomic data.

Assembly-free methods that map reads to reference genomes in order to estimate the relative abundance of known strains (Francis et al. 2013; Tu et al. 2014) are effective for well-characterized pathogens like *E. coli* that have thousands of sequenced genomes. However, such methods cannot detect strain-level variation for the vast majority of known species that currently have only a single sequenced representative. Other assembly-free approaches have been developed that use reads mapped to one or more reference genomes to identify SNPs (Schloissnig et al. 2013; Lieberman et al. 2014) and gene copy-number variants (Greenblum et al. 2015; Zhu et al. 2015; Scholz et al. 2016) of microbial populations. These approaches have not been integrated together and/or made available as software. Recently, several software tools have been developed (Luo et al. 2015; Sahl et al. 2015) that use SNP patterns to phylogenetically type strains, but these methods do not capture the gene content of these organisms and may not be able to resolve strains in communities with high population heterogeneity. Additionally, existing methods do not provide comprehensive up-to-date genomic databases of bacterial species, thus limiting their utility across different environments. Assembly-based methods (Nielsen et al. 2014; Cleary et al. 2015) that seek to reconstruct microbial genomes without using reference genomes are a powerful alternative to assembly-free methods. However, these often require many samples, struggle to deconvolve closely related strains, or require manual inspection.

To address these issues, we developed the *Metagenomic Intra-species Diversity Analysis System (MIDAS)*, which is a computational pipeline that quantifies bacterial species abundance and intra-species genomic variation from shotgun metagenomes. Our method integrates many features (for a comparison to existing methods, see Supplemental Table S1) and leverages a comprehensive database of >30,000 reference genomes. Given a shotgun metagenome, *MIDAS* rapidly and automatically quantifies gene content and identifies SNPs in bacterial species, which is accurate for populations with a minimum of 1 and 10x sequencing coverage, respectively. These statistics enable quantitative analysis of bacterial populations within and between metagenomic samples.

To demonstrate the utility of this approach, we used *MIDAS* to conduct novel strain-level analyses on two datasets. First, we applied *MIDAS* to stool metagenomes from 98 Swedish mothers and their infants and used rare SNPs to track vertical transmission and temporal stability of strains in infants over the first year of life. Second, we used *MIDAS* to quantify gene content of prevalent bacterial species in 198 globally distributed marine metagenomes and identified significant intra-species population structure associated with geographic location and environmental variables. These analyses reveal striking structures in microbial communities that are missed when metagenomes are analyzed at a coarser taxonomic resolution.

Results

Identification of bacterial species with a consistent definition and efficient algorithm

To quantify strain-level genomic variation broadly and accurately, we built a comprehensive database of 31,007 high-quality bacterial reference genomes obtained from the Pathosystems Resource Integration Center (PATRIC) (Wattam et al. 2014). We accurately clustered these genomes into species groups to avoid inconsistent, erroneous,

133 and incomplete annotations that afflict some microbial taxonomies (Mende et al. 2013),
134 and to expand and improve upon previous efforts to systematically delineate bacterial
135 species (Mende et al. 2013; Schloissnig et al. 2013; Varghese et al. 2015). Towards this goal,
136 we hierarchically clustered reference genomes using the average pairwise percent identity
137 across a panel of 30 universal genes (Figure 1a) that we selected from a panel of 112
138 candidates (Wu et al. 2013) (Supplemental Fig S1, Supplemental Table S2). We found that
139 the best gene families for identifying bacterial species were less conserved and more
140 widely distributed across the tree of life relative to other genes we tested (Supplemental
141 Fig S2). For example, many ribosomal gene families were too conserved to differentiate
142 closely related species (Supplemental Table S2). We applied a 96.5% nucleotide identity
143 cutoff, which produced genome-clusters that were highly concordant with a gold standard
144 definition of prokaryotic species based on 95% genome-wide average nucleotide identity
145 (Konstantinidis et al. 2006; Richter and Rossello-Mora 2009) (Supplemental Table S3). Our
146 procedure clustered the 31,007 bacterial genomes into 5,952 genome-clusters,
147 representing distinct bacterial species (Supplemental Table S4-S5). We inferred the
148 phylogenetic relationships of these species using a concatenated alignment of the 30
149 marker genes (Supplemental Fig S3). Because our algorithm uses a small set of highly
150 informative marker genes, rather than genome-wide sequence comparisons, it will be
151 efficient to update these genome-clusters as additional genomes are sequenced.
152
153 The genome-clusters we identified often differed from the PATRIC taxonomic labels (Figure
154 1b). Our procedure clustered 2,666 genomes (8.6% of total) that had not been previously
155 annotated at the species level and reassigned species labels for 3,035 genomes (9.8% of
156 total) to either (i) group them with genomes that were not labeled as the same species in
157 the reference taxonomy (N=1,380) or (ii) split them from genomes with the same label in
158 the reference taxonomy (N=1,655). Supporting our species definitions, we found that the
159 bacterial species we identified tended to have distinct functional repertoires, with only
160 0.05% of FIGfam protein families (Meyer et al. 2009) shared between genomes from
161 different species on average compared to >80% for pairs of genomes from the same
162 species. In previous work, Mende et al. conducted a similar procedure to cluster genomes
163 into species groups and found that the majority of disagreements with the NCBI taxonomy
164 were supported by the literature (Mende et al. 2013).
165
166 **Current reference genomes cover the majority of human-associated bacterial species**
167 **and highlight novel diversity in other environments**
168
169 We evaluated how comprehensively our reference database covers the abundance of
170 species present in different environments, as this is a requirement for conducting
171 reference-based strain-level analyses. Previous work has shown large gaps in diversity
172 between sequenced reference genomes and environmental microorganisms (Wu et al.
173 2009). To explore how well current genome sequences cover diversity present in
174 metagenomes from various environments, we developed a novel approach that estimates
175 the proportion of microbial genomes (including archaea and eukaryotes, but excluding
176 viruses) in a metagenome that contain a sequenced representative at the species level in a
177 reference database (Methods). This proportion, which we call *database coverage*, indicates
178 the degree to which species in a sample are known versus novel.
179
180 We applied this method to stool metagenomes from the Human Microbiome Project (HMP)
181 and four other studies of the human gut (Supplemental Table S6). We found that our
182 reference database of 5,952 bacterial species had high coverage of microbial communities

183 from the human body (Figure 1c). This included high database coverage of samples from
184 the skin (mean=83%), nasal cavity (mean=63%), urogenital tract (mean=62%), mouth
185 (mean=55%), and gastrointestinal tract (mean=49%). The human gut communities with
186 highest database coverage came from individuals in the United States, Europe, and China
187 that live urban lifestyles, which is consistent with a previous report (Sunagawa et al. 2013).
188 In contrast, gut microbiomes of individuals from Tanzania and Peru that live hunter-
189 gatherer and agricultural lifestyles had much higher levels of novel species with no
190 sequenced representative in our database. This finding extends the previous discoveries of
191 elevated levels of novel genera (Schnorr et al. 2014) and functions (Rampelli et al. 2015) in
192 the gut microbiome of African hunter-gatherers. Our analysis points to specific
193 phylogenetic gaps in the set of currently sequenced bacterial genomes. Gut communities
194 with lower database coverage (i.e., fewer species that have been sequenced) tended to have
195 higher levels of several genera including *Coprococcus*, *Subdoligranulum*, *Dorea*, and *Blautia*,
196 whereas well-characterized communities tended to have higher levels of the genus
197 *Bacteroides* (Supplemental Fig S4). We conclude that there is a clear bias of genome
198 sequencing to date towards species associated with hosts from industrialized countries.
199
200 In contrast to the human microbiome, a relatively small proportion of microbes present in
201 other environments were captured by our reference database (Figure 1c and Supplemental
202 Table S6). This included very low coverage for stool metagenomes from laboratory mice
203 (mean=4.3%), which was surprising since mice are often used as a model system for
204 studying the human microbiome. Coverage was also strikingly low in marine (means:
205 surface water=8.2%, deep chlorophyll maximum layer=6.9%, subsurface epipelagic mixed
206 layer=1.0%, mesopelagic zone=4.0%) and soil (means: desert=1.0%, forest=1.0%,
207 grassland=1.3%, tundra=1.1%) environments. These estimates emphasize the massive gap
208 that remains between the microbial diversity found in non-human environments and that
209 represented by sequenced bacterial reference genomes. Strain-level analyses can still be
210 performed for environments with low database coverage, but only for those species with
211 sequenced representatives.
212
213 **An integrated pipeline for quantifying intra-species genomic variation from shotgun**
214 **metagenomes**
215
216 We developed *MIDAS,* a software tool that processes shotgun metagenomes to sensitively
217 and automatically quantify species abundance and strain-level genomic variation for any of
218 the bacterial species in our database (Figure 2a and Methods). *MIDAS* was designed to be
219 fast, memory efficient, and to scale with the rapid increase in sequenced reference genomes
220 (Supplemental Fig S5). Using a single CPU, *MIDAS* processes ~5,000 reads per second and
221 requires ~3 gigabytes of RAM.
222
223 *MIDAS* first estimates the coverage and relative abundance of bacterial species by mapping
224 reads to a database of universal single-copy gene families (Supplemental Table S7).
225 Identifying species with sufficient coverage for gene content and SNP analyses directly
226 from the shotgun metagenome enables automatic selection of individual species for variant
227 quantification without any prior knowledge about a community's composition, and it
228 avoids computationally wasteful alignments to genes and genomes from sequenced
229 organisms that are not present in a community.
230
231 To quantify the gene content of individual species in each metagenome, *MIDAS* maps reads
232 to a pan-genome database. This database contains the set of non-redundant genes across

233 all sequenced genomes from each species. It is generated on the fly to include only the
234 subset of species with high sequencing coverage at universal single-copy genes in the
235 metagenome being analyzed. The coverages of genes in the pan-genome database are
236 normalized by the coverage of the universal single-copy gene families, yielding an
237 estimated copy number of a gene per cell of a given species in each sample. Additionally,
238 copy numbers are thresholded to predict gene presence-absence per sample.
239
240 To identify SNPs of individual species, *MIDAS* maps reads to a genome database. This
241 database contains one representative genome sequence per species, and it only includes
242 species with high sequencing coverage at universal single-copy genes in the metagenome
243 being analyzed. Representative genomes are selected in order to maximize their sequence
244 identity to all other genomes within the species. The core genome of each species is
245 identified directly from the data using nucleotide positions in the representative genome
246 that are at high coverage across multiple metagenomic samples (Supplemental Fig S6).
247 SNPs are quantified along the entire core genome, including at sites that are variable
248 between samples, but fixed within individual samples. Core genome SNPs are useful
249 because they occur in all strains of a species and facilitate comparative analyses.
250
251 *MIDAS* was validated using 20 mock metagenomes that we created by pooling Illumina
252 reads from completed genome sequencing projects (Supplemental Tables S8-9 and
253 Methods). These libraries are expected to contain sequencing errors and other
254 experimental artifacts found in real short-read sequencing data that might prevent
255 accurate estimation of species abundance and strain-level genomic variation. Using this
256 data, we found that *MIDAS* accurately estimated the relative abundance of bacterial species
257 (r^2=0.95), but slightly underestimated sequencing coverage (Figure 2b). *MIDAS* accurately
258 predicted the presence or absence of genes in species present with at least 1 to 3x
259 sequencing coverage (Figure 2c). Prediction accuracy was maximized at 0.96 for strains
260 with >3x coverage when using a threshold equal to 0.35x the coverage of universal single-
261 copy genes – lower thresholds resulted in lower specificity and higher thresholds resulted
262 in lower sensitivity. *MIDAS* also called SNPs at a low false-discovery rate, but required
263 between 5 to 10x coverage to identify the majority of SNPs present (Figure 2d).
264
265 **Species and strain-resolved analyses shed light on vertical transmission of human**
266 **gut microbiota**
267
268 We hypothesized that the large numbers of SNPs that *MIDAS* can identify from individual
269 metagenomes could be leveraged to detect bacterial strains unique to a host and
270 transmission of strains between mothers and their infants. An understanding of vertical
271 transmission is critical for determining the extent to which the microbiome – and by
272 extension microbiome-mediated phenotypes – are inherited. A recent study found
273 significant overlap in species between Swedish mothers and their infants over the first year
274 of life (Backhed et al. 2015). A large cohort study of UK twins estimated that abundances of
275 many microbial taxa are heritable (Goodrich et al. 2016). Neither study examined whether
276 strains are vertically transmitted, and recent work has shown that species-level analyses
277 alone are insufficient to resolve transmission events (Li et al. 2016). Transmission of
278 specific taxonomic groups has been resolved using multilocus sequence typing of infant
279 stool plus mother's breast milk (Martin et al. 2012) or stool (Makino et al. 2013). Other
280 studies have examined the development of the infant gut microbiome, including at the
281 strain level (Luo et al. 2015), 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.

To quantify strain transmission from mother to infant, we applied *MIDAS* to the Backhed et al. stool metagenomes from 98 mothers and their infants at 4 days, 4 months, and 12 months after birth (Backhed et al. 2015). We 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 3a and Supplemental Fig S7). Compared to infants, mothers had more diverse microbiomes that tended to harbor more unshared (i.e., unique to host) species (77% versus 48%, T-test $P<2.2e-16$). Despite this, we found a large number of shared species between infants and their mothers, which increased over time as diversity increased in the infants (Figure 3b) and did not strongly depend on birth mode ($P=0.52$) or breast-feeding at any stage (P-values = 0.04, 0.04, 0.33 at 4 days, 4 months, and 12 months). These species-level trends agree with the results of the original study that used different methods to identify species (Backhed et al. 2015). Surprisingly, we found nearly as many shared species between permuted mother-infant pairs where vertical transmission did not occur (Figure 3c), suggesting that increased similarity of species in a mother and her infant over its first year is unlikely the result of direct transmission.

To detect transmission of gut microbiota from mother to infant with high specificity and sensitivity, we developed a novel approach that uses SNPs output by *MIDAS* (Methods). First we identified species shared between mothers and their infants with >10x sequencing coverage, which is required for sensitive detection of SNPs (Figure 2d). Next, we identified rare SNPs within these species that were private to strains found in a mother and her infant. We refer to these SNPs as *marker alleles* because they serve as a marker for individual strains. To detect whether a transmission has occurred for a species, we quantified the percent of marker alleles found in a mother that were shared with her infant.

To validate that marker alleles could be used to track strains between hosts, we applied our method to stool metagenomes of healthy adults from the Human Microbiome Project (HMP) (Methods). As a positive control, we compared marker alleles of species between metagenomes from the same individual at the same time point (technical replicates), and as a negative control, we compared marker alleles of species between metagenomes from different unrelated individuals (non-replicates). As expected, we found high allele sharing (mean=79.5%) between technical replicates and low allele sharing between non-replicates (mean=1.01%) (Supplemental Fig S8). The fact that allele sharing was <100% in the technical replicates and >0% in the non-replicates is likely due to a combination of factors, including read sampling variation, small sample sizes, and sequencing errors. For example, marker alleles may be found in other individuals when sample sizes are increased. To define a transmission event, we selected a marker allele sharing cutoff of 5%, which clearly separated the positive and negative controls (sensitivity=99.8%, specificity=96.6%). High sensitivity and specificity was consistently observed across species we tested (Supplemental Table S10).

Strikingly, we found that marker alleles were commonly shared between mothers and vaginally born infants 4 days after birth (Figure 3d-e). There were no species present with >10x coverage in 15 C-section born infants and their mothers to assess transmission in these individuals. This likely reflects lower vertical transmission of the mother's gut

332 microbes, but we cannot directly test that hypothesis with the available data. On average
333 72% of marker alleles present in mother strains were found in vaginally born newborns,
334 which was only slightly less than the level of allele sharing observed from our positive
335 control. Furthermore, out of the 111 high-coverage species present in mothers and
336 newborns, 101 (91%) had greater than 5% marker allele sharing, indicating extensive
337 vertical transmission of gut microbiota shortly after birth. Commonly transmitted species
338 included *Bacteroides vulgatus* (25/28 mother-infant pairs with >5% marker allele sharing),
339 *Parabacteroides distasonis* (10/11), *Bifidobacterium adolescentis* (8/10), and *Escherichia*
340 *coli* (10/10) (Figure 3d). These are fairly different from the taxa with heritable abundances
341 in the UK twin study (Goodrich et al. 2016) (see Discussion).
342
343 While we detected high strain similarity 4 days after birth, mother and infant strains
344 significantly differed over time. Comparing strain-level SNPs in 4-month and 12-month
345 infants to their mothers, we observed a sharp decrease in marker allele sharing and
346 transmission rates (Figure 3d). Across all species, transmission rates decreased from 91%
347 at 4 days (101/111 shared species with >5% marker allele sharing), to 80% at 4 months
348 (131/163), to 55% at 12 months (172/313). C-section born infants tended to have fewer
349 vertically strains transmitted strains compared to vaginally born infants at four months
350 (Chi-square P=5e-8, 3/14 versus 128/149 shared species with >5% marker allele sharing)
351 and to a lesser extent at 12 months (Chi-square P=0.06, 13/34 versus 159/279). This trend
352 was in stark contrast to what we observed at the species level, where there was an increase
353 in the number of shared species over time and an increase in species-level compositional
354 similarity. Thus, while the species level composition of mothers and infants converged over
355 time, the strain level composition actually diverged.
356
357 We hypothesized that transmission rates decreased over time due to late colonization of
358 the infant gut by new strains from the environment and/or other hosts. If this were the
359 case, then we would expect that (i) infant strains that were distinct from the mother at 12
360 months had low abundance in the infant at earlier stages and (ii) strains transmitted from
361 the mother at 4 days persisted in the infant gut over one year. Supporting our hypothesis,
362 we found that the abundance of a species at 4 days was predictive of whether the strain of
363 that species was transmitted from the mother (Figure 4a). Specifically, strains with low
364 abundance at 4 days but high abundance at 12 months tended to be distinct from strains
365 found in the mother. In contrast, strains with high abundance at 4 days and high abundance
366 at 12 months were similar to strains found in the mother. Also supporting our hypothesis,
367 we found that the vast majority of strains that were transmitted from the mother at 4 days
368 persisted in the infants at 4 months (49/54 mother-infant pairs with >5% marker allele
369 sharing) and at 12 months (47/51). Because the mother's stool was only sequenced at 4
370 days after birth, we cannot rule out the possibility that late colonizing strains came from
371 the mother's gut but were not detected at the time of initial sampling. To address this issue,
372 we quantified the temporal stability of strains in 157 healthy adults from the HMP over a
373 time period of 300-400 days (Supplemental Fig S9). We found high marker allele sharing
374 (mean=77.0%) and transmission rates (96.2%), which suggests that maternal strains may
375 be quite stable over time, and agrees with previous work (Faith et al. 2013; Schloissnig et
376 al. 2013). Together these results suggest that bacteria from sources other than the mother's
377 gut increasingly colonize the infant gut over time.
378
379 We found that vertical transmission rates varied for different taxonomic groups of bacteria.
380 At one year after birth, the class Bacteroidia was enriched in vertical transmission events
381 (Chi-square P=2.56e-18), whereas the class Clostridia was depleted (Chi-square P=1.43e-

382 22). Similar results were observed at other taxonomic levels (Supplemental Table S11).
383 Examples of early colonizing Bacteroidia included *Bacteroides vulgatus*, *Bacteroides*
384 *uniformis*, and *Bacteroides ovatus* and examples of late colonizing Clostridia included
385 *Blautia wexlerae, Faecalibacterium prausnitzii*, *Eubacterium rectale,* and *Ruminococcus*
386 *bromii* (Figure 4b). The fact that Clostridia were rarely transmitted from mother to infant
387 may be due to the ability of members of this group to form spores and survive outside of
388 the host for longer periods of time (Browne et al. 2016). Together, these results highlight
389 differences in the inheritance of gut microbiota that may be linked to distinct modes and
390 timing of transmission between hosts.
391
392 **Global strain-level geography of prevalent marine bacteria**
393
394 Many bacterial species are distributed widely across the world's oceans (Sunagawa et al.
395 2015). Yet genomes of a given species sampled nearby each other can differ significantly in
396 their gene content (Kashtan et al. 2014). To explore the extent of population structure
397 across different marine bacterial species on a global scale, we used *MIDAS* to quantify gene
398 content for prevalent species in 198 marine metagenomes from 66 stations along the *Tara*
399 Oceans expedition (Sunagawa et al. 2015). Since we found that our database had relatively
400 low coverage of the cellular organisms present ocean samples (Figure 1c), we first
401 estimated relative abundance and coverage of bacterial species in each metagenome to
402 identify 30 species where gene content could be reliably estimated (coverage >3x across a
403 high percentage of samples). Among these species were several members of the genera
404 *Pelagibacter*, *Alteromonas*, *Synechococcus*, and *Marinobacter,* a large group of closely
405 related *Prochlorococcus* species, and several unnamed *Alphproteobacteria* species (Figure
406 5a). Reference pangenome sizes for these species ranged from 1,047 and 1,311 genes in the
407 streamlined genomes of SAR406 and SAR86 (each with 1 genome) to 6,427 genes in the
408 largest *Prochlorococcus* genome cluster (N=26 genomes) and 7,819 genes for *Alteromonas*
409 *macleodii* (N=4 genomes).
410
411 We discovered extensive variability of gene content for these prevalent species of marine
412 bacteria across the ocean metagenomes (Supplemental Table S12). Across all species, we
413 found an average of 318 genes that differed between samples, ranging from 144 genes in
414 *SAR86* to 700 in *Alteromonas marina*. We next quantified the percent of genes that were
415 different between samples using the Jaccard index and found that on average 19% of genes
416 differed between samples. This level of genomic variability was higher than the 13%
417 recently reported for human gut communities (Zhu et al. 2015), although this may be due
418 to methodological differences. Regardless, our estimate of 19% is almost certainly an
419 underestimate of the true level of gene content variation between populations, because
420 *MIDAS* cannot measure the variation of genes that are present in strains but absent from
421 sequenced reference genomes.
422
423 To explore how this variation correlated with geography and sampling depth, we
424 conducted a principal component analysis (PCA) of gene content for each bacterial species,
425 as has been done to study the geographic structure of human populations (Novembre et al.
426 2008). Strikingly, we found that the populations of many species clustered together by
427 ocean region based on the first two principal components of gene content, regardless of
428 sampling depth (Figure 5b). For example, populations of *Pelagibacter* (species id: 61513)
429 formed three discrete clusters corresponding to the Mediterranean Sea, South Atlantic
430 Ocean, and South Pacific Ocean, and each cluster contained samples from different water
431 layers. Similar results were obtained for many other species, including *Prochlorococcus*

(species id: 57810) and *SAR86* (species id: 59142) (Figure 5c). We found that the population structure of the 30 marine bacteria was highly consistent, regardless of the percent identity threshold used for defining pan-genome gene families (75-99% identity) (Supplemental Fig S10-11).

To evaluate the extent of gene content biogeography across species, we computed the correlation between PCA distances and geographic distances (Methods) and found significant distance-decay in gene content for the majority of species tested, including all 18 *Prochlorococcus* species (Figure 5b). Furthermore, this pattern was observed both in samples from the surface water layer and the deep chlorophyll maximum layer. A previous study found season to be a major driver of biodiversity patterns in the global ocean (Ladau et al. 2013). To explore whether season or other environmental variables were associated with strain-level population structure, we compared correlations of the first principal component of gene content (PC1) with geography and environmental variables (Supplemental Fig S12). 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 relatively little variation (4% on average) and was the most predictive covariate for only one *Prochlorococcus* species (species id: 60683).

A few species showed relatively little geographic structure. Instead they had gene content variation that correlated with depth or marine layer. The most striking example of this was an unnamed *Alphproteobacteria* species (species id: 44631) which contained two genomes in our database obtained via single-cell sequencing (Stepanauskas 2012). This species was predominantly found in the mesopelagic layer (below 200m) and increased in relative abundance with decreasing depth (Figure 5c). Looking only at mesopelagic samples, we found that the first principal component of gene content (PC1) was strongly correlated with depth (R^2=0.59) suggesting little mixing of strains of the *Alphproteobacteria* species across depths. We identified 266 genes positively correlated with depth and 316 genes that were negatively correlated with depth (FDR-corrected Spearman p-value < 0.01, Supplemental table S13). This 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. When we included samples from all marine layers, we found that samples from the mesopelagic and epipelagic zone clustered based on gene content and there was still a strong correlation between PC1 and depth (R^2=0.57) (Supplemental Fig S13).

Together, our 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, even though sampling location is not a strong predictor of species relative abundance (Sunagawa et al. 2015).

Discussion

We developed *MIDAS*, an integrated computational pipeline that quantifies bacterial strain-level gene content and SNPs, as well as species abundance, 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

481 compare hundreds of metagenomes to >30,000 reference genomes to identify genetic
482 variants present in the strains of each sample. Our publicly available software and data
483 resources will enable researchers to conduct large-scale population genetic analysis of
484 metagenomes.
485
486 This first version of *MIDAS* has several limitations. Since it currently relies on bacterial
487 reference genomes, *MIDAS* cannot quantify variation for novel species, novel genes, or
488 known species from other groups of microbes (e.g. archaea, eukaryotes, and viruses). To
489 accurately quantify strain-level gene content and SNPs, *MIDAS* requires greater than 1x and
490 10x sequencing coverage, respectively. This biases analyses towards the most abundant
491 and prevalent species in an environment. *MIDAS* was nonetheless able to capture the
492 majority of microbial species abundance across human body sites, making it well suited for
493 uncovering strain-level variation of human-associated bacteria. In contrast, other
494 environments appeared to be dominated by microbes missing from our reference database
495 of bacterial genomes sequenced to date. For this reason, it will be important to update the
496 database in the future as the number (Land et al. 2015) and diversity (Wu et al. 2009) of
497 microbial reference genomes continues to rapidly grow, as new experimental (Rinke et al.
498 2013) and computational (Nielsen et al. 2014) approaches uncover genome sequences of
499 uncultured microbes. It will also be useful to incorporate genomes from other domains of
500 life. Based on the design of our database and algorithm, *MIDAS* should scale with this
501 growth of reference data.
502
503 To illustrate the utility of *MIDAS*, we analyzed stool metagenomes from 98 mothers and
504 their infants over one year and used rare SNPs (i.e. marker alleles) to track transmission of
505 strains between hosts. Based on this analysis, we found extensive vertical transmission of
506 specific early colonizing bacteria shortly after birth, which largely persisted in the infant
507 for one year. In contrast, we found that late colonizing bacteria were often distinct from the
508 mother at the strain level, likely originating from the environment and/or other hosts.
509 Additionally we found that certain taxonomic groups, like Bacteroidia, tend to be vertically
510 transmitted, while others, like Clostridia do not, which suggests that only part of the gut
511 microbiome may be inherited. When the same metagenomes were analyzed at the species
512 level, these patterns of transmission were missed, and a false signal of increasing
513 transmission over time was detected due to convergence of the infant microbiome towards
514 a more diverse and adult-like species profile after weaning.
515
516 The bacterial taxa that tended to be transmitted from mother to infant in our analysis differ
517 from the taxa whose abundances were found to be heritable in a previous study of UK
518 twins (Goodrich et al. 2016). For example, we estimated low transmission rates for strains
519 of *Blautia* (Supplemental Table S11), but Goodrich et al. found that the abundance of
520 *Blautia* is highly heritable. Conversely, we estimated high transmission rates for strains of
521 *Bacteroides*, but this was one of the genera whose abundance was least heritable in the UK
522 twins. These differences between studies reveal some important distinctions. First,
523 heritability of taxon abundance does not necessarily imply vertical transmission. Rather,
524 twins and other related individuals may share a propensity to retain similar levels of taxa
525 that they both acquire from a variety of sources (i.e., different strains). On the other hand,
526 strains vertically transmitted at birth should result in heritability of presence if they are
527 retained. But they could be lost as the infant ages, or their abundances may not be heritable
528 even if their presence is. Finally, human genetics may not explain inter-individual
529 differences in abundance for taxa, such as *Bacteroides,* that are highly abundant and
530 prevalent, even if they are vertically transmitted.

Our 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, or the environment. One caveat of our analysis is that we did not distinguish which strains were transmitted to the infant from the mother in cases where mothers harbored multiple strains. Instead, we 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.

To explore bacterial population structure using gene content, we applied *MIDAS* to ocean samples from the *Tara* expedition. We found a number of prevalent and abundant bacterial species, which shows that our method can be applied to different environments, despite low database coverage. Based on these results, we 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 (Sunagawa et al. 2015). However, the gene content of other species found in the mesopelagic layer (200-1000m) were structured by depth. 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.

Microbiome research is in an era where metagenome-wide analyses can now pinpoint individual strains and genes that differ in presence or abundance between samples. Importantly, this level of resolution is not only revealing associations that are missed by analyses conducted at higher taxonomic levels, but also patterns that oppose those inferred from species abundance distributions. A striking example is our discovery that infants share more gut bacterial strains with their mothers at birth than later in their first year of life, despite the fact that the species composition of their microbiomes are becoming more similar as the infant ages. Without conducting a strain-level genomic analysis, one might incorrectly infer that vertical transmission of the gut microbiome is constant or increasing during the first year of life. Similarly, the high level of gene copy number variation that we observe in *Tara* oceans bacteria and its strong correlation with marine region in surface waters emphasizes functionally important differences in strains across global oceans that are missed when metagenomes are analyzed at the species level. Furthermore, it is clear that gene content of even the most prevalent and abundant marine bacterial species cannot be accurately inferred from the currently limited number of sequenced genomes for this environment. The same is true for laboratory mice, humans from non-industrialized countries, and soil. These results point to specific phylogenetic groups and environments that are highest priority for additional genome sequencing, culturing attempts, and functional assays.

Materials and Methods

579
580 **Sequence based identification of bacterial species**
581
582 We developed a procedure to cluster bacterial genomes into species groups based on the
583 pairwise percent identity across a set of universal gene families, which was inspired by
584 previous work (Mende et al. 2013). We began with 33,252 prokaryotic genomes
585 downloaded from PATRIC (Wattam et al. 2014) in March 2015. Next, we used HMMER3
586 (Eddy 2011) with an E-value threshold ≤1e-5 to identify protein homologs of 112 bacterial
587 universal gene families (Wu et al. 2013) across the genomes. The HMMER3 search took too
588 long for two gene families (B000042, B000044), which were dropped. When there were
589 multiple homologs of a gene family identified in a genome, we took the homolog with the
590 lowest E-value. We filtered out low quality genomes with fewer than 100 universal genes
591 identified (N=1,837) or with greater than 1,000 contigs (N=618), which left 31,007 high-
592 quality genomes (Supplemental Table S5).
593
594 Next, we used BLASTN (Altschul et al. 1990) to perform sequence alignment of each gene
595 family between all high-quality genomes. We filtered out local alignments where either the
596 query or target was covered by <70% of its length. We converted percent identities to
597 distances using the formula: $D_{ab} = (100 - P_{ab})/100$, where P_{ab} was the percent identity
598 of a gene between genomes *a* and *b*. This resulted in an undirected graph for each marker
599 gene family where nodes were genomes and edges were distances. We performed average-
600 linkage hierarchical clustering for each graph using the program MC-UPGMA (Loewenstein
601 et al. 2008). The output of MC-UPGMA is a tree, which we cut at different distance
602 thresholds (0.01 to 0.10). Each cut of the tree yielded a set of genome-clusters.
603
604 For validation, we compared each set of genome-clusters to average nucleotide identity
605 (ANI), which is considered to be a gold standard for delineating prokaryotic species
606 (Konstantinidis et al. 2006; Richter and Rossello-Mora 2009) but was too computationally
607 intensive to compute for all genome-pairs. Specifically, we used the procedure described by
608 Richter and Rossello-Mora (Richter and Rossello-Mora 2009) to compute ANI for >18,000
609 genome-pairs and labeled pairs of genomes with ANI ≥ 95% as members of the same
610 species and pairs of genomes with ANI < 95% as members of different species. We
611 compared these labels to our genome-clusters and classified each genome-pair into one of
612 the following categories: *True positive*: a clustered genome-pair with ANI ≥ 95%; *False*
613 *positive*: a clustered genome-pair with ANI < 95%; *False negative*: a split genome-pair with
614 ANI ≥ 95%; *True negative*: a split genome-pair with ANI < 95%. Using these classifications
615 we calculated the true positive rate (TPR), precision (PPV), and F1-score for each set of
616 genome-clusters corresponding to 90-99% identity between pairs of genomes for a given
617 marker gene (Supplemental Table S2).
618
619 Based on this evaluation, we identified a subset of 30 gene families that produced genome-
620 clusters that were in agreement with ANI, all with maximum F1-score > 0.93 across
621 thresholds. To increase clustering performance, we took the average pairwise distances
622 across these 30 gene-families and used these new distances to re-cluster genomes using
623 MC-UPGMA (Supplemental Table S3). We found that a distance cutoff of 0.035 (96.5%
624 nucleotide identity) maximized the F1-score at 0.98 and resulted in 5,952 genome-clusters
625 (Supplemental Table S3). Each genome-cluster was annotated by the most common
626 PATRIC Latin name within the cluster (Supplemental Table S4).
627
628 **Genomic database construction**

629
630 Genome-clusters (i.e. bacterial species) were leveraged in order to compile a
631 comprehensive genomic data resource used by *MIDAS*. First, we identified a representative
632 genome from each species to use for detecting core-genome SNPs. Each representative
633 genome was chosen in order to maximize its average nucleotide identity at the 30 universal
634 genes (Supplemental Table S2) to other members of the species. Next, we build a database
635 of 15 universal single-copy gene families (Supplemental Table S7) to use for estimating the
636 abundance of the species from a shotgun metagenome. Gene families were selected based
637 on their ability to accurately recruit metagenomic reads as well as being universal and
638 single-copy. Many of the 30 gene families for clustering genomes into species and the 15
639 gene families for quantifying species abundance from metagenomes were different and
640 were selected using distinct criteria. Next, we used USEARCH (Edgar 2010) to identify the
641 set of unique genes at 99% identity across all genomes within each species, which are used
642 by *MIDAS* for metagenomic pan-genome profiling. This procedure clustered 116,978,184
643 genes from the 31,007 genomes into 31,840,245 gene families. We further clustered these
644 genes at different levels of sequence identity (75-95% DNA identity) in order to identity *de*
645 *novo* gene families of varying size and diversity for downstream analyses. Functional
646 annotations for all genes were obtained from PATRIC and include FIGfams (Meyer et al.
647 2009), Gene Ontology (Consortium 2000), and KEGG Pathways (Kanehisa and Goto 2000).
648
649 **Species abundance estimation**
650
651 *MIDAS* uses reads mapped to 15 universal single-copy gene families to estimate the
652 abundance of the 5,952 bacterial species from a shotgun metagenome. These 15 gene
653 families were selected from a set of 112 phylogenetically informative bacterial gene
654 families (Wu et al. 2013) for their ability to accurately recruit metagenomic reads to the
655 correct species. To evaluate how informative different gene families are for estimation of
656 abundance, we simulated one hundred 100-bp reads from each of the 112 gene families in
657 each of the 5,952 species and used HS-BLASTN (Chen et al. 2015) to map these reads back
658 to a database that contained the full length gene sequences. To simulate the presence of
659 novel species and strains, we discarded alignments between reads and reference sequences
660 from the same species. Each read was assigned to a species based on its top hit.
661 Recruitment performance was measured using the F1-score. Based on this experiment, we
662 identified 15 universal single-copy gene families that were best able to accurately assign
663 the species from which metagenomic reads derived. Additionally, we identified the optimal
664 percent identity cutoffs for mapping reads to the database, which ranged from 94.5% to
665 98.00% identity depending on the gene family (Supplemental Table S7).
666
667 To perform taxonomic profiling, *MIDAS* aligns reads to the database of 15 universal gene
668 families with HS-BLASTN, discards local alignments that cover <70% of the read or
669 alignments that fail to satisfy the gene-specific species-level percent identity cutoffs, and
670 assigns each uniquely-mapped read to a species according to its best-hit. *MIDAS* assigns
671 non-uniquely mapped reads (i.e. identical alignment scores to genes from >1 species) using
672 probabilities estimated from uniquely mapped reads. These mapped reads are used to
673 estimate the coverage and relative abundance of each species.
674
675 **Gene content estimation**
676
677 To estimate gene content, *MIDAS* first uses the species abundance profile to identify
678 bacterial species with sufficient coverage (e.g. >1x). A pan-genome database is dynamically

679 built, which contains a set of non-redundant genes from these species. We used a 99%
680 sequence identity threshold to cluster similar genes such that any two genes that are <99%
681 similar were classified as distinct genes. Bowtie2 (Langmead and Salzberg 2012) is used to
682 locally map reads from the metagenome against the pan-genome database. Each read is
683 mapped a single time according to its best hit, and reads with an insufficient mapping
684 percent identity (default=94%), alignment coverage (default=70%), mapping quality
685 (default=20), or sequence quality (default=20) are discarded.
686
687 Mapped reads are used to compute the coverage of the genes clustered at 99% identity.
688 Since the 99% identity may result in many very similar gene families, *MIDAS* gives the
689 option of further clustering the gene families at lower sequence identities ranging from
690 75% to 95%. Aggregating enables quantification of gene families of varying size and
691 diversity, while maintaining mapping speed and sensitivity.
692
693 To estimate gene copy numbers in a bacterial population, gene coverages are normalized
694 by the median coverage across the 15 universal single-copy gene families (Supplemental
695 Table S7). Copy-number values are thresholded to produce gene presence/absence calls.
696 *MIDAS* merges these results across multiple metagenomic samples to produce gene content
697 matrices for all species, which facilitate comparative analyses across genes and
698 metagenomic samples.
699
700 **Identifying core genome SNPs**
701
702 To estimate core genome SNPs, *MIDAS* first uses the species abundance profile to identify
703 species with sufficient coverage (e.g. >10x). A representative genome database is
704 dynamically built, which contains a single genome per species that meets the coverage
705 requirement. The representative genome is a single genome chosen that has the greatest
706 nucleotide identity, on average, to other members of the species. Only a single genome is
707 needed for identifying the core genome, because this region should be present in all strains
708 of a species. Bowtie2 is used to globally map reads to the representative genome database.
709 Each read is mapped a single time according to its best hit, and reads with an insufficient
710 mapping percent identity (default=94%), alignment coverage (default=70%), mapping
711 quality (default=20), or sequence quality (default=20) are discarded. Additionally, bases
712 with low sequence quality scores are discarded (default=30). Samtools (Li et al. 2009) is
713 used to generate a pileup of nucleotides at each genomic position and generate VCF files.
714 VCF files are parsed to generate output files that report nucleotide variation statistics at all
715 genomic sites. To identify the core genome of a species, *MIDAS* uses the output from
716 multiple metagenomic samples to identify regions at consistently high coverage (e.g. >10x
717 coverage in 95% of samples) (Supplemental Fig S6). *MIDAS* then produces core genome
718 SNP matrices for all species, which facilitate comparative analyses of nucleotide variation
719 across genomic sites and metagenomic samples. *MIDAS* also gives the option of outputting
720 all SNPs, including those that are not in the core genome.
721
722 **Shotgun simulations and validation of *MIDAS* output**
723
724 To validate *MIDAS* we designed a series of realistic metagenomic simulations using reads
725 from completed genome-sequencing projects deposited in the NCBI Sequence Read Archive
726 (Leinonen et al. 2011) which we identified using the SRAdb tool (Yuelin Zhu 2013). We
727 used this data to construct 20 mock metagenomes, which each contained 100-bp Illumina
728 reads from 20 randomly selected bacterial genome projects (Supplemental Tables S8-9).

729 We only selected genome projects that corresponded to one of the 31,007 genomes present
730 in our reference database, and we used only one genome project per selected species. We
731 simulated libraries that contained 100x total genome coverage. The relative abundances of
732 the 20 genomes were exponentially distributed in each simulation (50%, 25%, 12%, 6.5%
733 etc.).
734
735 We compared the output of *MIDAS* to the known species abundance, gene content, and
736 SNPs in the simulated communities. To evaluate the accuracy of species abundance
737 estimation we compared the expected relative abundance and coverage to the simulated
738 relative abundance and coverage. To evaluate the accuracy of gene content estimation, we
739 ran *MIDAS* to estimate the copy-number of genes in the pan-genome of each species in each
740 simulation. We applied a cutoff to these values to predict gene presence/absence. True
741 positives (TP) were present genes predicted as present, false positives (FP) were absent
742 genes predicted as present, true negatives (TN) were absent genes predicted as absent, and
743 false negatives (FN) were present genes predicted as absent. Performance was measured
744 across a range of copy-number cutoffs using balanced accuracy: (TPR+TNR)/2, where
745 TPR=TP/(TP+ FN) and TNR=TN/(TN+FP). To evaluate the accuracy of core genome SNPs,
746 we ran *MIDAS* to estimate the frequency of nucleotide variants in the representative
747 genome of each species in each simulation. We predicted SNPs using the consensus allele at
748 each genomic position. True SNPs were identified by comparing genomes in the
749 simulations to the representative genomes used for read mapping with the program
750 MUMmer (Stefan Kurtz 2004), which identified 3,971,528 total true SNPs. True positives
751 were correctly called SNPs, false positives were incorrectly called SNPs, and false negatives
752 were SNPs that were not called due to insufficient coverage. We compared predicted SNPs
753 to true SNPs and measured performance using the true positive rate (TP/TP+FN) and
754 precision (TP/TP+FP).
755
756 **Assessing database coverage across different environments**
757
758 We estimated the species-level coverage of the *MIDAS* database across metagenomes from
759 different environments. Database coverage is defined as the percent (0 to 100%) of
760 genomes from cellular organisms in a community that have a sequenced representative at
761 the species level in the reference database. We estimated database coverage by (i)
762 computing the total coverage across all species in the *MIDAS* database by mapping
763 metagenomic reads to 15 universal single-copy genes and applying species-level mapping
764 thresholds, (ii) computing the coverage across all microbial species, including those absent
765 from the *MIDAS* reference database using the tool *MicrobeCensus* (Nayfach and Pollard
766 2015), and (iii) taking the ratio of these two quantities, multiplied by 100. We applied this
767 approach to metagenomes from human body sites (Consortium 2012), human stool (Qin et
768 al. 2012; Li et al. 2014; Obregon-Tito et al. 2015; Rampelli et al. 2015), baboon stool (Tung
769 et al. 2015), mouse stool (Xiao et al. 2015), ocean water (Sunagawa et al. 2015), and soil
770 from deserts, forests, grasslands, and tundra (Fierer et al. 2012). To identify possible
771 taxonomic groups that harbored novel species in the human gut, we performed Spearman
772 correlations between database coverage and the relative abundance of genera, estimated
773 using mOTU (Sunagawa et al. 2013).
774
775 **Tracking transmission of strains between hosts**
776
777 We used marker alleles to track transmission of strains between hosts. We defined a
778 marker allele as an allele at a genomic site that was present in only a single individual, or in

779 the case of the mother-infant dataset, a single mother-infant pair. For simplicity, we only
780 considered bi-allelic genomic sites. An allele was determined to be present in a sample if it
781 was supported by ≥3 reads and ≥10% of the total reads mapped at the genomic site. These
782 parameters were chosen to minimize the effect of sequencing errors and filter out low
783 frequency variants that could not be consistently detected between samples. Marker allele
784 sharing was computed as the percent of marker alleles in mother strains that were also
785 found in her infant. To minimize variation in marker allele sharing due to sampling, we
786 excluded individuals with fewer than 10 identified marker alleles for a species. We applied
787 this procedure to 66 species found in stool metagenomes from 98 Swedish mothers and
788 their infants (Backhed et al. 2015) as well as 123 American individuals from the Human
789 Microbiome Project (HMP) (Consortium 2012). We included the HMP samples in order to
790 increase sample sizes and therefore improve the specificity of marker alleles identified in
791 mothers and their infants. As a positive control to assess the sensitivity of our approach, we
792 quantified marker allele sharing for each species between pairs of technical replicates from
793 the HMP. As a negative control to assess specificity, we quantified marker allele sharing for
794 each species between pairs of unrelated individuals from the HMP, which were not used to
795 identify marker alleles. Based on these results, we defined a transmission event as >5%
796 marker allele sharing between a pair of individuals.
797
798 **Analysis of globally distributed marine metagenomes**
799
800 To assess the global population structure of marine bacteria, we analyzed 198 shotgun
801 metagenomes collected from the *Tara* Oceans expeditions that corresponded to
802 prokaryotic size fractions. We utilized up to 100 million reads per metagenome and
803 analyzed only one sequencing replicate per sample. We used *MIDAS* to quantify the relative
804 abundance of the 5,952 reference species, and based on these results identified 30 species
805 that occurred at >3x sequencing depth in the greatest number of metagenomes. The least
806 prevalent species was found in 23% of metagenomes. Next, we used *MIDAS* to quantify the
807 gene content of these species across metagenomic samples. Reads were mapped to the pan-
808 genome database, and reads with <94% alignment identity were discarded. Mapped reads
809 were used to compute the coverage of genes clustered at 95% identity. Gene coverages
810 were normalized by the coverage of 15 universal single copy genes to estimate gene copy
811 numbers. We estimated gene presence-absence by thresholding the gene copy numbers,
812 whereby any gene with a copy number <0.35 was considered to be absent.
813
814 To uncover population structure, we performed a principle component analysis of the gene
815 presence-absence matrix for each species. To assess the relationship between gene content
816 and geography, we first quantified the PCA distance and geographic distance between
817 metagenomic samples for each species. PCA distances were computed using the Euclidian
818 distance between samples based on the first two principle components. Geographic
819 distances were computed using the Great circle distance with the R package geosphere
820 (Hijmans 2016). Mantel tests were computed using the R package vegan (Oksanen et al.
821 2016) to correlate the PCA distances to the geographic distances. Up to one million
822 permutations were performed to assess significance.
823
824 # Figure legends
825

826 **Figure 1. Construction of bacterial species database and its coverage of microbial**
827 **communities across different environments**
828 **A)** 31,007 genomes were hierarchically clustered based on the pairwise identity across a
829 panel of 30 universal gene families. We identified 5,952 species groups by applying a
830 96.5% nucleotide identity cutoff across universal genes, which is equivalent to 95%
831 identity genome wide. **B)** Concordance of genome-cluster names and annotated species
832 names. Of the 31,007 genomes assigned to a genome-cluster, 5,701 (18%) disagreed with
833 the consensus PATRIC taxonomic label of the genome-cluster. Most disagreements are due
834 to genomes lacking annotation at the species level (47%). Other disagreements are because
835 a genome was split from a larger cluster with the same name (29%) or assigned to a
836 genome-cluster with a different name (24%). **C)** Coverage of the species database across
837 metagenomes from host-associated, marine, and terrestrial environments. Coverage is
838 defined as the percent (0 to 100%) of genomes from cellular organisms in a community
839 that have a sequenced representative at the species level in the reference database. Inset
840 panel shows the distribution of database coverage across human stool metagenomes from
841 six countries and two host lifestyles.
842
843 **Figure 2. An integrated pipeline for profiling species abundance and strain-level**
844 **genomic variation from metagenomes**
845 **A)** The *MIDAS* analysis pipeline. Reads are first aligned to a database of universal-single-
846 copy genes to estimate species coverage and relative abundance per sample. For species
847 with sufficient coverage, reads are next aligned to a pan-genome database of genes to
848 estimate gene coverage, copy-number, and presence/absence. Finally, reads are aligned to
849 a representative genome database to detect SNPs in the core genome. The core genome is
850 defined directly from the data by identifying high coverage regions across multiple
851 metagenomic samples. **(B-D)** To evaluate performance for each component of MIDAS, we
852 analyzed 20 mock metagenomes composed of 100-bp Illumina reads from microbial
853 genome-sequencing projects. Each community contained 20 organisms with exponentially
854 decreasing relative abundance. We tested the ability of MIDAS to estimate species coverage
855 and to predict genes and SNPs present in the strains of the mock communities compared to
856 the reference gene and genome databases. **B)** Species coverage is accurately estimated.
857 Each boxplot indicates the distribution of estimated genome coverages across 20 mock
858 communities for the top 8 most abundant species out of 20 analyzed. **C)** Gene
859 presence/absence is accurately predicted when genome coverage is above 1x and a gene
860 copy number cutoff of 0.35 is used. Accuracy = (Sensitivity + Specificity)/2; Sensitivity = (#
861 genes correctly predicted as present)/(# total genes present); Specificity = (# genes
862 correctly predicted as absent)/(# total genes absent). **D)** SNPs are detected with a low false
863 discovery rate and good sensitivity when genome coverage is above 10x. Sensitivity = (#
864 correctly called SNPs)/(# total SNPs); False Discovery Rate = (# incorrectly called
865 SNPs)/(# called SNPs).
866
867 **Figure 3. An increase in shared species but decrease in shared strains over time**
868 **between stool metagenomes from mothers and their infants**
869 **A)** Principal coordinate analysis of Bray-Curtis dissimilarity between species relative
870 abundance profiles of stool samples from mothers and infants at 4 days, 4 months, and 12
871 months following birth. Species composition of infant microbiomes is most similar to
872 mothers at 12 months. **B)** The number of shared species increases over time between
873 mothers and their own infants. **C)** This pattern for biological mother-infant pairs is similar
874 to that of unrelated mothers and infants (permuted pairs). **D)** In contrast, marker allele
875 sharing decreases over time between mothers and their infants for shared species with

876 >10x sequencing coverage, indicating highest strain similarity at 4 days. Allele sharing is
877 defined as the percent of marker alleles in the mother that are found in the infant. The
878 horizontal red line indicates the 5% marker allele threshold used for defining vertical
879 transmission events. **E)** Vertical transmissions for 20 bacterial species across mother-
880 infant pairs at three time points. A vertical transmission is defined as >5% marker allele
881 sharing between mother and infant. The phylogenetic tree on the left is constructed based
882 on a concatenated DNA alignment of 30 universal genes (Supplemental Fig S3). The tree
883 shows that species with more vertical transmission are phylogenetically clustered.
884
885 **Figure 4. Strains that colonize the infant gut late in the first year are rarely found in**
886 **the mother**
887 **A)** Vertical transmission of strains at 12 months versus their relative abundance at 4 days.
888 Each data point indicates a shared species found in a mother and her infant. The vertical
889 axis indicates whether a strain of the species was transmitted from the mother (y=1) or not
890 (y=0). The curve is a logistic regression fitted to data points. Histograms indicate the
891 distribution of relative abundance at 4 days for strains that were transmitted and not
892 transmitted. Strains with high relative abundance in infants at 4 days are frequently
893 transmitted from the mother. Strains that appear in the infant at later time points (i.e. low
894 abundance at 4 days) are rarely transmitted from the mother. **B)** Examples of early
895 colonizing species that are frequently transmitted vertically (*Bacteroides vulgatus,*
896 *Parabacteroides distasonis,* and *Bacteroides fragilis*) and late colonizing species that are
897 rarely transmitted vertically (*Eubacterium rectale, Faecalibacterium prausnitzii,* and *Blautia*
898 *wexlerae*). Points are colored based on vertical transmission. Gray indicates there was
899 insufficient sequencing coverage to quantify SNPs and determine whether the strain was
900 transmitted from the mother or not.
901
902 **Figure 5. Gene content and geography are correlated for many marine bacteria**
903 **A)** Prevalence of 50 bacterial species across 198 ocean metagenomes. Species identifiers
904 are indicated in parenthesis. Many marine species have sufficient sequencing depth and
905 prevalence for population genetic analyses. **B)** Correlation of PCA and geographic distance
906 between pairs of samples for different marine species. PCA distance was calculated using
907 the Euclidian distance between PC1 and PC2 of the gene presence/absence matrix.
908 Geographic distance was calculated using the great-circle distance between sampling
909 locations. Correlations and p-values were computed using the Mantel test with 1 million
910 permutations. Only one metagenome per location was included in the tests. The population
911 structure of marine bacteria, based on the first two principal components of gene content,
912 is correlated with geography for many species of bacteria. **C)** *Candidatus Pelagibacter* and
913 SAR86 are examples of marine bacterial species geographically structured populations
914 based on gene presence/absence. Each point indicates a bacterial population from a
915 different sample. Point colors indicate the marine region. Point shapes indicate the ocean
916 depth. Samples tended to cluster together based on ocean region, not ocean depth. **D)**
917 *Alpha proteobacterium* is an example of a species present in the mesopelagic zone where
918 depth is the major driver of population structure.
919
920

921 Software availability

922 *MIDAS* is implemented in Python and is freely available, along with documentation at:
923 **https://github.com/snayfach/MIDAS**. Source code is additionally included as

924 Supplemental Material. Our reference database of bacterial species and associated genomic
925 data resources are available at: **http://lighthouse.ucsf.edu/MIDAS**
926

927 Acknowledgements

928 We thank Jonathan Eisen and Dongying Wu for helpful discussions about marker genes and
929 phylogenetics and Joshua Ladau for discussions about microbial biogeography. This project
930 was supported by funding from NSF grant #DMS-1069303, Gordon & Betty Moore
931 Foundation grant #3300, the San Simeon Fund, and institutional funds from Gladstone
932 Institutes.
933

934 References

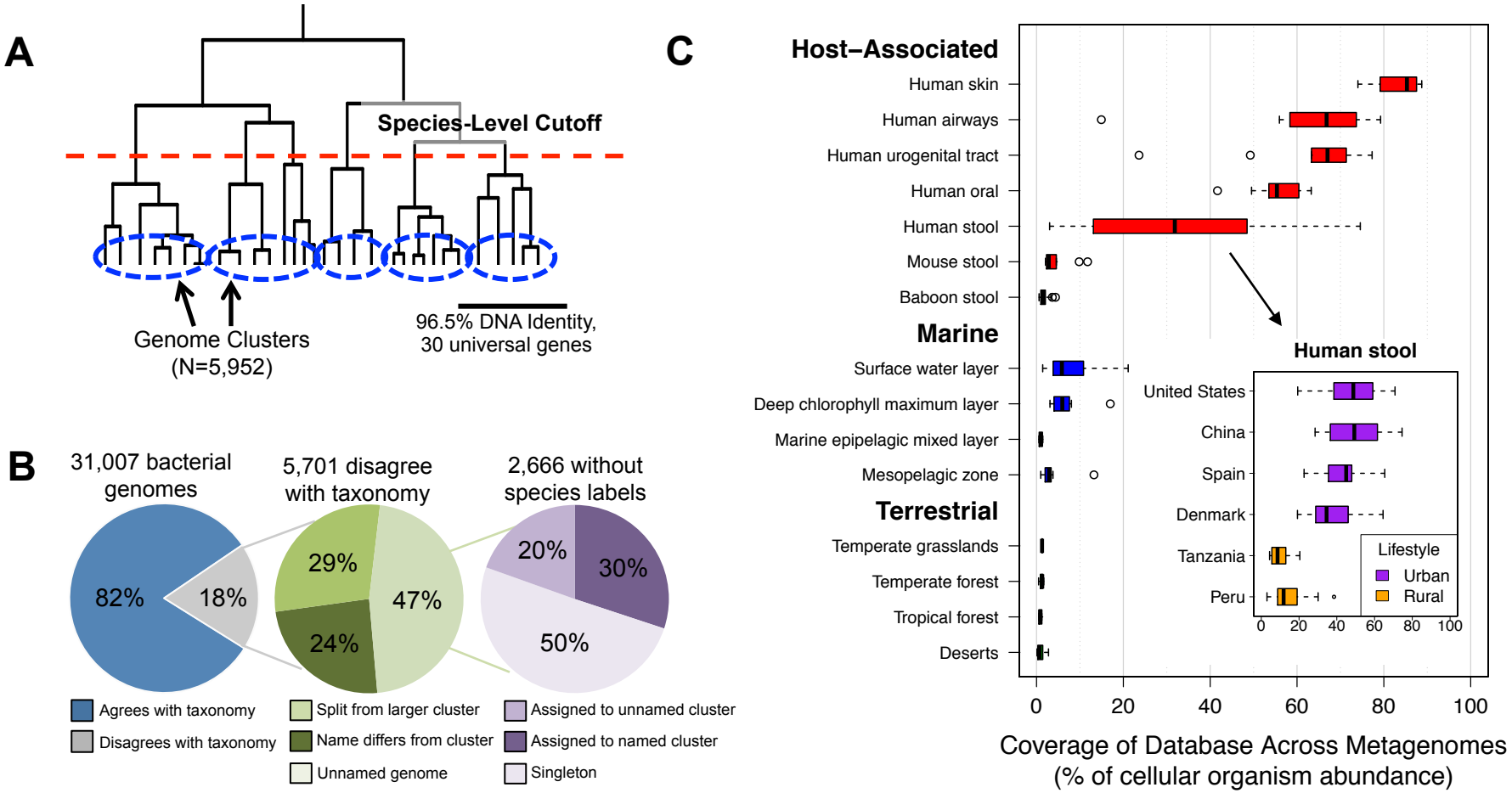
935

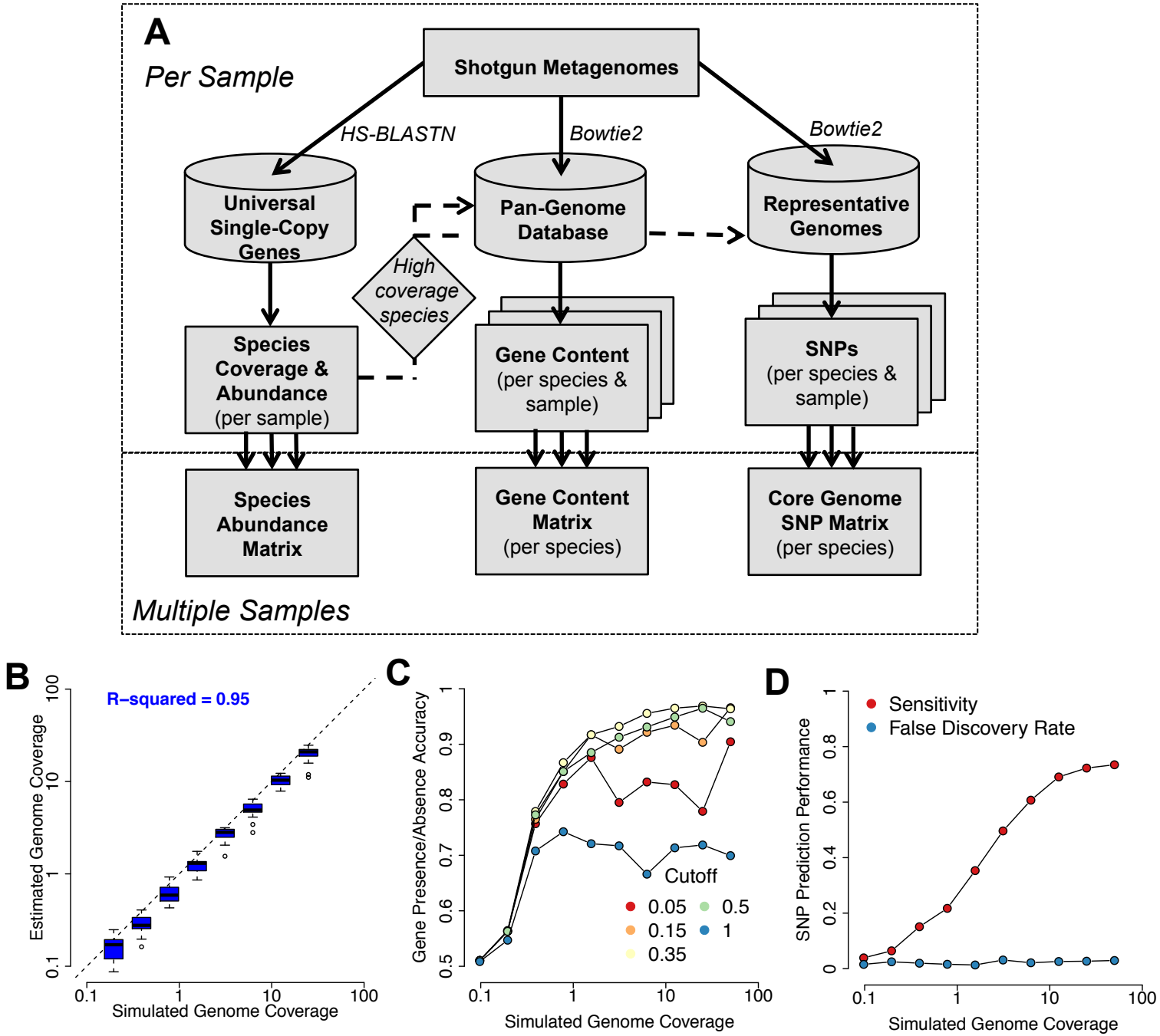
936 Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. 1990. Basic local alignment search
937 tool. *Journal of Molecular Biology* **215**(3): 403-410.
938 Backhed F, Roswall J, Peng Y, Feng Q, Jia H, Kovatcheva-Datchary P, Li Y, Xia Y, Xie H, Zhong
939 H et al. 2015. Dynamics and Stabilization of the Human Gut Microbiome during the
940 First Year of Life. *Cell host & microbe* **17**(5): 690-703.
941 Browne HP, Forster SC, Anonye BO, Kumar N, Neville BA, Stares MD, Goulding D, Lawley
942 TD. 2016. Culturing of 'unculturable' human microbiota reveals novel taxa and
943 extensive sporulation. *Nature* **533**(7604): 543-546.
944 Chen Y, Ye W, Zhang Y, Xu Y. 2015. High speed BLASTN: an accelerated MegaBLAST search
945 tool. *Nucleic Acids Res*.
946 Cleary B, Brito IL, Huang K, Gevers D, Shea T, Young S, Alm EJ. 2015. Detection of low-
947 abundance bacterial strains in metagenomic datasets by eigengenome partitioning.
948 *Nature biotechnology* **33**(10): 1053-1060.
949 Consortium TGO. 2000. Gene Ontology: tool for the unification of biology. *Nature genetics*
950 **25**.
951 Consortium THMP. 2012. A framework for human microbiome research. *Nature*
952 **486**(7402): 215-221.
953 Eddy SR. 2011. Accelerated Profile HMM Searches. *PLoS computational biology* **7**(10):
954 e1002195.
955 Edgar RC. 2010. Search and clustering orders of magnitude faster than BLAST.
956 *Bioinformatics* **26**(19): 2460-2461.
957 Faith JJ, Guruge JL, Charbonneau M, Subramanian S, Seedorf H, Goodman AL, Clemente JC,
958 Knight R, Heath AC, Leibel RL et al. 2013. The long-term stability of the human gut
959 microbiota. *Science* **341**(6141): 1237439.
960 Fierer N, Leff JW, Adams BJ, Nielsen UN, Bates ST, Lauber CL, Owens S, Gilbert JA, Wall DH,
961 Caporaso JG. 2012. Cross-biome metagenomic analyses of soil microbial
962 communities and their functional attributes. *Proceedings of the National Academy of*
963 *Sciences*.
964 Francis OE, Bendall M, Manimaran S, Hong C, Clement NL, Castro-Nallar E, Snell Q, Schaalje
965 GB, Clement MJ, Crandall KA et al. 2013. Pathoscope: species identification and
966 strain attribution with unassembled sequencing data. *Genome research* **23**(10):
967 1721-1729.
968 Goodrich JK, Davenport ER, Beaumont M, Jackson MA, Knight R, Ober C, Spector TD, Bell JT,
969 Clark AG, Ley RE. 2016. Genetic Determinants of the Gut Microbiome in UK Twins.
970 *Cell host & microbe* **19**(5): 731-743.

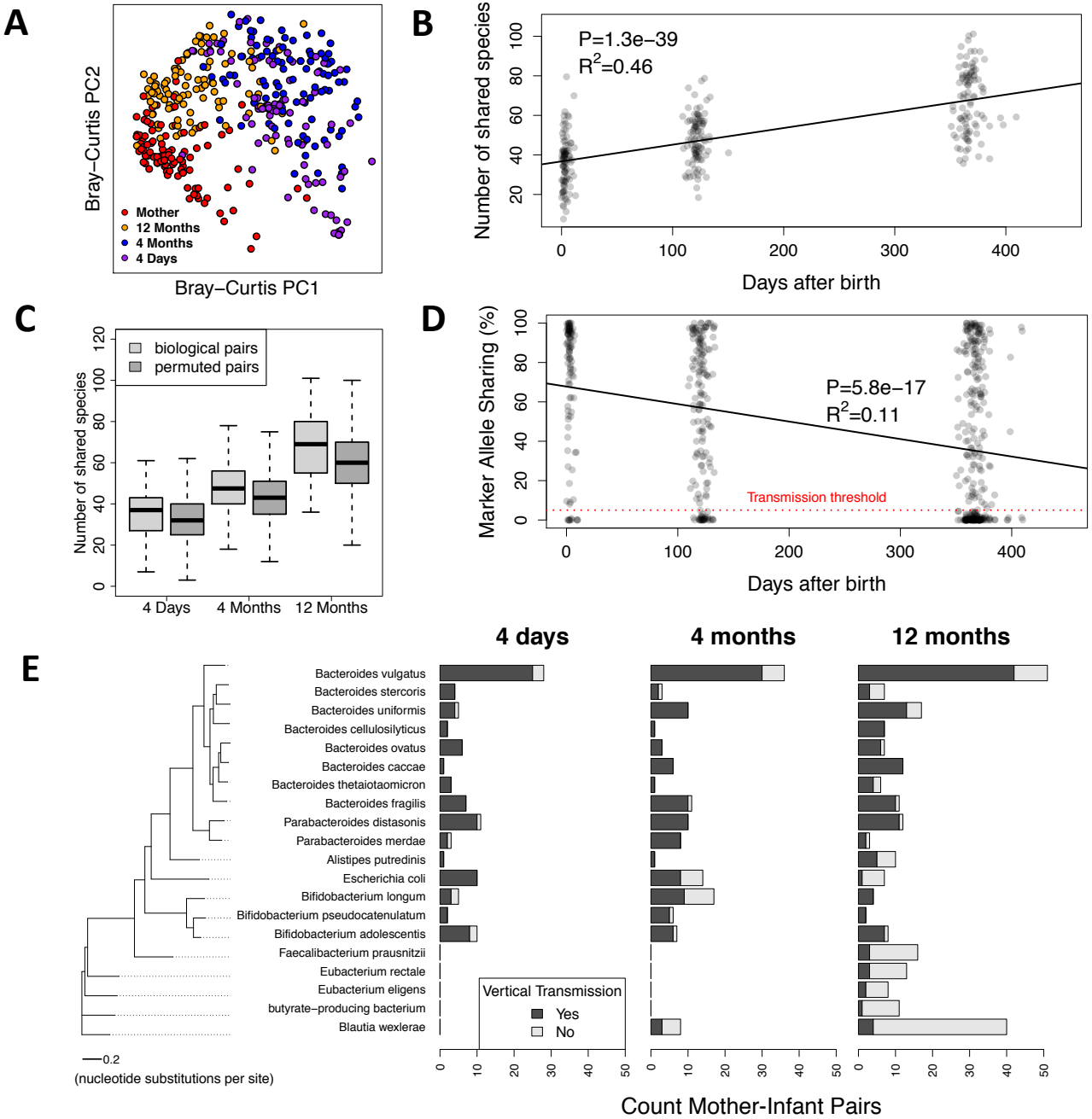
971 Greenblum S, Carr R, Borenstein E. 2015. Extensive strain-level copy-number variation
972 across human gut microbiome species. *Cell* **160**(4): 583-594.
973 Hijmans RJ. 2016. geosphere: Spherical Trigonometry.
974 Kanehisa M, Goto S. 2000. KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids*
975 *Research* **28**(1).
976 Kashtan N, Roggensack SE, Rodrigue S, Thompson JW, Biller SJ, Coe A, Ding H, Marttinen P,
977 Malmstrom RR, Stocker R et al. 2014. Single-Cell Genomics Reveals Hundreds of
978 Coexisting Subpopulations in Wild Prochlorococcus. *Science* **344**.
979 Konstantinidis KT, Ramette A, Tiedje JM. 2006. The bacterial species definition in the
980 genomic era. *Philosophical transactions of the Royal Society of London Series B,*
981 *Biological sciences* **361**(1475): 1929-1940.
982 Ladau J, Sharpton TJ, Finucane MM, Jospin G, Kembel SW, O'Dwyer J, Koeppel AF, Green JL,
983 Pollard KS. 2013. Global marine bacterial diversity peaks at high latitudes in winter.
984 *The ISME journal* **7**(9): 1669-1677.
985 Land M, Hauser L, Jun SR, Nookaew I, Leuze MR, Ahn TH, Karpinets T, Lund O, Kora G,
986 Wassenaar T et al. 2015. Insights from 20 years of bacterial genome sequencing.
987 *Functional & integrative genomics* **15**(2): 141-161.
988 Langmead B, Salzberg SL. 2012. Fast gapped-read alignment with Bowtie 2. *Nature methods*
989 **9**(4): 357-359.
990 Leinonen R, Sugawara H, Shumway M, International Nucleotide Sequence Database C. 2011.
991 The sequence read archive. *Nucleic Acids Res* **39**(Database issue): D19-21.
992 Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R,
993 Genome Project Data Processing S. 2009. The Sequence Alignment/Map format and
994 SAMtools. *Bioinformatics* **25**(16): 2078-2079.
995 Li J, Jia H, Cai X, Zhong H, Feng Q, Sunagawa S, Arumugam M, Kultima JR, Prifti E, Nielsen T
996 et al. 2014. An integrated catalog of reference genes in the human gut microbiome.
997 *Nature biotechnology* **32**(8): 834-841.
998 Li SS, Zhu A, Benes V, Costea PI, Hercog R, Hildebrand F, Huerta-Cepas J, Nieuwdorp M,
999 Salojarvi J, Voigt AY et al. 2016. Durable coexistence of donor and recipient strains
1000 after fecal microbiota transplantation. *Science* **352**(6285): 586-589.
1001 Lieberman TD, Flett KB, Yelin I, Martin TR, McAdam AJ, Priebe GP, Kishony R. 2014. Genetic
1002 variation of a bacterial pathogen within individuals with cystic fibrosis provides a
1003 record of selective pressures. *Nature genetics* **46**(1): 82-87.
1004 Loewenstein Y, Portugaly E, Fromer M, Linial M. 2008. Efficient algorithms for accurate
1005 hierarchical clustering of huge datasets: tackling the entire protein space.
1006 *Bioinformatics* **24**(13): i41-49.
1007 Luo C, Knight R, Siljander H, Knip M, Xavier RJ, Gevers D. 2015. ConStrains identifies
1008 microbial strains in metagenomic datasets. *Nature biotechnology*.
1009 Makino H, Kushiro A, Ishikawa E, Kubota H, Gawad A, Sakai T, Oishi K, Martin R, Ben-Amor
1010 K, Knol J et al. 2013. Mother-to-infant transmission of intestinal bifidobacterial
1011 strains has an impact on the early development of vaginally delivered infant's
1012 microbiota. *PloS one* **8**(11): e78331.
1013 Martin V, Maldonado-Barragan A, Moles L, Rodriguez-Banos M, Campo RD, Fernandez L,
1014 Rodriguez JM, Jimenez E. 2012. Sharing of bacterial strains between breast milk and
1015 infant feces. *Journal of human lactation : official journal of International Lactation*
1016 *Consultant Association* **28**(1): 36-44.
1017 Mende DR, Sunagawa S, Zeller G, Bork P. 2013. Accurate and universal delineation of
1018 prokaryotic species. *Nature methods* **10**(9): 881-884.
1019 Meyer F, Overbeek R, Rodriguez A. 2009. FIGfams: yet another set of protein families.
1020 *Nucleic Acids Res* **37**(20): 6643-6654.

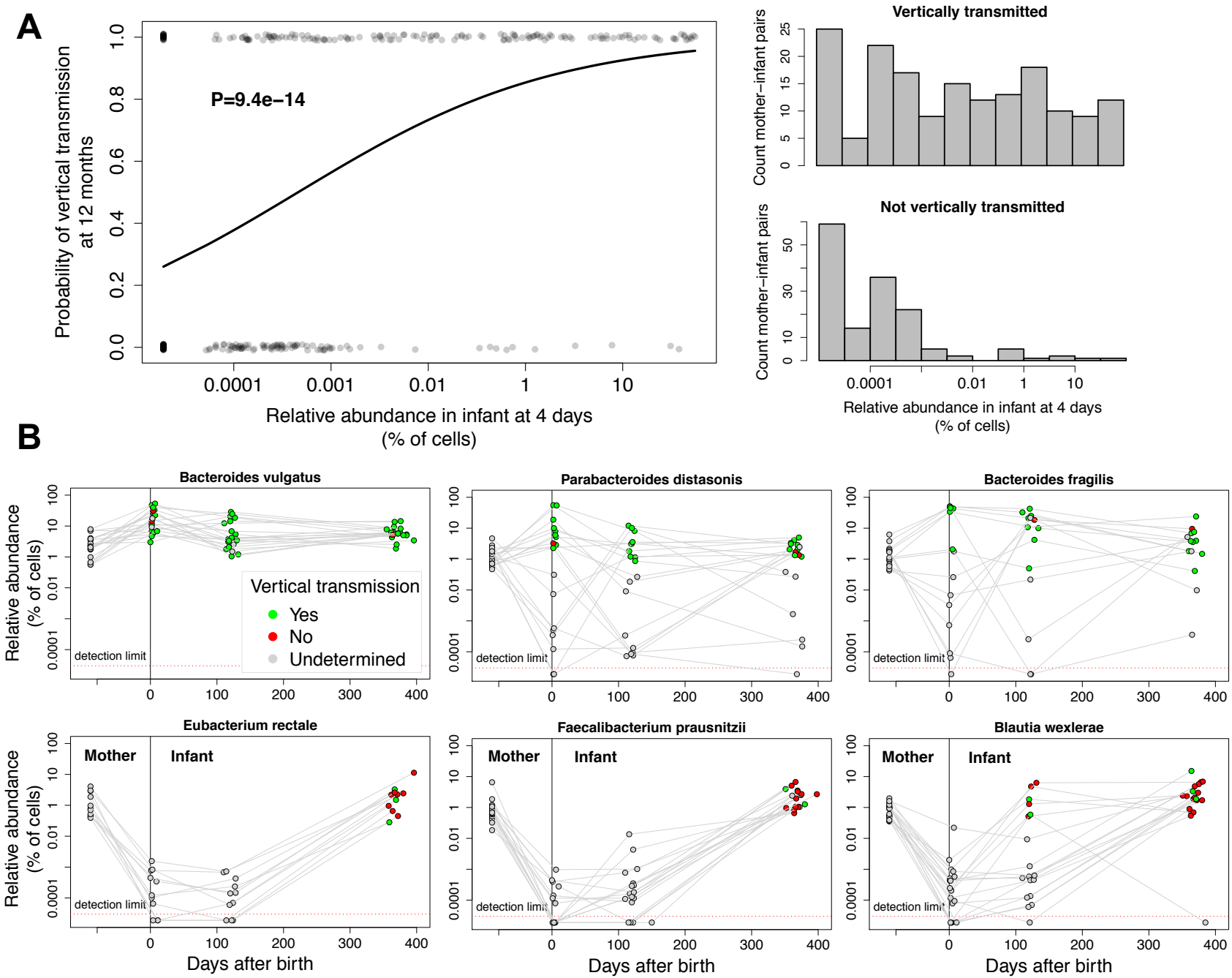
1021 Nayfach S, Pollard KS. 2015. Average genome size estimation improves comparative
1022 metagenomics and sheds light on the functional ecology of the human microbiome.
1023 *Genome biology* **16**(1): 51.
1024 Nielsen HB, Almeida M, Juncker AS, Rasmussen S, Li J, Sunagawa S, Plichta DR, Gautier L,
1025 Pedersen AG, Le Chatelier E et al. 2014. Identification and assembly of genomes and
1026 genetic elements in complex metagenomic samples without using reference
1027 genomes. *Nature biotechnology* **32**(8): 822-828.
1028 Novembre J, Johnson T, Bryc K, Kutalik Z, Boyko AR, Auton A, Indap A, King KS, Bergmann S,
1029 Nelson MR et al. 2008. Genes mirror geography within Europe. *Nature* **456**(7218):
1030 98-101.
1031 Obregon-Tito AJ, Tito RY, Metcalf J, Sankaranarayanan K, Clemente JC, Ursell LK, Zech Xu Z,
1032 Van Treuren W, Knight R, Gaffney PM et al. 2015. Subsistence strategies in
1033 traditional societies distinguish gut microbiomes. *Nature communications* **6**: 6505.
1034 Oksanen J, Blanchet F, Friendly M, Kindt R, Legendre P, McGlinn D, Minchin P, O'Hara RB,
1035 Simpson GL, Solymos P et al. 2016. vegan: Community Ecology Package.
1036 Qin J, Li Y, Cai Z, Li S, Zhu J, Zhang F, Liang S, Zhang W, Guan Y, Shen D et al. 2012. A
1037 metagenome-wide association study of gut microbiota in type 2 diabetes. *Nature*
1038 **490**(7418): 55-60.
1039 Rampelli S, Schnorr SL, Consolandi C, Turroni S, Severgnini M, Peano C, Brigidi P,
1040 Crittenden AN, Henry AG, Candela M. 2015. Metagenome Sequencing of the Hadza
1041 Hunter-Gatherer Gut Microbiota. *Current biology : CB* **25**(13): 1682-1693.
1042 Rasko DA, Webster DR, Sahl JW, Bashir A, Boisen N, Scheutz F, Paxinos EE, Sebra R, Chin CS,
1043 Iliopoulos D et al. 2011. Origins of the E. coli strain causing an outbreak of
1044 hemolytic-uremic syndrome in Germany. *The New England journal of medicine*
1045 **365**(8): 709-717.
1046 Richter M, Rossello-Mora R. 2009. Shifting the genomic gold standard for the prokaryotic
1047 species definition. *Proceedings of the National Academy of Sciences of the United*
1048 *States of America* **106**(45): 19126-19131.
1049 Rinke C, Schwientek P, Sczyrba A, Ivanova NN, Anderson IJ, Cheng JF, Darling A, Malfatti S,
1050 Swan BK, Gies EA et al. 2013. Insights into the phylogeny and coding potential of
1051 microbial dark matter. *Nature* **499**(7459): 431-437.
1052 Rosen MJ, Davison M, Bhaya D, Fisher DS. 2015. Fine-scale diversity and extensive
1053 recombination in a quasisexual bacterial population occupying a broad niche.
1054 *Science* **348**(6238).
1055 Sahl JW, Schupp JM, Rasko DA, Colman RE, Foster JT, Keim P. 2015. Phylogenetically typing
1056 bacterial strains from partial SNP genotypes observed from direct sequencing of
1057 clinical specimen metagenomic data. *Genome medicine* **7**(1): 52.
1058 Schloissnig S, Arumugam M, Sunagawa S, Mitreva M, Tap J, Zhu A, Waller A, Mende DR,
1059 Kultima JR, Martin J et al. 2013. Genomic variation landscape of the human gut
1060 microbiome. *Nature* **493**(7430): 45-50.
1061 Schnorr SL, Candela M, Rampelli S, Centanni M, Consolandi C, Basaglia G, Turroni S, Biagi E,
1062 Peano C, Severgnini M et al. 2014. Gut microbiome of the Hadza hunter-gatherers.
1063 *Nature communications* **5**: 3654.
1064 Scholz M, Ward DV, Pasolli E, Tolio T, Zolfo M, Asnicar F, Truong DT, Tett A, Morrow AL,
1065 Segata N. 2016. Strain-level microbial epidemiology and population genomics from
1066 shotgun metagenomics. *Nature methods* **13**(5): 435-438.
1067 Shapiro BJ, Friedman J, Cordero OX, Preheim SP, Timberlake SC, Szabo G, Polz MF, Alm EJ.
1068 2012. Population genomics of early events in the ecological differentiation of
1069 bacteria. *Science* **336**(6077): 48-51.

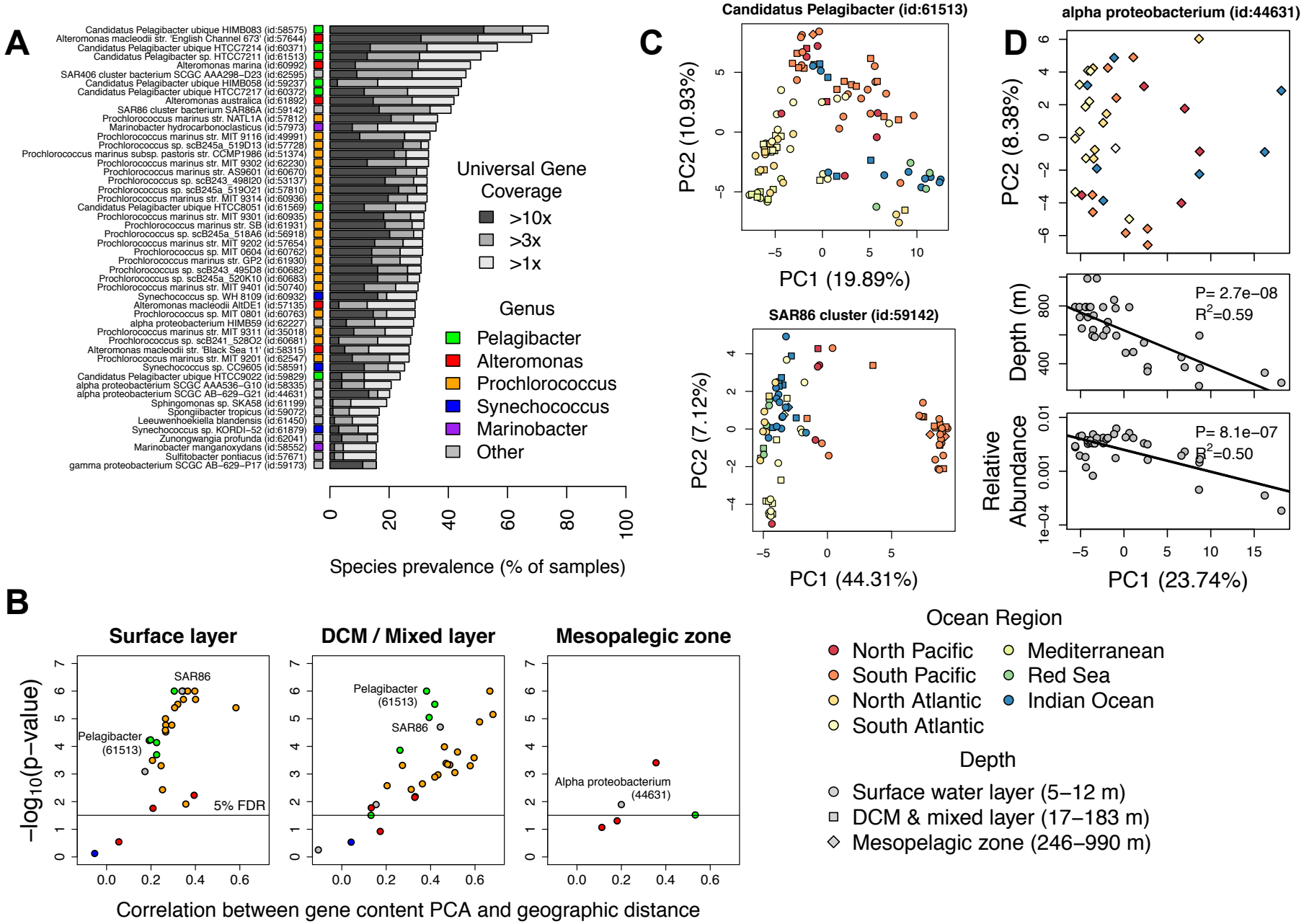
1070 Snitkin ES, Zelazny AM, Montero CI, Stock F, Mijares L, NISC-Comparative-Sequence-
1071 Program, Murray PR, Segre JA. 2011. Genome-wide recombination drives
1072 diversification of epidemic strains of Acinetobacter baumannii. *PNAS* **108**(33).
1073 Stefan Kurtz AP, Arthur L Delcher, Michael Smoot, Martin Shumway, Corina Antonescu and
1074 Steven L Salzberg. 2004. Versatile and open software for comparing large genomes.
1075 *Genome biology* **5**(2).
1076 Stepanauskas R. 2012. Single cell genomics: an individual look at microbes. *Current opinion*
1077 *in microbiology* **15**(5): 613-620.
1078 Sunagawa S, Coelho LP, Chaffron S, Kultima JR, Labadie K, Salazar G, Djahanschiri B, Zeller
1079 G, Mende DR, Alberti A et al. 2015. Ocean plankton. Structure and function of the
1080 global ocean microbiome. *Science* **348**(6237): 1261359.
1081 Sunagawa S, Mende DR, Zeller G, Izquierdo-Carrasco F, Berger SA, Kultima JR, Coelho LP,
1082 Arumugam M, Tap J, Nielsen HB et al. 2013. Metagenomic species profiling using
1083 universal phylogenetic marker genes. *Nature methods* **10**(12): 1196-1199.
1084 Tu Q, He Z, Zhou J. 2014. Strain/species identification in metagenomes using genome-
1085 specific markers. *Nucleic Acids Res* **42**(8): e67.
1086 Tung J, Barreiro LB, Burns MB, Grenier JC, Lynch J, Grieneisen LE, Altmann J, Alberts SC,
1087 Blekhman R, Archie EA. 2015. Social networks predict gut microbiome composition
1088 in wild baboons. *eLife* **4**.
1089 Varghese NJ, Mukherjee S, Ivanova N, Konstantinidis KT, Mavrommatis K, Kyrpides NC, Pati
1090 A. 2015. Microbial species delineation using whole genome sequences. *Nucleic Acids*
1091 *Res*.
1092 Wattam AR, Abraham D, Dalay O, Disz TL, Driscoll T, Gabbard JL, Gillespie JJ, Gough R, Hix
1093 D, Kenyon R et al. 2014. PATRIC, the bacterial bioinformatics database and analysis
1094 resource. *Nucleic Acids Res* **42**(Database issue): D581-591.
1095 Wu D, Hugenholtz P, Mavromatis K, Pukall R, Dalin E, Ivanova NN, Kunin V, Goodwin L, Wu
1096 M, Tindall BJ et al. 2009. A phylogeny-driven genomic encyclopaedia of Bacteria and
1097 Archaea. *Nature* **462**(7276): 1056-1060.
1098 Wu D, Jospin G, Eisen JA. 2013. Systematic identification of gene families for use as
1099 "markers" for phylogenetic and phylogeny-driven ecological studies of bacteria and
1100 archaea and their major subgroups. *PloS one* **8**(10): e77033.
1101 Xiao L, Feng Q, Liang S, Sonne SB, Xia Z, Qiu X, Li X, Long H, Zhang J, Zhang D et al. 2015. A
1102 catalog of the mouse gut metagenome. *Nature biotechnology* **33**(10): 1103-1108.
1103 Yuelin Zhu RMS, Paul S Meltzer, Sean R Davis. 2013. SRAdb: query and use public next-
1104 generation sequencing data from within R. *BMC bioinformatics* **14**(1): 19.
1105 Zhu A, Sunagawa S, Mende DR, Bork P. 2015. Inter-individual differences in the gene
1106 content of human gut bacterial species. *Genome biology* **16**: 82.
1107


A
Species-Level Cutoff
Genome Clusters
(N=5,952)
96.5% DNA Identity,
30 universal genes
B
31,007 bacterial genomes
5,701 disagree with taxonomy
2,666 without species labels
82%
18%
29%
47%
24%
20%
30%
50%
Agrees with taxonomy
Disagrees with taxonomy
Split from larger cluster
Name differs from cluster
Unnamed genome
Assigned to unnamed cluster
Assigned to named cluster
Singleton
C
Host–Associated
Human skin
Human airways
Human urogenital tract
Human oral
Human stool
Mouse stool
Baboon stool
Marine
Surface water layer
Deep chlorophyll maximum layer
Marine epipelagic mixed layer
Mesopelagic zone
Terrestrial
Temperate grasslands
Temperate forest
Tropical forest
Deserts
Human stool
United States
China
Spain
Denmark
Tanzania
Peru
Lifestyle
Urban
Rural
0
20
40
60
80
100
Coverage of Database Across Metagenomes
(% of cellular organism abundance)


A
Per Sample
Shotgun Metagenomes
HS-BLASTN
Bowtie2
Bowtie2
Universal Single-Copy Genes
Pan-Genome Database
Representative Genomes
High coverage species
Species Coverage & Abundance (per sample)
Gene Content (per species & sample)
SNPs (per species & sample)
Species Abundance Matrix
Gene Content Matrix (per species)
Core Genome SNP Matrix (per species)
Multiple Samples
B
R−squared = 0.95
Estimated Genome Coverage
Simulated Genome Coverage
C
Gene Presence/Absence Accuracy
Cutoff
0.05
0.15
0.35
0.5
1
Simulated Genome Coverage
D
Sensitivity
False Discovery Rate
SNP Prediction Performance
Simulated Genome Coverage


A
Bray–Curtis PC2
Bray–Curtis PC1
Mother
12 Months
4 Months
4 Days
B
P=1.3e−39
R2=0.46
Number of shared species
Days after birth
C
biological pairs
permuted pairs
Number of shared species
4 Days
4 Months
12 Months
D
P=5.8e−17
R2=0.11
Transmission threshold
Marker Allele Sharing (%)
Days after birth
E
4 days
4 months
12 months
Bacteroides vulgatus
Bacteroides stercoris
Bacteroides uniformis
Bacteroides cellulosilyticus
Bacteroides ovatus
Bacteroides caccae
Bacteroides thetaiotaomicron
Bacteroides fragilis
Parabacteroides distasonis
Parabacteroides merdae
Alistipes putredinis
Escherichia coli
Bifidobacterium longum
Bifidobacterium pseudocatenulatum
Bifidobacterium adolescentis
Faecalibacterium prausnitzii
Eubacterium rectale
Eubacterium eligens
butyrate–producing bacterium
Blautia wexlerae
Vertical Transmission
Yes
No
0.2
(nucleotide substitutions per site)
Count Mother-Infant Pairs


A
P=9.4e−14
Probability of vertical transmission at 12 months
Relative abundance in infant at 4 days (% of cells)
Vertically transmitted
Not vertically transmitted
Count mother−infant pairs
B
Bacteroides vulgatus
Parabacteroides distasonis
Bacteroides fragilis
Eubacterium rectale
Faecalibacterium prausnitzii
Blautia wexlerae
Relative abundance (% of cells)
detection limit
Vertical transmission
Yes
No
Undetermined
Mother
Infant
Days after birth


A
Candidatus Pelagibacter ubique HIMB083 (id:58575)
Alteromonas macleodii str. 'English Channel 673' (id:57644)
Candidatus Pelagibacter ubique HTCC7214 (id:60371)
Candidatus Pelagibacter sp. HTCC7211 (id:61513)
Alteromonas marina (id:60992)
SAR406 cluster bacterium SCGC AAA298–D23 (id:62595)
Candidatus Pelagibacter ubique HIMB058 (id:59237)
Candidatus Pelagibacter ubique HTCC7217 (id:60372)
Alteromonas australica (id:61892)
SAR86 cluster bacterium SAR86A (id:59142)
Prochlorococcus marinus str. NATL1A (id:57812)
Marinobacter hydrocarbonoclasticus (id:57973)
Prochlorococcus marinus str. MIT 9116 (id:49991)
Prochlorococcus sp. scB245a_519D13 (id:57728)
Prochlorococcus marinus subsp. pastoris str. CCMP1986 (id:51374)
Prochlorococcus marinus str. MIT 9302 (id:62230)
Prochlorococcus marinus str. AS9601 (id:60670)
Prochlorococcus sp. scB243_498I20 (id:53137)
Prochlorococcus sp. scB245a_519O21 (id:57810)
Prochlorococcus marinus str. MIT 9314 (id:60936)
Candidatus Pelagibacter ubique HTCC8051 (id:61569)
Prochlorococcus marinus str. MIT 9301 (id:60935)
Prochlorococcus marinus str. SB (id:61931)
Prochlorococcus sp. scB245a_518A6 (id:56918)
Prochlorococcus marinus str. MIT 9202 (id:57654)
Prochlorococcus sp. MIT 0604 (id:60762)
Prochlorococcus marinus str. GP2 (id:61930)
Prochlorococcus sp. scB243_495D8 (id:60682)
Prochlorococcus sp. scB245a_520K10 (id:60683)
Prochlorococcus marinus str. MIT 9401 (id:50740)
Synechococcus sp. WH 8109 (id:60932)
Alteromonas macleodii AltDE1 (id:57135)
Prochlorococcus sp. MIT 0801 (id:60763)
alpha proteobacterium HIMB59 (id:62227)
Prochlorococcus marinus str. MIT 9311 (id:35018)
Prochlorococcus sp. scB241_528O2 (id:60681)
Alteromonas macleodii str. 'Black Sea 11' (id:58315)
Prochlorococcus marinus str. MIT 9201 (id:62547)
Synechococcus sp. CC9605 (id:58591)
Candidatus Pelagibacter ubique HTCC9022 (id:59829)
alpha proteobacterium SCGC AAA536–G10 (id:58335)
alpha proteobacterium SCGC AB–629–G21 (id:44631)
Sphingomonas sp. SKA58 (id:61199)
Spongiibacter tropicus (id:59072)
Leeuwenhoekiella blandensis (id:61450)
Synechococcus sp. KORDI–52 (id:61879)
Zunongwangia profunda (id:62041)
Marinobacter manganoxydans (id:58552)
Sulfitobacter pontiacus (id:57671)
gamma proteobacterium SCGC AB–629–P17 (id:59173)
Universal Gene Coverage
>10x
>3x
>1x
Genus
Pelagibacter
Alteromonas
Prochlorococcus
Synechococcus
Marinobacter
Other
0
20
40
60
80
100
Species prevalence (% of samples)
C
Candidatus Pelagibacter (id:61513)
PC2 (10.93%)
PC1 (19.89%)
SAR86 cluster (id:59142)
PC2 (7.12%)
PC1 (44.31%)
D
alpha proteobacterium (id:44631)
PC2 (8.38%)
Depth (m)
P= 2.7e−08
R2=0.59
Relative Abundance
P= 8.1e−07
R2=0.50
PC1 (23.74%)
B
Surface layer
DCM / Mixed layer
Mesopalegic zone
−log10(p−value)
SAR86
Pelagibacter (61513)
5% FDR
Alpha proteobacterium (44631)
Correlation between gene content PCA and geographic distance
Ocean Region
North Pacific
South Pacific
North Atlantic
South Atlantic
Mediterranean
Red Sea
Indian Ocean
Depth
Surface water layer (5–12 m)
DCM & mixed layer (17–183 m)
Mesopelagic zone (246–990 m)