Published in final edited form as:
Nat Biotechnol. 2018 January ; 36(1): 61–69. doi:10.1038/nbt.4037.

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**[1,2,3], **Xue-Song Zhang**[4], **Austin Davis-Richardson**[5], **Ronald Canepa**[5], **Eric W. Triplett**[5], **Jeremiah J. Faith**[1,2,3], **Robert Sebra**[1,2], **Eric E. Schadt**[1,2], and **Gang Fang**[1,2,#]

[1]Department of Genetics and Genomic Sciences, Icahn School of Medicine at Mount Sinai, New York, NY 10029, USA

[2]Icahn Institute for Genomics and Multiscale Biology, Icahn School of Medicine at Mount Sinai, New York, NY 10029, USA

[3]Immunology Institute, Icahn School of Medicine at Mount Sinai, New York, NY 10029, USA

[4]Department of Medicine, New York University School of Medicine, New York 10016, USA

[5]Department of Microbiology and Cell Science, Institute of Food and Agricultural Sciences, University of Florida, Gainesville, FL, 32611, USA

Abstract

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

[#]Address correspondence to: gang.fang@mssm.edu.

Editors summary
Similar bacterial species and strains in metagenomes are resolved and linked to the plasmids that they harbor using DNA methylation analysis.

Accession codes
All nucleotide sequences assembled in this study have been deposited 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.

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.T., J.F. R.S., E.S. and G.F. contributed to experimental design. I.M., X. Z., A.D-R., R.C., M.L., E.T. and J.F. conducted the experiments. G.D. and R.S. designed and conducted sequencing. J.B., S.Z., E.T., J.F., R.S., E.S. and G.F. analysed the data. J.B. and GF wrote the manuscript with inputs and comments from all co-authors. G.F. conceived and supervised the project.

Competing Financial Interests
E.E.S. is on the scientific advisory board of Pacific Biosciences. J.B. and G.F. are inventors of a U.S. Provisional patent application (No. 62/525,908) that describes the method for methylation binning.

validated 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.

INTRODUCTION

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 unmethylated[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 (HGT) 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 identify methylated motifs. We show that combination of sequence features based on composition and coverage with methylation motifs can improve genome segregation and linking of 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 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] (Online Methods).

The sensitivity and specificity of a motif methylation score are a function of the number of IPD values comprising the score (Fig. 2a; Online Methods). 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].

Methylation scores for multiple motifs are compiled into methylation profiles. The methylated motifs included in the profile are 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 are retained for inclusion in the methylation profiles (Supplementary Methods). Filtering resulted in profiles of between 7-38 motifs for the metagenomic samples that we analysed (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; Online Methods). All sequencing data from this study is available through NCBI BioProject PRJNA404082. Following metagenomic assembly of the combined reads (Supplementary Table 3; Online Methods), our motif filtering procedure identified 16 N6-methyladenine (6mA) motifs from the metagenomic contigs based solely on methylation scores, 14 (87.5%) of which were exact matches to the true methylated motifs (as validated by independent methylation analysis of each species prior to 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 specification of the true motif. Hierarchical clustering of the motif methylation scores for the largest contigs from each species reveals 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] (Online Methods), which has previously been used to visualize metagenomic sequence composition features[13,14]. The 2D map of methylation features generated by t-SNE reveals contigs that are 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 either t-SNE to generate a scatter plot of 5-mer frequency features alone (Supplementary Fig. 1a; Online Methods) or a scatter plot of contig coverage values vs. GC-content (Supplementary Fig. 1b; Online Methods). Two small, high-coverage *Collinsella aerofaciens* contigs (putative plasmids) in the coverage vs. 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* are missing certain methylated motif sites (Supplementary Fig. 2) and therefore cluster 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.

Next, we analysed methylation profiles of contigs assembled from SMRT sequencing of a fecal microbiota sample isolated from an adult mouse (Online Methods; Supplementary Table 2). 16S rRNA gene amplicon sequencing (Online Methods) 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 in the assembly (Supplementary Table 3) and visualized the methylation landscape using t-SNE (Fig. 2e). Contigs were annotated using Kraken[10] (Supplementary Table 5; Online Methods).

We identified nine distinct contig bins using 38 methylation features in the murine gut microbiota sample. Seven bins assigned to the order *Bacteroidales* share high ANI with each other (81-91% ANI), but at values suggesting inter- rather than intraspecies relationships[32] (Supplementary Table 6; Supplementary Methods). In eight of nine bins, alignment of reads to the binned contigs revealed uniform coverage values within each bin (Supplementary Table 6; Online Methods), suggesting that the bins correspond 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 has 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; Online Methods) and visualizing the contigs in a scatter plot of sample coverage versus coverage from a related sample (Supplementary Fig. 3b; Online Methods). 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 approach described by Albertsen et al[16]. An additional analysis of infant gut microbiome sequencing (Online Methods; Supplementary Table 2) demonstrated how methylation profiles can complement sequence composition features to resolve contigs from two mixed strains of *Bacteroides dorei* (Supplementary Methods; Supplementary Figs. 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 other 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; Supplementary Fig. 5), we combined methylation profiles of contigs in each CONCOCT bin (Online Methods). 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 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 Figs. 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 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 among 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 5-mer frequency statistics of plasmids and chromosomes of their bacterial hosts (Online Methods), 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 the 5.5 kb plasmid pHel3 (GenBank MG214727) from *Escherichia coli* DH5α into *E. coli* CFT073 and *Helicobacter pylori* JP26 (Online Methods), 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 is 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 datasets, we first simulated communities of between 20-200 members by sampling methylomes of SMRT sequenced bacterial chromosomes and plasmids from the REBASE database[43] (Online Methods). 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) decreases in larger synthetic communities (Fig. 3c) and is more pronounced when multiple strains of a species are present. Similar trends were observed when only including 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 (Online Methods), we found that, on average, 90% of 35 kb sequences contain at least 75% of the 6mA motifs found in the host genome, and that 90% of 60 kb sequences capture 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 thirteen antibiotic resistance genes. By comparing the methylome of *K. pneumoniae* str. 234-12 with nine other similar species and 24 other *K. pneumoniae* strains (Online Methods), we found the methylation profile of *K. pneumoniae* str. 234-12 to be unique among the examined genomes (Supplementary Figs. 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 (Online Methods) 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 (Online Methods), 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 not incorrectly mapped to the wrong host, but 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 nineteen MGE contigs in the adult mouse gut microbiome assembly (Supplementary Table 3) between 7-132 kb, of which ten are fully circularized and nine are conjugative transposons (encoding at least five genes annotated as conjugative transposon-related) (Online Methods). Conjugative transposons have an important role in HGT 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[45] (Supplementary Methods). Of the nineteen identified MGE contigs, eight had methylation profiles that could be conclusively matched to the previously identified methylation bins containing genomes from the order *Bacteroidales* (Table 1; Online Methods). 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 are highly species-specific and resilient to random sequencing errors. For instance, despite having very low sequence coverage that precludes assembly (Supplementary Fig. 7), unassembled reads from *Rhodobacter sphaeroides* form a distinct cluster when read-level 5-mer frequency profiles are visualized using t-SNE (Figs. 4a and 4b). Unsurprisingly, species segregation improves with increasing read lengths (Supplementary Figs. 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 (Online Methods) 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 result 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 (Online Methods) across four 6mA motifs present at high density in the genome: GATC, GAGG, TGCA, and CATG[46] (Supplementary Table 8). 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 (Online Methods) 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 removed much of the sequencing and IPD errors that occur in longer motifs in raw reads (Online Methods). Bulk assembly of the mixture of error-corrected reads resulted in many chimeric contigs and very few contigs that are specific to a strain (Fig. 4f), but binning the reads by methylation profiles across four differentiating motifs (Fig. 4g; Supplementary Table 9) prior to 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 (Figs. 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 approach 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; 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.

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 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 10kb 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, Carsbad, CA). A Life Sciences Reporting Summary is available for this study.10kb 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]

pHel3 plasmid transformation into three species

The *E. coli-H. pylori* shuttle plasmid pHel3[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 hours. *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 condition for 48 hours. 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). 2kb 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 10kb 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 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 minutes and 30 seconds followed by decreasing the temperature by 0.1° 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 hours at 30°C and then held at 4°C until ready for magbead loading, prior to sequencing. The magnetic bead-loading step was conducted at 4°C for 60 minutes 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-minute 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 $V = 512$) for each sequence (contig or single-molecule read), *i*, denoted $\boldsymbol{Z_i} = \left(Z_{i,1}, \ldots, Z_{i,V}\right)$. 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*[58] and SMRT reads were aligned to contigs during the HGAP3[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 $Y_i = \left(Y_{i,1}, \ldots, Y_{i,N}\right)$, where *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 log-transforming the ratio of each subread IPD value to the mean of all IPD values in the subread, corrects for any potential slowing of polymerase kinetics over the course of an entire read (which can consists of multiple subreads)[28,42]. Each normalized IPD (nIPD) value in the subread is calculated as follows:

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

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

$$\mathrm{R}_{\mathrm{ij}}^{\mathrm{o}}{=}\frac{1}{\sum_{\mathrm{s}=1}^{\mathrm{S}}\mathrm{M}_{\mathrm{s}}}\sum_{\mathrm{s}=1}^{\mathrm{S}}\sum_{\mathrm{m}=1}^{\mathrm{M}_{\mathrm{s}}}\mathrm{nIPD}_{\mathrm{ms}}$$

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, $\mathrm{R}_{\mathrm{ij}}^{\mathrm{o}}$, a corresponding set of control kinetics scores, $\mathrm{R}_{\mathrm{i}}^{\mathrm{c}}$. These control kinetics scores are motif-matched and calculated similar to $\mathrm{R}_{\mathrm{ij}}^{\mathrm{o}}$ using a sampling of SMRT sequencing unaligned reads (N=20,000) known to be free of any methylation:

$$\mathrm{R}_{\mathrm{ij}}{=}\mathrm{R}_{\mathrm{ij}}^{\mathrm{o}}-\mathrm{R}_{\mathrm{i}}^{\mathrm{c}}$$

As no methylated motifs were detected after sequencing an isolate of *Ruminococcus gnavus*, this data served as the non-methylated control set for calculating values of $\mathrm{R}_{\mathrm{i}}^{\mathrm{c}}$. These non-methylated control values are used for the motif filtering procedure, but not for the final calculation of methylation profiles. Because the dimensionality reduction with t-SNE calculates a Euclidian 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:

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

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, $\mathrm{C}_{\mathrm{i}}^{\mathrm{c}}$, are generated using a sample of aligned reads (N=20,000) known to be free of methylation and subtracted from the observed kinetics scores, $\mathrm{C}_{\mathrm{ij}}^{\mathrm{o}}$, in order to remove the baseline kinetics variation stemming from local sequence context:

$$\mathrm{C}_{\mathrm{ij}}=\mathrm{C}_{\mathrm{ij}}^{\mathrm{o}}-\mathrm{C}_{\mathrm{i}}^{\mathrm{c}}$$

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 > 1.7 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 >100bp 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 Euclidian 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 Euclidian norm of the difference between vectors:

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

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 analysed and 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 Fig. 3c represent the different results obtained by varying the multi-strain content of the mock communities. The same procedure was again used to analyse 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 Fig. 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, …, 100kb. 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 Fig. 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 362kb 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 20kb 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 (<200kb) 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 twenty 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; 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 Fig. 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 150x 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 prior to 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 prior to 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) prior to 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 subpopulation described by high-dimensional features. The high-dimensional matrix of features (e.g. *k*-mer frequencies, methylation scores, or a combination) 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 $\mathcal{O}\left(N^2\right)$) to $\mathcal{O}\left(N \log N\right)$), 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 twenty 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.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

We thank Martha Lewis for her assistance in DNA extraction and Ali 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 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.

References

1. Cho I, Blaser MJ. The human microbiome: at the interface of health and disease. Nat Rev Genet. 2012; 13:260–270. [PubMed: 22411464]
2. The Human Microbiome Project Consortium. Structure, function and diversity of the healthy human microbiome. Nature. 2012; 486:207–214. [PubMed: 22699609]
3. Janda JM, Abbott SL. 16S rRNA gene sequencing for bacterial identification in the diagnostic laboratory: pluses, perils, and pitfalls. J Clin Microbiol. 2007; 45:2761–4. [PubMed: 17626177]
4. Qin J, et al. A human gut microbial gene catalogue established by metagenomic sequencing. Nature. 2010; 464:59–65. [PubMed: 20203603]
5. Tyson GW, et al. Community structure and metabolism through reconstruction of microbial genomes from the environment. Nature. 2004; 428:37–43. [PubMed: 14961025]
6. Modi SR, Lee HH, Spina CS, Collins JJ. Antibiotic treatment expands the resistance reservoir and ecological network of the phage metagenome. Nature. 2013; 499:219–22. [PubMed: 23748443]
7. Luo C, et al. ConStrains identifies microbial strains in metagenomic datasets. Nat Biotechnol. 2015; 33:1045–1052. [PubMed: 26344404]
8. Kuleshov V, et al. Synthetic long-read sequencing reveals intraspecies diversity in the human microbiome. Nat Biotechnol. 2015; 34:64–69. [PubMed: 26655498]
9. Brady A, Salzberg SL. Phymm and PhymmBL: metagenomic phylogenetic classification with interpolated Markov models. Nat Methods. 2009; 6:673–6. [PubMed: 19648916]
10. Wood DE, Salzberg SL. Kraken: ultrafast metagenomic sequence classification using exact alignments. Genome Biol. 2014; 15:R46. [PubMed: 24580807]
11. Saeed I, Tang SL, Halgamuge SK. Unsupervised discovery of microbial population structure within metagenomes using nucleotide base composition. Nucleic Acids Res. 2012; 40
12. Iverson V, et al. Untangling genomes from metagenomes: revealing an uncultured class of marine Euryarchaeota. Science. 2012; 335:587–90. [PubMed: 22301318]
13. Laczny C, Pinel N, Vlassis N, Wilmes P. Alignment-free Visualization of Metagenomic Data by Nonlinear Dimension Reduction. Sci Rep. 2014; :1–12. DOI: 10.1038/srep04516
14. Laczny CC, et al. VizBin - an application for reference-independent visualization and human-augmented binning of metagenomic data. Microbiome. 2015; :1–7. DOI: 10.1186/s40168-014-0066-1 [PubMed: 25621171]
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. 2013; 23:111–20. [PubMed: 22936250]
16. Albertsen M, et al. Genome sequences of rare, uncultured bacteria obtained by differential coverage binning of multiple metagenomes. Nat Biotechnol. 2013; 31:533–8. [PubMed: 23707974]
17. Nielsen HB, et al. Identification and assembly of genomes and genetic elements in complex metagenomic samples without using reference genomes. Nat Biotechnol. 2014; 32
18. Alneberg J, et al. Binning metagenomic contigs by coverage and composition. Nat Methods. 2014; 11
19. Marbouty M, et al. Metagenomic chromosome conformation capture (meta3C) unveils the diversity of chromosome organization in microorganisms. Elife. 2014; 3:e03318. [PubMed: 25517076]
20. Burton JN, Liachko I, Dunham MJ, Shendure J. Species-Level Deconvolution of Metagenome Assemblies with Hi-C-Based Contact Probability Maps. G3 (Bethesda). 2014; 4:1339–1346. [PubMed: 24855317]
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. 2017; 3:e1602105. [PubMed: 28232956]
22. Flusberg BA, et al. Direct detection of DNA methylation during single-molecule, real-time sequencing. Nat Methods. 2010; 7:461–5. [PubMed: 20453866]
23. Eid J, et al. Real-time DNA sequencing from single polymerase molecules. Science (80-). 2009; 323:133–138.

24. Casadesús J, Low D. Epigenetic gene regulation in the bacterial world. Microbiol Mol Biol Rev. 2006; 70:830–56. [PubMed: 16959970]

25. Blow MJ, et al. The Epigenomic Landscape of Prokaryotes. PLOS Genet. 2016; 12:e1005854. [PubMed: 26870957]

26. Kobayashi I, Nobusato A, Kobayashi-Takahashi N, Uchiyama I. Shaping the genome--restriction-modification systems as mobile genetic elements. Curr Opin Genet Dev. 1999; 9:649–656. [PubMed: 10607611]

27. Conlan S, et al. Single-molecule sequencing to track plasmid diversity of hospital-associated carbapenemase-producing Enterobacteriaceae. Sci Transl Med. 2014; 6:254ra126.

28. Schadt EE, et al. Modeling kinetic rate variation in third generation DNA sequencing data to detect putative modifications to DNA bases. Genome Res. 2013; 23:129–41. [PubMed: 23093720]

29. Beaulaurier J, et al. Single molecule-level detection and long read-based phasing of epigenetic variations in bacterial methylomes. Nat Commun. 2015; 6:7438. [PubMed: 26074426]

30. van der Maaten L, Hinton G. Visualizing Data using t-SNE. J Mach Learn Res. 2008; 9:2579–2605.

31. van der Maaten L. Accelerating t-sne using tree-based algorithms. J Mach Learn Res. 2014; 15:3221–3245.

32. Kim M, Oh H, Park S, 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. 2014; 64:346–351. [PubMed: 24505072]

33. Parks DH, Imelfort M, Skennerton CT, Hugenholtz P, Tyson GW. CheckM: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. Genome Res. 2015; 25:1043–55. [PubMed: 25977477]

34. Uchimura Y, et al. Complete Genome Sequences of 12 Species of Stable Defined Moderately Diverse Mouse Microbiota 2. Genome Announc. 2016; 4:4–5.

35. Ormerod KL, et al. Genomic characterization of the uncultured Bacteroidales family S24-7 inhabiting the guts of homeothermic animals. Microbiome. 2016; 4:36. [PubMed: 27388460]

36. Xiao L, et al. A catalog of the mouse gut metagenome. Nat Biotechnol. 2015; 33:1103–1108. [PubMed: 26414350]

37. Wannemuehler MJ, Overstreet A, Ward DV, Phillips J. Draft Genome Sequences of the Altered Schaedler Flora, a Defined Bacterial Community from Gnotobiotic Mice. Genome Announc. 2014; 2:1–2.

38. Imelfort M, et al. GroopM: An automated tool for the recovery of population genomes from related metagenomes. PeerJ. 2014; 2:e409v1. [PubMed: 24949232]

39. Kang DD, Froula J, Egan R, Wang Z. MetaBAT, an efficient tool for accurately reconstructing single genomes from complex microbial communities. PeerJ. 2015; 3:e1165. [PubMed: 26336640]

40. Slater FR, Bailey MJ, Tett AJ, Turner SL. Progress towards understanding the fate of plasmids in bacterial communities. FEMS Microbiol Ecol. 2008; 66:3–13. [PubMed: 18507680]

41. Thomas CM, Nielsen KM. Mechanisms of, and barriers to, horizontal gene transfer between bacteria. NatRevMicrobiol. 2005; 3:711–721.

42. Fang G, et al. Genome-wide mapping of methylated adenine residues in pathogenic *Escherichia coli* using single-molecule real-time sequencing. Nat Biotechnol. 2012; 30:1232–9. [PubMed: 23138224]

43. Roberts RJ, Vincze T, Posfai J, Macelis D. REBASE-a database for DNA restriction and modification: Enzymes, genes and genomes. Nucleic Acids Res. 2015; 43:D298–D299. [PubMed: 25378308]

44. Coyne MJ, et al. Evidence of Extensive DNA Transfer between *Bacteroidales* Species within the Human Gut. MBio. 2014; 5:e01305–14. [PubMed: 24939888]

45. Chin CS, et al. Nonhybrid, finished microbial genome assemblies from long-read SMRT sequencing data. Nat Methods. 2013; 10:563–9. [PubMed: 23644548]

46. Krebes J, et al. The complex methylome of the human gastric pathogen *Helicobacter pylori*. Nucleic Acids Res. 2013; :1–18. DOI: 10.1093/nar/gkt1201 [PubMed: 23143271]

47. Clarke J, et al. Continuous base identification for single-molecule nanopore DNA sequencing. Nat Nanotechnol. 2009; 4:265–270. [PubMed: 19350039]

48. Fuller CW, et al. Real-time single-molecule electronic DNA sequencing by synthesis using polymer-tagged nucleotides on a nanopore array. Proc Natl Acad Sci. 2016; 113:5233–5238. [PubMed: 27091962]

49. Rand AC, et al. Mapping DNA methylation with high-throughput nanopore sequencing. Nat Methods. 2017; 14:411–413. [PubMed: 28218897]

50. Lan F, Demaree B, Ahmed N, Abate AR. Single-cell genome sequencing at ultra-high-throughput with microfluidic droplet barcoding. Nat Biotechnol. 2017; 35:640–646. [PubMed: 28553940]

51. Rousseeuw PJ. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. J Comput Appl Math. 1987; 20:53–65.

52. Caporaso JG, et al. QIIME allows analysis of high-throughput community sequencing data. Nat Publ Gr. 2010; 7:335–336.

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 U S A. 2008; 105:16731–6. [PubMed: 18936492]

54. Livanos AE, et al. Antibiotic-mediated gut microbiome perturbation accelerates development of type 1 diabetes in mice. Nat Microbiol. 2016; 1:16140. [PubMed: 27782139]

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. 1998; 257:519–528. [PubMed: 9563837]

56. Zhang XS, Blaser MJ. Natural transformation of an engineered *Helicobacter pylori* strain deficient in type II restriction endonucleases. J Bacteriol. 2012; 194:3407–3416. [PubMed: 22522893]

57. Leonard MT, et al. The methylome of the gut microbiome: disparate Dam methylation patterns in intestinal *Bacteroides dorei*. Front Microbiol. 2014; 5:361. [PubMed: 25101067]

58. Langmead B, Trapnell C, Pop M, Salzberg SL. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biol. 2009; 10:R25. [PubMed: 19261174]

59. Feng Z, et al. Detecting DNA modifications from SMRT sequencing data by modeling sequence context dependence of polymerase kinetic. PLoS Comput Biol. 2013; 9:e1002935. [PubMed: 23516341]

60. Kurtz S, et al. Versatile and open software for comparing large genomes. Genome Biol. 2004; 5:R12. [PubMed: 14759262]

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. 2015; 3:e00742–15. [PubMed: 26159529]

62. Müllner D. fastcluster: Fast Hierarchical, Agglomerative. J Stat Softw. 2013; 53:1–18.

63. van der Walt S, Colbert SC, Varoquaux G. The NumPy Array: A Structure for Efficient Numerical Computation. Comput Sci Eng. 2011; 13:22–30.

64. Hunt M, et al. Circlator: automated circularization of genome assemblies using long sequencing reads. Genome Biol. 2015; 16:294. [PubMed: 26714481]

65. Krumsiek J, Arnold R, Rattei T. Gepard: A rapid and sensitive tool for creating dotplots on genome scale. Bioinformatics. 2007; 23:1026–1028. [PubMed: 17309896]

66. Aziz RK, et al. The RAST Server: Rapid Annotations using Subsystems Technology. BMC Genomics. 2008; 9:75. [PubMed: 18261238]

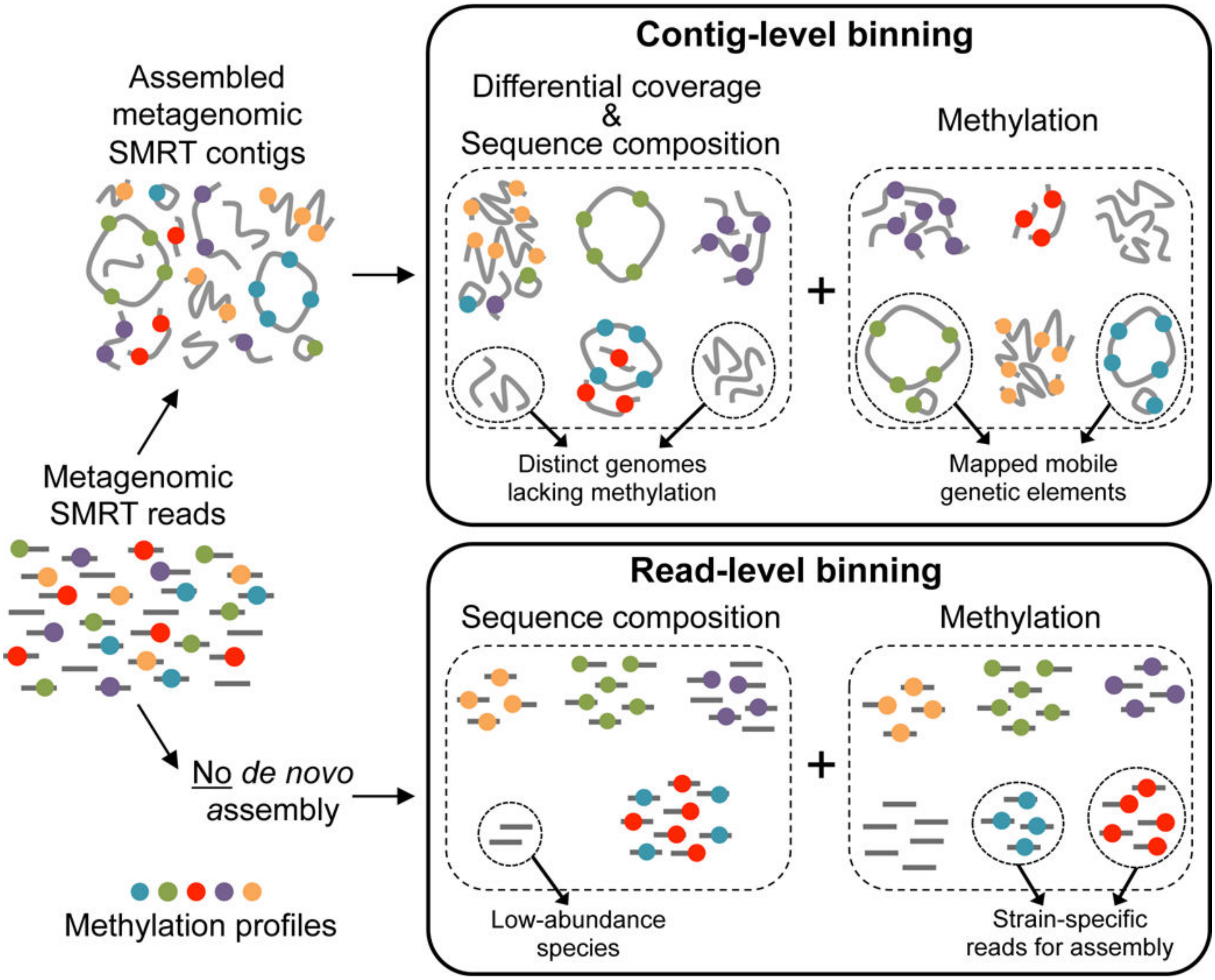
Assembled metagenomic SMRT contigs
Metagenomic SMRT reads
No de novo assembly
Methylation profiles
Contig-level binning
Differential coverage & Sequence composition
Methylation
+
Distinct genomes lacking methylation
Mapped mobile genetic elements
Read-level binning
Sequence composition
Methylation
+
Low-abundance species
Strain-specific reads for assembly

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 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.

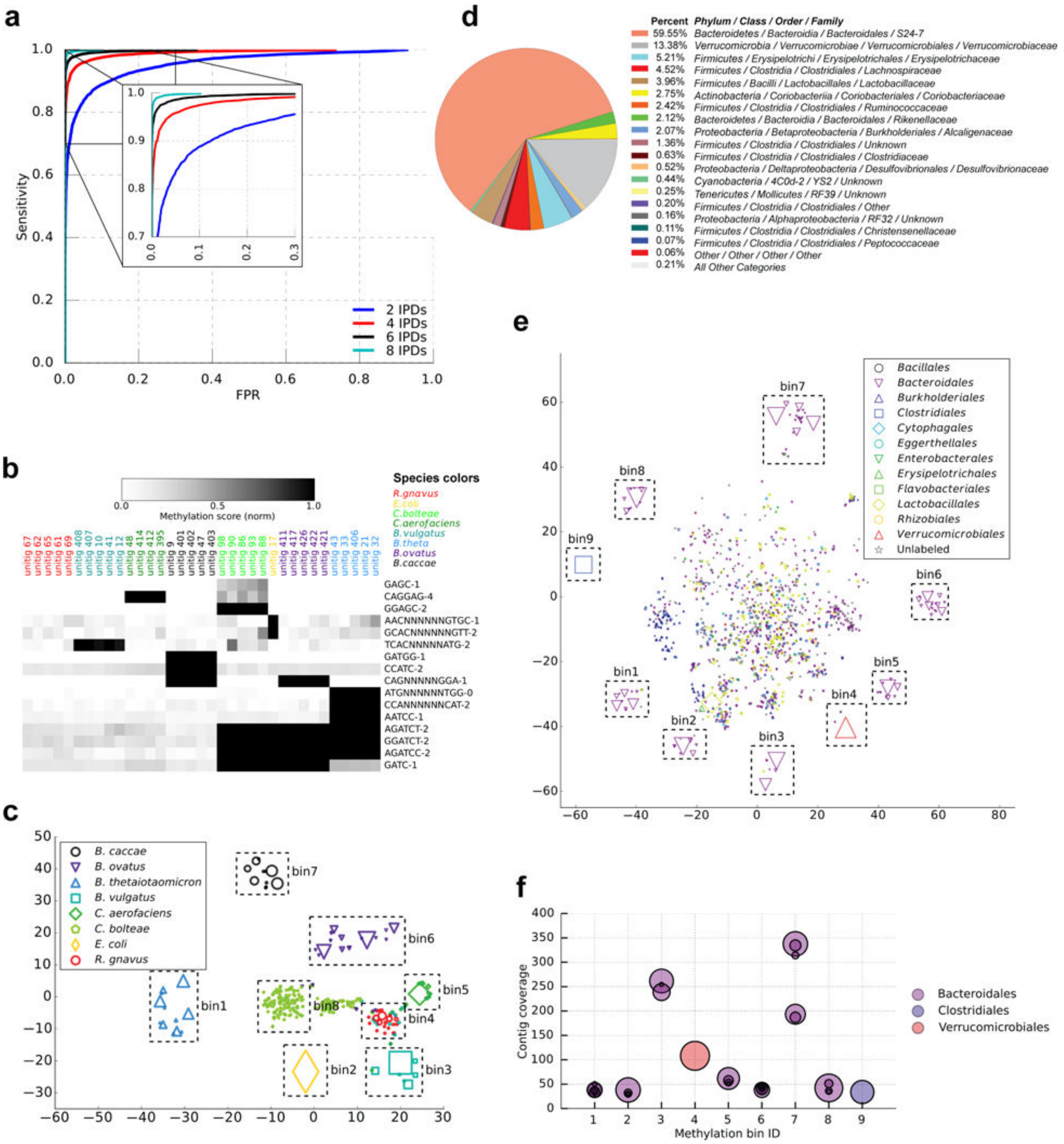

Figure 2. Metagenomic binning by methylation profiles

(a) Receiver operating characteristic (ROC) curve illustrating the power to classify a contig as methylated or non-methylated regarding a specific sequence motif, as a function of the number IPD values available for the motif sites on the contig. (b) Heatmap of contig-level methylation scores for fourteen 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 fourteen 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 (>100kp to exclude small MGEs) in each of the nine bins identified by methylation binning.

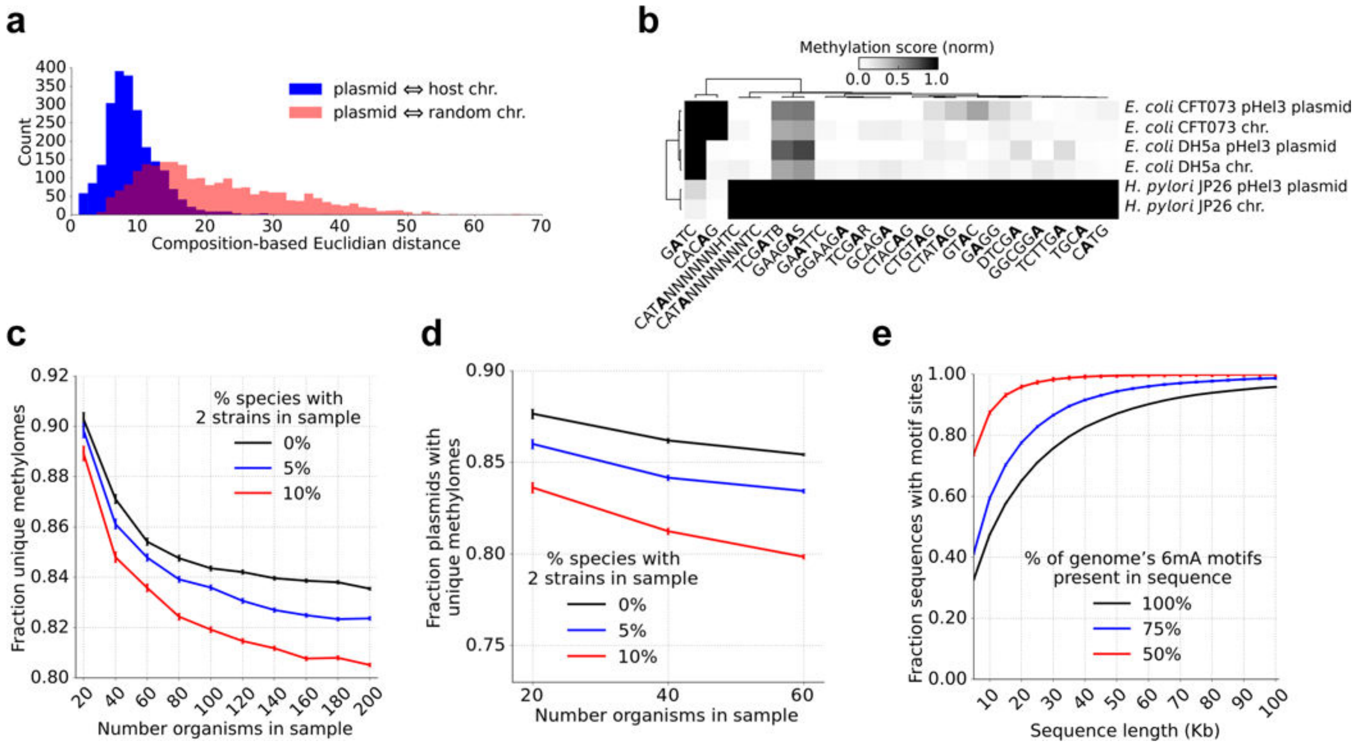

Figure 3. Methylation profiles can link plasmids to the chromosomal DNA of their host species.
(a)

Histogram of sequence-based Euclidian 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 twenty motifs matches the host from which it was isolated. **(c)** Simulation analysis (1000 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 (1000 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.

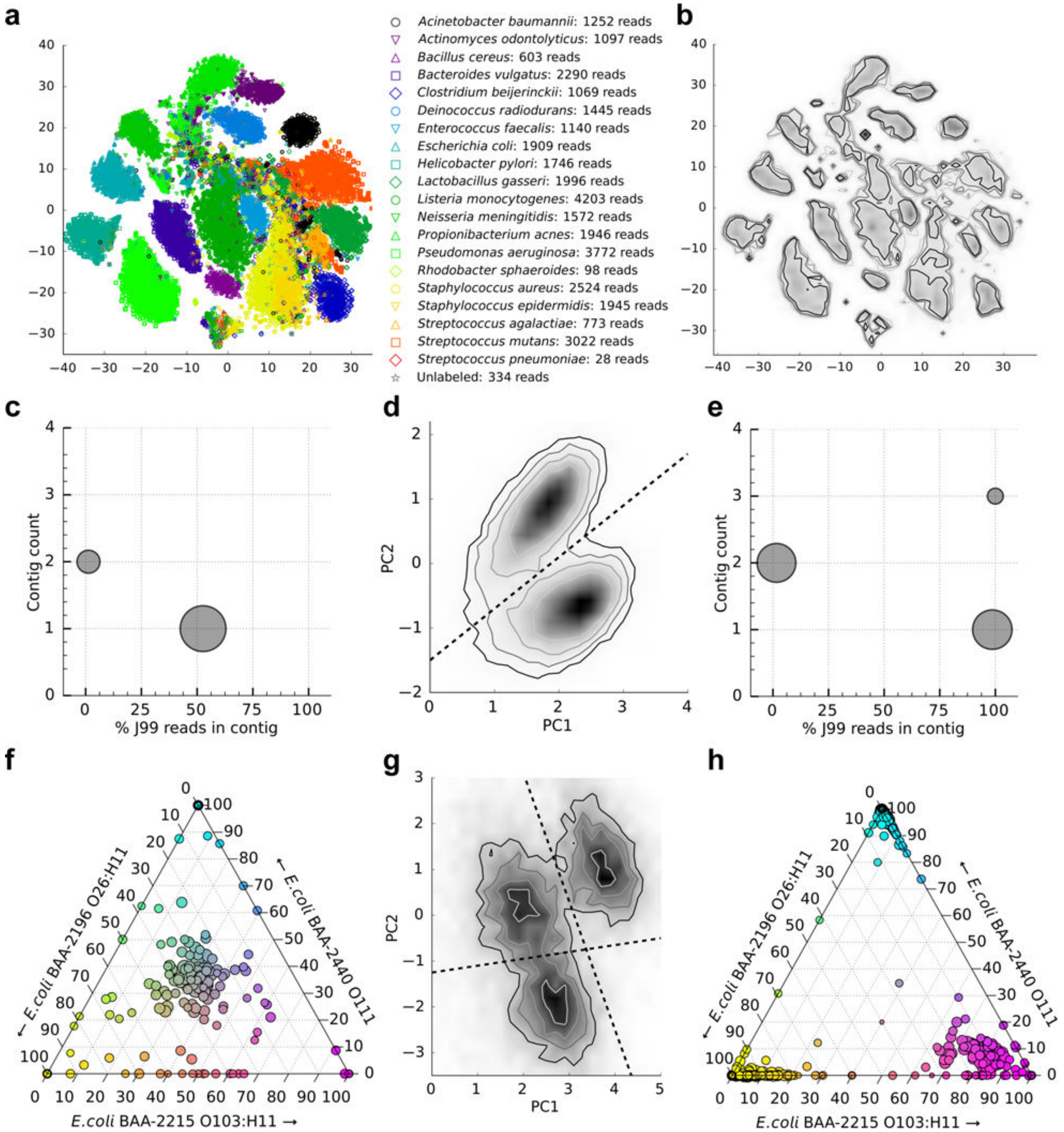

Figure 4. Binning SMRT reads using composition and DNA methylation profiles

(a) 5-mer frequency-based binning of assembled contigs and raw reads (length>15 kb) from the HMP 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.

Table 1

Genomes binned from adult mouse gut microbiome using DNA methylation profiles

Annotation of binned contigs was conducted using Kraken. 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 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 mobile genetic elements (MGE) are those with matching methylation profiles to the specified methylation bin.

	Binning statistics				**Annotation**	**Bin validation**		**Methylation summary**		
Bin	**Num. 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	4027504	1128400	1089244	*Bacteroidales* (97.5)	98.68	2.26	ACCG**A**G CC**A**SNNNNNNNATGT	1.85 2.01	12.7kb plasmid, 19.1kb conjugative transposon
2	9	3496584	2164130	2164130	*Bacteroidales* (97.1)	77.48	2.01	CTGC**A**G	2.43	None found
3	7	3853295	2087314	2087314	*Bacteroidales* (98.0)	99.43	1.13	TC**A**GNNNNNCCTC CC**A**GNNNNNNVTGG CC**A**GNNNNNNRTGG	1.62 2.22 2.50	None found
4	5	2759439	2712836	2712836	*Verrucomicrobiales* (98.3)	97.96	0.68	G**A**TTNNNNNCAGT G**A**TTNNNNNNAGT	3.11 2.93	None found
5	10	3378404	1873721	1873721	*Bacteroidales* (100.0)	97.55	1.76	AGC**A**NNNNNNRTC G**A**CNNNNNNTGCT	1.98 2.27	None found
6	16	4441324	1159367	764722	*Bacteroidales* (100.0)	98.36	1.26	ATGC**A**T CC**A**NNNNNTCG AAC**A**GC	1.76 1.93 2.80	None found
7	22	6207805	2165375	1643203	*Bacteroidales* (98.5)	98.24	21.52	GGC**A**GC GTG**A**TG	2.22 2.00	24.7kb plasmid, 14.7kb plasmid, 23.2kb conjugative transposon
8	14	3913657	2565370	2565370	*Bacteroidales* (98.2)	97.22	2.77	AG**A**TGA AG**A**TG G**A**TGGY AG**A**TGT	2.21 1.94 1.94 1.72	14.3kb plasmid, 15.8kb plasmid, 21.1kb conjugative transposon

Binning statistics					Annotation	Bin validation		Methylation summary		
Bin	Num. 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
								KAG**A**TG	2.08	
								TAG**A**TG	1.96	
								TG**A**TGG	1.71	
								G**A**TGG	1.81	
9	1	2021078	2021078	2021078	*Clostridiales* (100.0)	99.19	0.00	CGA**A**G	2.46	None found
								GA**A**GNNNNNACGT	2.18	
								TGM**A**GG	2.48	
								CG**A**GNNNNNNCCTT	1.69	
								ACC**A**TC	2.20	