Exploring high-quality microbial genomes by assembly of linked-reads with high barcode specificity using deep learning

Zhenmiao Zhang[1], Hongbo Wang[1], Chao Yang[1], Yufen Huang[2,3], Zhen Yue[2,4], Yang Chen[5], Lijuan Han[6], Aiping Lyu[7], Xiaodong Fang[2,3,5,*], and Lu Zhang[1,*]

[1]Department of Computer Science, Hong Kong Baptist University, Hong Kong, China
[2]BGI Genomics, BGI-Shenzhen, Shenzhen, China
[3]BGI-Shenzhen, Shenzhen, China
[4]BGI-Sanya, BGI-Shenzhen, Sanya, China
[5]State Key Laboratory of Dampness Syndrome of Chinese Medicine, The Second Affiliated Hospital of Guangzhou University of Chinese Medicine, Guangzhou, China
[6]Department of Scientific Research, Kangmeihuada GeneTech Co.,Ltd (KMHD), Shenzhen, China
[7]School of Chinese Medicine, Hong Kong Baptist University, Hong Kong, China
[*]To whom correspondence should be addressed, email: fangxd@genomics.cn, ericluzhang@hkbu.edu.hk

ABSTRACT

De novo assembly of metagenomic sequencing data plays an essential role in elucidating the genomes of unculturable microbes. Linked-reads, in which short-reads are linked together by barcodes that mark a long original DNA fragment, are a promising method for cost-effective metagenome assembly. Recently, the original linked-read sequencing platform from 10X genomics was discontinued; however, single-tube Long Fragment Read (stLFR) and Transposase Enzyme-Linked Long-read Sequencing (TELL-Seq) are another two linked-read sequencing platforms, which are designed with high barcode specificity and have the potential to efficiently deconvolve complex microbial communities.

We developed Pangaea, a metagenome assembler that assembles linked-reads with high barcode specificity using deep learning. It adopts a fast binning strategy to group linked-reads using a variational autoencoder, followed by rescue of low-abundance microbes with multi-thresholding reassembly. We sequenced a 20-strain-mixed mock community using 10x, stLFR, and TELL-Seq, and stool samples from two healthy human subjects using stLFR. We compare the performance of Pangaea with Athena, Supernova, and metaSPAdes. For the mock community, we observed that the assemblies from Pangaea on stLFR and TELL-Seq linked-reads achieved substantially better contiguity than the assemblies on 10x linked-reads, indicating that barcode specificity is a critical factor in metagenome assembly. We also observed Pangaea outperformed the other three tools on both stLFR and TELL-Seq linked-reads. For the human gut microbiomes, Pangaea still achieved the highest contiguity and considerably more near-complete metagenome-assembled genomes (NCMAGs) than the other assemblers. For the two human stool samples, Pangaea generated more NCMAGs than metaFlye on PacBio long-reads, as well as two complete and circular NCMAGs, demonstrating its ability to generate high-quality microbial reference genomes.

Introduction

Metagenome assembly is a common approach used to reconstruct microbial genomes from culture-free metagenomic sequencing data[1]. Inexpensive short-read sequencing approaches have been widely applied to generate high-quality microbial reference genomes from large cohorts of human gut metagenomic sequencing data[2–4], but the short read lengths (100-300 bp) limit their ability to achieve complete genomes, or to resolve intra-species repetitive regions and inter-species conserved regions[5]. Alternatively, long-read sequencing technologies, such as Oxford Nanopore long-reads[6], PacBio continuous (CLR)[7] and HiFi[8] long-reads, have shown superiority to short-reads in generating complete and circular microbial genomes from metagenomic sequencing data[9–11]. However, the low base quality, high cost, and requirement of a large amount of input DNAs still prevent long-read sequencing from being applied to population-scale or clinical studies. In our previous study[12], we observed that long-reads generated fewer high-quality metagenome-assembled genomes (MAGs) than short-reads due to insufficient sequencing depth, suggesting that considerable loss of information occurred when using long-reads in current sequencing settings, despite the assemblies having high contiguity.

Linked-read sequencing technologies combine merits of both short- and long-read sequencing, providing low base errors and long-range DNA information. They tag identical barcodes to short-reads derived from the same long DNA fragment.

Before its discontinuation, 10x Chromium was the most widely used linked-read sequencing platform, generating contigs with high contiguity and producing more near-complete metagenome-assembled genomes (NCMAGs; **Methods**) than short-read sequencing[13]. Two assemblers have been developed for metagenome assembly using 10x linked-reads: (i) Athena[13], which fills the gaps between contigs by recruiting the co-barcoded reads for local assembly, and (ii) cloudSPAdes[14], which reconstructs the long DNA fragments in the assembly graph by solving the shortest superstring problem to improve contiguity. Although 10x linked-reads have shown significant potential for metagenome assembly, the inherent technical issues complicate the deconvolution of complex microbial communities. 10x Chromium assigns long DNA fragments into droplets through a microfluidic system, wherein the number of fragments per droplet or barcode ($N_{F/B}$) follows a Poisson distribution. In other words, reads with identical barcodes may be derived from long DNA fragments ($N_{F/B}$=16.61; **Supplementary Note 1**) corresponding to different microbes. This characteristic would introduce off-target reads in local assembly (for Athena) and complicate the reconstruction of long fragments in the assembly graph (for cloudSPAdes). In addition, 10x Genomics has discontinued support for its genome product, posing a pressing requirement to devise alternative linked-read sequencing platforms with high barcode specificity.

Recently, MGI and Universal Sequencing Technology (UST) released their linked-read sequencing platforms, single-tube Long Fragment Read (stLFR)[15] and Transposase Enzyme-Linked Long-read Sequencing (TELL-Seq)[16], respectively. These platforms have shown comparable performance to 10x Chromium in human variant phasing and genome assembly[15,16]. However, unlike 10x Chromium, the reactions in stLFR and TELL-Seq occur in polymerase chain reaction (PCR) tubes without the need for expensive instrumentation. Essentially, the barcoding reactions occur on billions of microbeads in a single tube, leading to high barcode specificity ($N_{F/B}$=1.54 for stLFR, $N_{F/B}$=4.26 for TELL-Seq; Figure 1 a; **Supplementary Note 1**). MetaTrass[17] was recently developed for metagenome assembly on stLFR linked-reads; this tool groups the linked-reads by taxonomic annotation and applies Supernova[18] to assemble the genome of each identified species. Such a reference-based assembly tool is sensitive to the qualities of reference genomes and thus has a restricted ability to discover novel species. There is still a lack of an efficient tool that could fully exploit the high barcode specificity and long-range DNA information of stLFR and TELL-Seq linked-reads to improve the *de novo* metagenome assembly.

In this paper, we present Pangaea (Figure 1 b), a metagenome assembler developed to assemble linked-reads with high barcode specificity using deep learning. Pangaea is inspired by the reads binning strategy, which has been proven to facilitate short-read metagenome assembly[19–21]. However, existing short-read binning tools are unable to process millions of short-reads, and the features used for binning are unstable. To cope with these problems, Pangaea applies barcode binning instead of grouping short-reads and uses the *k*-mer frequencies and tetranucleotide frequency (TNF) from the co-barcoded reads as the inputs to a variational autoencoder (VAE) for learning low-dimensional latent embeddings in linear time. Pangaea applies a random projection hashing based k-means algorithm (RPH-kmeans[22]) to the latent embeddings to group the co-barcoded reads. In practice, the linked-reads from the same species could be assigned to different bins, which may result in poor assembly for low-abundance microbes. Therefore, Pangaea adopts a multi-thresholding reassembly strategy to refine the contigs by reassembling the linked-reads with different abundance thresholds (**Methods**).

To validate our approach, we sequenced a 20-strain-mixed mock community (ATCC-MSA-1003) using 10x Chromium, TELL-Seq, and stLFR to generate linked-reads, and compared the performance of Pangaea with that of two linked-read *de novo* assemblers – Athena and Supernova, and a short-read assembler – metaSPAdes. We found that Pangaea achieved substantially better contiguity than the second-best assembler Athena on both stLFR and TELL-Seq linked-reads. Pangaea on stLFR and TELL-Seq linked-reads produced contigs with higher contiguity than the other linked-read assemblers on 10x linked-reads. We also sequenced human gut microbiomes from two stool samples (S1 and S2) using stLFR and showed that Pangaea significantly outperformed Athena, Supernova, and metaSPAdes. Further, Pangaea produced significantly more near-complete metagenome-assembled genomes (NCMAGs) than the other three assemblers and generated two complete and circular microbial genomes that were not found in the assemblies generated by the other assemblers. Pangaea even produced many more NCMAGs than metaFlye on PacBio CLR long-reads from the human stool samples.

Results

Metagenome assembly of linked-reads using Pangaea

Pangaea is designed for metagenome assembly of linked-reads with high barcode specificity (e.g. stLFR and TELL-Seq linked-reads) using deep learning (Figure 1 b; **Methods**). It collects all of the linked-reads with the same barcodes and represents them using *k*-mer frequencies and TNFs (**Methods**) to overcome feature instability in grouping short-reads individually. These features enable the VAE to represent barcodes in low-dimensional latent space, where the latent variable follows a standard Gaussian distribution (**Methods**; **Supplementary Note 2**). We observed that the number of barcodes per species was highly correlated with the species' abundance, which may lead to the VAE being dominated by high-abundance species. Pangaea adopts a weighted sampling strategy to balance the number of co-barcoded reads from different species in each training batch based on their *k*-mer frequencies (**Methods**). It groups the co-barcoded reads in the latent space using RPH-kmeans, which is

scalable to large binning tasks by random projection hashing. Pangaea assembles the linked-reads in each cluster independently using MEGAHIT. We observed that linked-reads from the same microbe may be dispersed into multiple bins, which may substantially influence the assembly of low-abundance microbes. Thus, we designed a multi-thresholding reassembly approach (Figure 1 b) to improve the assemblies of low-abundance microbes. This approach aligns the linked-reads to the contigs and removes them if they are from contigs with abundances above certain thresholds (**Methods**). The remaining reads are used to reassemble the contigs of low-abundance microbes, which are finally combined with the contigs assembled for each cluster, and the local assembly contigs from Athena (**Methods**).

Barcode specificity of linked-reads is critical for metagenome assembly

We validated Pangaea on a mock microbial community, ATCC-MSA-1003, containing 20 strains mixed at different loadings (from 0.02% to 18%; **Supplementary Table 1**). ATCC-MSA-1003 was sequenced using stLFR and TELL-Seq linked-read sequencing, yielding 132.95 Gb and 173.28 Gb raw reads, respectively (**Supplementary Table 2**; **Methods**). We also used 10x linked-reads of ATCC-MSA-1003 from our previous study[23] (**Supplementary Table 2**). To assess barcode specificity, we aligned the linked-reads to the reference genomes, reconstructed the physical long fragments, and calculated $N_{F/B}$ (**Supplementary Note 1**; **Methods**). stLFR linked-reads yielded the lowest $N_{F/B}$ ($N_{F/B}$=1.54), and TELL-Seq linked-reads yielded a slightly higher number ($N_{F/B}$=4.26), although both numbers were much lower than that obtained from 10x linked-reads ($N_{F/B}$=16.61).

The contigs of Pangaea from the stLFR and TELL-Seq linked-reads had substantially higher N50s (24.84 times on average; Table 1; Figure2 c) and overall higher NA50s (11.81 times on average; Table 1) than the contigs of Athena and Supernova from 10x linked-reads. For the 15 strains with abundance $\geq$ 0.18% (**Supplementary Table 3**), Pangaea on stLFR and TELL-Seq linked-reads also achieved significantly higher per-strain NA50 (Figure 2 f) and NGA50 (Figure 2 i) than the assemblies of Athena and Supernova from 10x linked-reads. For the remaining 5 strains with abundance = 0.02%, Pangaea on stLFR (average genome fraction: 33.50%) and TELL-Seq (average genome fraction: 23.57%) linked-reads obtained much higher genome fractions than Athena (average genome fraction: 3.99%) and Supernova (average genome fraction: 6.29%) on 10x linked-reads (**Supplementary Table 4**). These results suggest that the linked-read technologies with high barcode specificity, stLFR and TELL-Seq, produce better metagenome assemblies with Pangaea than 10x Chromium.

Pangaea generated high-quality assembly on ATCC-MSA-1003

We compared Pangaea with Athena, Supernova, and metaSPAdes on the barcode-stripped TELL-Seq and stLFR linked-reads of the ATCC-MSA-1003 mock community (**Supplementary Table 2**). For TELL-Seq (Table 1; Figure 2 a), Pangaea achieved the highest N50 (1195.44 Kb) and NA50 (601.41 Kb) when compared with the statistics achieved by Athena (N50: 466.50 Kb; NA50: 361.57 Kb), Supernova (N50: 102.76 Kb; NA50: 97.31 Kb), and metaSPAdes (N50: 112.34 Kb; NA50: 105.63 Kb). When considering those 15 strains with abundance $\geq$ 0.18% (**Supplementary Table 3**), Pangaea still generated a significantly higher per-strain NA50 (Figure 2 d; **Methods**) and NGA50 (Figure 2 g) than Athena (NA50: p-value = 8.36e-3; NGA50: p-value = 8.36e-3), Supernova (NA50: p-value = 3.05e-4; NGA50: p-value = 3.05e-4), and metaSPAdes (NA50: p-value = 6.10e-5; NGA50: p-value = 6.10e-5). A comparable trend was observed for the assemblies of stLFR and TELL-Seq linked-reads (Table 1; Figure 2 b, e, and h), suggesting that Pangaea using linked-reads with high barcode specificity significantly improves contiguity compared to the other assemblers. For the strains with the lowest abundance (0.02%), the assemblies of Pangaea had much higher genome fractions than those of Athena (8.12 times on average) and Supernova (54.64 times on average) on stLFR and TELL-Seq linked-reads (**Supplementary Table 4**).

Pangaea generated high-quality assembly on the human gut microbiomes

We collected DNAs from two healthy Chinese individual stool samples and sequenced their gut microbiomes (S1 and S2) using stLFR, obtaining 136.6 Gb and 131.6 Gb raw reads, respectively (**Supplementary Table 2**; **Supplementary Figure 1**; **Methods**). The assemblies generated by Pangaea had the highest total assembly length among all of the benchmarked assemblers on both S1 (Pangaea = 488.79 Mb, Athena = 469.28 Mb, Supernova = 311.97 Mb, metaSPAdes = 452.60 Mb; Table 1) and S2 (Pangaea = 414.46 Mb, Athena = 393.69 Mb, Supernova = 290.60 Mb, metaSPAdes = 374.17 Mb; Table 1). Moreover, Pangaea achieved substantially higher N50s than the other three assemblers for both S1 (1.44 times of Athena; 1.06 times of Supernova; 4.50 times of metaSPAdes; Table 1) and S2 (1.61 times of Athena; 2.64 times of Supernova; 8.18 times of metaSPAdes; Table 1).

We grouped the contigs into MAGs and annotated NCMAGs for comparison (**Methods**). Pangaea generated 24 and 18 NCMAGs from S1 and S2 (Figure 3 a and e), which were much more than those generated by Athena (S1: 13 and S2: 12; Figure 3 a and e), Supernova (S1: 14 and S2: 10; Figure 3 a and e), and metaSPAdes (S1: 0 and S2: 1; Figure 3 a and e). Counting of the NCMAGs at different minimum values of N50 revealed that Pangaea obtained more NCMAGs than the other three assemblers at almost all N50 thresholds (Figure 3 b and f), demonstrating the high contiguity of NCMAGs generated by Pangaea. Pangaea also outperformed the other assemblers by counting the NCMAGs at different maximum values of the

abundance (Figure 3 c and g). Especially when the N50 of NCMAG was larger than 1 Mb (Figure 3 d and h), Pangaea achieved substantially more NCMAGs (S1: 8, S2: 4) than any of the other three assemblers at all abundance thresholds, while the second best assembler Athena only produced 3 and 1 NCMAG on S1 and S2, respectively.

Pangaea achieved high quality MAGs for different microbes

We annotated the MAGs using kraken2 with a custom database built from the Nucleotide (NT) database of the National Center for Biotechnology Information (NCBI; **Methods**). There were 43 selected microbes (S1: 26 and S2: 17) annotated from Pangaea's MAGs; 39 of them (S1: 23 and S2: 16) achieved the highest N50 (Figure 4) and 24 microbes had two-fold higher N50 than the second best assemblers (S1: 16 and S2: 8). For the remaining 4 microbes for which Pangaea did not get the highest N50 (Figure 4), it generated comparable N50s with the second best assembler on *Alistipes indistinctus*, *Oscillospiraceae*, and *Ruminococcus bicirculans* from S1; and achieved a lower N50 but substantially higher completeness than Supernova on *Roseburia hominis* from S2 (Pangaea: completeness = 96.54%, contamination = 0.48%; Supernova: completeness = 64.91%, contamination = 0.00%; **Supplementary Table 5**).

Moreover, the MAGs generated by Pangaea had higher MAG quality than those generated by Athena, Supernova, and metaSPAdes. There were 11 microbes (S1: 7 and S2: 4; Figure 4) that had NCMAGs generated by Pangaea where all the other assemblers only generated MAGs with lower quality or could not generate the matching MAGs. Pangaea generated the NCMAGs with N50s over 1 Mb for 13 microbes, where Athena, Supernova and metaSPAdes generated 4 ,1 and 0 microbes, respectively (Figure 4). In addition, Pangaea generated four unique microbes (*Bacteriophage* sp. and *Dialister* from S1, *Prevotella copri*, and *Parabacteroides* from S2) that were not generated by any other assemblers, and two of them were represented by NCMAGs (*Dialister* from S1 and *Parabacteroides* from S2; Figure 4). These results demonstrate that our barcode binning and assembly approach has the capability to recover high-quality unique genomes that might be lost by the off-the-shelf tools.

Strong collinearities between NCMAGs and their corresponding reference genomes

We aligned the NCMAGs assembled by Pangaea to their closest reference genomes to examine their collinearities (Figure 5; **Supplementary Figure 2**; **Methods**). The NCMAGs from Pangaea and their closest reference genomes had high alignment identities (average 98.04%), a stable alignment fraction (average 87.17%), and strong collinearity (Figure 5; **Supplementary Table 6**), suggesting that Pangaea generated assemblies with high base accuracy.

Inversions and genome rearrangements relative to reference sequences appeared in some bacterial genomes for both S1 and S2, including *A. communis* (S1; **Supplementary Figure 2** a), *Siphoviridae* sp. (S1; **Supplementary Figure 2** j), *Alistipes* sp. (S2; Figure 5 b) and *A. indistinctus* (S2; Figure 5 h). Pangaea-assembled NCMAGs for *Alistipes* sp. from both S1 and S2 (Figure 5 a and b) had comparable total sequence lengths (S1: 2.84 Mb and S2: 2.75 Mb; **Supplementary Table 5**), but better N50 was achieved in S1 (N50: 2344.71 Kb for S1 and 513.55 Kb for S2; **Supplementary Table 5**). This result might be due to the different abundance of *Alistipes* sp. in these two samples (read depth: 210.87x for S1 and 69.82x for S2; **Supplementary Table 5**).

Pangaea could generate NCMAGs with higher quality and larger N50 than the other assemblers, such as *Sutterella wadsworthensis* from S1 and P. *copri* from S2 (Figure 5 i and d **Supplementary Table 5**). Further, evaluation of the read depths and GC-skew of the MAGs revealed that Pangaea recovered the regions with extremely low read depths and high GC-skew, such as the region at approximately 1,100 Kb of *R. hominis* from S2 (Figure 5 f). This indicates that Pangaea has potential to reveal hard-to-assemble genomic regions.

Pangaea generated complete and circular MAGs

We next examined if there existed completed and circularized genomes in NCMAGs from the four tools using the circularization module in Lathe[24] (**Methods**). We found that only Pangaea generated two circular NCMAGs, which were annotated as *B. adolescentis* and *Myoviridae* sp. (Figure 5 e and g), respectively. For both of the two microbes, Pangaea generated a gapless contig with perfect collinearity with the closest reference genomes (Figure 5 e and g).

Athena generated three and two contigs for *B. adolescentis* and *Myoviridae* sp., with substantially lower contig N50 than those of the contigs obtained by Pangaea (*B. adolescentis*: Pangaea = 2167.94 Kb, Athena = 744.54 Kb; *Myoviridae* sp.: Pangaea = 2137.66 Kb, Athena = 1709.63 Kb; **Supplementary Table 5**). Supernova and metaSPAdes could only generate incomplete MAGs or could not assemble these two species, and the completeness of their candidate MAGs was significantly lower than that of MAGs generated by Pangaea (**Supplementary Table 5**).

Pangaea generated more NCMAGs than PacBio long-read sequencing

We compared the assemblies from Pangaea with those from metaFlye on PacBio CLR long-reads of S1 and S2 (**Supplementary Table 2**; **Supplementary Figure 1**; **Methods**). Although metaFlye generated contigs with higher N50s, Pangaea produced a substantially greater total assembly length for both S1 (Pangaea = 488.19 Mb, metaFlye = 243.88 Mb) and S2 (Pangaea =

414.46 Mb, metaFlye = 256.78 Mb; **Supplementary Table 7**). Moreover, Pangaea generated significantly more NCMAGs than metaFlye (Pangaea = 42, metaFlye = 16; Figure 3 i), especially those with N50s < 1 Mb (Pangaea = 30, metaFlye = 4; Figure 3 j and l) and read depths < 300x (Pangaea = 26, metaFlye = 0; Figure 3 k), whereas Pangaea and metaFlye obtained comparable numbers of NCMAGs with N50 > 1 Mb (Pangaea = 12, metaFlye = 12; Figure 3 j).

Discussion

Short-read sequencing of short metagenomic fragments has proven to be an important approach for analyzing human gut microbiota from large sequencing cohorts. However, its lack of long-range information makes assembling conserved sequences, intra- and inter-species repeats, and ribosomal RNAs (rRNAs) difficult[5]. As a result, it has limitations in producing complete microbial genomes. Cost-effective linked-read sequencing platforms, which attach barcodes to short-reads to provide long-range DNA connectedness, have achieved great success in improving contiguity in metagenome assembly[13,14]. Unlike 10x linked-reads, stLFR[15] and TELL-Seq linked-reads[16] have high barcode specificity, but a dedicated assembler that could make full use of high barcode specificity to improve metagenome assembly is lacking.

In this study, we developed Pangaea to improve metagenome assembly of linked-reads with high barcode specificity based on deep learning. Pangaea includes two key steps: co-barcoded read binning and multi-thresholding reassembly for low-abundance microbes. Inspired by long-read binning tools, Pangaea considers the co-barcoded linked-reads as long-reads and extracts their *k*-mer frequencies and TNFs for linked-read clustering. This strategy significantly reduces the complexity in metagenome sequencing and makes the assembly more efficient. Because clustering is sensitive to data sparsity and noise[25], Pangaea represents the input features in low-dimensional latent space using a VAE, which has been proven to be successful in contig binning. We also designed a weighted sampling strategy to generate a balanced training set for microbes with different abundances. The low-abundance species may have only a few reads in raw sequencing data, assembly of which greatly relies on the binning accuracy. Losing a small number of reads could result in fragmented contigs and low genome coverage. Pangaea adopts a multi-thresholding reassembly strategy to rescue the incorrectly assigned reads from low-abundance microbes.

Several studies have attempted to apply the read binning strategy to short-read / short fragment metagenomic sequencing[19–21], but it is exceedingly difficult in practice. The fragments are too short to allow the extraction of stable sequence abundance and composition features from the individual reads. Therefore, existing reads binning tools have to identify the overlap between every pair of reads for binning. However, the millions or even billions of short-reads make the overlap-based reads binning algorithm extremely slow and highly memory intensive. Overlap Graph-based Read clustEring (OGRE) was developed to improve the computational performance of reads binning, but it still consumed 2,263 CPU hours even for the low-complexity dataset of CAMI[19]. We tested OGRE on our mock community sequenced by stLFR (664.77M read pairs) and observed that OGRE crashed due to insufficient memory if 100 threads were applied. If fewer threads were applied, the binning time would become extremely long. In comparison, Pangaea with 100 threads only took 64.06 hours in real time and 514.63 hours in CPU time and consumed 281.99 Gb of random-access memory (RAM) to group and assemble the linked-reads from the mock community sequenced by stLFR.

The VAE was successfully applied to contigs and long-reads binning[25,26] and showed better binning performance than classical dimensional reduction algorithms such as principal component analysis (PCA; **Supplementary Figure 3**). For clustering linked-reads in the latent space of VAE, the classical k-means was not optimized to process the highly imbalanced metagenomic data due to its instability to choose proper initial centroids. We adopted RPH-kmeans[22] that used a random projection hashing strategy to solve this problem and was also time-efficient when dealing with large datasets. We applied RPH-kmeans, k-means and the Gaussian mixture model to group co-barcoded reads in the latent space of VAE using stLFR linked-reads from the mock community (**Supplementary Figure 4**). We observed RPH-k-means achieved a better overall F1 score and adjusted rand index (ARI) than the other two algorithms. Large k may result in higher binning precision and lower recall (**Supplementary Figure 5**). However, the values of k have little effect on the final assembly (**Supplementary Figure 6**), suggesting the performance of Pangaea was robust to the number of clusters.

We compared Pangaea with two other linked-read assemblers, Athena[13] and Supernova[18]. Athena was developed for 10x linked-reads and improved contiguity by local assembly. Linked-read sequencing technologies with high barcode specificity, such as stLFR and TELL-Seq, can reduce off-target reads in local assembly but might not strongly influence the performance of Athena, which was only designed with a focus on connecting the contigs with sufficient depths. Supernova was originally designed for human genome assembly, with some internal parameters optimized for assembling diploid genomes. Another tool, cloudSPAdes[14], was developed for metagenome assembly of linked-reads with comparable performance to Athena. cloudSPAdes was not included for comparison because it required over 2 TB of memory to assemble stLFR linked-reads from our mock community. This was probably because stLFR includes considerably more barcodes than 10x Chromium, which could overload and crash cloudSPAdes.

Long-read sequencing has received increasing attention due to its ability to generate complete microbial genomes from complex communities. However, it is limited by a high cost and huge amount of initial DNA. In contrast, linked-read

sequencing is cost-effective and only requires a tiny amount of input DNA, and can thus be a complementary solution to long-read sequencing. In our experiments, we found that long-read assemblies had 61.90% fewer NCMAGs than linked-read assemblies from Pangaea, indicating that important microbes might be lost due to insufficient long-read sequencing depth. Similar observations have been reported in previous studies[24]. Although stLFR ($N_{F/B}$=1.54) and TELL-Seq ($N_{F/B}$=4.26) linked-reads had high barcode specificity in the mock community, we observed that a considerable fraction of barcodes still contained more than one fragment (stLFR = 37.02%, TELL-Seq = 72.95%), which could complicate the deconvolution of barcodes for existing linked-read assemblers. We believe that further protocol improvement for these technologies (e.g. increasing the number of beads) may further improve their metagenome assembly performance.

Methods

DNA preparation and linked-read sequencing

For the mock community, the microbial DNAs were extracted directly from the 20 Strain Staggered Mix Genomic Material (ATCC MSA-1003) without size selection using a QIAamp DNA stool mini kit (Qiagen, Valencia, CA, USA). For the human gut microbiomes, microbial DNAs from stool samples of two individuals (S1 and S2) were extracted using the QIAamp DNA stool mini kit (Qiagen) and size-selected using a BluePippin instrument targeting the size range of 10-50 Kb according to the manufacturer's protocol. The stLFR libraries were prepared using the stLFR library prep kit (16 RXN), followed by 2×100 paired-end short-read sequencing using BGISEQ-500. The TELL-Seq library for the mock community was prepared using the TELL-Seq Whole Genome Sequencing library prep kit, followed by 2×146 paired-end sequencing on an Illumina sequencing system.

Extract *k*-mer frequencies and TNFs from co-barcoded linked-reads

We extracted *k*-mer frequencies and TNFs from the co-barcoded linked-reads if their total lengths were longer than 2 Kb to ensure feature stability. The *k*-mer frequencies were calculated from the histogram of global *k*-mer occurrences, which followed a Poisson distribution with the mean equals the microbial abundance. We adopted $k = 15$ the same as previous studies[27,28] and built a 15-mer frequency table using all reads and stored it in an *unordered_map* data structure of C++ for fast searching. We removed all 15-mers with frequencies higher than 4,000 (to avoid duplicated sequences) and divided the remaining 15-mers into 400 bins with equal sizes. For each barcode, we sheared the co-barcoded linked-reads into 15-mers and assigned them to these 400 bins. We calculated the number of *k*-mers falling in each bin and generated a count vector with 400 dimensions as the *k*-mer frequencies of candidate barcode. A TNF vector was constructed by calculating the frequencies of all 136 non-redundant 4-mers for co-barcoded linked-reads. The *k*-mer frequencies and TNF vectors were L1-normalized to eliminate the bias introduced by the different lengths of co-barcoded linked-reads.

Binning co-barcoded linked-reads with a VAE

The normalized *k*-mer frequencies (X_A) and TNF vectors (X_T) were concatenated into a vector with 536 dimensions as the input to a VAE (Figure 1 b; **Supplementary Note 2**). The encoder of VAE consisted of two fully connected layers with 512 hidden neurons, and each layer was followed by batch normalization[29] and a dropout layer[30] (P = 0.2). The output of the last layer was fed to two parallel latent layers with 32 hidden neurons for each to generate μ and σ of a Gaussian distribution $N(\mu, \sigma^2)$, from which the embedding Z was sampled. The decoder also contained two fully connected hidden layers of the same size as the encoder layers to reconstruct the input vectors ($\hat{X}_A$ and $\hat{X}_T$) from the latent embedding Z. We applied the *softmax* activation function on $\hat{X}_A$ and $\hat{X}_T$ to achieve the normalized output vectors, because the input features X_A and X_T were both L1-normalized. The loss function (*Loss*) was defined as the weighted sum of three components: the reconstruction loss of *k*-mer frequencies (L_A), the reconstruction loss of TNF vectors (L_T), and the Kullback-Leibler divergence loss (L_{KL}) between the latent and prior standard Gaussian distributions. We adopted cross-entropy loss for L_A and L_T to deal with probability distributions, and all of the loss terms were formularized as follows:

$$L_A = \sum \ln(\hat{X}_A + 10^{-9}) X_A \tag{1}$$

$$L_T = \sum \ln(\hat{X}_T + 10^{-9}) X_T \tag{2}$$

$$L_{KL} = -\sum \frac{1}{2}(1 + \ln\sigma - \mu^2 - \sigma) \tag{3}$$

$$Loss = w_A L_A + w_T L_T + w_{KL} L_{KL} \tag{4}$$

where the weights of the three loss components were $w_A = \alpha/\ln(dim(X_A))$, $w_T = (1-\alpha)/\ln(dim(X_T))$, and $w_{KL} = \beta/dim(Z)$. We adopted 0.1 and 0.015 for the parameters α and β, respectively. The VAE was trained with early stopping to reduce the training time and avoid overfitting. We used the RPH-kmeans[22] algorithm with random projection hashing to group the

co-barcoded linked-reads using their latent embeddings obtained from μ. The numbers of clusters for the mock community, and the human gut microbiomes were set to 15 and 30, respectively.

Balancing the training dataset with weighted sampling

We designed a weighted sampling strategy to balance the training set of co-barcoded linked-reads from microbes with different abundances. Theoretically, the abundances of co-barcoded linked-reads can be estimated by a Poisson distribution from global k-mer occurrences. In practice, the distribution is not perfect due to sequencing errors. To infer the abundances from k-mer frequencies, we designed a time-efficient heuristic function, $H(X_A) = 1/Max(X_A)^2$, to estimate the abundances and used $1/H$ as the sampling weight for co-barcoded linked-reads. The sampling weights were automatically used by the WeightedRandomSampler of PyTorch to create a balanced dataset in each training batch.

Multi-thresholding reassembly for low-abundance microbes

We designed a multi-thresholding reassembly strategy (Figure 1 b) to improve the assembly qualities of low-abundance microbes by collecting the reads from the same microbes that were misclustered into different bins. We assembled all reads (contigs$_{ori}$) using metaSPAdes (v3.15.3)[31] and the reads from each cluster (contigs$_{bin}$) using MEGAHIT (v1.2.9)[32]. Then, we aligned all of the linked-reads to contigs$_{bin}$ using BWA (v0.7.17)[33] to calculate the read depth for each contig. The read depth was calculated using "jgi_summarize_bam_contig_depths" in MetaBat2 (v2.12.1)[34]. We next extracted the linked-reads that could not be mapped to the contigs$_{bin}$ with read depth $> t_i$ and assembled them using metaSPAdes (v3.15.3) with contigs$_{ori}$ as the –"--untrusted-contigs". We repeated this procedure with a range of thresholds ($T = \{t_i | i = 1, 2, ...\}$) and collected the resultant contigs as contigs$_{low}$. Finally, we used metaFlye (v2.8) "--subassemblies"[35] to merge contigs$_{bin}$, contigs$_{low}$, and the local assembly contigs from contigs$_{ori}$ to integrate the merits of all three contig sets. We also used quickmerge (v0.3)[36] to optimize the resulting contigs with the contigs from Athena using contigs$_{ori}$[13], as they were observed to be complementary. We used $T = \{10, 30, 50, 70, 90\}$, $T = \{10, 30\}$ and $T = \{10, 30, 50, 70\}$ for ATCC-MSA-1003, S1, and S2, respectively.

Circularization of Pangaea assembly

We used the circularization module of Lathe[24] to analyze the assemblies from Pangaea by regarding the contigs before metaFlye merging as pseudo long-reads (including contigs$_{bin}$, contigs$_{low}$, and contigs$_{ori}$ after local assembly). These pseudo long-reads were used to circularize the merged contigs and generate the final assembly of Pangaea.

Reconstructing physical long fragments based on reference genomes

We reconstructed the physical long fragments from linked-reads of the mock community to calculate $N_{F/B}$. The linked-reads were mapped to the reference genomes using BWA (v0.7.17)[33] with option "-C" to retain the barcode information in the alignment file, followed by sorting based on read alignment coordinates using SAMtools (v1.9)[37]. We connected the co-barcoded reads into long fragments if their coordinates were within 10 Kb on the reference genome. Each fragment was required to include at least two read pairs and to be no shorter than 1 Kb.

Assembly of 10x, TELL-Seq, and stLFR linked-reads and PacBio CLR long-reads

The 10x, stLFR, and TELL-Seq sequencing datasets were demultiplexed to generate raw linked-reads using Long Ranger (v2.2.0)[38], stLFR_read_demux[39] and LRTK[40], respectively. 10x and TELL-Seq linked-reads were assembled using metaSPAdes (v3.15.3)[31], Athena (v1.3)[13], and Supernova (v2.1.1)[18]. stLFR link-reads were assembled using metaSPAdes (v3.15.3), Athena (v1.3), and stlfr2supernova_pipeline[41] from BGI, because Supernova does not accept raw stLFR linked-reads as input. The scaffolds produced by Supernova were broken into contigs at successive Ns that were longer than 10 bp before evaluation. PacBio CLR long-reads from the two human gut microbiomes were assembled using metaFlye (v2.8)[35].

Benchmarking on the mock microbial community

The reference genomes of ATCC-MSA-1003 were downloaded from the NCBI reference databases (**Supplementary Table 1**). The contigs assembled from the mock community were assessed using MetaQUAST (v5.0.2)[42], with the option "--fragmented --min-alignment 500 --unique-mapping" to enable the alignment of fragmented reference genomes and discard ambiguous alignments. The p-values of differences in the NA50 and NGA50 of different assemblers were obtained using the Wilcoxon signed-rank test performed by the *wilcox.test* function of R with "paired=TRUE".

Contig binning and quality evaluation

We aligned the linked-reads to the contigs generated from the assemblers using BWA (v0.7.17)[33] and calculated the read depths using "jgi_summarize_bam_contig_depths" in MetaBat2 (v2.12.1)[34]. The contigs with read depths were binned into MAGs using MetaBat2 (v2.12.1) with default parameters. CheckM[43] was used to report the completeness and contamination of the MAGs. ARAGORN (v1.2.38)[44] and barrnap (v0.9)[45] were used to annotate the transfer RNAs (tRNAs) and rRNAs (5S, 16S,

and 23S rRNAs), respectively. According to the standard of minimum information about MAGs[46], we classified the MAGs into near-complete (completeness $> 90\%$, contamination $< 5\%$, and could detect 5S, 16S, and 23S rRNAs and at least 18 tRNAs), high-quality (completeness $> 90\%$, contamination $< 5\%$), medium-quality (completeness $\geq 50\%$, contamination $< 10\%$), and low-quality (the others).

Annotation of the MAGs and the closest reference genomes

The contigs were annotated using kraken2[47] with the custom database built from the NT database of NCBI. We used the "--fast-build" option of kraken2-build to reduce the database construction time. Subsequently, the "assign_species.py" script from "metagenomics_workflows"[13,24] was used to annotate MAGs as species (if the fraction of contigs belonging to the species was more than 60%) or genus (otherwise) based on contig annotations. The closest reference genomes of the NCMAGs that can be annotated at species-level were identified using GTDB-Tk (v2.1.0)[48], which also reported the alignment identities and alignment fractions between them.

Availability of data and materials

The 10x linked-reads of ATCC-MSA-1003 mock community was downloaded from NCBI run SRR12283286. The stLFR and TELL-Seq sequencing data of ATCC-MSA-1003 was uploaded to NCBI BioProject PRJNA875547. The stLFR sequencing data of the two human gut microbiomes was deposited in China National GeneBank (CNGB) project CNP0003432. Codes of Pangaea and all the command lines are available at https://github.com/ericcombiolab/Pangaea.

References

1. Yang, C. *et al.* A review of computational tools for generating metagenome-assembled genomes from metagenomic sequencing data. *Comput. Struct. Biotechnol. J.* **19**, 6301–6314 (2021).
2. Almeida, A. *et al.* A new genomic blueprint of the human gut microbiota. *Nature* **568**, 499–504 (2019).
3. Li, J. *et al.* An integrated catalog of reference genes in the human gut microbiome. *Nat. biotechnology* **32**, 834–841 (2014).
4. Almeida, A. *et al.* A unified catalog of 204,938 reference genomes from the human gut microbiome. *Nat. biotechnology* **39**, 105–114 (2021).
5. Yuan, C., Lei, J., Cole, J. & Sun, Y. Reconstructing 16s rrna genes in metagenomic data. *Bioinformatics* **31**, i35–i43 (2015).
6. Jain, M., Olsen, H. E., Paten, B. & Akeson, M. The oxford nanopore minion: delivery of nanopore sequencing to the genomics community. *Genome biology* **17**, 1–11 (2016).
7. Rhoads, A. & Au, K. F. Pacbio sequencing and its applications. *Genomics, proteomics & bioinformatics* **13**, 278–289 (2015).
8. Wenger, A. M. *et al.* Accurate circular consensus long-read sequencing improves variant detection and assembly of a human genome. *Nat. biotechnology* **37**, 1155–1162 (2019).
9. Vollger, M. R. *et al.* Long-read sequence and assembly of segmental duplications. *Nat. methods* **16**, 88–94 (2019).
10. 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 biology* **20**, 1–18 (2019).
11. Moss, E. L., Maghini, D. G. & Bhatt, A. S. Complete, closed bacterial genomes from microbiomes using nanopore sequencing. *Nat. biotechnology* **38**, 701–707 (2020).
12. Zhang, Z., Yang, C., Fang, X. & Zhang, L. Benchmarking de novo assembly methods on metagenomic sequencing data. *bioRxiv* (2022).
13. Bishara, A. *et al.* High-quality genome sequences of uncultured microbes by assembly of read clouds. *Nat. biotechnology* **36**, 1067–1075 (2018).
14. Tolstoganov, I., Bankevich, A., Chen, Z. & Pevzner, P. A. cloudspades: assembly of synthetic long reads using de bruijn graphs. *Bioinformatics* **35**, i61–i70 (2019).
15. Wang, O. *et al.* Efficient and unique cobarcoding of second-generation sequencing reads from long dna molecules enabling cost-effective and accurate sequencing, haplotyping, and de novo assembly. *Genome research* **29**, 798–808 (2019).
16. Chen, Z. *et al.* Ultralow-input single-tube linked-read library method enables short-read second-generation sequencing systems to routinely generate highly accurate and economical long-range sequencing information. *Genome research* **30**, 898–909 (2020).

17. Qi, Y. *et al.* Metatrass: High-quality metagenomic taxonomic read assembly of single-species based on co-barcoding sequencing data and references. *bioRxiv* (2021).

18. Weisenfeld, N. I., Kumar, V., Shah, P., Church, D. M. & Jaffe, D. B. Direct determination of diploid genome sequences. *Genome research* **27**, 757–767 (2017).

19. Balvert, M., Luo, X., Hauptfeld, E., Schönhuth, A. & Dutilh, B. E. Ogre: overlap graph-based metagenomic read clustering. *Bioinformatics* **37**, 905–912 (2021).

20. Wang, Y., Leung, H. C., Yiu, S.-M. & Chin, F. Y. Metacluster 5.0: a two-round binning approach for metagenomic data for low-abundance species in a noisy sample. *Bioinformatics* **28**, i356–i362 (2012).

21. Girotto, S., Pizzi, C. & Comin, M. Metaprob: accurate metagenomic reads binning based on probabilistic sequence signatures. *Bioinformatics* **32**, i567–i575 (2016).

22. Xie, K., Huang, Y., Zeng, F., Liu, Z. & Chen, T. scaide: clustering of large-scale single-cell rna-seq data reveals putative and rare cell types. *NAR genomics bioinformatics* **2**, lqaa082 (2020).

23. Zhang, L. *et al.* A comprehensive investigation of metagenome assembly by linked-read sequencing. *Microbiome* **8**, 1–11 (2020).

24. Moss, E. L., Maghini, D. G. & Bhatt, A. S. Complete, closed bacterial genomes from microbiomes using nanopore sequencing. *Nat. biotechnology* **38**, 701–707 (2020).

25. Nissen, J. N. *et al.* Improved metagenome binning and assembly using deep variational autoencoders. *Nat. biotechnology* **39**, 555–560 (2021).

26. Wickramarachchi, A. & Lin, Y. Binning long reads in metagenomics datasets using composition and coverage information. *Algorithms for Mol. Biol.* **17**, 1–15 (2022).

27. Wickramarachchi, A., Mallawaarachchi, V., Rajan, V. & Lin, Y. Metabcc-lr: meta genomics b inning by c overage and c omposition for l ong r eads. *Bioinformatics* **36**, i3–i11 (2020).

28. Ruan, J. & Li, H. Fast and accurate long-read assembly with wtdbg2. *Nat. methods* **17**, 155–158 (2020).

29. Ioffe, S. & Szegedy, C. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *International conference on machine learning*, 448–456 (PMLR, 2015).

30. Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I. & Salakhutdinov, R. Dropout: a simple way to prevent neural networks from overfitting. *The journal machine learning research* **15**, 1929–1958 (2014).

31. Nurk, S., Meleshko, D., Korobeynikov, A. & Pevzner, P. A. metaspades: a new versatile metagenomic assembler. *Genome research* **27**, 824–834 (2017).

32. Li, D., Liu, C.-M., Luo, R., Sadakane, K. & Lam, T.-W. Megahit: an ultra-fast single-node solution for large and complex metagenomics assembly via succinct de bruijn graph. *Bioinformatics* **31**, 1674–1676 (2015).

33. Li, H. Aligning sequence reads, clone sequences and assembly contigs with bwa-mem. *arXiv preprint arXiv:1303.3997* (2013).

34. Kang, D. D. *et al.* Metabat 2: an adaptive binning algorithm for robust and efficient genome reconstruction from metagenome assemblies. *PeerJ* **7**, e7359 (2019).

35. Kolmogorov, M. *et al.* metaflye: scalable long-read metagenome assembly using repeat graphs. *Nat. Methods* **17**, 1103–1110 (2020).

36. Chakraborty, M., Baldwin-Brown, J. G., Long, A. D. & Emerson, J. Contiguous and accurate de novo assembly of metazoan genomes with modest long read coverage. *Nucleic acids research* **44**, e147–e147 (2016).

37. Li, H. *et al.* The sequence alignment/map format and samtools. *Bioinformatics* **25**, 2078–2079 (2009).

38. Marks, P. *et al.* Resolving the full spectrum of human genome variation using linked-reads. *Genome research* **29**, 635–645 (2019).

39. Wang, O. *et al.* stLFR_read_demux. https://github.com/stLFR/stLFR_read_demux (2019).

40. Yang, C., Zhang, Z., Liao, H. & Zhang, L. Lrtk: A unified and versatile toolkit for analyzing linked-read sequencing data. *bioRxiv* (2022).

41. stlfr2supernova_pipeline. https://github.com/BGI-Qingdao/stlfr2supernova_pipeline.

42. Mikheenko, A., Saveliev, V. & Gurevich, A. Metaquast: evaluation of metagenome assemblies. *Bioinformatics* **32**, 1088–1090 (2016).

43. Parks, D. H., Imelfort, M., Skennerton, C. T., Hugenholtz, P. & Tyson, G. W. Checkm: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. *Genome research* **25**, 1043–1055 (2015).

44. Laslett, D. & Canback, B. Aragorn, a program to detect trna genes and tmrna genes in nucleotide sequences. *Nucleic acids research* **32**, 11–16 (2004).

45. Seemann, T. barrnap. https://github.com/tseemann/barrnap (2018).

46. Bowers, R. M. *et al.* Minimum information about a single amplified genome (misag) and a metagenome-assembled genome (mimag) of bacteria and archaea. *Nat. biotechnology* **35**, 725–731 (2017).

47. Wood, D. E., Lu, J. & Langmead, B. Improved metagenomic analysis with kraken 2. *Genome biology* **20**, 1–13 (2019).

48. Chaumeil, P.-A., Mussig, A. J., Hugenholtz, P. & Parks, D. H. Gtdb-tk: a toolkit to classify genomes with the genome taxonomy database (2020).

Acknowledgements

We thank Tom Chen and Yong Wang from Universal Sequencing Technology for providing the TELL-Seq sequencing data of the ATCC-MSA-1003 mock community. We thank Arend Sidow for his comments to improve manuscript language and structure. We also thank the Research Committee of Hong Kong Baptist University and the Interdisciplinary Research Clusters Matching Scheme for their kind support of this project.

Funding

The design of the study and the collection, analysis, and interpretation of the data were partially supported by the Science Technology and Innovation Committee of Shenzhen Municipality, China (SGDX20190919142801722). This research was partially supported by the Hong Kong Research Grant Council Early Career Scheme (HKBU 22201419), HKBU IRCMS (No. IRCMS/19-20/D02), and the Guangdong Basic and Applied Basic Research Foundation (No. 2019A1515011046 and No. 2021A1515012226).

Author contributions statement

LZ and XDF conceived the study. ZMZ and LZ designed the Pangaea algorithms. ZMZ implemented the Pangaea software. LZ and ZMZ conceived the experiments. ZMZ conducted the experiments. ZMZ and LZ analyzed the results. HBW drew and analyzed the circos plots. CY generated and analyzed the statistics of the three types of linked-reads. ZMZ and LZ wrote the manuscript. XDF, YFH, ZY, YC and LJH sequenced the stLFR linked-reads. APL revised the paper and supported the project. All authors reviewed the manuscript.

Supplementary information

Supplementary Text and Figures: Supplementary Notes 1-2, Supplementary Figures 1-6, and Supplementary Tables 1, 2, and 7

Supplementary Table 3: The per-strain NA50 and NGA50 for the 15 strains with abundance $\geq 0.18\%$ assembled by different assembly tools using stLFR, TELL-Seq, and 10x linked-reads from the mock community.

Supplementary Table 4: The genome fractions for the five strains with 0.02% abundance assembled by different assembly tools using stLFR, TELL-Seq, and 10x linked-reads from the mock community.

Supplementary Table 5: The statistics of the MAGs generated from the two human gut microbiomes by all assemblers.

Supplementary Table 6: The closest reference genomes of the NCMAGs assembled by Pangaea with species level annotations, and their alignment statics reported using GTDB-Tk (v2.1.0).

	ATCC-MSA-1003 (stLFR)			
	Pangaea	**Athena**	**Supernova**	**metaSPAdes**
Total assembly length	**58,259,182**	52,159,846	35,226,545	57,225,487
Genome fraction (%)	**85.05**	77.12	52.21	83.99
Longest alignment	**2,853,175**	2,281,647	1,105,108	883,552
Overall N50	**1,833,445**	875,747	243,194	132,556
Overall NA50	**732,394**	677,911	215,052	125,586
Strain average NGA50	**677,348.60**	575,370.80	137,023.30	133,977.90
Strain average NA50	**677,716.10**	576,620.90	145,645.60	134,476.55
	ATCC-MSA-1003 (TELL-Seq)			
	Pangaea	**Athena**	**Supernova**	**metaSPAdes**
Total assembly length	59,488,595	**60,847,375**	56,748,937	60,648,311
Genome fraction (%)	80.01	81.99	76.46	**82.46**
Longest alignment	**4,968,167**	4,968,084	1,096,372	776,102
Overall N50	**1,195,435**	466,498	102,757	112,342
Overall NA50	**601,408**	361,569	97,312	105,630
Strain average NGA50	**795,662.60**	485,195.60	121,656.85	118,390.75
Strain average NA50	**803,392.25**	483,734.00	123,276.80	119,252.65
	ATCC-MSA-1003 (10x)			
	Pangaea stLFR	**Pangaea TELL-Seq**	**Athena**	**Supernova**
Total assembly length	58,259,182	59,488,595	52,159,979	**89,828,047**
Genome fraction (%)	**85.05**	80.01	77.20	75.08
Longest alignment	2,853,175	**4,968,167**	2,278,020	974,529
Overall N50	**1,833,445**	1,195,435	596,076	32,128
Overall NA50	**732,394**	601,408	437,889	30,194
Strain average NGA50	677,348.60	**795,662.60**	338,609.40	89,993.55
Strain average NA50	677,716.10	**803,392.25**	337,372.75	93,097.65
	Human gut microbiome (S1)			
	Pangaea	**Athena**	**Supernova**	**metaSPAdes**
Total assembly length	**488,785,611**	469,284,964	311,971,769	452,598,342
Longest contig	2,394,379	2,394,379	**2,400,768**	697,064
Overall N50	**64,394**	44,759	60,619	14,325
	Human gut microbiome (S2)			
	Pangaea	**Athena**	**Supernova**	**metaSPAdes**
Total assembly length	**414,455,014**	393,685,495	290,599,879	374,166,135
Longest contig	**3,264,340**	1,903,088	1,152,844	443,089
Overall N50	**192,489**	119,620	72,947	23,546

Table 1. Assembly statistics for different assemblers on the ATCC-MSA-1003 mock community and human gut microbiomes. The highest values are in bold.

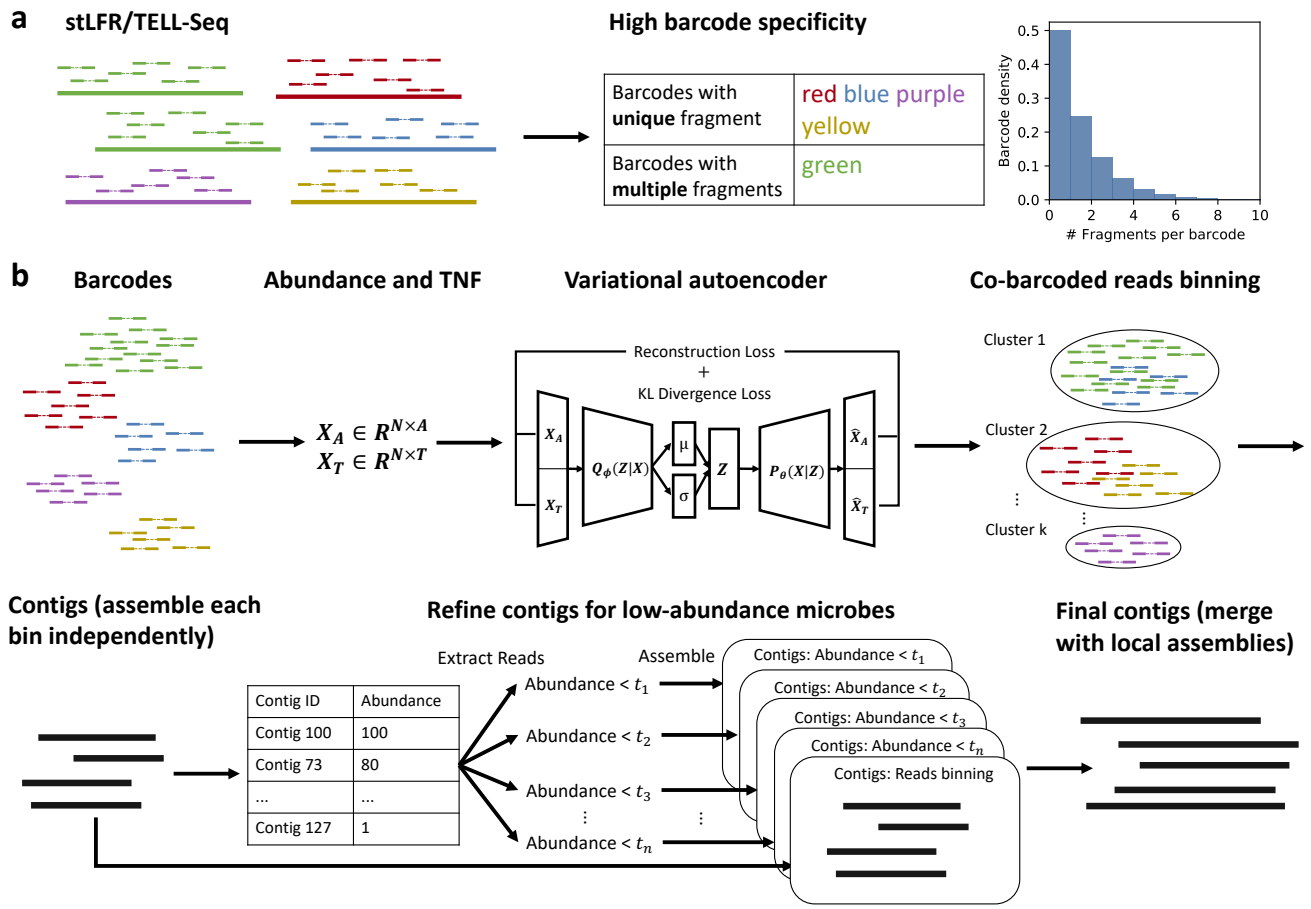

Figure 1. Workflow of Pangaea on stLFR and TELL-Seq linked-reads. **(a)** High barcode specificity for stLFR and TELL-Seq linked-reads. **(b)** Pangaea extracts the k-mer frequencies and TNF features from co-barcoded reads. The features are concatenated and used to represent data in low-dimensional latent space using a variational autoencoder. The embeddings of co-barcoded reads are clustered by RPH-kmeans. Pangaea assembles the linked-reads from each bin independently and adopts a multi-thresholding reassembly strategy to improve the assemblies for low-abundance microbes.

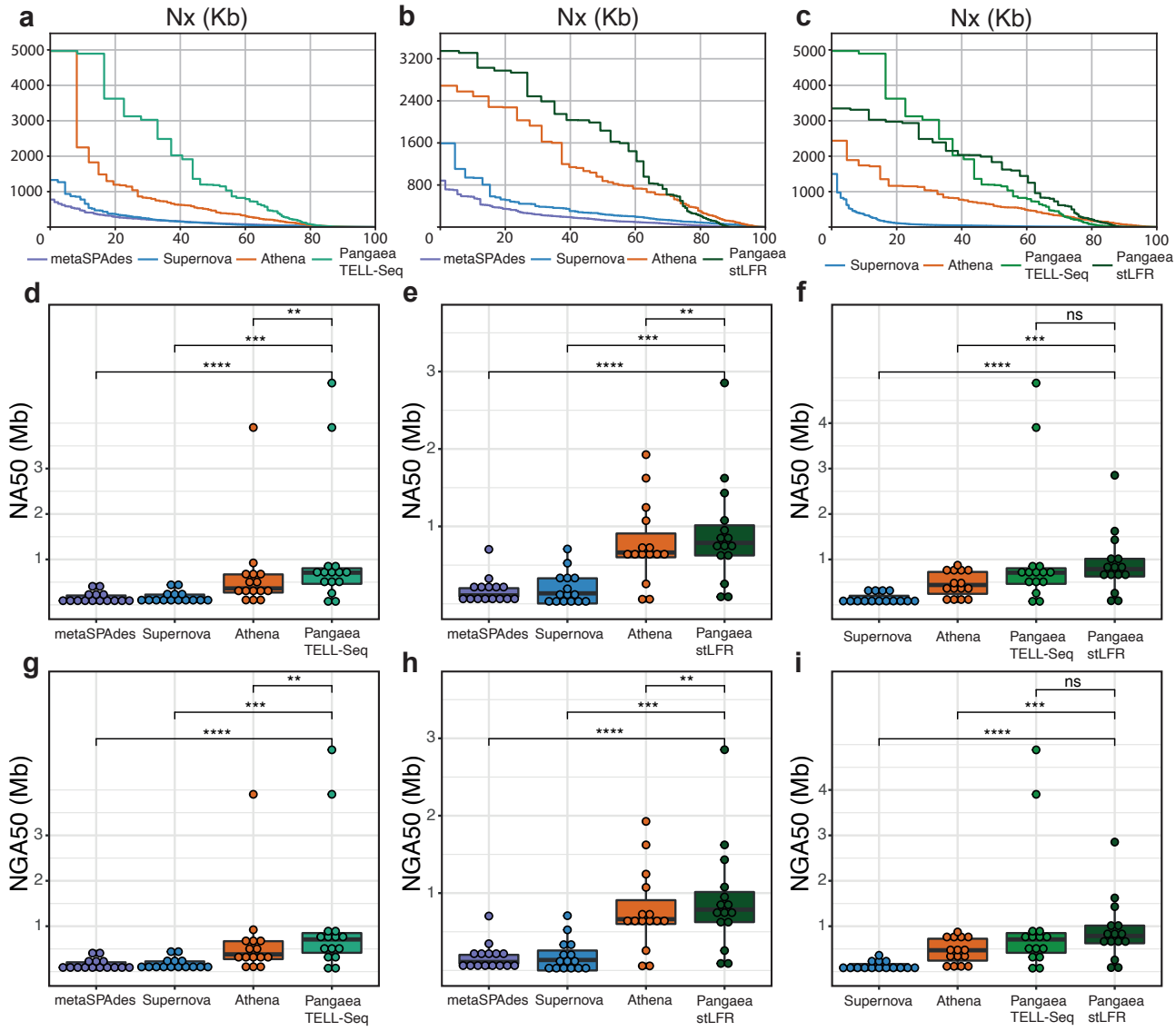

Figure 2. Nx, with x ranging from 0 to 100, on TELL-Seq (a), stLFR (b), and 10x linked-reads (c). NA50 and NGA50 for the 15 strains with abundance $\geq 0.18\%$ assembled by metaSPAdes, Supernova, Athena, and Pangaea using the TELL-Seq (d and g), stLFR (e and h) and 10x (f and i) linked-reads from the ATCC-MSA-1003 mock community. The p-values are reported using the wilcox.test function of R with "paired=TRUE". **, ***, and ns denote $p < 0.01$, $p < 0.001$, and no statistically significant difference, respectively.

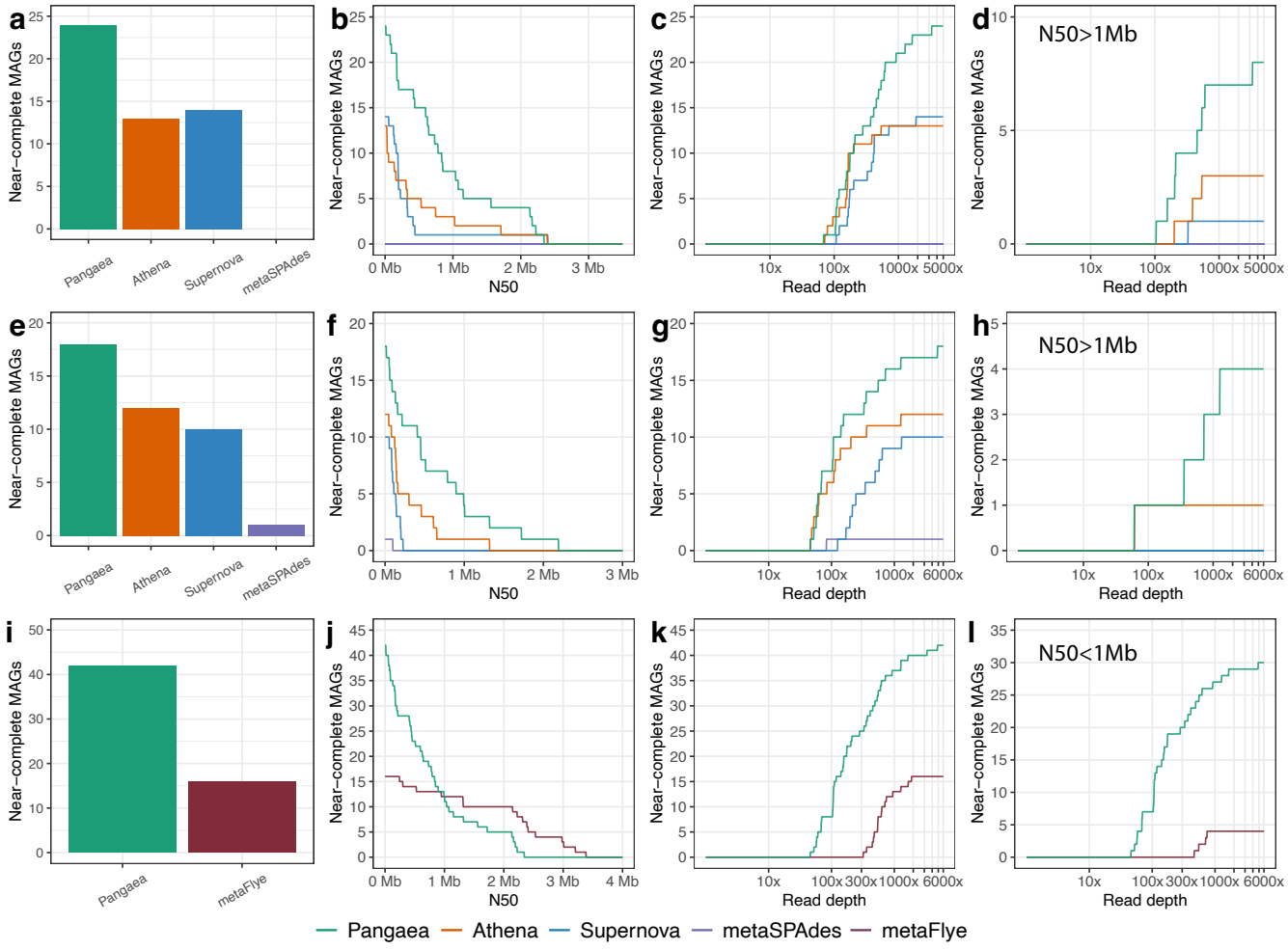

Figure 3. The number of NCMAGs generated using different assembly tools (a, e, and i). The number of NCMAGs by thresholding minimum N50 (b, f, and j) and maximum read depth (c-d, g-h, and k-l). We compared the performances of Pangaea, Athena, Supernova, and metaSPAdes on stLFR linked-reads from S1 (a-d) and S2 (e-h). The performance of Pangaea on stLFR linked-reads was also compared with that of metaFlye on PacBio CLR long-reads (i-l).

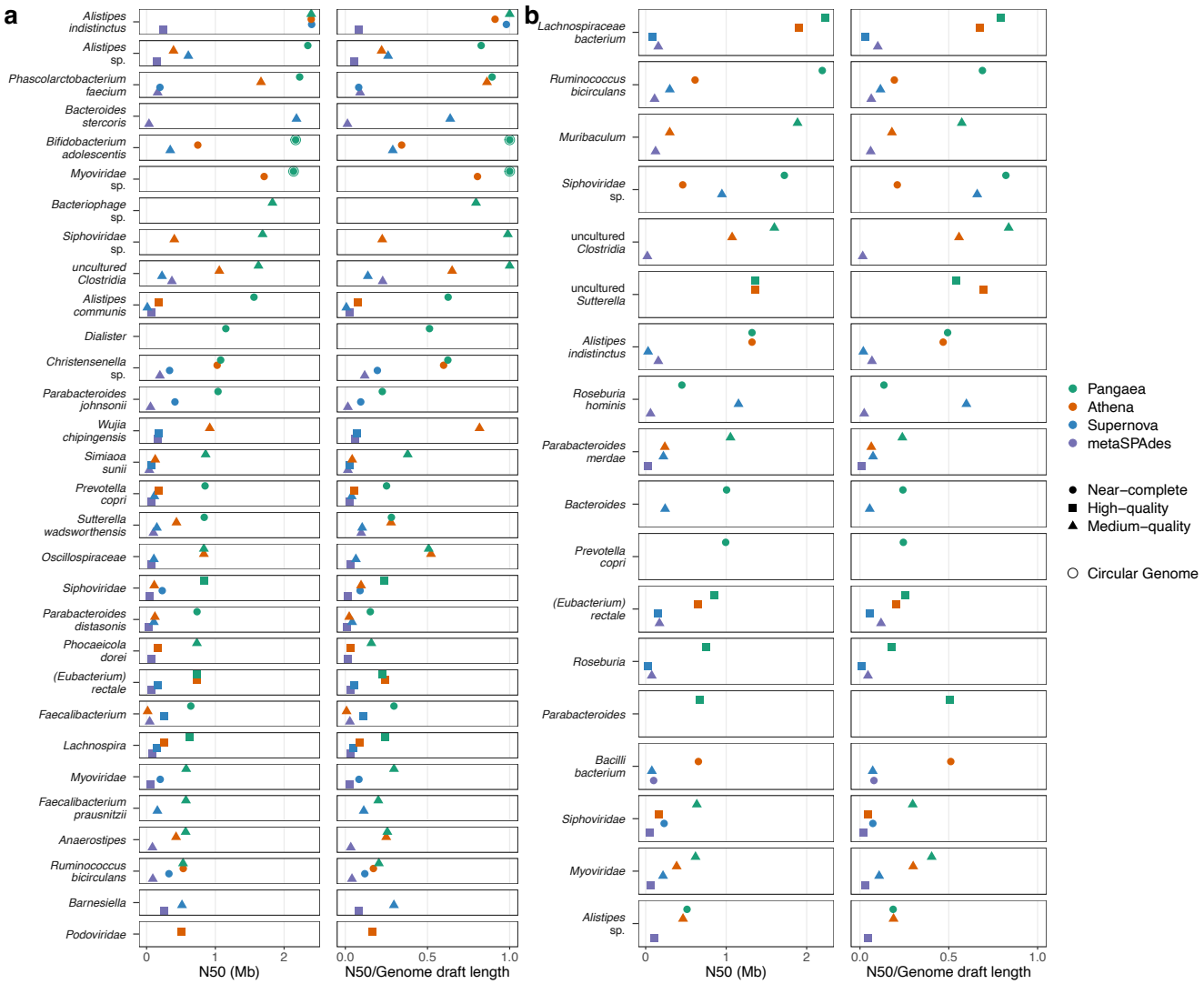

Figure 4. The annotated microbes of the MAGs produced by Pangaea, Athena, Supernova, and metaSPAdes from S1 (a) and S2 (b). The microbes were shown here if the N50s of their corresponding MAGs were larger than 500 Kb by any assembler. If the same microbe was annotated by more than one MAG from the same assembler, the one with the highest N50 was selected.

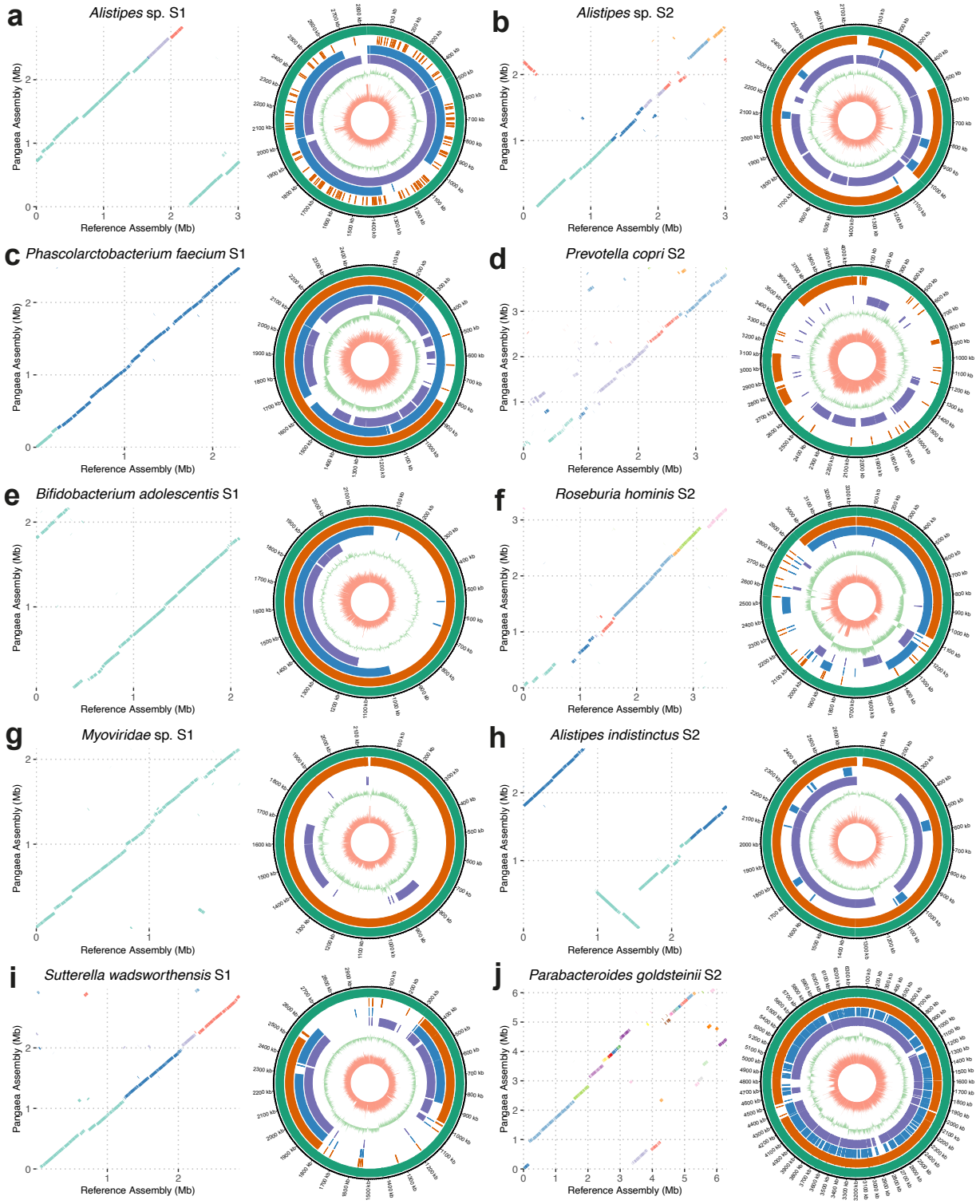

Figure 5. Genome collinearity analysis between the selected NCMAGs produced by Pangaea and their closest reference genomes (dot plots), and comparison of the corresponding MAGs produced by different assemblers (circos plots) from S1 and S2. e and g are species for which Pangaea obtained complete and circular MAGs. Colors are used in the dot plots to distinguish different contigs. The six rings in the circos plots from outside to inside denote the Pangaea MAGs (green), Athena MAGs (orange), Supernova MAGs (blue), metaSPAdes MAGs (purple), GC-skew of Pangaea MAGs, and read depth of Pangaea MAGs, respectively. If the same species was annotated by more than one MAG from the same assembler, the one with the highest N50 is shown here. The remaining NCMAGs produced by Pangaea are shown in **Supplementary Figure 2**.