Supplementary Methods

Additional details of methylated motifs detection in metagenomic sequencing
An initial motif-filtering step is necessary to reduce the space of motifs down to only those that have a significant methylation score in the metagenomic mixture. First, due to memory considerations and because a motif could theoretically describe any arbitrary string of bases, we must define the maximum motif length and allowable base configuration of motifs in our initial query space. For this study, we considered all possible 4mers, 5mers, and 6mers, for a total of 7,680 contiguous motifs. For bipartite motifs, where a string of non-specific Ns is bookended by sets of specific bases (e.g. CCANNNNCCAT), we considered several common configurations often found in prokaryotes. All combinations of the following were considered: 3 or 4 specific bases (beginning), 5 or 6 non-specific Ns (middle), and 3 or 4 specific bases (end). This adds an additional 194,560 possible bipartite motifs to space of motifs to consider for the initial filtering step, for a total of 202,240 motifs. The exact same method can be used to further incorporate 7-mer and 8-mer motifs.

Next, we reduce the motif query space by randomly sampling a small number of reads (N=20,000) from the mixture and removing from further analysis all motifs that do not return a methylation score above a chosen threshold (1.7) on at least one contig in the assembly (or on at least twenty unaligned reads for read-level binning). Despite choosing a lenient threshold to include many variations of the truly modified motif, this typically reduces the number of motifs to be included in the further analysis by multiple orders of magnitude. A further step searches for multiple specifications representing a single degenerate motif that, if identified, replaces the individual specifications in the final set of motifs. The remaining motifs need not exactly match the most parsimonious versions of the methylated motifs, but they nonetheless will carry some methylation signature that is useful for binning the sequences through subsequent dimensionality reduction analysis. Put another way, the number of motifs that remain after filtering is not usually critically important as long as the set of remaining motifs captures the most significant differences between methylation profiles. This property contrasts with existing methods for methylation motif discovery that attempt to identify the single most parsimonious version of a motif.

Power analysis of contig methylation classification
In order to assess the power of methylation scores to distinguish a contig methylated at a motif sites (case) from a contig that is not methylated at that motif (control), we sampled 15,000 normalized IPD (nIPD) values from GATC sites on each of two large assembled contigs from our mixture of eight bacterial species. The case was the 4.6Mb contig representing the E. coli chromosome, while the second 0.7Mb contig (control) represents a large assembled portion of
the R. gnnavus genome, which does not contain any methylated motifs based on SMRT sequencing data (Supplementary Table 2). We then used the two sets of 15,000 nIPD values as pools from which to sample 2, 4, 6, and 8 values for both the case and control. The nIPD values were used to construct methylation scores for GATC on both the case and control contigs, for each of the four specified nIPD sampling numbers (2, 4, 6, and 8). This process was repeated 10,000 times to create a receiver operating characteristic (ROC) curve (Fig. 2a) showing the effect of the number of nIPD values on creating methylation scores that can distinguish a methylated contig/motif from a non-methylated contig/motif.

Genome size parameter values used for HGAP3 assembly
The expected genome size parameter is used to determine the optimum number of long seed reads and was adjusted based on the expected complexity of the metagenome. Specifically, the genome size was set to 40Mb for the synthetic mixture of eight bacterial species assembly, 100Mb for the adult mouse gut microbiome assembly, 66Mb for the 20-member HMP assembly, 20Mb for the combined infant gut microbiome samples A and B assembly, 1.6Mb for the combined and separate H. pylori strain assemblies, 4.6Mb for the combined and separate E. coli strain assemblies, and 20Mb for the infant gut microbiome sample A assembly.

Genome-genome similarity
To assess the sequence similarity between two reference genomes, average nucleotide identity (ANI) was calculated using the web-based portal at http://enve-omics.ce.gatech.edu/ani/.

Length-weighted processing of large contigs with t-SNE
The long reads used in this study often result in a bacterial genome being represented by a small number of very large contigs. The t-SNE dimensionality reduction algorithm places data points in low-dimensional space based on the local similarities in the original high-dimensional space. Species with few large contigs that are represented by only a few points in the high-dimensional space do not contribute significantly to the objective function of the t-SNE algorithm. To adjust for this bias from different contig sizes, we use a length-weighted representation of all large contigs over 50 kbp in length so that each large contig is represented in the matrix of features not by one row, but by N rows, where N is the contig length divided by 50 kbp (i.e. the number of digested 50kb 'sub-contigs'). The features (column values) for each 50 kbp sub-contig, either k-mer frequency or methylation scores, are the same values that were computed for the original large contig.
Annotation of methylation bins in adult mouse gut microbiome

Querying the contig sequences in each bin against a manually curated set of 591 publicly available mouse gut microbial references\(^3-6\) (Supplementary Table 11) revealed significant reference hits with eight of the nine bins (Supplementary Fig. 13; Supplementary Table 6), providing further support that the bins identified using methylation profiles represent the genomes of distinct organisms. Bin4 and bin5 have high-quality, nearly full-length matches with the finished genomes for Akkermansia mucinophilia YL-44 (average nucleotide identity (ANI) = 98.94\%) and Parabacteroides sp. YL-27 (ANI = 98.43\%), respectively\(^5\). The remaining six bins have high-quality matches with genome assemblies of species that have been identified in the mouse gut in other studies but lack finished reference sequences. Three of these six bins have full-length matches with three draft assemblies of uncultured members of the Bacteroidales S24-7 family: bin1 matches Bacteroidales bacterium M1 (ANI = 98.63\%), bin3 matches Bacteroidales bacterium M12 (ANI = 98.45\%), and bin8 matches Bacteroidales bacterium M2 (ANI = 98.24\%)\(^4\). The final three bins have high-quality matches with three unidentified metagenomic species (MGS) previously binned in a large study of mouse gut microbiomes\(^3\): bin2 matches MGS:0161 (ANI = 99.41\%), bin8 matches MGS:0004 (ANI = 99.38\%), and bin9 matches MGS:0305 (ANI = 99.96\%).

Because the only other family of Bacteroidales identified in the sample by 16S sequencing was the family Rikenellaceae at 2.12\% abundance, it is likely that these seven highly contiguous genome bins all belong to the poorly characterized S24-7 family of Bacteroidales that dominated the 16S abundance profile for the sample (Fig. 2d). We did observe quality alignment of the bin5 contigs to the reference for Parabacteroides sp. YL-27, which is classified as belonging to the closely related Bacteroidales family Tannerellaceae, but there is some apparent divergence in the alignment that raises doubts about it being an exact match (Supplementary Fig. 13).

Binning mouse gut assembly with CONCOCT

bowtie2\(^7\) was used to align reads from 100 publicly available mouse gut microbiome sequencing data sets\(^3\) to the contigs assembled from the adult mouse gut microbiome (Supplementary Table 3). CONCOCT\(^8\) was used to call bins of these contigs using normalized coverage values from the alignments and 4-mer frequency features.

Re-assembly of sequences in each methylation bin

In each methylation bin, the reads aligning to each binned contig were re-assembled with the HGAP3 assembler\(^9\) using a genomeSize parameter modified to reflect the total number of contig bases in each bin.
Integrating methylation and composition for strain resolution in infant gut microbiome

We used epigenetic information to segregate contigs assembled from highly similar strains that would be otherwise indistinguishable using k-mer frequency-based methods. We examined two sets of infant gut microbiota obtained from stool samples of children who were selected for sequencing based on a high genetic risk for development of T1D.

Interestingly, it has been observed that the particular species of Bacteroides that dominates the composition of both samples, Bacteroides dorei, often spikes in relative abundance prior to onset of T1D in children, making it an important species to understand and potentially monitor during early adolescence. 16S sequencing showed that the two samples contained two distinct strains of B. dorei: Sample A consisted of 63.7% B. dorei str. 105 (CP007619), while Sample B contained 47.9% B. dorei str. 439 (CP008741). Despite a high sequence similarity of 99.43% ANI between the two B. dorei strains (Online Methods), each strain has a unique set of methylated sequence motifs and therefore a unique methylation profile.

We collected SMRT sequencing data for the two microbiome samples from a previous study (Supplementary Table 2) and performed a metagenomic de novo assembly (Supplementary Table 3) using a combination of both gut samples to generate a mixture of contigs from both B. dorei strains in our output set of metagenomic contigs. Lacking any labeling for these contigs, we applied the sequence annotation tool Kraken for labeling of all non-B. dorei contigs and an alignment-based labeling approach for distinguishing the two B. dorei strains (Online Methods). We first conducted composition-based binning using 5-mer frequency profiles, followed by t-SNE dimensionality reduction (Supplementary Fig. 4a). The map has five distinct clusters of contigs, four of which consist primarily of a combination of contigs from multiple species or strains. This suggests that composition-based binning is insufficient to segregate the two strains of B. dorei due to their high sequence similarity. Notably, composition-based binning also fails to segregate Bacteroides fragilis from Bacteroides thetaiotaomicron, Bifidobacterium breve from Bifidobacterium longum, and Alistipes finegoldii from Alistipes shahii.

Motif filtering identified seven motifs with significant methylation scores on at least one contig in the assembly: GGATCA, GATCA, TTGCAGA, GATC, CTCAT, GAATC, and GGATC. The resulting t-SNE map constructed using methylation profiles alone (Supplementary Fig. 4b) resolves the contigs into four clusters. In contrast to the k-mer frequency-based map and as a consequence of their unique methylation profiles, the two strains of B. dorei are very well segregated in the methylation-based binning analysis. However, methylation-based binning
alone did not fully segregate all other species due to an insufficient diversity of methylated motifs among them. This suggests that both methylation-based and composition-based binning methods can complement each other to compensate for the shortcomings of each approach. By combining k-mer frequency and methylation profiles, both reduced separately by t-SNE to 2D, into a single matrix with four columns, we can again use t-SNE to reduce the matrix and generate a 2D scatter plot (Supplementary Fig. 4c). This approach succeeds in separating the two strains of *B. dorei* from each other, *B. fragilis* from *B. thetaiotaomicron*, and *B. breve* from *B. longum*. Only the two species from the Alistipes genus remain convoluted in the combined map, due to high sequence similarity and likely identical methylomes. Again using a silhouette coefficient to assess the contig clustering, we find that while composition-based binning alone results in a silhouette coefficient of 0.03, the integration with methylation-based binning increases the coefficient to 0.41, demonstrating that contig methylation profile can help deconvolute contigs with high sequence similarity.

**Labeling *Bacteroides dorei* contigs by strain**  
In the infant gut microbiome t-SNE maps showing the combined assemblies of samples A and B (Supplementary Fig. 4a-c), all contigs other than those labeled as belonging to *B. dorei* were annotated using Kraken. The contigs belonging to the two *B. dorei* strains, however, were manually labeled by first aligning the reads from the combined samples to the fully assembled references for each *B. dorei* strain (strain 105: CP007619; strain 439: CP008741). The contig-labeling assignments were determined by examining the reads aligning to the either of the *B. dorei* references and counting how many of these reads aligned to each of the assembled contigs. For example, if the majority of the reads aligning to a contig aligned to the strain 105 reference, the contig was labeled as belonging to strain 105. However, if the majority aligned to the strain 439 reference, the contig was labeled as belonging to strain 439.

**Read-level 5-mer frequency binning of HMP mock community B**  
Sequence composition features and t-SNE have been used as features for metagenomic binning of assembled contigs\(^4\), but it was unclear whether this would also apply to long, unassembled SMRT reads that have a high single-pass error rate (typically ~13% for SMRT sequencing)\(^5\). To evaluate this, we used a modified version of a synthetic 20-member Mock Community B created as part of the Human Microbiome Project (HMP)\(^6\), where the DNA was mixed prior to sequencing and the reads were downsampled to create a more realistic abundance profile (Online Methods; Supplementary Fig. 14; Supplementary Table 2).

We first visualized and annotated (Online Methods) just the contigs in the t-SNE map of 5-mer frequency features for both unassembled reads and assembled
contigs, revealing a clean separation of contigs from species for which there is a significant number of assembled bases (Supplementary Fig. 7). We observed similar results using contigs assembled from the original Mock Community B with an unmodified, even abundance profile, indicating that contig separation was not biased due to poor assembly results of the low-abundance species (Supplementary Fig. 15). Next, we visualized and annotated the unassembled reads in the t-SNE map of 5-mer frequency features using Kraken (Online Methods), showing read clusters are highly species-specific and highly resilient to the random sequencing errors. Despite being poorly represented in the set of assembled contigs, *Rhodobacter sphaeroides* is clearly present as a distinct cluster of unassembled reads (Fig. 4a). We confirmed that read-level composition features stabilize with additional sequence length, finding that shorter reads generated less distinct clusters than longer reads (Supplementary Figs. 8a,b). Furthermore, an unlabeled 2D histogram provides an overview of global community complexity absent any sequence annotation (Fig. 4b), making it possible to identify a set of novel sequences from a particular cluster for further investigation.

**Comparison with synthetic long reads**
Recent advances in library preparation protocols for Illumina sequencing have made it possible to generate synthetic long reads of several kilobases in length. The read lengths of synthetic long reads can approach those generated by SMRT sequencing, yet important differences between the technologies have implications for their specific applications in metagenomics and therefore warrant a detailed investigation. Because the capability to infer methylation events is a unique strength of SMRT sequencing as studied above, we instead focus here on other aspects of the two techniques and emphasize their potential complementarity.

The read lengths and high accuracy of synthetic reads have enabled researchers to phase substrain-level bacterial haplotypes in metagenomic samples. By aligning synthetic long reads to contigs generated through de novo metagenomic assembly, the study revealed the presence of multiple genotypes within the same strain. A prerequisite for substrain haplotyping with synthetic long reads is a metagenomic assembly that serves as a reference for the read alignment. Kuleshov *et al.* acknowledge that SMRT reads are more likely to result in large draft assemblies, and indeed point out that contigs assembled from SMRT reads are significantly larger than those assembled using synthetic long reads, even when the latter was supplemented by traditional short reads.

Given the multi-kb read lengths and high accuracy of synthetic long reads, we sought to understand why they resulted in more fragmented and less comprehensive assemblies than did SMRT reads. To this end, we aligned both the synthetic long reads sequenced from the 20-member HMP Mock Community B (staggered abundance; HM-277D) and the SMRT reads from the same community to their reference genomes. The microbial
DNA HM-277D (HMP Mock Community B) was obtained from BEI Resources and was sequenced in a previous study (SRR2822454) by Kuleshov et al.\textsuperscript{17} using the Illumina TruSeq protocol.

Because the SMRT reads were sequenced from a different version of the HMP Mock Community B (even abundance; HM-276D), we downsampled the aligned reads so that total numbers of aligned bases for each organism were roughly equal for both sequencing technologies. The HMP Mock Community B SMRT reads (\textit{Supplementary Table 2}) were aligned using \textit{blasi}\textsuperscript{18} with default parameters and “-bestn 1 -sam” options. The synthetic long reads were aligned using \textit{bwa-mem}\textsuperscript{19} with default parameters. The aligned reads were analysed to count the total number of bases aligned to each reference for each technology. For each reference, the smaller number of aligned bases was chosen as the target number of aligned bases and the file with the larger number of aligned bases was selected for downsampling. The target fraction is calculated by dividing the target number of aligned bases by the original number of bases. The following \textit{samtools}\textsuperscript{20} command was used to generate the downsampled file:

\begin{verbatim}
    samtools view -s 1.[target_frac] -h -b original.bam > downsampled.bam
\end{verbatim}

The results of this downsampling are summarized in \textit{Supplementary Table 10}.

Despite considering approximately the same number of aligned bases for each technology, SMRT reads covered a higher percentage of genome positions in 17 of the 20 species and matched the percentage of genome positions covered by synthetic long reads in the remaining three species (\textit{Supplementary Fig. 10a}; \textit{Supplementary Table 10}). In several cases, the increases in genome coverage over synthetic long reads were dramatic: SMRT sequencing of \textit{D. radiodurans}, \textit{A. odontolyticus}, \textit{E. faecalis}, and \textit{S. pneumoniae} covered an additional 67.1\%, 69.2\%, 90.0\%, and 91.2\% of their genomes, respectively. The genomes with the highest GC-content (\textit{R. sphaeroides}, 68.8\% GC; \textit{D. radiodurans}, 66.6\% GC; \textit{P. aeruginosa}, 66.6\% GC; \textit{A. odontolyticus}, 65.4\% GC) were among those that saw significant increases in genome coverage with SMRT reads compared to synthetic long reads (\textit{Supplementary Table 10}). This observation is consistent with previous studies showing that the PCR amplification of DNA fragments required for synthetic long read sequencing is sensitive to genomic GC-content and can result in significant coverage biases (i.e. highly non-uniform sequence coverage)\textsuperscript{17,21,22}.

SMRT sequencing, however, is an amplification-free protocol and is not subject to GC bias\textsuperscript{23,24}, resulting in more uniform coverage profiles across genomes (\textit{Supplementary Fig. 10b}). Further illustrating this phenomenon are three small regions from the genomes of \textit{S. agalactiae}, \textit{S. aureus}, and \textit{P. aeruginosa} (\textit{Supplementary Fig. 11}), which are representative of many of the genomes in the mock community (\textit{Supplementary Fig. 12}). The synthetic long reads coverage profiles consist of peaks and valleys, representing over- and under-amplified DNA fragments, respectively. Some of the valleys result in complete
coverage dropouts, across which genome assembly becomes impossible. The SMRT sequencing protocol, on the other hand, results in much more uniform coverage profiles and fewer coverage dropouts, making it more amenable to metagenomic assembly and more likely to result in chromosome-scale contigs.

Two additional sources of systematic error in the synthetic long reads, resulting from dilution and sub-assembly steps in the protocol, make it more difficult to assemble high abundance species and regions containing tandem repeats. These steps are unique to synthetic long reads and do not apply to SMRT sequencing, which might further contribute to the superiority of SMRT reads for generating large metagenomic assemblies. The strengths of synthetic long reads, however, lie in their ability to call (and phase) local genomic features, such as single nucleotide variants (SNVs) or short insertions and deletions. Overall, this suggests a complementary strategy for maximizing assembly quality with SMRT sequencing and leveraging synthetic long reads for variant calling and haplotyping.

Supplementary References

1. Rodriguez-r, L. M. & Konstantinidis, K. T. The enveomics collection: a toolbox for specialized analyses of microbial genomes and metagenomes. *PeerJ Prepr.* (2016).
2. van der Maaten, L. & Hinton, G. Visualizing Data using t-SNE. *J. Mach. Learn. Res.* 9, 2579–2605 (2008).
3. Xiao, L. et al. A catalog of the mouse gut metagenome. *Nat. Biotechnol.* 33, 1103–1108 (2015).
4. Ormerod, K. L. et al. Genomic characterization of the uncultured Bacteroidales family S24-7 inhabiting the guts of homeothermic animals. *Microbiome* 4, 36 (2016).
5. Uchimura, Y. et al. Complete Genome Sequences of 12 Species of Stable Defined Moderately Diverse Mouse Microbiota 2. *Genome Announc.* 4, 4–5 (2016).
6. Wannemuehler, M. J., Overstreet, A., Ward, D. V & Phillips, J. Draft Genome Sequences of the Altered Schaedler Flora, a Defined Bacterial Community from Gnotobiotic Mice. *Genome Announc.* 2, 1–2 (2014).
7. 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).
8. Alneberg, J. et al. Binning metagenomic contigs by coverage and composition. *Nat. Methods* 11, (2014).
9. Chin, C.-S. et al. Nonhybrid, finished microbial genome assemblies from long-read SMRT sequencing data. *Nat. Methods* 10, 563–9 (2013).
10. Leonard, M. T. et al. The methylome of the gut microbiome: disparate Dam methylation patterns in intestinal Bacteroides dorei. *Front. Microbiol.* 5, 361 (2014).
11. Kukko, M. et al. Dynamics of diabetes-associated autoantibodies in young children with human leukocyte antigen-conferred risk of type 1 diabetes recruited from the general population. *J. Clin. Endocrinol. Metab.* **90**, 2712–2717 (2005).

12. Davis-Richardson, A. G. et al. Bacteroides dorei dominates gut microbiome prior to autoimmunity in Finnish children at high risk for type 1 diabetes. *Front. Microbiol.* **5**, 1–11 (2014).

13. Wood, D. E. & Salzberg, S. L. Kraken: ultrafast metagenomic sequence classification using exact alignments. *Genome Biol.* **15**, R46 (2014).

14. Laczny, C., Pinel, N., Vlassis, N. & Wilmes, P. Alignment-free Visualization of Metagenomic Data by Nonlinear Dimension Reduction. *Sci. Rep.* **1**–12 (2014). doi:10.1038/srep04516

15. Dröge, J. & Mchardy, A. C. Taxonomic binning of metagenome samples generated by next-generation sequencing technologies. *Brief. Bioinform.* **13**, 646–655 (2012).

16. Consortium, T. H. M. P. Structure, function and diversity of the healthy human microbiome. *Nature* **486**, 207–214 (2012).

17. Kuleshov, V. et al. Synthetic long-read sequencing reveals intraspecies diversity in the human microbiome. *Nat. Biotechnol.* **34**, 64–69 (2015).

18. Chaisson, M. & Tesler, G. Mapping single molecule sequencing reads using basic local alignment with successive refinement (BLASR): application and theory. *BMC Bioinformatics* (2012).

19. Li, H. & Durbin, R. Fast and accurate long-read alignment with Burrows-Wheeler transform. *Bioinformatics* **26**, 589–95 (2010).

20. Li, H. et al. The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078–9 (2009).

21. Kuleshov, V. et al. Whole-genome haplotyping using long reads and statistical methods. *Nat. Biotechnol.* **32**, (2014).

22. McCoy, R. C. et al. Illumina TruSeq synthetic long-reads empower de novo assembly and resolve complex, highly-repetitive transposable elements. *PLoS One* **9**, (2014).

23. Shin, S. C. et al. Advantages of Single-Molecule Real-Time Sequencing in High-GC Content Genomes. *PLoS One* **8**, (2013).

24. Chaisson, M. J. P. et al. Resolving the complexity of the human genome using single-molecule sequencing. *Nature* **517**, 608–611 (2015).
Supplementary Figure 1: Binning contigs from 8-species mock community. (a) t-SNE scatter plot of 5-mer composition profiles for contigs and (b) scatter plot of contig GC-content vs. contig coverage.
Supplementary Figure 2: Shorter contigs contain fewer methylated motif sites. After de novo assembly of reads from a mixture of eight bacterial species, the contigs belonging to C. bolteae were isolated. As the contig length decreases, it becomes less common for the contig to contain IPD values from the full diversity of motif sites that are methylated in C. bolteae, making it increasingly difficult to segregate smaller contigs based on contig methylation patterns alone.
Supplementary Figure 3: Composition and coverage-based binning methods applied to adult mouse gut microbiome assembly. (a) Contig GC-content vs. coverage for adult mouse gut microbiome assembly, and (b) contig coverage plotted against the contig coverage using sequencing from a related sample.
Supplementary Figure 4: Infant gut microbiome contigs binned by sequence composition and methylation profiles. (a) t-SNE map of 5-mer frequency features for contigs assembled from a mixture of two infant microbiome samples. Several clusters contain a mixture of species from the same genus. (b) t-SNE map of methylation features for the same contigs. (c) t-SNE map of the same contigs binned by both 5-mer frequency and methylation profiles (Online Methods), which resolve the contigs into mostly species-specific clusters. Kraken annotation relies on an existing reference database (Online Methods) and is therefore incomplete; contigs not generating a database hit are marked Unlabeled. Contigs <10kb are omitted.
Supplementary Figure 5: CONCOCT bins of the mouse gut microbiome. Taxonomic composition of the 29 bins identified by CONCOCT in the mouse gut metagenomic assembly. Taxonomy is based on contig-level annotations by Kraken.
Supplementary Figure 6: Heatmaps of methylation profiles for *K. pneumoniae*. (a) Hierarchical clustering of all known methylated motifs in REBASE for *K. pneumoniae* strain 234-12 and nine other species whose chromosomes have smaller sequence distance to the *K. pneumoniae* strain 234-12 plasmid (horizontal red bars) than its own host chromosome. (b) Hierarchical clustering of all motifs in REBASE for 25 strains of *K. pneumoniae*. The strains contain 17 unique methylation motifs, including CCAYNNNNNTCC that is observed solely in *K. pneumoniae* strain 234-12.
Supplementary Figure 7: Sequence composition t-SNE map of modified HMP mock community B. 5-mer frequency-based binning of assembled contigs and raw reads (length>15kb) from the log-abundance HMP mock community. Only the contigs are labeled (raw reads represented underneath contigs by density map) and the sum of assembled bases for each Kraken-annotated species is included in the legend.
Supplementary Figure 8: 5-mer frequency-based binning of unaligned reads from the modified HMP mock community. (a) read lengths between 5-10kb, and (b) read lengths between 10-15kb. The shorter read lengths result in more diffuse and overlapping clusters due to the increased variation in 5-mer frequency metrics on these shorter reads.
Supplementary Figure 9: t-SNE map of read-level methylation profiles for two *H. pylori* strains. 2D map of reads from each of the *H. pylori* strains, 26695 and J99, analyzed in the multi-strain synthetic mixture. 2D map generated using t-SNE, where the only features used in dimensionality reduction are methylation profiles of the reads.
Supplementary Figure 10: Comparison of abundance-matched SMRT vs. synthetic long read (SLR) sequencing coverage. (a) Human Microbiome Project Mock Community B members in decreasing order of GC content in genome. The percentage of the reference positions covered by SLRs is consistently lower than the percentage covered by abundance-matched SMRT reads. (b) Coverage variation for alignments of abundance-matched SLR and SMRT reads. A significant number of bases in SLRs are aligned in the same regions, creating dramatic peaks in coverage. SMRT reads largely lack these peaks and have a more uniform coverage profile.
Supplementary Figure 11: Examples of uneven coverage in SLR. Uneven coverage by synthetic long reads in a 40 kb region of the *S. agalactiae* genome (top), a 40 kb region of the *S. aureus* genome (middle), and a 50 kb region of the *P. aeruginosa* genome (bottom).
Supplementary Figure 12: Genomewide coverage of SLR and SMRT reads for all genomes in HMP mock community B. Genome-wide coverage of abundance-matched synthetic long reads (red lines) and SMRT reads (blue lines). Regions with zero coverage are highlighted for synthetic long reads (pink) and SMRT reads (light blue).
Supplementary Figure 13: Reference matches for bins identified from methylation profiles in mouse gut microbiome. Dot plot visualizations created using mummerplot that show the top reference alignment for bins isolated from the mouse gut microbiome metagenomic assembly using only methylation profiles. See Supplementary Table 6 for details of these alignments and the matching reference sequences.
Supplementary Figure 14: Modified relative abundances in HMP mock community B.
Relative abundances of the 20-species in the Human Microbiome Project Mock Community B modified to follow a log-curve distribution.
Supplementary Figure 15: Sequence composition t-SNE map of unmodified HMP mock community B. 5-mer frequency-based binning of assembled contigs and raw reads (length>15kb) from the even-abundance HMP mock community. Only the contigs are labeled (raw reads represented underneath contigs by density map) and the sum of assembled bases for each Kraken-annotated species is included in the legend.