Exploring high-quality microbial genomes by assembling short-reads with long-range connectivity

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

[1]Department of Computer Science, Hong Kong Baptist University, Hong Kong, China

[2]BGI Research, Shenzhen 518083, China

[3]BGI Research, Sanya 572025, China

[4]State Key Laboratory of Dampness Syndrome of Chinese Medicine, the Second Affiliated Hospital of Guangzhou University of Chinese, Guangzhou, China

[5]Department of Scientific Research, Kangmeihuada GeneTech Co., Ltd (KMHD), Shenzhen, China

[6]School of Chinese Medicine, Hong Kong Baptist University, Hong Kong, China

[7]Institute for Research and Continuing Education, Hong Kong Baptist University, Shenzhen, China

[†]These authors contributed equally to this work

[*]To whom correspondence should be addressed: ericluzhang@hkbu.edu.hk

Abstract

Despite long-read sequencing enables to generate complete genomes of unculturable microbes, its high cost hinders its widespread application in large cohorts. An alternative method is to assemble short-reads with long-range connectivity, which can be a cost-effective way to generate high-quality microbial genomes. We developed Pangaea to improve metagenome assembly using short-reads with physical or virtual barcodes. It adopts a deep-learning-based binning algorithm to assemble the co-barcoded reads with similar sequence contexts and abundances to improve assemblies of high- and medium-abundance microbes. Pangaea also leverages a multi-thresholding reassembly strategy to refine assembly for low-abundance microbes. We benchmarked Pangaea with linked-reads and a combination of short- and long-reads from mock communities and human gut metagenomes. Pangaea achieved significantly higher contig continuity as well as more near-complete metagenome-assembled genomes (NCMAGs) than the existing assemblers. Pangaea was also observed to generate three complete and circular NCMAGs on the human gut microbiomes.

Introduction

Metagenome assembly is one of the main steps to reconstruct microbial genomes from culture-free metagenomic sequencing data [1]. Cost-effective short-read sequencing technologies have been widely applied to generate high-quality microbial reference genomes from large cohorts of human gut microbiomes [2, 3, 4]. However, the short read length (100bps-300bps) may not allow us to resolve intra-species repetitive regions and inter-species conserved regions [5] or to achieve complete microbial genomes. The emerging long-read sequencing technologies such as PacBio continuous long-read sequencing (PacBio CLR) [6], Oxford Nanopore sequencing (ONT) [7] and PacBio HiFi sequencing [8], have shown their superiority to short-read sequencing in generating metagenome-assembled genomes (MAGs) with high continuity or producing complete and circular MAGs using the long-range connectivity they provided [9, 10, 11]. Despite potential benefits, the high cost of long-read sequencing makes deep sequencing impracticable and continues to hinder its application in population-scale or clinical studies [12]. In our previous research [13], long-reads generated fewer high-quality MAGs than short-reads due to insufficient sequencing depth. As an alternative way to deep long-read sequencing, some studies [14, 15, 16, 13] suggested utilizing cost-effective short-reads with long-range connectivity for metagenome assembly. The long-range connectivity could be derived from physical barcodes (e.g., linked-reads) or virtual links by other long-fragment sequencing technologies (e.g. long-reads).

Linked-read sequencing attaches identical barcodes (physical barcodes) to the short-reads if they are derived from the same long DNA fragment. Before the discontinuation, 10x Chromium was the

most widely used linked-read sequencing technology, generating contigs with high continuity and producing more near-complete metagenome-assembled genomes (NCMAGs; **Methods**) than short-read sequencing [13, 14]. However, co-barcoded short-reads of 10x have a high chance of being derived from multiple DNA fragments (average number of fragments per barcode [$N_{F/B}$] is 16.61; **Methods** and **Supplementary Note 1**), which may complicate the deconvolution of complex microbial communities. Recently, MGI and Universal Sequencing Technology released their linked-read sequencing technologies, namely single-tube Long Fragment Read (stLFR) [17] and Transposase Enzyme-Linked Long-read Sequencing (TELL-Seq) [18]. The barcoding reactions of these technologies occur on billions of microbeads, leading to much higher barcode specificity ($N_{F/B} = 1.54$ for stLFR; $N_{F/B} = 4.26$ for TELL-Seq; **Supplementary Note 1**). The co-barcoded short-reads of these technologies are more likely to come from the same genomic regions.

Several assemblers have been developed for assembling linked-reads: (i) Athena [14], which fills the gaps between contigs by recruiting the 10x co-barcoded reads for local assembly; (ii) cloudSPAdes [19], which reconstructs the long DNA fragments in the assembly graph for solving the shortest superstring problem to improve contig continuity; (iii) Supernova [20], which was developed for diploid human genome assembly by allowing two paths in megabubble structures based on a series of modification of assembly graph using 10x co-barcoded reads; (iv) MetaTrass [21], which groups the stLFR co-barcoded reads by reference-based taxonomic annotation and applies Supernova to assemble the genome of each identified species. With the exception of MetaTrass, all the other three tools were developed for 10x linked-reads with low barcode specificity. We exclude MetaTrass for comparison in this paper because it relies on the available microbial reference genomes and thus has a limited capability to discover novel species. There is still a lack of efficient tools that could fully exploit the long-range connectivity of short-reads from high-specificity barcodes to improve *de novo* metagenome assembly.

Long-range connectivity could also be provided by other long-fragment sequencing technologies, which can be used in conjunction with short-reads for hybrid assembly. The hybrid assembly is typically performed by combining high-depth short-reads with shallow-depth long-reads. This approach takes advantage of the high base quality of short-reads for contig assembly and the long-range connectivity of long-reads for contig extension. Two hybrid assemblers were developed for metagenome assembly: (i) hybridSPAdes [16] maps long-reads to the assembly graph from short-reads and utilizes the long-range connectivity to resolve uneven path depth and repetitive sequences in the graph; (ii) OPERA-MS [15] aligns long-reads to the contigs assembled from short-reads to construct a scaffold graph and groups contigs based on microbial reference genomes followed by gap filling in each cluster.

Previous studies showed that read subsampling was an effective strategy for assembling large complex metagenomic datasets [22]. It could improve the assembly of high-abundance microbes [23],

but result in poor quality in assembling low-abundance microbes [24] due to insufficient reads. In a metagenome community, microbes may have highly uneven abundances [25]. Sequences obtained from low-abundance microbes can result in spurious edges with shallow depths in the assembly graph. Distinguishing these connections from edges originating from sequencing errors becomes a particularly challenging task [26]. We introduce Pangaea, a metagenome assembler, to improve metagenome assembly with three modules. Firstly, Pangaea applies co-barcoded short-read clustering and assembly to reduce the complexity of metagenomic sequencing data. The read binning strategy has been proven advantageous in metagenome assembly [27, 28, 29]. It could be a more sophisticated read downsampling strategy to improve the assemblies of high- and medium-abundance microbes. However, the existing tools are impractical when it comes to handling millions of short-reads within acceptable time and memory limitations. Pangaea groups co-barcoded short-reads rather than grouping independent short-reads, as the co-barcoded reads have highly specific barcodes (**Methods**). Short-reads that belong to the same cluster have lower complexity than the original dataset and are assembled separately. Secondly, Pangaea adopts a multi-thresholding reassembly step to refine the assembly of low-abundance microbes using different abundance thresholds to handle the uneven abundances of microbes (**Methods**). The data from high-abundance microbes are gradually removed from the assembly graph, thus the sequences from various levels of low-abundance microbes are preserved. Athena [14] demonstrated that enhancing contig continuity can be achievable by combining the original short-read assembly with contig local assembly based on co-barcoded reads. Thirdly, Pangaea integrates the assemblies from the above two modules, original short-read assembly and local assembly to improve contig continuity (Methods). We benchmarked Pangaea using linked-reads with physical barcodes from linked-read sequencing of the mock and human gut metagenomes, and evaluated its generalizability using short-reads with virtual barcodes generated from their alignments on long-reads (**Methods**). For linked-reads, we compared Pangaea with two short-read assemblers (metaSPAdes [30] and MEGAHIT [31]) and three linked-read assemblers (cloudSPAdes, Supernova and Athena). We found Pangaea achieved substantially better contig continuity and more NCMAGs than the other tools on all datasets. It also generated three complete and circular microbial genomes for the three real complex microbial communities. For short-reads with virtual barcodes, Pangaea could substantially improve contig continuity and generate better assemblies for both high- and low-abundance microbes than the short-read and hybrid assemblers.

Results

Workflow of Pangaea

Pangaea is a *de novo* metagenome assembler designed for short-reads with long-range connectivities represented by their attached barcodes (**Figure 1 a**). The barcodes can be generated physically (from linked-read sequencing with high barcode specificity, e.g., stLFR and TELL-Seq) or virtually (from long-reads; **Methods**). The co-barcoded short-reads are considered to be derived from the same long DNA fragments. The core algorithms of Pangaea are designed to reduce the complexity of metagenomic sequencing data (for high- and medium-abundance microbes) and deal with the uneven abundances of involved microbes (for low-abundance microbes) based on the barcodes with high specificity. Pangaea contains three main modules: (i) Co-barcoded read binning. This module is intended to reduce metagenomic complexity before assembly and is mainly used to improve the assembly of high- and medium-abundance microbes from real-world datasets with high complexity. Pangaea collects the co-barcoded short-reads and represents them using k-mer histograms and TNFs (**Methods**) to overcome feature instability in grouping short-reads individually. These features enable Variational Autoencoder (VAE) to represent co-barcoded short-reads in low-dimensional latent space, where the latent variable follows a standard Gaussian distribution (**Methods**; **Supplementary Note 2**). Pangaea adopts a weighted sampling strategy in VAE training to balance the number of co-barcoded short-reads from microbes with different abundances (**Methods**). The k-mer histogram (X_A) for each specific barcode can be modeled as the mixture of two Poisson distributions, one represents the erroneous k-mers with extremely low frequency [32], and the other represents the true k-mer abundances [33]. We observed there was a negative correlation between $max(X_A)$ and the abundance of the co-barcoded reads (**Supplementary Figure 1**). Co-barcoded reads from low-abundance microbes should have a higher chance of being sampled, thus we used $max(X_A)^2$ as the sampling weight of each barcode (**Methods**). Pangaea groups co-barcoded short-reads in the latent space using RPH-kmeans [34], which is beneficial if the bins have uneven sizes (**Methods**) and assembles the short-reads in each bin independently. (ii) Multi-thresholding reassembly. This module refines the assembly of low-abundance microbes by removing the sequences of high-abundance microbes from the assembly graph gradually (**Figure 1 b**; **Methods**). This module aligns short-reads to contigs and removes them if they are from the contigs with abundances above certain thresholds (**Methods**). The remaining reads are used to reassemble the contigs of low-abundance microbes. (iii) Ensemble assembly. This module is to eliminate the impact of mis-binning on final assembly results. Athena [14] shows the integration of original short-read assembly and local assembly can improve metagenome assembly using the overlap-layout-consensus (OLC) strategy. Following the same strategy, Pangaea additionally incorporates the

assemblies from the previous two modules (**Methods**). Pangaea would finally integrate the ensemble assembly with the assembly from the available assembly tools for the short-reads with virtual barcode using quickmerge (**Methods**). This is optional for the assembly of short-reads with physical barcodes.

Co-barcoded read binning increases NCMAGs from medium- and high-abundance microbes

Co-barcoded read binning in the first module is a core step of Pangaea. We demonstrate this module significantly improved the assembly for high- and medium-abundance microbes, particularly for the real complex microbial communities. We sequenced the stLFR (132.95Gb) and TELL-Seq (173.2Gb) linked-reads (**Methods**) of the ATCC-MSA-1003 mock community containing 20 strains (whose reference genomes are available; **Supplementary Table 1**) with loadings from 0.02% to 18% to evaluate co-barcoded read binning on the low-complexity mock dataset. We also prepared three real human gut microbiome datasets (S1, S2, and S3) with stLFR sequencing (**Supplementary Table 2**; **Methods**) to validate the algorithm on real complex microbial communities.

As there were no available binning tools for linked-reads, we compared the VAE-based binning algorithm in Pangaea with two existing long-read binning tools, METABCC-LR (v2.0.0) [33] and LRBinner (v2.1) [35] on ATCC-MSA-1003. In order to enable these long-read binning tools on linked-reads, the co-barcoded reads were connected using a single "N" into long-reads. LRBinner failed to finish the binning task within three weeks with 100 threads, so we excluded it for further comparison. The VAE-based binning algorithm achieved a higher overall F1 score and adjusted rand index (ARI) than METABCC-LR (Panagea: F1 = 0.6144, ARI = 0.3764; METABCC-LR: F1 = 0.5887; ARI = 0.1704; **Supplementary Table 3**). The VAE-based binning particularly improved the binning performance for low-abundance microbes (abundance: 0.02%), where it achieved an F1 score of 0.6073 and ARI of 0.4624. The performance of METABCC-LR was almost the same as random binning (**Supplementary Table 3**).

We compared the metagenome assemblies of Panagea with (ASM_B) and without ($ASM_{\neg B}$) read binning to investigate the impact of co-barcoded read binning on final assembly results. For TELL-Seq of ATCC-MSA-1003, ASM_B had a higher overall NA50 (649.67Kb) than $ASM_{\neg B}$ (601.67Kb; **Supplementary Table 4**). Considering the individual strains, ASM_B achieved higher NGA50s for 7 out of the 10 strains with abundances higher than 1% (**Supplementary Table 4**). For the human gut microbiomes, ASM_B obtained comparable N50s and more NCMAGs than $ASM_{\neg B}$ on the three samples (S1: $ASM_B = 24$, $ASM_{\neg B} = 20$; S2: $ASM_B = 17$, $ASM_{\neg B} = 11$; S3: $ASM_B = 9$, $ASM_{\neg B} = 8$). These NCMAGs generated by ASM_B were commonly observed from high-abundance microbes (average depths of NCMAGs: S1 = 526.6X; S2 = 211.19X; S3 = 256.52X). Next, we evaluated the

impact of bin number on the binning performance and final assembly of ATCC-MSA-1003. We observed a large bin number resulted in high binning precision and low recall (**Supplementary Figure 2**). The bin number is a trade-off between generating bins with low complexities (large bin number) or keeping more reads from the same microbes in the same bin (small bin number). Although the bin numbers would influence the read binning performance, the assembly results were robust if bin numbers varied in a certain range (e.g. 10-20 in ATCC-MSA-1003 with 20 strains; **Supplementary Figure 3**). In general, datasets with high complexity should be assigned a large bin number. We used 15 and 30 as bin numbers for the datasets from ATCC-MSA-1003 and the human gut microbiomes, respectively.

Multi-thresholding reassembly improves assemblies of low-abundance microbes

Pangaea improves assemblies of low-abundance microbes by gradually removing the reads of high-abundance microbes from the assembly graph using multiple abundance thresholds (represented by T). We compared assemblies using different settings of T and found T = {10, 30} was the best choice (**Supplementary Note 3**), which was used in the subsequent experiments. We compared the assemblies with and without multi-thresholding reassembly and found this module could increase the NGA50s of 5 strains with abundances less than 1% (out of 6 strains, covered by more 50% genomes; **Supplementary Table 5**). It could generate more sequences from the long contigs (>10Kb) of 6 strains with abundances less than 1% (out of 7 strains with contigs longer than 10Kb; **Supplementary Table 5**). For the three human gut microbiomes, the assembly with multi-thresholding reassembly could also generate two more NCMAGs.

Barcode specificity is critical for linked-read assembly

We sequenced ATCC-MSA-1003 using stLFR, TELL-Seq linked-read sequencing and collected 10x linked-reads [36] (**Supplementary Table 2**; **Methods**) to evaluate the influence of barcode specificity in metagenome assembly. Linked-reads of stLFR yielded the lowest $N_{F/B}$ ($N_{F/B} = 1.54$), and TELL-Seq linked-reads yielded a slightly higher value ($N_{F/B} = 4.26$). Both values were much lower than that obtained from 10x linked-reads ($N_{F/B} = 16.61$; **Supplementary Note 1**).

The contigs of Pangaea from stLFR and TELL-Seq linked-reads had substantially higher N50 (22.42 times on average; Figure2 a) and higher overall NA50 (11.55 times on average; **Supplementary Table 6**) than the contigs of Athena and Supernova from 10x linked-reads. For the 15 strains with abundance higher than 0.1% (**Supplementary Table 7**), Pangaea on stLFR and TELL-Seq linked-reads also achieved significantly higher per-strain NA50 (**Figure 2 d**) and NGA50 (**Figure 2 g**) than the assemblies of Athena and Supernova using 10x linked-reads. For the remaining 5 strains

with abundances of 0.02%, Pangaea on stLFR (average genome fraction: 33.49%) and TELL-Seq (average genome fraction: 23.55%) 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 8**). These results suggest that short-reads with high barcode specificity could produce better metagenome assemblies using Pangaea.

Pangaea generated high-quality assemblies for both high- and low-abundance microbes on mock linked-read datasets

We used two mock linked-read datasets (ATCC-MSA-1003 and ZYMO) to compare Pangaea to Athena, Supernova, cloudSPAdes, MEGAHIT, and metaSPAdes. For ATCC-MSA-1003, we benchmarked Pangaea on both TELL-Seq and stLFR linked-reads. For TELL-Seq (**Table 1**; **Figure 2 b**), Pangaea achieved the highest N50 (1,360.32Kb) and overall NA50 (649.47Kb) when compared with the statistics achieved by Athena (N50: 466.50Kb; NA50: 361.57Kb), Supernova (N50: 102.76Kb; NA50: 97.31Kb), cloudSPAdes (N50: 127.42Kb; NA50: 118.16Kb), MEGAHIT (N50: 128.07Kb; NA50: 112.51Kb) and metaSPAdes (N50: 112.34Kb; NA50: 105.63Kb). The contig N50 and overall NA50 for Pangaea on stLFR linked-reads were akin to the results on TELL-Seq linked-reads (**Table 1**; **Figure 2 c**).

We simulated 143.04Gb stLFR linked-reads for the ZymoBIOMICS^{TM} Microbial Community Standard II (Log Distribution) [37] (ZYMO; **Supplementary Table 2**; **Methods**), which consists of 10 strains with highly imbalanced abundances (**Supplementary Table 9**). Pangaea also generated substantially higher N50, overall NA50, NA50 per strain, and NGA50 per strain than the other five assemblers on ZYMO (**Table 1**). These observations suggest Pangaea could significantly improve contig continuity using linked-reads with high barcode specificity.

When considering those 15 strains with high abundance (higher than 0.1%) on TELL-Seq linked-reads of ATCC-MSA-1003 (**Supplementary Table 7**), Pangaea still generated significantly higher strain NA50s (**Figure 2 e**) and NGA50s (**Figure 2 h**) than Athena (NA50: p-value = 1.22e-4; NGA50: p-value = 6.10e-5), Supernova (NA50: p-value = 3.05e-4; NGA50: p-value = 3.05e-4), cloudSPAdes (NA50: p-value = 6.10e-5; NGA50: p-value = 6.10e-5), MEGAHIT (NA50: p-value = 6.10e-5; NGA50: p-value = 6.10e-5) and metaSPAdes (NA50: p-value = 6.10e-5; NGA50: p-value = 6.10e-5). A comparable trend was observed on the assemblies of stLFR linked-reads of ATCC-MSA-1003 (**Table 1**; **Figure 2 f and i**). For the 5 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 8**), indicating more genomic sequences could be assembled by Pangaea for low-abundance microbes.

For ZYMO, Pangaea and Athena always generated better assemblies than the other tools on strains

with different abundances (**Supplementary Table 10**). For the two strains (*Escherichia coli* and *Salmonella enterica*) with abundances between 0.01% and 0.1%, Pangaea achieved consistently higher NA50 (1.68 times on average) and NGA50 (1.68 times on average) than Athena (**Supplementary Table 10**). Although all assemblers could not generate contigs covering half of the genome of *Lactobacillus fermentum* (abundance between 0.001% and 0.01%), Pangaea still obtained the highest genome fraction (43.64%), which was substantially higher than the values of the other assemblers considering short-read barcodes (**Supplementary Table 10**). The strains with abundance below 0.001% were unable to be assembled by all assemblers (**Supplementary Table 10**).

Pangaea generated high-quality assemblies on the human gut microbiomes

We collected DNAs from three healthy Chinese individual stool samples (S1, S2, and S3) and sequenced 136.60Gb (S1), 131.59Gb (S2) and 50.74Gb (S3) stLFR linked-reads for their gut microbiomes (**Supplementary Table 2**; **Supplementary Figure 4**; **Methods**). The assemblies generated by Pangaea achieved the highest total assembly length for all three samples (**Table 1**). Moreover, Pangaea achieved substantially higher N50s than all the other assemblers for both S1 (1.44 times of Athena, 1.06 times of Supernova, 4.71 times of MEGAHIT, 4.50 times of metaSPAdes; **Table 1**) and S2 (1.57 times of Athena, 2.58 times of Supernova, 8.00 times of MEGAHIT, 7.99 times of metaSPAdes; **Table 1**). For S3, Pangaea generated much more sequences than the other assemblers (contig length>1Kb; **Methods**), making it unfair to compare their N50 values directly. We transformed their assembly length to be comparable by only considering contigs longer than 5Kb. This led to a significant improvement in the contig N50 of Pangaea, which became the best one (Pangaea: 275.65Kb, Athena: 157.81Kb, Supernova: 105.54Kb, cloudSPAdes: 44.59Kb, MEGAHIT: 33.26Kb, metaSPAdes: 36.34Kb; **Table 1**).

We grouped the contigs into MAGs and identified NCMAGs (**Methods**) to evaluate the performance of metagenome assembly. Pangaea generated NCMAGs (**Figure 3 a, e and i**) of 24, 17 and 9 for S1, S2 and S3, which were much more than those generated by Athena (S1: 13, S2: 12, S3: 7), Supernova (S1: 14, S2: 10, S3: 6), cloudSPAdes (only available on S3: 0 due to its low computational efficiency), MEGAHIT (S1: 0, S2: 1, S3: 0) and metaSPAdes (S1: 0, S2: 1, S3: 0). By calculating the number of NCMAGs at different minimum values of N50, we found Pangaea obtained more NCMAGs than the other assemblers at almost all N50 thresholds (**Figure 3 b, f and j**). Pangaea also outperformed the other assemblers with respect to the number of NCMAGs at different maximum read depth thresholds (**Figure 3 c, g and k**). Especially for the NCMAGs with N50s larger than 1Mb, Pangaea achieved substantially more NCMAGs (S1: 8, S2: 4, S3: 5; **Figure 3 d, h and l**) than the other assemblers at all read depth thresholds, while the second best assembler (Athena) only produced

3, 1 and 2 NCMAGs on S1, S2 and S3, respectively (**Figure 3 d, h and l**).

Pangaea generated high-quality genomes of annotated microbes on linked-reads

We annotated MAGs using Kraken2 [38] with the Nucleotide database of the National Center for Biotechnology Information (**Methods**). Total 61 microbes (S1: 26, S2: 19, S3: 16; **Figure 4**) were annotated from Pangaea's MAGs; 56 of them (S1: 24, S2: 16, S3: 16; **Figure 4**) achieved the highest N50 and 33 microbes had two-fold higher N50s comparing to the second best assemblers (S1: 16, S2: 8, S3: 9; **Figure 4**). Among the remaining 5 microbes that Pangaea did not achieve the highest N50 (**Figure 4**), Pangaea generated comparable N50s with the best assembler on *Alistipes indistinctus* and *Ruminococcus bicirculans* from S1, and *Bacilli bacterium* from S2; and achieved a lower N50 but substantially higher completeness than Supernova on *Roseburia hominis* (Pangaea: completeness = 96.54%, contamination = 0.48%; Supernova: completeness = 64.91%, contamination = 0.00%; **Supplementary Table 11**) and *uncultured Clostridia* from S2 (Pangaea: completeness = 98.25%, contamination = 7.02%; Athena: completeness = 76.78%, contamination = 1.34%; **Supplementary Table 11**).

Moreover, Pangaea generated more NCMAGs for the annotated microbes. There were 13 microbes (S1: 7, S2: 2, S3: 4; **Figure 4**) that could be assembled as NCMAGs from Pangaea, whereas all the other assemblers only generated MAGs with lower quality or could not generate the matching MAGs. For the three human gut microbiomes , Pangaea identified 17 annotated microbes from the NCMAGs with N50s greater than 1Mb, where Athena, Supernova, cloudSPAdes, MEGAHIT and metaSPAdes only identified 6, 1, 0, 0, and 0 microbes, respectively (**Figure 4**). In addition, Pangaea recognized 6 unique microbes (*Bacteriophage* sp. and *Dialister* from S1, and *uncultured bacterium*, *Muribaculum gordoncarteri*, *Parabacteroides* and *Prevotella copri* from S2) that were not found by any other assemblers, and *Dialister* from S1 was represented by NCMAG from Pangaea (**Figure 4**).

Strong collinearities between NCMAGs and their closest reference genomes

We compared the NCMAGs that can be annotated as species with their closest reference genomes to evaluate their collinearities (**Methods**). The NCMAGs generated by different assemblers and their closest reference genomes had comparable average alignment identities (Pangaea: 98.16%, Athena: 98.12%, Supernova: 98.31%, MEGAHIT: 98.7%, metaSPAdes: 98.8%) and average alignment fractions (Pangaea: 87.9%, Athena: 88.6%, Supernova: 88.4%, MEGAHIT: 88%, metaSPAdes: 90%; **Supplementary Table 12**), while Pangaea produced significantly more species-level NCMAGs than the other assemblers (Pangaea: 29, Athena: 21, Supernova: 14, cloudSPAdes: 0, MEGAHIT: 1,

metaSPAdes: 1; **Supplementary Table 12**).

The NCMAGs assembled by Pangaea had high collinearities with their closest reference genomes (**Figure 5**; **Supplementary Figure 5**). Some NCMAGs showed inversions and rearrangements in comparison to the reference sequences, including *Alistipes communis* (S1; **Supplementary Figure 5 a**), *Desulfovibrio desulfuricans* (S1; **Supplementary Figure 5 c**) and *Siphoviridae* sp. (S2; **Figure 5 g**). Pangaea assembled NCMAGs for *Siphoviridae* sp. from both S2 and S3 (**Figure 5 g and h**). The two NCMAGs had comparable total sequence lengths (S2: 2.20Mb and S3: 2.39Mb; **Supplementary Table 11**) and better N50 was achieved in S3 (N50: 1.16Mb for S2 and 2.39Mb for S3; **Supplementary Table 11**). Note that the NCMAG of *Siphoviridae* sp. from S3 is a single-contig NCMAG, but it's not a circular contig which might be because it does not include all single-copy genes (completeness: 91.28%; **Supplementary Table 11**).

Pangaea could generate NCMAGs for annotated species with better quality and larger N50 values than the other assemblers, such as *Alistipes* sp. from S1 and *Collinsella aerofaciens* from S2 (**Figure 5 d and e**; **Supplementary Table 11**). The 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 480Kb of *Faecalibacterium prausnitzii* from S3 (**Supplementary Figure 5 q**). This indicates Pangaea has the potential to reconstruct hard-to-assemble genomic regions.

Pangaea generated complete and circular MAGs using linked-reads

We examined if there existed complete and circular genomes in NCMAGs based on the circularization module in Lathe [39] (**Methods**). We found that only Pangaea generated three circular NCMAGs, which were annotated as *Bifidobacterium adolescentis* (S1), *Myoviridae* sp. (S1) and *Wujia chipingensis* (S3). For each of the three microbes, Pangaea produced a gapless contig with perfect collinearity with the closest reference genomes (**Figure 5 a, b and i**).

Athena generated three and two contigs for *B. adolescentis* (from S1) and *Myoviridae* sp. (from S1), with substantially lower contig N50 values than the contigs obtained by Pangaea for these two species (*B. adolescentis*: Pangaea = 2,167.94Kb, Athena = 744.54Kb; *Myoviridae* sp.: Pangaea = 2,137.66Kb, Athena = 1,709.63Kb; **Supplementary Table 11**). Supernova, MEGAHIT 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 11**). For *W. chipingensis* from S3, Pangaea was the only assembler that obtained NCMAG and got a significantly higher contig N50 than the other assemblers (Pangaea: 2,846.41Kb, Athena: 639.65Kb, Supernova: 254.31Kb, cloudSPAdes: 135.56Kb, MEGAHIT: 134.14Kb, metaSPAdes: 146.44Kb; **Supplementary Table 11**).

Pangaea generated high-quality assemblies on short-reads with virtual barcodes from long-reads

We evaluated the generalizability of Pangaea on hybrid assembly in terms of attaching virtual barcodes to short-reads. We attached identical barcodes to short-reads if they were aligned to the same long-reads, which is another way to offer long-range connectivity of short-reads (**Methods**). We used the short-reads alignment with 1.27Gb PacBio CLR [13] (ATCC-MSA-1003) and 2.48Gb Nanopore GridION (ZYMO) [37] long-reads (**Supplementary Table 2**) to substitute the barcodes of TELL-Seq (ATCC-MSA-1003) and stLFR (ZYMO) linked-reads. We adjusted the workflow of Pangaea to integrate with hybridSPAdes (namely Pangaea-Hybridspades) and OPERA-MS (namely Pangaea-Operams; **Methods**). We compared Pangaea-Hybridspades and Pangaea-Operams to hybridSPAdes, OPERA-MS, Athena and metaSPAdes, and observed that virtual barcodes could prominently increase contig N50 and NA50 (**Table 2**; **Figure 6 a**). For example, the N50 and NA50 of Pangaea-Hybridspades were 1.7 and 1.31 times higher than hybridSPAdes while the corresponding values of OPERA-MS were increased by 4.43 and 3.58 folds (**Table 2**).

For ATCC-MSA-1003, Pangaea-Hybridspades achieved the highest N50s (average N50: 903.64Kb) and NA50s (average NA50: 701.48Kb; **Supplementary Table 13**) for the 10 high- and medium-abundance strains with abundance higher than 1%. It also generated significantly higher NA50s and NGA50s than the other assemblers for the 15 strains with abundances higher than 0.1% (**Figure 6 b and c**). For the 5 strains with the lowest abundance (0.02%), all assemblers got poor NA50s and NGA50s; Pangaea-Operams achieved much more sequences from the contigs longer than 10Kb compared to OPERA-MS (Pangaea-Operams: 141.76Kb, OPERA-MS: 64.19Kb; **Supplementary Table 13**), suggesting Pangaea could improve contig continuity for low-abundance microbes.

For ZYMO, Pangaea-Hybridspades also generated significantly higher average NA50 (Pangaea-Hybridspades: 1.98Mb, hybridSPAdes: 887.00Kb; **Supplementary Table 13**) and NGA50 (Pangaea-Hybridspades: 1.98Mb, hybridSPAdes: 887.00Kb; **Supplementary Table 13**) for the two strains with abundances higher than 1% (*Listeria monocytogenes* and *Pseudomonas aeruginosa*) than hybridSPAdes. For the two strains (*Escherichia coli* and *Salmonella enterica*) with abundance between 0.01% and 0.1%, Pangaea-Hybridspades and hybridSPAdes had comparable contig NA50 and NGA50; while Pangaea-Operams produced better NA50 (1.66 times on average) and NGA50 (1.71 times on average) than OPERA-MS (**Supplementary Table 13**). All the assemblers generated low-quality assemblies for the strain with an abundance between 0.001% and 0.01% (*Lactobacillus fermentum*). Pangaea-Hybridspades (19.77Kb) and Pangaea-Operams (54.91Kb) achieved much more sequences from the contigs longer than 10Kb than hybridSPAdes (13.69Kb) and OPERA-MS (37.41Kb), respectively (**Supplementary Table 13**). The strains with abundances below 0.001% cannot be assembled

for all assemblers (**Supplementary Table 13**). These observations demonstrate that Pangaea could improve the assemblies for both high-abundance and low-abundance microbes by leveraging virtual barcodes from long-reads.

Evaluation of running time and maximum memory usage

We compared the computational performance (CPU time, Real time, and Maximum Resident Set Size [RSS]) of the benchmarked assemblers on ZYMO (**Figure 7**; **Methods**). Both Athena and Pangaea required the assemblies from the other tools, so we only considered their additional processing time and memory. MEGAHIT was the fastest assembler with the lowest maximum RSS (**Figure 7**). metaSPAdes and cloudSPAdes consumed substantially higher CPU times than the other assemblers (**Figure 7 a**). The real time (wall clock time) used by metaSPAdes, cloudSPAdes, Supernova, Athena and hybridSPAdes were comparable and significantly higher than those consumed by MEGAHIT, Pangaea, OPERA-MS, Pangaea-Operams and Pangaea-Hybridspades (**Figure 7 b**). Supernova used the highest maximum RSS on ZYMO (**Figure 7 c**). The maximum RSSs used by metaSPAdes, cloudSPAdes, and hybridSPAdes were close to each other, and much higher than those needed by the other assemblers (**Figure 7 c**). These results revealed Pangaea could improve the assembly quality of the existing assemblers in a reasonable time and using a relatively low maximum RSS.

Discussion

Short-read sequencing has proven to be an important approach for analyzing human gut microbiota from large sequencing cohorts. However, its lack of long-range DNA connectivity makes assembling conserved sequences, intra- and inter-species repeats, and ribosomal RNAs (rRNAs) difficult [5]. It has limitations in producing complete microbial genomes and long-read sequencing is relatively expensive to be applied to large cohorts. Cost-effective linked-read sequencing technologies, which attach barcodes to short-reads to provide long-range DNA connectivity, have achieved great success in improving contig continuity in metagenome assembly [14, 19]. Unlike 10x linked-reads, stLFR [17] and TELL-Seq linked-reads [18] have high barcode specificity, but a dedicated assembler that could make full use of this characteristic to improve metagenome assembly is lacking. Besides linked-reads, the long-range connectivity of short-reads could also be provided by virtual barcodes from other independent sequencing technologies (e.g., long-reads).

In this study, we developed Pangaea to improve metagenome assembly by leveraging long-range connectivity from linked-reads and long-reads. It considers the co-barcoded reads as long DNA fragments and extracts their k-mer histograms and TNFs for co-barcoded read binning. This strategy

significantly reduces the complexity of metagenomic sequencing data and makes the assembly more efficient. Because sequence clustering is sensitive to data sparsity and noise, Pangaea represents the input features in a low-dimensional latent space using VAE. We also designed a weighted sampling strategy to generate a balanced training set for microbes with different abundances. This module primarily advantages microbes with high- and medium-abundance, because they are more robust to mis-binning and usually have sufficient data for assembly in the corresponding bins. Pangaea adopts a multi-thresholding reassembly strategy to rescue the reads from low-abundance microbes. It eliminates short-reads from high-abundance microbes in the assembly graph gradually to differentiate the spurious edges from low-abundance microbes and sequencing errors. In the third module, we merged the assemblies from different strategies due to their complementary nature of each other.

The previous study [40] showed that co-assembly with multiple samples could improve completeness and decrease the contamination of MAGs. However, the co-assembly of metagenomes derived from large samples poses a practical challenge due to time and computational limitations. Because the sequencing data from all the individual samples need to be merged before co-assembly. Read binning is a strategy that enables co-assembly on large datasets by producing smaller and simpler subsets of reads, thereby facilitating the assembly process. Several studies have attempted to apply read binning to short-read metagenomic sequencing [29, 28, 27], but it is exceedingly difficult in practice. The fragments of short-reads are too short to allow the extraction of stable sequence abundance and composition features from the individual reads. Existing read binning tools have to identify the overlap between each pair of reads for binning. However, the millions or even billions of short-reads make the overlap-based read binning algorithm extremely slow and highly memory intensive. Overlap Graph-based Read clustEring (OGRE) [27] was developed to improve the computational performance of read binning, but it still consumed 2,263 CPU hours even for the low-complexity dataset of CAMI [27]. We evaluated OGRE on stLFR linked-reads of ATCC-MSA-1003 (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 (more than two weeks). Pangaea with 100 threads only took 64.06 hours in real time, 514.63 hours in CPU time, and consumed 281.99GB of maximum RSS to group and assemble this linked-reads dataset.

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 relatively high cost for large cohort studies. In contrast, linked-read and hybrid sequencing (deep short-read and shallow long-read sequencing) techniques are cost-effective. Linked-read sequencing 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 60.98% fewer NCMAGs than linked-read assemblies from Pangaea

(**Supplementary Note 4**), indicating that some microbes might be lost due to insufficient long-read sequencing depth. Similar observations have been reported in previous studies [39]. Although stLFR and TELL-Seq linked-reads had high barcode specificity in ATCC-MSA-1003, 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 the 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 sequencing for linked-read sequencing

For ATCC-MSA-1003, 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 three individuals (S1, S2, and S3) 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 ATCC-MSA-1003 was prepared using the TELL-Seq Whole Genome Sequencing library prep kit, followed by 2×146 paired-end sequencing on an Illumina sequencing system.

Simulate stLFR linked-reads for ZYMO

We downloaded reference genomes of the 10 strains of the ZymoBIOMICS^{TM} Microbial Community Standard II (Log Distribution) [37], and used LRTK (v1.7) [41] to simulate stLFR linked-reads with the same strain composition as the ZYMO mock community.

Generate virtual barcodes to short-reads from long-reads

We attached the same long-read indexes as virtual barcodes to short-reads if they were aligned to the same long-read. We mapped the short-reads to long-reads using BWA-MEM (v0.7.17) [42] and removed the spurious alignments if the minimum aligned nucleotides were below 60bps. If a short-read was aligned to multiple long-reads, we randomly chose one of the long-read indexes as its barcode.

Extract k-mer histogram and TNFs from co-barcoded reads

We extracted k-mer histograms and TNFs from the co-barcoded reads if their total lengths were longer than 2Kb to ensure feature stability. A k-mer histogram was calculated based on global k-mer occurrences and could reflect the abundance features of the microbial genome [33]. We adopted $k = 15$ as used in the previous studies [33, 43], and calculated the global 15-mer frequencies using the whole dataset. We removed 15-mers with frequencies higher than 4,000 (to avoid repetitive sequences) and divided the global frequency distribution into 400 bins with equal sizes (the ith bin denoted frequencies between $10 * i - 10$ and $10 * i$). We collected the co-barcoded reads for each barcode and divided these reads into 15-mers, which were assigned to the 400 bins based on their global frequencies. We calculated the number of 15-mers allocated to each bin and generated a count vector with 400 dimensions as the k-mer histogram of the specific barcode. A TNF vector was constructed by calculating the frequencies of all 136 non-redundant 4-mers from co-barcoded reads. The k-mer histogram and TNF vector were normalized to eliminate the bias introduced by the different lengths of co-barcoded reads.

Binning co-barcoded reads with a VAE

The normalized k-mer histogram (X_A) and TNF vector (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 [44] and a dropout layer (P = 0.2) [45]. The output of the last layer was fed to two parallel latent layers with 32 hidden neurons for each to generate μ and σ for a Gaussian distribution $\mathcal{N}(\mu, \sigma^2)$, 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 normalized. The loss function ($Loss$) was defined as the weighted sum of three components: the reconstruction loss of k-mer histogram (L_A), the reconstruction loss of TNF vectors (L_T), and the Kullback-Leibler divergence loss (L_{KL}) between the latent and the prior standard Gaussian distributions. We adopted cross-entropy loss for L_A and L_T to deal with probability distributions, and formularized the loss terms 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 α and β, respectively (**Supplementary Note 3**). The VAE was trained with early stopping to reduce the training time and avoid overfitting. We used the RPH-kmeans [34] algorithm with random projection hashing to group the co-barcoded reads using their latent embeddings obtained from μ. We set the number of clusters equal to or smaller than the actual number of microbes in the microbial community. For the two mock communities of ZYMO and ATCC-MSA-1003, the numbers of clusters were set to 10 and 15, respectively. For all three real human gut microbiomes, we used 30 as the number of clusters.

Weighted sampling for VAE training

We designed weighted sampling to balance the training set of co-barcoded reads from microbes with different abundances. Considering that the highest value of X_A is negatively related to the abundance of the co-barcoded reads (**Supplementary Figure 1**), we used a heuristic function $max(X_A)^2$ as the sampling weight for the barcode of the co-barcoded reads. The square was to make the low-abundance microbes have a much higher sampling weight. The calculated sampling weights were automatically used by the WeightedRandomSampler of PyTorch to construct a balanced training dataset.

Multi-thresholding reassembly for low-abundance microbes

For co-barcoded read binning, we assembled reads in each cluster independently using MEGAHIT (v1.2.9) [31]. We designed a multi-thresholding reassembly strategy (**Figure 1 b**) to improve the assembly qualities of low-abundance microbes by recollecting the reads from the low-abundance microbes that were misclustered into different bins using read depth thresholds. To calculate the read depth of the contigs assembled from each read cluster (denoted as contigs$_{bin}$), we aligned the input reads to contigs$_{bin}$ using BWA-MEM (v0.7.17) [42] and calculate the read depth for each contig using the "jgi_summarize_bam_contig_depths" program in MetaBat2 (v2.12.1) [46]. To collect the reads of low-abundance microbes, we extracted the reads that could not be mapped to the high-depth contigs (with read depth $> t_i$) in contigs$_{bin}$ and assembled them using the standard short-read assembler,

metaSPAdes (v3.15.3) [30]. This step can be substituted by MEGAHIT to reduce the running time, and can also integrate available contigs to guide the path resolution in metaSPAdes to make the assembly of low-abundance microbes more efficient (e.g., using the contigs assembled by metaSPAdes from the whole read dataset as the input to "--untrusted-contigs" of metaSPAdes, which is used in our experiments and optional for Pangaea). We repeated this procedure with a range of thresholds ($T = \{t_i | i = 1, 2, ...\}$) producing contigs$_{low}$. We chose $T = \{10, 30\}$ for all the experiments, which worked well for both the mock communities and the human gut microbiomes.

Ensemble assembly

We use ensemble assembly to avoid incomplete metagenome assembly caused by the mis-binning of previous modules. For linked-read assembly, the ensemble strategy includes two steps: (i) we use contigs$_{bin}$ (contigs assembled from each read bin), contigs$_{low}$ (contigs from multi-thresholding re-assembly), contigs$_{local}$ (contigs from the local assembly of Athena) and contigs$_{ori}$ (contigs assembled from short-reads by metaSPAdes [v3.15.3]) using an OLC-based assembler, metaFlye (v2.8) with the "--subassemblies" parameter; (ii) quickmerge (v0.3) [47] was used to merge the contigs from step (i) and Athena contigs. For the assembly of short-reads with virtual barcodes, we substituted the metaSPAdes in step (i) and Athena in step (ii) with the corresponding hybrid assemblies (contigs generated from hybridSPAdes [v3.15.3] [16] or OPERA-MS [v0.8.3] [15]). Step (ii) is optional for linked-read assembly since Athena is already integrated in Step (i).

Detecting circular contigs

We adopted the circularization module of Lathe [39] to detect circular contigs in all the assemblies. The module needs long-reads as input which is not available for linked-read assembly, so we modified the alignment and assembly parameters in the circularization module to accept contigs as input, and merged the contigs$_{ori}$, contigs$_{bin}$, contigs$_{low}$, and contigs$_{local}$ as "pseudo long-reads" for running it.

Reconstructing physical long fragments based on reference genomes

We reconstructed the physical long fragments from linked-reads of ATCC-MSA-1003 to calculate $N_{F/B}$. The linked-reads were mapped to the reference genomes using BWA-MEM (v0.7.17) [42] with option "-C" to retain the barcode information in the alignment file, followed by sorting based on read alignment coordinates using SAMtools (v1.9) [48]. We connected the co-barcoded reads into long fragments if their coordinates were within 10Kb on the reference genome. Each fragment was required to include at least two read pairs and to be no shorter than 1Kb.

Metagenome assembly of the other assemblers on different datasets

The 10x, stLFR, and TELL-Seq sequencing datasets were demultiplexed to generate raw linked-reads using Long Ranger (v2.2.0) [49], stLFR_read_demux (Git version 3ecaa6b) [17] and LRTK (Git version 28012df) [41], respectively. The linked-reads were assembled using metaSPAdes (v3.15.3) [30], MEGAHIT (v1.2.9) [31], cloudSPAdes (v3.12.0-dev) [30], Athena (v1.3) [14] and Supernova (v2.1.1) [20]. Supernova does not accept raw stLFR linked-reads as input, so we applied stlfr2supernova_pipeline (https://github.com/BGI-Qingdao/stlfr2supernova_pipeline; Git version 95f0848) to convert the barcode format of stLFR. The scaffolds produced by Supernova were broken into contigs at successive "N"s longer than 10 before evaluation. The datasets with both short- and long-reads were assembled using OPERA-MS (v0.8.3) [15], hybridSPAdes (v3.15.3) [16], metaSPAdes (v3.15.3, only short-reads) and Athena (v1.3, short-reads with virtual barcodes). The PacBio CLR long-reads from S1 and S2 were assembled using metaFlye (v2.8) [50] with the "--pacbio-raw" parameter. All the assemblers were run with default parameters. For measuring computational performance, we set the threads of all the assemblers to 100 (if the assembler had this option), and used the command "/usr/bin/time -v" to report the system time, user time, and maximum RSS consumed by the programs.

Benchmarking on the mock microbial communities

The reference genomes of ATCC-MSA-1003 and ZYMO were downloaded from the NCBI reference databases (**Supplementary Table 1**) and the previous study on ZYMO [37], respectively. The contigs assembled for the two mock communities were assessed using MetaQUAST (v5.0.2) [51], with the option "-m 1000 --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 MAG quality evaluation

We aligned the linked-reads (or short-reads) to the contigs using BWA-MEM (v0.7.17) [42] and calculated the read depths using "jgi_summarize_bam_contig_depths" in MetaBat2 (v2.12.1) [46]. The contigs with read depths were binned into MAGs using MetaBat2 (v2.12.1) with default parameters. CheckM (v1.1.2) [52] was used to report the completeness and contamination of the MAGs. ARAGORN (v1.2.38) [53] and barrnap (v0.9) [54] were used to annotate the transfer RNAs (tRNAs) and rRNAs (5S, 16S, and 23S rRNAs), respectively. According to standard criteria of the minimum information about MAGs [55], we classified the MAGs into near-complete (completeness > 90%, con-

tamination $< 5\%$, and could be detected 5S, 16S, and 23S rRNAs, and at least 18 tRNAs), high-quality (completeness $> 90\%$, and contamination $< 5\%$), medium-quality (completeness $\geq 50\%$, and contamination $< 10\%$), and low-quality (the other MAGs).

Annotation of the MAGs and the closest reference genomes

The contigs were annotated using Kraken2 (v2.1.2) with the custom database built from the NT database of NCBI (Aug 20, 2022). We used the "--fast-build" option of kraken2-build to reduce the database construction time. Subsequently, the "assign_species.py" script from "metagenomics_workflows" [14, 39] 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; database version r207) [56], which also reported the alignment identities and alignment fractions between them.

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 the manuscript's 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 open project of BGI-Shenzhen, Shenzhen 518000, China (BGIRSZ20220012), the Hong Kong Research Grant Council Early Career Scheme (HKBU 22201419), HKBU Start-up Grant Tier 2 (RC-SGT2/19-20/SCI/007), HKBU IRCMS (No. IRCMS/19-20/D02), the Guangdong Basic and Applied Basic Research Foundation (No. 2021A1515012226) and Shenzhen Science and Technology Innovation Commission (SZSTI) - Shenzhen Virtual University Park (SZVUP) Special Fund Project (No. 2021Szvup135).

Abbreviations

MAG: Metagenome-Assembled Genome

NCMAG: Near-Complete Metagenome-Assembled Genome

OGRE: Overlap Graph-based Read clustEring

PacBio CLR: PacBio Continuous Long-Reads

RSS: Resident Set Size

rRNA: Ribosomal RNA

stLFR: single-tube Long Fragment Read

TELL-Seq: Transposase Enzyme-Linked Long-read Sequencing

TNF: TetraNucleotide Frequency

tRNA: Transfer RNA

UST: Universal Sequencing Technology

VAE: Variational AutoEncoder

Availability of data and materials

The 10x linked-reads of the ATCC-MSA-1003 mock community were downloaded from NCBI SRA accession SRR12283286. The stLFR and TELL-Seq sequencing data of ATCC-MSA-1003 was uploaded to NCBI BioProject PRJNA875547. The stLFR sequencing data of the three human gut microbiomes was deposited in China National GeneBank (CNGB) project CNP0003432. The PacBio CLR long-reads of ATCC-MSA-1003, S1 and S2 were downloaded from NCBI SRA accessions SRR12371719, SRR19505636 and SRR19505632, respectively. The ONT long-reads of ZYMO were available from the NCBI SRA accession ERR3152366. Codes of Pangaea and all the command lines are available at https://github.com/ericcombiolab/Pangaea. The version of the source code used in the manuscript was uploaded to https://doi.org/10.5281/zenodo.7996172.

Ethics approval and consent to participate

Not applicable

Competing interests

LJH is an employee of Kangmeihuada GeneTech Co., Ltd (KMHD).

Consent for publication

Not applicable

Authors' contributions

LZ conceived the study. ZMZ, LZ and JX designed the Pangaea algorithms. ZMZ and JX implemented the Pangaea software. LZ and ZMZ conceived the experiments. ZMZ and JX conducted the experiments. ZMZ, LZ and JX 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.

Additional Files

Supplementary Text and Figures

Supplementary Notes 1-4, Supplementary Figures 1-5, and Supplementary Tables 1-3, 6, and 9.

Supplementary Table 4

The impact of VAE-based binning algorithm adopted by Pangaea on the final assembly results using TELL-Seq linked-reads of the ATCC-MSA-1003 mock community.

Supplementary Table 5

The impact of different ranges of abundance thresholds (Ts) on the final assembly using the TELL-Seq linked-reads of the ATCC-MSA-1003 mock community.

Supplementary Table 7

The per-strain NA50 and NGA50 generated by different assembly tools for the 15 strains with abundances higher than 0.1% using TELL-Seq, stLFR, and 10x linked-reads from ATCC-MSA-1003.

Supplementary Table 8

The genome fractions generated by different assembly tools for the five strains with abundances of 0.02% using TELL-Seq, stLFR, and 10x linked-reads from ATCC-MSA-1003.

Supplementary Table 10

The genome fractions, per-strain NA50 and NGA50 generated by different assembly tools for the 10 strains using the stLFR linked-reads from ZYMO.

Supplementary Table 11

The statistics of MAGs generated for the three human gut microbiomes by all assemblers.

Supplementary Table 12

The closest reference genomes of the NCMAGs with species-level annotations generated by all assemblers and their alignment statistics reported by GTDB-Tk (v2.1.0).

Supplementary Table 13

The genome fractions, total assembly length for contigs longer than 10Kb, per-strain NA50 and NGA50 by different assemblers using short- and long-reads from the two mock communities.

References

[1] Yang, C. *et al.* A review of computational tools for generating metagenome-assembled genomes from metagenomic sequencing data. *Computational and Structural Biotechnology Journal* **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. *Nature biotechnology* **32**, 834–841 (2014).

[4] Almeida, A. *et al.* A unified catalog of 204,938 reference genomes from the human gut microbiome. *Nature 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] Rhoads, A. & Au, K. F. PacBio sequencing and its applications. *Genomics, proteomics & bioinformatics* **13**, 278–289 (2015).

[7] 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).

[8] Wenger, A. M. *et al.* Accurate circular consensus long-read sequencing improves variant detection and assembly of a human genome. *Nature biotechnology* **37**, 1155–1162 (2019).

[9] Vollger, M. R. *et al.* Long-read sequence and assembly of segmental duplications. *Nature 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. *Nature biotechnology* **38**, 701–707 (2020).

[12] Amarasinghe, S. L. *et al.* Opportunities and challenges in long-read sequencing data analysis. *Genome biology* **21**, 1–16 (2020).

[13] Zhang, Z., Yang, C., Veldsman, W. P., Fang, X. & Zhang, L. Benchmarking genome assembly methods on metagenomic sequencing data. *Briefings in Bioinformatics* **24**, bbad087 (2023).

[14] Bishara, A. *et al.* High-quality genome sequences of uncultured microbes by assembly of read clouds. *Nature biotechnology* **36**, 1067–1075 (2018).

[15] Bertrand, D. *et al.* Hybrid metagenomic assembly enables high-resolution analysis of resistance determinants and mobile elements in human microbiomes. *Nature biotechnology* **37**, 937–944 (2019).

[16] Antipov, D., Korobeynikov, A., McLean, J. S. & Pevzner, P. A. hybridSPAdes: an algorithm for hybrid assembly of short and long reads. *Bioinformatics* **32**, 1009–1015 (2016).

[17] 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).

[18] 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).

[19] Tolstoganov, I., Bankevich, A., Chen, Z. & Pevzner, P. A. cloudSPAdes: assembly of synthetic long reads using de Bruijn graphs. *Bioinformatics* **35**, i61–i70 (2019).

[20] 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).

[21] Qi, Y. *et al.* MetaTrass: A high-quality metagenome assembler of the human gut microbiome by cobarcoding sequencing reads. *iMeta* **1**, e46 (2022).

[22] Martin, S. *et al.* Capturing variation in metagenomic assembly graphs with MetaCortex. *Bioinformatics* **39**, btad020 (2023).

[23] Hug, L. A. *et al.* Critical biogeochemical functions in the subsurface are associated with bacteria from new phyla and little studied lineages. *Environmental microbiology* **18**, 159–173 (2016).

[24] Cattonaro, F., Spadotto, A., Radovic, S. & Marroni, F. Do you cov me? effect of coverage reduction on metagenome shotgun sequencing studies. *F1000Research* **7** (2018).

[25] Ghurye, J. S., Cepeda-Espinoza, V. & Pop, M. Focus: microbiome: metagenomic assembly: overview, challenges and applications. *The Yale journal of biology and medicine* **89**, 353 (2016).

[26] Yorki, S. *et al.* Comparison of long-and short-read metagenomic assembly for low-abundance species and resistance genes. *Briefings in Bioinformatics* **24**, bbad050 (2023).

[27] Balvert, M., Luo, X., Hauptfeld, E., Schönhuth, A. & Dutilh, B. E. OGRE: Overlap Graph-based metagenomic Read clustEring. *Bioinformatics* **37**, 905–912 (2021).

[28] 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).

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

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

[31] 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).

[32] Laehnemann, D., Borkhardt, A. & McHardy, A. C. Denoising DNA deep sequencing data–high-throughput sequencing errors and their correction. *Briefings in bioinformatics* **17**, 154–179 (2016).

[33] Wickramarachchi, A., Mallawaarachchi, V., Rajan, V. & Lin, Y. MetaBCC-LR: meta genomics binning by coverage and composition for long reads. *Bioinformatics* **36**, i3–i11 (2020).

[34] 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 and bioinformatics* **2**, lqaa082 (2020).

[35] Wickramarachchi, A. & Lin, Y. Binning long reads in metagenomics datasets using composition and coverage information. *Algorithms for Molecular Biology* **17**, 1–15 (2022).

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

[37] Nicholls, S. M., Quick, J. C., Tang, S. & Loman, N. J. Ultra-deep, long-read nanopore sequencing of mock microbial community standards. *Gigascience* **8**, giz043 (2019).

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

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

[40] Mattock, J. & Watson, M. A comparison of single-coverage and multi-coverage metagenomic binning reveals extensive hidden contamination. *Nature Methods* 1–4 (2023).

[41] Yang, C., Zhang, Z., Liao, H. & Zhang, L. LRTK: A unified and versatile toolkit for analyzing linked-read sequencing data. *bioRxiv* 2022–08 (2022).

[42] Li, H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. *arXiv preprint arXiv:1303.3997* (2013).

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

[44] 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).

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

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

[47] 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).

[48] Li, H. *et al.* The sequence alignment/map format and SAMtools. *Bioinformatics* **25**, 2078–2079 (2009).

[49] Marks, P. *et al.* Resolving the full spectrum of human genome variation using Linked-Reads. *Genome research* **29**, 635–645 (2019).

[50] Kolmogorov, M. *et al.* metaFlye: scalable long-read metagenome assembly using repeat graphs. *Nature Methods* **17**, 1103–1110 (2020).

[51] Mikheenko, A., Saveliev, V. & Gurevich, A. MetaQUAST: evaluation of metagenome assemblies. *Bioinformatics* **32**, 1088–1090 (2016).

[52] 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).

[53] Laslett, D. & Canback, B. ARAGORN, a program to detect tRNA genes and tmRNA genes in nucleotide sequences. *Nucleic acids research* **32**, 11–16 (2004).

[54] Seemann, T. barrnap. https://github.com/tseemann/barrnap (2018).

[55] Bowers, R. M. *et al.* Minimum information about a single amplified genome (MISAG) and a metagenome-assembled genome (MIMAG) of bacteria and archaea. *Nature biotechnology* **35**, 725–731 (2017).

[56] 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 (2019).

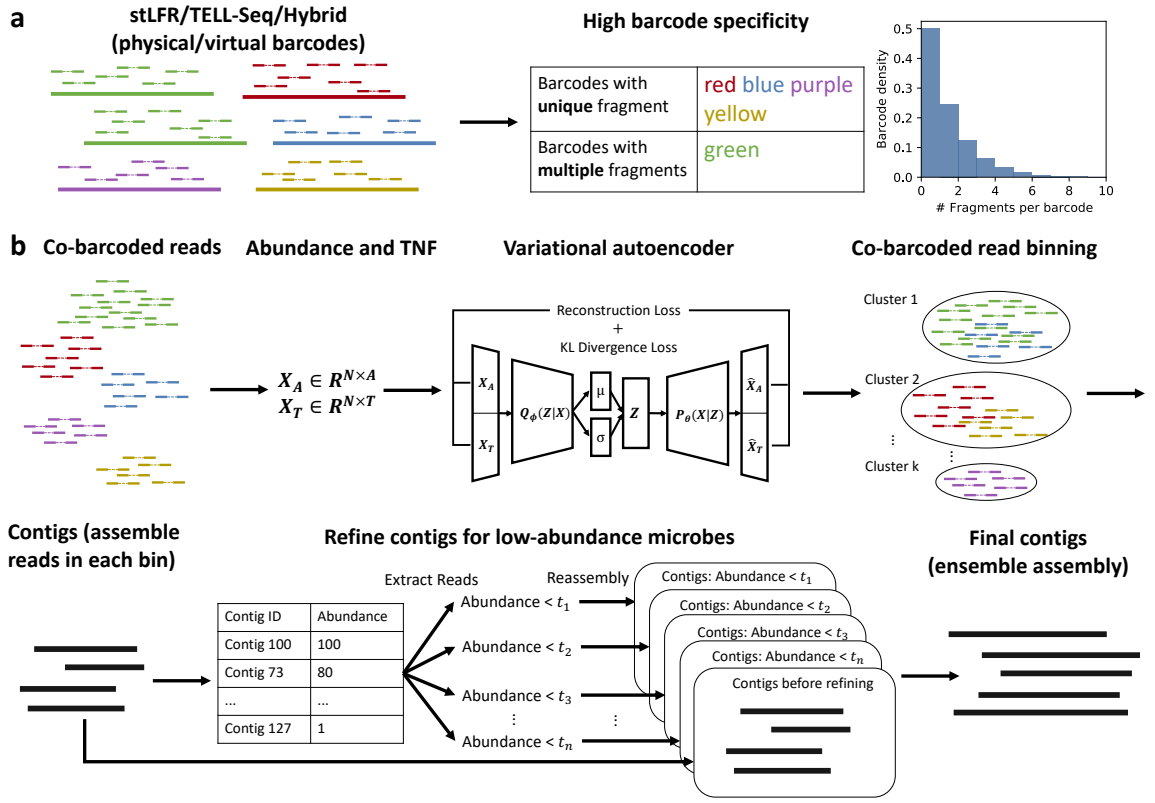

Figure 1: Workflow of Pangaea on stLFR and TELL-Seq linked-reads (physical barcodes) and short-reads with virtual barcodes from long-reads. (**a**) High barcode specificity for stLFR and TELL-Seq linked-reads. (**b**) Pangaea extracts features including k-mer histograms and TNFs from co-barcoded reads. The features are concatenated and used to represent reads in low-dimensional latent space using a variational autoencoder. The embeddings of co-barcoded reads are clustered by RPH-kmeans. Pangaea assembles the reads from each bin independently and adopts a multi-thresholding reassembly strategy to improve the assemblies for low-abundance microbes. Ensemble assembly integrates the contigs from different strategies using OLC algorithm.

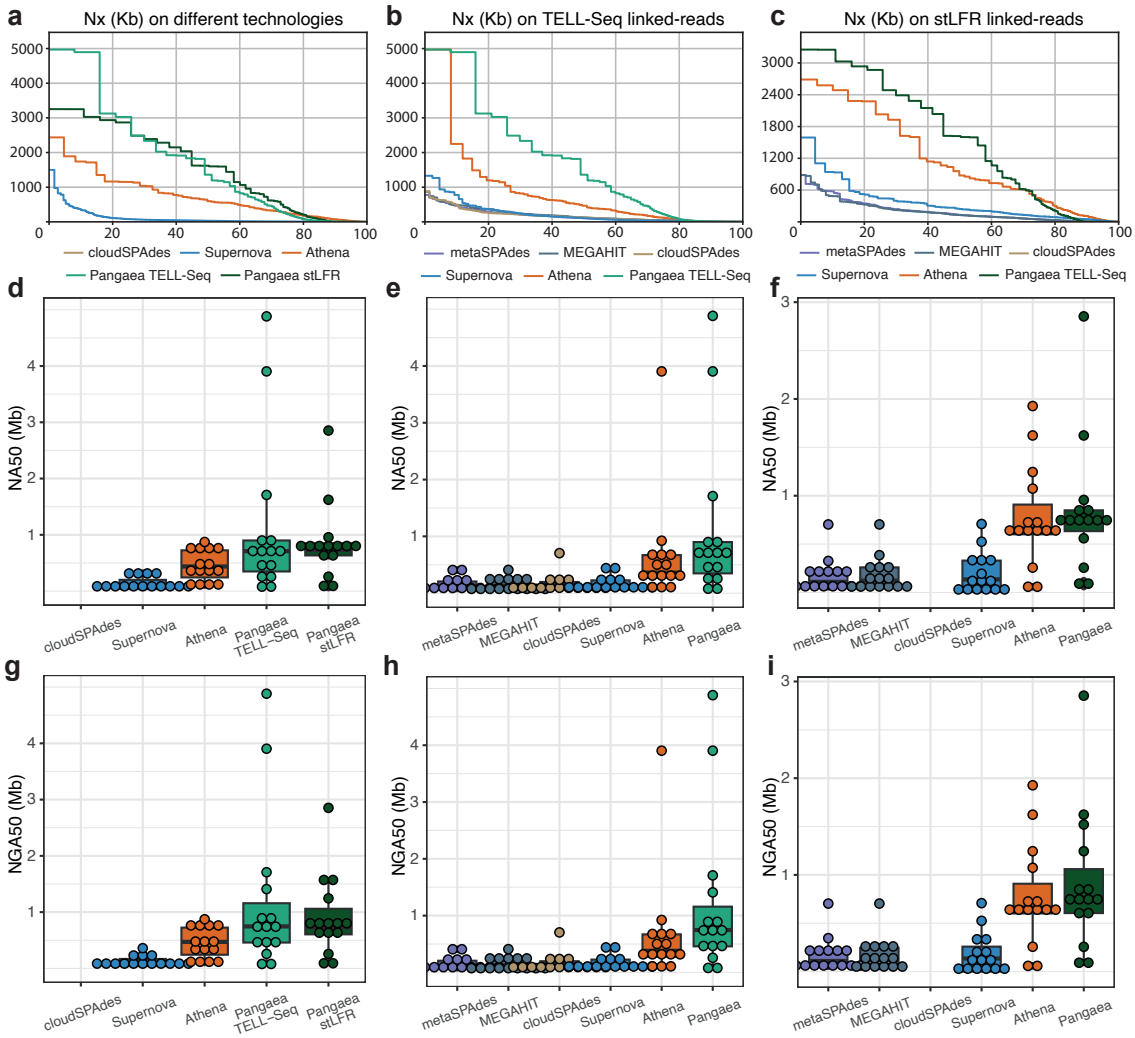

Figure 2: Nx, with x ranging from 0 to 100, for different linked-read sequencing technologies (**a**; except Panagea, all assemblers were applied to 10x linked-reads), TELL-Seq linked-reads (**b**), and stLFR linked-reads (**c**) from ATCC-MSA-1003 mock community. NA50 and NGA50 for the 15 strains with abundances higher than 0.1% assembled for different linked-read sequencing technologies (**d and g**; except Panagea, all assemblers were applied to 10x linked-reads), TELL-Seq linked-reads (**e and h**), and stLFR linked-reads (**f and i**) on ATCC-MSA-1003. cloudSPAdes was unavailable for 10x linked-reads and stLFR linked-reads, as it requires extremely large memory (>1TB) on these datasets.

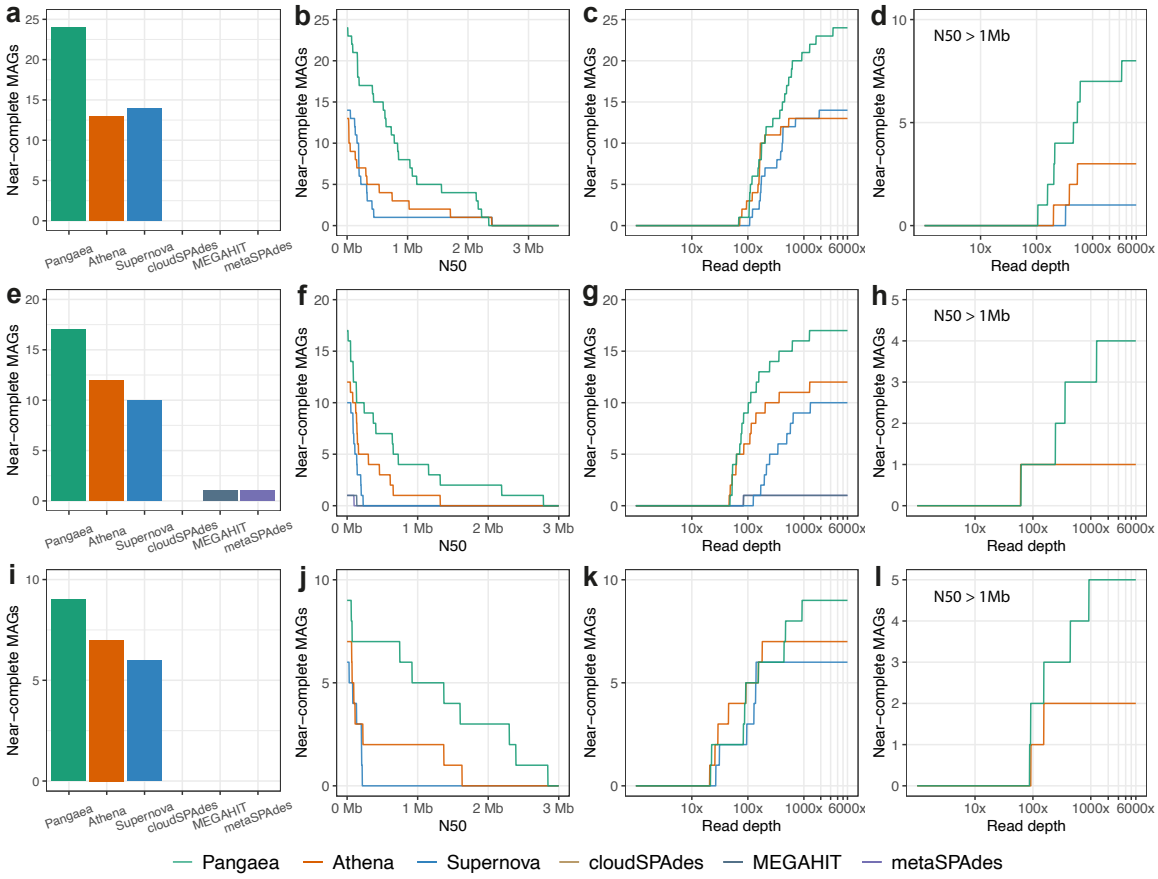

Figure 3: The number of NCMAGs generated by different assembly tools (**a, e, and i**). The number of NCMAGs by thresholding minimum N50 (**b, f, and j**) and maximum read depth (**c-d, g-h, and k-l**). We compared the performances of Pangaea, Athena, Supernova, cloudSPAdes, MEGAHIT and metaSPAdes on stLFR linked-reads from S1 (**a-d**), S2 (**e-h**) and S3 (**i-l**). cloudSPAdes was not available for S1 and S2 because it did not finish assembly within two weeks using 100 threads.

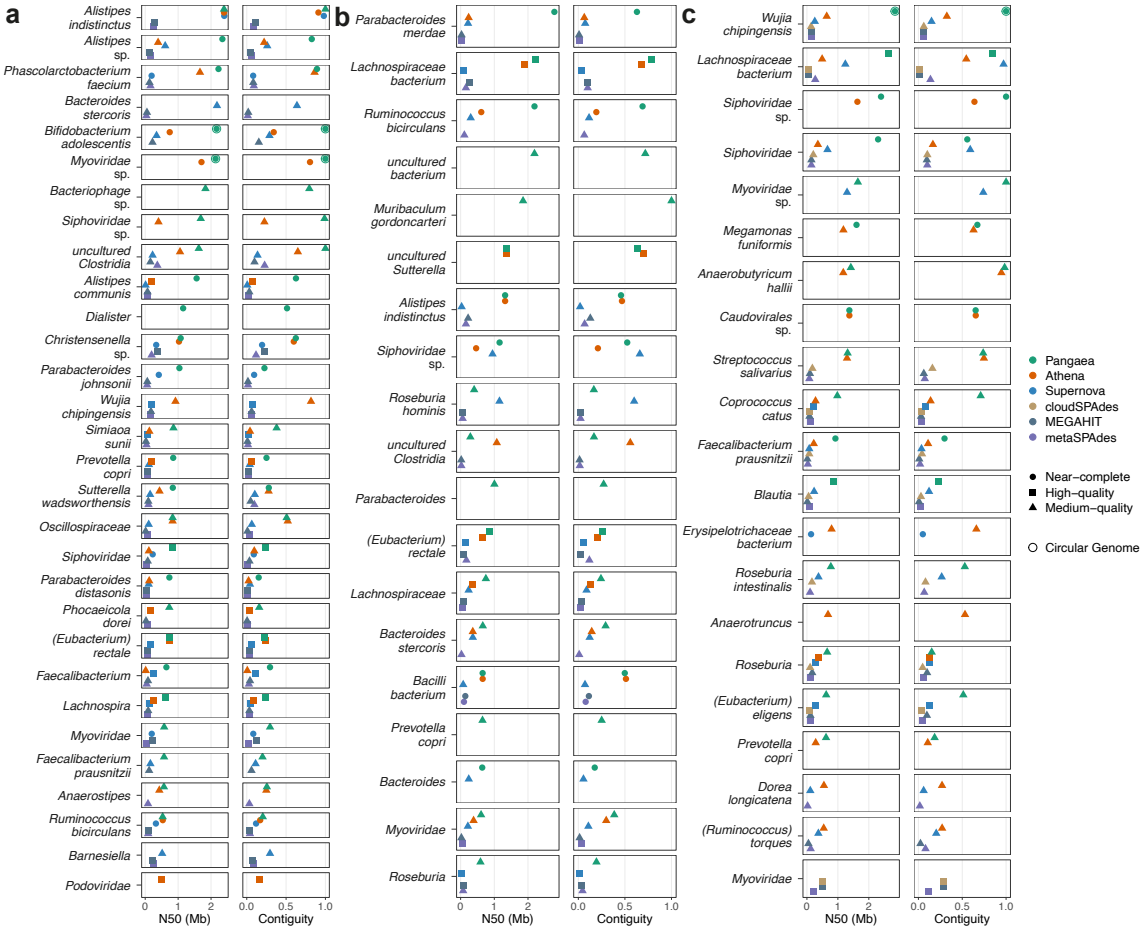

Figure 4: The annotated microbes for the MAGs produced by Pangaea, Athena, Supernova, cloudSPAdes, MEGAHIT and metaSPAdes from S1 (**a**), S2 (**b**) and S3 (**c**). The microbes were shown here if the N50s of their corresponding MAGs were larger than 500Kb by any assembler. If the same microbe was annotated by more than one MAG from the same assembler, the one with the highest N50 was selected. cloudSPAdes was not available for S1 and S2 because it did not finish assembly within two weeks using 100 threads.

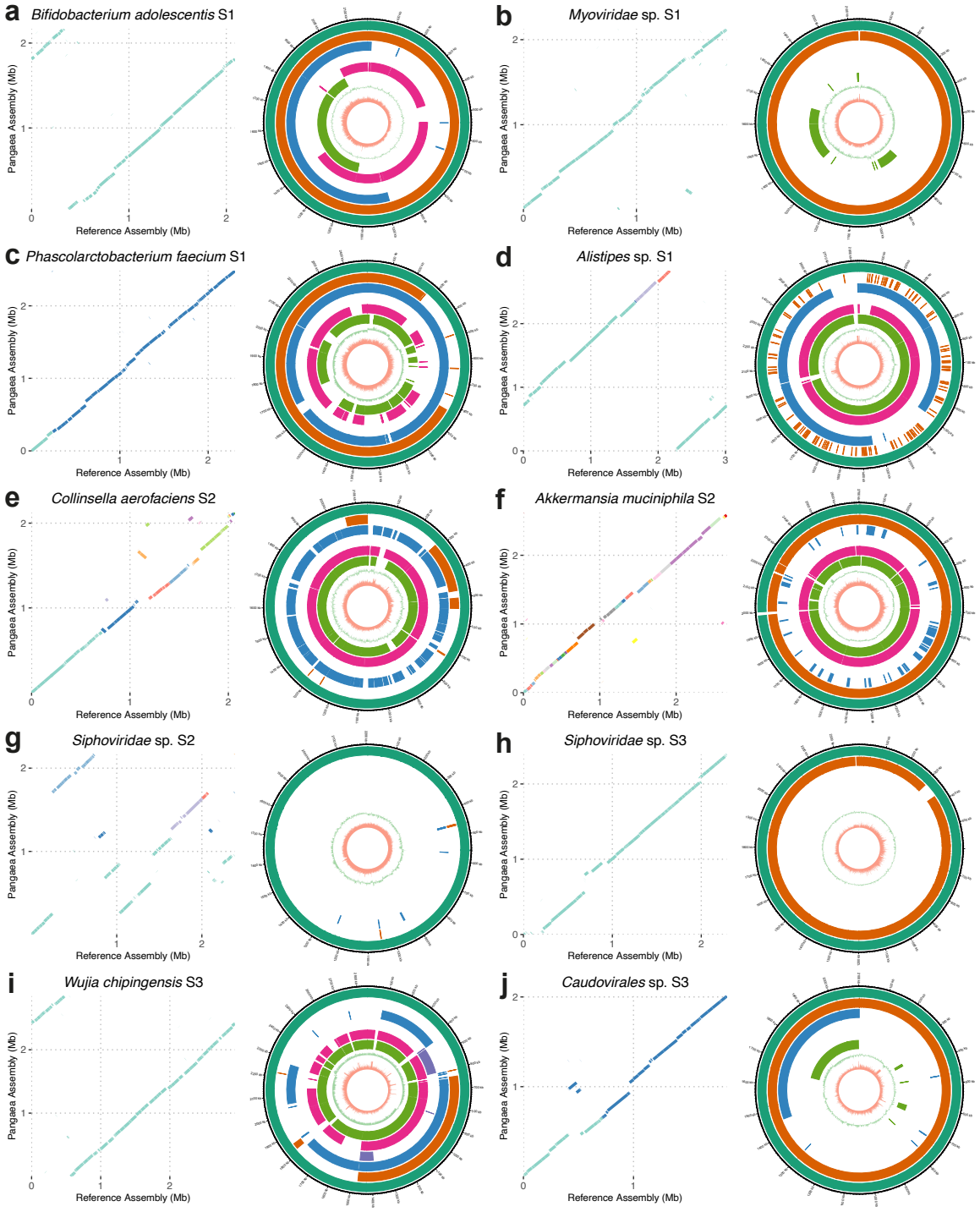

Figure 5: Genome collinearity analysis between the selected NCMAGs produced by Pangaea and their closest reference genomes (dot plots), and comparison of the corresponding MAGs produced by different assemblers (circos plots) from S1, S2 and S3. **a, b and i** are species for which Pangaea obtained complete and circular MAGs. Colors are used in the dot plots to distinguish different contigs. The eight rings in the circos plots from outside to inside denote the Pangaea MAGs (dark green), Athena MAGs (orange), Supernova MAGs (blue), cloudSPAdes MAGs (purple), MEGAHIT MAGs (pink), metaSPAdes MAGs (light green), 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 5**.

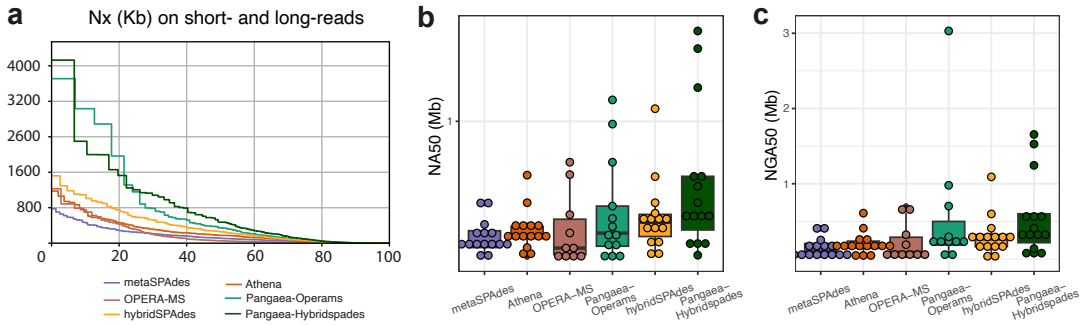

Figure 6: Nx with x ranging from 0 to 100 (**a**), NA50 of the 15 strains with abundance > 0.1% (**b**), and NGA50 of the 15 strains with abundance > 0.1% (**c**), assembled by different assemblers using short-reads (metaSPAdes), short-reads with virtual barcodes (Athena, Pangaea-Operams and Pangaea-Hybridspades), and using short- and long-reads (OPERA-MS and hybridSPAdes) on ATCC-MSA-1003 mock community.

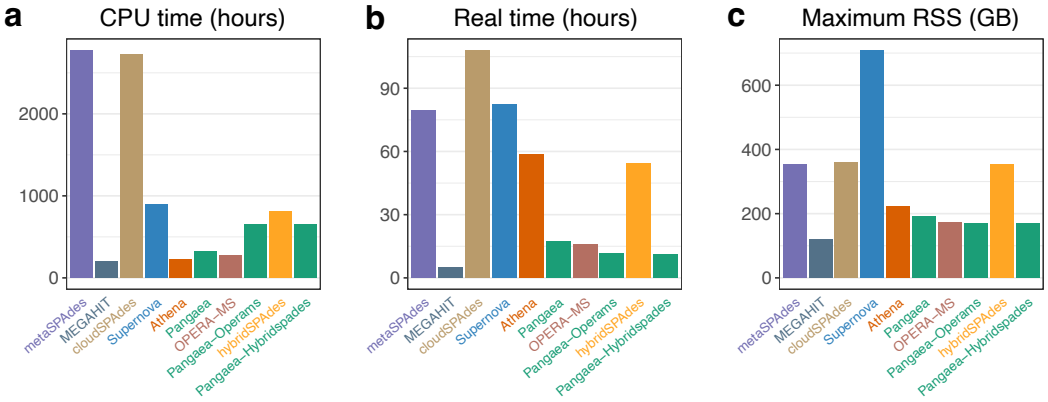

Figure 7: The CPU time (**a**), real time (**b**) and maximum resident set size (RSS; **c**) of the assemblers using short-reads (metaSPAdes and MEGAHIT), linked-reads (cloudSPAdes, Supernova, Athena and Pangaea) and using short- and long-reads (OPERA-MS, Pangaea-Operams, hybridSPAdes and Pangaea-Hybridspades) on ZYMO mock community.

Table 1: Assembly statistics for different assemblers using the linked-reads datasets on the mock communities and human gut microbiomes. The highest values are in bold.

			ZYMO			
	Pangaea	**Athena**	**Supernova**	**cloudSPAdes**	**MEGAHIT**	**metaSPAdes**
Total assembly length	**36,094,665**	35,254,802	25,103,660	35,843,842	35,358,538	35,511,020
Genome fraction (%)	**49.24**	48.14	34.27	47.99	48.38	48.46
Longest alignment	**2,839,942**	2,717,703	768,486	1,012,282	1,079,942	847,644
Overall N50	**1,094,665**	761,749	288,912	210,427	124,248	191,688
Overall NA50	**1,072,622**	760,284	226,947	208,894	116,738	191,688
NGA50 per strain	**1,083,616**	938,270	132,015	246,262	177,318	179,371
NA50 per strain	**1,066,276**	927,278	136,210	245,562	179,245	183,935
			ATCC-MSA-1003 (stLFR)			
	Pangaea	**Athena**	**Supernova**	**cloudSPAdes**	**MEGAHIT**	**metaSPAdes**
Total assembly length	**59,484,233**	52,159,846	35,226,545	-	55,506,708	57,225,487
Genome fraction (%)	**84.43**	77.12	52.21	-	81.99	83.99
Longest alignment	**2,853,278**	2,281,647	1,105,108	-	883,580	883,552
Overall N50	**1,619,916**	875,747	243,194	-	127,879	132,556
Overall NA50	**731,990**	677,911	215,052	-	116,995	125,586
NGA50 per strain	**677,353**	575,371	137,023	-	133,877	133,978
NA50 per strain	**628,059**	576,621	145,646	-	140,463	134,477
			ATCC-MSA-1003 (TELL-Seq)			
	Pangaea	**Athena**	**Supernova**	**cloudSPAdes**	**MEGAHIT**	**metaSPAdes**
Total assembly length	61,990,266	60,847,375	56,748,937	**62,316,993**	60,291,592	60,648,311
Genome fraction (%)	**82.63**	81.99	76.46	82.44	82.10	82.46
Longest alignment	**4,968,123**	4,968,084	1,096,372	884,364	867,473	776,102
Overall N50	**1,360,322**	466,498	102,757	127,419	128,069	112,342
Overall NA50	**649,672**	361,569	97,312	118,159	112,513	105,630
NGA50 per strain	**887,107**	485,196	121,657	129,531	121,938	118,391
NA50 per strain	**838,457**	483,734	123,277	129,001	122,932	119,253
			Human gut microbiome (S1)			
	Pangaea	**Athena**	**Supernova**	**cloudSPAdes**	**MEGAHIT**	**metaSPAdes**
Total assembly length	**488,785,611**	469,284,964	311,971,769	-	459,128,709	452,598,342
Longest contig	2,394,379	2,394,379	**2,400,768**	-	721,029	697,064
Overall N50	**64,394**	44,759	60,619	-	13,670	14,325
			Human gut microbiome (S2)			
	Pangaea	**Athena**	**Supernova**	**cloudSPAdes**	**MEGAHIT**	**metaSPAdes**
Total assembly length	**408,819,148**	393,685,495	290,599,879	-	381,215,035	374,166,135
Longest contig	**2,877,256**	1,903,088	1,152,844	-	480,921	443,089
Overall N50	**188,161**	119,620	72,947	-	23,524	23,546
			Human gut microbiome (S3, contig > 1,000bp)			
	Pangaea	**Athena**	**Supernova**	**cloudSPAdes**	**MEGAHIT**	**metaSPAdes**
Total assembly length	**323,679,273**	270,721,423	222,438,535	284,200,415	301,723,162	298,374,739
Longest contig	**4,769,781**	4,570,813	1,384,668	526,768	526,790	592,456
Overall N50	105,969	**114,886**	79,377	19,102	15,376	15,081
			Human gut microbiome (S3, contig > 5,000bp)			
	Pangaea	**Athena**	**Supernova**	**cloudSPAdes**	**MEGAHIT**	**metaSPAdes**
Total assembly length	**249,885,090**	237,573,809	192,606,058	196,090,881	209,538,475	203,743,361
Longest contig	**4,769,781**	4,570,813	1,384,668	526,768	526,790	592,456
Overall N50	**275,648**	157,810	105,541	44,589	33,264	36,343

Table 2: Assembly statistics for different assemblers using short-reads with virtual barcodes from long-reads on the ZYMO and ATCC-MSA-1003. The highest values are in bold.

	ZYMO					
	Pangaea-Hybirdspades	**hybridSPAdes**	**Pangaea-Operams**	**OPERA-MS**	**Athena**	**metaSPAdes**
Total assembly length	35,806,901	35,517,607	**35,917,521**	35,334,818	34,567,225	35,511,020
Genome fraction (%)	48.47	**48.63**	48.62	48.28	47.34	48.46
Longest alignment	2,350,555	1,799,936	**3,579,300**	3,579,284	1,282,835	847,644
Overall N50	**444,276**	289,924	347,153	131,207	237,784	191,688
Overall NA50	**342,859**	287,773	301,599	124,358	232,219	191,688
NGA50 per strain	704,255	383,552	**757,488**	681,358	210,268	179,371
NA50 per strain	697,544	387,101	**756,021**	683,735	214,809	183,935
	ATCC-MSA-1003					
	Pangaea-Hybirdspades	**hybridSPAdes**	**Pangaea-Operams**	**OPERA-MS**	**Athena**	**metaSPAdes**
Total assembly length	**61,432,598**	60,794,505	53,145,042	44,017,454	60,077,880	60,648,311
Genome fraction (%)	82.69	**82.93**	63.25	57.87	81.57	82.46
Longest alignment	2,824,147	1,518,615	**3,027,146**	885,414	1,175,563	776,102
Overall N50	**475,582**	255,665	345,524	55,544	176,953	112,342
Overall NA50	**338,892**	236,258	236,263	49,863	173,018	105,630
NGA50 per strain	**412,875**	229,492	311,080	105,557	157,166	118,391
NA50 per strain	**410,176**	229,866	219,172	104,942	158,550	119,253