nature biotechnology

ARTICLES

https://doi.org/10.1038/s41587-020-00797-0

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

Matthew R. Olm[1,2,8], Alexander Crits-Christoph[2], Keith Bouma-Gregson[3], Brian A. Firek[4], Michael J. Morowitz[4] and Jillian F. Banfield[1,5,6,7]✉

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 uses metagenomic paired reads to profile intra-population genetic diversity (microdiversity) across whole genomes and compares microbial populations in a microdiversity-aware manner, greatly increasing the accuracy of genomic comparisons 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 by cesarean section harbor *Klebsiella* with significantly higher nucleotide diversity than infants delivered vaginally, potentially reflecting acquisition from hospital rather than maternal microbiomes. Genomic loci that show diversity in individual infants include 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.

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]. Early approaches for studying genetic variation in microbial populations involved isolating a multitude of cells from the same population and performing phenotypic analysis and/or genome sequencing. More recent efforts have used 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. Although 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 their constituent species 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 with the same sequence that has been modified by reads from another sample[18–21] (Supplementary Fig. 1). 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 (for example, 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 nonsynonymous 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

[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. [8]Present address: Department of Microbiology and Immunology, Stanford University School of Medicine, Stanford, CA, USA. ✉e-mail: jbanfield@berkeley.edu

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 5 years. inStrain is available as an open-source Python program on GitHub (https://github.com/MrOlm/inStrain) and documentation is available both in the supplementary materials (Supplementary Software Manual 1) 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 do not 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; ref. [23]) of all base pairs with at least 5× coverage (Fig. 1a). We chose 5× as the default minimum because it is the lowest coverage at which minor alleles under 50% frequency can be reliably detected (Supplementary Fig. 2), and this value can be adjusted by the user. Both biallelic 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, nonsynonymous or intergenic based on gene annotations, and linkage disequilibrium is calculated between SNVs that are connected by at least 20 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 and patterns of linkage disequilibrium, and to report other findings (Fig. 1b–f). All data generated during an inStrain run are 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 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 pairwise 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 (5× by default) are identified. Only these positions are considered in the popANI and conANI calculations. Second, the number of 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) (Fig. 2a). This consideration of shared minor alleles greatly increases the accuracy of population-level comparisons (Fig. 2) with the following limitations: (1) genomic positions within a read length of scaffold ends have reduced accuracy due to difficulties with read mapping (Supplementary Fig. 3); (2) sequencing depth must be sufficient to detect minor alleles for them to be considered in popANI calculations (Supplementary Fig. 2); and (3) the number of distinct genotypes shared between samples is not enumerated; for each pair of samples that a reference genome is present in, a single popANI value is calculated.

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 (Fig. 2b). All four methods performed well on this test, which does not consider microdiversity, although 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 with each other in a pairwise manner (Fig. 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 represent either 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 nonfixed 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 the other tools. Using the previously reported rate of 0.9 SNSs accumulated per genome per year in the gut

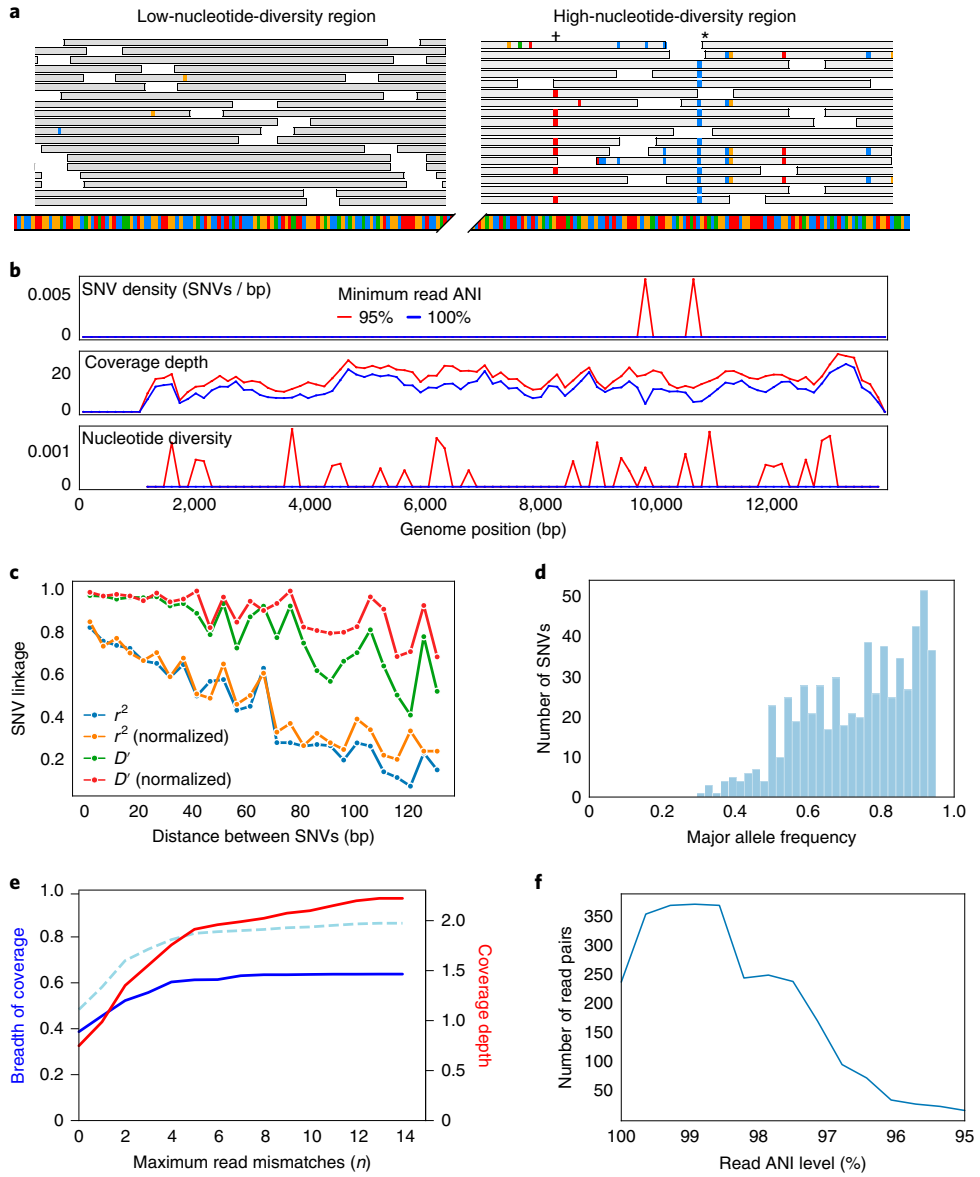

Fig. 1 | inStrain measures population-level diversity from metagenomic data. a, Examples of metagenomic reads (gray boxes) mapping to genomic regions with low and high nucleotide diversity. Mismatches to the reference genome are represented by small colored marks on the reads, and the reference genome 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 mismapping. Positions with nucleotide diversity and no SNV density are those where diversity exists but is not high enough to call an SNV. **c**, Metrics of SNV linkage (r^2 and D'; see Methods) versus distance between SNVs; linkage decay (as shown here) is a common signal of recombination. **d**, Distribution of the major allele frequencies of biallelic SNVs (the Site Frequency Spectrum). Alleles with major frequencies below 50% are the result of multiallelic sites. The lack of distinct puncta suggests 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.

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 (Supplementary Table 1). 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 were also used to assess the ability of 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

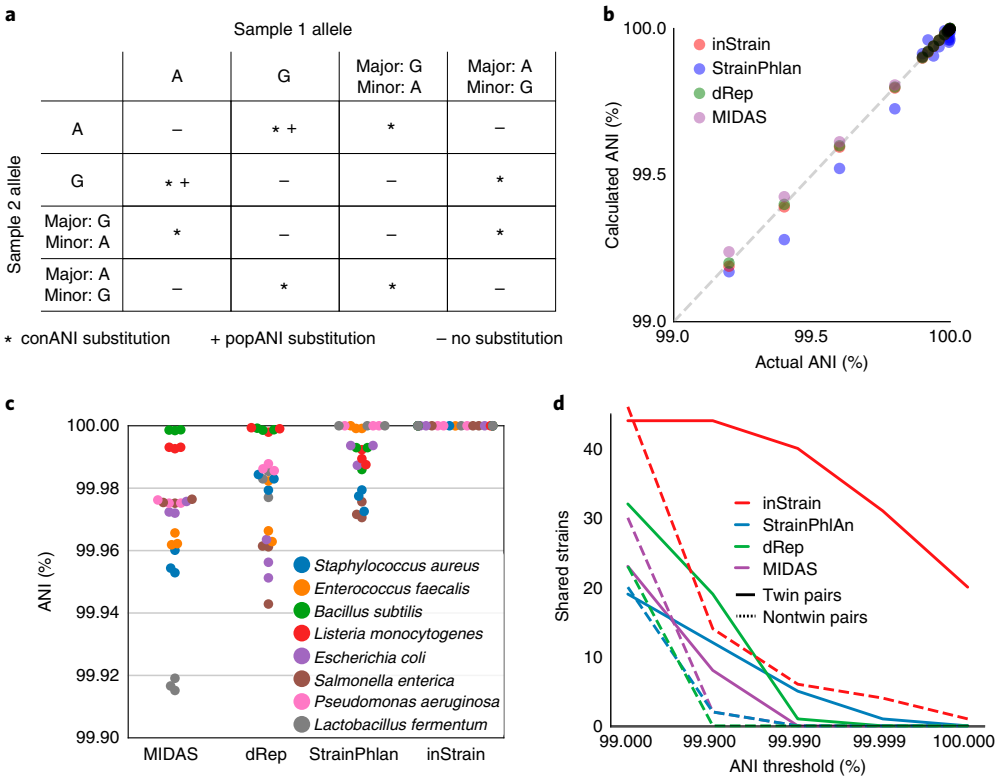

Fig. 2 | inStrain accurately discriminates between closely related strains. a, Table demonstrating the circumstances under which conANI and popANI substitutions will be called. ConANI substitutions are called whenever the consensus base differs, and popANI substitutions are only called when there is 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 with 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.

taxa) (Supplementary Table 2). The UHGG reference genomes had between 93.9% and 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) (Supplementary Table 1), 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 four tools. We then compared the number of strains found to be shared by twins and nontwins 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 any of the other tools (Fig. 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 for strain tracking.

Finally, we re-analyzed a previously generated dataset to compare data from inStrain with data from isolate-based sequencing[1]. We focused on individual S01, from which (1) 123 colonies of *Bacteroides fragilis* were isolated and sequenced from nine fecal samples collected over 2 years; and (2) metagenomic sequencing of the same fecal samples resulted in detection of a *B. fragilis* genome at 34× coverage (metagenomic data from all samples were 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 (Supplementary Table 3). If the isolate-detected mutations are considered ground truth (although 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 mutation). While the results of the methods were broadly consistent, the discrepancies between them may be due to shifting allele frequencies in the *B. fragilis* population during the 2 years that sampling occurred, as the isolate genomes were sampled evenly from all samples but most metagenomic reads came from the two samples where *B. fragilis* 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

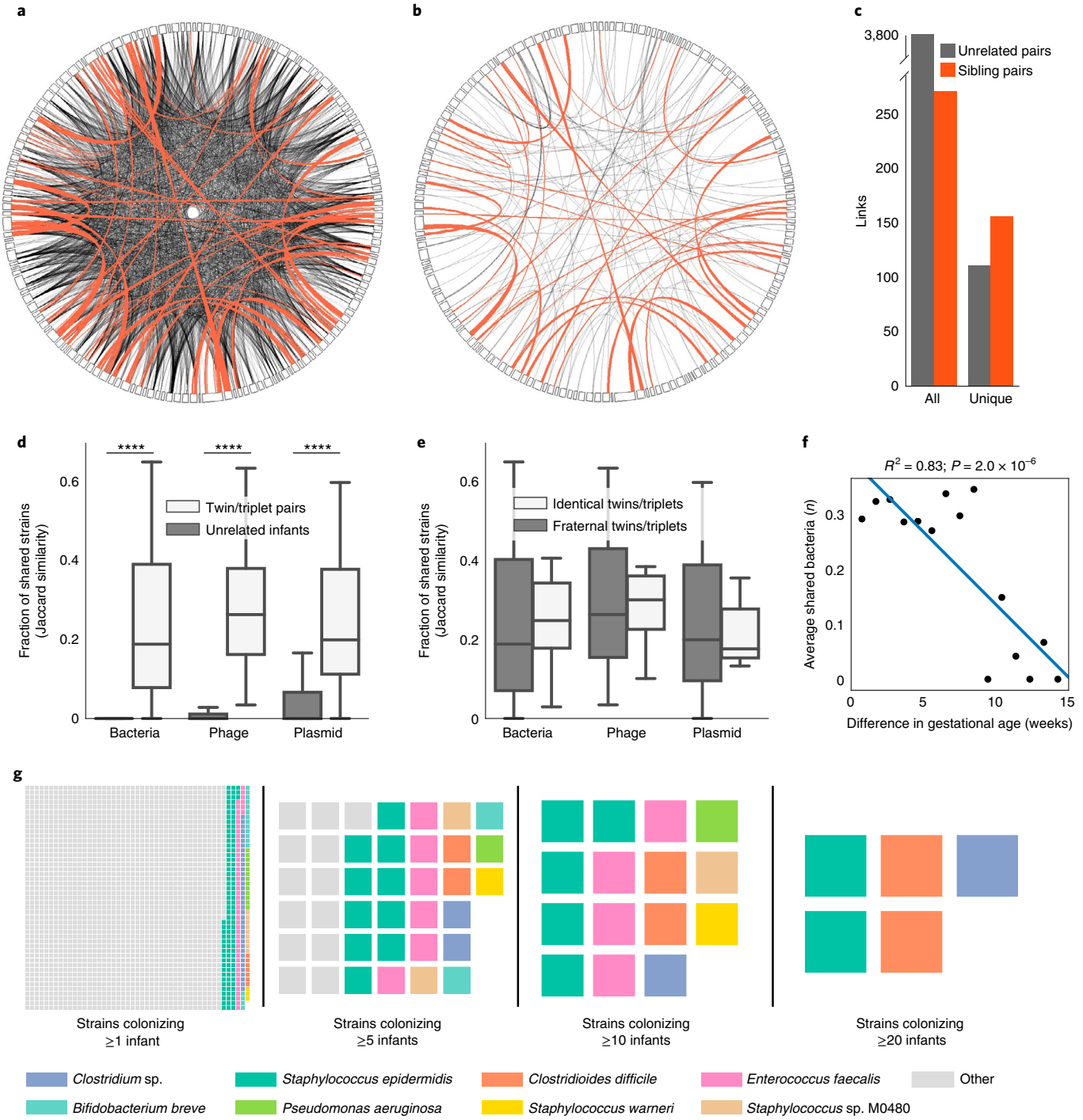

Fig. 3 | Siblings share significantly more microbial strains at birth than unrelated infants. a,**b**, A link is drawn for each strain shared between pairs of infants (represented by rectangles along the circumferences). Links between sibling pairs are drawn in red; links between unrelated infants are drawn in gray. 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 and triplet pairs ($n = 38$) share significantly more strains of bacteria ($P = 1.1 \times 10^{-22}$), phages ($P = 5.5 \times 10^{-26}$) and plasmids ($P = 9.8 \times 10^{-16}$) than unrelated pairs ($n = 12{,}842$). **e**, Identical twin pairs ($n = 6$) do not share significantly more strains of bacteria ($P = 0.87$), phages ($P = 0.88$) or plasmids ($P = 0.93$) than fraternal twin pairs ($n = 32$); *P* values for **d** and **e** from two-sided Wilcoxon rank-sum test; ****$P < 1 \times 10^{-15}$. **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. Data in **d** and **e** are represented as boxplots where the middle line is the median, the lower and upper box segments correspond to the first and third quartiles, the upper whisker extends from the hinge to the largest value no further than 1.5 × IQR from the hinge (where IQR is the interquartile range) and the lower whisker extends from the hinge to the smallest value at most 1.5 × IQR from the hinge, while data beyond the end of the whiskers are outlying points that are not depicted.

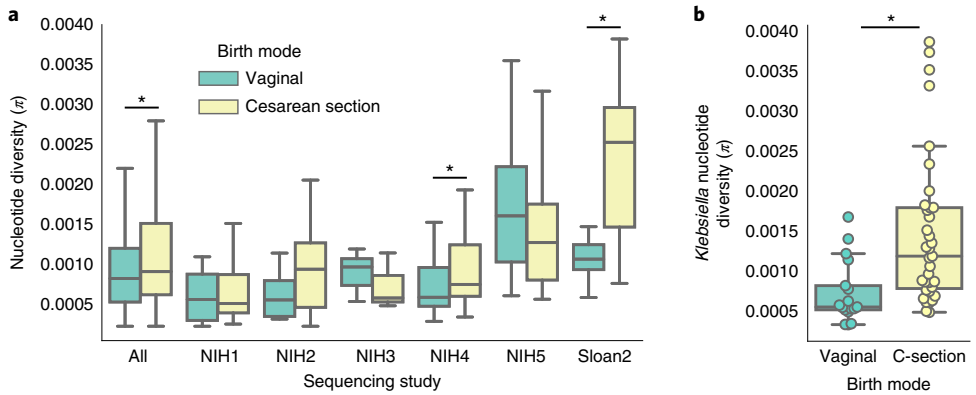

Fig. 4 | Analysis of the microdiversity of premature infant colonists. a, Infants born via cesarean section had bacterial colonists with higher nucleotide diversity than those delivered vaginally in the NIH4 campaign ($P=0.0002$), in the Sloan2 campaign ($P=0.0002$) and overall ($P=0.005$), but not in the NIH1 ($P=0.4$), NIH2 ($P=0.054$), NIH3 ($P=0.4$) and NIH5 ($P=0.07$) campaigns ($n=449$, 30, 54, 3, 210, 127 and 52 bacterial colonists of cesarean-section-born infants and $n=156$, 16, 10, 17, 67, 32 and 14 bacterial colonists of vaginally born infants overall and in the NIH1, NIH2, NIH3, NIH4, NIH5 and Sloan2 campaigns, respectively) (two-sided Wilcoxon rank-sum test with Benjamini–Hochberg *P* value correction for testing each campaign and overall; $^*P<0.05$). **b**, Organisms of the genus *Klebsiella* have significantly higher nucleotide diversity ($P=0.02$) in infants born via cesarean section (C-section) ($n=35$ organisms) than those delivered vaginally ($n=17$ organisms) (two-sided Wilcoxon rank-sum test with Benjamini–Hochberg *P* value correction[51] for testing each microbial species and genus present in both vaginally and cesarean-section-born infants; $^*P<0.05$). Data are represented as boxplots where the middle line is the median, the lower and upper box segments correspond to the first and third quartiles, the upper whisker extends from the hinge to the largest value no further than 1.5×IQR from the hinge and the lower whisker extends from the hinge to the smallest value at most 1.5×IQR from the hinge, while data beyond the end of the whiskers are outlying points that are not depicted (**a**) or depicted as points beyond the whiskers (**b**).

NICU[25]. The dataset includes samples from six individual sampling campaigns, and 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 8,000 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 dereplicated at 98% ANI to form 'subspecies' groups and representative genomes from each subspecies were combined into a single mapping database consisting of 2,266 genomes to reduce multi-mapped reads (Supplementary Fig. 4). 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 subspecies 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 (Fig. 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) (Supplementary Fig. 5). 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 (gray lines in Fig. 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 Fig. 3a). Further, the majority of bacterial strains that were identified in two and only two infants were shared between sibling pairs (Fig. 3b,c). Similar patterns were identified for bacteriophage and plasmid colonists (Supplementary Table 4).

The majority of bacterial strains (specific definition of 'strain' provided above) identified in this study were detected in only a single infant (1,818 of 3,044 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.* (Fig. 3g). These frequently encountered strains may have specific adaptations that enable them to survive in the NICU. Alternatively, they may be acquired from healthcare workers that commonly interact with these infants.

Overall, siblings shared significantly more strains of bacteria, bacteriophages and plasmids than unrelated infant pairs (Fig. 3d). However, among siblings, MZ twins shared no more strains than DZ twins and TZ triplets (Fig. 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 (Supplementary Fig. 6). 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 (Fig. 3f and Supplementary Fig. 6). In combination, the results point to the roles of infant physiology, sibling status and calendar date of birth (that is, 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 subspecies of bacteria, 26.9±1.5 subspecies of bacteriophages and 7.4±0.3 subspecies of plasmids per infant (mean±s.e.m.; colonization defined as detection of genome at ≥5× depth coverage across ≥50% of the genome) (Supplementary Table 5). 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 (Supplementary Fig. 6), but mea-

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	7.93×10^{-20}	PF00886.18	Ribosomal protein S16
	S. epidermidis	16419	9.51×10^{-15}	PF02597.19	ThiS family
	K. pneumoniae	15325	4.33×10^{-10}	PF02617.16	ATP-dependent Clp protease adapter protein ClpS
Phage	*Escherichia*	223	9.39×10^{-06}	PF08775.9	ParB family
	E. coli	205	0.00011027	PF02924.13	Bacteriophage λ 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	6.23×10^{-43}	PF05345.11	Putative immunoglobulin domain
	K. pneumoniae	15332	2.38×10^{-34}	PF00465.18	Iron-containing alcohol dehydrogenase
	E. faecalis	23792	4.03×10^{-32}	PF13731.5	WxL domain surface cell wall-binding
Phage	*Escherichia*	226	1.25×10^{-13}	PF03400.12	IS1 transposase
	E. coli	243	6.41×10^{-12}	PF03406.12	Phage tail fiber repeat
	E. faecalis	293	1.18×10^{-08}	PF01183.19	Glycosyl hydrolases family 25
Plasmid	Unknown	358	1.95×10^{-28}	PF00665.25	Integrase core domain
	Bacilli	11	4.84×10^{-25}	PF03432.13	Relaxase/mobilization nuclease domain
	Clostridium	374	9.71×10^{-14}	PF02782.15	FGGY family of carbohydrate kinases, C-terminal domain

sured 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 (Supplementary Fig. 7). This is likely due to differences in the read error profiles associated with the sequencing platforms[26]. Notably, bacterial nucleotide diversity was not associated with sequencing depth in any campaign (Supplementary Fig. 7d). 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 phages in four of six campaigns, whereas plasmids had the lowest nucleotide diversity in four of six campaigns (Supplementary Fig. 7). Relative to other bacteria, Proteobacteria had significantly higher and Firmicutes significantly lower nucleotide diversity in three of six and in four of six campaigns, respectively (Supplementary Table 4). Approximately 75% of premature infants were born via cesarean section (118 of 160), and their bacterial colonists had significantly higher nucleotide diversity than vaginally delivered infants in the NIH4 and Sloan2 cohorts and overall (Fig. 4a). This effect was particularly striking for *Klebsiella* (Fig. 4b), and the difference in *Klebsiella* microdiversity remained significant even when excluding infants in the NIH4 and Sloan2 cohorts (Supplementary Fig. 7).

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 housekeeping genes such as ribosomal protein S16 in bacteria and ParB in bacteriophage (where it maintains circular lysogens[27]), as well as genes with more interesting functions, including a plasmid-encoded polymyxin resistance protein, which is predicted to confer resistance to polymyxin antibiotics[28], and bacteriophage λ head decoration protein D, which stabilizes the expansion of the capsid after genome packaging[29]. Among the genes with significantly higher microdiversity than the average gene are a bacterially encoded gene with an immunoglobulin domain (which can be involved in cell adhesion and invasion[30]) and a bacteriophage gene encoding tail fibers (which are often involved in host cell recognition[31]). Interestingly, both the immunoglobulin 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 *E. faecalis* bacteriophage (subspecies 482_10.ph) that was present at high coverage depth (>20×) and breadth of coverage (>80%) in 44 infants in our cohort (Supplementary Table 5). We identified 410 loci with SNSs fixed between infants, 679 loci with SNVs with multiple alleles in the same infant and 1,062 loci where both were observed (Fig. 5d). 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. Of the intra-infant SNVs, 18% were found to be polymorphic in at least three different infants, indicating an overlap in variants across infants (Supplementary Table 6). Genomic regions and genes with a substantial number of intra-infant SNVs had correspondingly more inter-infant substitutions (Fig. 5a,b).

Seven of the 51 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 (Fig. 5a,c). The predicted distal tail gene, which might play a role in host specificity, was also found to have an

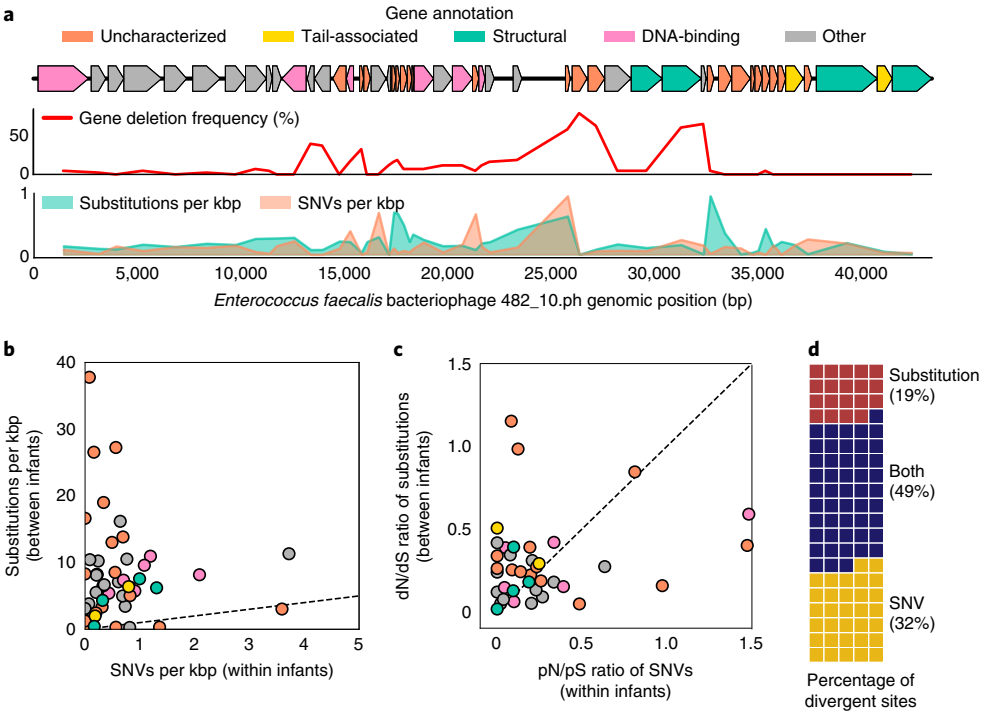

Fig. 5 | Tracking specific genetic differences within and between populations of an *E. faecalis* bacteriophage. a, Frequencies of gene deletions, substitutions and SNVs for all genes across an *E. faecalis* bacteriophage genome identified in 44 infants. Genes are colored based on their annotations. **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 nonsynonymous to synonymous substitutions (dN/dS) and ratios of nonsynonymous 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.

intra-infant pN/pS ratio of 0 (6 synonymous SNVs and 0 nonsynonymous 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 (Fig. 5c). 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[15,19,32–35]. 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 (Fig. 2) used in recent high-profile publications to quantify the ecologically critical process of microbiome transmission[14,36]. 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 with DZ twins can be ascribed to genetic effects[37]. Although studies of adult twins have consistently found some microbial taxa to be more commonly identified in MZ than DZ twins[38–41], diet and lifestyle preferences have also been shown to be more similar in MZ twins than DZ twins[42–44], presenting potential for confounding effects. In contrast with earlier studies, all subjects in the current study were housed in the same NICU for the entirety of the sampling time. Our findings, based on demonstrably robust methods, indicate that MZ twins shared no more strains of bacteria, bacteriophages 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[45,46]. The hospitalized premature infants in this study were all given prophylactic antibiotics immediately after birth and were 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 the following: (1) Family-specific sources. Strains present in two and only two infants were significantly more likely to be shared between siblings (Fig. 3), highlighting the role of strain sources such as shared visitors and/or parents in infant colonization. (2) The hospital environment. Nonsibling 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 (*Enterococcus faecium*, *Staphylococcus aureus*, *Klebsiella pneumoniae*, *Acinetobacter baumannii*, *Pseudomonas aeruginosa* and *Enterobacter* spp.) 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 who interact with many infants. (3) 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 (for example, formula feeding versus breast milk). (4) Unique sources. The majority of strains identified were found in only a single infant, demonstrating that even in a highly cleaned environment such as 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 cesarean 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[47]) than in the maternal microbiome. The general increase in nucleotide diversity and dN/dS ratios of genes involved in cell–cell interactions compared with other functions indicates that these genes are likely under diversifying selection. Identification of housekeeping 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 (Fig. 5). 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[48–50]. 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 and population-level phenotypes, and to track ecologically relevant processes such as strain migration and in situ evolution.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41587-020-00797-0.

Received: 31 January 2020; Accepted: 11 December 2020;
Published online: 18 January 2021

References

1. Zhao, S. et al. Adaptive evolution within gut microbiomes of healthy people. *Cell Host Microbe* **25**, 656–667.e8 (2019).
2. Schloissnig, S. et al. Genomic variation landscape of the human gut microbiome. *Nature* **493**, 45–50 (2012).
3. Simmons, S. L. et al. Population genomic analysis of strain variation in *Leptospirillum* group II bacteria involved in acid mine drainage formation. *PLoS Biol.* **6**, e177 (2008).
4. Eppley, J. M., Tyson, G. W., Getz, W. M. & Banfield, J. F. Genetic exchange across a species boundary in the archaeal genus *Ferroplasma*. *Genetics* **177**, 407–416 (2007).
5. Good, B. H., McDonald, M. J., Barrick, J. E., Lenski, R. E. & Desai, M. M. The dynamics of molecular evolution over 60,000 generations. *Nature* https://doi.org/10.1038/nature24287 (2017).
6. Ignacio-Espinoza, J. C., Ahlgren, N. A. & Fuhrman, J. A. Long-term stability and Red Queen-like strain dynamics in marine viruses. *Nat. Microbiol.* https://doi.org/10.1038/s41564-019-0628-x (2019).
7. Bendall, M. L. et al. Genome-wide selective sweeps and gene-specific sweeps in natural bacterial populations. *ISME J.* **10**, 1589–1601 (2016).
8. Delmont, T. O. et al. Single-amino acid variants reveal evolutionary processes that shape the biogeography of a global SAR11 subclade. *eLife* **8**, e46497 (2019).
9. Garud, N. R., Good, B. H., Hallatschek, O. & Pollard, K. S. Evolutionary dynamics of bacteria in the gut microbiome within and across hosts. *PLoS Biol.* **17**, e3000102 (2019).
10. Smillie, C. S. 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).
11. Siranosian, B. A., Tamburini, F. B., Sherlock, G. & Bhatt, A. S. Acquisition, transmission and strain diversity of human gut-colonizing crAss-like phages. *Nat. Commun.* **11**, 280 (2020).
12. Crits-Christoph, A., Olm, M. R., Diamond, S., Bouma-Gregson, K. & Banfield, J. F. Soil bacterial populations are shaped by recombination and gene-specific selection across a grassland meadow. *ISME J.* **14**, 1834–1846 (2020).
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).
14. Shao, Y. et al. Stunted microbiota and opportunistic pathogen colonization in caesarean-section birth. *Nature* **574**, 117–121 (2019).
15. Korpela, K. et al. Selective maternal seeding and environment shape the human gut microbiome. *Genome Res.* https://doi.org/10.1101/gr.233940.117 (2018).
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).
17. Olm, M. R., Brown, C. T., Brooks, B. & Banfield, J. F. 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).
18. Truong, D. T., Tett, A., Pasolli, E., Huttenhower, C. & Segata, N. Microbial strain-level population structure and genetic diversity from metagenomes. *Genome Res.* **27**, 626–638 (2017).
19. Nayfach, S., Rodriguez-Mueller, B., Garud, N. & Pollard, K. S. An integrated metagenomics pipeline for strain profiling reveals novel patterns of bacterial transmission and biogeography. *Genome Res.* **26**, 1612–1625 (2016).
20. Brito, I. L. et al. Transmission of human-associated microbiota along family and social networks. *Nat. Microbiol.* **4**, 964–971 (2019).
21. Costea, P. I. et al. metaSNV: a tool for metagenomic strain level analysis. *PLoS ONE* **12**, e0182392 (2017).
22. Truong, D. T. et al. MetaPhlAn2 for enhanced metagenomic taxonomic profiling. *Nat. Methods* **12**, 902–903 (2015).
23. Nei, M. & Li, W. H. Mathematical model for studying genetic variation in terms of restriction endonucleases. *Proc. Natl Acad. Sci. USA* **76**, 5269–5273 (1979).
24. Almeida, A. et al. A unified catalog of 204,938 reference genomes from the human gut microbiome. *Nat. Biotechnol.* https://doi.org/10.1038/s41587-020-0603-3 (2020).
25. Olm, M. R. et al. Necrotizing enterocolitis is preceded by increased gut bacterial replication, *Klebsiella*, and fimbriae-encoding bacteria. *Sci. Adv.* **5**, eaax5727 (2019).
26. Schirmer, M., D'Amore, R., Ijaz, U. Z., Hall, N. & Quince, C. Illumina error profiles: resolving fine-scale variation in metagenomic sequencing data. *BMC Bioinf.* **17**, 125 (2016).
27. Lobocka, M. & Yarmolinsky, M. P1 plasmid partition: a mutational analysis of ParB. *J. Mol. Biol.* **259**, 366–382 (1996).
28. Fu, W. et al. First structure of the polymyxin resistance proteins. *Biochem. Biophys. Res. Commun.* **361**, 1033–1037 (2007).
29. 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).

30. Bodelón, G., Palomino, C. & Fernández, L. Á. Immunoglobulin domains in *Escherichia coli* and other enterobacteria: from pathogenesis to applications in antibody technologies. *FEMS Microbiol. Rev.* **37**, 204–250 (2013).
31. Tétart, F., Repoila, F., Monod, C. & Krisch, H. M. Bacteriophage T4 host range is expanded by duplications of a small domain of the tail fiber adhesin. *J. Mol. Biol.* **258**, 726–731 (1996).
32. 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).
33. 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).
34. Eren, A. M. et al. Anvi'o: an advanced analysis and visualization platform for 'omics data. *PeerJ* **3**, e1319 (2015).
35. Brito, I. L. & Alm, E. J. Tracking strains in the microbiome: insights from metagenomics and models. *Front. Microbiol.* **7**, 712 (2016).
36. 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).
37. Goodrich, J. K. et al. Genetic determinants of the gut microbiome in UK twins. *Cell Host Microbe* **19**, 731–743 (2016).
38. Lim, M. Y. et al. The effect of heritability and host genetics on the gut microbiota and metabolic syndrome. *Gut* **66**, 1031–1038 (2017).
39. Turpin, W. et al. Association of host genome with intestinal microbial composition in a large healthy cohort. *Nat. Genet.* **48**, 1413–1417 (2016).
40. Davenport, E. R. et al. Genome-wide association studies of the human gut microbiota. *PLoS ONE* **10**, e0140301 (2015).
41. Goodrich, J. K., Davenport, E. R., Clark, A. G. & Ley, R. E. The relationship between the human genome and microbiome comes into view. *Annu. Rev. Genet.* **51**, 413–433 (2017).
42. 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).
43. 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).
44. Vinkhuyzen, A. A. E., van der Sluis, S., de Geus, E. J. C., Boomsma, D. I. & Posthuma, D. Genetic influences on 'environmental' factors. *Genes Brain Behav.* **9**, 276–287 (2010).
45. Faith, J. J. et al. The long-term stability of the human gut microbiota. *Science* **341**, 1237439–1237439 (2013).
46. Ding, T. & Schloss, P. D. Dynamics and associations of microbial community types across the human body. *Nature* **509**, 357–360 (2014).
47. Shin, H. et al. The first microbial environment of infants born by C-section: the operating room microbes. *Microbiome* **3**, 59 (2015).
48. Thévenon, S. & Couvet, D. The impact of inbreeding depression on population survival depending on demographic parameters. *Anim. Conserv.* **5**, 53–60 (2002).
49. Oh, J., Byrd, A. L., Park, M., Kong, H. H. & Segre, J. A. Temporal stability of the human skin microbiome. *Cell* **165**, 854–866 (2016).
50. Jovel, J. et al. Characterization of the gut microbiome using 16S or shotgun metagenomics. *Front. Microbiol.* **7**, 459 (2016).
51. 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).

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/nonsynonymous 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 Supplementary Software Manual 1.

Dependencies. inStrain requires Samtools[52] for interacting with .bam and .sam files, Prodigal[53] for annotating open reading frames (ORFs) 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 (1) a nucleotide sequence or a set of nucleotide sequences in fasta format, and (2) 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 and so on).

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. (1) 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. (2) 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). For example, if pairs have a median insert size of 500 bp, by default all pairs with insert sizes over 1,500 bp will be excluded. (3) 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 (that is, whether the read maps equally well to multiple genomic locations). The read in the pair with the higher mapQ is used for the pair. (4) Pairs must be above a user-defined minimum nucleotide identity value. For example, if reads in a pair are 100 bp 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: (1) average coverage; (2) median coverage; (3) standard deviation of coverage; (4) number of bases with 0 coverage; (5) breadth of coverage (the fraction of bases that are covered by at least a single read); (6) minCov breadth of coverage (the fraction of bases that are covered by at least the number of reads required to call an SNV, which is 5 by default); and (7) 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 expected breadth $= 1 - e^{0.883 \times \text{coverage}}$. If the breadth is substantially lower than the expected breadth, it indicates that reads are mapping only to a specific region of the scaffold (for example, a transposon, prophage or other mobile element).

The nucleotide diversity (π; ref. [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 three criteria. (1) There must be at least a user-defined minimum number of reads mapping to the position. By default this is 5. (2) More than a user-defined percentage of reads must have a variant base at that position. By default this is 5%. (3) 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 Phred Q30 score (probability of an incorrect base call 1 in 1,000) 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 1×10^{-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. Biallelic SNVs are positions where there are two alleles present in the reads at a position. Multiallelic 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 nonsynonymous (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 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. r^2 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 (1) 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; (2) mapping_info.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; (3) SNVs.tsv, which lists the location, reference base counts, variant base counts, number of alleles and other information about each identified SNV; (4) 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; (5) gene_info.tsv, which lists the coverage, nucleotide diversity and related metrics for each ORF; and (6) genome_info.tsv, which lists the metrics described in scaffold_info.tsv on a genome level (rather than a scaffold level). Finally, inStrain uses this information to generate a number of figures, examples of which are shown in Fig. 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 with the same nucleotide sequence(s). These comparisons are performed in a pairwise manner and follow a series of four steps. (1) 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. (2) Each position identified in step 1 is classified following the logic depicted in Fig. 2a as 'no SNV', 'consensus SNV' or 'population SNV'. If both samples have no SNVs called at a 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. (3) ConANI is calculated as (1 − number of consensus SNVs)/'compared_bases_count' (calculated in step 1), and popANI is calculated as (1 − number of population SNVs)/'compared_bases_count'. (4) Datatables are made listing the metrics calculated above on a scaffold-by-scaffold level as well as on 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 (Fig. 2b) were performed by using SNP Mutator[60] to introduce a known number of mutations into a reference genome (*E. coli* strain SQ88; RefSeq accession number GCF_000988385.1) and comparing the mutated genomes with the original reference genome. For dRep, mutated genomes were compared with the reference genome using dRep on default settings. For inStrain, MIDAS and StrainPhlAn, Illumina reads were simulated for all genomes at 20× coverage using pIRS[61]. For inStrain, synthetic reads were mapped back to the reference genome using Bowtie 2 (ref. [62]), profiled using 'inStrain profile' under default settings, and compared using 'inStrain compare' under default settings. For StrainPhlAn, synthetic reads were profiled with Metaphlan2 (ref. [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 Supplementary Table 1. 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 1 in 100 bases (see above paragraph for how this was made). We generated four versions of this genome made up of scaffolds of length 1 kbp, 10 kbp, 100 kbp or 1 Mbp, and evaluated the ability of 'inStrain profile' to detect the known mutations (Supplementary Fig. 3).

Isolate-based comparisons (Fig. 2c) were performed based on the ZymoBIOMICS Microbial Community Standards product (catalog no. 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 were 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 Supplementary Table 1. 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 assessed using all Zymo reads concatenated together, and raw values are available in Supplementary Table 2.

Twin-based comparisons (Fig. 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 and 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 dereplicated bacterial genomes assembled and binned from each infant (available from ref. [25]) were compared in a pairwise manner using dRep under default settings. For inStrain, 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 Supplementary Table 1.

To compare the biallelic mutations identified by isolate-based sequencing with those identified by metagenomic inStrain analysis, we downloaded all metagenomes and isolate genomes from individual S01 using the Sequence Read Archive (SRA) links provided in ref. [1]. We only considered the nine 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 *B. fragilis* reference genome (NC_003228.3) using Bowtie 2. 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 Supplementary Table 3.

To determine the sequencing coverage needed to detect minor alleles with 95% probability (Supplementary Fig. 2), 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 Phred Q30 Illumina errors). For each allele frequency (AF) between 5% and 50% (5%, 6%, 7% and so on), we iterated over each coverage value (c) between 1× and 150× 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 nonlinear least-squares fitting (Supplementary Fig. 2).

Calling, detection and profiling of subspecies of bacteria, bacteriophages and plasmids. Genomes of bacteria, bacteriophages, plasmids and eukaryotes were previously binned from the infants in 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 with each other using dRep version 2.2.0 under default settings, all bacteriophage genomes were compared with 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 with 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 scaffolds with taxonomic classifications specifying 'Eukarya' were marked as 'likely human' and excluded from further analysis. Information about subspecies is available in Supplementary Table 5.

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 subspecies genomes concatenated together using Bowtie 2 with default settings. 'inStrain profile' was run on all resulting mapping files with default settings. Detection of a subspecies in a sample was defined as that genome being present with ≥0.5 minCov breadth (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 subspecies abundance is available in Supplementary Table 5.

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 Supplementary Table 4. 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 was 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 Supplementary Table 4.

The number of strains shared between infants was visualized in Fig. 3a,b 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 nontwin pairs, as calculated using the Python module scipy.stats.ranksums[55]. Associations between the number of strains shared between infants and their differences in birth day, birth weight and gestational age were 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).

Nucleotide diversity analysis. The coReads inStrain analysis described above resulted in a total of 8,336 subspecies/infant pairs in which a subspecies genome was detected at 5× coverage across at least 50% of the genome (Supplementary Fig. 4). 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 to generate P values (Fig. 4 and Supplementary Figs. 6 and 7).

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 was compared with 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[51]. Genes were annotated based on Pfam database hidden Markov models[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. *E. faecalis* bacteriophage subspecies 482_10.ph was identified with at least 80% breadth of coverage and 20× coverage depth in the coReads of 44 infants. ORFs were called using Prodigal in metagenome mode, and genes were annotated using USEARCH to search against the UniRef100 database. Gene categories (tail-associated, structural and so on) were assigned based on manual inspection of the resulting database hits. The gene map presented in Fig. 5a was generated using the Python module 'dna_features_viewer'.

Biallelic 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 numbers of synonymous and nonsynonymous sites in each gene were determined using methods from the script 'dnds_from_drep.py'[69]. dN/dS was calculated using the formula ((nonsynonymous substitutions/nonsynonymous sites)/(synonymous substitutions/synonymous sites)), and pN/pS was calculated using the formula ((nonsynonymous SNVs/nonsynonymous 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 5× coverage; the coverage required to call an SNV) of the gene for each infant the gene was identified in. Genes with a minCov 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.

Reporting Summary. Further information on research design is available in the Nature Research Reporting Summary linked to this article.

Data availability

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.

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/.

References

52. Li, H. et al. The sequence alignment/map format and SAMtools. *Bioinformatics* **25**, 2078–2079 (2009).
53. Hyatt, D. et al. Prodigal: prokaryotic gene recognition and translation initiation site identification. *BMC Bioinf.* **11**, 119 (2010).
54. McKinney, W. et al. Data structures for statistical computing in python. in *Proc. 9th Python in Science Conf.* **445**, 51–56 (2010).
55. Jones, E., Oliphant, T. & Peterson, P. SciPy: open source scientific tools for Python (SciPy Developers, 2001); http://scipy.org
56. Harris, C. R. et al. Array programming with NumPy. *Nature* **585**, 357–362 (2020).
57. Hunter, J. D. Matplotlib: a 2D graphics environment. *Comput. Sci. Eng.* **9**, 90–95 (2007).
58. Waskom, M. et al. mwaskom/seaborn: v0.11.1. https://doi.org/10.5281/ZENODO.592845 (2020).
59. VanLiere, J. M. & Rosenberg, N. A. Mathematical properties of the r^2 measure of linkage disequilibrium. *Theor. Popul. Biol.* **74**, 130–137 (2008).
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).
62. Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nat. Methods* **9**, 357–359 (2012).
63. Cock, P. J. A. et al. Biopython: freely available Python tools for computational molecular biology and bioinformatics. *Bioinformatics* **25**, 1422–1423 (2009).
64. Peng, Y., Leung, H. C. M., Yiu, S. M. & Chin, F. Y. L. IDBA-UD: a de novo assembler for single-cell and metagenomic sequencing data with highly uneven depth. *Bioinformatics* **28**, 1420–1428 (2012).
65. Delcher, A. L., Phillippy, A., Carlton, J. & Salzberg, S. L. Fast algorithms for large-scale genome alignment and comparison. *Nucleic Acids Res.* **30**, 2478–2483 (2002).
66. Bushnell, B., Rood, J. & Singer, E. BBMerge—accurate paired shotgun read merging via overlap. *PLoS ONE* **12**, e0185056 (2017).
67. Krzywinski, M. et al. Circos: an information aesthetic for comparative genomics. *Genome Res.* **19**, 1639–1645 (2009).
68. El-Gebali, S. et al. The Pfam protein families database in 2019. *Nucleic Acids Res.* https://doi.org/10.1093/nar/gky995 (2018).
69. Olm, M. R. et al. Consistent metagenome-derived metrics verify and delineate bacterial species boundaries. *mSystems* https://doi.org/10.1128/mSystems.00731-19 (2020).

Acknowledgements

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

Author contributions

M.R.O., M.J.M. and J.F.B. designed the study. M.R.O. performed metagenomic analyses. M.R.O., A.C.-C. and K.B.-G. contributed to software development and population genomic analyses. B.A.F. performed all DNA extractions. M.R.O. and J.F.B. wrote the manuscript and all authors contributed to manuscript revisions.

Competing interests

J.F.B. is a founder of Metagenomi.

Additional information

Supplementary information is available for this paper at https://doi.org/10.1038/s41587-020-00797-0.

Correspondence and requests for materials should be addressed to J.F.B.

Reprints and permissions information is available at www.nature.com/reprints.

nature research

Corresponding author(s): Jillian Banfield

Last updated by author(s): November 24, 2020

Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see Authors & Referees and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed	
☐	☒	The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☐	☒	A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	☒	The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☐	☒	A description of all covariates tested
☐	☒	A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐	☒	A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☒	☐	For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☒	☐	For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	☐	For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☒	☐	Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	New data was not collected in this study
Data analysis	samtools version v1.9 Prodigal version V2.6.3 Bowtie2 version 2.3.4.1 SNP Mutator version 0.2.0 pIRS version 2.0.0 MetaPhlAn version 2.7.7 IDBA-UD version 1.1.3 NUCmer version 3.1 dRep version 2.2.0 cicros version 0.69 usearch version v10.0 Details of how software was used is available in the methods section

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors/reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on 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

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	This study made use of 1,163 previously reported fecal metagenomes from 160 premature infants. All publicly available fecal metagenomes collected from infants from the University of Pittsburgh Medical Center Magee-Womens Hospital were included in this study
Data exclusions	No data was excluded from this study
Replication	No wet-lab experiments were performed in this study
Randomization	Experimental groups were not generated in this study
Blinding	Blinding is not relevant to this study which primarily related to a novel bioinformatics algorithm

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☒	☐ Antibodies
☒	☐ Eukaryotic cell lines
☒	☐ Palaeontology
☒	☐ Animals and other organisms
☒	☐ Human research participants
☒	☐ Clinical data

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☒	☐ MRI-based neuroimaging