nature biotechnology

ARTICLES

https://doi.org/10.1038/s41587-021-01130-z

Generating lineage-resolved, complete metagenome-assembled genomes from complex microbial communities

Derek M. Bickhart[1,10], Mikhail Kolmogorov[2,10], Elizabeth Tseng[3], Daniel M. Portik[3], Anton Korobeynikov[4], Ivan Tolstoganov[4], Gherman Uritskiy[5], Ivan Liachko[6], Shawn T. Sullivan[6], Sung Bong Shin[7], Alvah Zorea[8], Victòria Pascal Andreu[9], Kevin Panke-Buisse[1], Marnix H. Medema[9], Itzhak Mizrahi[8], Pavel A. Pevzner[2] ✉ and Timothy P. L. Smith[7] ✉

Microbial communities might include distinct lineages of closely related organisms that complicate metagenomic assembly and prevent the generation of complete metagenome-assembled genomes (MAGs). Here we show that deep sequencing using long (HiFi) reads combined with Hi-C binning can address this challenge even for complex microbial communities. Using existing methods, we sequenced the sheep fecal metagenome and identified 428 MAGs with more than 90% completeness, including 44 MAGs in single circular contigs. To resolve closely related strains (lineages), we developed MAGPhase, which separates lineages of related organisms by discriminating variant haplotypes across hundreds of kilobases of genomic sequence. MAGPhase identified 220 lineage-resolved MAGs in our dataset. The ability to resolve closely related microbes in complex microbial communities improves the identification of biosynthetic gene clusters and the precision of assigning mobile genetic elements to host genomes. We identified 1,400 complete and 350 partial biosynthetic gene clusters, most of which are novel, as well as 424 (298) potential host–viral (host–plasmid) associations using Hi-C data.

The creation of reference-quality, species-level assemblies from metagenome communities is exceedingly difficult. In particular, generating a complete genome assembly of a microbe closely related to a more abundant member of the community has been an elusive goal. Previous short-read studies have resulted in high-quality MAGs[1] only after extensive polishing and manual curation of initial contigs[2]. However, if a community contains thousands of organisms at different levels of abundance, manual curation of each MAG to achieve reference quality is extremely laborious. The prevalence of structurally variant strains in a sample might also impede contiguous assembly of microbial chromosomes[3]. To compensate, multiple samples may be sequenced with the idea that the prevalence of a unique strain is higher in one sample, thereby allowing disambiguation from other lineages[3,4]. Recent metagenome assembly improvements based on long reads[5–7] or linked reads[8] have resulted in some successes in the creation of closed circle MAGs, but the goal of assembling all distinct members of complex microbial communities as circular contigs has been elusive. A major source of discontinuity in metagenome assembly appears to be from the prevalence of high-sequence-identity orthologous genes and operons that can inhibit complete assembly[8]. Furthermore, nearly all short-read and long-read assembly algorithms typically collapse the variant features into a single representation that does not reflect the true strain- or species-level diversity of a subpopulation within the community[9,10]. For example, the metaFlye assembler improved reconstruction of complex environmental metagenomes using long reads[10] but produced collapsed representations of similar bacterial strains. These consensus assemblies might include various artifacts arising from the variation collapsing procedure—for example, frame shifts—complicating downstream analysis[11]. Ambiguity resulting from the metagenomic assembly of both short reads and error-prone long reads[12] has precluded first-pass characterization of microbial strains.

Microbial lineages are frequently defined using standard taxonomy terms, but genetic variation within related populations of microbes can cause substantial differences in observed phenotypes, such as pathogenicity or virulence. Thresholds of average nucleotide identity (ANI) among sequenced genomes have been proposed to differentiate between species (97%) and strains (99.999%)[13], but these estimates are based on comparisons of shared sequence that might not always reflect actual inherited genomic DNA. We, therefore, adopted the term 'lineage', similarly to another study[14], to describe a clonal subpopulation derived from a single ancestral genotype that can be distinguished from other populations in the same sample using discrete haplotypes.

Binning methods were developed to address issues with assembly fragmentation and to organize contigs into candidate MAGs based on assumptions of shared sequence composition[15]

[1]USDA Dairy Forage Research Center, Madison, WI, USA. [2]Department of Computer Science and Engineering, University of California - San Diego, La Jolla, CA, USA. [3]Pacific Biosciences, Menlo Park, CA, USA. [4]Center for Algorithmic Biotechnology, St. Petersburg State University, St. Petersburg, Russia. [5]Amazon, Seattle, WA, USA. [6]Phase Genomics, Seattle, WA, USA. [7]USDA Meat Animal Research Center, Clay Center, NE, USA. [8]Department of Life Sciences and the National Institute for Biotechnology in the Negev, Ben Gurion University of the Negev, Beer Sheba, Israel. [9]Bioinformatics Group, Wageningen University, Wageningen, Netherlands. [10]These authors contributed equally: D. M. Bickhart, M. Kolmogorov. ✉e-mail: ppevzner@ucsd.edu; tim.smith2@usda.gov

Table 1 | Assembly quality statistics

Assembly	Contigs	Assembly length (Mbp)	Contig N50 (Kbp)	High-quality draft contigs[a]	Circular contigs[b]	Circular + high-quality draft contigs
HiFi	57,259	3,424	280	123	49	44
pCLR1	48,338	2,985	185	54	21	18
pCLR2	48,790	3,008	187	65	28	26
pCLR3	56,456	2,978	181	64	26	22

[a]Contigs that were determined to have more than 90% SCG completeness and less than 5% SCG redundancy. [b]Contigs larger than 1 Mbp in size that were predicted to be circular by the metaFlye assembler.

or orthologous linkage data[16]. The presence of single copy genes (SCGs) expected to be in all bacterial and archaeal lineages has been proposed as a measure of the completeness and redundancy within these bins[17]. High-quality draft MAGs have been defined as having over 90% of the expected count of SCGs with less than 5% redundancy of their prevalence and containing assemblies of ribosomal RNA (rRNA) and transfer RNA (tRNA) genes[1]. However, bacterial and archaeal lineages might contain substantial accessory gene content[18] that is not assessed using these metrics. Bins are often generalized to represent distinct microbial taxonomic units in a sample, but they are rarely assumed to accurately represent true, genetically distinct microbial populations in a sample[14,19], and precise definitions for individual, highly resolved MAGs remain contextual to each study. Similarly to one of these studies[14], we focused on generating separate representative reference genomes for distinct microbial lineages within an individual metagenome, which we define as 'lineage-resolved MAGs'[1,20]. We further extend the term to 'lineage-resolved complete MAGs' for assemblies that have high degrees of SCG completeness (>90%) and low degrees of SCG redundancy (<10%). Tools have been developed to identify or separate lineage-resolved complete MAGs from metagenomic bins post hoc, but these tools often rely on co-assembly data, assembly graphs or various statistical methods to overcome biases in read alignments to estimate strains from observed genetic variant data and, therefore, require more curation to properly disentangle lineages from MAGs[3,14,21,22]. Furthermore, these workflows are designed primarily to identify strain lineages from alignments of short-read data and do not capture variant linkage data from longer-read datasets. A recent attempt to adapt uncorrected long reads to this purpose requires the use of manual curation and a priori estimates of strain numbers to achieve optimal results[23]. An intuitive and automated method to generate lineage-resolved high-quality MAGs is needed for analysis of more complex metagenome communities to reduce the time required to validate results.

Highly accurate HiFi reads from the Pacific Biosciences platform have error rates below 1%[24], providing an opportunity to improve assembly quality[25] and resolve both haplotypes of diploid genomes[26,27]. They were used to generate the first complete human genome assembly and opened the era of 'complete genomics'[28]. This technology could be suitable for assembling the highly repetitive orthologous genomic features present in metagenomes into lineage-resolved high-quality MAGs, enabling 'complete metagenomics'. Here we present a proof-of-principle study for the application of HiFi sequencing to complex microbiomes using extremely deep sequencing of a fecal sample from a parasite-infested lamb, combined with Hi-C data from the same sample. We document and quantify the improvement in assembly of MAGs with HiFi reads and present a computational approach called MAGPhase to phase alternative single-nucleotide polymorphism (SNP) haplotypes in these MAGs to provide finer resolution of descendant lineage variation in the sample. We further show that HiFi assemblies greatly improve precision in assigning mobile genetic elements to host genomes and inference of complete biosynthetic gene clusters from metagenomic data.

Results

Assembly of the sheep gut microbiome. HiFi and short-read data were generated from a fecal sample of an adult sheep (Methods). The short and HiFi reads comprised 154 and 255 total Gbp in 1,024,375,790 and 22,118,393 reads, respectively, with the latter representing higher depth of coverage than most previous reports of long-read metagenome assembly. Classification of reads with Kaiju[29] revealed a slight decrease in representation of Gram-negative lineages in the HiFi dataset compared to short-read data, although the subreads from which the HiFi data were generated did not reflect this decrease, suggesting that the process of HiFi filtering might be responsible for the effect[30]. Assembly of HiFi reads with metaFlye resulted in a total of 57,259 contigs with a contig N50 of 279 kb, including 127 contigs that fit the criteria of a high-quality draft[1] MAG, among which 44 (35%) represented closed circles in the metagenome assembly graph (Table 1).

Complete MAGs enabled by assembly with HiFi reads. We hypothesized that the use of HiFi reads decreased ambiguity in resolving structural complexity in microbial genomes and led to improved assembly completeness. We confirmed this hypothesis using an experimental design (Extended Data Fig. 1a) that allowed for an apples-to-apples comparison of HiFi and CLR reads by extracting subreads from the original HiFi reads to generate a series of 'pseudo-CLR' (pCLR) datasets (Methods). The average pCLR contig was longer than the average HiFi contig in all superkingdoms except the eukaryotes (Extended Data Fig. 1b). However, the total assembly length of pCLR contigs was lower than the HiFi assembly in all categories except the unassigned, 'no-hit' lineage (Extended Data Fig. 1c). In the archaea and bacteria annotated contigs, the pCLR assemblies had an average of 61 high-quality draft genomes with an average of 22 predicted circular high-quality genomes, representing a 48% and 50% reduction, respectively, compared to the HiFi assembly (Table 1 and Extended Data Fig. 1d).

Binning the HiFi contigs with Hi-C linkage data (Methods) resulted in 428 complete MAGs (defined as MAGs with >90% SCG completeness and <10% SCG contamination), which is the largest number of reference-quality MAGs reported from a single sample, to our knowledge (Supplementary Table 1). Of the HiFi assembly complete MAGs, 319 fit all of the criteria for high-quality draft MAGs specified by Bowers et al.[1] (Supplementary Note, 'pCLR assemblies' provides further comparisons between pCLR and HiFi assemblies). A cumulative assembly length plot suggested that a larger proportion of complete MAGs in the HiFi dataset were of low relative abundance (with coverage below 10×) compared to MAGs in the pCLR assemblies (Fig. 1a). Comparisons of bin SCG completeness and average depth of coverage also indicated that the HiFi assembly had more low-coverage complete MAGs than the pCLR assemblies (Fig. 1b). The contrast between HiFi and pCLR assemblies was more pronounced in bins that had more than 90% SCG completeness (Fig. 1c), where the pCLR assemblies contained mainly bins with more than 10× coverage and as much as 1,000× coverage compared to the HiFi complete MAGs. The distribution of coverage for complete MAGs is consistent with the hypothesis

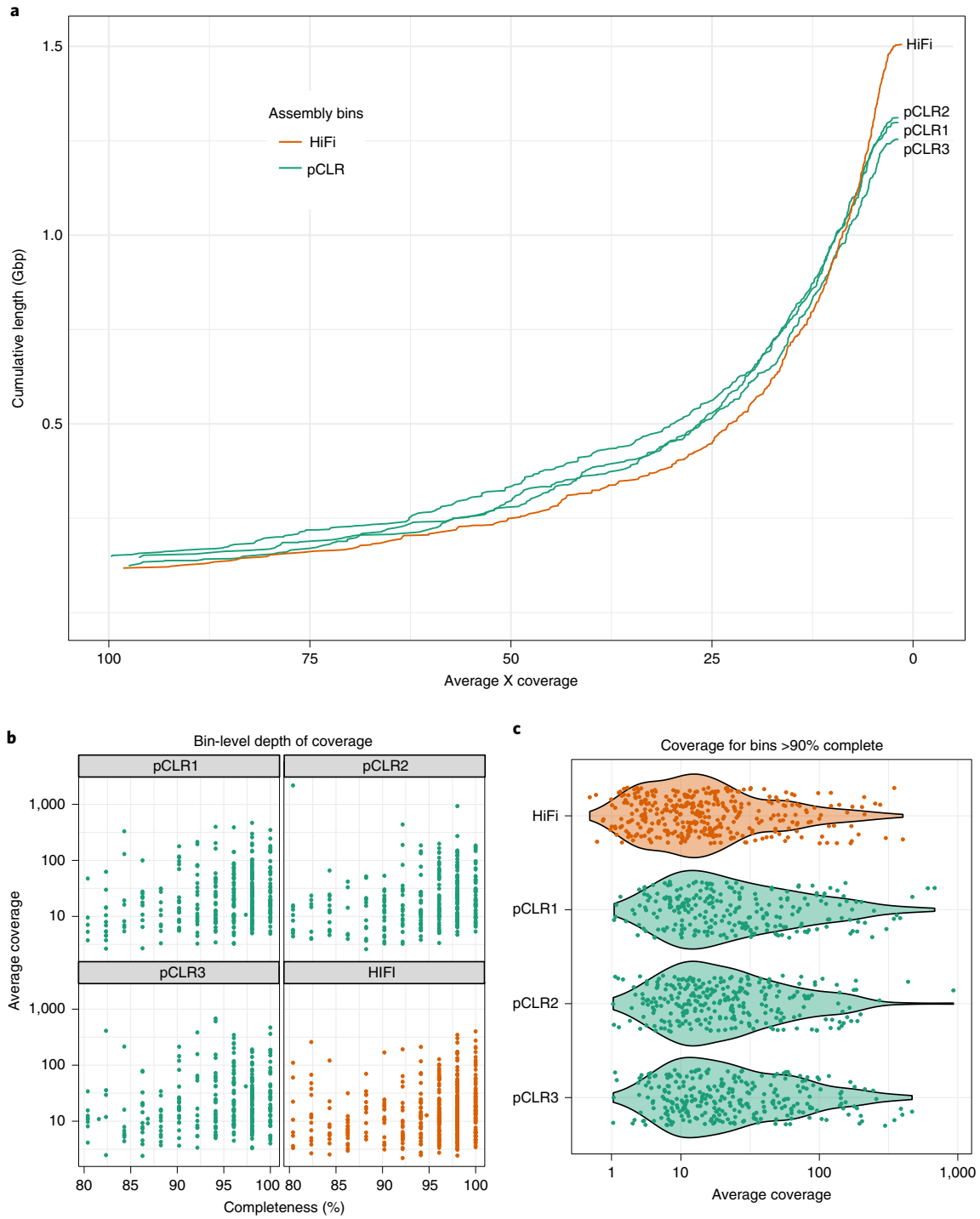


Fig. 1 | HiFi complete MAGs assembled low relative abundant species in the sample. Cumulative length of assembled HiFi bins. (**a**) peaks at lower depths of coverage at a faster rate than the cumulative lengths of pCLR bins, suggesting that lower-abundance taxa were more likely to be assembled by HiFi reads. Comparisons of average short-read coverage against SCG completeness estimates (**b**) for high-quality bins revealed a substantial number of HiFi bins below the 10× coverage threshold compared to the pCLR datasets. This is particularly enhanced in the >90% completion category, where the average coverage of the HiFi bins is lower than that of each pCLR assembly, and several HiFi bins have less than 1× average short-read coverage as opposed to pCLR bins with no equivalent coverage profile.

that HiFi assembly resolved pCLR bins into higher-resolution, lower-coverage bins that had been compressed into single bins in the pCLR assembly.

Identification of resolved lineages within MAG bins. We tested the hypothesis that HiFi metaFlye assembly had increased the total number of MAGs (as compared to pCLR assemblies) in part by separating distinct lineages into individual assemblies within metagenomes. HiFi and pCLR complete MAGs were classified into predicted phylogeny using GTDB-TK[31], resulting in 197 and 187 distinct genera classifications and 15 and 14 distinct phyla classifications, respectively (Extended Data Fig. 2). There were 22 genera

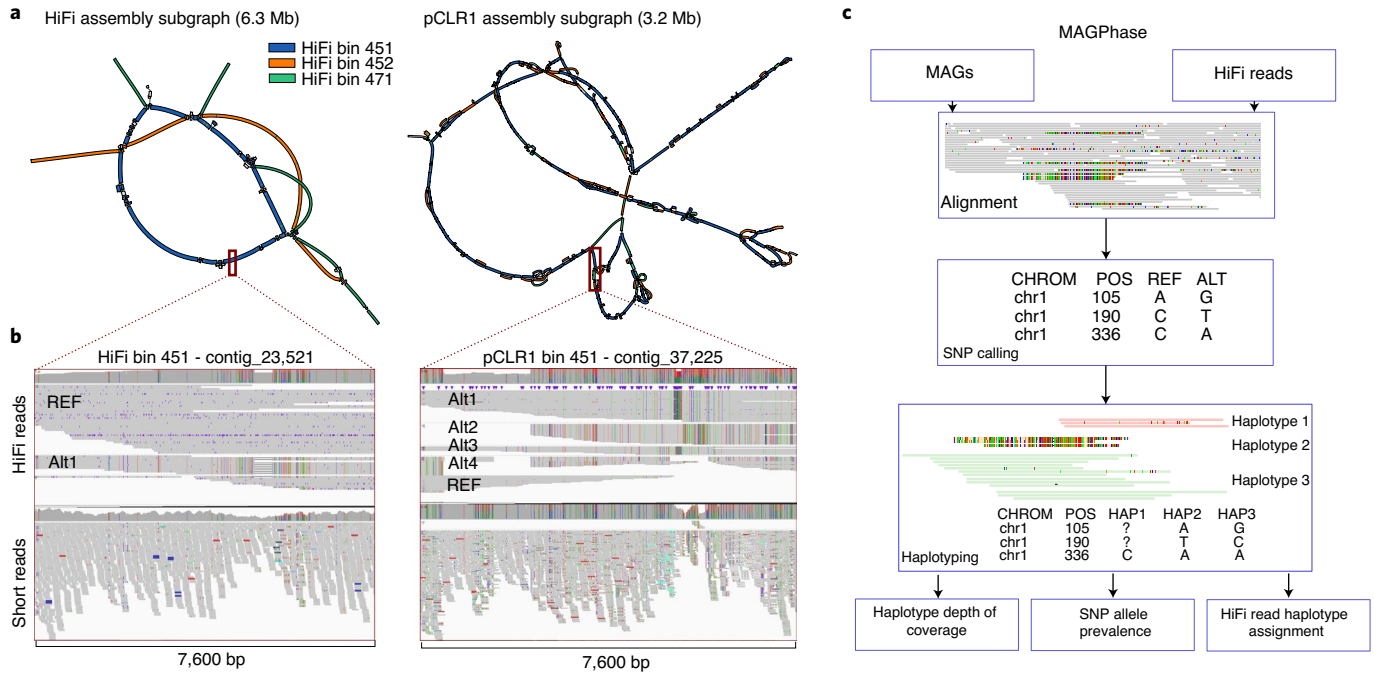


Fig. 2 | Phased SNP haplotype detection in metagenomes with MAGPhase. Lineage-resolved MAGs (**a**) in the HiFi assembly often corresponded to two or more compressed bins in the pCLR assemblies. In this example, we show comparative alignments of HiFi bins (colored according to the legend) on superset graphs of three HiFi bins (left-most graph) and a single pCLR1 bin (right-most graph). pCLR graph alignments show bifurcation and trifurcation of sequence into bubbles that were otherwise condensed in the final assembly. Dark red tinted boxes correspond to IGV plots in **b**. IGV plots of specific loci within these bins (**b**) show the power of this method to easily distinguish among haplotypes without the need for extensive statistical post hoc analysis. Comparative alignments of HiFi reads to HiFi bin 451 show only one alternate allele, whereas the equivalent region in the pCLR1 bin 451 shows as many as four alternate alleles (labeled on the figures). To automate this analysis in future datasets, we provide the MAGPhase software package (**c**). MAGPhase identifies candidate SNP variant sites within MAGs and phases them into longer haplotypes using the length and high accuracy of HiFi reads.

unique to the HiFi dataset, compared to eight among all three pCLR datasets, and one phylum unique to HiFi bins (Supplementary Figs. 1–3). Several cases where the HiFi assembly had more assembled bins for a taxon than the pCLR assemblies were also identified (Supplementary Table 2), including the *Clostridia* class, which had a single bin in the pCLR MAGs but three assembled bins in the HiFi assembly. These three bins had estimated average nucleotide identity (eANI) between 93% and 95%, suggesting that they are separate assemblies of related organisms within this class and possibly represent different species within genus or strains within species (Supplementary Tables 3 and 4). Separation of MAGs within the HiFi dataset was evident in comparisons of alignments of contigs to the assembly graphs (Fig. 2a), which identified heterogenous regions of alignment in the pCLR collapsed MAGs and in Mash[32] *k*-mer profile comparisons that revealed that all three HiFi MAGs had greater than a 90% predicted eANI from a single bin in each pCLR assembly. This suggested that the pCLR assemblies had collapsed the distinct components of the separate HiFi MAGs into single bins. This conclusion was supported by read depth coverage plots, which indicated very uneven coverage in the pCLR bin, averaging approximately 45-fold consistent with a collapse of multiple related strains present at variable abundance (Extended Data Fig. 3). In contrast, the resolved HiFi bins had very even coverage at approximately 10×, 20× and 33× (Extended Data Fig. 4), which suggested that the assembly had resolved species- or strain-level lineages across this range of abundance. This outcome has considerable implications in the use of read coverage in resolving strain lineages from metagenomes.

A total of 15, 10 and 11 pCLR MAGs were found to be condensed orthologs of 31, 23 and 25 HiFi bins in the pCLR1-3 assemblies, respectively (Supplementary Figs. 4–6). We identified 18 MAGs within the HiFi assembly that are likely species- or strain-resolved assemblies using a nearest neighbor distance analysis with a low eANI pairwise distance cutoff (≥93% distance) compared to six in the pCLR assemblies (Supplementary Table 5). These HiFi MAGs had solitary representatives in the pCLR assemblies, suggesting that differences in sequence content and structural variation are likely to be lost in assemblies of error-prone reads.

MAGPhase separation of lineage-resolved high-quality MAGs. The demonstration that HiFi assembly can resolve some sublineages even at the stage of initial contigs motivated us to investigate whether we could resolve additional HiFi bins into lineage-resolved complete MAGs using SNP variant data, as attempted previously[21]. We identified MAGs that had SNP variation above that expected from read error rates within SCG regions. Alignments of short reads could not distinguish true polymorphic sites, particularly in highly repetitive or orthologous gene regions (Fig. 2b), so we developed a computational approach to resolve lineages in metagenomes. The ability to distinguish structural variant subtypes within a MAG required an ability to simultaneously consider depth of coverage and haplotype information. This problem has similarities to phasing isoforms of transcripts in the context of variable expression from parental alleles in gene expression studies, so we adapted the phasing algorithm of the IsoPhase workflow[33,34] into a tool called MAGPhase to identify SNPs, detect reads supporting these SNPs and use these reads for phasing SNPs across identified SCG regions in each MAG (Fig. 2c). To avoid potential false-positive SNP haplotypes due to errors in reads, we only call variants in SCG regions that have at least ten spanning HiFi reads and are prevalent at substantial proportions of read depth (Methods).

Table 2 | MAGPhase haplotyping results

	HiFi	pCLR1	pCLR2	pCLR3
Total complete MAGs	428	345	345	315
Average contig count per MAG	8.3	10.8	11.4	11.9
Zero haplotype MAGs[a]	220	130	136	89
Percent polymorphic MAGs[b]	48.5%	62.3%	60.6%	71.7%
Average haplotype variant length[c]	20.1	27.1	24.1	34.0
Average haplotype genomic length[d]	1,151.3	1,106.1	1,057.8	961.2
Maximum haplotype genomic length	336,899	463,082	480,257	493,333
Average haplotype alleles per locus[e]	4.18	4.72	4.43	5.06
Maximum haplotype alleles per locus	25	59	54	60

[a]MAG that did not have detectable SNP haplotypes that could be linked with HiFi reads. [b]Number of MAGs that had at least one detectable alternative SNP haplotype allele. [c]Haplotype variant length represents the count of polymorphic SNP loci within an identified haplotype. [d]Haplotype genomic length was defined as the distance in bases on the assembled contig from the first polymorphic SNP site to the final site. [e]A haplotype allele was defined as a unique permutation of polymorphic SNP variants that were found to be linked by direct observation on consensus overlap of HiFi reads.

Phased SNP haplotypes were identified in each target region, and the maximum number of haplotype alleles was counted for each MAG to assess the upper boundary for SCG variation in each MAG. Most MAGs in pCLR bins had multiple haplotype alleles (average of 219, 65% of total; Supplementary Table 6), suggesting that they represent confounded lineages. By contrast, most HiFi MAGs (220, two-fold more than pCLR assemblies) had zero identified alternate haplotype alleles, suggesting that many lineages were well resolved by the HiFi assembly or did not have detectable polymorphic subpopulations in the sample (Table 2). Polymorphic HiFi MAGs were found to exhibit as many as 25 unique haplotype alleles within SCG regions, suggesting that MAGPhase can identify localized regions of genetic drift. This is further supported by the fact that, among 48 HiFi haplotype loci with more than ten unique alleles, we found that 40% (122/305 haplotypes) differed from the original reference sequence by three or fewer bases, suggesting fixation of neutral mutations in subpopulations[35]. Median coverage of the alternative alleles in these hotspot regions was an average of five HiFi reads across the length of the haplotype, suggesting that these are likely true variable sites rather than artifacts of coincidental alignment of erroneous positions in HiFi reads to these loci. However, we cannot rule out the possibility that some putative variant sites result from inaccurate read alignment.

Comparisons of aligned short reads to polymorphic HiFi MAGs revealed limitations in the use of short reads for strain heterogeneity assessments. For example, alignment of the lineage-resolved Clostridia MAGs identified 7, 1 and 0 alternative haplotype loci on HiFi bins 451, 452 and 471, respectively (Extended Data Fig. 4). Clear variant patterns in individual HiFi reads aligning to these regions showed the power of using these reads for phasing haplotypes from metagenome bins (Fig. 2b). These patterns were not readily apparent or were heavily fragmented in the short-read alignments to the HiFi bins (Fig. 2b). Read pileups in lineage-resolved complete HiFi MAGs and orthologous pCLR collapsed MAGs were instructional in determining how read mapping could be used in downstream variant calling workflows. For example, visual determination of haplotypes was trivial when comparing orthologous regions of HiFi and pCLR MAGs 451 (Fig. 2b), where a large insertion identified by MAGPhase is clearly visible in read pileups in the region. The pCLR1 MAG has four distinguishable haplotype alleles that are difficult to resolve, consistent with the properties of a collapsed assembly, whereas HiFi MAG 451 can be separated into two lineage-resolved complete MAGs using these identified haplotypes. We identified 35 and 32 additional complete HiFi MAGs that had only 1 or 2 identified alternative SNP haplotypes that could be separated into an additional 70 and 96 lineage-resolved complete MAGs, respectively. However, 220 of our complete MAGs had zero identified haplotypes without any need for manual curation and, therefore, fit the criteria of lineage-resolved complete MAGs by default (hereafter referred to as 'de novo lineage-resolved'). The 220 de novo lineage-resolved complete MAGs show the lack of need for extensive post hoc editing. We emphasize that short-read alignments failed to consistently identify variants within identified haplotype alleles, regardless of the quality of the underlying MAG.

The paucity of consistent signal and the reduced power to link variants into haplotypes limits the use of short reads for variant phasing in metagenome communities. The prevalence of many ambiguous short-read alignments with a mapping quality score of 0 (MapQ0) in haplotype regions suggests that they are highly repetitive in the overall assembly and do not provide sufficient unique sequence for short-read alignment. The percentage of short-read MapQ0 alignments out of the total were 7%, 9% and 17% for bins 451, 452 and 471, respectively, suggesting that large portions of these bins would be otherwise intractable to short-read variant profiling. Analysis across the entirety of the HiFi assembly in 5-kb windows identified that 18% of the assembly is covered by windows that have ratios of MapQ0 alignments to total alignments greater than 0.5. Naturally occurring variation is unlikely to be detected in these windows by short-read alignments due to mapping ambiguity. By contrast, only ~2% of the length of the HiFi assembly had high MapQ0 windows, suggesting that 98% of the assembly contains sufficient unique sequence for HiFi read alignment (Supplementary Fig. 7).

Effect of HiFi accuracy on the generation of complete MAGs. The proportion of lineage-resolved complete MAGs in this complex sample was estimated by counting the number of HiFi reads aligning to each MAG completeness category of the HiFi assembly. These alignments revealed that 5.7% of all HiFi reads mapped to lineage-resolved complete MAGs, with 18% and 7% aligning to complete MAGs and other, lower-quality MAGs, respectively (Fig. 3a). Most reads (83%) mapped to contigs assigned to bacterial taxonomy (Fig. 3b), among which 7% of HiFi read alignments were to lineage-resolved complete MAGs. However, most alignments (63%) were to bins that did not meet minimal standards (>50% SCG completeness) for MAG generation. These data suggest that our de novo lineage-resolved complete MAGs do not represent the most abundant lineages in our dataset and might, instead, be stratified by homogeneity of genome sequence in the sample that enables these lineages to be more easily assembled at lower depths of coverage.

The depth of sequence required to characterize the genomes extracted from a complex sample is not usually known a priori. Our study design employed unusually deep HiFi sequencing to reduce the limitation of sequence depth on the analysis of the microbiome while supporting an analysis of the relationship between sequence depth and the generation of lineage-resolved complete MAGs. To explore this, we performed multiple assemblies of downsampled read sets. Total assembly size continued to increase up to the full dataset, possibly reflecting assembly of lower abundance microbes or eukaryote genomes, but the rate of accumulation diminished as the total depth was approached (Fig. 3). The number of total MAGs and complete MAGs continued to increase with depth, although the count of complete MAGs was less constrained by depth (45% of the count at 40-Gbp depth) (Supplementary Note, 'Analyzing effect

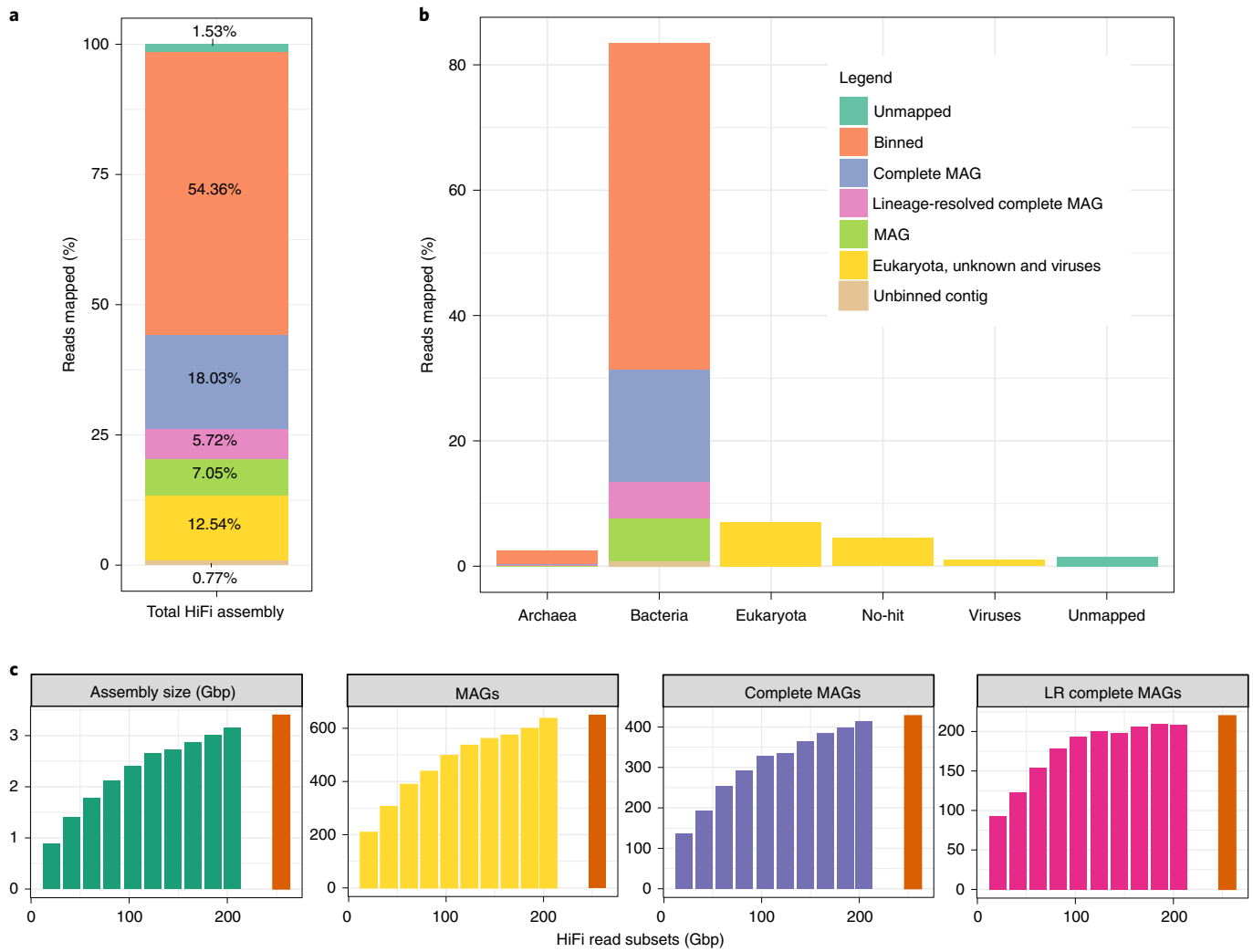


Fig. 3 | MAG representation and assembly at different depths of coverage. HiFi read alignment percentages (**a**) were tabulated based on the underlying quality category of MAGs. This revealed that a small proportion of the reads were unmapped (dark green) or mapped to unbinned contigs (brown). Most reads aligned to contigs that were placed into discrete bins by Bin3c but did not meet minimal completeness criteria (they had <50% SCG completeness) for analysis (orange). Alignments to lineage-resolved complete MAGs (pink) were a smaller proportion of total alignments than originally expected, with complete MAGs (>90% completeness and <10% contamination; blue) and other MAGs (>50% completeness; green) representing a substantial percentage of alignments instead. Breakdowns of read alignments to different contig-level taxonomic assignments (**b**) revealed that most lineage-resolved complete MAGs were of bacterial origin, as expected. Downsampling of HiFi reads in 20-Gbp portions (**c**) revealed that smaller target depths of coverage could still result in substantial counts of complete and lineage-resolved complete MAGs. Each downsampled dataset is compared to the full HiFi dataset (dark orange) in each category. LR, lineage-resolved.

of HiFi accuracy and sequencing depth'; Supplementary Tables 8 and 9). We also tested the efficacy of polishing HiFi assemblies using short reads, finding that it does not substantially improve sequence accuracy and suggesting that this step can be omitted in the future (Supplementary Fig. 8).

Improvements in functional genetics analysis. The advantages of HiFi reads in functional annotation of a metagenome were assessed by predicting biosynthetic gene clusters (BGCs) that are notoriously difficult to identify in fragmented assemblies[36]. We identified 1,400 complete and 350 partial BGCs in the HiFi assembly using antiSMASH[37]. To our knowledge, this represented the largest number of complete BCGs ever reported in metagenomic assemblies (Extended Data Fig. 5). Nearly all identified BGCs were classified as novel, illustrating the capabilities of long reads for exploration of novel natural products (Supplementary Note, 'Identification of biosynthetic gene clusters').

Improved resolution of mobile DNA association analysis. Contigs assigned to viral (Fig. 4a and Extended Data Figs. 6–8) or plasmid (Fig. 4d) taxonomy were evaluated for putative bacterial hosts using Hi-C links and partial long-read alignments by application of a previously published workflow[6], described in detail in Supplementary Note, 'Mobile DNA association analysis'. Using the SCAPP[38] plasmid assembly tool, we identified 5,528 candidate plasmid contigs within the HiFi assembly and identified 298 plasmid–contig associations. We predicted six plasmid contigs and 25 candidate bacterial hosts (Supplementary Fig. 1), in which one plasmid was predicted to inhabit members of 13 different bacterial genera, suggesting inter-genera mobility of this plasmid. We also predicted associations between plasmid contigs and three genera of archaea, including *Methanobrevibacter* and *Methanosphaera*, which were previously not known to carry naturally occurring plasmids[30]. These results underscore the value of combined HiFi assembly and Hi-C contact mapping for assignment of mobile genetic elements to host organisms.

Discussion

The goal of metagenome assembly is to generate reference genomes for the majority of organisms that comprise the sample. Our data suggest that both short and long error-prone reads produce collapsed assemblies that require extensive manual curation to resolve into reference-quality resources. By contrast, metaFlye assemblies using HiFi reads generate many lineage-resolved complete MAGs from complex metagenomes without the need for curation (Fig. 2), including organisms prevalent at lower relative abundance in the community (Fig. 1a,c). Compared to recent work using assemblies of long-read sequence data polished with short reads, we found proportional improvements in recovery of rRNA and tRNA genes (319 MAGs containing full complements in our dataset) and circular high-quality contigs (44 in our dataset) from a single sample without the need for curation[4]. These complete MAGs appear to be resolved with respect to structural variation and orthologous gene sequence compared to closely related (>90% eANI) lineages. Sketch-based comparisons revealed that HiFi MAGs (23–31 MAGs, 6–7% of total) that were condensed into collapsed assemblies in the pCLR datasets were found to be poor representatives of the actual genomic sequence of the organisms based on read alignment metrics and variant phasing analysis (Fig. 2b). The data suggest that metagenome assemblies of long error-prone reads most likely collapse contigs of closely related genomes present in the sample into single inaccurate representations.

MAGPhase detects discrete haplotypes and identifies variant lineages more efficiently and correctly than existing short-read-based strain-resolution algorithms that rely on multiple sample observations and statistical variant linkage analysis to determine potential microbial lineages[13,14]. HiFi reads provide suitable accuracy and length that enabled identification of phased haplotypes of up to 309 SNPs and phase variants across segments as large as 300 kbp in our HiFi MAGs (Table 2). Rather than limiting analysis of microbial lineages to ANI thresholds that might be biased due to short-read alignment inaccuracy, HiFi reads allow for detection of haplotypes segregating in a sample that have as low as 2% (five reads out of 300) relative abundance of the reference MAG haplotype. The MAGPhase workflow uses HiFi reads for haplotype analysis on metagenome assemblies (https://github.com/Magdoll/MagPhase) and enables visual scrutiny of discovered haplotypes, as illustrated by IGV alignment diagrams. We have included tools to label these diagrams and enable visual verification of predicted SNP haplotypes (https://github.com/njdbickhart/SheepHiFiManuscript). Even when using MAGs produced by long error-prone reads (pCLR assemblies) as a reference, MAGPhase can still produce discernable SNP haplotypes that could be used to identify descendant lineages (Fig. 2b), suggesting that it might have value applied to existing non-HiFi long-read assemblies. Recent work has attempted to realize strain-separated MAGs using post hoc, short-read linkage statistics dereplication[13,39]. This approach might be more cost-effective, but it comes at the cost of accuracy and effort in post hoc dereplication, whereas our method generates a large proportion of lineage-resolved MAGs de novo.

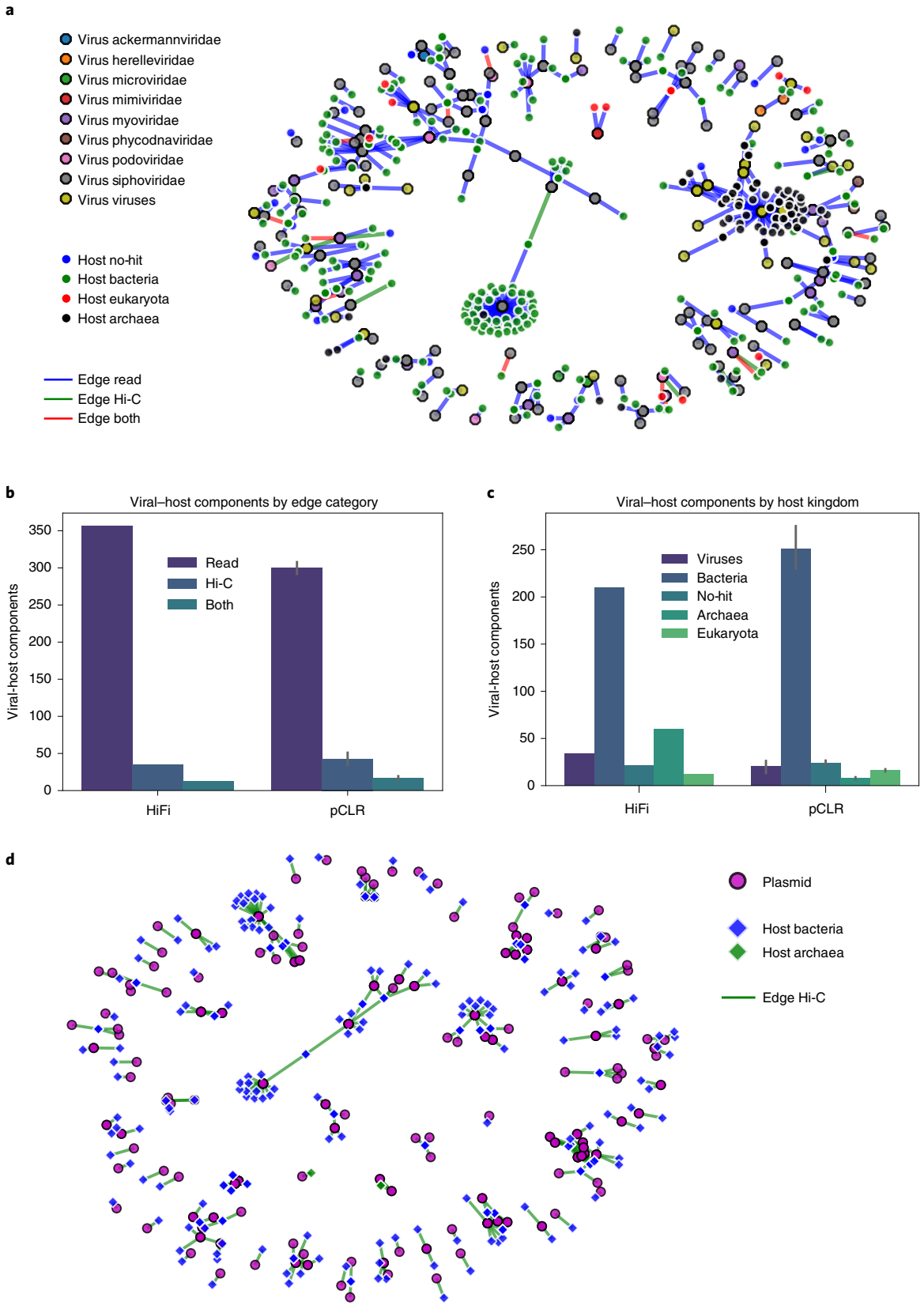
The ability to assemble low-abundance members of the microbial community and to resolve highly related, descendant lineages present in the same sample is dependent on the depth of sequencing. We performed deep sequencing of a complex community to set a benchmark for comparing lower depth coverages (Fig. 3c). Subsequent downsampling experiments evaluated the tradeoff of coverage versus resolution of species and strain-level assembly, revealing that over 300 complete MAGs can be assembled from our sample with only 100 Gbp of HiFi reads. We found that a large proportion of these MAGs are de novo lineage-resolved complete MAGs in our sample (93–210 MAGs in all replicates). This tally might vary based on differences in sample composition in other environments, but it reveals a constraint on genome variation in these lineages that was previously difficult to identify. The de novo assembly of lineage-resolved MAGs with HiFi reads is more likely to occur in lineages without large structural variants or genomic islands that cannot be spanned by current HiFi reads, such that lineages containing features like phage integrations will likely still produce collapsed representations of multiple genomes. We hypothesise that better strain-level resolution of MAGs could be achieved before assembly by combining the structural variation information from the metaFlye assembly graph with phased SNP haplotypes generated by MAGPhase.

Long-read technologies are capable of generating reads many kilobases in length and might eventually be able to encompass entire microbial genomes in single reads. HiFi reads are presently limited to an average of 20 kbp or fewer and are more expensive to generate than equivalent amounts of short reads and long error-prone reads. However, the principal limitation of read length in metagenomic assembly is currently the DNA extraction methodologies used to ensure comprehensive representation of the microbial community, which generally produce DNA fragments fewer than 10 kbp[6,7], making the read length capability of other platforms moot. We also showed that HiFi assemblies are not substantially improved by polishing with short reads, so the complexity and cost relative to other experimental designs are reduced (Extended Data Fig. 1). We noted a modest (~20%) depletion in representation of Gram-negative taxa in HiFi reads compared to short reads; however, this difference did not appear in the subreads from the cells that generated the HiFi reads, indicating that the effect is related to the conversion of subreads to HiFi reads[30]. We show that MAGPhase is capable of resolving lineages using HiFi read alignments to assemblies of differing quality (Fig. 2c), so low-pass sequencing with HiFi reads could potentially be used in tandem with previously assembled references in future surveys. We note that improvements in long-read accuracy have extended to other platforms[40] and that our methods should be equally applicable to these datasets. However, issues imposed by current methods for metagenome DNA extraction will continue to limit the size of reads achievable by all sequencing technologies.

Several biological insights were made possible by the use of HiFi reads. Use of the antiSMASH[37] detection tool identified 40% more BGCs in the HiFi assembly than the highest count in the next best pCLR assembly. The antiSMASH results identified 19 BGCs that show high similarity to a recently identified class of gene clusters

Fig. 4 | HiFi reads improve associations of mobile genetic elements with candidate host species. A network plot of predicted host-virus associations (**a**) identified through HiFi read overlaps (blue), Hi-C links (green) and both data types (red) revealed viral genomes that have broad host specificity. In addition, the HiFi assembly was better able to identify candidate viral–archaeal associations than those detected in the pCLR datasets. Viral–host associations were predominantly identified through HiFi read alignments (**b**), and the HiFi assembly had a higher proportion of this evidence (356 associations) compared to the average pCLR assembly (mean, 251.34). Each pCLR bar ($n=3$) indicates the mean value for each category (Read: 251.34; Hi-C: 43.09; Both: 16.21), and error bars indicate the standard deviation (Read: 25.10; Hi-C: 19.93; Both: 6.84). Highlighting the difference in domain detection between the assemblies, more viral–archaeal links (**c**) were identified in the HiFi assembly than in the pCLR assemblies. Each pCLR bar ($n=3$) indicates the mean value for each category (Archaea: 7.67; Bacteria: 251.25; Eukaryota: 16.14; Viruses: 20.40; no-hit: 23.86), and error bars indicate the standard deviation (Archaea: 2.07; Bacteria: 35.06; Eukaryota: 1.77; Vruses: 11.22; no-hit: 4.02). Using Hi-C link data, we were also able to identify candidate hosts for assembled plasmid sequence (**d**) in the HiFi assembly.

encoding the production of proteasome inhibitors from the human gut microbiota[41], indicating that these functions might be of similar importance for host colonization in ruminants as they are in humans. Additionally, we identified several novel associations of mobile genetic elements in our sample using a combination of Hi-C linkage data and HiFi read alignment overlaps. We detected archaeal–virus association links ($n = 60$) with high complexity (diameter = 7) primarily through HiFi read overlaps. Host–plasmid



analysis using Hi-C links also identified broad host specificity for assembled, circular plasmids. In total, we identified 424 and 298 potential host–viral and host–plasmid links, respectively, which represents one of the most substantial associations of mobile element activity in a single sample to date.

To our knowledge, no previous study has reported lineage-resolved high-quality MAGs at the strain level in a complex metagenome. Our analysis suggests that biological insights can be gained through the use of long (>5-kb) reads with suitably low (~1%) error rates capable of spanning orthologous genomic regions and resolving species-level and strain-level haplotypes into separate assemblies. Assembly with metaFlye, binning with Hi-C and phasing with MAGPhase can produce strain-level MAGs with minimal manual curation. Resulting lineage-resolved complete MAGs are a step toward 'complete metagenomics'—isolate-quality genome assemblies for microbial organisms from complex metagenome samples.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41587-021-01130-z.

Received: 4 May 2021; Accepted: 13 October 2021;
Published online: 03 January 2022

References

1. Bowers, R. M. et al. Minimum information about a single amplified genome (MISAG) and a metagenome-assembled genome (MIMAG) of bacteria and archaea. *Nat. Biotechnol.* **35**, 725–731 (2017).
2. Chen, L.-X., Anantharaman, K., Shaiber, A., Eren, A. M. & Banfield, J. F. Accurate and complete genomes from metagenomes. *Genome Res.* **30**, 315–333 (2020).
3. Pasolli, E. et al. Extensive unexplored human microbiome diversity revealed by over 150,000 genomes from metagenomes spanning age, geography, and lifestyle. *Cell* **176**, 649–662 (2019).
4. Singleton, C. M. et al. Connecting structure to function with the recovery of over 1000 high-quality metagenome-assembled genomes from activated sludge using long-read sequencing. *Nat. Commun.* **12**, 2009 (2021).
5. Vollger, M. R. et al. Long-read sequence and assembly of segmental duplications. *Nat. Methods* **16**, 88–94 (2019).
6. Bickhart, D. M. et al. Assignment of virus and antimicrobial resistance genes to microbial hosts in a complex microbial community by combined long-read assembly and proximity ligation. *Genome Biol.* **20**, 153 (2019).
7. Moss, E. L., Maghini, D. G. & Bhatt, A. S. Complete, closed bacterial genomes from microbiomes using nanopore sequencing. *Nat. Biotechnol.* **38**, 701–707 (2020).
8. Zhang, L. et al. A comprehensive investigation of metagenome assembly by linked-read sequencing. *Microbiome* **8**, 156 (2020).
9. Nurk, S., Meleshko, D., Korobeynikov, A. & Pevzner, P. A. metaSPAdes: a new versatile metagenomic assembler. *Genome Res.* **27**, 824–834 (2017).
10. Kolmogorov, M. et al. metaFlye: scalable long-read metagenome assembly using repeat graphs. *Nat. Methods* **17**, 1103–1110 (2020).
11. Watson, M. & Warr, A. Errors in long-read assemblies can critically affect protein prediction. *Nat. Biotechnol.* **37**, 124–126 (2019).
12. Latorre-Pérez, A., Villalba-Bermell, P., Pascual, J. & Vilanova, C. Assembly methods for nanopore-based metagenomic sequencing: a comparative study. *Sci. Rep.* **10**, 13588 (2020).
13. Olm, M. R. et al. inStrain profiles population microdiversity from metagenomic data and sensitively detects shared microbial strains. *Nat. Biotechnol.* **39**, 727–736 (2021).
14. Quince, C. et al. STRONG: metagenomics strain resolution on assembly graphs. *Genome Biol.* **22**, 214 (2021).
15. Kang, D. D., Froula, J., Egan, R. & Wang, Z. MetaBAT, an efficient tool for accurately reconstructing single genomes from complex microbial communities. *PeerJ* **3**, e1165 (2015).
16. Burton, J. N., Liachko, I., Dunham, M. J. & Shendure, J. Species-level deconvolution of metagenome assemblies with Hi-C–based contact probability maps. *G3 (Bethesda)* **4**, 1339–1346 (2014).
17. Parks, D. H. et al. Recovery of nearly 8,000 metagenome-assembled genomes substantially expands the tree of life. *Nat. Microbiol.* **2**, 1533–1542 (2017).
18. Lapierre, P. & Gogarten, J. P. Estimating the size of the bacterial pan-genome. *Trends Genet.* **25**, 107–110 (2009).
19. Vicedomini, R., Quince, C., Darling, A. E. & Chikhi, R. Strainberry: automated strain separation in low-complexity metagenomes using long reads. *Nat. Commun.* **12**, 4485 (2021).
20. O'Brien, J. D. et al. A Bayesian approach to inferring the phylogenetic structure of communities from metagenomic data. *Genetics* **197**, 925–937 (2014).
21. Quince, C. et al. DESMAN: a new tool for de novo extraction of strains from metagenomes. *Genome Biol.* **18**, 181 (2017).
22. Nicholls, S. M. et al. On the complexity of haplotyping a microbial community. *Bioinformatics* https://doi.org/10.1093/bioinformatics/btaa977 (2020).
23. Vicedomini, R., Quince, C., Darling, A. E. & Chikhi, R. Strainberry: automated strain separation in low-complexity metagenomes using long reads. *Nat. Commun.* **12**, 4485 (2021).
24. Wenger, A. M. et al. Accurate circular consensus long-read sequencing improves variant detection and assembly of a human genome. *Nat. Biotechnol.* **37**, 1155–1162 (2019).
25. Ebert, P. et al. Haplotype-resolved diverse human genomes and integrated analysis of structural variation. *Science* **372**, eabf7117 (2021).
26. Garg, S. et al. Chromosome-scale, haplotype-resolved assembly of human genomes. *Nat. Biotechnol.* **39**, 309–312 (2021).
27. Porubsky, D. et al. Fully phased human genome assembly without parental data using single-cell strand sequencing and long reads. *Nat. Biotechnol.* **39**, 302–308 (2021).
28. Miga, K. H. et al. Telomere-to-telomere assembly of a complete human X chromosome. *Nature* **585**, 79–84 (2020).
29. Menzel, P., Ng, K. L. & Krogh, A. Fast and sensitive taxonomic classification for metagenomics with Kaiju. *Nat. Commun.* **7**, 11257 (2016).
30. Kolmogorov, M. Supporting data for the manuscript 'Generation of lineage-resolved complete metagenome-assembled genomes in complex microbial communities'. https://doi.org/10.5281/zenodo.5138306 (2021).
31. Chaumeil, P.-A., Mussig, A. J., Hugenholtz, P. & Parks, D. H. GTDB-Tk: a toolkit to classify genomes with the Genome Taxonomy Database. *Bioinformatics* **36**, 1925–1927 (2020).
32. Ondov, B. D. et al. Mash: fast genome and metagenome distance estimation using MinHash. *Genome Biol.* **17**, 132 (2016).
33. Wang, B. et al. Variant phasing and haplotypic expression from long-read sequencing in maize. *Commun. Biol.* **3**, 1–11 (2020).
34. Tseng, E. cDNA_cupcake v24.0.0. https://github.com/Magdoll/cDNA_Cupcake
35. Nei, M. & Rooney, A. P. Concerted and birth-and-death evolution of multigene families. *Annu. Rev. Genet.* **39**, 121–152 (2005).
36. Meleshko, D. et al. BiosyntheticSPAdes: reconstructing biosynthetic gene clusters from assembly graphs. *Genome Res.* **29**, 1352–1362 (2019).
37. Blin, K. et al. antiSMASH 5.0: updates to the secondary metabolite genome mining pipeline. *Nucleic Acids Res.* **47**, W81–W87 (2019).
38. Pellow, D. et al. SCAPP: an algorithm for improved plasmid assembly in metagenomes. *Microbiome* **9**, 144 (2021).
39. He, C. et al. Genome-resolved metagenomics reveals site-specific diversity of episymbiotic CPR bacteria and DPANN archaea in groundwater ecosystems. *Nat. Microbiol.* **6**, 354–365 (2021).
40. Wick, R. R., Judd, L. M. & Holt, K. E. Performance of neural network basecalling tools for Oxford Nanopore sequencing. *Genome Biol.* **20**, 129 (2019).
41. Guo, C.-J. et al. Discovery of reactive microbiota-derived metabolites that inhibit host proteases. *Cell* **168**, 517–526 (2017).

Methods

Long-read DNA sequencing and subread extraction. A fecal sample was taken from a young (<1 year old) wether lamb of the Katahdin breed. The animal died while on pasture and was diagnosed postmortem with combined *Strongyloides* and coccidial infection. The sample was acquired postmortem following the USDA ARS IACUC protocol 137.0 during routine necropsy to determine cause of death. The sample had a watery texture consistent with diarrhea, and apparent parasite eggs were observed within the sample, which was transferred to a 50-ml tube, mixed to make as homogenous as possible and aliquoted into 1.5-ml microfuge tubes. DNA was extracted in small batches from approximately 0.5 g per batch using the QIAamp PowerFecal DNA Kit, as suggested by the manufacturer (Qiagen), with moderate bead beating and sheared using a Digilab Genomic Solutions HydroShear instrument. The sheared DNA was size-selected to approximately 9–18 kb on a SageELF instrument to a final target size that varied from 9 kbp to 16 kbp, followed by library preparations using the SMRTbell Template Prep Kit version 1.0, as described[20]. Sequence data were collected over time and included 46 SMRT cells on a Sequel instrument using ten library preparations, with 24 cells of v2 chemistry and average inserts of 9–10 kbp and 22 cells of v3 chemistry and average inserts of 14 kbp. An additional eight cells representing individual library preparations were sequenced on a Sequel II instrument using v1.0 chemistry and average inserts of 14 kbp. Subreads and circular consensus sequencing (CCS) reads were generated using SMRTLink software version 6.0 CCS protocol and default settings. An average of 35% of subreads per cell were converted to CCS-corrected reads (range, 1–63%). This resulted in 255 Gbp of total CCS reads from both the Sequel I (45 Gbp of the total) and Sequel II (210 Gbp) sequencing runs. A subset of these data (46 Sequel I SMRT cells) representing 18% (45 Gbp) of the total dataset was previously assembled as validation data in the metaFlye assembler publication[10]. The Sequel II dataset was filtered after CCS correction to retain only reads that fit HiFi quality standards (3+ full length passes and average read quality scores >Q20). A small proportion of our Sequel I dataset (4,350 reads, 0.02% of the total number of CCS reads) consisted of reads that did not meet HiFi read quality standards (average Q scores above 20), as this dataset had been filtered with a previous version of the SMRTLink software. These reads were retained as they comprised a very small proportion of the total dataset.

Subreads were extracted from the converted CCS reads to provide a suitable comparison between uncorrected and corrected long-read datasets. First, all of the constitutive subreads of the CCS reads were identified from subread BAM files. Using a custom script (https://github.com/njdbickhart/SheepHiFiManuscript/blob/main/downsampling/extractPacbioCLRFromCCSData.py), the second, third and fourth subreads were separately extracted into FASTQ files designated pCLR1, pCLR2 and pCLR3, respectively (the first subread does not typically encompass the complete DNA fragment, so it was discarded). Statistics on subread lengths from the Sequel I and Sequel II datasets are shown in Supplementary Figs. 10 and 11, respectively. Due to sequence read falloff in later subreads, the pCLR3 dataset was truncated relative to the pCLR1 and pCLR2 datasets. In the Sequel I dataset, a small proportion of reads (51 reads, ~0.001%) did not have a fourth (pCLR3) subread, making the third replicate dataset smaller than the others. This resulted in a reduction of 104 Mbp of sequence in this dataset (48.763 Gbp) compared to the pCLR1 or pCLR2 extracted subreads (48.830 Gbp). The original CCS reads were organized into a dataset hereafter referred to as the 'HiFi' reads, and the three subread replicates were labeled pCLR1-3 in the chronological order in which they were sequenced in the subread BAM files.

Short-read sequencing and Hi-C library preparation. An approximately 2-g subsample of frozen homogenized fecal material was provided to Phase Genomics for Hi-C contact map construction using their ProxiMeta service. The restriction endonucleases Sau3AI and MluCI were used to generate separate Hi-C sequencing libraries, as previously described[42]. Using a total of 107 million paired-end reads from both, Hi-C libraries were generated for analysis. A separate portion of the fecal sample was saved for short-read 'whole genome shotgun' (WGS) DNA sequencing, which was performed by Phase Genomics using their ProxiMeta service. TruSeq PCR-free libraries were created from this sample as previously described[43] and were sequenced on an Illumina NextSeq 500. A total of 149 Gbp of WGS short reads were generated from this sample.

Genome assembly, read alignment and binning. Reads from the HiFi and pCLR datasets were assembled into contigs using the metaFlye[10] genome assembler, version 2.7-b1646 for HiFi reads and version 2.7.1-b1590 for pCLR reads. The assembler was run in metagenome mode ('—meta') flag, and the '—pacbio-hifi' and '—pacbio-raw' data prefix flags were used for input HiFi reads and the pCLR reads, respectively. We note that the '—pacbio-hifi' input designation only uses reads that have average error rates below 1% for the disjointig and contig phases of the metaFlye workflow. This means that only HiFi quality reads (Q20+) were used to generate the initial graphs and final contigs of the HiFi assembly. However, all input reads were used in the consensus polishing step of metaFlye. The pCLR assemblies were polished with two rounds of Pilon[44] correction using the previously generated short-read dataset. The HiFi assembly was compared against the pCLR assemblies without post hoc polishing. Contigs shorter than 1,000 bp in all assemblies were removed from further analysis. Circular contigs were identified from metaFlye assembly reports. Short reads were aligned to the assemblies using BWA MEM[45] using default settings. HiFi reads were aligned using minimap2 (ref. [46]) with the '-x asm 20' preset setting as recommended by the developers. Window-based alignment analysis was conducted by using custom Python scripts (https://github.com/njdbickhart/SheepHiFiManuscript/blob/main/magphase_workflow/getBAMMapQ0Ratios.py).

Hi-C read pairs were aligned to each assembly using BWA MEM with the '-5SP' flag to disable attempts to pair reads according to normal Illumina paired-read settings. Resulting BAM files from Hi-C reads were sorted by read name. Hi-C alignments were used in the bin3c[47] binning pipeline to generate a set of bins for each assembly.

Contig and MAG quality assessment. Contig-level quality was assessed by CheckM[17] run on all contigs longer than 1 Mb in size in each assembly. Bin quality was determined by DAS Tool[48] SCG metrics generated from the solitary input of bin3c binning metrics. The rRNA genes were predicted from each MAG using Barrnap (https://github.com/tseemann/barrnap), and tRNA gene content was determined using tRNAscan-SE[49]. MAGs were classified as 'complete' for bins that had more than 90% SCG completeness and less than 10% SCG contamination estimates from the DAS Tool quality assessment data. The MISAG/MIMAG quality standards[1] were assessed from the combined output of DAS Tool, Barrnap and tRNAscane-SE where appropriate, with high-quality draft MAGs meeting minimal SCG metrics of more than 90% complete and less than 5% contaminated, with the presence of the 5S, 16S and 23S genes and at least 18 tRNA genes. The total set of MAGs derived from DAS Tool output automatically met the Bowers et al.[1] criteria for medium-quality draft (≥50% complete and <10% contaminated). We refer to this set as 'MAGs' (without prefix adjectives) for ease of reference in this article.

Optimizing metaFlye for long and accurate reads. The metaFlye algorithm was initially designed for assembling long error-prone reads and featured limited support of the long and accurate reads[10]. In this study, we performed several modifications of the algorithm and added a new '–pacbio-hifi' option for improved assembly of long and accurate reads (incorporated into the updated metaFlye package, version 2.8).

First, we added a homopolymer-compressed alignment scoring scheme, similarly to the approach implemented in HiCanu[50]. Because the vast majority of errors in HiFi reads are indels inside homopolymer regions, the modified scoring scheme computes alignments with higher sensitivity and thus simplifies the repeat graph by separating inexact copies of genomic repeats.

Second, we implemented a minimizer indexing scheme[46,51] as a replacement to the solid *k*-mer approach[10]. The minimizer indexing substantially reduces the memory footprint and reduces running time (metaFlye took 3,900 CPU hours and 662 Gb of RAM to assemble the entire HiFi dataset).

Third, we modified the disjointig assembly algorithm to limit the number of occasional breaks in contiguity as follows. The original algorithm is using only the alignments longer than the MinimumOverlap parameter to assemble disjointigs and build a repeat graph. There is a tradeoff in selecting this parameter: a higher parameter value results in a less tangled graph and more contiguous assembly, whereas a lower value results in a more fragmented graph and breaks in assembly (because genomes with low depth of coverage might not have enough reads with sufficient overlap length). To address this tradeoff, the modified version of the disjointing assembly algorithm may use alignments shorter than the MinimumOverlap parameter (as short as 1 kbp) under the following extra conditions: (1) there are no alternative alignments longer than MinimumOverlap, and (2) the alignment region is not contained inside a repetitive region. Putative repetitive regions are determined from the alignment pileup profile.

Taxonomic assignment. We distinguish between contig-level and bin-level taxonomic classification to show differences in pCLR/HiFi assembly quality and assign representative taxonomy of the final polished bins, respectively. Contigs were assigned to candidate taxa using the BlobTools version 1.0 (ref. [48]) taxify pipeline, using models from the UniProt (release 2017_07) database, as described previously[6]. Contigs that did not meet the BlobTools threshold for taxonomic assignment, or were identified as belonging to faulty database entries (for example, the 'Cetacean' lineage), were labeled as 'no-hit' taxa. Viral contigs identified from this analysis were used in subsequent virus association analysis (see below). Predicted viral contigs were separately verified using the CheckV[52] pipeline using the 'end-to-end' workflow, multi-threaded, and with normal settings.

The GTDB-TK version 1.0 (ref. [31]) 'classify_wf' workflow was used to assign candidate taxonomic affiliation to all assembled bins. Default GTDB-TK settings were used, with the only exception being the setting of the '—pplacer_cpus' argument to '1', as recommended by the authors. In cases where GTDB-TK was unable to assign a taxonomic lineage, a consensus of contig-level assignments from the BlobTools taxify pipeline was used to assign candidate taxonomic affiliation for the bin. The prevalence of three or more contigs in the MAG, indicating the same species-level taxonomy, were used when possible. In the case of 'ties' among contig-level taxonomic consensus, the final taxonomic consensus was resolved to the lowest possible level (that is, genus or family).

Read-level classification was performed on the Illumina, HiFi and raw PacBio subread datasets with Kaiju[29]. The 'kaiju_db_nr_euk_2021-02-24' database was the reference for all classifications, and Kaiju was run with default parameters in all cases. Output was converted to tabular format with the 'kaiju2krona' utility, and KronaTools version 2.8 (ref. [53]) was used to visualize this output using the 'ktImportText' utility. All three classifications were consolidated into a single interactive plot for easy comparison, which can be found at the following URL: https://zenodo.org/record/5138306/files/hifi_illumina_sub_reads.krona.html?download=1.

Orthologous MAG identification and read alignment. We first sought to identify orthologous bins among each of the pCLR assemblies and the HiFi assembly to provide direct comparisons among similar assembled taxonomic groups. To identify orthologous bins, we used Mash version 2.2 (ref. [32]) sketches of all HiFi bins as a reference against queries of all pCLR bins. Mash sketch settings were -s 100,000 and -k 21, with all other settings left at the default. The Mash 'dist' command was used with a cutoff of 0.10 distance to identify orthologous MAGs, which is approximately equivalent to an average nucleotide identity of 90% between hits. Mash distance values were inverted (1.0 – dist) and relabeled as eANI values for consistency in comparisons. Multiple reference and query hits were allowed and retained for future comparisons.

HiFi reads were realigned to the HiFi and pCLR assembly bins using minimap2 (ref. [46]) as previously described, and alignment files were converted to BAM file format using SAMtools[54]. To reduce the possibility of supplementary or split-read alignments affecting downstream variant calling, we filtered these alignments from the HiFi read BAM files. We then used these filtered alignment files for variant calling and haplotype identification using MAGPhase.

MAGPhase lineage resolution. The MAGPhase algorithm attempts to identify full-length SNP variant haplotypes in a greedy fashion within a given set of genomic coordinates. First, all input read alignment BAM files are stripped of supplementary (flag status = 2,048) alignments to exclude the majority of chimeric reads. By default, only genomic coordinates that have at least ten full-length reads (10× coverage) are considered for variant calling. Initial SNP variants used in read phasing are identified from HiFi read alignment pileups using pysam (https://pysam.readthedocs.io/en/latest/api.html). Due to the lower error rates of HiFi reads, a SNP variant calling strategy similar to short-read variant callers[54] is used. Initially, all candidate SNP variants are tallied within all reads that overlap the genomic coordinate range. To distinguish between potential errors in reads and SNPs, we model expected errors at a rate of 0.5% and test observed variant coverages against an expected error variant coverage using the Fisher exact test. To correct for further multiple hypothesis testing across an entire region, we employ a Benjamini–Hochberg[55] procedure to estimate modified *P* values. Only SNP variants that meet our corrected threshold *P* value (default value, 0.1) are kept for subsequent phasing.

Once a set of candidate SNP variants is called, MAGPhase attempts to phase them into haplotypes based on their presence in HiFi reads. The entire length of an alternate haplotype is imputed using the linkage of previously identified SNP variants on individual HiFi reads that span the region in a stepwise fashion using the read pileups. HiFi reads are first decomposed into their variant position status at each of the predefined, filtered variant loci. A directed, acyclic graph of variant sites is formed through observation of filtered variant alleles on each HiFi read. This graph is then used to compose candidate haplotype 'strings' of variant positions. The assignment of reads to each haplotype is retained for later output, and read counts supporting each variant position are used to estimate relative coverage for the alternative haplotype allele. Missing variant information (due, primarily, to erroneous bases in HiFi reads or alignment of chimeric reads) is recovered through a round of imputation using the IsoPhase algorithm[33]. Imputation is performed when a previously filtered alternative variant base is found at a candidate variant site to attempt to resolve the read into its true haplotype affiliation. The implementation of our imputation method relies on calculation of exact match scoring between a read's observed haplotype allele affiliation and previously identified haplotypes in the region. If the read's score is fewer than three exact variant matches from the next best haplotype allele or has a score that is tied with the best and second-best matching haplotypes, the read's haplotype state is considered to be ambiguous. A single read alignment with no variant state that matches to any previously observed haplotype is discarded. Any haplotypes with variant states that could not be recovered using this imputation algorithm are denoted with question marks (?). The final haplotype dataset consists of equal-length haplotypes of SNP variants observed in the genomic region and their associated read counts.

To reduce the potential expansion of haplotype counts due to recombination, we phased HiFi reads within identified SCG regions of each HiFi and pCLR bin. CCS reads that extended over the edges of SCG regions were included in haplotype phasing, so if two SCG regions were within short distances from each other, phased variant haplotypes could extend further. Partially imputed haplotypes (haplotypes that contained question marks) were excluded from analysis, as these could have resulted from chimeric read alignments or base call errors on selected SNP variant sites within the haplotype. Haplotypes were considered alternative alleles based on read depth, with lower-depth haplotypes considered to be alternatives to the highest-read-depth allele at that loci. Haplotypes that included fewer than three SNPs were filtered, as these tended to have lower counts of read alignments and higher alternate allele haplotype counts. If a MAG was found to have no SNP variants that fit the read depth statistical requirements, it was considered to be a 'lineage-resolved' MAG. MAGs that had unfiltered SNP variants that were otherwise unable to be assigned to haplotypes with three or more SNPs were not considered to be lineage resolved and were labeled as 'polymorphic'. Read depth and read clustering were assessed through custom Python scripts (https://github.com/njdbickhart/SheepHiFiManuscript/blob/main/magphase_workflow/plotMagPhaseOutput.py) and IGV[56] plots.

HiFi read downsampling. To simulate smaller subsets of HiFi reads, we employed a progressive downsampling strategy using recursive sampling. We sampled from the Sequel II HiFi read dataset, as this dataset benefited from the latest advances in chemistry for PacBio sequencing and is likely to be most analogous to the data used in the future by other groups. In brief, this method attempts to bin reads into their lowest deciles using a recursive test against the target percentile with values generated from the default Perl random number generator. Reads are placed into all deciles to which their random number value fits, without replacement. In practical terms, this means that all reads within the lower decile samples (for example, 10%) are present in the top (90%) decile. Actual percentages of read counts and total base pairs were compared to expected decile categories and were found to be almost perfectly matched (Supplementary Table 8). This algorithm is implemented in base Perl (version 5.8+) at the following URL: https://github.com/njdbickhart/SheepHiFiManuscript/blob/main/downsampling/downsampleFastaForAssembly.pl. Each downsampled dataset was assembled with metaFlye using the same settings described above. Bins were generated using the full Hi-C sequence dataset and bin3c, and lineage-resolved complete MAGs were identified using MAGPhase. The HiFi reads from each downsampled decile were aligned to their associated assembly to determine variant sites in the MAGPhase pipeline.

Polishing HiFi assembly using short reads. Pilon[44] polishing of the HiFi assembly was accomplished using the short-read sequence dataset in a single iteration with the '—fix indels' setting. A single iteration was chosen, as the first polishing iteration usually results in the largest magnitude of changes[57]. The polished and unpolished HiFi assemblies were then assessed using the IDEEL pipeline[58]. Comparisons between counts of open reading frame length ratios between the two assemblies were conducted with the qqplot function in the R stats (version 4.0.2) package and plotted with base R.

Biosynthetic gene cluster prediction and functional annotation. The HiFi and pCLR assemblies were used as input for the BGC prediction tool antiSMASH[37] (version 5), with Prodigal[59] as the default gene prediction tool. The predicted BGCs were classified into six BGC classes: RiPP, NRPS, Terpene, PKS, Saccharide and Others. Also, from the annotated GenBank files, BGCs were classified into either 'Partial' (when they were found on a contig edge) or 'Complete' (when this was not the case). Finally, predicted BGCs with fewer than 50% of the genes having hits to the best KnownClusterBlast hit, which is obtained from searching all BGCs in the MIBiG database, version 2.0 (ref. [60]), were considered 'novel'. When this condition was not satisfied, the BGCs were classified into the 'Known' group.

Virus and plasmid association analysis. Viral contigs were identified from BlobTools taxonomic assignment for use in the association analysis. Genome completeness of these viral contigs was estimated by the CheckV 1.0 'end_to_end' workflow[40] (Supplementary Table 7). Given the potential novelty of assembled viral genomes in this dataset, the 'Not-Determined' and 'Medium' completeness viral contigs were not filtered before the association analysis. CCS read overlaps and Hi-C link data were used to identify potential host–viral associations, as previously described[6]. In brief, read overlap data consisted of CCS reads that partially mapped to both viral and non-viral contigs. Associative Hi-C links consisted of cases where the number of inter-contig Hi-C pair alignments between viral and non-viral contigs were three standard deviations above the average count for all contigs. Both datasets were compared for overlap, and network plots were generated using the Python NetworkX version 2.5 module. The analysis workflow and network plotting were automated using the following script: https://github.com/njdbickhart/SheepHiFiManuscript/blob/main/viral_association/viralAssociationPipeline.py.

Plasmids were identified using the SCAPP workflow[38] with the metaFlye HiFi assembly graph ('gfa' file) and aligned short-read BAM files to the final, polished assembly FASTA file[38]. The default settings were used apart from the setting of the '-k/–max_kmer' value to '0' to disable *k*-mer-based tokenization of sequence reads. SCAPP plasmid nodes were filtered if they were shorter than 5 kb or longer than 1 Mb in length before alignment. Plasmid node orthologs in each main assembly were identified through minimap2 (ref. [46]) alignments and were removed before alignment. Hi-C reads were aligned to this modified reference using bwa MEM[45], and alignment files were converted to BAM format using SAMtools[54]. The alignment file was used in the aforementioned viral association workflow script to identify substantial links between candidate plasmids and host contigs.

Contig-level annotation via the BlobTools[48] taxify pipeline was used to classify each candidate host by kingdom. Networks were visualized using the Python NetworkX version 2.5 module.

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

Data availability

The HiFi sheep dataset, Hi-C reads and WGS short reads are available on National Center of Biotechnology Information BioProject PRJNA595610 at accession IDs SRX7628648, SRX10704191 and SRX7649993, respectively. Whole-metagenome assemblies and MAG bins for the pCLR and HiFi datasets are available at https://doi.org/10.5281/zenodo.4729049. The 'kaiju_db_nr_euk_2021-02-24' database was used for Kaiju classification (https://kaiju.binf.ku.dk/server). The '2017-07' version of the UniProt database was used for BlobTools classification (https://ftp.uniprot.org/pub/databases/uniprot/previous_major_releases/release-2017_07/).

Code availability

The MAGPhase script and codebase are part of the https://github.com/Magdoll/cDNA_Cupcake GitHub repository. Scripts to replicate the analysis of the manuscript and to implement the MAGPhase workflow are located at this centralized repository: https://github.com/njdbickhart/SheepHiFiManuscript (ref. [61]). A listing of all analysis software packages used in this study can be found in Supplementary Table 10.

References

42. Press, M. O. et al. Hi-C deconvolution of a human gut microbiome yields high-quality draft genomes and reveals plasmid-genome interactions. Preprint at https://www.biorxiv.org/content/10.1101/198713v1 (2017).
43. Bentley, D. R. et al. Accurate whole human genome sequencing using reversible terminator chemistry. *Nature* **456**, 53–59 (2008).
44. Walker, B. J. et al. Pilon: an integrated tool for comprehensive microbial variant detection and genome assembly improvement. *PLoS ONE* **9**, e112963 (2014).
45. Li, H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. Preprint at https://arxiv.org/abs/1303.3997 (2013).
46. Li, H. Minimap and miniasm: fast mapping and de novo assembly for noisy long sequences. *Bioinformatics* **32**, 2103–2110 (2016).
47. DeMaere, M. Z. & Darling, A, E.bin3C: exploiting Hi-C sequencing data to accurately resolve metagenome-assembled genomes. *Genome Biol.* **20**, 46 (2019).
48. Laetsch, D. R. & Blaxter, M. L. BlobTools: interrogation of genome assemblies. *F1000Research* **6**, 1287 (2017).
49. Chan, P. P. & Lowe, T. M. tRNAscan-SE: searching for tRNA genes in genomic sequences. *Methods Mol. Biol.* **1962**, 1–14 (2019).
50. Nurk, S. et al. HiCanu: accurate assembly of segmental duplications, satellites, and allelic variants from high-fidelity long reads. *Genome Res.* **30**, 1291–1305 (2020).
51. Chin, C.-S. & Khalak, A. Human genome assembly in 100 minutes. Preprint at https://www.biorxiv.org/content/10.1101/705616v1 (2019).
52. Nayfach, S. et al. CheckV assesses the quality and completeness of metagenome-assembled viral genomes. *Nat. Biotechnol.* **39**, 578–585 (2020).
53. Ondov, B. D., Bergman, N. H. & Phillippy, A. M. Interactive metagenomic visualization in a web browser. *BMC Bioinformatics* **12**, 385 (2011).
54. Li, H. et al. The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078–2079 (2009).
55. Benjamini, Y. & Hochberg, Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Stat. Soc. Ser. B Methodol.* **57**, 289–300 (1995).
56. Robinson, J. T. et al. Integrative Genomics Viewer. *Nat. Biotechnol.* **29**, 24–26 (2011).
57. Chen, Z., Erickson, D. L. & Meng, J. Polishing the Oxford Nanopore long-read assemblies of bacterial pathogens with Illumina short reads to improve genomic analyses. *Genomics* **113**, 1366–1377 (2021).
58. Stewart, R. D. et al. Compendium of 4,941 rumen metagenome-assembled genomes for rumen microbiome biology and enzyme discovery. *Nat. Biotechnol.* **37**, 953–961 (2019).
59. Hyatt, D. et al. Prodigal: prokaryotic gene recognition and translation initiation site identification. *BMC Bioinformatics* **11**, 119 (2010).
60. Kautsar, S. A. et al. MIBiG 2.0: a repository for biosynthetic gene clusters of known function. *Nucleic Acids Res.* **48**, D454–D458 (2020).
61. Bickhart, D. M. SheepHiFiManuscript. https://doi.org/10.5281/zenodo.5120910 (2021).

Acknowledgements

We thank K. McClure, K. Kuhn, B. Lee, J. Carnahan and W. Thompson for technical support. D.M.B. was supported by appropriated USDA CRIS Project 5090-31000-026-00-D. T.P.L.S. and S.B.S. were supported by appropriated USDA CRIS Project 3040-31000-100-00D. I.L., S.T.S. and G.U. were supported, in part, by NIH grants R44AI150008 and R44AI162570 to Phase Genomics. I.M. was supported by grants from the European Research Council (no. 640384) and from the Israel Science Foundation (no. 1947/19). M.K. and P.A.P. were supported by NSF/MCB-BSF grant 1715911. V.P.A. was supported by the US Defense Advanced Research Projects Agency's Living Foundries program award HR0011-15-C-0084. A.K. and I.T. were supported by St. Petersburg State University (grant ID PURE 73023672). K.P. was supported by appropriated USDA CRIS Project 5090-21000-071-000-D. We thank P. J. Weimer for helpful comments and suggestions on the manuscript. The USDA does not endorse any products or services. Mentioning of trade names is for information purposes only. The USDA is an equal opportunity employer.

Author contributions

T.P.L.S. and D.M.B. conceived the project, with extensive modifications introduced on the advice of I.L. and P.A.P. S.B.S and T.P.L.S. were responsible for collecting the sample and generating the sequence data. D.B. and M.K. produced the assemblies and conducted a large proportion of reported analysis. G.U. and S.T.S. conducted analysis related to Hi-C linkage data. V.P.A. and M.H.M. identified biosynthetic gene clusters in the dataset. D.M.B., A.Z. and I.M. identified mobile genetic elements in the sample. E.T. developed the MAGPhase algorithm, with input from D.M.B. D.M.B., T.P.L.S., M.K. and P.A.P. wrote the manuscript. All authors read and contributed to the final manuscript.

Competing interests

The authors declare the following competing interests: M.H.M. is a co-founder of Design Pharmaceuticals and a member of the scientific advisory board of Hexagon Bio. E.T. and D.M.P. are employees of Pacific Biosciences. G.U. is an employee of Amazon. S.T.S. and I.L. are co-founders and the CTO and CEO, respectively, of Phase Genomics. The remaining authors declare no competing interests.

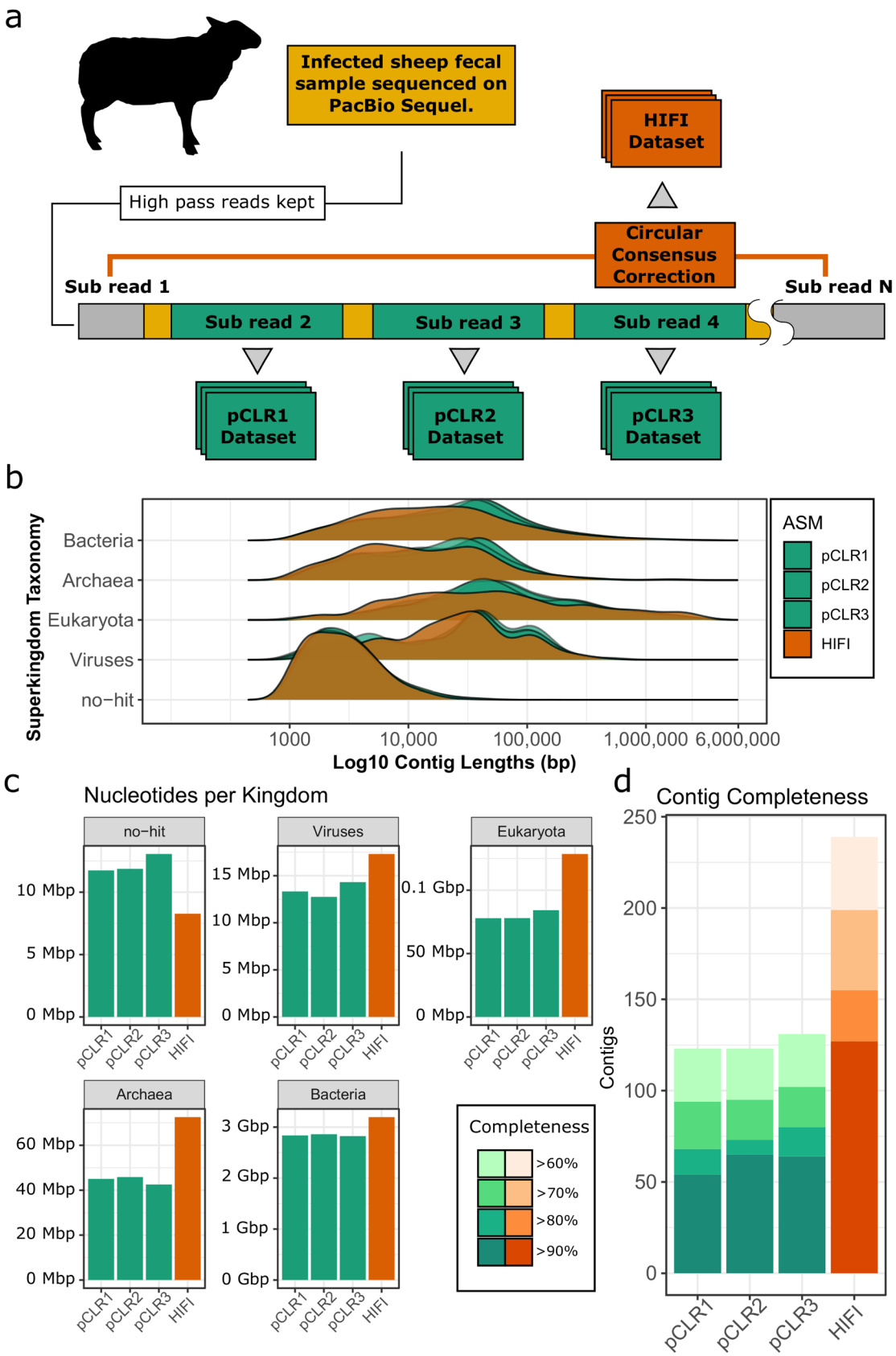
Additional information

Extended data is available for this paper at https://doi.org/10.1038/s41587-021-01130-z.

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41587-021-01130-z.

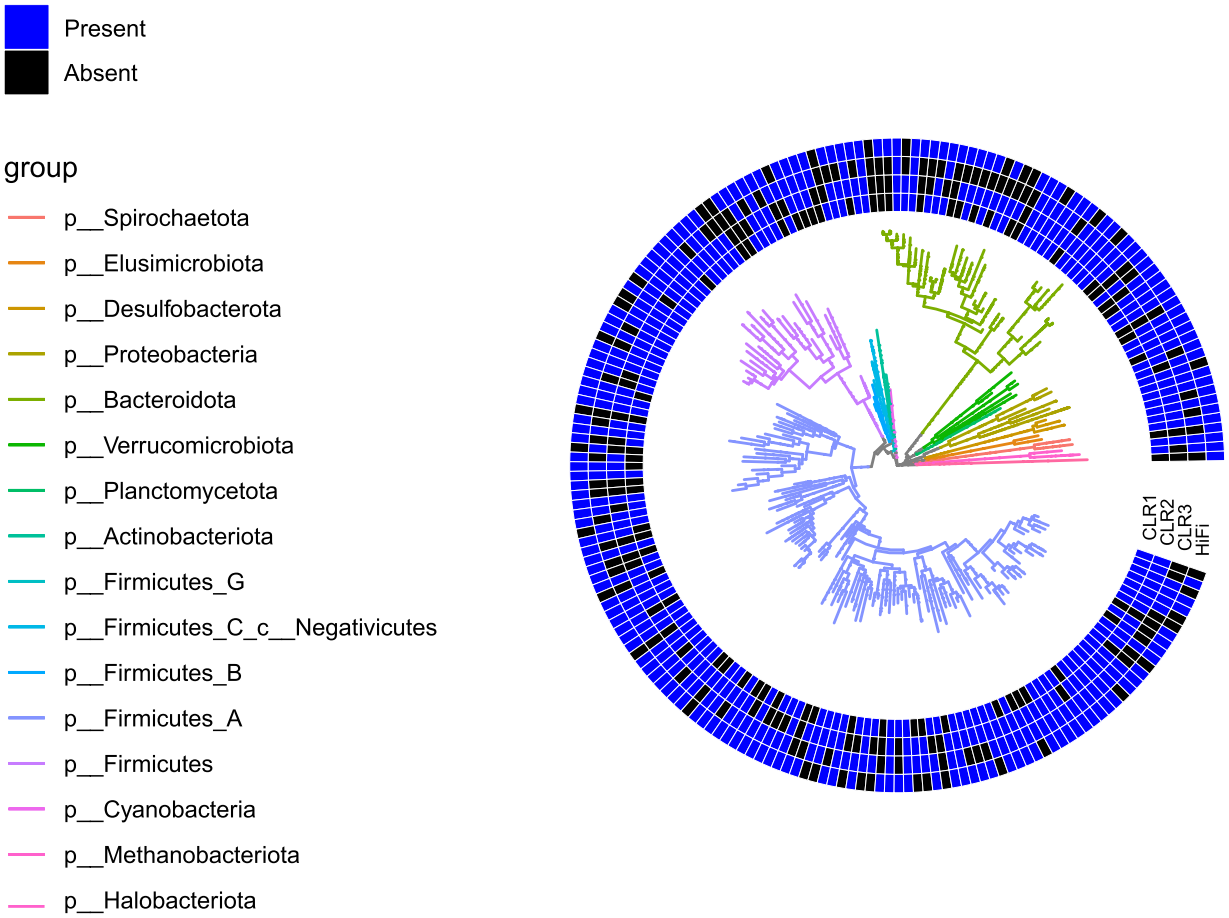
Correspondence and requests for materials should be addressed to Pavel A. Pevzner or Timothy P. L. Smith.

Peer review information *Nature Biotechnology* thanks Mads Albertsen, C. Titus Brown and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.

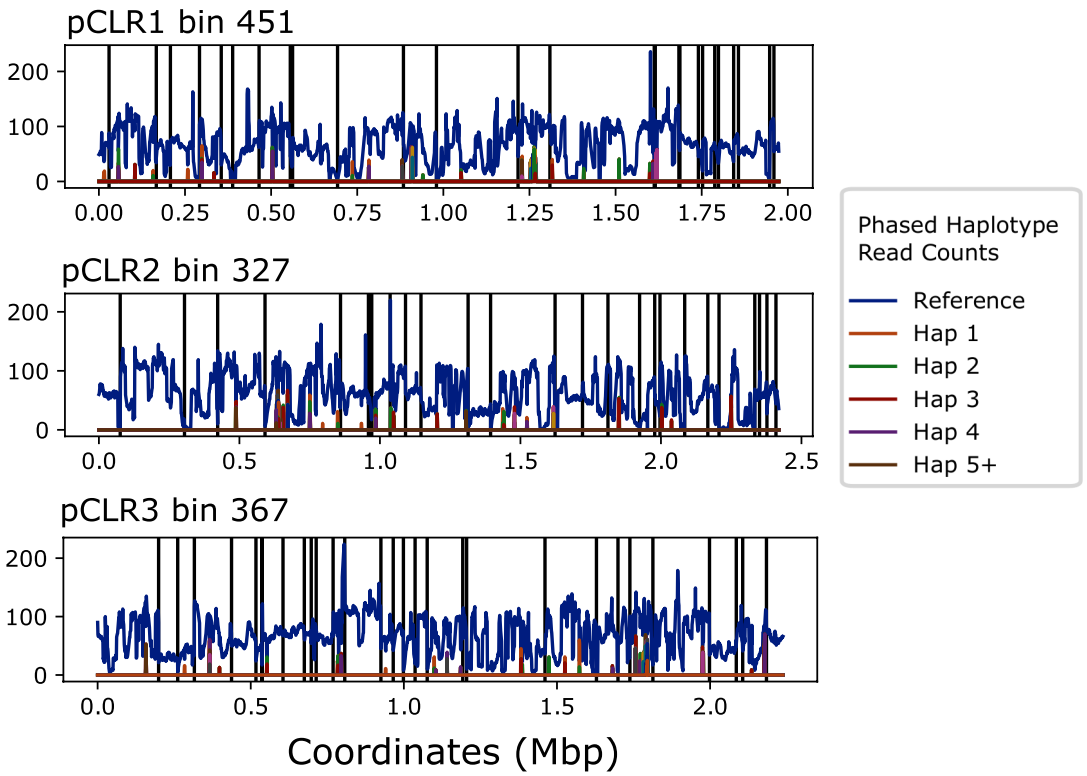


Extended Data Fig. 1 | See next page for caption.

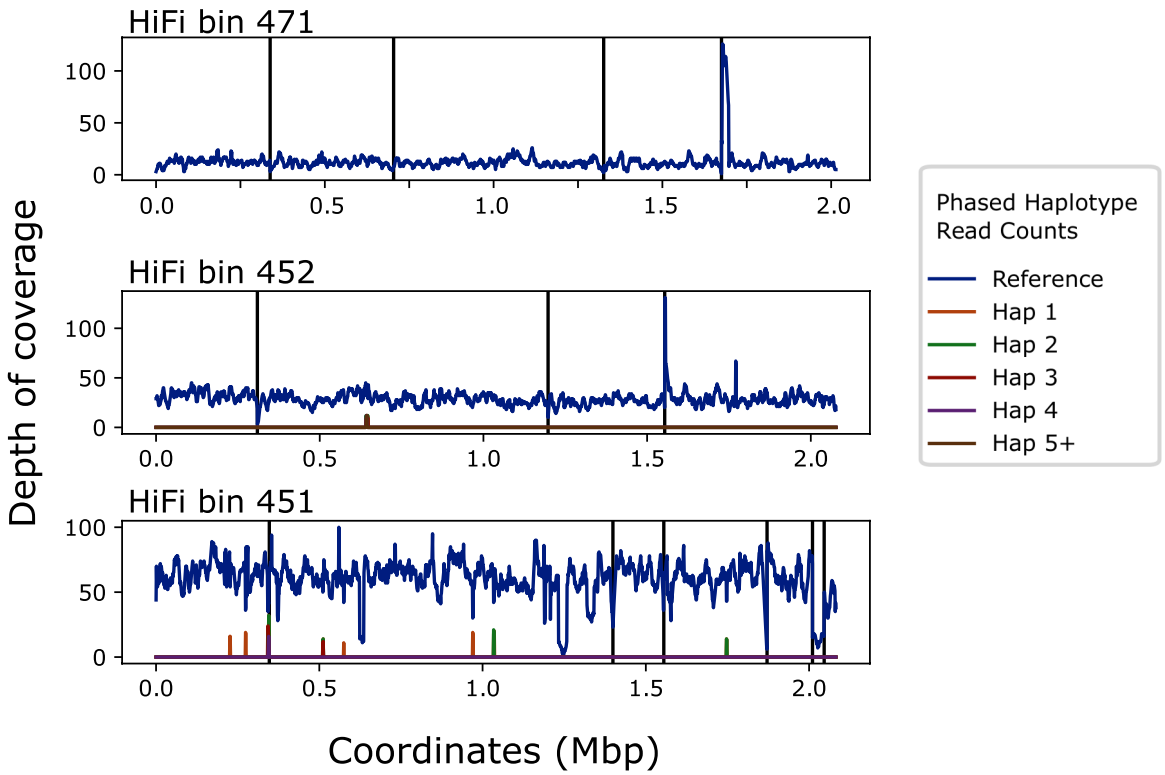
Extended Data Fig. 1 | Contig-level comparison of pCLR and HiFi assemblies. a. Strategy for generating the read sets for the three pCLR and the HiFi assemblies. b. Comparison of contig length distributions in the four assemblies demonstrating a tendency for pCLR assembly to create longer contigs. c. Comparison of the total length of each assembly after separation of contigs into predicted Superkingdoms demonstrating an increased length from HiFi assembly among assigned Superkingdom and reduced length in unassigned bin. d. Comparison of the completeness of pCLR and HiFi assemblies based on the presence of >90% expected single-copy genes with <5% redundancy.



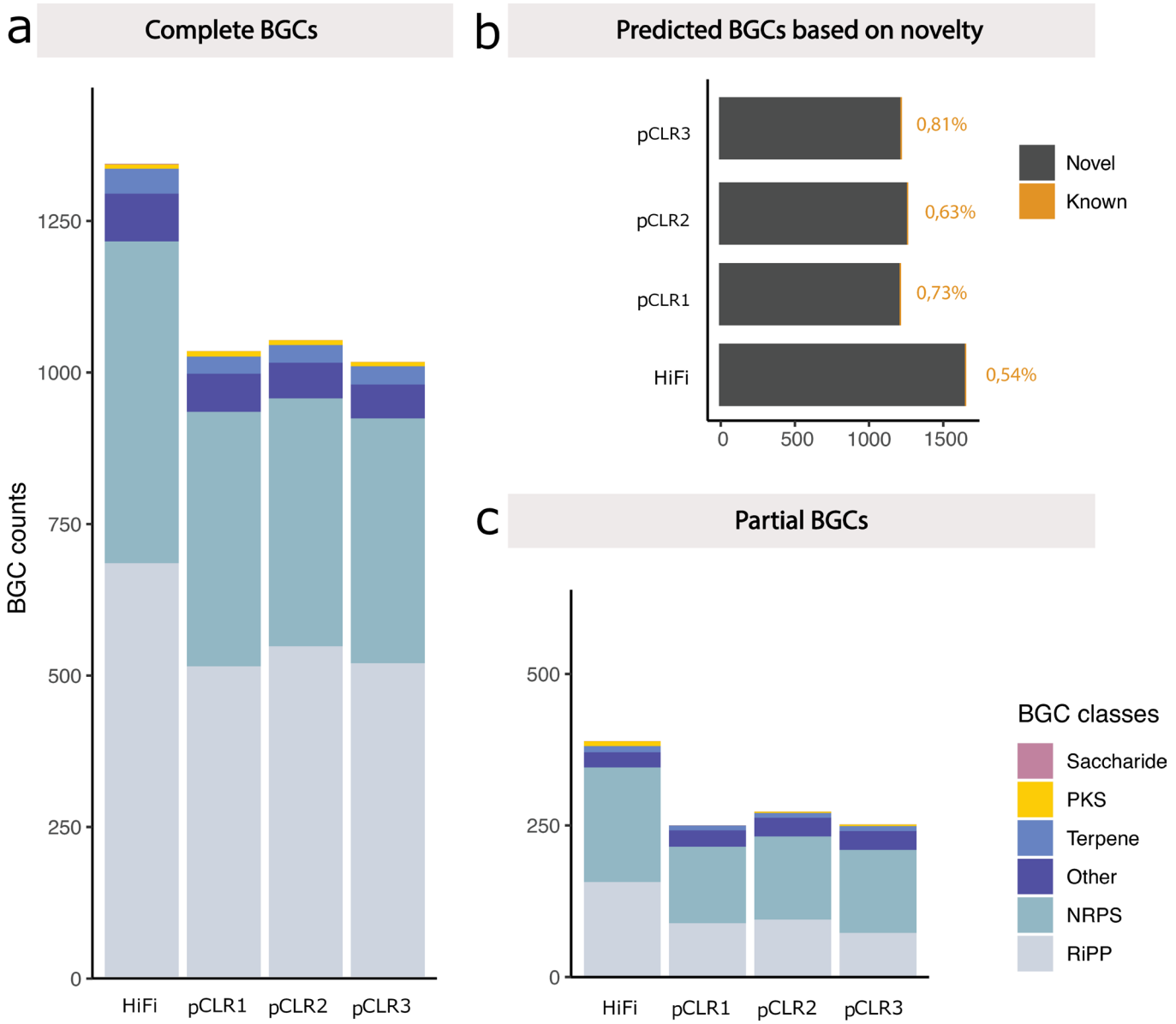
Extended Data Fig. 2 | Assembled MAG taxonomy. A circular dendrogram showing the presence (blue) and absence (black) of GTDB-TK assigned taxonomy to Assembly bins for the HiFi (outermost ring) and CLR (innermost rings, descending) assemblies. Branch nodes were consolidated to Genus-level affiliations when possible. Branch colors were assigned based on Phylum-level classification, with the exception of the Firmicutes, which was sub-divided into separate classes due to its increased diversity relative to other Phyla.



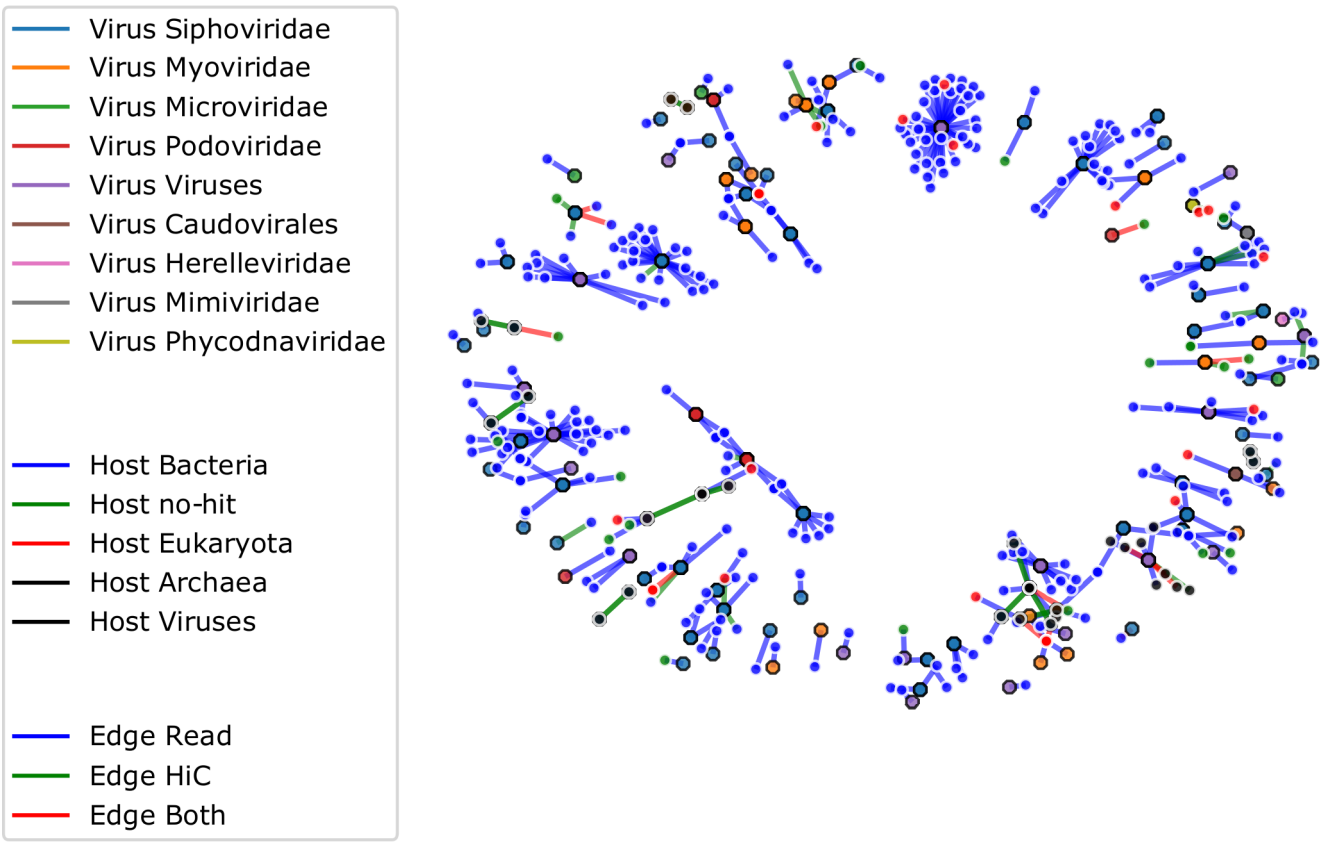
Extended Data Fig. 3 | Read depth across orthologous, collapsed pCLR bins. Each bin from separate, replicate pCLR assemblies corresponds to all three HiFi bins displayed in Supplementary Figure 6. Read depth that can be attributed to the reference sequence is labeled in blue, whereas phased alternative haplotypes identified via MAGPhase are labelled in alternating colors (see legend). Contig ends are denoted by vertical black bars and the x-axis represents the total length of the entire MAG with contigs placed randomly from end-to-end.



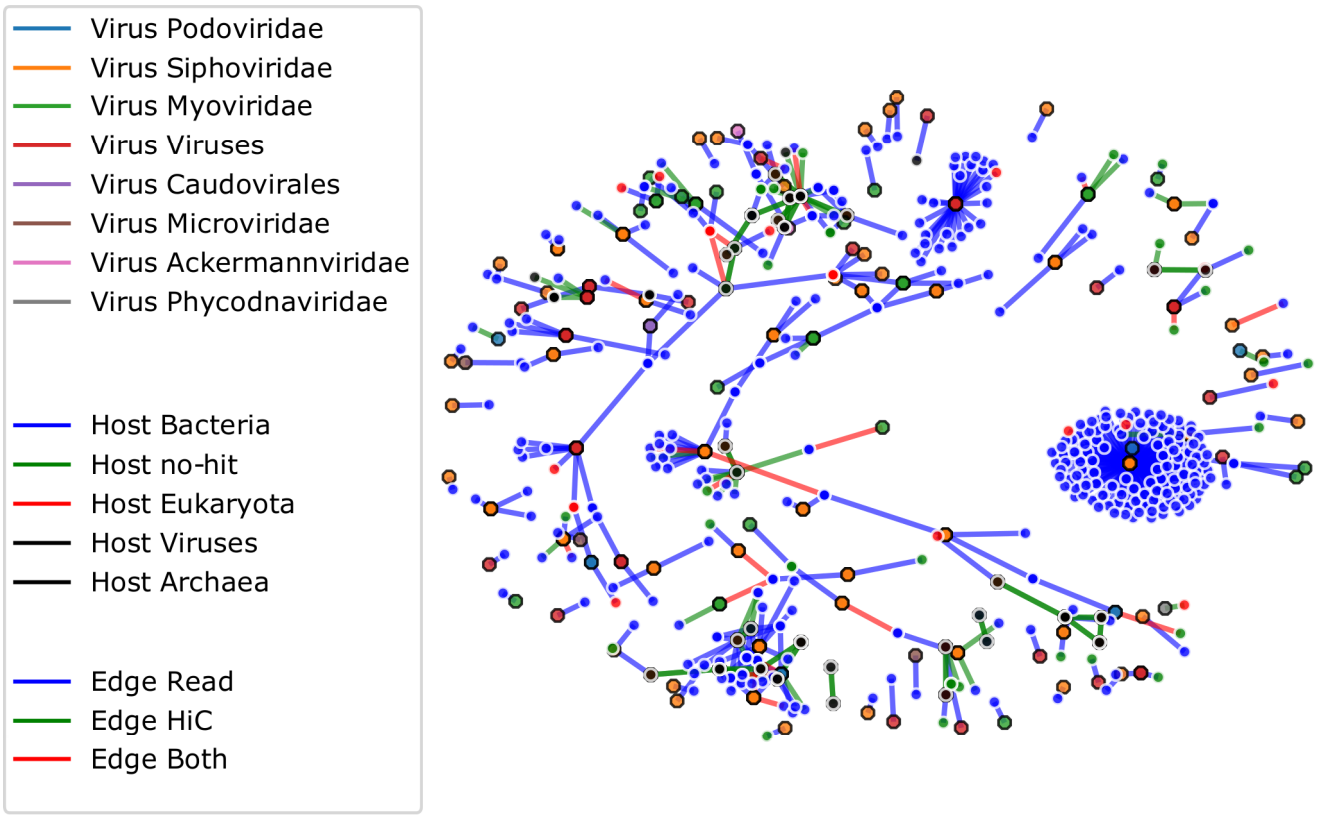
Extended Data Fig. 4 | Read depth across three closely related HiFi Complete MAGs. Read depth that can be attributed to the reference sequence is labeled in blue, whereas phased alternative haplotypes identified via MAGPhase are labelled in alternating colors (see legend). Contig ends are denoted by vertical black bars and the x-axis represents the total length of the entire MAG with contigs placed randomly from end-to-end.



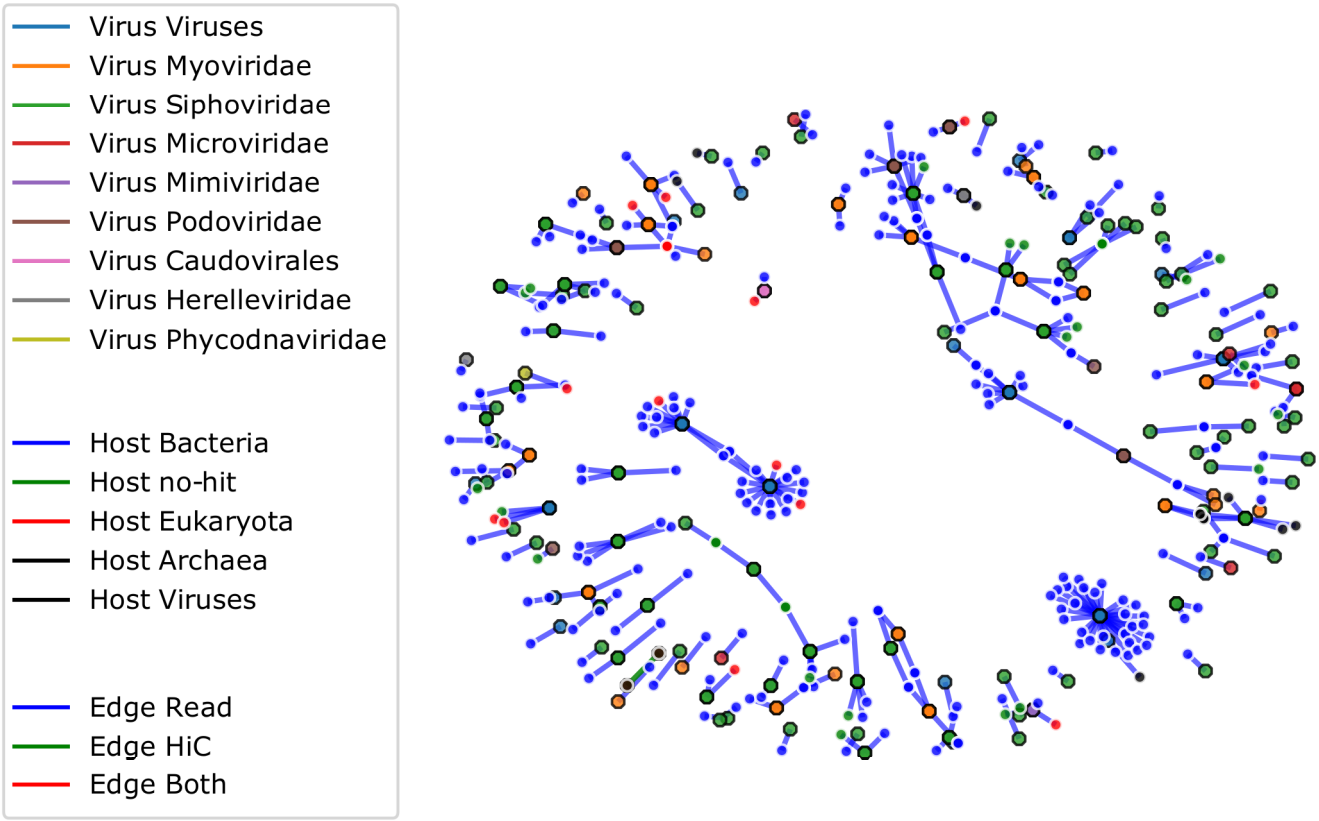
Extended Data Fig. 5 | Biosynthetic Gene Cluster Analysis. The HiFi assembly revealed approximately 25% more complete Biosynthetic Gene Clusters (BGCs) than the average pCLR assembly (a). This increase was manifested in all identified BGC classes (colors in legend) and was not exclusive to one particular class. As found in other metagenome assembly datasets, the majority of identified BGCs were novel in all assemblies (b), but the HiFi assembly had a higher proportion of novel BGCs than the other assemblies. Additionally, the HiFi assembly contained more partial BGCs (c) of any assembly.



Extended Data Fig. 6 | CLR1 viral association network plot. Viral contigs identified from Blobtools-assigned taxonomy estimates are represented as hexagonal nodes with black borders, whereas non-viral host contigs are represented as circular nodes with white borders. Edges represent associations identified for each connection, with colors representing the identification of partial HiFi read overlap (blue), Hi-C read links (green) or both types of data (red), respectively.



Extended Data Fig. 7 | CLR2 Viral association network plot. Viral contigs identified from Blobtools-assigned taxonomy estimates are represented as hexagonal nodes with black borders, whereas non-viral host contigs are represented as circular nodes with white borders. Edges represent associations identified for each connection, with colors representing the identification of partial HiFi read overlap (blue), Hi-C read links (green) or both types of data (red), respectively.



Extended Data Fig. 8 | CLR3 Viral association network plot. Viral contigs identified from Blobtools-assigned taxonomy estimates are represented as hexagonal nodes with black borders, whereas non-viral host contigs are represented as circular nodes with white borders. Edges represent associations identified for each connection, with colors representing the identification of partial HiFi read overlap (blue), Hi-C read links (green) or both types of data (red), respectively.

nature portfolio

Corresponding author(s): Pavel A. Pevzner, and Timothy P.L. Smith

Last updated by author(s): Oct 11, 2021

Reporting Summary

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

Statistics

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

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

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

Software and code

Policy information about availability of computer code

Data collection	Software used in data collection is listed in Supplementary Table 10
Data analysis	Third party software used in data analysis is also listed in Supplementary Table 10. Open source software developed by the authors has been made available via Github repositories (cDNACupcake: https://github.com/Magdoll/cDNA_Cupcake; Manuscript scripts: https://github.com/njdbickhart/SheepHiFiManuscript). The version of manuscript scripts used to analyze the data presented is hosted on Zenodo via this DOI: 10.5281/zenodo.5120910

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

Data

Policy information about availability of data

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

- Accession codes, unique identifiers, or web links for publicly available datasets
- A description of any restrictions on data availability
- For clinical datasets or third party data, please ensure that the statement adheres to our policy

The HiFi Sheep dataset, Hi-C reads and WGS short-reads are available on NCBI Bioproject PRJNA595610 at accession ids SRX7628648, SRX10704191, and SRX7649993, respectively. Whole metagenome assemblies and MAG bins for the pCLR and HiFi datasets are available at the following DOI: https://doi.org/10.5281/zenodo.4729049. The "kaiju_db_nr_euk_2021-02-24" database was used for kaiju classification (https://kaiju.binf.ku.dk/server). The "2017-07" version of the Uniprot database was used for Blobtools classification (https://ftp.uniprot.org/pub/databases/uniprot/previous_major_releases/release-2017_07/).

Field-specific reporting

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

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

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

Life sciences study design

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

Sample size	A single sample was sequenced to high depths of coverage using PacBio HiFi data. The expense of data generation at the time precluded the sequencing of other samples, but the depth of coverage was sufficient to create several replicate samples by downsampling to test thresholds of effective coverage.
Data exclusions	Only HiFi reads that did not meet quality thresholds automatically determined by the SMRTLink software suite (PacBio, USA) were excluded from analysis.
Replication	The authors performed a downsampling experiment to estimate coverage thresholds for the assembly of high quality MAGs. This was conducted using deciles of the original dataset, and all downsampled assemblies were processed using an identical workflow to determine the influence of coverage on assembly quality. Each downsampled decile was created only once, with read counts and read bases tabulated to ensure that the dataset was appropriately sectioned into fractions of ten. All downsampled datasets were within 0.1% of the expected fraction of reads, indicating successful fractionation of the data.
Randomization	Samples were not randomized given the small sample size.
Blinding	Blinding was not relevant to this study as it did not compare samples under different treatment effects and it only sequenced one sample to extreme depths of coverage. No aspect of the study could be made anonymous to preclude observer bias in the interpretation of results.

Reporting for specific materials, systems and methods

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

Materials & experimental systems

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

Methods

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

Animals and other organisms

Policy information about studies involving animals; ARRIVE guidelines recommended for reporting animal research

Laboratory animals	This study did not involve laboratory animals.
Wild animals	This study did not involve wild animals.
Field-collected samples	A single fecal sample was collected postmortem from a Katahdin sheep lamb that died on pasture at the United States Department of Agriculture's Meat Animal Research Center in Clay Center, NE, USA. The lamb was born at 4.5 kg as a twin May 12, 2018 in a lambing barn on an operational livestock ranch. The animal was monitored by animal husbandry staff and checked by veterinarian at 7 days for health with no problems observed. The lamb was released onto pasture with his twin and the maternal ewe on May 20, 2018. The pasture environment contains controls for coyote predation and was lush between May and July; at no time was feed limited by pasture conditions and unlimited water was available through pumped water tanks. The pasture was located in a temperate part of the United States with daily low and high temperatures between 0 and 39 C during the life of the animal. The animal was weaned from the ewe on July 16, 2018 at 13.2 kg, and given prophylactic treatment with Cydectin and Valbazen for worm parasites endemic to local pasture before moving to a feedlot setting where a finishing ration and unlimited water was provided. A generalized pneumonia of low severity was observed by animal handling staff on July 31, 2018 and treatment with oxytetracycline injection (BioMycin 200) administered. The animal appeared to recover but two weeks later was found dead, which autopsy August 13 (three

	months of age) revealed substantial worm parasite infestation diagnosed as combined strongyloides and coccidia infection. The sample used in the experiment was collected from the lower intestine of the animal during the autopsy to determine cause of death.
Ethics oversight	The United States Department of Agriculture's Agricultural Research Service (ARS) provided ethical guidance in the collection of this sample. Specifically, the sample was collected postmortem following the USDA ARS IACUC protocol #137.0 during a routine necropsy to determine cause of death.

Note that full information on the approval of the study protocol must also be provided in the manuscript.