Title

inStrain profiles population microdiversity from metagenomic data and sensitively detects shared microbial strains

Journal
Nature Biotechnology, 39(6)

ISSN
1087-0156

Authors
Olm, Matthew R
Crits-Christoph, Alexander
Bouma-Gregson, Keith
et al.

Publication Date
2021-06-01

DOI
10.1038/s41587-020-00797-0

Peer reviewed

Published in final edited form as:
Nat Biotechnol. 2021 June ; 39(6): 727–736. doi:10.1038/s41587-020-00797-0.

InStrain enables population genomic analysis from metagenomic data and sensitive detection of shared microbial strains

Matthew R. Olm[1,2],
Alexander Crits-Christoph[2],
Keith Bouma-Gregson[3],
Brian Firek[4],
Michael J. Morowitz[4],
Jillian F. Banfield[1,5,6,7,✣]

[1]Department of Earth and Planetary Science, University of California, Berkeley, CA, USA

[2]Department of Plant and Microbial Biology, University of California, Berkeley, CA, USA

[3]Office of Information Management and Analysis, California State Water Resources Control Board, Sacramento, CA, USA

[4]Department of Surgery, University of Pittsburgh School of Medicine, Pittsburgh, PA, USA

[5]Department of Environmental Science, Policy, and Management, University of California, Berkeley, CA, USA

[6]Earth Sciences Division, Lawrence Berkeley National Laboratory, Berkeley, CA, USA

[7]Chan Zuckerberg Biohub, San Francisco, CA, USA

Abstract

Coexisting microbial cells of the same species often exhibit genetic variation that can affect phenotypes ranging from nutrient preference to pathogenicity. Here we present inStrain, a program that utilizes metagenomic paired reads to profile intra-population genetic diversity (microdiversity) across whole genomes and compare populations in a microdiversity-aware manner, dramatically increasing genomic comparison accuracy when benchmarked against existing methods. We use inStrain to profile >1,000 fecal metagenomes from newborn premature infants and find that siblings share significantly more strains than unrelated infants, although identical twins share no more strains than fraternal siblings. Infants born via cesarean section harbored *Klebsiella* with significantly higher nucleotide diversity than infants delivered vaginally, potentially reflecting acquisition from hospital versus maternal microbiomes. Genomic loci showing diversity within an

[✣]Corresponding author: jbanfield@berkeley.edu.

Conflict of Interest
The authors declare no conflict of interest.

Code availability
InStrain is available as an open-source python program on GitHub (https://github.com/MrOlm/inStrain) and documentation is online at https://instrain.readthedocs.io/en/latest/.

infant included variants found between other infants, possibly reflecting inoculation from diverse hospital-associated sources. InStrain can be applied to any metagenomic dataset for microdiversity analysis and rigorous strain comparison.

Main

There is genetic heterogeneity within all microbial populations. Genetic polymorphisms rapidly arise through *de novo* mutation, and variant frequencies change over time due to drift, selection, or linked selection. It is estimated that billions to trillions of bacterial genetic mutations are generated *de novo* every day in the microbiome of an individual adult human [1], and these differences can be clinically relevant. For example, just three point mutations can confer antibiotic resistance in *Enterobacteriaceae* [2]. Studying genetic variation in microbial populations has historically involved isolating a multitude of cells from the same population and performing phenotypic analysis and/or genome sequencing. Genome-resolved metagenomic analysis, which involves extracting and sequencing DNA directly from the environment and using computational tools to assemble and bin the resulting DNA sequences into genomes *in silico*, presents an attractive high-throughput alternative to this process. While complete haplotypes within a population cannot be precisely determined with short-read sequencing (due to the inability to associate variant loci across the genome), this technique allows simultaneous analysis of all taxa in microbial communities, identification of genetic variants and their frequencies in the species populations that comprise them, and measurement of the overall heterogeneity within these populations. Metagenomic analysis has been used to reveal fine-scale evolutionary mechanisms [3–5], dynamics [6–12], and strain level metabolic variation that could contribute to strain selection [1,13].

Many fundamental questions in human microbiome research relate to the transmission of microbial populations between individuals, including how we are seeded by microbes early in life [14–16]. However, intra-population diversity (genetic variation within a population) presents challenges for such analyses. Sequence comparisons are usually performed by aligning consensus genomes assembled from different samples [1,17] or by modifying a reference genome using mapped reads and comparing it to the same sequence that has been modified by reads from another sample [18–21] (Supplemental Figure S1). These methods represent each population based on the most common alleles, which can lead to erroneous results. For example, if sample 1 contains a single nucleotide variant (SNV) A at 20% frequency and T (the consensus choice) at 80% frequency, and sample 2 has A at 100% frequency, comparing the consensus genome of both samples will fail to identify the variant shared by both populations. Furthermore, alleles at intermediate frequencies (e.g. 30% - 70%) can be stochastically detected above or below 50% due to random sampling, resulting in chimeric consensus sequences. As natural microbial populations can have many polymorphic sites, genomic comparison methods that consider the genetic diversity are needed, as are standardized methods that are easy to use and that are applicable to all metagenomic studies.

Here we present inStrain, a program that profiles population microdiversity from metagenomic short read alignments and performs microdiversity-aware genomic comparisons. This includes calculating nucleotide diversity and linkage disequilibrium, identifying SNVs (including non-synonymous and synonymous variants), and reporting accurate coverage, depth, and breadth. We demonstrate that inStrain performs strain-level comparisons with higher accuracy and sensitivity than leading tools. To demonstrate the value of inStrain for microbiome studies, we apply inStrain to a large collection of previously sequenced infant fecal microbiomes to reveal patterns of microbiome microdiversity and strain sharing among infants born in the same neonatal intensive care unit (NICU) over a period of five years. inStrain is available as an open-source python program on GitHub (https://github.com/MrOlm/inStrain) and documentation is available both in the supplemental materials (Supplemental Document S1) and online at https://instrain.readthedocs.io/en/latest/.

Results

inStrain measures population-level diversity from metagenomic data

InStrain profiles the microdiversity of any DNA sequence dataset that consists of paired short reads that are mapped to a genome assembled from a metagenome or from a cultured isolate. Functionality can be broken into three major steps:

Step 1) Read filtering.—To increase the likelihood that mapped read pairs originate from organisms belonging to the same population a series of filters are applied. For each read pair aligned to the reference genome (*de novo* assembled from the same sample or a genome from another source) the mapQ score, average nucleotide identity (ANI) of the pair to the reference genome, and the insert size between aligned reads are calculated. Read pairs that don't pass adjustable quality cutoffs are removed, as are all unpaired reads. The exclusive use of pairs doubles the number of bases used to calculate the read ANI and mapQ score, increasing their accuracy and substantially increasing the span of genome analyzed. This reduces mismapping at repeat regions or regions conserved in multiple genomes. Other software tools, such as StrainPhlAn and MetaPhlAn [18,22], treat pairs of reads as separate observations and can assign each read pair to a different population, contrary to the strong expectation from Illumina sequencing protocols that a pair originates from a single DNA molecule.

Step 2) Calculation of nucleotide diversity, SNVs, and linkage.—For each gene, scaffold, and/or genome, inStrain calculates the mean, median, and standard deviation of the depth of coverage (number of reads per base-pair), breadth of coverage (percentage of reference base pairs covered by at least one read), expected breadth of coverage (given the average depth of coverage, the breadth of coverage that would be expected if reads were evenly spread across the genome), and average nucleotide diversity (π; [23]) of all base-pairs with at least 5x coverage (Figure 1a). 5x was chosen as the default minimum because it is the lowest coverage at which minor alleles under 50% frequency can be reliably detected (Supplemental Figure S2), and this value can be adjusted by the user. Both bialllelic and multiallelic SNVs and their frequencies are identified and annotated at positions where phred30 quality filtered reads differ from the reference genome and at positions where multiple bases are simultaneously detected at levels above the expected sequencing error rate. SNVs are classified as synonymous, non-synonymous, or intergenic based on gene annotations, and linkage disequilibrium is calculated between SNVs that are connected by at least twenty read-pairs.

Step 3) Generation of tables and figures.—Tables are generated that describe how many reads were removed by each filter described in *Step 1* and enumerate all metrics described in *Step 2*. Figures are generated for each genome to document SNV allele frequencies, genome-wide nucleotide diversity, patterns of linkage disequilibrium, and to report other findings (Figure 1b–f). All data generated during an inStrain run is stored in a space-efficient manner and can be used to quickly re-generate plots and tables with different parameters.

Microdiversity-aware ANI calculations (popANI) increase accuracy of strain discrimination

Most existing strain-comparison pipelines compare microbes in different samples based on their consensus genomes. In contrast, inStrain considers both major and minor alleles during genomic comparison. This new microdiversity-aware ANI metric is referred to as “popANI” (population-level ANI), and it is reported alongside consensus-based ANI (“conANI”). Both metrics are calculated in a pair-wise manner for samples that have been profiled using the methods described above. First, all positions of the genome at or above a minimum coverage threshold in both samples (5x by default) are identified. Only these positions are considered in the popANI and conANI calculations. Second, the number positions that differ in allelic composition between the samples is enumerated. For conANI, if the consensus base differs between the two samples a substitution is called. For popANI, a substitution is called at a site only if both samples share no alleles (either major or minor) (Figure 2a). This consideration of shared minor alleles dramatically increases the accuracy of population-level comparisons (Figure 2) with the following limitations: i) genomic positions within a read length of scaffold ends have reduced accuracy due to difficulties with read mapping (Supplemental Figure S3), ii) sequencing depth must be sufficient to detect minor alleles for them to be considered in popANI calculations (Supplemental Figure S2), and iii) the number of distinct genotypes shared between samples is not enumerated; a single popANI value is generated for each reference genome present in multiple samples.

We benchmarked inStrain’s strain comparison method against three existing common tools: dRep, which calculates genome-wide ANI [17], StrainPhlAn [18], which aligns short reads to a marker gene database (0.3% of the genome in the case of *Escherichia coli*) and compares the consensus maker genes in multiple samples, and MIDAS [19], which aligns short reads to a reference genome database and compares the single-nucleotide substitutions (SNSs) identified in each sample. We first compared the ability of each method to report the ANI between genomes with a known number of *in silico* mutations (Figure 2b). All four methods performed well on this test, which does not consider microdiversity, though dRep, inStrain, and MIDAS had lower errors in the ANI calculation than StrainPhlAn overall (0.00001%, 0.002%, 0.006% and 0.03%, respectively; average discrepancy between the true and calculated ANI). This is likely because dRep, inStrain, and MIDAS compare positions from across the entire genome (99.99998%, 99.7%, and 85.8% of the genome, respectively) and StrainPhlAn does not.

We next used each tool to compare metagenomes derived from defined bacterial communities. The ZymoBIOMICS Microbial Community Standard, which contains cells from eight bacterial species at defined abundances, was divided into three aliquots and subjected to DNA extraction, library preparation, and metagenomic sequencing. Each strain comparison tool was then used to compare bacterial species in each sample to each other in a pairwise manner (Figure 2c). As all genomic comparisons originate from the same defined community of microbes, each tool should report 100% ANI for all genomic comparisons. Deviations from this ideal either represent errors in sequence alignment or the presence of microdiversity that is likely present because cultures have been maintained in the laboratory. MIDAS, dRep, StrainPhlAn, and inStrain reported average ANI values of 99.97%, 99.98%, 99.990% and 99.999998%, respectively, with inStrain reporting average popANI values of 100% for 23 of the 24 comparisons and 99.99996% for one comparison. The difference in performance arises because the Zymo cultures contain non-fixed nucleotide variants that inStrain uses to confirm population overlap but that confuse the consensus sequences reported by dRep, StrainPhlAn, and MIDAS.

We used the Zymo data to establish a threshold for the detection of “same” versus “different” strains. The thresholds for MIDAS, dRep, StrainPhlAn, and inStrain, calculated based on the lowest average ANI across all 24 sequence comparisons, were 99.92% ANI, 99.94% ANI, 99.97% ANI, and 99.99996% ANI, respectively. Thus, inStrain can be used for detection of identical microbial strains with a stringency that is substantially higher than either other tool. Using the previously reported rate of 0.9 SNSs accumulated per genome per year in the gut microbiome of healthy human adults [1], in this test MIDAS is able to discriminate between strains that have diverged for at least 3,771 years, dRep for 2,528 years, StrainPhlAn for 1,307 years, and inStrain for 2.2 years (Supplemental Table S1). Stringent thresholds are useful for strain tracking, as strains that have diverged for hundreds to thousands of years are clearly not linked by a recent transmission event.

The Zymo data was also used to assess the ability inStrain to detect and compare organisms in the absence of sample-specific reference genomes. By mapping reads to all 4,644 representative genomes in the Unified Human Gastrointestinal Genome (UHGG) collection [24], inStrain detected the eight bacterial taxa known to be present in each of the three Zymo metagenomes. When using the recommended 50% genome breadth cutoff, these were the only eight taxa detected in each case with inStrain. MIDAS and Metaphlan2 detected 15 and 11 taxa in addition to the true community members, respectively, yet neither tool reports genome breadth or any other metric to filter out these erroneous results (besides relative abundance, which limits the ability to detect genuine low-abundance taxa) (Supplemental Table S2). The UHGG reference genomes had between 93.9% - 99.6% ANI to the organisms present in the Zymo samples. InStrain comparisons based on these genomes were still highly accurate (average 99.9998% ANI, lowest 99.9995% ANI, limit of detection 32.2 years) (Supplemental Table S1), highlighting that inStrain can be used with reference genomes from databases when sample-specific reference genomes cannot be assembled.

To compare the ability of the four methods to detect strains shared by twin premature infants, the microbiomes of six infants were processed according to the best recommended practice for each of the three tools. We then compared the number of strains found to be shared by twins and non-twins over a range of ANI thresholds. All methods identified significantly more strain sharing among twin pairs than pairs of unrelated infants, as expected, and inStrain remained sensitive at substantially higher ANI thresholds than either of the other tools (Figure 2d). We attribute the reduced ability of StrainPhlAn and MIDAS to identify shared strains to their reliance on consensus-based ANI measurements. We know that microbiomes can contain multiple coexisting strains, and when two or more strains of a species are in a sample at similar abundance levels it can lead to pileups of reads from multiple strains and chimeric sequences. The popANI metric is designed to account for this complexity. In combination, the reduced ability of previously available tools to detect truly shared strains and their inability to perform with the precision needed to use high ANI thresholds limit their utility of the for strain tracking.

Finally, we re-analyzed a previously generated dataset to compare data from inStrain to that from isolate-based sequencing [1]. We focused on individual S01, from which i) 123 colonies of *Bacteroides fragilis* were isolated and sequenced from 9 fecal samples collected over two years, and ii) metagenomic sequencing of the same fecal samples resulted in detection of a *B. fragilis* genome at 34x coverage (metagenomic data from all samples was analyzed together to increase sequencing depth). 2,477 biallelic mutations were identified among isolate genomes (mutations present in 20% - 80% of genomes), 8,164 biallelic mutations were identified by inStrain analysis of metagenomic data, and 903 were identified by both methods (Supplemental Table S3). If the isolate-detected mutations are considered ground truth (though in reality these may suffer from cultivation biases), inStrain performed with 36.5% sensitivity (percentage of isolate biallelic mutations identified by inStrain) and 99.8% specificity (percentage of genomic loci correctly identified as not having a biallelic mutations). While broadly consistent, the discrepancies between the methods may be due shifting allele frequencies in the *B. fragilis* population during the two years that sampling occurred, as the isolate genomes were sampled evenly from all samples but most metagenomic reads came from the two samples where it was most abundant.

Siblings share significantly more microbial strains at birth than unrelated infant pairs

We next applied inStrain to 1,163 fecal metagenomes from 160 premature infants born into the same neonatal intensive care unit [25]. The dataset includes samples from six individual sampling campaigns, involved the enrollment of 6 sets of monozygotic twins (MZ; identical), 20 sets of dizygotic twins (DZ; fraternal) and 3 sets of trizygotic (TZ) triplets, and over eight thousand *de novo* genomes from bacteria, bacteriophage, and plasmid colonists. Organisms that may have been introduced through contamination were removed based on their presence in sequenced negative controls, each genome set was de-replicated at 98% ANI to form “sub-species” groups, and representative genomes from each sub-species were combined into a single mapping database consisting of 2,266 genomes in order to reduce multi-mapped reads (Supplemental Figure S4). All metagenomes were mapped to this dereplicated genome set and inStrain was used to profile the microdiversity of each mapping. In all cases where a sub-species was detected in multiple infants with over 50% breadth of coverage, inStrain was used to compare strains.

A threshold of 99.999% popANI was chosen as the threshold to define bacterial, bacteriophage, and plasmid strains as being the same “strain” based on the Zymo experiment (Figure 2c) and on analysis of comparisons between subspecies present in the same infant over time (based on the assumption that strain genotypes from samples collected within days or weeks of each other typically represent the same strain) (Supplemental Figure S5). Thus, to be classified as the same strain, two populations must have no fixed differences within this margin of error. Of the 109,731 comparisons made, 4,103 (grey lines in Figure 3a) indicated that infants shared bacterial strains. Of these, 268 cases revealed sharing between pairs of siblings (despite sibling pair comparisons comprising only 0.3% of all comparisons; red lines in Figure 3a). Further, the majority of bacterial strains that were identified in two and only two infants were shared between sibling pairs (Figure 3b,c). Similar patterns were identified for bacteriophage and plasmid colonists (Supplemental Table S4).

The majority of bacterial strains (specific definition of “strain” provided above) identified in this study were detected in only a single infant (1818 of 3044 strains). The most frequently colonizing strain (*Staphylococcus epidermidis* 158.2.ba_7) was identified in samples from 49 of the 160 infants. Six of the seven other most frequently colonizing species were also Firmicutes, and many are known for their role in nosocomial infections, including *Clostridioides difficile* and *Enterococcus faecalis*. *Pseudomonas aeruginosa*, a frequently colonizing Proteobacterium, is also implicated in nosocomial infections. Twelve strains colonized more than ten infants, including five strains of *S. epidermidis*, three strains of *E. faecalis,* two strains of *C. difficile*, and one strain each of *P. aeruginosa* and *Clostridium sp.* (Figure 3g). These frequently encountered strains may have specific adaptations that enable them to survive in the neonatal intensive care unit (NICU). Alternatively, they may be acquired from health care workers that commonly interact with these infants.

Overall, siblings shared significantly more strains of bacteria, bacteriophage, and plasmids than unrelated infant pairs (Figure 3d). However, among siblings, monozygotic (MZ) twins shared no more strains than dizygotic (DZ) twins and trizygotic (TZ) triplets (Figure 3e). Infants born at more chronologically similar times shared significantly more strains of bacteriophages and plasmids, supporting the role of the hospital room environment in shaping initial bacteriophage and plasmid strain acquisition (Supplemental Figure S6). Infants born with similar gestational ages and birth weights also shared significantly more strains of bacteria, bacteriophages, and plasmids than those with different ages and weights (Figure 3f; Supplemental Figure S6). In combination, the results point to the role of infant physiology, sibling status, and calendar date of birth (i.e., similar date of residence in the NICU) in strain acquisition.

Nucleotide diversity of the premature infant microbiome

Over the sampling time-series in this study (generally the first few months of life) we detected an average of 17.8 ± 0.7 sub-species of bacteria, 26.9 ± 1.5 sub-species of bacteriophage, and 7.4 ± 0.3 sub-species of plasmids per infant (mean ± SEM; colonization defined as detection of genome at >5x depth coverage across ≥ 50% of the genome)

(Supplemental Table S5). As the 160 infants were sampled over six different campaigns, each using a unique combination of library preparation methodology, Illumina machine for sequencing, and institutional sequencing center, we first tested for effects related to sampling campaign. Infants of the same campaign were not more likely to share strains (Supplemental Figure S6), but measured nucleotide diversity among colonists varied significantly between the six different sampling campaigns, primarily driven by differences in library preparation methodology and the DNA sequencing machine used (Supplemental Figure S7). This is likely due to differences in the read error profiles associated with the sequencing platforms [26]. Importantly, bacterial nucleotide diversity was not associated with sequencing depth in any campaign (Supplemental Figure S7d). We thus analyzed each cohort separately for relationships between microdiversity and infant metadata, allowing us to validate the consistency of inStrain when run using different sequencing methodologies.

Bacteria had significantly higher nucleotide diversity than plasmids and phage in 4/6 campaigns, whereas plasmids had the lowest nucleotide diversity in 4/6 campaigns (Supplemental Figure S7). Relative to other bacteria, Proteobacteria had significantly higher and Firmicutes significantly lower nucleotide diversity in 3/6 and in 4/6 campaigns, respectively (Supplemental Table S4). Approximately 75% of premature infants were born via cesarean section (118/160), and their bacterial colonists had significantly higher nucleotide diversity than vaginally delivered infants in the NIH4 and Sloan2 cohorts and overall (Figure 4a). This effect was particularly striking for *Klebsiella* (Figure 4b), and the difference in *Klebsiella* microdiversity remained significant even when excluding infants in the NIH4 and Sloan2 cohorts (Supplemental Figure S7).

Finally, we performed a statistical test to identify genes with significantly different microdiversity than other genes in the genome (Table 1). Genes with significantly lower microdiversity include house-keeping genes like ribosomal protein S16 in bacteria and ParB in bacteriophage (where it is used to maintain circular lysogens [28]), as well as genes with more interesting functions including a plasmid-encoded polymyxin resistance protein, which is predicted to confer resistance to polymyxin antibiotics [29], and bacteriophage lambda head decoration protein D, which stabilizes the expansion of the capsid after genome packaging [30]. Among the genes with significantly higher microdiversity than the average gene are a bacterial-encoded gene with an immunoglobulin (Ig) domain (which can be involved in cell adhesion and invasion [31]) and a bacteriophage gene encoding tail fibers (which are often involved in host cell recognition [32]). Interestingly, both the Ig domain protein and tail fiber protein are involved in host interaction.

Tracking specific genetic variants within and between populations

To investigate the relationship between the diversity of a population within a single infant (intra-infant diversity) and the diversity of populations of the same subspecies in multiple different infants (inter-infant diversity), we performed a detailed analysis of an Enterococcus faecalis bacteriophage (subspecies 482_10.ph) that was present at high coverage depth (>20x) and breadth of coverage (>80%) in 44 infants in our cohort (Supplemental Table S5). We identified 410 loci with single-nucleotide substitutions (SNSs) fixed between infants, 679 loci with single-nucleotide variants (SNVs) with multiple alleles in the same infant, and 1062 loci where both were observed (Figure 6d). Intra-infant SNVs that were also observed as inter-infant SNSs could be ascribed to mixing of variants that are found alone in other individuals, and were thus excluded from further analysis to focus on intra-infant SNVs that presumably arose via *de novo* mutation. 18% of intra-infant SNVs were found to be polymorphic in at least 3 different infants, indicating an overlap in variants across infants (Supplemental Table S6). Genomic regions and genes with a substantial number of intra-infant SNPs had correspondingly more inter-infant substitutions (Figure 6a,b).

Seven of the fifty-one genes annotated on the *E. faecalis* bacteriophage genome had *dN/dS* ratios over 0.5, including five proteins of unknown function, a DnaB replication initiation homolog, and a predicted distal tail gene (Figure 6a,c). The predicted distal tail gene, which might play a role in host specificity, was also found to have an intra-infant *pN/pS* ratio of 0 (6 synonymous SNVs and 0 non-synonymous SNVs), possibly indicating selection for variation between but not within individual populations. Multiple small hypothetical proteins also had high *dN/dS* ratios, one of which was only present in ~50% of infants (Figure 6c). The relaxed purifying selection indicated by high *dN/dS* ratios and the variable presence of these genes may indicate an accessory or vestigial function, although adaptation can also be a driver of increased dN/dS ratios in some contexts

Discussion

InStrain is an integrated and versatile program for profiling the microdiversity of organisms from metagenomic data. Its ability to perform microdiversity-aware genomic comparisons offers several advantages over existing pipelines, including the consideration of major and minor alleles, thus accounting for the presence of coexisting strains. Because it uses sample-assembled genomes and full paired-read information there is greatly increased confidence that reads are aligned correctly, which improves the high resolution comparisons being made based on entire genomes. Many of these capabilities have been successfully implemented individually in previous studies [15,19,33–36]. However, their simultaneous integration into a well-documented and easy to use pipeline allows substantially more rigorous detection of near-identical strains than the existing commonly used pipelines (Figure 2) used in recent high-profile publications to quantify the ecologically critical process of microbiome transmission [14,37]. The method substantially increases the stringency of evidence for strain sharing and thus identification of the factors that determine the extent to which this occurs.

Twin studies have previously been used to elucidate relationships between host genetics and human microbiome composition, with the basic premise being that because twins are reared together and share similar environments, increased microbiome similarity between MZ twins compared to DZ twins can be ascribed to genetic effects [38]. Although studies of adult twins have consistently found some microbial taxa to be more commonly identified in MZ than DZ twins [39–42], diet and lifestyle preferences have also been shown to be more similar in MZ twins than DZ twins [43–45], presenting significant potential for confounding effects. In contrast to prior studies, all subjects in the current study were housed in the same NICU for the entirety of sampling time. Our findings, based on new and demonstrably more robust methods, indicate that MZ twins shared no more strains of bacteria, bacteriophage, or plasmids than DZ twins. This points to a minimal role of human genetics in early life strain colonization.

Initial colonists are believed to have an outsized role in microbiome development [46,47]. The hospitalized premature infants in this study were all given prophylactic antibiotics immediately after birth, housed in isolettes that maintained separation from other infants, and ~75% were born by cesarean section. These factors likely limited their exposure to microbes from the mother, other family members, and the external home environment. The patterns of strain-sharing among infants in this study suggest the importance of *i) Family-specific sources*. Strains present in two and only two infants were significantly more likely to be shared between siblings (Figure 3), highlighting the role of strain sources such as shared visitors and/or parents in infant colonization. *ii) The hospital environment*. Non-sibling infants born at similar times chronologically shared more strains of bacteriophages and plasmids than those born further apart, indicating that the local hospital microbiome plays a role in strain acquisition. The identification of strains of ESKAPE pathogens (known for their antibiotic resistance and ability to cause nosocomial infections) colonizing large numbers of infants further points to the hospital room as an important source of initial strains. These highly-colonizing strains may have been dispersed in part by healthcare workers that interact with many infants. *iii) Infant physiology*. Infants with similar physiological properties such as gestational age and birth weight shared significantly more strains, potentially due to differences in the development of the human immune system, the development of the physical gut environment, clinical treatment, or nutrition (e.g., formula feeding vs. breast milk). *iv) Unique sources*. The majority of strains identified were found in only a single infant, demonstrating that even in a highly-cleaned environment like the NICU, initial microbiota acquisition is a largely individualized process.

It is difficult to distinguish microbiome diversity that is evolved *in situ* from that introduced by immigration [1,9]. In this study of newborn infants we found evidence that initial bacterial microdiversity can be related to mode of acquisition; *Klebsiella* had higher levels of nucleotide diversity in infants born via C-section than those born vaginally, suggesting that there is a more abundant and/or diverse pool of *Klebsiella* strains in the operating room (where *Enterobacteriaceae* have previously been identified [48]) than in the maternal microbiome. The general increase in nucleotide diversity and *dN/dS* ratios of genes involved in cell-cell interactions compared to other functions indicates that these genes are likely under diversifying selection. Identification of house-keeping genes with lower than average nucleotide diversity demonstrates the utility of inStrain for identifying genes under purifying selection (Table 1).

By reporting and classifying all gene variants, inStrain enables locus-specific analyses of the genetic differences within and between populations. Further, as inStrain also does not rely upon reference databases or conserved bacterial marker genes, it is capable of tracking genetic variation in bacteriophages and plasmids. For example, applying inStrain to a highly prevalent *E. faecalis* bacteriophage confirmed a relationship between the diversity within individual infants and the subspecies diversity overall, and identified specific genes with divergent *dN/dS* ratios and variable presence (Figure 6). Specifically, we found evidence that nonsynonymous changes in a tail fiber gene are purged within infants (possibly to maintain infectivity), yet selected for between infants (suggestive of variation in bacterial host immunity).

Diversity is a hallmark of stable and healthy human microbiomes [49–51]. While microbial diversity is typically measured by quantifying the number and evenness of microbial species or genera present in a sample, the detected microbial taxa represent larger populations of cells with within-population genetic heterogeneity. Microdiversity may increase the likelihood of harboring a fit genotype as conditions change. Alternatively, an overall wider gene variant pool may reflect adaptation to spatial variation in local environmental conditions. InStrain allows scientists to easily measure and analyze population microdiversity. In existing and future metagenomic sequencing-based projects, there is the potential to improve our understanding of relationships between microbial population diversity and resilience, stability, population-level phenotypes and to track ecologically relevant processes such as strain migration and *in situ* evolution.

Methods

InStrain implementation

InStrain is an open-source Python package for analysis of genomes for population comparisons, reporting of gene coverage and breadth, SNV calling with gene localization and synonymous/non-synonymous identification, and calculation of population genetics parameters including nucleotide diversity and linkage disequilibrium. It is implemented as a set of interrelated modules, the basic functionality of which are described below. Full documentation is available online (https://instrain.readthedocs.io) and provided in Supplemental Document S1.

Dependencies

InStrain requires Samtools [52] for interacting with .bam and .sam files, Prodigal [53] for annotating open reading frames, and a number of publicly available python modules that are bundled and automatically installed with inStrain for statistical analysis, efficient data storage, and figure generation (including Pandas [54], SciPy [55], Numpy [56], Matplotlib [57], and Seaborn [58]). All other functions are implemented natively in Python.

Program input

The required input to inStrain is i) a nucleotide sequence or a set of nucleotide sequences in fasta format, and ii) a mapping file in .sam or .bam format [52] documenting where reads align to the nucleotide sequence. The fasta file can be a set of genomes assembled from a sample of interest, a set of reference genomes acquired from an online database, or a single genome sequence of interest. The mapping file can be created using any number of publicly available programs, allowing the user flexibility in how the mapping should be performed given the study design and specific type of reads that were sequenced (Illumina, PacBio, nanopore, etc.).

Read filtering

When calling SNVs in a metagenomic context, it is most important to consider whether mapped reads truly belong to the population of interest. Careful filtering of reads in the .bam file is performed to reduce the probability of reads being erroneously mapped. i) All unpaired reads are removed by default, and filters are applied to pairs of reads in combination. This behavior can be modified by the user to specify a privileged set of reads that do not need to be paired (such as long-reads or merged reads), or to retain all reads and treat unpaired-reads as pairs. ii) Paired reads must be mapped in the proper orientation within an expected insert size. The minimum insert distance can be adjusted by the user, and the maximum insert distance is a user-specified multiple of the median insert distance (3 by default). E.g., if pairs have a median insert size of 500bp, by default all pairs with insert sizes over 1500bp will be excluded. iii) Pairs must have a user-defined minimum mapQ score. MapQ scores represent both the number of mismatches in the read mapping and how unique that mapping is (i.e. whether the read maps equally well to multiple genomic locations). The read in the pair with the higher mapQ is used for the pair. iv) Pairs must be above a user-defined minimum nucleotide identity value. For example, if reads in a pair are 100bp each, and each read has a single mismatch, the ANI of that pair would be 0.99. Only reads that pass this set of four filters are used in the following analysis.

Calculating coverage and nucleotide diversity

The coverage of (number of reads aligned to) each position in the provided nucleotide sequence file is calculated using Samtools [52]. This information is used to calculate the following for each genome, scaffold, and gene in the input nucleotide sequence file: i) average coverage, ii) median coverage, iii) standard deviation of coverage, iv) number of bases with 0 coverage, v) breadth of coverage (the fraction of bases that are covered by at least a single read), vi) unmasked breadth of coverage (the fraction of bases that are covered by at least the number of reads required to call a SNV, which is 5 by default), and vii) expected breadth of coverage. Given the calculated average coverage value, the expected breadth of coverage is the breadth that would be expected if reads were evenly distributed along the genome. It is calculated based on the empirically determined function: breadth = −1 * e^(0.883 * coverage) + 1. If the breadth is significantly lower than the expected breadth, it indicates that reads are mapping only to a specific region of the scaffold (e.g. a transposon, prophage, or other mobile element).

The nucleotide diversity (π; [23]) of each position is calculated using the formula: nucleotide diversity = 1 − [(number of “A” bases / total bases) ^2 + (number of “C” bases / total bases) ^2 (number of “T” bases / total bases) ^2 (number of “G” bases / total bases) ^2]. This information is used to calculate the average nucleotide diversity and median nucleotide diversity for each genome, scaffold, and gene in the input nucleotide sequence file.

Identifying SNVs and linkage

SNVs are identified on filtered reads based on 3 criteria. i) There must be at least a user-defined minimum number of reads mapping to the position. By default this is 5. ii) More than a user-defined percentage of reads must have a variant base at that position. By default this is 5%. iii) The number of reads with the variant base must be higher than a null model given the coverage of the base. The null model describes the probability that the number of true reads that support a variant base could be due to random mutation error, assuming Q30 score for each base. The null model can be adjusted to account for technologies with different sequencing error rates, and the false discovery rate given the null model can be adjusted as well (by default it is set at 1e-6, or one false-positive SNV in a million).

All SNVs are further classified based on the number of alleles and the reference base at the SNV position. Reference SNVs are positions where a single allele is present in the reads, and the allele is different from the input sequence base. Bi-alleleic SNVs are positions where there are two alleles present in the reads at a position. Multi-allelic SNVs are positions where there are more than two alleles present. Population SNVs are positions where the reference base is not one of the detected alleles, regardless of the number of alleles detected. SNVs are further classified as non-synonymous (the SNV causes an amino acid change), synonymous (the SNV does not cause an amino acid change), or intergenic (the SNV is not in an open reading frame (ORF)) based on user-provided ORFs.

Metrics of linkage disequilibrium are calculated between pairs of SNV locations that are both present on a user-defined number of read pairs (20 by default). Only pairwise biallelic haplotypes are examined, and additional alleles are ignored. R2 and D' are calculated using all available reads as described previously [59], and also calculated using a rarefied number of reads to account for how differences in coverage between sites may impact these metrics.

Reporting and storing results

A number of output datatables are created after all calculations are complete. These include i) scaffold_info.tsv, which lists the coverage, breadth, nucleotide diversity, number of identified SNVs, and other related metrics for each sequence in the input nucleotide sequence, ii) read_report.tsv, which lists the number of reads that pass and fail each of the read-filtering steps described above on a sequence-by-sequence basis, iii) SNVs.tsv, which lists the location, reference base counts, variant base counts, number of alleles, and other information about each identified SNV, iv) linkage.tsv, which lists the R^2, D', distance, position, and other information about each pair of SNVs linked by a sufficient number of read pairs, v) gene_info.tsv, which lists the coverage, nucleotide diversity, and related metrics for each ORF, vi) genomeWide_scaffold_info.tsv, which lists the metrics described in scaffold_info.tsv on a genome level (rather than a scaffold level), and vii) SNV_mutation_types.tsv, which list the gene location and predicted mutation type (synonymous, nonsynonymous, or intergenic) of each SNV. Finally, inStrain uses this information to generate a number of figures, examples of which are shown in Figure 1.

In addition to the tables described above, a large amount of auxiliary data is generated and stored upon completion of inStrain. This includes base-by-base coverage and nucleotide diversity of each location, graphs generated during calculation of linkage, and the lengths of all input nucleotide sequences. This information is stored in a directory structure called an "inStrain profile", and can be programmatically accessed using the provided API. It also allows future operations to be rapidly run on an existing inStrain project.

Comparing inStrain profiles

InStrain performs strain-level comparisons by comparing inStrain profile objects that were created by mapping different sets of reads to the same nucleotide sequence(s). These comparisons are performed in a pair-wise manner and follow a series of four steps. i) All positions in which both inStrain profiles have at least the minimum coverage to call SNVs (5 by default) are identified. The percentage of bases that fit this criteria (referred to as “compared_bases_count”) is reported as “percent_genome_compared”, representing the percentage of the sequence that will be compared in the following steps. ii) Each position identified in step i is classified following the logic depicted in **Table 2a** as either “no SNV”, “consensus SNV”, or “population SNV”. If both samples have no SNVs called at position, or if both samples have the same major allele at a position, no SNV is called. If samples have different major alleles, a “consensus SNV” is called. If samples share no alleles at a position, major or minor, a “population SNV” is called. iii) ConANI is calculated as: (1 − number of consensus SNVs) / “compared_bases_count” (calculated in step i), and popANI is calculated as: (1 − number of population SNVs) / “compared_bases_count”. iv) Datatables are made listing the metrics calculated above on a scaffold-by-scaffold level as well as a genome-by-genome level. Dendrograms visualizing the strain-level relationships between groups of genomes are also generated using Seaborn and Matplotlib.

Benchmarking inStrain

Synthetic comparisons (Figure 2b) were performed by using SNP Mutator [60] to introduce a known number of mutations into a reference genome (*Escherichia coli* strain SQ88; RefSeq accession number GCF_000988385.1) and comparing the mutated genomes to the original reference genome. For dRep, mutated genomes were compared to the reference genome using dRep on default settings. For inStrain, MIDAS, andStrainPhlAn, Illumina reads were simulated for all genomes at 20x coverage using pIRS [61]. For inStrain, synthetic reads were mapped back to the reference genome using Bowtie 2 [62], profiled using “inStrain profile” under default settings, and compared using “inStrain compare” under default settings. For StrainPhlan, synthetic reads profiled with Metaphlan2 [22], resulting marker genes were aligned using StrainPhlan, and the ANI of resulting nucleotide alignments was calculated using the class “Bio.Phylo.TreeConstruction.DistanceCalculator(‘identity’)” from the BioPython python package [63]. Raw values from this analysis are available in Supplemental Table S1. For MIDAS, synthetic reads were provided to the program directly using the “run_midas.py species” command, and compared using the “run_midas.py snps” command. The ANI of the resulting comparisons was calculated as “[mean(sample1_bases, sample2_bases) − count_either] / mean(sample1_bases, sample2_bases)”.

To measure the impact of genome fragmentation on inStrain we used the GCF_000988385.1 genome with mutations introduced to one in one hundred bases (see above paragraph for how this was made). We generated four versions of this genome made up of scaffolds of length 1kb, 10kb, 100kb, or 1Mbp, and evaluated the ability of “inStrain profile” to detect the known mutations (Supplemental Figure S3).

Isolate-based comparisons (Figure 2c) were performed based on the ZymoBIOMICS Microbial Community Standards product (Catalog #D6300). Three samples were prepared from aliquots of this mixture of cells in which DNA extraction, library preparation, and *in silico* sequence trimming and analysis were performed separately. For dRep, reads from each sample were assembled independently using IDBA-UD [64], binned into genomes based off of alignment to the provided reference genomes (https://s3.amazonaws.com/zymo-files/BioPool/ZymoBIOMICS.STD.refseq.v2.zip) using nucmer [65], and compared using dRep on default settings. For StrainPhlAn, reads from Zymo samples profiled with Metaphlan2, resulting marker genes were aligned using StrainPhlan, and the ANI of resulting nucleotide alignments was calculated as described above. For MIDAS, reads from Zymo samples were provided to MIDAS directly and the ANI of sample comparisons was calculated as described above. For inStrain, reads from Zymo samples were aligned to the provided reference genomes using Bowtie 2, profiled using "inStrain profile" under default settings, and compared using "inStrain compare" under default settings. "popANI" values were used for inStrain. Eukaryotic genomes were excluded from this analysis, and raw values are available in Supplemental Table S1. To evaluate inStrain when using genomes from public databases, all reference genomes from the UHGG collection were downloaded and concatenated into a single .fasta file. Reads from the Zymo sample were mapped against this database and processed with inStrain as described above. The ability of each method to detect genomes was performed using all Zymo reads concatenated together, and raw values are available in Supplemental Table S2.

Twin-based comparisons (Figure 2d) were performed on three randomly chosen sets of twins that were sequenced during a previous study [25]. For StrainPhlAn, all reads sequenced from each infant were concatenated and profiled using Metaphlan2, compared using StrainPhlAn, and the ANI of resulting nucleotide alignments was calculated as described above. For MIDAS, all reads sequenced from each infant were concatenated and profiled with MIDAS, and the ANI of species profiled in multiple infants was calculated as described above. For dRep, all de-replicated bacterial genomes assembled and binned from each infant (available from [25]) were compared in a pairwise manner using dRep under default settings. For inStain, strain-sharing from these six infants was determined using the methods described below. ANI values from all compared genomes and the number of genomes shared at a number of ANI thresholds are available for all three methods in Supplemental Table S1.

To compare the biallelic mutations identified by isolate-based sequencing to those identified by metagenomic inStrain analysis, we downloaded all metagenomes and isolate genomes from individual S01 using the SRA links provided in [1]. We only considered the 9 fecal samples for which metagenomic and isolate sequencing data were available. Read files were validated using SRA-tools "vdb-validate", singleton reads were removed and reads were trimmed using "repair.sh" and "bbduk.sh" from BBTools [66], and reads were mapped to the *Bacteroides fragilis* reference genome (NCBI:txid272559) using Bowtie2. Mapping files from all metagenomes were merged using samtools, and inStrain was run on the resulting .bam files and all isolate .bam files with default settings. Biallelic positions among isolate genomes were defined as those where inStrain identified a particular consensus base in at least 24 (20%), but no more than 98 (80%) of the 123 isolate genomes. Biallelic positions for the metagenome sample were defined as those where inStrain reported "allele_count = 2" in the output table. All SNVs identified in this analysis are available in Supplemental Table S3.

To determine the sequencing coverage needed to detect minor alleles with 95% probability (Supplemental Figure S2) we used binomial statistics and the null model described above (which establishes the minimum number of observations to detect an allele beyond levels expected by phred30 illumina errors). For each allele frequency (AF) between 5% and 50% (5%, 6%, 7%, etc.), we iterated over each coverage value (c) between 1x and 150x and used the null model to determine the minimum number of reads (n) needed to detect an allele of AF frequency and c coverage. We used the “scipy.stats.binom” package [55] to determine the probability of observing n minor alleles, given c observations and an AF probability of observing a minor allele. We used this information to determine the minimum coverage needed to detect minor alleles of each frequency with a 95% probability, and used the package “scipy.optimize.curve_fit” to fit an exponential curve using non-linear least squares fitting (Supplemental Figure S2).

Calling, detection, and profiling of sub-species of bacteria, bacteriophage, and plasmids

Genomes of bacteria, bacteriophage, plasmid, and eukaryotes were previously binned from the infants comprising this study, as described previously [25], and downloaded from the link https://doi.org/10.6084/m9.figshare.c.4740080.v1. To generate a single genome set, all bacterial genomes were compared to each other using dRep version 2.2.0 under default settings, all bacteriophage genomes were compared to each other using the command “dRep dereplicate -- S_algorithm ANImf -nc .5 -l 10000 -N50W 0 -sizeW 1 --noQualityFiltering --clusterAlg single”, and all plasmid genomes were compared to each other using the same command as bacteriophages. Genomes with ANI >= 98% were classified as the same subspecies, and the genome with the highest score (as determined by dRep) was chosen as the representative genome from each subspecies. Bacteriophage and plasmid genomes with taxonomic classifications specifying “Eukarya” were marked as “likely human” and excluded from further analysis. Information about sub-species is available in Supplemental Table S5.

Reads from each individual fecal sample, reads from each infant concatenated together (referred to as “coReads”), and reads from all negative extraction control samples concatenated together were mapped to all representative sub-species genomes concatenated together using Bowtie 2 with default settings. “InStrain profile” was run on all resulting mapping files with default settings. Detection of a sub-species in a sample was defined as that genome being present with >= 0.5 unmaskedBreadth (meaning that at least half of the bases in the genome were covered by at least 5 reads). Mappings from coReads were used for all analyses unless otherwise specified. Subspecies detected in the negative extraction control sample, and genomes detected significantly more often in one of the six individual sampling campaigns were marked “likely contaminant” and excluded from further analysis. Information on sub-species abundance is available in Supplemental Table S5.

Identification of strains and associations with metadata

Strain-level comparisons were performed between subspecies detected in multiple samples from the same infant over time-series sampling, and strain-level comparisons were performed between subspecies detected in the coReads of multiple infants. For within-infant subspecies comparisons, all subspecies detected in multiple individual samples from an infant (as described above) were compared using “inStrain compare”. Raw values are available in Supplemental Table S4. For between-infant subspecies comparisons, subspecies that were detected in coRead samples from multiple infants (or the coRead sample consisting of all negative extraction controls) were compared using “inStrain compare” with default settings. A distance matrix then created for each subspecies based on popANI values, and this matrix was used to cluster subspecies into a number of individual strains using ‘average’ hierarchical clustering with a threshold of 99.999% ANI with the scipy cluster package [55]. Strains that were present in the reads from the negative extraction control, and strains from subspecies that were filtered out using the methods described above were removed from further analysis. Raw comparison values and strain identities are available in Supplemental Table S4.

The number of strains shared between infants was visualized in Figure 3ab using Circos [67]. The strain-level Jaccard distance between infants was calculated according to the formula: Jaccard similarity = number of strains shared by both infants / number of strains present in either infant. P-values for Jaccard similarity are based on the two-tailed Wilcoxon rank-sum statistic between all twin pairs and all non-twin pairs, as calculated using the python module scipy.stats.ranksums [55]. Associations between the number of strains shared between infants and their difference in birth day, birth weight, and gestational age was determined by first binning the metadata variable into windows of size 20 (birth weight, gestational age) or 1 (gestational age) and calculating the average number of strains shared between infants within that window. Siblings were excluded from this analysis. P-values and R^2 values are based on linear least-squares regression, with the two-sided p-value reported for a hypothesis test whose null hypothesis is that the slope is zero (calculated using the python module scipy.stats.linregress).

The visualization in Figure 3g was created by first identifying the eight bacterial species with the highest colonizing strain, and then assigning a specific color to each strain within these eight species that colonized at least five infants. For each value on the x-axis, the y-axis displays the proportional count of the total strains detected in infants by strains that colonized at least that value of infants.

Nucleotide diversity analysis

The coReads inStrain analysis described above resulted in a total of 8,336 subspecies / infant pairs in which the subspecies genome was detected at 5x coverage across at least 50% of the genome (Supplemental Figure S4). The two-tailed Wilcoxon rank-sum statistic (as implemented in Scipy [55]) was used to compare the nucleotide diversity of different sets of genomes and generate p-values (Figure 4; Supplemental Figure S6; Supplemental Figure S7).

Gene-based nucleotide analysis

InStrain was used on default settings to profile genes for all detected subspecies in individual samples and coReads, using gene annotations provided by Prodigal [53] run in metagenome mode on original assemblies. Genes with significantly different coverage and/or nucleotide diversity than the rest of genes on the genome were identified using data from coReads profiling of subspecies. For each genome present in at least three infants, the coverage / nucleotide diversity of each gene on the genome across all infants in which the subspecies was present were compared to the coverage / nucleotide diversity of all other genes on the genome across all infants in which the subspecies was present using the Wilcoxon rank-sum statistic (as implemented in Scipy). P-values were corrected to q-values to account for multiple hypothesis testing using Benjamini-Hochberg p-value correction [27]. Genes were annotated based on pFam database HMMs [68]. For display in Table 1, only genes with pFam annotations that did not include the words "Uncharacterized" or "unknown" in the description were retained, all genes with significant differences in coverage (in addition to nucleotide diversity) were excluded, and a maximum of one gene from each taxonomic annotation was allowed for inclusion in each quadrant of high/low microdiversity and organism type.

Tracking specific nucleotide variants

Enterococcus faecalis bacteriophage subspecies 482_10.ph was identified with at least 80% breadth of coverage and 20x coverage depth in the coReads of 44 infants. Open reading frames were called using Prodigal in metagenome mode, and genes were annotated using USEARCH to search against the UniRef100 database. Gene categories (tail-associated, structural, etc.) were assigned based on manual inspection of the resulting database hits. The gene map presented in Figure 6a was generated using the python module "dna_features_viewer".

Bi-allelic SNVs (intra-infant variants) were identified based on the results of "inStrain profile_genes", where the resulting "SNV_mutation_types" table was subset to SNVs with an allele_count of 2. Substitutions (inter-infant variants) were identified from the "SNVs" table resulting from the operation "inStrain profile", where the table was subset to SNVs with an allele_count of 1. The number of genomic locations where an SNV was identified in at least one infant, where a substitution was identified in at least one infant, and where both were identified in at least one infant was displayed in a waffle plot using the python module "PyWaffle".

Synonymous and nonsynonymous variants were identified using inStrain, and the total number of synonymous and nonsynonymous sites in each gene was determined using methods from the script "dnds_from_drep.py" [69]. *dN/dS* was calculated using the formula [(non-synonymous substitutions / non-synonymous sites) / (synonymous substitutions / synonymous sites)], and *pN/pS* was calculated using the formula [(non-synonymous SNVs / non-synonymous sites) / (synonymous SNVs / synonymous sites)]. The number of substitutions per kbp and the number of SNVs per kbp were calculated by dividing the total number of substitutions / SNVs identified in each gene in all infants by the sum of the length of the gene times the masked breadth (the percentage of the gene with at least 5x coverage; the coverage required to call a SNV) of the gene for each infant the gene was identified in. Genes with a masked breadth ≥ 50% were defined as being present, and the gene deletion frequency was calculated as the percentage of infants where the gene was not present.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgements

This research was supported by the National Institutes of Health (NIH) under award RAI092531A, the Alfred P. Sloan Foundation under grant APSF-2012-10-05, a National Science Foundation Graduate Research Fellowship to M.O. under Grant No. DGE 1106400, and Chan Zuckerberg Biohub. The study was approved by the University of Pittsburgh Institutional Review Board (Protocol PRO10090089).

Data availability

The authors declare that the data supporting the findings of this study are available within the paper and its supplementary information files. Reads from infant samples are available under BioProject PRJNA294605 (SRA studies SRP052967, SRP114966, and SRP012558; and SRA accessions SRR5405607 to SRR5406014), reads from Zymo samples are available under BioProject PRJNA648136), and *de novo* assembled genomes are available at https://doi.org/10.6084/m9.figshare.c.4740080.v1 [25]

References

1. Zhao S et al. Adaptive Evolution within Gut Microbiomes of Healthy People. Cell Host Microbe 25, 656–667.e8 (2019). [PubMed: 31028005]
2. Schloissnig S et al. Genomic variation landscape of the human gut microbiome. Nature 493, 45–50 (2012). [PubMed: 23222524]
3. Simmons SL et al. Population genomic analysis of strain variation in Leptospirillum group II bacteria involved in acid mine drainage formation. PLoS Biol. 6, e177 (2008). [PubMed: 18651792]
4. Eppley JM, Tyson GW, Getz WM & Banfield JF Genetic exchange across a species boundary in the archaeal genus ferroplasma. Genetics 177, 407–416 (2007). [PubMed: 17603112]
5. Good BH, McDonald MJ, Barrick JE, Lenski RE & Desai MM The dynamics of molecular evolution over 60,000 generations. Nature (2017) doi:10.1038/nature24287.
6. Ignacio-Espinoza JC, Ahlgren NA & Fuhrman JA Long-term stability and Red Queen-like strain dynamics in marine viruses. Nat Microbiol (2019) doi:10.1038/s41564-019-0628-x.
7. Bendall ML et al. Genome-wide selective sweeps and gene-specific sweeps in natural bacterial populations. ISME J. 10, 1589–1601 (2016). [PubMed: 26744812]
8. Delmont TO et al. Single-amino acid variants reveal evolutionary processes that shape the biogeography of a global SAR11 subclade. Elife 8, (2019).
9. Garud NR, Good BH, Hallatschek O & Pollard KS Evolutionary dynamics of bacteria in the gut microbiome within and across hosts. PLoS Biol. 17, e3000102 (2019). [PubMed: 30673701]
10. Smillie CS et al. Strain Tracking Reveals the Determinants of Bacterial Engraftment in the Human Gut Following Fecal Microbiota Transplantation. Cell Host Microbe 23, 229–240.e5 (2018). [PubMed: 29447696]
11. Siranosian BA, Tamburini FB, Sherlock G & Bhatt AS Acquisition, transmission and strain diversity of human gut-colonizing crAss-like phages. Nat. Commun. 11, 280 (2020). [PubMed: 31941900]
12. Crits-Christoph A, Olm MR, Diamond S, Bouma-Gregson K & Banfield JF Soil bacterial populations are shaped by recombination and gene-specific selection across a grassland meadow. The ISME Journal vol. 14 1834–1846 (2020). [PubMed: 32327732]
13. Sharon I et al. Time series community genomics analysis reveals rapid shifts in bacterial species, strains, and phage during infant gut colonization. Genome Res. 23, 111–120 (2013). [PubMed: 22936250]

14. Shao Y et al. Stunted microbiota and opportunistic pathogen colonization in caesarean-section birth. Nature 574, 117–121 (2019). [PubMed: 31534227]

15. Korpela K et al. Selective maternal seeding and environment shape the human gut microbiome. Genome Res. gr.233940.117+ (2018) doi:10.1101/gr.233940.117.

16. Brooks B et al. Strain-resolved analysis of hospital rooms and infants reveals overlap between the human and room microbiome. Nat. Commun. 8, 1814 (2017). [PubMed: 29180750]

17. Olm MR, Brown CT, Brooks B & Banfield JF dRep: a tool for fast and accurate genomic comparisons that enables improved genome recovery from metagenomes through de-replication. ISME J. 11, 2864–2868 (2017). [PubMed: 28742071]

18. Truong DT, Tett A, Pasolli E, Huttenhower C & Segata N Microbial strain-level population structure and genetic diversity from metagenomes. Genome Res. 27, 626–638 (2017). [PubMed: 28167665]

19. Nayfach S, Rodriguez-Mueller B, Garud N & Pollard KS An integrated metagenomics pipeline for strain profiling reveals novel patterns of bacterial transmission and biogeography. Genome Res. 26, (2016).

20. Brito IL et al. Transmission of human-associated microbiota along family and social networks. Nat Microbiol 4, 964–971 (2019). [PubMed: 30911128]

21. Costea PI et al. metaSNV: A tool for metagenomic strain level analysis. PLoS One 12, e0182392 (2017). [PubMed: 28753663]

22. Truong DT et al. MetaPhlAn2 for enhanced metagenomic taxonomic profiling. Nat. Methods 12, (2015).

23. Nei M & Li WH Mathematical model for studying genetic variation in terms of restriction endonucleases. Proceedings of the National Academy of Sciences vol. 76 5269–5273 (1979).

24. Almeida A et al. A unified catalog of 204,938 reference genomes from the human gut microbiome. Nat. Biotechnol. (2020) doi:10.1038/s41587-020-0603-3.

25. Olm MR et al. Necrotizing enterocolitis is preceded by increased gut bacterial replication, Klebsiella, and fimbriae-encoding bacteria. Science Advances 5, eaax5727 (2019). [PubMed: 31844663]

26. Schirmer M, D’Amore R, Ijaz UZ, Hall N & Quince C Illumina error profiles: resolving fine-scale variation in metagenomic sequencing data. BMC Bioinformatics 17, 125 (2016). [PubMed: 26968756]

27. Yekutieli D & Benjamini Y Resampling-based false discovery rate controlling multiple test procedures for correlated test statistics. J. Stat. Plan. Inference 82, 171–196 (1999).

28. Lobocka M & Yarmolinsky M P1 plasmid partition: a mutational analysis of ParB. J. Mol. Biol. 259, 366–382 (1996). [PubMed: 8676375]

29. Fu W et al. First structure of the polymyxin resistance proteins. Biochem. Biophys. Res. Commun. 361, 1033–1037 (2007). [PubMed: 17686460]

30. Yang F et al. Novel fold and capsid-binding properties of the λ-phage display platform protein gpD. Nat. Struct. Biol. 7, 230–237 (2000). [PubMed: 10700283]

31. Bodelón G, Palomino C & Fernández LÁ Immunoglobulin domains in Escherichia coli and other enterobacteria: from pathogenesis to applications in antibody technologies. FEMS Microbiol. Rev. 37, 204–250 (2013). [PubMed: 22724448]

32. Tétart F, Repoila F, Monod C & Krisch HM Bacteriophage T4 host range is expanded by duplications of a small domain of the tail fiber adhesin. J. Mol. Biol. 258, 726–731 (1996). [PubMed: 8637004]

33. Vatanen T et al. Genomic variation and strain-specific functional adaptation in the human gut microbiome during early life. Nat Microbiol 4, 470–479 (2019). [PubMed: 30559407]

34. Yassour M et al. Strain-Level Analysis of Mother-to-Child Bacterial Transmission during the First Few Months of Life. Cell Host Microbe 24, 146–154.e4 (2018). [PubMed: 30001517]

35. Eren AM et al. Anvi’o: an advanced analysis and visualization platform for ‘omics data. PeerJ 3, e1319 (2015). [PubMed: 26500826]

36. Brito IL & Alm EJ Tracking strains in the microbiome: insights from metagenomics and models. Front. Microbiol. 7, (2016).

37. Ferretti P et al. Mother-to-Infant Microbial Transmission from Different Body Sites Shapes the Developing Infant Gut Microbiome. Cell Host Microbe 24, 133–145.e5 (2018). [PubMed: 30001516]

38. Goodrich JK et al. Genetic Determinants of the Gut Microbiome in UK Twins. Cell Host Microbe 19, 731–743 (2016). [PubMed: 27173935]

39. Lim MY et al. The effect of heritability and host genetics on the gut microbiota and metabolic syndrome. Gut 66, 1031–1038 (2017). [PubMed: 27053630]

40. Turpin W et al. Association of host genome with intestinal microbial composition in a large healthy cohort. Nat. Genet. 48, 1413–1417 (2016). [PubMed: 27694960]

41. Davenport ER et al. Genome-Wide Association Studies of the Human Gut Microbiota. PLoS One 10, e0140301 (2015). [PubMed: 26528553]

42. Goodrich JK, Davenport ER, Clark AG & Ley RE The Relationship Between the Human Genome and Microbiome Comes into View. Annu. Rev. Genet. 51, 413–433 (2017). [PubMed: 28934590]

43. Spor A, Koren O & Ley R Unravelling the effects of the environment and host genotype on the gut microbiome. Nat. Rev. Microbiol. 9, 279–290 (2011). [PubMed: 21407244]

44. Teucher B et al. Dietary patterns and heritability of food choice in a UK female twin cohort. Twin Res. Hum. Genet. 10, 734–748 (2007). [PubMed: 17903115]

45. Vinkhuyzen AAE, van der Sluis S, de Geus EJC, Boomsma DI & Posthuma D Genetic influences on ‘environmental’ factors. Genes Brain Behav. 9, 276–287 (2010). [PubMed: 20050926]

46. Faith JJ et al. The Long-Term Stability of the Human Gut Microbiota. Science 341, 1237439–1237439 (2013). [PubMed: 23828941]

47. Ding T & Schloss PD Dynamics and associations of microbial community types across the human body. Nature 509, 357–360 (2014). [PubMed: 24739969]

48. Shin H et al. The first microbial environment of infants born by C-section: the operating room microbes. Microbiome 3, (2015).

49. Thévenon S & Couvet D The impact of inbreeding depression on population survival depending on demographic parameters. Animal Conservation vol. 5 53–60 (2002).

50. Oh J, Byrd AL, Park M, Kong HH & Segre JA Temporal Stability of the Human Skin Microbiome. Cell 165, 854–866 (2016). [PubMed: 27153496]

51. Jovel J et al. Characterization of the Gut Microbiome Using 16S or Shotgun Metagenomics. Front. Microbiol. 7, (2016).

52. Li H et al. The sequence alignment/map format and SAMtools. Bioinformatics 25, 2078–2079 (2009). [PubMed: 19505943]

53. Hyatt D et al. Prodigal: prokaryotic gene recognition and translation initiation site identification. BMC Bioinformatics 11, 119 (2010). [PubMed: 20211023]

54. McKinney W & Others. pandas: a foundational Python library for data analysis and statistics. Python for High Performance and Scientific Computing 14, (2011).

55. Jones E, Oliphant T & Peterson P SciPy: Open source scientific tools for Python. URL http://scipy.org (2001).

56. Developers N NumPy. NumPy Numpy. Scipy Developers 31 (2013).

57. Hunter JD Matplotlib: A 2D graphics environment. Comput. Sci. Eng. 9, 90–95 (2007).

58. Waskom M Seaborn: statistical data visualization — seaborn 0.7.1 documentation. (2012).

59. VanLiere JM & Rosenberg NA Mathematical properties of the r2 measure of linkage disequilibrium. Theor. Popul. Biol. 74, 130–137 (2008). [PubMed: 18572214]

60. Davis S et al. CFSAN SNP Pipeline: an automated method for constructing SNP matrices from next-generation sequence data. PeerJ Comput. Sci. 1, e20 (2015).

61. Hu X et al. pIRS: Profile-based Illumina pair-end reads simulator. Bioinformatics 28, 1533–1535 (2012). [PubMed: 22508794]

62. Langmead B & Salzberg SL Fast gapped-read alignment with Bowtie 2. Nat. Methods 9, 357–359 (2012). [PubMed: 22388286]

63. Cock PJA et al. Biopython: freely available Python tools for computational molecular biology and bioinformatics. Bioinformatics 25, 1422–1423 (2009). [PubMed: 19304878]

64. Peng Y, Leung HCM, Yiu SM & Chin FYL IDBA-UD: a de novo assembler for single-cell and metagenomic sequencing data with highly uneven depth. Bioinformatics 28, 1420–1428 (2012). [PubMed: 22495754]

65. Delcher AL, Salzberg SL & Phillippy AM Using MUMmer to identify similar regions in large sequence sets. Curr. Protoc. Bioinformatics Chapter 10, Unit 10.3 (2003).

66. Bushnell B, Rood J & Singer E BBMerge--accurate paired shotgun read merging via overlap. PLoS One 12, e0185056 (2017). [PubMed: 29073143]

67. Krzywinski M et al. Circos: an information aesthetic for comparative genomics. Genome Res. 19, 1639–1645 (2009). [PubMed: 19541911]

68. El-Gebali S et al. The Pfam protein families database in 2019. Nucleic Acids Res. (2018) doi:10.1093/nar/gky995.

69. Olm MR et al. Consistent Metagenome-Derived Metrics Verify and Delineate Bacterial Species Boundaries. mSystems 5, (2020).

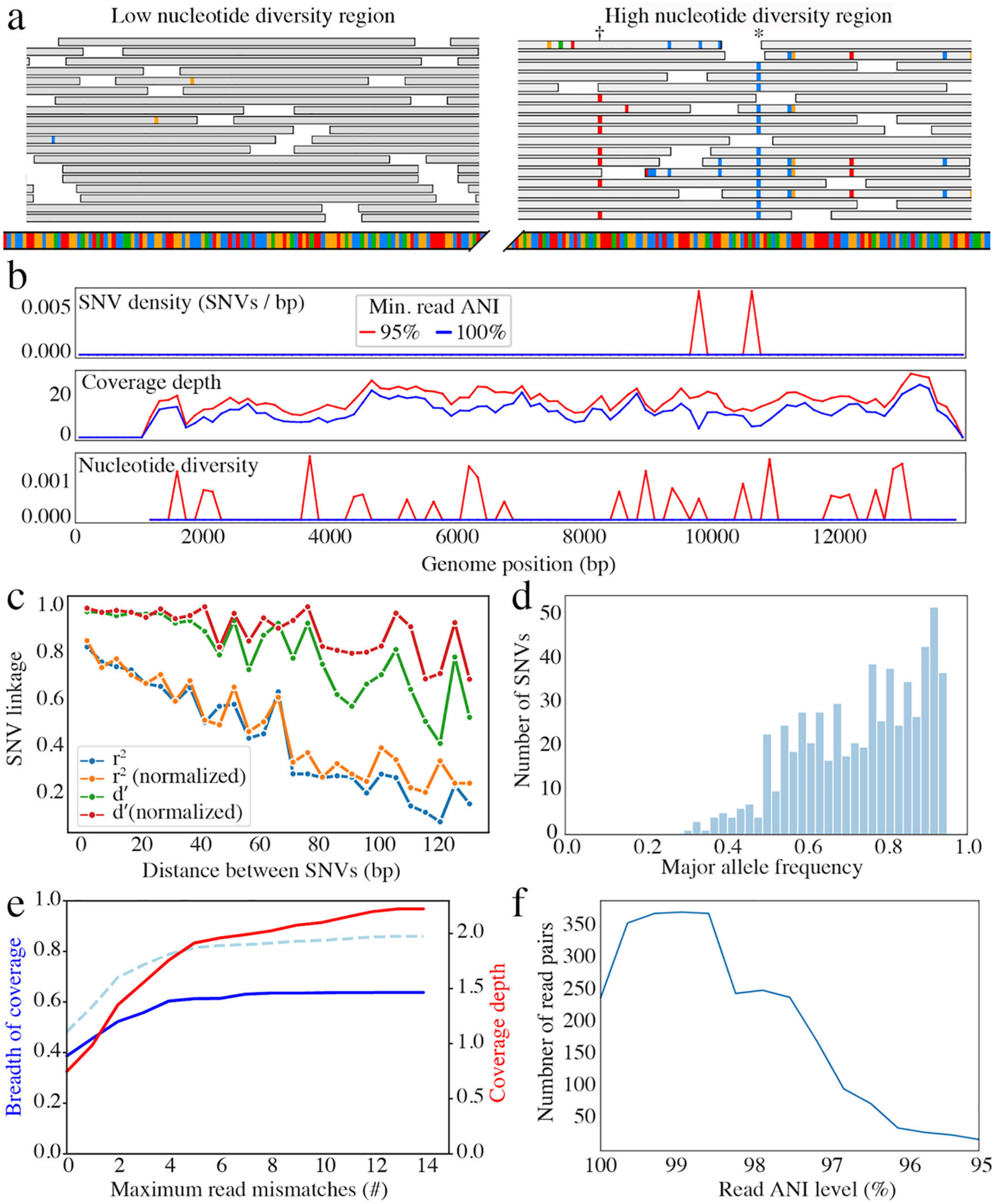
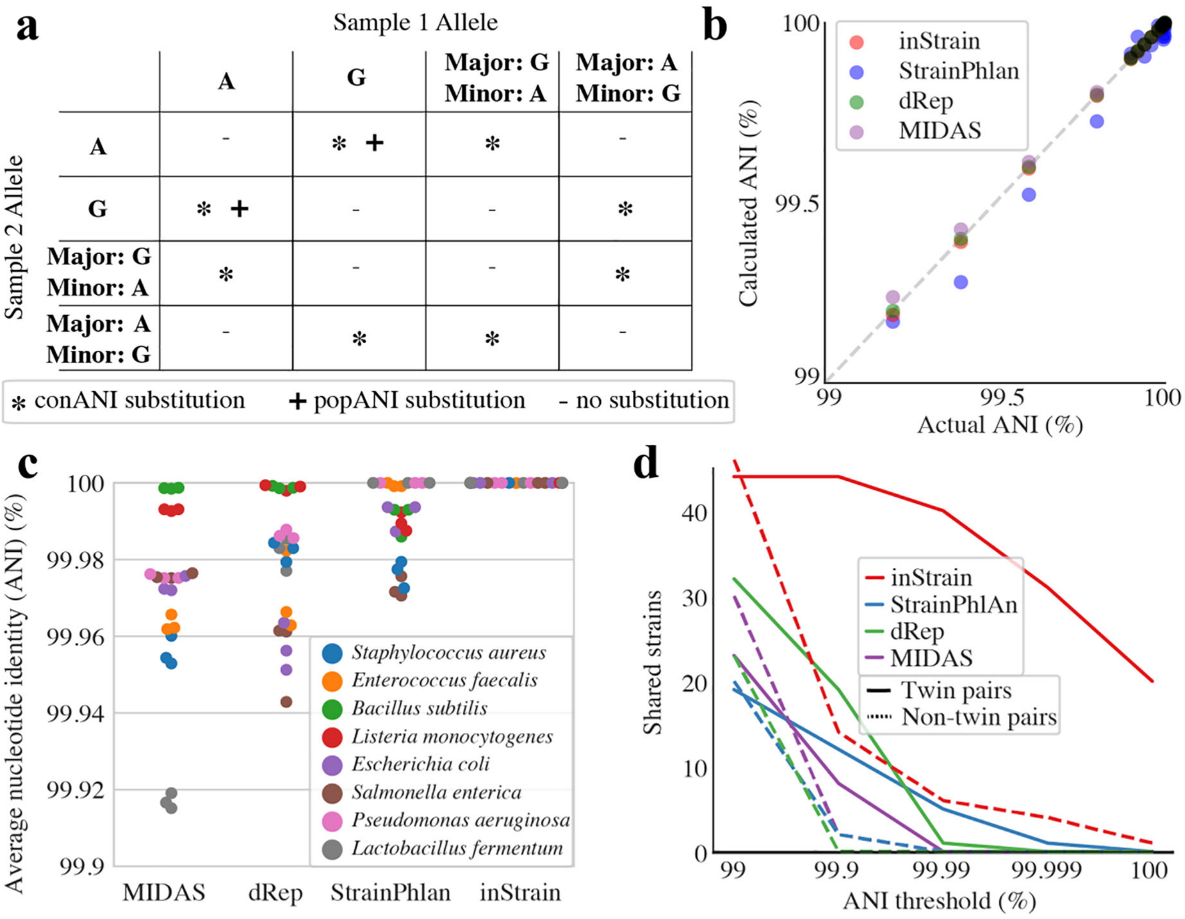


Figure 1. InStrain measures population-level diversity from metagenomic data.

a) Examples of metagenomic reads (grey boxes) mapping to genomic regions with low and high nucleotide diversity. Mismatches to the reference genome are represented by small colored marks on the reads, and the reference genome is represented below the reads.

b-f) Examples of figures automatically generated by inStrain. **b**) SNV density, coverage, and nucleotide diversity across a bacteriophage genome. Spikes in nucleotide diversity and SNV density do not correspond with increased coverage, indicating that the signals are not due to read mis-mapping. Positions with nucleotide diversity and no SNV-density are those where diversity exists but is not high enough to call a SNV c) Metrics of SNV linkage vs. distance between SNVs; linkage decay (as shown here) is a common signal of recombination. **d**) Distribution of the major allele frequencies of bi-allelic SNVs (the Site Frequency Spectrum). Alleles with major frequencies below 50% are the result of multiallelic sites. The lack of distinct puncta suggest that more than a few distinct strains are present. **e**) Breadth of coverage (blue line), coverage depth (red line), and expected breadth of coverage given the depth of coverage (dotted blue line) versus the minimum ANI of mapped reads. Coverage depth continues to increase while breadth plateaus, suggesting that all regions of the reference genome are not present in the reads being mapped. **f**) Distribution of read pair ANI levels when mapped to a reference genome; this plot suggests that the reference genome is >1% different than the mapped reads.



Sample 2 Allele \ Sample 1 Allele	A	G	Major: G Minor: A	Major: A Minor: G
A	-	* +	*	-
G	* +	-	-	*
Major: G Minor: A	*	-	-	*
Major: A Minor: G	-	*	*	-

* conANI substitution + popANI substitution - no substitution

Figure 2. InStrain accurately discriminates between closely related strains.

a) Table demonstrating the circumstances under which conANI and popANI substitutions will be called. ConANI substitutions are called whenever the consensus base differs, and popANI substitutions are only called when there is no allelic overlap between samples. **b**) Synthetic mutations were introduced to a reference genome of *E. coli* obtained from RefSeq to generate variant genomes with specific ANI differences from the reference genome, and four tools were used to compare the variant genomes to the reference genome. dRep, inStrain, and MIDAS consistently reported accurate ANI values, while StrainPhlAn was inaccurate by a median of 0.03% ANI. **c**) A mock community of bacterial cells was sequenced in biological triplicate and compared using four tools. InStrain performed best in correctly identifying that the genomes were identical in all three samples. **d**) The fecal microbiomes of three sets of twins were compared using each of the four tools, and the number of bacterial genomes with ANI values above a range of thresholds is plotted for pairs of twins (which are expected to share more strains) and pairs of unrelated infants. InStrain remained sensitive at higher ANI thresholds than the other three tools.

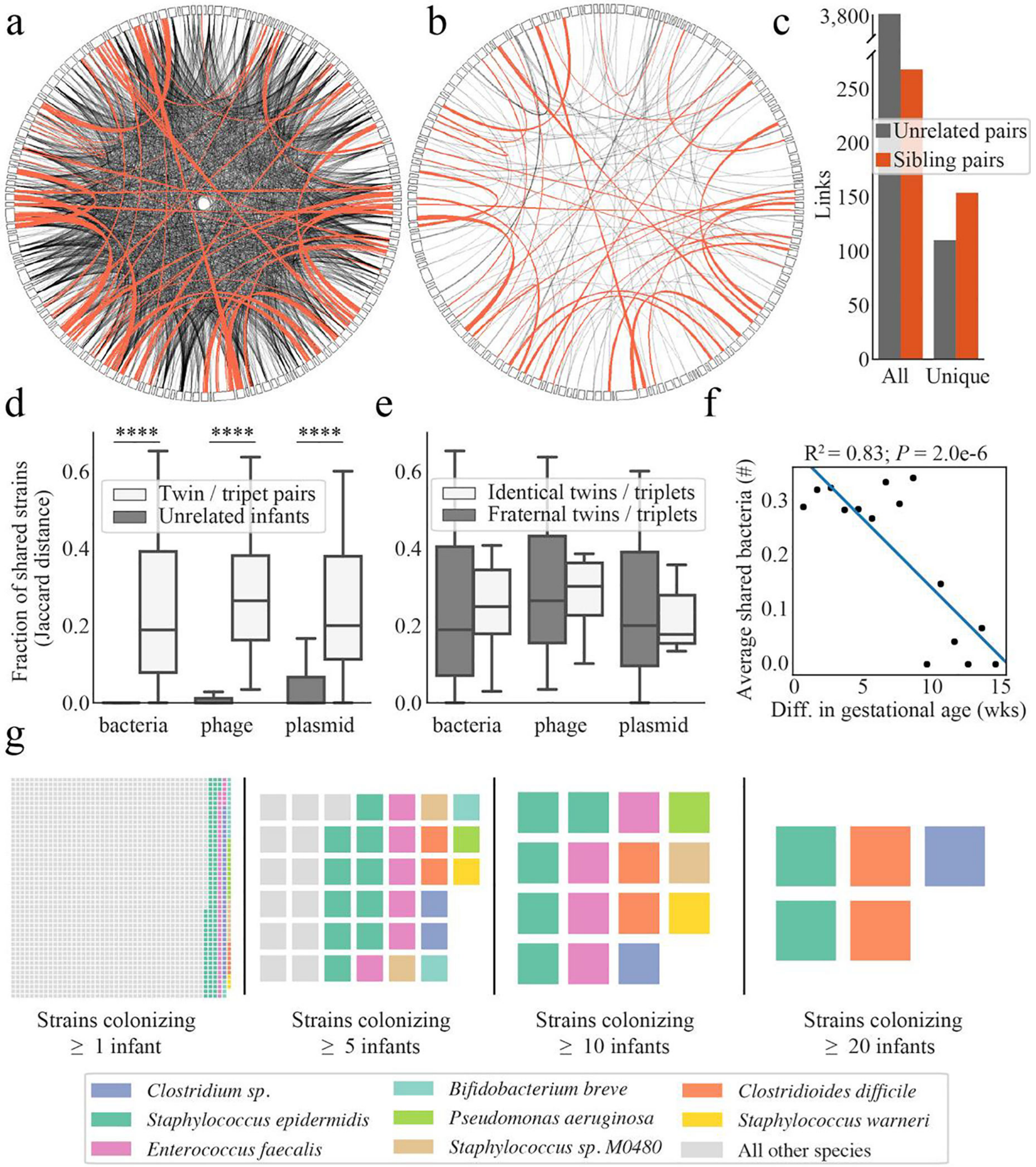


Figure 3. Siblings share significantly more microbial strains at birth than unrelated infants.

a,b) A link is drawn for each strain shared between pairs of infants (represented by rectangles along the circumferences). Links between sibling pairs are drawn in red, links between unrelated infants are drawn in grey. Diagrams are made displaying all strains (a) and only strains that are uniquely in two and only two infants **(b). c)** Enumeration of links drawn in **(a)** and **(b). d)** Twin pairs share significantly more strains of all domains than unrelated pairs (**** = $p < 1e-15$; two-sided Wilcoxon rank-sum test). **e**) Identical twin pairs do not share significantly more strains than fraternal twin pairs. **f)** Infants born more closely in gestational age share significantly more bacterial strains. **g)** Most strains colonize only a single infant, but some strains colonize many more. For each minimum number of infants colonized, a box is drawn for each strain that colonizes at least that many infants. Boxes are colored based on the species identity of each strain.

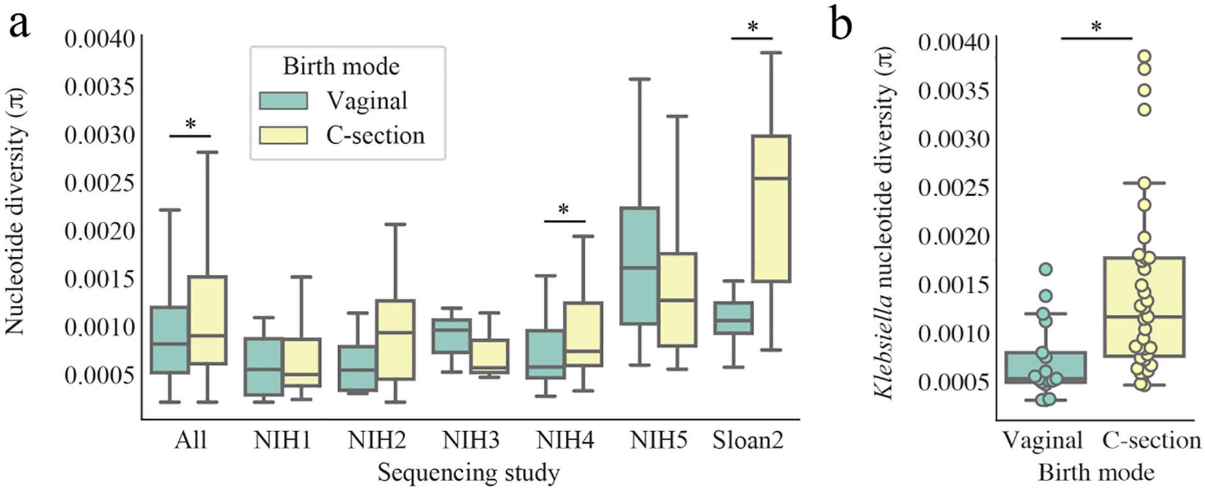


Figure 4. Analysis of the microdiversity of premature infant colonists.

a) Overall and among two of the six individual study cohorts, infants born via C-section had host microbes with higher nucleotide diversity than those delivered vaginally (* = $p < 0.05$; two-sided Wilcoxon rank-sum test). **b)** Organisms of the genus *Klebsiella* have significantly higher nucleotide diversity in infants born via C-section than those delivered vaginally (* = p < 0.05; two-sided Wilcoxon rank-sum test with Benjamini-Hochberg p-value correction [27] for testing each microbial species and genus present in both vaginal and C-section born infants).

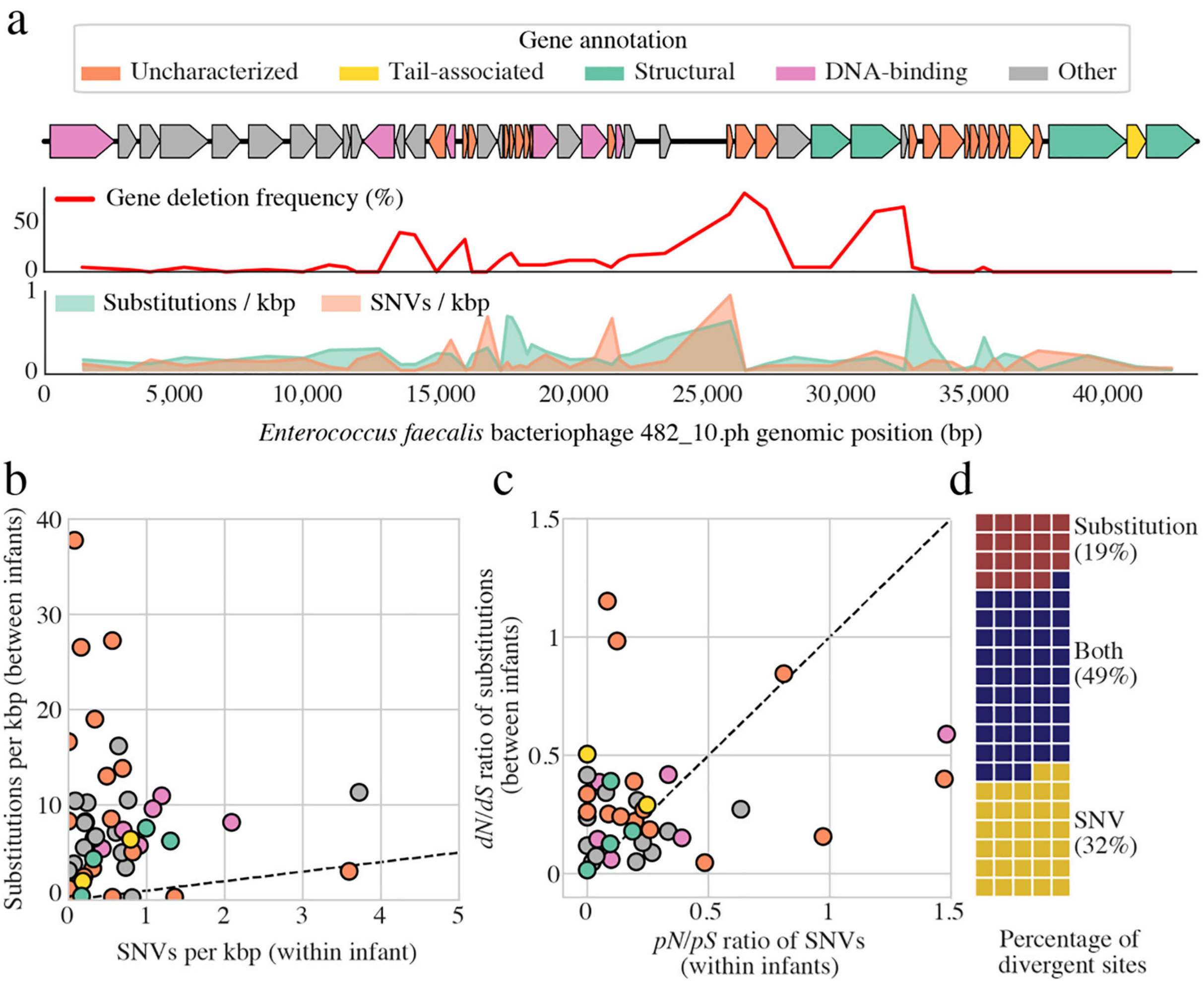


Figure 6. Tracking specific genetic differences within and between populations of an *E. faecalis* bacteriophage.

a) Frequencies of gene deletions, substitutions, and SNVs for all genes across an *E. faecalis* bacteriophage genome identified in 44 infants. Genes are colored based on their annotations. **b)** Frequency of observed substitutions (fixed differences between pairs of infants) in each gene versus frequency of SNVs (positions with multiple alleles in an individual infant at positions that are never observed as fixed differences). **c)** Ratios of non-synonymous to synonymous substitutions (*dN/dS*) and ratios of non-synonymous to synonymous population-level variants (*pN/pS*) for each gene. **d**) Classification of variant sites observed across infants only as substitutions, only as SNVs, and as both.

Table 1.

Genes with significantly higher or lower microdiversity than the rest of the genome.

Type	Taxonomy	Gene ID	q-value	pFam	Description
Low microdiversity					
bacteria	*E. faecalis*	23754	7.93E-20	PF00886.18	Ribosomal protein S16
	S. epidermidis	16419	9.51E-15	PF02597.19	ThiS family
	K. pneumoniae	15325	4.33E-10	PF02617.16	ATP-dependent Clp protease adaptor protein ClpS
phage	*Escherichia*	223	9.39E-06	PF08775.9	ParB family
	E. coli	205	0.00011027	PF02924.13	Bacteriophage lambda head decoration protein D
	Phietavirus	334	0.00018497	PF00692.18	dUTPase
plasmid	*Bacilli*	0	0.00013818	PF02388.15	FemAB family
	K. aerogenes	281	0.00034062	PF11183.7	Polymyxin resistance protein PmrD
	S. epidermidis	290	0.00043106	PF01479.24	S4 domain
High microdiversity					
bacteria	*S. epidermidis*	15627	6.23E-43	PF05345.11	Putative Ig domain
	K.pneumoniae	15332	2.38E-34	PF00465.18	Iron-containing alcohol dehydrogenase
	E. faecalis	23792	4.03E-32	PF13731.5	WxL domain surface cell wallbinding
phage	*Escherichia*	226	1.25E-13	PF03400.12	IS1 transposase
	E. coli	243	6.41E-12	PF03406.12	Phage tail fiber repeat
	E. faecalis	293	1.18E-08	PF01183.19	Glycosyl hydrolases family 25
plasmid	Unknown	358	1.95E-28	PF00665.25	Integrase core domain
	Bacilli	11	4.84E-25	PF03432.13	Relaxase/Mobilisation nuclease domain
	Clostridium	374	9.71E-14	PF02782.15	FGGY family of carbohydrate kinases, C-terminal domain