nature biotechnology

ARTICLES

Metagenomic binning and association of plasmids with bacterial host genomes using DNA methylation

John Beaulaurier[1,2], Shijia Zhu[1,2], Gintaras Deikus[1,2], Ilaria Mogno[1–3], Xue-Song Zhang[4], Austin Davis-Richardson[5], Ronald Canepa[5], Eric W Triplett[5], Jeremiah J Faith[1–3], Robert Sebra[1,2,6], Eric E Schadt[1,2,6] & Gang Fang[1,2]

Shotgun metagenomics methods enable characterization of microbial communities in human microbiome and environmental samples. Assembly of metagenome sequences does not output whole genomes, so computational binning methods have been developed to cluster sequences into genome 'bins'. These methods exploit sequence composition, species abundance, or chromosome organization but cannot fully distinguish closely related species and strains. We present a binning method that incorporates bacterial DNA methylation signatures, which are detected using single-molecule real-time sequencing. Our method takes advantage of these endogenous epigenetic barcodes to resolve individual reads and assembled contigs into species- and strain-level bins. We validate our method using synthetic and real microbiome sequences. In addition to genome binning, we show that our method links plasmids and other mobile genetic elements to their host species in a real microbiome sample. Incorporation of DNA methylation information into shotgun metagenomics analyses will complement existing methods to enable more accurate sequence binning.

Despite growing appreciation for the role of microbial communities in human health[1,2], comprehensive characterization of microbiomes remains difficult. Culture-independent sequencing of clinical and environmental samples has revealed the immense diversity of microbial life. Unlike 16S rRNA gene sequencing[3], whole metagenome shotgun sequencing[4] can identify chromosomes, plasmids, and bacteriophages[5,6]. This approach also enables better phylogenetic resolution than 16S rRNA gene amplicon sequencing[7,8].

Shotgun-sequenced metagenomes are diverse and complex, meaning that the sequenced reads and assembled contigs are challenging to interpret. Reference genome sequences of cultivated organisms can help with metagenome annotation[9,10], but sequences from bacteria lacking cultivated relatives are segregated into putative taxa and species with 'binning' methods. Unsupervised binning methods do not require data from reference genomes.

Sequence composition features can be used to bin sequences[11–14], but often fail to segregate sequences from very similar genomes[11,13]. Coverage features that are based on similar abundance profiles across multiple samples provide a powerful means of binning assembled contigs[15–18]. However, they cannot effectively bin mobile genetic elements (MGEs), especially plasmids that replicate separately from bacterial chromosomes. Chromosomal interaction maps discerned using Hi-C can link assembled contigs, including plasmids[19–21], but cannot distinguish between closely related organisms due to high sequence similarity and uneven Hi-C link densities[20].

DNA methylation in bacteria and archaea is catalyzed by DNA methyltransferases (MTases) that add methyl groups to nucleotides in a highly sequence-specific manner. Some sequence motifs in DNA molecules are almost 100% methylated whereas other motifs remain non-methylated[22–25]. A survey of 230 diverse bacterial and archaeal genomes found evidence of DNA methylation in 93% of genomes, with a diverse array of methylated motifs (834 distinct motifs; average of three motifs per organism)[25]. Horizontal gene transfer of MGEs containing MTase genes is the main driver of diversity in bacterial methylomes[25–27]. Importantly, the full genetic complements of a cell (chromosomes and MGEs) are methylated by MTases and therefore share the same set of methylated motifs. These motifs often differ among species and strains[24,25], making it possible to use combinations of methylated motifs (*endogenous epigenetic barcode*) for metagenomic binning.

We develop a method that uses single-molecule, real-time (SMRT) sequencing of metagenomic DNA to capture methylated motifs. We show that methylation motifs can be combined with composition and coverage features to improve genome segregation and link MGEs to their host chromosomes.

RESULTS

Methylation profiles in metagenome sequences

As with sequence composition or differential coverage profiles, which normalize *k*-mer frequencies across *k*-mers or normalize coverage

[1]Department of Genetics and Genomic Sciences, Icahn School of Medicine at Mount Sinai, New York, New York, USA. [2]Icahn Institute for Genomics and Multiscale Biology, Icahn School of Medicine at Mount Sinai, New York, New York, USA. [3]Immunology Institute, Icahn School of Medicine at Mount Sinai, New York, New York, USA. [4]Department of Medicine, New York University School of Medicine, New York, New York, USA. [5]Department of Microbiology and Cell Science, Institute of Food and Agricultural Sciences, University of Florida, Gainesville, Florida, USA. [6]Sema4, a Mount Sinai venture, Stamford, Connecticut, USA. Correspondence should be addressed to G.F. (gang.fang@mssm.edu).

Received 13 September, 2016; accepted 13 November, 2017; published online 11 December 2017; doi:10.1038/nbt.4037

values across samples, respectively, DNA methylation can be used as a feature to bin sequences. In the case of methylation profiles, each sequence has a feature set consisting of DNA methylation scores across motifs (**Fig. 1**). The methylation score for a given motif on a contig reflects the extent to which all instances of that motif are methylated and is calculated using inter-pulse duration (IPD) values that measure the time it takes a DNA polymerase to translocate from one nucleotide to the next during SMRT sequencing[22,28,29].

The sensitivity and specificity of a motif methylation score are a function of the number of IPD values making up the score (**Fig. 2a**). The IPD count for each motif is determined by both the number of motif sites on the contig, which is generally larger for shorter motifs, and the number of reads aligning to the contig, as each read contributes independent IPD measurements[22].

We compiled methylation scores for multiple motifs into methylation profiles. The methylated motifs included in the profile were determined using a motif filtering approach that we developed for this study. After assessing the methylation scores for all possible motifs in a subset of the metagenomic sequencing data, only those motifs with evidence of methylation in at least one of the assembled contigs were retained for inclusion in the methylation profiles (**Supplementary Methods**). Filtering resulted in profiles of 7–38 motifs for the metagenomic samples that we analyzed (**Supplementary Table 1**). It is the combination of methylated motifs in this set of filtered motifs that provides the discriminative power for methylation binning. The code for motif filtering and methylation binning (**Supplementary Code**) is available at https://github.com/fanglab/mbin.

Binning assembled contigs using methylation profiles

To evaluate DNA methylation profiles as features for metagenomic binning, we first created a synthetic metagenomic mixture of SMRT sequencing reads from eight separately sequenced bacterial species (**Supplementary Table 2**). All sequencing data from this study are available through NCBI BioProject PRJNA404082. Following metagenomic assembly of the combined reads (**Supplementary Table 3**), our motif filtering procedure identified 16 N6-methyladenine (6mA) motifs from the metagenomic contigs based solely on methylation scores (**Supplementary Table 1**), 14 (87.5%) of which were exact matches to the true methylated motifs (as validated by independent methylation analysis of each species before mixing). The remaining two motifs, GAGC and TCACNNNNNATG, are closely related to the true motifs, GGAG and CACNNNNNATG: instances of the detected GAGC motif that are preceded by a guanine are expected to be methylated, while all instances of TCACNNNNNATG are expected to be methylated as they are further specifications of the true motif. Hierarchical clustering of the motif methylation scores for the largest contigs from each species revealed unique methylation profiles for each species across the detected 16 motifs (**Fig. 2b**).

To visualize and interpret methylation features across multiple metagenomic contigs, we used the dimensionality reduction algorithm t-distributed stochastic neighbor embedding (t-SNE)[30,31], which has previously been used to visualize metagenomic sequence composition features[13,14]. The two-dimensional (2D) map of methylation features generated by t-SNE revealed contigs that were well clustered at the species level (**Fig. 2c**). We conservatively picked eight bins in the 2D map and assessed binning quality by aligning the binned contigs to reference genome sequences. We found >98% completeness in 7 of 8 bins (76.91% in the *Clostridium bolteae* bin) and <1% contamination in 7 of 8 bins (4.28% in the *Ruminococcus gnavus* bin) (**Supplementary Table 4**). Notably, four species from the *Bacteroides* genus showed better separation than was possible using

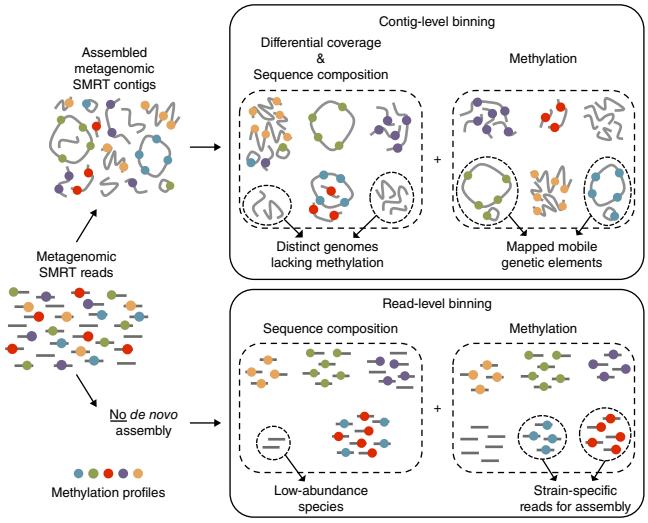

Figure 1 Overview of metagenomic binning using DNA methylation detected in SMRT long reads. Given a set of metagenomic shotgun SMRT sequencing reads, one can either assemble them into contigs for contig-level binning or can directly perform read-level binning without *de novo* assembly. A widely used approach for unsupervised binning of metagenomic contigs uses coverage (and its covariance across multiple samples) and sequence composition profiles, but these can be complemented by methylation profiles to better segregate contigs with similar sequence composition and coverage covariance, as well as to map mobile genetic elements to contigs from their host bacterium in the microbiome sample. Read-level binning by sequence composition can isolate reads from low-abundance species that do not assemble into contigs, while read-level binning by methylation profiles can segregate reads from multiple strains for the purpose of separate, strain-specific *de novo* genome assemblies. These methylation and composition features can be combined with abundance features to maximize binning resolution.

either t-SNE to generate a scatter plot of 5-mer frequency features alone (**Supplementary Fig. 1a**) or a scatter plot of contig coverage values versus GC-content (**Supplementary Fig. 1b**). Two small, high-coverage *Collinsella aerofaciens* contigs (putative plasmids) in the coverage versus GC-content plot illustrate how the coverage values of plasmids can differ dramatically from those of their host chromosomes, rendering coverage-based binning methods unable to identify the plasmid host in metagenomic samples.

Some small contigs were too short (e.g., <20 kb) to contain all of the motif sites in the methylation profile, which can lead to imperfect clustering if methylation of the missing motifs is a major discriminating feature between clusters. For example, several small contigs from *Clostridium bolteae* were missing certain methylated motif sites (**Supplementary Fig. 2**) and therefore clustered more closely with *Ruminococcus gnavus*, one of the rare species lacking methylation[25]. In such cases, complementary discriminative features, like sequence composition or coverage, should be leveraged and integrated.

Next, we analyzed methylation profiles of contigs assembled from SMRT sequencing of a fecal microbiota sample isolated from an adult mouse (**Supplementary Table 2**). 16S rRNA gene amplicon sequencing showed that the sample was of low- to medium-complexity and dominated by an unknown number of organisms from the S24-7 family of the order *Bacteroidales* (**Fig. 2d**; SRX3160950). We applied motif filtering to detect 38 methylated motifs (**Supplementary Table 1**) in the assembly (**Supplementary Table 3**) and visualized the methylation

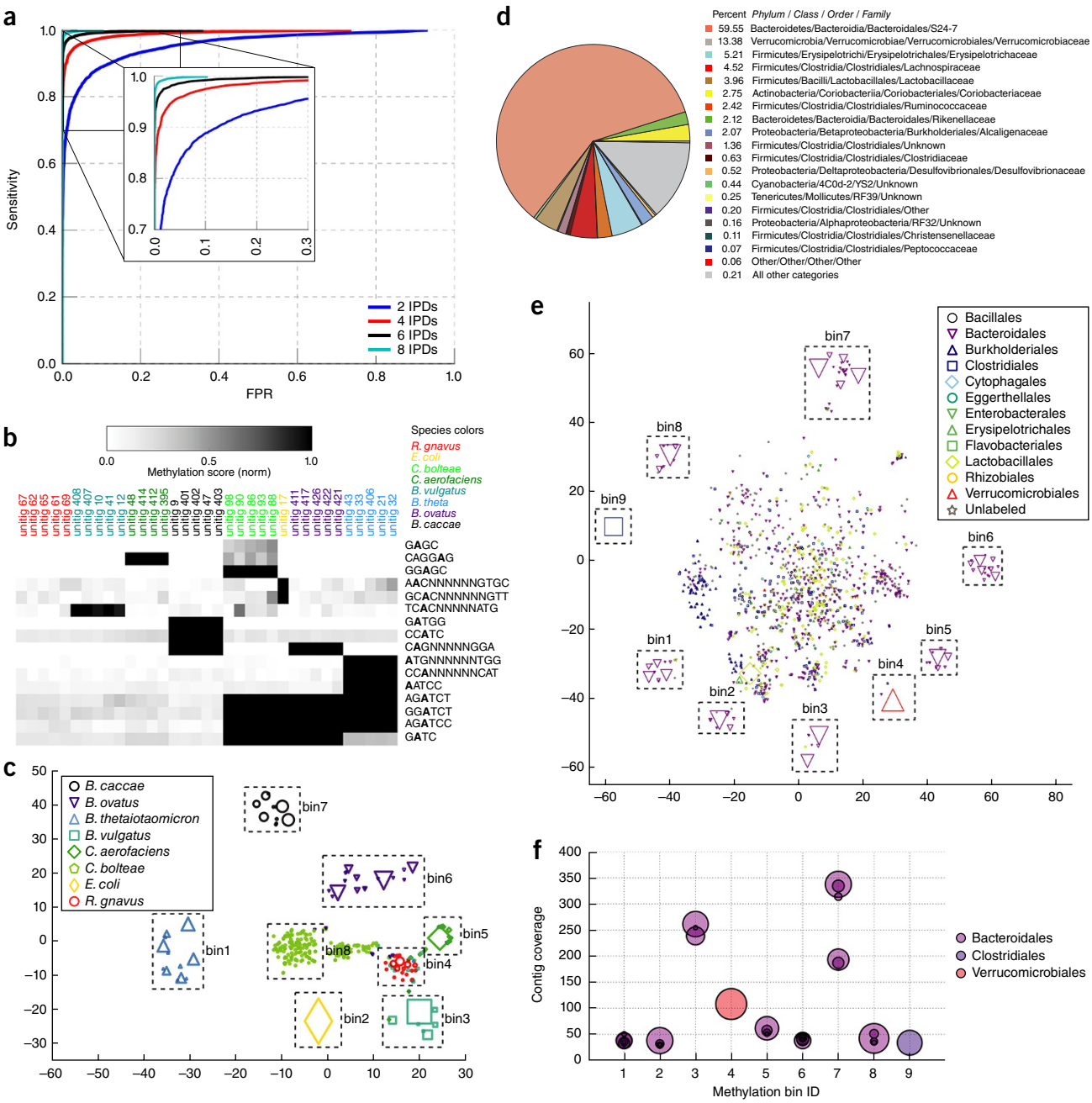

Figure 2 Metagenomic binning by methylation profiles. (**a**) Receiver operating characteristic (ROC) curve illustrating the power to classify a contig as methylated (N6-methyladenine, 6mA) or non-methylated regarding a specific sequence motif, as a function of the number of IPD values available for the motif sites on the contig. FPR, false-positive rate. (**b**) Heatmap of contig-level methylation scores for 16 motifs on a set of contigs from a metagenomic assembly of eight bacterial species. Contigs from each species possess distinct methylation profiles across the selected motifs. (**c**) t-SNE scatter plot of contig-level methylation scores across 16 selected motifs, with manually selected bins marked by boxes. Cluster silhouette coefficients[51] were computed for the contigs from the four *Bacteroides* species. The coefficients (–1 indicates complete mixing, while 1 indicates complete separation) were 0.53 using methylation features and t-SNE, 0.14 using 5-mer frequency features and t-SNE (**Supplementary Fig. 1a**), and –0.03 using plotted coverage vs. GC-content values (**Supplementary Fig. 1b**). (**d**) Family-level annotation of 16S rRNA gene amplicon sequencing reads from an adult mouse gut microbiome by QIIME[52]. (**e**) t-SNE projection of metagenomic contigs assembled from SMRT reads of an adult mouse gut microbiome, organized according to differing methylation profiles across 38 sequence motifs in the sample. Labeled bins denote genome-scale assemblies with distinct methylation profiles (**Table 1**). (**f**) Coverage values for contigs (>100 kp to exclude small MGEs) in each of the nine bins identified by methylation binning.

landscape using t-SNE (**Fig. 2e**). Contigs were annotated using Kraken[10] (**Supplementary Table 5**).

We identified nine distinct contig bins using 38 methylation features in the murine gut microbiota sample. Seven bins assigned to the order *Bacteroidales* shared high average nucleotide identity (ANI) with each other (81–91% ANI), but at values suggesting inter- rather than intraspecies relationships[32] (**Supplementary Table 6** and **Supplementary Methods**). In eight of nine bins, alignment of reads

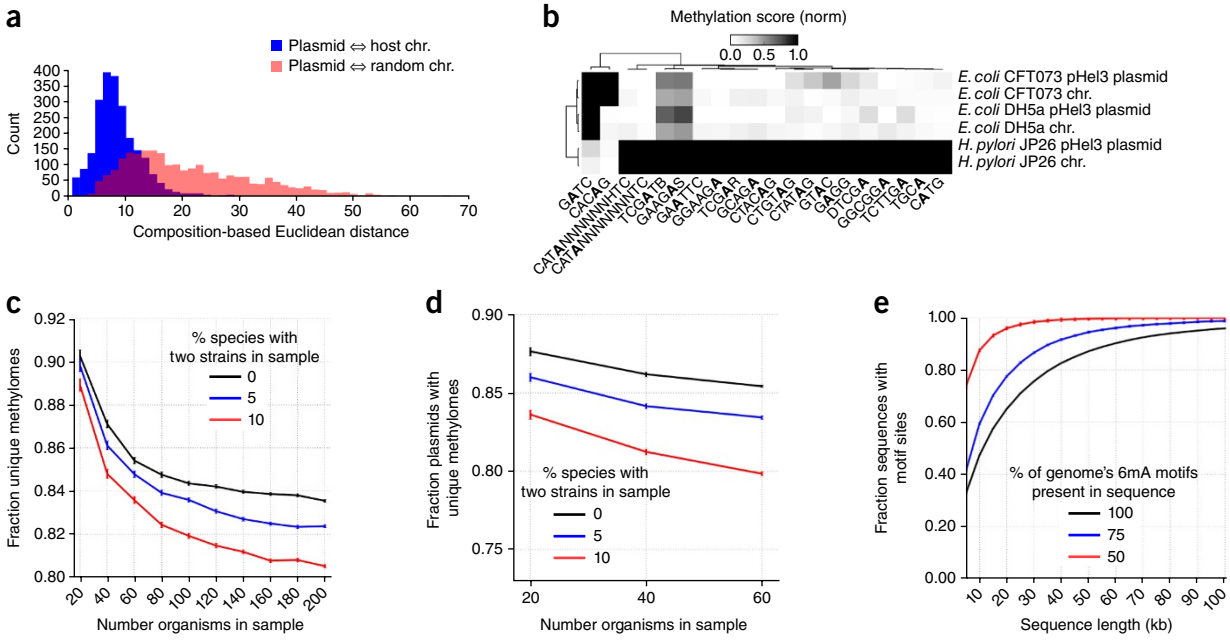

Figure 3 Methylation profiles can link plasmids to the chromosomal DNA of their host species. (**a**) Histogram of sequence-based Euclidean distance between 5-mer frequency vectors of plasmid and chromosome sequences, showing the distance between plasmids and their host chromosome (blue; based on 2,278 bacterial plasmids and their known hosts), as well as the distance between plasmid and randomly sampled chromosomes from other species (red). (**b**) Heatmap showing methylation profiles for the pHel3 plasmid and its three hosts: *E. coli* CFT073, *E. coli* DH5α, and *H. pylori* JP26. The methylation profile of pHel3 across 20 motifs matches the host from which it was isolated. (**c**) Simulation analysis (1,000 iterations) using 878 SMRT sequenced bacterial genomes in the REBASE database showing expected number of genomes with a unique 6mA methylome as a function of community size and presence of multi-strain species in the community. (**d**) Simulation analysis (1,000 iterations) using 155 SMRT sequenced genomes with known plasmids in the REBASE database showing expected number of genomes with a unique 6mA methylome as a function of community size and presence of multi-strain species in the community. (**e**) Simulation analysis (500 iterations) using 878 SMRT sequenced genomes in the REBASE database showing the expected sequence lengths required to capture at least one instance of the methylation motifs in a genome. As expected, capturing at least one instance of some, but not all, of the methylation motifs reduces the required sequence length.

to the binned contigs revealed uniform coverage values within each bin (**Supplementary Table 6**), suggesting that the bins corresponded to individual genomes (**Fig. 2f**). The split coverage values in bin7 suggest the presence of two genomes. CheckM[33], a bin validation tool that uses single-copy gene counts to assess genome completeness and contamination, found >97% completeness in eight of the nine bins. Bin7 had substantial contamination, in accordance with the observed split coverage (**Table 1**). We validated the eight highly complete genome bins by identifying high-quality sequence matches with several publicly available mouse gut microbial references[34–37] (**Supplementary Methods**).

We next explored whether coverage and composition features could resolve the same nine bins obtained from the mouse gut microbiota. We applied a variety of strategies for binning with these more standard features, including visualizing the contigs in a scatter plot of coverage versus GC-content (**Supplementary Fig. 3a**) and visualizing the contigs in a scatter plot of sample coverage versus coverage from a related sample (**Supplementary Fig. 3b**). Although several genomes were binned using these approaches, other genomes, including multiple genomes annotated as belonging to the order *Bacteroidales* (**Fig. 2e**), were not clearly resolved, showing that incorporation of methylation profiles can improve binning. For example, higher-complexity samples could benefit from methylation profiles as a means of refining differential coverage bins, analogous to the strategy described by Albertsen *et al.*[16]. An additional analysis of infant gut microbiome sequencing (**Supplementary Table 2**) demonstrated how methylation profiles can complement sequence composition features to resolve contigs from two mixed strains of *Bacteroides dorei* (**Supplementary Methods** and **Supplementary Fig. 4a–c**).

In addition to using contig-level methylation profiles as features for binning, methylation scores can also be used to detect methylated motifs in bins called by coverage- or composition-based binning tools[18,38,39]. After using CONCOCT[18] to bin assembled contigs in our adult mouse gut microbiome sample (**Supplementary Methods** and **Supplementary Fig. 5**), we combined methylation profiles of contigs in each CONCOCT bin. By pooling the IPD values across all contigs in each bin, we identified eight additional bin-level motifs that were not detected on individual contigs (**Supplementary Table 7**). This integrative approach for motif discovery in metagenomic samples is most helpful when short, poorly assembled contigs can be successfully binned using composition and coverage, but are too short for standard contig-level methylation motif discovery.

Our results confirm that methylation profiles can be used to resolve genomes (**Fig. 2e**) that cannot be completely resolved by composition and coverage features (**Supplementary Fig. 3a,b**). However, composition and coverage features are effective at resolving other population structures missed by methylation profiles, such as bins containing genomes from the orders *Lactobacillales* and *Burkholderiales* (**Supplementary Table 7**). Complete resolution of the full genomic architecture of more complex communities will likely require the integration of all of these complementary binning features.

Linking mobile genetic elements and host chromosomes

Plasmids can encode antibiotic resistance genes, virulence factors, or metabolic pathways and it is imperative to understand their contribution to microbiome functions[40,41]. These small (typically 1–200 kb), circular, and mobile DNA elements can transfer between

Table 1 Genomes binned from adult mouse gut microbiome using DNA methylation profiles

	Binning statistics				Annotation	Bin validation		Methylation summary		
Bin	No. contigs	Total bases (bp)	Largest contig (bp)	Contig N50 (bp)	Taxonomic order (% binned bases with specified annotation)	Completeness (%)	Contamination (%)	Significant motifs	Mean contig methylation score	Mapped MGEs
1	14	4,027,504	1,128,400	1,089,244	*Bacteroidales* (97.5)	98.68	2.26	ACCG**A**G	1.85	12.7 kb plasmid,
								CC**A**SNNNNNNATGT	2.01	19.1 kb conjugative transposon
2	9	3,496,584	2,164130	2,164,130	*Bacteroidales* (97.1)	77.48	2.01	CTGC**A**G	2.43	None found
3	7	3,853,295	2,087,314	2,087,314	*Bacteroidales* (98.0)	99.43	1.13	TC**A**GNNNNNCCTC	1.62	None found
								CC**A**GNNNNNNVTGG	2.22	
								CC**A**GNNNNNNRTGG	2.50	
4	5	2,759,439	2,712,836	2,712,836	*Verrucomicrobiales* (98.3)	97.96	0.68	G**A**TTNNNNNCAGT	3.11	None found
								G**A**TTNNNNNNAGT	2.93	
5	10	3,378,404	1,873,721	1,873,721	*Bacteroidales* (100.0)	97.55	1.76	AGC**A**NNNNNNRTC	1.98	None found
								G**A**CNNNNNNTGCT	2.27	
6	16	4,441,324	1,159,367	764,722	*Bacteroidales* (100.0)	98.36	1.26	ATGC**A**T	1.76	
								CC**A**NNNNNTCG	1.93	None found
								AAC**A**GC	2.80	
7	22	620,7805	2,165,375	1,643,203	*Bacteroidales* (98.5)	98.24	21.52	GGC**A**GC	2.22	24.7kb plasmid,
								GTG**A**TG	2.00	14.7kb plasmid,
										23.2kb conjugative transposon
8	14	3,913,657	2,565,370	2,565,370	*Bacteroidales* (98.2)	97.22	2.77	AG**A**TGA	2.21	
								AG**A**TG	1.94	
								G**A**TGGY	1.94	14.3kb plasmid,
								AG**A**TGT	1.72	15.8kb plasmid,
								KAG**A**TG	2.08	21.1kb conjugative transposon
								TAG**A**TG	1.96	
								TG**A**TGG	1.71	
								G**A**TGG	1.81	
9	1	2,021,078	2,021,078	2,021,078	*Clostridiales* (100.0)	99.19	0.00	CGA**A**G	2.46	
								GA**A**GNNNNNACGT	2.18	
								TGM**A**GG	2.48	None found
								CG**A**GNNNNNNCCTT	1.69	
								ACC**A**TC	2.20	

Annotation of binned contigs was conducted using Kraken[10]. The taxonomic order with the largest percentage of binned bases assigned to that order is reported for each bin. Assembly validation was done using CheckM[33] and reflected the presence or absence of a set of single-copy marker genes. Significant motifs are those with a mean methylation score across binned contigs greater than 1.6 (28/38 motifs detected from contigs in this assembly are significant in these bins). Mapped MGEs are those with matching methylation profiles to the specified methylation bin.

host bacteria by conjugation or natural transformation, making them important mediators of horizontal gene transfer. Plasmid replication can be independent of chromosomal replication, meaning that the sequence coverages of a plasmid and its host chromosome typically differ. Furthermore, by comparing the 5-mer frequency statistics of plasmids and the chromosomes of their bacterial hosts, we found that the sequence composition profiles can also differ (**Fig. 3a**), making such features unreliable for linking a plasmid to its host in metagenomic samples.

Plasmid and chromosomal DNA of the bacterial host are methylated by the same set of MTases[42], resulting in matching methylation profiles. To confirm this, we transformed *Escherichia coli* CFT073 and *Helicobacter pylori* JP26 with the 5.5-kb plasmid pHel3 (GenBank MG214727) from *E. coli* DH5α, then sequenced both plasmid and genomic DNA prepared from each of the three bacterial hosts. In each case, SMRT sequencing (**Supplementary Table 2**) showed that pHel3 was marked by the methylation profile of its host strain (**Fig. 3b**).

In order to determine whether methylation profiles can be used to map plasmids to their hosts in metagenomics data sets, we first simulated communities of 20–200 members by sampling methylomes of SMRT sequenced bacterial chromosomes and plasmids from the REBASE database[43]. Unambiguous plasmid mapping in a microbiome sample requires that the plasmid and host chromosome have unique methylomes. As expected, the number of unique methylomes (expressed as a fraction of total community members) decreased in larger synthetic communities (**Fig. 3c**) and this decrease was more pronounced when multiple strains of a species were present in a community. Similar trends were observed when including only the methylomes of organisms that have at least one known plasmid (**Fig. 3d**). Large plasmids are more likely to contain instances of the motifs that are required to match plasmid and chromosome methylation profiles. By extracting nucleotide substrings of various lengths from random positions in known reference sequences in REBASE, we found that, on average, 90% of 35-kb sequences contained at least 75% of the 6mA motifs found in the host genome, and that 90% of 60-kb sequences captured 100% of the 6mA motifs (**Fig. 3e**). This means that larger, rather than smaller, plasmids are more likely to be correctly mapped to their host by methylation-assisted binning.

Furthermore, a notable entry in the REBASE database is the virulent 234-12 strain of *Klebsiella pneumoniae* and its 362-kb plasmid pKpn23412-362, which encodes 13 antibiotic resistance genes. By comparing the methylome of *K. pneumoniae* str. 234-12 with nine other similar species and 24 other *K. pneumoniae* strains, we found

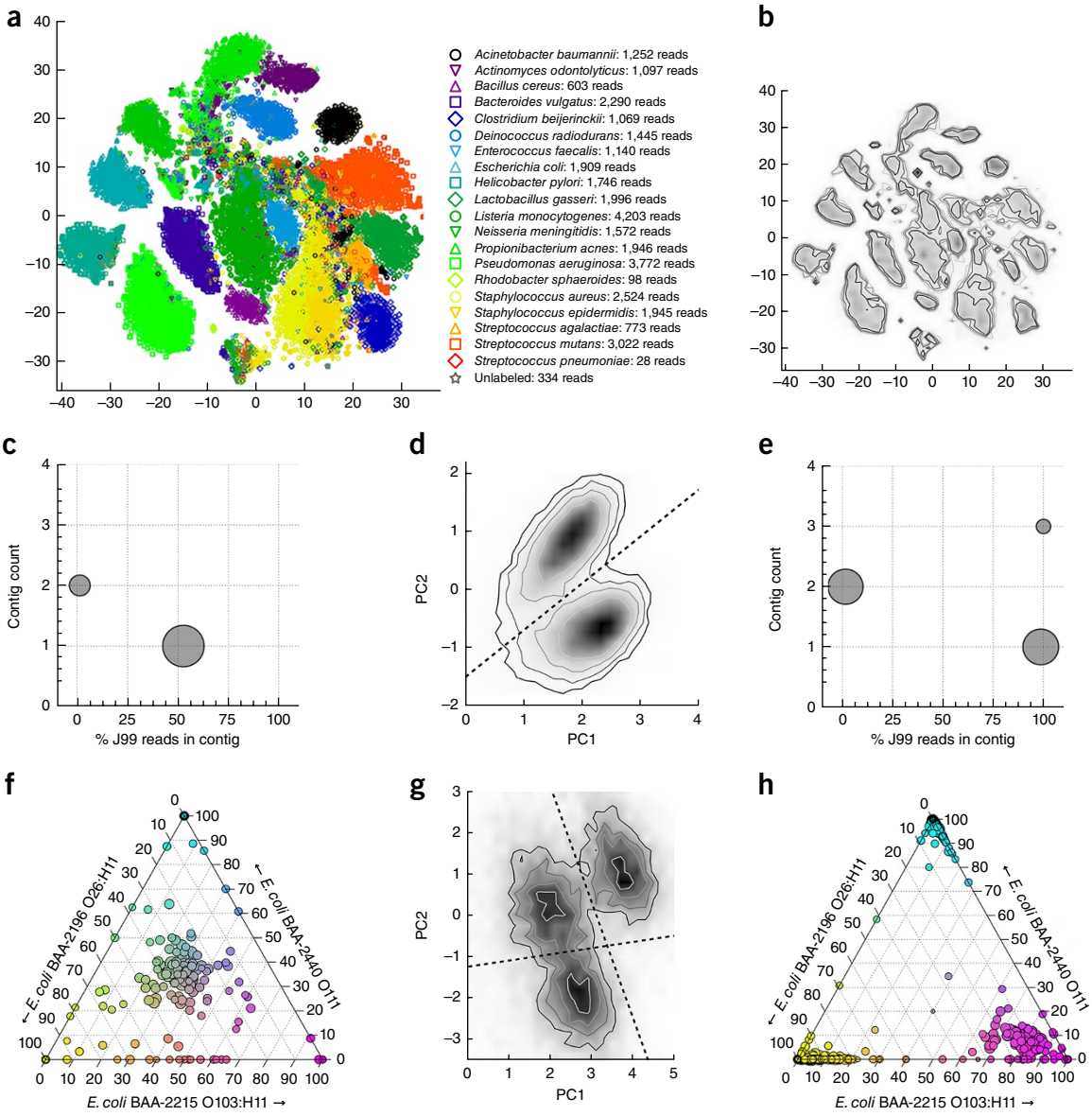

Figure 4 Binning SMRT reads using composition and DNA methylation profiles. (**a**) 5-mer frequency-based binning of assembled contigs and raw reads (length >15 kb) from the Human Microbiome Project mock community, where only the unassembled reads are labeled. Reads from the low-abundance species *R. sphaeroides* form a distinct cluster near the coordinates (–8,–22). (**b**) The 2D histogram of contigs and unassembled reads, corresponding to **a**; this 2D histogram lacks labeling but nevertheless includes many highly species-specific subpopulations. (**c**) Combined assembly of a synthetic mixture of reads from *H. pylori* strains J99 and 26995 results in one small contig containing mostly reads from strain 26695 and one large, highly chimeric contig. (**d**) Read-level methylation profiles for unassembled reads from the synthetic mixture are separated by principal component analysis (PCA) into discrete, strain-specific clusters. (**e**) Separate assembly of reads that were segregated using methylation profiles results in large, highly strain-specific contigs. (**f**) Combined assembly of a synthetic mixture of reads from *E. coli* strains BAA-2196 O26:H11, BAA-2215 O103:H11, and BAA-2440 O111 results in many chimeric contigs that contain reads from all three strains. (**g**) Reads from the synthetic mixture were aligned to the *E. coli* K12 MG1655 reference in order to correct sequencing errors and the read-level methylation profiles were separated by PCA into strain-specific clusters. (**h**) Separate assembly of reads segregated by methylation profiles as demonstrated in **g** results in a dramatic reduction of chimerism in the assembled reads.

the methylation profile of *K. pneumoniae* str. 234-12 to be unique among the examined genomes (**Supplementary Fig. 6a**,**b**), making it possible to identify it as the host of pKpn23412-362 among similar strains when they co-exist in a microbiome sample.

We next identified six putative plasmid sequences of 4–44 kb in the contigs assembled from our mock community of eight bacterial species (**Supplementary Table 3**). By comparing methylation profiles of these sequences with those of chromosomal contigs, we were able to correctly assign these plasmids to their hosts in four of the six cases, including the only previously characterized plasmid in the group, *B. thetaiotaomicron* plasmid p5482 (GenBank accession AY171301.1). The remaining two putative plasmids were too short (<10 kbp) to contain sufficient motif sites for conclusive mapping, consistent with the REBASE simulation analysis (**Fig. 3e**), showing that only 40% of 10-kb sequences are expected to contain instances of all motifs methylated by the host.

Finally, we identified 19 MGE contigs in the adult mouse gut microbiome assembly (**Supplementary Table 3**) between 7 and 132 kb, of which 10 were fully circularized and 9 were conjugative transposons (encoding at least five genes annotated as conjugative transposon-related) (Online Methods). Conjugative transposons have an important role in horizontal gene transfer and the spread of antibiotic resistance genes in *Bacteroidales*, having been shown to transfer between multiple *Bacteroidales* species in the human gut[44]. Thirteen of these MGEs were discovered by re-assembling the reads mapping to contigs in each bin using HGAP3 (ref. 45; **Supplementary Methods**). Of the 19 identified MGE contigs, 8 had methylation profiles that could be conclusively matched to the previously identified methylation bins containing genomes from the order *Bacteroidales* (**Table 1**). These eight linked MGEs included five putative circular plasmids of <50 kb containing an origin of replication, as well as three conjugative transposons.

Binning unassembled SMRT reads

Although it has been shown that visualizing sequence composition features of assembled contigs using t-SNE can be effective for binning contigs[13], we found that sequence composition features are also well suited for segregating long, unassembled SMRT reads. After combining sequences from both the contigs and unassembled reads previously sequenced from a 20-member mock community (**Supplementary Table 2**), we visualized and labeled the reads in the t-SNE map of 5-mer frequency features for all sequences (**Supplementary Methods**). Read clusters in the map were highly species-specific and resilient to random sequencing errors. For instance, despite having very low sequence coverage that precluded assembly (**Supplementary Fig. 7**), unassembled reads from *Rhodobacter sphaeroides* formed a distinct cluster when read-level 5-mer frequency profiles were visualized using t-SNE (**Fig. 4a**,**b**). Unsurprisingly, species segregation improved with increasing read lengths (**Supplementary Fig. 8a**,**b**).

In addition to sequence composition features, unassembled SMRT reads also contain methylation features that could help address some of the challenges posed by multi-strain species in metagenomic samples. To explore whether methylation binning could be extended to the level of unassembled reads, we constructed two synthetic mixtures of reads from (1) two strains of *H. pylori* and (2) three strains of *E. coli* (**Supplementary Table 2**). Despite the high sequence similarity of the strains in each mixture (93.65% ANI for two *H. pylori* strains and >99% ANI for three *E. coli* strains) (**Supplementary Methods**), the different MTases they encode resulted in distinct sets of methylated motifs. Assembly of the *H. pylori* mixture containing reads from strains J99 and 26695 resulted in one small contig from strain 26695 and another large chimeric contig (**Fig. 4c**). We used read-level methylation profiles across four 6mA motifs present at high density in the genome: G$\underline{A}$TC, G$\underline{A}$GG, TGC$\underline{A}$, and C$\underline{A}$TG[46] (**Supplementary Table 8**). Principal component analysis (PCA) of the methylation profiles revealed a bimodal Gaussian distribution of reads (**Fig. 4d**) that was more amenable to separation than the map generated by t-SNE (**Supplementary Fig. 9**). Separate assembly of each bin resulted in contigs with improved contiguity, including chromosome-scale contigs for both strains, and minimal chimerism (**Fig. 4e**). Finally, we applied a slightly modified approach to the mixture of *E. coli* strains, where an additional error-correction step (Online Methods) removed much of the sequencing and IPD errors that occur in longer motifs in raw reads. Bulk assembly of the mixture of error-corrected reads resulted in many chimeric contigs and very few contigs that were specific to a strain (**Fig. 4f**), but binning the reads by methylation profiles across four differentiating motifs (**Fig. 4g**; **Supplementary Table 9**) before assembly resulted in a substantial increase in the purity of contigs (**Fig. 4h**).

DISCUSSION

We report that microbial DNA methylation can be exploited as endogenous epigenetic barcodes to complement coverage and composition features to improve metagenomic binning. Notably, methylation motifs can link mobile genetic elements to their host genomes in microbial samples and improve strain-level resolution of metagenomes.

We used our approach to bin nine genomes, several of which were previously poorly characterized, in an adult mouse gut microbiome. We also linked eight assembled MGEs to these genomes based on matching methylation profiles. Furthermore, we show that unassembled reads in metagenomics samples can be binned using methylation profiles. This holds promise for simplifying multi-strain assembly, although it typically requires read lengths of at least 10–15 kb, depending on the methylome complexity. We expect our approach to be well suited for analyzing low-to-medium complexity communities, while the value added by methylation binning in higher complexity samples will largely be a function of sequencing depth, assembly quality, and methylome uniqueness of a particular microbiome sample.

Multiple factors should be taken into account before attempting to bin genomes using methylation in high-complexity samples, such as adult human gut or environmental samples. The most important factor is the degree of methylome uniqueness, that is, the fraction of methylomes with unique combinations of methylated motifs in a sample. As the number of genomes in a microbiome sample increases, the expected level of methylome uniqueness typically declines (**Fig. 3c**,**d**) and, consequently, the discriminative resolution of methylation binning decreases. In high-complexity samples, methylation profiles are therefore better suited to refine bins called by coverage and/or composition features, similar to the binning refinement strategy described by Albertsen *et al.*[16]. High-complexity samples may contain multiple co-existing strains, which present challenges for assembly tools and therefore often lack high-quality contigs for methylation binning, although read-level methylation profiles can potentially improve multi-strain assemblies.

The presence of low-abundance organisms in a community presents additional challenges for methylation binning, as it is difficult to detect methylated motifs from the small contigs that are typically assembled from such genomes. However, this can be complemented by the use of binning assignments from coverage- and composition-based binning tools, such as CONCOCT[18]. Phasing IPD information from all contigs in a bin makes it possible to detect additional methylated motifs. If organism abundance is too low for genomic assembly, the only solution is additional sequencing depth. Despite the relatively higher cost of SMRT sequencing, we anticipate that technological advances will continue to bring down the cost per base as read lengths and total yields increase. Improved metagenomic assembly algorithms specially designed for long reads should result in higher quality assemblies and larger contigs that are more amenable to methylation analysis. Motif discovery on unassembled reads remains challenging, but longer reads could make this more feasible in the future.

Although our study focused mainly on 6mA motifs, improved detection of other methylation events, like 5-methylcytosine (5mC) and N4-methylcytosine (4mC), will expand the set of motifs that can be included in methylation profiles. Such improvements, as well as decreases in the input DNA requirement, promise to broaden the metagenomic application space for third-generation technologies.

SMRT sequencing libraries with long insert sizes improve contiguity in metagenomic assemblies, but the size selection procedure may filter out certain MGEs like small plasmids and phages. Integrating additional sequencing from rolling circle amplification libraries might highlight small, circular sequences that are lost during size-selection steps or do not fully circularize in the metagenomic assembly.

Beyond metagenomic binning, methylation profiles could be used for monitoring the transmission of plasmids and bacteriophages between hosts across multiple time points or conditions, such as antibiotic treatment[6]. Additionally, *de novo* detection of methylation motifs in microbial communities may help to reveal mechanisms of epigenetic regulation in uncultured bacteria, and identify novel MTases and restriction enzymes for use in research.

Although our study focused on SMRT sequencing, our framework applies to other third-generation sequencing technologies capable of detecting bacterial DNA methylation, such as Oxford Nanopore[47] or possibly Genia[48]. The Minion instrument from Oxford Nanopore is an intriguing option, although efforts to develop robust methylation detection methods are ongoing[49]. Synthetic long-read technologies can be useful for interrogating complex communities, but lack methylation signatures and are subject to coverage biases that impede genomic assembly (**Supplementary Methods**, **Supplementary Figs. 10–12** and **Supplementary Table 10**). By integrating second- and third-generation sequencing with complementary analyses like Hi-C intrachromosomal maps[19–21] or single-cell techniques[50], we expect researchers to gain an increasingly complete understanding of the genomic and epigenomic landscape of microbial communities.

METHODS

Methods, including statements of data availability and any associated accession codes and references, are available in the online version of the paper.

Note: Any Supplementary Information and Source Data files are available in the online version of the paper.

ACKNOWLEDGMENTS

We thank M. Lewis for her assistance in DNA extraction and A. Bashir for his guidance in computational matters. We also thank those who contributed to the generation of the publically available SMRT sequencing data for the 20-member Mock Community B. The work is funded by R01 GM114472 (G.F.) from the National Institutes of Health and Icahn Institute for Genomics and Multiscale Biology. G.F. is a Nash Family Research Scholar. This work was also supported in part through the computational resources and staff expertise provided by the Department of Scientific Computing at the Icahn School of Medicine at Mount Sinai.

AUTHOR CONTRIBUTIONS

J.B. and G.F. designed the methods. J.B. developed the software package for all the proposed computational analyses. J.B., E.W.T., J.J.F. R.S., E.E.S. and G.F. contributed to experimental design. I.M., X.-S.Z., A.D.-R., R.C., E.W.T. and J.J.F. conducted the experiments. G.D. and R.S. designed and conducted sequencing. J.B., S.Z., E.W.T., J.J.F., R.S., E.E.S. and G.F. analyzed the data. J.B. and G.F. wrote the manuscript with inputs and comments from all co-authors. G.F. conceived and supervised the project.

COMPETING FINANCIAL INTERESTS

The authors declare competing financial interests: details are available in the online version of the paper.

1. Cho, I. & Blaser, M.J. The human microbiome: at the interface of health and disease. *Nat. Rev. Genet.* **13**, 260–270 (2012).
2. Human Microbiome Project Consortium. Structure, function and diversity of the healthy human microbiome. *Nature* **486**, 207–214 (2012).
3. Janda, J.M. & Abbott, S.L. 16S rRNA gene sequencing for bacterial identification in the diagnostic laboratory: pluses, perils, and pitfalls. *J. Clin. Microbiol.* **45**, 2761–2764 (2007).
4. Qin, J. *et al.* A human gut microbial gene catalogue established by metagenomic sequencing. *Nature* **464**, 59–65 (2010).
5. Tyson, G.W. *et al.* Community structure and metabolism through reconstruction of microbial genomes from the environment. *Nature* **428**, 37–43 (2004).
6. Modi, S.R., Lee, H.H., Spina, C.S. & Collins, J.J. Antibiotic treatment expands the resistance reservoir and ecological network of the phage metagenome. *Nature* **499**, 219–222 (2013).
7. Luo, C. *et al.* ConStrains identifies microbial strains in metagenomic datasets. *Nat. Biotechnol.* **33**, 1045–1052 (2015).
8. Kuleshov, V. *et al.* Synthetic long-read sequencing reveals intraspecies diversity in the human microbiome. *Nat. Biotechnol.* **34**, 64–69 (2016).
9. Brady, A. & Salzberg, S.L. Phymm and PhymmBL: metagenomic phylogenetic classification with interpolated Markov models. *Nat. Methods* **6**, 673–676 (2009).
10. Wood, D.E. & Salzberg, S.L. Kraken: ultrafast metagenomic sequence classification using exact alignments. *Genome Biol.* **15**, R46 (2014).
11. Saeed, I., Tang, S.L. & Halgamuge, S.K. Unsupervised discovery of microbial population structure within metagenomes using nucleotide base composition. *Nucleic Acids Res.* **40**, e34 (2012).
12. Iverson, V. *et al.* Untangling genomes from metagenomes: revealing an uncultured class of marine Euryarchaeota. *Science* **335**, 587–590 (2012).
13. Laczny, C.C., Pinel, N., Vlassis, N. & Wilmes, P. Alignment-free visualization of metagenomic data by nonlinear dimension reduction. *Sci. Rep.* **4**, 4516 (2014).
14. Laczny, C.C. *et al.* VizBin - an application for reference-independent visualization and human-augmented binning of metagenomic data. *Microbiome* **3**, 1–7 (2015).
15. 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).
16. Albertsen, M. *et al.* Genome sequences of rare, uncultured bacteria obtained by differential coverage binning of multiple metagenomes. *Nat. Biotechnol.* **31**, 533–538 (2013).
17. Nielsen, H.B. *et al.* Identification and assembly of genomes and genetic elements in complex metagenomic samples without using reference genomes. *Nat. Biotechnol.* **32**, 822–828 (2014).
18. Alneberg, J. *et al.* Binning metagenomic contigs by coverage and composition. *Nat. Methods* **11**, 1144–1146 (2014).
19. Marbouty, M. *et al.* Metagenomic chromosome conformation capture (meta3C) unveils the diversity of chromosome organization in microorganisms. *eLife* **3**, e03318 (2014).
20. Burton, J.N., Liachko, I., Dunham, M.J. & Shendure, J. Species-level deconvolution of metagenome assemblies with Hi-C-based contact probability maps. *G3 (Bethesda)* **4**, 1339–1346 (2014).
21. Marbouty, M., Baudry, L., Cournac, A. & Koszul, R. Scaffolding bacterial genomes and probing host-virus interactions in gut microbiome by proximity ligation (chromosome capture) assay. *Sci. Adv.* **3**, e1602105 (2017).
22. Flusberg, B.A. *et al.* Direct detection of DNA methylation during single-molecule, real-time sequencing. *Nat. Methods* **7**, 461–465 (2010).
23. Eid, J. *et al.* Real-time DNA sequencing from single polymerase molecules. *Science* **323**, 133–138 (2009).
24. Casadesús, J. & Low, D. Epigenetic gene regulation in the bacterial world. *Microbiol. Mol. Biol. Rev.* **70**, 830–856 (2006).
25. Blow, M.J. *et al.* The epigenomic landscape of prokaryotes. *PLoS Genet.* **12**, e1005854 (2016).
26. Kobayashi, I., Nobusato, A., Kobayashi-Takahashi, N. & Uchiyama, I. Shaping the genome--restriction-modification systems as mobile genetic elements. *Curr. Opin. Genet. Dev.* **9**, 649–656 (1999).
27. Conlan, S. *et al.* Single-molecule sequencing to track plasmid diversity of hospital-associated carbapenemase-producing Enterobacteriaceae. *Sci. Transl. Med.* **6**, 254ra126 (2014).
28. Schadt, E.E. *et al.* Modeling kinetic rate variation in third generation DNA sequencing data to detect putative modifications to DNA bases. *Genome Res.* **23**, 129–141 (2013).
29. Beaulaurier, J. *et al.* Single molecule-level detection and long read-based phasing of epigenetic variations in bacterial methylomes. *Nat. Commun.* **6**, 7438 (2015).
30. van der Maaten, L. & Hinton, G. Visualizing data using t-SNE. *J. Mach. Learn. Res.* **9**, 2579–2605 (2008).
31. van der Maaten, L. Accelerating t-sne using tree-based algorithms. *J. Mach. Learn. Res.* **15**, 3221–3245 (2014).
32. Kim, M., Oh, H.S., Park, S.C. & Chun, J. Towards a taxonomic coherence between average nucleotide identity and 16S rRNA gene sequence similarity for species demarcation of prokaryotes. *Int. J. Syst. Evol. Microbiol.* **64**, 346–351 (2014).
33. Parks, D.H., Imelfort, M., Skennerton, C.T., Hugenholtz, P. & Tyson, G.W. CheckM: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. *Genome Res.* **25**, 1043–1055 (2015).
34. Uchimura, Y. *et al.* Complete genome sequences of 12 species of Stable Defined Moderately Diverse Mouse Microbiota 2. *Genome Announc.* **4**, e00951–16 (2016).
35. Ormerod, K.L. *et al.* Genomic characterization of the uncultured Bacteroidales family S24-7 inhabiting the guts of homeothermic animals. *Microbiome* **4**, 36 (2016).
36. Xiao, L. *et al.* A catalog of the mouse gut metagenome. *Nat. Biotechnol.* **33**, 1103–1108 (2015).
37. Wannemuehler, M.J., Overstreet, A.M., Ward, D.V. & Phillips, G.J. Draft genome sequences of the altered schaedler flora, a defined bacterial community from gnotobiotic mice. *Genome Announc.* **2**, e00287–14 (2014).

38. Imelfort, M. *et al.* GroopM: an automated tool for the recovery of population genomes from related metagenomes. *PeerJ* **2**, e603 (2014).
39. Kang, D.D., Froula, J., Egan, R. & Wang, Z. MetaBAT, an efficient tool for accurately reconstructing single genomes from complex microbial communities. *PeerJ* **3**, e1165 (2015).
40. Slater, F.R., Bailey, M.J., Tett, A.J. & Turner, S.L. Progress towards understanding the fate of plasmids in bacterial communities. *FEMS Microbiol. Ecol.* **66**, 3–13 (2008).
41. Thomas, C.M. & Nielsen, K.M. Mechanisms of, and barriers to, horizontal gene transfer between bacteria. *Nat. Rev. Microbiol.* **3**, 711–721 (2005).
42. Fang, G. *et al.* Genome-wide mapping of methylated adenine residues in pathogenic *Escherichia coli* using single-molecule real-time sequencing. *Nat. Biotechnol.* **30**, 1232–1239 (2012).
43. Roberts, R.J., Vincze, T., Posfai, J. & Macelis, D. REBASE—a database for DNA restriction and modification: enzymes, genes and genomes. *Nucleic Acids Res.* **43**, D298–D299 (2015).
44. Coyne, M.J., Zitomersky, N.L., McGuire, A.M., Earl, A.M. & Comstock, L.E. Evidence of extensive DNA transfer between *bacteroidales* species within the human gut. *MBio* **5**, e01305–e01314 (2014).
45. Chin, C.-S. *et al.* Nonhybrid, finished microbial genome assemblies from long-read SMRT sequencing data. *Nat. Methods* **10**, 563–569 (2013).
46. Krebes, J. *et al.* The complex methylome of the human gastric pathogen Helicobacter pylori. *Nucleic Acids Res.* **42**, 2415–2432 (2014).
47. Clarke, J. *et al.* Continuous base identification for single-molecule nanopore DNA sequencing. *Nat. Nanotechnol.* **4**, 265–270 (2009).
48. Fuller, C.W. *et al.* Real-time single-molecule electronic DNA sequencing by synthesis using polymer-tagged nucleotides on a nanopore array. *Proc. Natl. Acad. Sci. USA* **113**, 5233–5238 (2016).
49. Rand, A.C. *et al.* Mapping DNA methylation with high-throughput nanopore sequencing. *Nat. Methods* **14**, 411–413 (2017).
50. Lan, F., Demaree, B., Ahmed, N. & Abate, A.R. Single-cell genome sequencing at ultra-high-throughput with microfluidic droplet barcoding. *Nat. Biotechnol.* **35**, 640–646 (2017).
51. Rousseeuw, P.J. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *J. Comput. Appl. Math.* **20**, 53–65 (1987).
52. Caporaso, J.G. *et al.* QIIME allows analysis of high-throughput community sequencing data. *Nat. Methods* **7**, 335–336 (2010).

ONLINE METHODS

Code availability. The software supporting all proposed methods (**Supplementary Code**) is implemented in Python and is available with full documentation at http://www.github.com/fanglab/mbin.

Culture conditions for bacteria from eight-species mixture and purification. *Bacteroides caccae* American Type Culture Collection (ATCC) 43185, *Bacteroides ovatus* ATCC 8483, *Bacteroides thetaiotaomicron* VPI-5482, *Bacteroides vulgatus* ATCC 8492, *Collinsella aerofaciens* ATCC 25986, *Clostridium bolteae* ATCC BAA-613, and *Ruminococcus gnavus* ATCC 29149 were grown individually in 10 ml of supplemented Brain-heart infusion broth[53] in an anaerobic chamber from Coy Laboratory Products. *Escherichia coli* MG1655 was grown aerobically in 5 ml of LB broth. Construction of the 10-kb DNA libraries for SMRT sequencing was performed according to the manufacturer's instructions.

Mouse gut microbiome DNA purification and library preparation. A male 6-week-old NOD/shiltj mouse (no. 001976, Jackson Labs) was housed in a Specific Pathogen Free (SPF) room at New York University Langone Medical Center (NYUMC). At week 12 of life, the mouse was placed into a clean plastic container in a fume hood and its fresh fecal pellets were collected in sterilized microcentrifuge tubes and frozen at −80 °C. Fecal DNA was extracted using PowerSoil DNA isolation kit (MoBio Labs, Carlsbad, CA). A **Life Sciences Reporting Summary** is available for this study. 10-kb library preparation for SMRT sequencing was performed according to the manufacturer's instructions. The bacterial 16S rRNA gene V4 regions were amplified and libraries constructed as previously described by Livanos *et al.*[54].

Three species transformation by pHel3 plasmid. The *E. coli–H. pylori* shuttle plasmid pHel3 (ref. 55) was electroporated from *E. coli* strain DH5α to strain CFT073 using MicroPulser following procedures recommended by the manufacturer (Bio-Rad Lab., Hercules, CA). The same plasmid was also introduced from *E. coli* strain DH5α into *H. pylori* strain JP26 by natural transformation as previously described[56]. *E. coli* DH5α carrying pHel3 and CFT073 carrying pHel3 were grown in Luria–Bertani (LB) medium with kanamycin (Km; 50 μg/ml) at 37 °C for 24 h. *H. pylori* JP26 carrying pHel3 were grown in Brucella broth (BB) medium supplemented with 10% newborn calf serum (NBCS) and Km (10 μg/ml) at 37 °C in microaerophilic conditions for 48 h. Bacterial cell pellets of *E. coli* or *H. pylori* cultures were collected by centrifugation, genomic DNA of each culture was purified using Wizard Genomic DNA Purification Kit (Promega, Madison, WI), and plasmid DNA of each culture was purified using QIAprep Spin Miniprep Kit (QIAgen, Valencia, CA). 2-kb library preparation for SMRT sequencing genomic and plasmid DNA for each culture was performed according to the manufacturer's instructions.

Three *E. coli* strains for synthetic mixture. Genomic DNA for the three strains of *E. coli*, BAA-2196, BAA-2215, and BAA-2440, were purchased from ATCC and construction of the 10-kb DNA libraries for SMRT sequencing was performed according to the manufacturer's instructions.

Infant gut microbiome samples. DNA was isolated from stool samples taken from two Finnish children. The donor of Sample A (containing *B. dorei* str. 105) was 13.5 months of age, while Sample B (containing *B. dorei* str. 439) was obtained from a child at 3.3 months of age. Full details on sample isolation and DNA extraction are provided by Leonard *et al.*[57]. A summary of the SMRT sequencing statistics can be found in **Supplementary Table 2**.

Sequencing. For SMRT sequencing, primer was annealed to size-selected SMRTbells with the full-length libraries (80 °C for 2 min and 30 s followed by decreasing the temperature by 0.1 °C to 25 °C). The polymerase-template complex was then bound to the P6 enzyme using a ratio of 10:1 polymerase to SMRTbell at 0.5 nM for 4 h at 30 °C and then held at 4 °C until ready for magbead loading, before sequencing. The magnetic bead-loading step was conducted at 4 °C for 60 min per manufacturer's guidelines. The magbead-loaded, polymerase-bound, SMRTbell libraries were placed onto the RSII machine at a sequencing concentration of 125–175 pM and configured for a 240-min continuous sequencing run. For 16s rRNA gene amplicon sequencing, sequencing of the 16S V4 region was performed using the Illumina MiSeq platform as previously described by Livanos *et al.*[54]

Sequence composition features. All *k*-mer frequency metrics in this study used a *k*-mer size of 5. Counts of pairs of 5-mers that are reverse complements of each other were combined, resulting in a vector of 5-mer composition features (length $\boldsymbol{V} = 512$) for each sequence (contig or single-molecule read), *i*, denoted $\boldsymbol{Z}_i = (Z_{i,1}, \ldots, Z_{i,V})$. Following the procedure described by Alneberg *et al.*[18], we add a small pseudo-count to each 5-mer count to ensure all counts are non-zero, then normalize by the total number of 5-mers in the sequence and $\log_2$-transform the normalized values:

$$\boldsymbol{Z}'_i = \ln\left(\frac{Z_{i,j}+1}{V}\right)$$

The script *create_kmer_freq_vectors.py* (**Supplementary Code**) calculates *k*-mer frequency vectors for sequences in an input fasta file. Alternatively, GC-content metrics simply reflect the fraction of cytosine or guanine nucleotides in a DNA sequence.

Contig coverage features. All contig coverage features represent the read depth assessed by aligning reads to assembled contigs. Illumina reads were aligned to contigs using *bowtie2* (ref. 58) and SMRT reads were aligned to contigs during the HGAP3 (ref. 45) assembly process. For a single sample, each contig has a single coverage value. Contig coverage values from two samples are leveraged by plotting coverage values from each sample on the *x* and *y* axes. If using additional samples, coverage profiles are built for each contig, *i*, into a vector of *N* coverage features, denoted by $\boldsymbol{Y}_i = (Y_{i,1}, \ldots, Y_{i,N})$, where $\boldsymbol{N}$ is the number of samples.

Motif methylation scoring. The contig- and read-level polymerase kinetics scores are calculated using the inter-pulse duration (IPD) values provided in the SMRT sequencing reads[22]. Subread normalization, done by subtracting the subread mean of log-transformed IPD values from each log-transformed IPD value, corrects for any potential slowing of polymerase kinetics over the course of an entire read (which can consist of multiple subreads)[28,42]. Each normalized IPD (nIPD) value in the subread is calculated as follows:

$$\mathrm{nIPD} = \ln \mathrm{IPD} - \frac{1}{\mathrm{N}}\sum_{\mathrm{k}=1}^{\mathrm{N}} \ln \mathrm{IPD_k}$$

where the subread is *N* bases long and therefore contains *N* IPD values. To calculate the observed read-level methylation score ($\mathrm{R^o}$) for motif *i* on read *j*, $\mathrm{R^o_{ij}}$, we take the mean of all nIPD values from all sites of motif *i* across all subreads of read *j*:

$$\mathrm{R^o_{ij}} = \frac{1}{\sum_{\mathrm{s}=1}^{\mathrm{S}} \mathrm{M_s}} \sum_{\mathrm{s}=1}^{\mathrm{S}} \sum_{\mathrm{m}=1}^{\mathrm{M_s}} \mathrm{nIPD_{ms}}$$

where each of the *S* subreads in the read contains M_s motif sites. Longer subreads typically contain more distinct sites of a given motif and generate more reliable methylation scores.

Kinetic variation in the polymerase activity exists even in the absence of methylated bases and is highly correlated with the local nucleotide context surrounding the polymerase as it processes along the template[59]. To account for this baseline variation and remove it from the final methylation score, we subtract from our observed kinetics scores, $\mathrm{R^o_{ij}}$, a corresponding set of control kinetics scores, $\mathrm{R^c_i}$. These control kinetics scores are motif-matched and calculated similar to $\mathrm{R^o_{ij}}$ using a sampling of SMRT sequencing unaligned reads ($N = 20{,}000$) known to be free of any methylation:

$$\mathrm{R_{ij}} = \mathrm{R^o_{ij}} - \mathrm{R^c_i}$$

As no methylated motifs were detected after sequencing an isolate of *Ruminococcus gnavus*, these data served as the non-methylated control set for calculating values of $\mathrm{R^c_i}$. These non-methylated control values are used for the motif filtering procedure, but not for the final calculation of methylation

doi:10.1038/nbt.4037

profiles. Because the dimensionality reduction with t-SNE calculates a Euclidean distance between two points (i.e., two methylation profiles), the subtraction of a constant (control) vector from both methylation profiles has no effect on their pairwise distances.

Contig-level methylation scores (C) for motif i on contig j, C_{ij}, are calculated in a similar manner. The difference is that the scores take into account not just the subreads from a single read, but rather all subreads that align to the contig:

$$C_{ij}^{o} = \frac{1}{\sum_{s=1}^{S^*} M_s} \sum_{s=1}^{S^*} \sum_{m=1}^{M_s} \mathrm{nIPD}_{ms}$$

where each of the S^* subreads that align to the contig contain M_s motif sites. Similar to the read-level methylation scores, matching control kinetics scores, C_i^c, are generated using a sample of aligned reads ($N = 20{,}000$) known to be free of methylation and subtracted from the observed kinetics scores, C_{ij}^{o}, in order to remove the baseline kinetics variation stemming from local sequence context:

$$C_{ij} = C_{ij}^{o} - C_i^c$$

As with the read-level methylation scoring, non-methylated control values are used only during the motif filtering procedure but not in the final contig-level methylation scores. Much like the read-level methylation assessment, the reliability of the motif score on a contig increases with the number of motif sites on the contig. Typically, short motifs are present at higher density in the genome than longer, more complex motifs, although exceptions to this rule exist. Therefore, while even the shortest contigs in an assembly are able to return reliable methylation scores for short motifs, longer contigs are usually required to accurately assess the methylation status of more complex motifs. A default methylation score of zero is assigned if no instances of the motif occur on the read or contig.

The optional parameter–cross_cov_bins in the mBin program accepts a file containing contig assignments to bins (in the format *contig_name,bin_id*) identified from coverage- and composition-based binning tools, such as CONCOCT[18] or MetaBAT[39]. If this parameter is specified, the IPD values used to calculate each contig-level methylation score are aggregated based on binning assignment and bin-level methylation scores are calculated.

Motif filtering for methylation-based clustering. Methylated motifs are identified from the entire space of all possible motifs conforming to a predefined set of allowable motif configurations. This study considered all 7,680 possible 4-mer, 5-mer, and 6-mer contiguous motifs (e.g., CTGCAG), as well as 194,560 bipartite motifs (e.g., CATNNNNNCTC). A subset of available reads ($N = 20{,}000$) are sampled and methylation scores are compiled for each of the 202,240 motifs. Only those motifs with methylation scores > 1.6 on at least one contig are retained. Finally, multiple specifications of a motif are replaced by a single degenerate motif using IUPAC nucleotide codes. See the **Supplementary Methods** for additional details.

Combined *k*-mer frequency and methylation score vectors. The combination of *k*-mer frequency and methylation scores used to segregate contigs in the combined infant gut microbiome samples A and B (**Supplementary Fig. 4c**) was done by z-score transforming both feature matrices after each had been reduced to 2D using t-SNE. The two 2D matrices of z-scores were then combined and the resulting 4D matrix of z-scores was subjected to a second round of t-SNE to generate the final 2D map.

Bin validation and annotation. We applied CheckM[33] to assess the genome completeness and contamination in binned genomes. After writing the contig sequences in each bin to a fasta file in a directory called *bins*, we ran the following CheckM command:

checkm lineage_wf -t 8 -x fasta bins/ out

For species annotation, a database of 591 reference genomes isolated from the mouse gut was compiled from four recent studies[34–37] (**Supplementary Table 11**). Bin-level fasta files for the 541 genomes identified in Xiao *et al.*[36] were created from binned gene sequences using the script *write_xiao_MGS_bin_fastas.py* (**Supplementary Code**) after downloading the data files located at https://genome.jgi.doe.gov/pages/dynamicOrganismDownload.jsf?organism=IMG_3300005806. After compiling the database of all 591 reference sequences, we ran *blastn* to identify which of the references had significant matches with the contigs in the nine bins identified using methylation profiles. Alignments >100 bp in length with >97% identity were considered significant. For each bin, the reference genomes were ranked based on the percentage of the total binned contig sequences that were covered by a significant hit with the reference. We then used the *mummer* package[60] to align the highest ranked matching references to the contigs in each bin and visualized the alignments (**Supplementary Fig. 13**) with the *mummer* package.

Plasmid and chromosome sequence composition distances. The empirical distribution of Euclidean distances between the plasmids and randomly selected bacteria was constructed by iterating over all plasmids in REBASE[43], randomly selecting a "host" bacterium for each plasmid, and calculating the 5-mer frequency vector (as described in **Sequence composition features**) of the plasmid, $\boldsymbol{Z}'_p$, and of the largest chromosome of the selected bacterium, $\boldsymbol{Z}'_c$. The distance, d, between each pair of vectors $\boldsymbol{Z}'_p$ and $\boldsymbol{Z}'_c$ was computed as the Euclidean norm of the difference between vectors:

$$d = \sqrt{\sum_{i=1}^{V} (\boldsymbol{Z}'_{c,i} - \boldsymbol{Z}'_{p,i})^2}$$

Survey of methylome uniqueness in simulated communities. Methylation motifs were gathered for each of the 878 SMRT sequenced bacterial genomes stored in the REBASE database[43] and mock communities of N species were constructed, where $N = 20, 40, 60, \ldots, 200$ and each community was created 1,000 times by randomly selecting from the 878 organisms. For each mock community, the methylation motifs for each constituent organism were analyzed and the number of organisms with a unique methylome in the community was returned, reported as the fraction of total organisms in the community. Multiple curves in **Figure 3c** represent the different results obtained by varying the multi-strain content of the mock communities. The same procedure was again used to analyze only those 155 organisms in REBASE that are known to host at least one plasmid sequence. Mock communities of N species were again constructed, where $N = 20, 40, 60$, and each community was created 1,000 times by randomly selecting from the 155 organisms. Multiple curves in **Figure 3d** represent the different results obtained by varying the multi-strain content of the mock communities.

Survey of methylation motif content in simulated sequences. For each SMRT-sequenced genome in the REBASE database[43], 500 genomic sequences were simulated by extracting nucleotide substrings of length L from random positions in the known reference sequence, where $L = 5, 10, 15, \ldots, 100$ kb. Given the known methylation motifs for each genome, the number of sequences containing the motifs was returned, reported as the fraction of the 500 total simulated sequences. Multiple curves in **Figure 3e** represent the different results obtained by varying the percentage of the genome's methylation motifs that are required to be present on each sequence. For instance, the 75% curve represents the number of simulated sequences that contain at least one instance of at least three quarters of the genome's total set of methylation motifs.

Methylome analysis of *Klebsiella pneumoniae* strain. We examined the REBASE[43] entry for a virulent and antibiotic-resistant strain 243-12 of *Klebsiella pneumoniae* (GenBank CP011313) that was isolated from a patient during a 2011 outbreak in Germany[61] and hosted a single 362-kb plasmid named pKpn23412-362 (GenBank CP011314). We then compared the methylome of *K. pneumoniae* str. 234-12 to those of nine other bacterial genomes listed in REBASE, all of which had more similar chromosome sequence composition to plasmid pKpn23412-362 (see **plasmid and chromosome sequence composition distances**) than did the true host *K. pneumoniae* str. 234-12 chromosome. The methylated motifs of plasmid pKpn23412-362, *K. pneumoniae* str. 234-12, and the nine other bacterial species were represented in a matrix where 0 and 1 represented unmethylated and methylated motifs, respectively. Another matrix was created using all 25 strains of *K. pneumoniae* listed in REBASE. Using the Python packages fastcluster[62] and SciPy[63], both matrices

were subject to 2-dimensional hierarchical clustering to evaluate methylome similarities across species and strains.

Matching plasmid and host methylation profiles. We defined a confident mapping of a plasmid to a host if contigs accounting for >75% of the host genome contained (1) the same methylated motifs (i.e., motifs with methylation score ≥ 1.6 calculated from ≥ 10 IPD values) that are found on the plasmid, and (2) no additional methylated motifs.

Identification of MGE contigs in metagenomic assembly. A combination of two methods was used to identify circular contigs in metagenomic assemblies: (1) a custom script aligned the 20-kb sequences at the beginning and end of contigs to look for evidence of circularization (**Supplementary Code**), and (2) the freely available program Circlator[64] was used with default parameters. Contigs identified as circularized were then manually checked using Gepard[65] to look for visual evidence of circularization, as opposed to signs of mis-assembly. Small (<200 kb) contigs were classified as conjugative transposons if they contained at least five genes encoding conjugative transposon-related genes, according to gene annotations generated by RAST[66].

Synthetic metagenomic communities *Eight-species synthetic mixture.* SMRT reads were obtained separately from eight individual bacterial species (**Supplementary Table 2**) and the reads were mixed, without any labeling, by combining one SMRT cell of sequencing from each species to create a synthetic metagenomic mixture at similar relative abundances. Read labels were applied for evaluation purposes only after all binning procedures were completed.

Human Microbiome Project Mock Community B. Sequencing data from 49 SMRT cells was downloaded from https://github.com/PacificBiosciences/DevNet/wiki/Human_Microbiome_Project_MockB_Shotgun. In order to simulate a more realistic mixture of the 20 species in the HMP mock community, we downsampled the raw sequencing reads to impose relative species abundances that follow a natural log decay curve (**Supplementary Fig. 14** and **Supplementary Table 2**). We first determined the species identity for all reads by aligning the reads to reference assemblies for each species. After determining the species mappings for all reads (excluding those with ambiguous alignments), we then selected reads from each species to impose our desired relative abundances. The alignment and labeling procedures were used strictly for data downsampling and were not part of the read-level binning procedure. Reads in their original abundances were assembled to verify that the contig binning in **Supplementary Figure 7** was due to sequence composition differences, not due to poor assembly of the downsampled reads (**Supplementary Fig. 15**).

Multi-strain mixture of Helicobacter pylori. Two strains of *H. pylori*, str. 26695 (NC_000915) and str. J99 (NC_000921), were sequenced separately using a Pacific Biosciences RSII instrument as part of a previous study[29]. In order to generate matching 150× sequence coverage for each strain, reads were downsampled to 35,093 and 30,043 reads for strains 26695 and J99, respectively (**Supplementary Table 2**). All reads were combined before binning and assembly without knowledge of their strain of origin. Strain chimerism was assessed by mapping strain labels back to assembled reads after assembly.

Multi-strain mixture of Escherichia coli. Three strains of *E. coli*, BAA-2196 O26:H11, BAA-2215 O103:H11, and BAA-2440 O111, were sequenced separately using a Pacific Biosciences RSII instrument (see Online Methods section entitled **Three *E. coli* strains for synthetic mixture**). The synthetic, multi-strain mixture was created by combining a single SMRT cell from each of these separate sequencing runs (**Supplementary Table 2**). All reads were combined before binning and assembly without knowledge of their strain of origin. Strain chimerism was assessed by mapping strain labels back to assembled reads after assembly. In order to prevent sequencing errors from corrupting the IPD signatures for longer methylation motifs, we conducted an error-correction step by aligning the raw reads from each strain to the *E. coli* K12 MG1655 reference sequence (RefSeq accession NC_000913.3) before constructing read-level methylation scores for each motif.

t-SNE embedding for dimensionality reduction. t-SNE is a non-linear algorithm that is designed to preserve local pairwise distances, contrasting linear methods that capture global variance, such as principal components analysis (PCA). This makes t-SNE well suited for complex microbiome communities with subpopulations described by high-dimensional features. The high-dimensional matrix of features (e.g. *k*-mer frequencies, methylation scores, or a combination thereof) for all sequences was subjected to the Barnes–Hut implementation of t-distributed stochastic neighbor embedding (t-SNE)[31]. The Barnes–Hut approximation of t-SNE reduces the computational complexity from $\mathcal{O}(N^2)$ to $\mathcal{O}(N \log N)$, making it feasible to generate 2D maps of hundreds of thousands of metagenomic sequences containing hundreds of features. All runs used the default parameters for perplexity (30) and theta (0.5). Large assembled contigs (>50 kb) are represented in the high-dimensional matrices by multiple 'sub-contigs' in order to give them more weight during minimization of the t-SNE objective function (**Supplementary Methods**).

Metagenomic assembly. All metagenomic assemblies in this study used the hierarchical genome-assembly process (HGAP3)[45]. With the exception of the parameter specifying the expected genome size to be assembled, all default parameters were used. See **Supplementary Methods** for the *genomeSize* parameter values used for each assembly.

Metagenomic annotations using Kraken. Kraken version 0.10.5-beta[10] was configured to use two databases. The database used to annotate sequences from the Human Microbiome Project (HMP)[2] Mock Community B consisted of reference sequences for the 20 known species included in the mock community (**Supplementary Table 2**). All other Kraken annotations used a database consisting of the RefSeq complete set of bacterial/archaeal genomes (using "–download-library bacteria") and draft assemblies of five *Bacteroides dorei* strains. Database construction from these libraries and all Kraken annotations used default parameters. Bin-level annotations (**Table 1** and **Supplementary Table 7**) reflect the Kraken annotation (the taxonomic order) assigned to the largest percentage of contig bases in each bin.

Life Sciences Reporting Summary. Further information on experimental design is available in the Life Sciences Reporting Summary.

Data availability. All nucleotide sequences assembled in this study have been deposited at DDBJ/ENA/GenBank: 8-species synthetic mixture assembly (accession code PDZQ01000000), adult mouse gut microbiome assembly (PDYJ01000000), human infant gut microbiome assembly (PDYI01000000), pHel3 plasmid (MG214727). All sequencing data generated in this study are available under NCBI BioProject PRJNA404082. Source data files for **Figures 1–4** are available online.

53. Sokol, H. *et al. Faecalibacterium prausnitzii* is an anti-inflammatory commensal bacterium identified by gut microbiota analysis of Crohn disease patients. *Proc. Natl. Acad. Sci. USA* **105**, 16731–16736 (2008).
54. Livanos, A.E. *et al.* Antibiotic-mediated gut microbiome perturbation accelerates development of type 1 diabetes in mice. *Nat. Microbiol.* **1**, 16140 (2016).
55. Heuermann, D. & Haas, R. A stable shuttle vector system for efficient genetic complementation of *Helicobacter pylori* strains by transformation and conjugation. *Mol. Gen. Genet.* **257**, 519–528 (1998).
56. Zhang, X.S. & Blaser, M.J. Natural transformation of an engineered *Helicobacter pylori* strain deficient in type II restriction endonucleases. *J. Bacteriol.* **194**, 3407–3416 (2012).
57. Leonard, M.T. *et al.* The methylome of the gut microbiome: disparate Dam methylation patterns in intestinal *Bacteroides dorei*. *Front. Microbiol.* **5**, 361 (2014).
58. Langmead, B., Trapnell, C., Pop, M. & Salzberg, S.L. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol.* **10**, R25 (2009).
59. Feng, Z. *et al.* Detecting DNA modifications from SMRT sequencing data by modeling sequence context dependence of polymerase kinetic. *PLOS Comput. Biol.* **9**, e1002935 (2013).
60. Kurtz, S. *et al.* Versatile and open software for comparing large genomes. *Genome Biol.* **5**, R12 (2004).

doi:10.1038/nbt.4037

61. Becker, L. *et al.* Complete genome sequence of a CTX-M-15-producing *Klebsiella pneumoniae* outbreak strain from multilocus sequence type 514. *Genome Announc.* **3**, e00742–e15 (2015).
62. Müllner, D. fastcluster: Fast hierarchical, agglomerative. *J. Stat. Softw.* **53**, 1–18 (2013).
63. van der Walt, S., Colbert, S.C. & Varoquaux, G. The NumPy Array: a structure for efficient numerical computation. *Comput. Sci. Eng.* **13**, 22–30 (2011).
64. Hunt, M. *et al.* Circlator: automated circularization of genome assemblies using long sequencing reads. *Genome Biol.* **16**, 294 (2015).
65. Krumsiek, J., Arnold, R. & Rattei, T. Gepard: a rapid and sensitive tool for creating dotplots on genome scale. *Bioinformatics* **23**, 1026–1028 (2007).
66. Aziz, R.K. *et al.* The RAST Server: rapid annotations using subsystems technology. *BMC Genomics* **9**, 75 (2008).

nature research

Corresponding author(s): Gang Fang

☐ Initial submission ☐ Revised version ☒ Final submission

Life Sciences Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form is intended for publication with all accepted life science papers and provides structure for consistency and transparency in reporting. Every life science submission will use this form; some list items might not apply to an individual manuscript, but all fields must be completed for clarity.

For further information on the points included in this form, see Reporting Life Sciences Research. For further information on Nature Research policies, including our data availability policy, see Authors & Referees and the Editorial Policy Checklist.

Experimental design

1. Sample size

Describe how sample size was determined.

This applies to simulation analysis to examine methylome uniqueniss in a microbiome sample with different complexity (number of species, and strains). We subsampled the update to date largest collection of bacterial methylome data (>800 unique species) into different levels of complexity as described in detail in Methods.

2. Data exclusions

Describe any data exclusions.

No sample exclusion. Sequencing data outlier removal clearly described in Methods.

3. Replication

Describe whether the experimental findings were reliably reproduced.

Yes.

4. Randomization

Describe how samples/organisms/participants were allocated into experimental groups.

N/A

5. Blinding

Describe whether the investigators were blinded to group allocation during data collection and/or analysis.

N/A

Note: all studies involving animals and/or human research participants must disclose whether blinding and randomization were used.

6. Statistical parameters

For all figures and tables that use statistical methods, confirm that the following items are present in relevant figure legends (or in the Methods section if additional space is needed).

n/a	Confirmed
☐	☒ The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement (animals, litters, cultures, etc.)
☐	☒ A description of how samples were collected, noting whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	☒ A statement indicating how many times each experiment was replicated
☐	☒ The statistical test(s) used and whether they are one- or two-sided (note: only common tests should be described solely by name; more complex techniques should be described in the Methods section)
☐	☒ A description of any assumptions or corrections, such as an adjustment for multiple comparisons
☐	☒ The test results (e.g. *P* values) given as exact values whenever possible and with confidence intervals noted
☐	☒ A clear description of statistics including central tendency (e.g. median, mean) and variation (e.g. standard deviation, interquartile range)
☐	☒ Clearly defined error bars

See the web collection on statistics for biologists for further resources and guidance.

Nature Biotechnology: doi:10.1038/nbt.4037

Software

Policy information about availability of computer code

7. Software

Describe the software used to analyze the data in this study.	In Supplementary Code. Also the software supporting all proposed methods is implemented in Python and is available with full documentation at http://www.github.com/fanglab/mbin

For manuscripts utilizing custom algorithms or software that are central to the paper but not yet described in the published literature, software must be made available to editors and reviewers upon request. We strongly encourage code deposition in a community repository (e.g. GitHub). *Nature Methods* guidance for providing algorithms and software for publication provides further information on this topic.

Materials and reagents

Policy information about availability of materials

8. Materials availability

Indicate whether there are restrictions on availability of unique materials or if these materials are only available for distribution by a for-profit company.	No restriction

9. Antibodies

Describe the antibodies used and how they were validated for use in the system under study (i.e. assay and species).	N/A

10. Eukaryotic cell lines

a. State the source of each eukaryotic cell line used.	N/A
b. Describe the method of cell line authentication used.	N/A
c. Report whether the cell lines were tested for mycoplasma contamination.	N/A
d. If any of the cell lines used are listed in the database of commonly misidentified cell lines maintained by ICLAC, provide a scientific rationale for their use.	N/A

Animals and human research participants

Policy information about studies involving animals; when reporting animal research, follow the ARRIVE guidelines

11. Description of research animals

Provide details on animals and/or animal-derived materials used in the study.	A male 6-weekold NOD/shiltj mice (no. 001976, Jackson Labs) was housed in a Specific Pathogen Free (SPF) room at New York University Langone Medical Center (NYUMC). At the week 12 of life, the mouse were placed into a clean plastic container in a fume hood, and its freshly fecal pellets were collected in sterilized microcentrifuge tubes and frozen at –80°C. Fecal DNA was extracted using PowerSoil DNA isolation kit (MoBio Labs, Carsbad, CA).

Policy information about studies involving human research participants

12. Description of human research participants

Describe the covariate-relevant population characteristics of the human research participants.	N/A. No human subject recruited for this study.

Nature Biotechnology: doi:10.1038/nbt.4037