MetaTrass: High-quality **Meta**genomic **T**axonomic **R**ead **A**ssembly of **S**ingle-**S**pecies based on co-barcoding sequencing data and references

Yanwei Qi[a,b,c#], Shengqiang Gu[a,d#], Yue Zhang[a], Lidong Guo[a,d], Mengyang Xu[a,b,c,e], Xiaofang Cheng[e,f], Ou Wang[e,f], Jianwei Chen[a], Xiaodong Fang[e,g], Xin Liu[a,b,c], Li Deng[a,b,c*], and Guangyi Fan[a,b,c,e*]

[a]BGI-Qingdao, BGI-Shenzhen, Qingdao 266555, China

[b]State Key Laboratory of Agricultural Genomics, BGI-Shenzhen, Shenzhen 518083, China

[c]China National GeneBank, BGI-Shenzhen, Shenzhen 518120, China

[d]College of Life Sciences, University of Chinese Academy of Sciences, Beijing 100049, China

[e]BGI-Shenzhen, Shenzhen 518083, China

[f]MGI, BGI-Shenzhen, Shenzhen 518083, China

[g]BGI Genomics, BGI-Shenzhen, Shenzhen 518083, China

[#]These authors contributed equally to this work.

[*]Corresponding authors: Guangyi Fan (fanguangyi@genomics.cn) and Li Deng (dengli1@genomics.cn).

Abstract

With the development of sequencing technologies and computational analysis in metagenomics, the genetic diversity of non-conserved region has been receiving increasing attention to unravel complex microbial communities. However, it remains a challenge to obtain enough microbial genome drafts at a high resolution from a microbial community sample. In this work, we presented MetaTrass, a reference-guided assembling pipeline, which exploited both the public microbe reference genomes and long-range co-barcoding information, to assemble high-quality draft genomes from metagenomic co-barcoding reads. By applying this approach to single tube long fragment reads (stLFR) datasets of four human faeces samples, MetaTrass could generate more high-quality genome drafts (>90% completeness, <5%

contamination) with longer contiguity and higher resolution in comparison with the common combination strategies of genome assembling and binning. Total of 178 high-quality genomes were assembled by MetaTrass, comparing to 58 high-quality genomes assembled by the combination strategies. These high-quality genomes paved the way of the genetic diversity and evolution analysis among different samples. Thus, MetaTrass will facilitate the study of spatial and dynamics of complex microbial communities at high resolution. The open-source code of MetaTrass is available at https://github.com/BGI-Qingdao/MetaTrass.

Introduction

Through sequencing and analyzing the DNA of microbial communities directly from the environment, metagenomics has shown its important role in advancing the study of uncultured microbiomes [1, 2]. Comprehensive databases of metagenome-assembled genomes are massively expanded to completely understand the genomic taxonomic structure of different microbiome communities according to genetic similarity [3, 4]. The progresses in metagenomics have shed new light on the study of spatial distribution and dynamics of complex microbial communities from human gut [5, 6].

Based on the function mining of high-quality strain-resolved genomes, it is realized that genotypic difference among strains from the same species strongly correlated with their phenotype difference [7, 8]. The importance of genetic diversity within a species have been intensively studied in the field of pathogenicity, and many species with both pathogenic and commensal strains have been found [9, 10]. Indeed, the percentage of conserved genes shared between strains within a species is as low as 40% [11], so the large part of non-conservation region is thought as the genetic origin of phenotypic diversity. Thus, complete genome drafts from a microbiome sample at species resolution will enable a more comprehensive studies of intra-species genome diversity, but it is still a challenge to generate high-quality genomes from metagenomic datasets.

Most of current approaches to analyze the microbiome communities are based on high-throughput and low-cost next-generation sequencing (NGS) reads. Many highly modularized computational tools have been developed such as genome assembler, genome binner, taxonomic binner and taxonomic profiler [12, 13]. The combination of genome assembler and binner has been commonly used to generate metagenome-assembled genomes. A large number of short reads from a microbial community are firstly assembled to generate longer sequences by metagenomic assemblers with consideration of the uneven coverage depth of different microbial species [14-16]. Then the assembled sequences are grouped into more comprehensive genome drafts by genome binner based on similar *K-mer* composition and read coverage [17-19]. However, the limited NGS read length makes it impossible to address inherent complexity of microbial sample caused by long repeats and uneven abundance among different species.

Many sequencing technologies with long-range information accompanied with specialized computational tools are promised to overcome the problem of long repeats and uneven abundance. Third-generation single-molecule real-time sequencing (TGS) technologies developed by Pacific Biosciences and Oxford Nanopore can produce contiguous reads with length up to 100 kb, and show great potential to generate complete genomes from both cultured and uncultured microbial communities [20-22]. Using the chromatin-level contact probability information generated by high-throughput chromosome conformation capture (Hi-C) technology, more high-quality genome bins can be retrieved from short-read genome drafts and the contiguity can also be significantly improved [23]. The co-abundance of species in multiple samples with the common *K-mer* composition are also used to improve the capability to retrieve high-quality genome bins for NGS datasets [24]. However, there are limitations for these approaches. The high sequencing error rates in TGS long read hamper distinction between true variants and sequencing errors. An effective contact map with Hi-C library can only be established for a draft genome with preferable contiguity. Constructing co-abundance in multiple samples may ignore the genome characteristics of a single sample and also increase the sequencing cost.

The co-barcoding sequencing library [25-29], an improved short-read sequencing with long-range genomic information, can provide an alternative way to analyze metagenomes accurately and quantitatively. For different co-barcoding libraries such as BGI's stLFR library[28], 10X Genomics' linked-reads library[30] and Illumina's contiguity preserving transposase sequencing library [26], the differences in the total barcode number and the short-read coverage of high weight DNA molecules (HWMs) have a great impact on their powers in the downstream analysis [31-34]. The co-barcoding correlation between assembled draft sequences and barcode distribution on the assembled graph have been successfully applied to both single genome [35-37] and metagenome assembly [27, 38, 39]. However, the inherent complexity arising from long repeat sequences and uneven abundance among different species is still unsolved to construct draft sequences or assembly graphs in current strategies, making them unstable or difficult to generate enough high-quality genome drafts for complex microbial communities.

In this work, we introduced a pipeline named MetaTrass to obtain high-quality genome drafts using references of microbiome from public databases and co-barcoding information from stLFR read sets. In our strategy, the co-barcoding information is used not only to improve the assemblies by implementing co-barcoding assemblies, but also to simplify the dataset before assembling using references with the help of taxonomic binning. We apply MetaTrass to stLFR datasets of a mock metagenome community and four real gut microbiome communities to evaluated its capability of producing high-quality genome drafts with high contiguity and high taxonomic resolution, with comparing to the common combination strategies. In addition, the draft genomes with taxonomic information at species resolution obtained in taxonomic binning would be convenient to make further use of these assemblies.

Datasets and methods

Datasets

A mock microbial and four gut microbial communities were analyzed to evaluate the efficiency of MetaTrass. The mock microbial community (ZymoBIOMICS^{TM} Microbial Community DNA Standard) consists of 8 isolated bacteria with the abundance of about 12% and 2 fungi with the abundance of about 2%. The four gut microbial DNA samples include three faeces from healthy volunteers and one faeces from a patient volunteer with inflammatory bowel disease. The stLFR libraries were constructed according to the standard protocol [28]. The DNA samples were firstly sheared into long fragments, and then the long fragments were captured into a magnetic microbead with a unique barcode sequence. Finally, each long fragment was broken and hybridized with the unique barcode by the Tn5 transposase on the surface of the microbead. The stLFR libraries of the mock and the patient sample were sequenced on BGISEQ500 platform, and those of healthy samples were sequenced on MGISEQ2000 platform. The read length in the read pair was 100 b for all datasets. The mock and the three healthy sample libraries were allocated to a half lane individually, and a total of about 50 Gb raw reads were generated. The faeces library of the patient was allocated to a full lane, and about 100 Gb raw reads were generated. The barcodes were extracted from the end of the second read in a read pair and then replaced by their numerical symbols in the read names in the fastq file with an in-house script. SOAPfilter_v2.2 with parameters (*-y -F CTGTCTCTTATACACATCTTAGGAAGACAAGCACTGACGACATGA -R TCTGCTGAGTCGAGAACGTCTCTGTGAGCCAAGGAGTTGCTCTGG -p -M 2 -f -1 -Q 10*) was used to clean the low-quality raw reads with adaptors, excessive confused bases, and high duplications. Finally, 55.65 Gb clean data were retained for the mock microbiome, 34.48 Gb for the first healthy sample (H_Gut_Meta01), 35.33 Gb for the second one (H_Gut_Meta02), 37.88 Gb for the third one (H_Gut_Meta03), and 97.20 Gb for the patient sample (P_Gut_Meta01).

Taxonomic binning

We adopted Kraken2 (version 2.0.9-beta) [40] to classify stLFR reads into different species. Firstly, different customized databases were constructed to analyze the mock microbial community and gut microbial communities, respectively. Then the corresponding stLFR reads were classified with default parameters. The references attached to the ZYMO product were used to construct the taxonomic database of the mock sample. The Kraken2 database of the Unified Human Gastrointestinal Genomes (UHGG) collection [3] was used to study the gut samples, and totally 4542 representative genomes at the species level were included.

Co-barcoding reads refining

Since a taxonomic tree of references was constructed to reduce the number of multiple hits of a *K-mer* from repeat sequences of different species in Kraken2, many reads were classified into the lowest common ancient (LCA) rank higher than its responding species. Some works try to reallocated these reads to species by statistical inferences using the coverage depth of unique region of a species or co-barcoding information [41, 42]. In MetaTrass pipeline, the co-barcoding correlation between taxonomic reads of a species and reads classified into high LCA rank were used to retrieve datasets classified into high LCA rank to the species. Since sufficient read coverage is required for a complete genome assembling, only the read sets of one species with an abundance higher than 10× were refined by co-barcoding information. The abundance of each species was roughly calculated according to the taxonomic read coverage on the reference. Meanwhile, we set a data size threshold of the refined read set to reduce the computational consumption for species with an extreme high abundance (e.g. 300×). In the co-barcoding reads refining, reads were collected according to the barcode properties including the number of reads in the taxonomic read set (Num_T) and the ratio of these reads to the total reads (Ratio_T). All barcodes were firstly extracted from the taxonomic read set as candidates. Then, the candidates ranked first by Num_T from largest to smallest and then by Ratio_T for those with the same Num_T. Finally, the reads with barcode of Ratio_T larger than 0.1 were chose based on the barcode rank. Paired-end reads were extracted by Seqtk (version 1.3-r114-dirty) according to the barcode-related read names from the fastq file of clean reads. Note that although Ratio_T was set to reduce obviously false positive reads caused by the collision of long fragments from different species in the same microbead, there were still some false positive reads. Sequences assembled by these reads would be filtered as following description in section 2.5.

Co-barcoding reads assembling

In our pipeline, the read set of a single-species with abundance larger than 10× was assembled by Supernova (version 2.1.1), which is a co-barcoding de novo assembler for single large eukaryotic genomes with high performances. Supernova is designed for linked-reads of 10X Genomics, which have different barcode sequences and formats from stLFR reads. Thus, the stLFR reads were converted into linked-reads fastq files with an in-house script. Additionally, the parameter --*accept-extreme-coverage* was set as *yes* to adapt to different coverage depths.

Sequences purifying

With the increasing number of microbial genomes, the similarity between whole genomes including alignment fraction (AF) and average nucleotide identity (ANI) have been adopted to circumscribe species [3, 4]. Similar parameters of AF and ANI between assembled contigs and the reference were used to purify the sequences assembled by the refined co-barcoding read sets. ANI was calculated independently for each alignment. AF was defined as the ratio of total length of alignments with ANI larger than a threshold to the total length of the contig. In our pipeline, we set ANI threshold as 90%, and AF threshold as 50%. The alignments between contigs and references were generated by QUAST (version 5.0.2) [43] with default parameters, except the identity threshold to obtain valid alignment was set as 90%.

Combination strategies of genome assembling and binning

In a standard analysis of NGS metagenomic dataset, the combination of de novo assembling with binning was adopted to get genomes for different species. This strategy can also be applied to the analysis of co-barcoding dataset. We compared different combination strategies to MetaTrass by analyzing the mock and four gut samples. In our test, the stLFR co-barcoding reads were assembled by NGS assemblers including IDBA-UD (version 1.1.3), MEGAHIT (version 1.1.3), and MetaSPAdes (version 3.10.1) or co-barcoding assemblers including Supernova [35], Athena (version 1.3.0) [27], and CloudSPAdes (version 3.13.1) [38]. Then, all these draft assemblies were binned by two genome binners, MetaBAT2 (version 2.12.1)[19] and Maxbin2.0 (version 2.2.5) [18]. Since CloudSPAdes and Athena were also not designed for stLFR reads, we made an appropriate format conversion before assembling with an in-house script where Longranger (version 2.2.2) [44] was used. In genome assembling, Supernova was run with parameters as those have been adopted in MetaTrass, IDBA-UD, MEGAHIT, MetaSPAdes, Athena and CloudSPAdes were run with default parameters. All the assembling results were deposited into CNGB Sequence Archive (CNSA) [45] (https://db.cngb.org/cnsa/) of China National GeneBank DataBase (CNGBdb) [46] with accession number CNP0002163. In genome binning, MetaBAT2 and Maxbin2.0 were run with default parameters.

Evaluations

Both reference-based and reference-free assessments were used to evaluate the quality of assemblies obtained by different strategies. For the mock microbial community with definite references, the reference-based tool QUAST was used to evaluate contiguity and accuracy of metagenomic assemblies. In QUAST assessments, minimap2 was used to map assemblies to references and get valid alignments with the identity threshold of 95%. Then, the statistics such as genome fraction, NG50/NGA50, and number of misassemblies were assessed from the alignments with default parameters. For the real gut microbial communities, the reference-free tool CheckM (version 1.1.2) [47] were run with default parameters to evaluate completeness and contamination of each genome from metagenomic assemblies in addition to the QUAST. Following the guidance proposed in CheckM, we defined a high-quality assembly if it has >90% completeness and <5% contamination and a medium-quality assembly if it has >50% completeness and <10% contamination and not meets a high-quality criterion. In addition, the statistics of each genome such as N50, genome size, and taxonomic rank were also obtained by CheckM, where the taxonomic rank was used to demonstrate the resolution of a genome bin.

Variant and phylogenetic analysis

All the high-quality genomes assembled by MetaTrass were used to call variants for the four gut samples. We aligned each genome to the corresponding reference using minimap2 (2.17-r974-dirty) with parameters (*-x asm5*) to prevent an alignment extending to regions with diversity >5%. Samtools (version 1.9) [48] and Paftools were used to convert the bam file of initial unsorted alignments into a paf file of sorted alignments. We identified variants using the "call" module in Paftools with parameters (*-L 10000*) to filter the alignments shorter than 10,000 bp. SNVs only referred to substitutions, did not include single-base insertions or deletions. Insertions or deletions with length shorter than 50 bp were defined as small indels, and the others were large indels. The position and sequence information of a variant between different genomes were compared to determine whether a variant was shared by species in different samples.

We used the "classify_wf" function of GTDB-tk (version 0.3.1)[49] to conduct taxonomic annotation of the genome bins obtained by the common strategies with default parameters. Considering the procedure of UHGG database construction[4], genome bins were assigned at the species level if the AF to the closed species representative genomes was higher than 30% and ANI was higher than 95%. We used FastTree (version 2.1.10) [50] to build maximum-likelihood phylogenetic trees of the high-quality genomes assembled by MetaTrass. The input of protein sequence alignments was produced by GTDB-Tk using marker gene set of 120 bacterial and 122 archaeal. Interactive Tree of Life (iTOL version 4.4.2) [51] was used to visualize and annotate trees.

Results and discussions

MetaTrass pipeline

In this work, we developed an assembling pipeline named MetaTrass to combine the references and long-range co-barcoding information of stLFR library (Figure 1). In a co-barcoding library construction, the HWMs sheared from DNA samples are firstly distributed into different isolated partitions, and then short-read fragments from the HWM in the same partition are labeled with a unique barcode sequence, finally the co-barcoded fragments are sequenced by standard short-read sequencing platforms (Figure 1a-d). In stLFR libraries, the large number of barcodes and the low collision rate of HWMs in one microbead was distinguishing character of stLFR from other co-barcoding libraries [28]. In our pipeline, the references corresponding to a sample were firstly used to build a taxonomic database by Kraken2 [40]. Using the taxonomic database, the metagenomic reads were classified into different taxonomic read sets for each species. Since the phylogenetic relations among references were used in Kraken2, reads from repeat regions would be classified into a higher rank as false negative (Figure 1a). About 10% reads were classified into ranks higher than the species for the four human gut datasets (Table S1). The co-barcoding correlation between the taxonomic read set and the false negative read set were used to refine the final set for a target species. Reads with the barcodes appeared in the taxonomic reads were extracted according to barcode properties for each species. The barcodes containing more reads classified into the species are more likely to retain the long-range genomic information, so a paired-end read with a barcode with more reads and a higher ratio of taxonomic reads were prior to be chose. The refined read sets of each species were independently assembled by Supernova. Several long fragments from different species genomes shared the same barcode in real stLFR libraries (Figure S1), thus involving some false positive reads from non-target species in the assembly process. Finally, the initial assembly was purified according to the AF and ANI values of alignments between the assembly and references. The comprehensive use of co-barcoding information and references in our approach could reduce the false negative effects of taxonomic binning and the false positive effects of co-barcoding read refining.

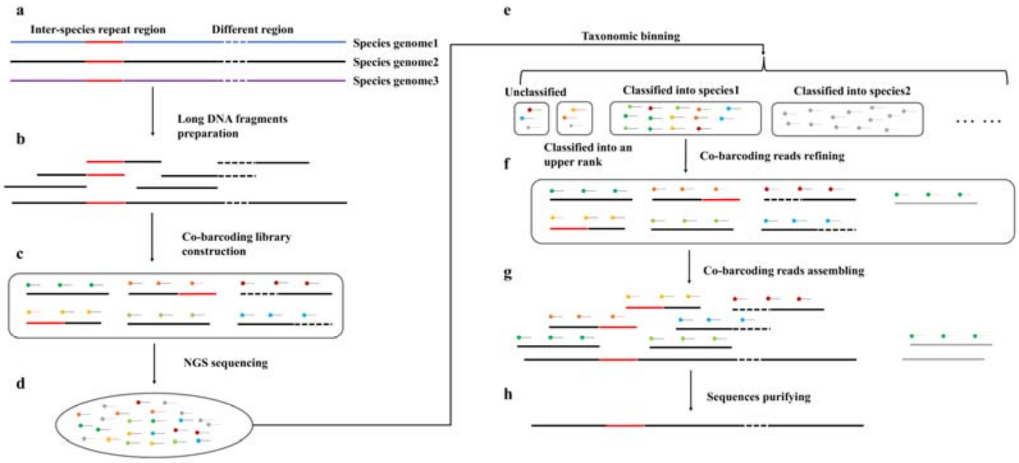

Figure 1. Scheme of co-barcoding library and MetaTrass pipeline. a-d) The co-barcoding correlation between unique region of a species and repeats (marked as red color) or differences (marked as dash line) in co-barcoding libraries. e-h) The subprocesses in MetaTrass include taxonomic binning, co-barcoding reads refining, co-barcoding reads assembling and sequences purifying.

Assembly of the mock microbiome

The strategy, where the reads with long range information were firstly binned and then assembled, have been widely adopted to assemble haplotype genomes for eukaryotes with large sizes [32, 52]. But it has been rarely used to assemble metagenomes. We firstly applied MetaTrass to assemble stLFR read sets of the mock microbial community. Totally, up to 99.4% percent of reads the data were assigned into different datasets of species (Table S2). For testing the efficiency of our strategy, we compared it with the common mixed assembling strategies (Figure 2a). Besides the MetaTrass, the stLFR reads were also mixed assembled by the NGS and co-barcoding assemblers including IDBA-UD, MegaHit, Supernova, CloudSPAdes, and Athena. Additionally, the optimal mixed assemblies of ONT reads and Illumina NGS reads in Nicholls's work [53] were also used to make a comparison, where the ONT result was assembled by WTDBG and the NGS result was assembled by SPAdes. All the assembling results for each species were evaluated by QUAST to assess genome fraction, contiguity and accuracy. The draft genome of each species in a mixed assembly was extracted with our purifying method.

Overall, our pipeline was superior in the production of accurate genomes with high genome fractions and long contiguity (Figure 2). As a co-barcoding assembler designed for a single genome, Supernova incompletely assembled two species *Enterococcus faecalis* and *Lactobacillus fermentum* with low genome fractions 17.7% and 8.9% in the mixed assembly. However, both species were properly recovered in MetaTrass, indicating that the assembling complexity caused by uneven abundances was reduced by taxonomic binning. All the assemblies by MetaTrass showed high genome integrity as those by NGS and co-barcoding assemblers designed for metagenome, which were higher than those of ONT assemblies. Compared to NGS assemblers, the assembler using long range information from co-barcoding datasets obtained draft assemblies with significantly better contiguity. As a result, there were no genomes with NG50 larger than 1 Mb in the assemblies by NGS assemblers, while many long-contiguity assemblies were built by co-barcoding and TGS long reads assemblers. MetaTrass generated seven draft genomes with NG50 around 2 Mb, which was the maximum for all. Furthermore, the accuracy was also guaranteed by MetaTrass, which obtained the most assemblies with NGA50 around 2 Mb. Compared to ONT assemblies, assemblies by MetaTrass had less errors, close to those of NGS assemblies (Figure S2). The average mismatch number and indel number per 100 kb in assemblies with NGS reads were about 60 and 10. Both of them were much lower than 171 and 267 in the ONT assemblies.

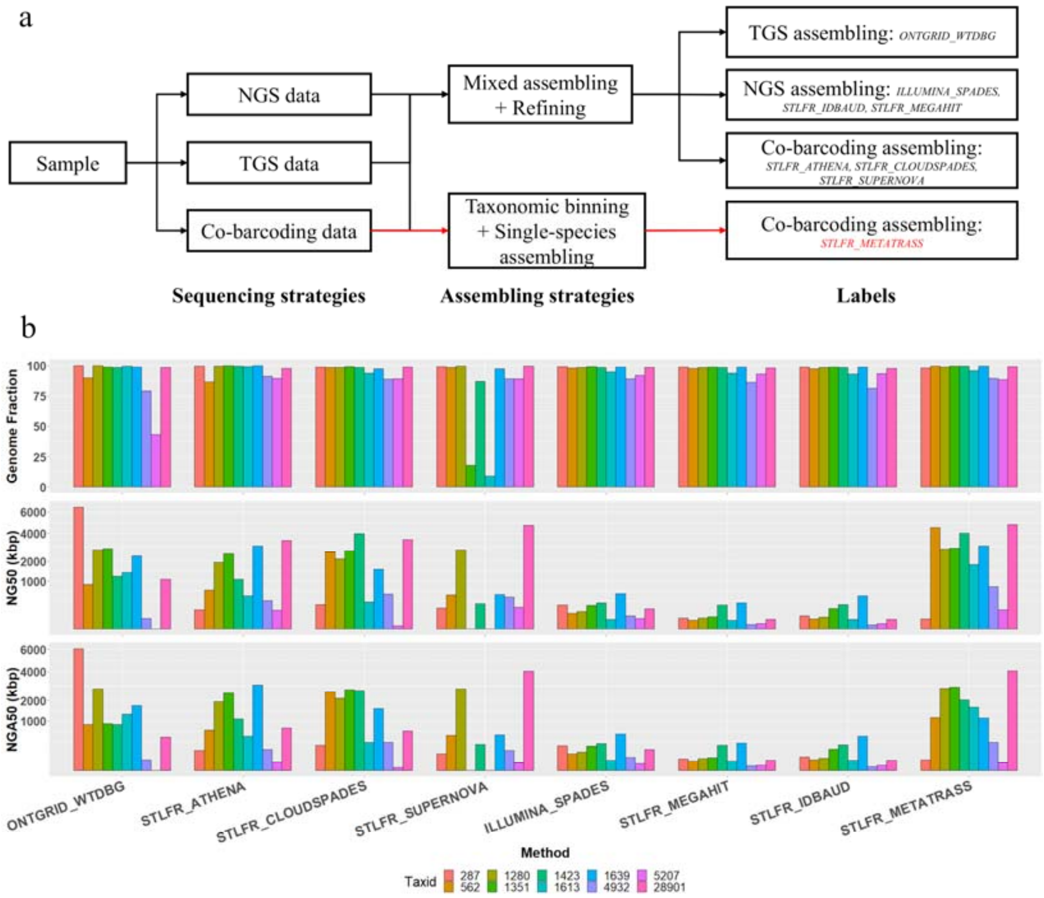

Figure 2. a) All assemblies with different sequencing and assembling strategies. b) Genome fraction, NG50 and NGA50 evaluated by QUAST for assemblies of different reads assembled by different assemblers.

Assembly of four human gut microbiome

To evaluate the robustness of our approach to samples from natural microbial community, we applied MetaTrass to stLFR reads of four human faeces samples. The comprehensive genome references of UHGG were used to classify NGS reads by Kraken2, and the community compositions were estimated by the taxonomic reads at different taxonomy ranks (Figure S3-S6). Based on the taxonomic reads at phylum assigned by Kraken (Figure S3), the three healthy samples had a similar microbial community, where the major microbiomes were from *Firmicutes A* phylum. This was different from the patient sample where *Proteobacteria* dominated. It was consistent with the previous observation that a bloom of *Proteobacteria* is strongly correlated with the enteric diseases caused by dysbiosis in gut microbiota [54]. The abundance at species level was assessed with a simple coverage-approximation method to identify the species assembled in following steps. The total numbers of species with an abundance higher than 10× were 113, 108, 93, and 158 in H_Gut_Meta01, H_Gut_Meta02, H_Gut_Meta03, and P_Gut_Meta01 samples, respectively.

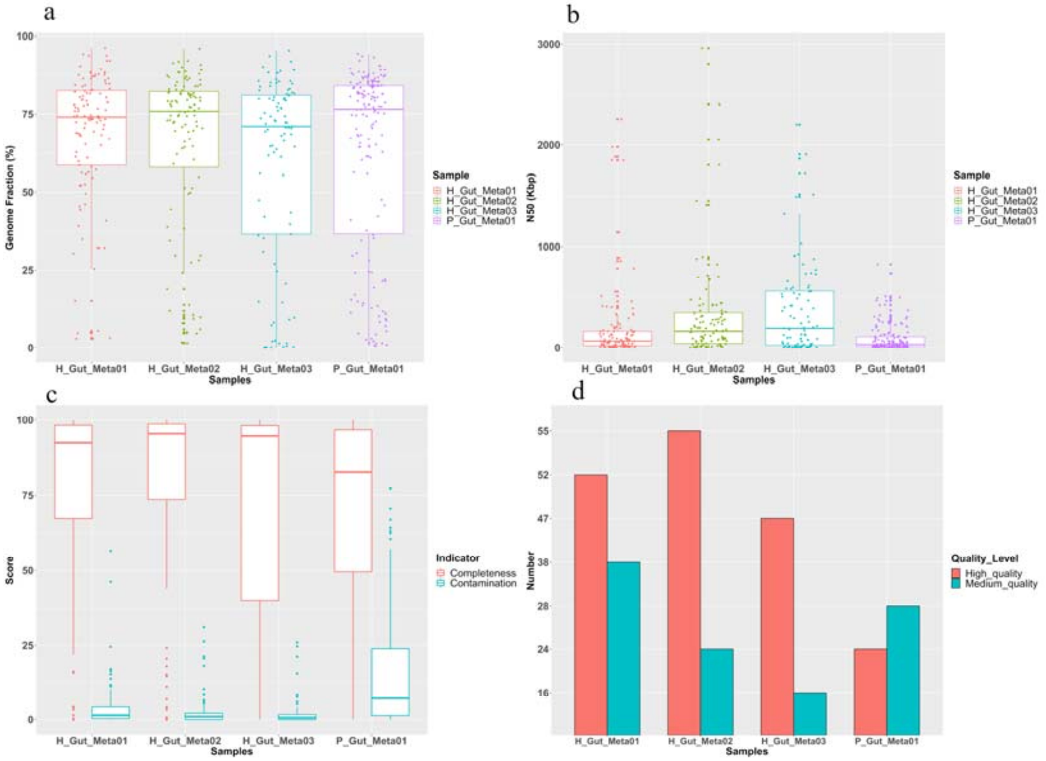

Figure 3. The QUAST and CheckM evaluations of all assemblies by MetaTrass for the four human gut samples. a) Genome fraction. b) Scaffold N50. c) Box plot of completeness and contamination. d) Number of high- and medium-quality genomes.

The genome fraction of an assembly to the reference was usually used to evaluate the completeness in single genome assembly. The genome fraction evaluated by QUAST for all samples widely range from 0% to 90%, and the distributions of H_Gut_Meta01 and H_Gut_Meta02 were more concentrated than those of H_Gut_Meta03 and P_Gut_Meta01 (Figure 3a). However, more than half of the assembled genomes were with a genome fraction higher than 50 for all the samples. Considering the large genetic diversity between sample genomes and the references

[7], these results indicated that our pipeline was able to assemble complete genomes for a large proportion of species with an abundance larger than 10×. The genetic diversity was also proved by the significant differences in genome fraction and the ratio of assembled length to the reference length among all the four samples (Figure S7). The distributions of genomes N50 were generally dispersed, and the medians of H_Gut_Meta02 and H_Gut_Meta03 were obviously higher than those of H_Gut_Meta01 and P_Gut_Meta01 (Figure 3b). Nevertheless, the third quartiles in the box plots for all the samples were larger than 100 kb, demonstrating that our pipeline had a strong capability to generate draft genome with high contiguity. Note that for these three healthy samples a plenty of ultra-long genome drafts (N50>1 Mb) were obtained, which may provide possibilities to study the large genome difference in the microbiome.

Considering the intra-species genetic diversity, we also evaluated the quality of metagenomic assemblies based on the conserved marker genes by CheckM. The completeness medians of three healthy samples were larger than 92, and the contamination medians were smaller than 2 (Figure 3c). However, the completeness of the patient sample was about 83, and the contamination median was about 7 (Figure S8). Meanwhile, a great number of the high- and medium-quality genomes were assembled by MetaTrass for all the four samples (Figure 3d). 52 high-quality and 37 medium-quality genomes were produced for H_Gut_Meta01, and 55 and 24 for H_Gut_Meta02, and 47 and 16 for H_Gut_Meta03, respectively. These numbers of high-quality genomes were obviously larger than that of the patient sample, where 24 high- and 28 medium-quality genomes was assembled.

Comparison to common combination strategies

To further evaluate our approach's efficiency, we compared it with common combination strategies of assembling tools and genome binning tools as listed in the section 2.6 by analyzing the four gut datasets. It should be noted that currently there are still no genome binning tools to directly exploit the co-barcoding information. By counting the number of bins with completeness larger than 50 in the bin with at least one conserved marker genes (Table S3), we observed that MetaBAT2 clustered different draft assemblies into more bins than Maxbin2.0 for all the samples. Maxbin2.0 preferred to produce more bins with completeness higher than 50% than MetaBAT2, while MetaTrass outperform both. Especially for P_Gut_Meta01, the optimal combination between Supernova and Maxbin2.0 obtained 66 bins with completeness higher than 50 which was significantly less than 117 obtained by MetaTrass.

By comprehensively analyzing the completeness, contamination and taxonomy rank of each bin, we assessed MetaTrass and common strategies in the ability to get high- and medium-quality genomes and resolution of taxonomy rank (Figure 4). For different samples, the best combination to produce the optimal results was different. The optimal combinations were MetaSPAdes and Maxbin2.0, Supernova and MetaBAT2, MetaSPAdes and MetaBAT2, and Athena and MetaBAT2 for H_Gut_Meta01, H_Gut_Meta02, H_Gut_Meta03, and P_Gut_Meta01, respectively. For all the four samples, the optimal results of the common strategies were still inferior those of MetaTrass. For the example of H_Gut_meta01, the combination of MetaSPAdes and Maxbin2.0 produced a total of 41 high- and medium-quality genomes which was significantly less than 90 obtained by MetaTrass. There were only 3 genomes out of total 18 high-quality genome with a taxonomic rank lower than order rank, but 15 out of 52 for MetaTrass. Comparing the strategies only using NGS read information, the combination strategies of co-barcoding assembler and binner showed no obvious advantages in generating genomes with high quality and resolution, but MetaTrass was significantly superior to them. These results demonstrated that the usage of co-barcoding information in MetaTrass was more efficient and accurate than those in a mixed assembling.

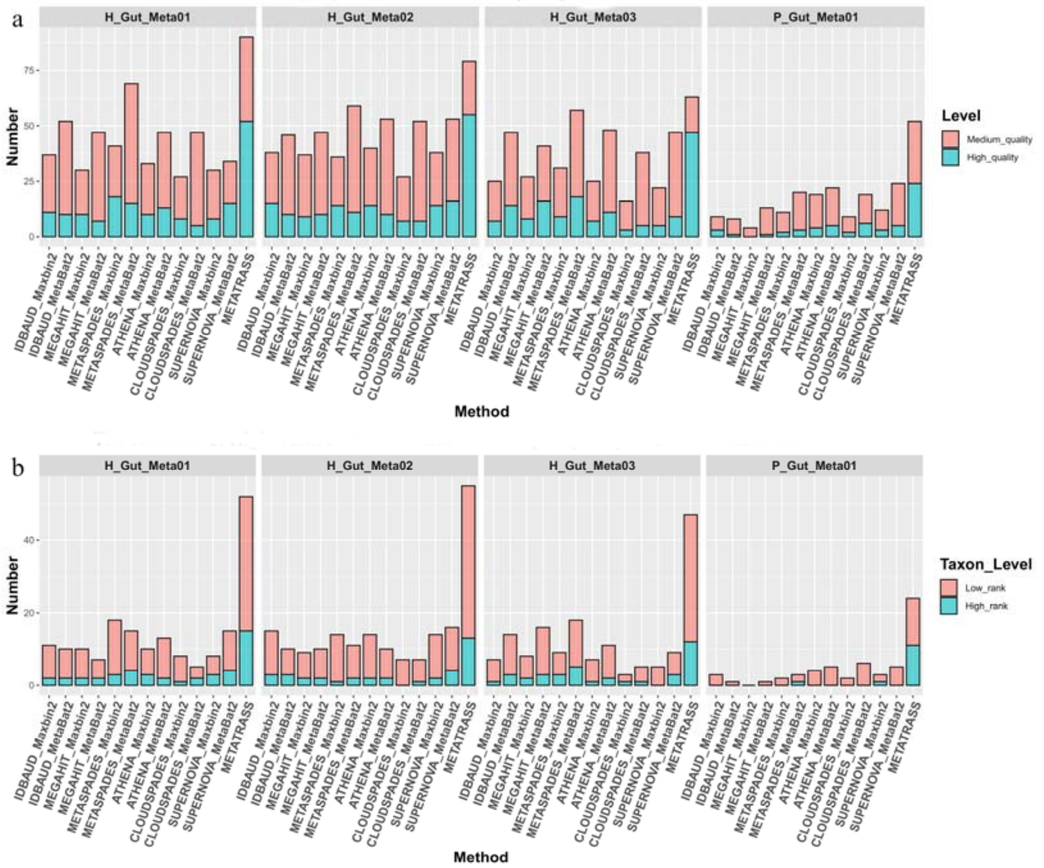

Figure4. a) Number of high- and medium-quality genomes assembled in different methods. b) Number of high-quality genomes with high- and low-rank in different methods.

The human gut microbiome composition attracts much attention due to its strong correlation with personality traits [55]. To compare the microbiome composition structures of the high-quality genomes with different methods, we uniformly classified the high-quality genome bins into species using GTDB-tk. Using the large number of high-quality genomes obtained by MetaTrass, the phylogenetic trees of these genomes were constructed with corresponding N50 information for all the samples. Meanwhile, the high-quality genome bins by common strategies appeared in MetaTrass results were marked in red color in the middle heat map. The structure of the phylogenetic tree of genomes assembled by MetaTrass could give a more comprehensive insight of the composition structure than those of the common strategies. From the different trees in Figure 5 and Figure S9-S11, the numbers of orders with high-quality genomes assembled by MetaTrass were 9, 11, 7, and 7 for H_Gut_Meta01, H_Gut_Meta02, H_Gut_Meta03, and P_Gut_Meta01, respectively. Notably, we did obtain some orders with more than 5 high-quality genomes, which might provide convenience to study the microbiome structure at the genome-wide scale. For the sample of H_Gut_Meta01 (Figure 5), there were 27 high-quality genomes classified into *Lachnospirales* order and 14 genomes into *Oscillospirales* order. These two orders were exactly the dominating orders according to the taxonomic abundance distribution. Similar results were obtained for the other two healthy samples (Figure S9 and S10), indicating that the microbiome with higher sequencing coverage could be better assembled in MetaTrass. For P_Gut_Meta01, the orders with more than 5 high-quality genomes were *Enterobacterales* and *Actinomycetales* (Figure S11). The obvious difference between the healthy and patient samples was consistent with the different microbiome compositions observed in the taxonomic binning. For all the common strategies, most of their high-quality genomes were also successfully assembled by MetaTrass. For H_Gut_Meta01 (Figure 5), a total of 137 genome bins were assembled by different combination strategies, while only 25 of them were not included in the results of MetaTrass. From the heat maps, most of the common strategies assembled draft genomes for each order, but the total numbers in each order were small. The maximal number of genomes in one order was 6 and obtained by the combination of Supernova and MetaBAT2 for *Lachnospirales*. The species in the same order obtained by different common strategies were different and randomly distributed in the phylogenetic tree. Moreover, 146 of 179 high-quality genomes were with N50 larger than 100 kb, demonstrating that MetaTrass had a strong ability to produce assemblies with long contiguity using co-barcoding information.

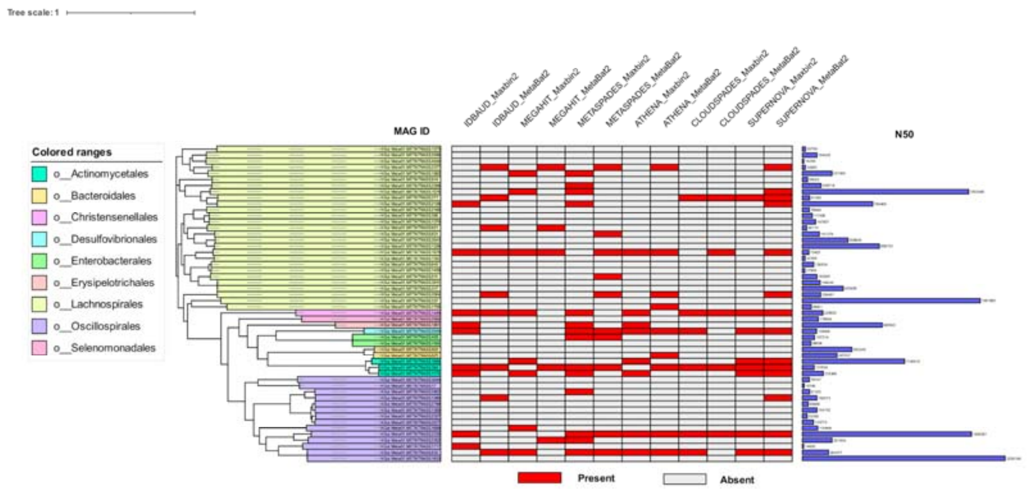

Figure 5. Phylogenetic tree of the high-quality genomes assembled by MetaTrass for H_Gut_Meta01. Distribution of the high-quality genomes assembled by other methods were colored as red in the middle heat map. N50 of each high-quality genome was shown in the left part.

Genetic diversity in different samples

Different types of variants in gut microbiomes are strongly associated with host health, and related researches focused on the genetic origin of phenotypic difference among people of different regions or health status [56, 57]. By aligning draft genomes to the references, we called variants for high-quality genomes for each species in different samples, including single nucleotide variants (SNV), small indels, and large indels. For different types of variants, the numbers of SNV were significantly larger than those of the small and large indels for all the four samples (Figure S12). The numbers of different variants were close for the three healthy samples, but obviously larger than those of the patient sample. This might come from fewer alignments for the patient sample. To remove the effects of the total aligned length in comparisons, the SNV density was also calculated. The SNV density of the patient sample was obviously higher than those of healthy samples (Figure S12d). The SNV density median was about 21, but those was about 9 for the healthy samples. The difference of SNV density could exist in samples from different continents reported in UHGG database [4], or from different physiological status. Meanwhile, the distribution of

SNV densities was diffuse for all the samples, indicating that the contribution to the genomic diversity were different for different species in a microbial community.

For the taxonomic information of high-quality genomes from different samples, we found 15 species shared by three samples, where14 species appeared in the three healthy sample but only one species of *Escherichia* appeared in the patient sample and two healthy samples. This facilitated our investigation of the genetic diversity between species from different samples. The SNV density in different samples and intersection of variants between different samples for each species in three healthy sample were analyzed. Different species showed different SNV densities, but the SNV density of the share species demonstrated few differences for almost all cases (Figure 6a). From Figure 6b to 6d, the number of unique and shared variants in different types significantly fluctuated for different species, but their difference among samples showed great consistency. The number of unique variants of H_Gut_Meta03 was obviously larger than that of H_Gut_Meta01 or H_Gut_Meta02. The numbers of shared variants between H_Gut_Meta03 and other two samples were smaller than those between H_Gut_Meta01 and H_Gut_Meta02 for all the species. H_Gut_Meta03 shared close number of variants with H_Gut_Meta01 and H_Gut_Meta02. H_Gut_Meta01 and H_Gut_meta02 shared 94498 same SNVs, which was about 10 times of shared SNVs between H_Gut_Meta01 and H_Gut_Meta03. The huge genetic diversity between H_Gut_Meta03 and other samples might come from the difference in host metabolic state. Notably, the ratio of large indels shared by all three samples to the total number was much smaller than those of SNVs and small indels. This result indicated that large variants are more specific than small variants, which could be potentially used in the study of their correlations with host health [56].

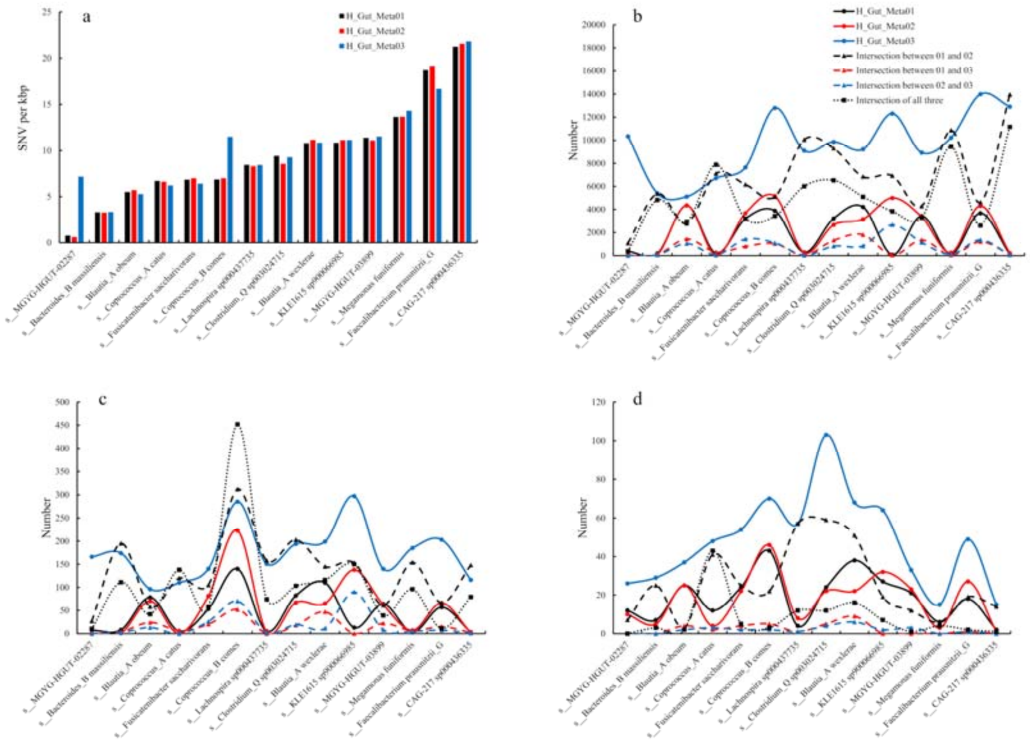

Figure 6. SNV density and number of unique and shared variants for each species appearing in all three healthy samples. a is the SNV density. b, c and d were the number of SNVs, small and large indels, respectively.

Conclusions

Co-barcoding sequencing reads have shown its great potential in de novo genome assembly, but the complexity caused by uneven abundance and inter-species repeats in metagenomic assembling makes it unstable and inefficient. In this work, we developed a tool to get high-quality genomes with high taxonomic resolutions by combining the co-barcoding information with public references. Compared with the common combination strategies, our pipeline generated more high-quality genomes for microbiome datasets with different complexities. Meanwhile, plenty of draft genomes were also assembled with NG50 larger than 1 Mb, and some of which were even longer than that of the references for both mock and human gut datasets. For all the four real gut datasets, 178 draft genomes with high completeness and low contamination were generated according to the evaluation by CheckM, but their genome fractions relative to the references were low. This indicated that there existed significant differences between the sample genomes and the reference genomes.

In MetaTrass, the co-barcoding information was also used to reduce the false negative reads in taxonomic binning by refining, not only for assembling. Using the co-barcoding correlation between classified reads at the species level and others, we could retrieve the datasets in repetitive regions. For the patient sample, the number of high-quality genomes with long contiguity assembled by MetaTrass was significantly larger than that without co-barcoding reads refining (Figure S13). Thus, the co-barcoding information should have the potential to extract reads from regions other than the references.

The efficiency of our pipeline was dependent on the co-barcoding information quality of the stLFR dataset including the read coverage and length of an HWM. The read coverage of refined read set was higher than that of taxonomic read set, but still lower than that of all aligned reads (Table S4). This result indicated that there were still some false negative reads without correlations due to the low coverage or short length. On the other hand, when we used co-barcoding information to reduce false negative, some false positive reads were introduced at the same time. The results of the purifying module demonstrated that these sequences could be filtered effectively, but would increase the computational consumption in the co-barcoding genome assembly. Thus, improvements on co-barcoding library and more deliciated positive reducing algorithm would enhance the performance of MetaTrass.

With an increasing number of studies on the correlation between dramatic phenotype and variants in the non-conserved genome region of species, the high-quality genomes from metagenomic sequencing were strongly required to completely understand the genetic diversity of microbial community at high resolutions. The application of MetaTrass in human gut samples showed promise of generating high-quality genomes for real complex microbial community at a high resolution. With the increasing number of microbial reference genomes and the development of co-barcoding sequencing library, the assembling strategy in MetaTrass based on the combination of co-barcoding sequencing library and references will be extended to more metagenomic studies of different microbial communities.

Acknowledgements

This research was supported by the National Key Research and Development Program of China (Grant No. 2018YFD0900301-05).

References

1. Schloss PD, Handelsman J. **Metagenomics for studying unculturable microorganisms: cutting the Gordian knot**. *Genome Biol.* 2005; **6**(8):1-4.

2. Qin J, Li R, Raes J, Arumugam M, Burgdorf KS, Manichanh C, Nielsen T, Pons N, Levenez F, Yamada T *et al.* **A human gut microbial gene catalogue established by metagenomic sequencing**. *Nature.* 2010; **464**(7285):59-65.

3. Parks DH, Chuvochina M, Chaumeil P-A, Rinke C, Mussig AJ, Hugenholtz P. **A complete domain-to-species taxonomy for Bacteria and Archaea**. *Nat Biotechnol.* 2020; **38**(9):1079-1086.

4. Almeida A, Nayfach S, Boland M, Strozzi F, Beracochea M, Shi ZJ, Pollard KS, Sakharova E, Parks DH, Hugenholtz P *et al.* **A unified catalog of 204,938 reference genomes from the human gut microbiome**. *Nat Biotechnol.* 2021; **39**(1):105-114.

5. Sheth RU, Li M, Jiang W, Sims PA, Leong KW, Wang HH. **Spatial metagenomic characterization of microbial biogeography in the gut**. *Nat Biotechnol.* 2019; **37**(8):877-883.

6. Martino C, Shenhav L, Marotz CA, Armstrong G, McDonald D, Vázquez-Baeza Y, Morton JT, Jiang L, Dominguez-Bello MG, Swafford AD *et al.* **Context-aware dimensionality reduction deconvolutes gut microbial community dynamics**. *Nat Biotechnol.* 2021; **39**(2):165-168.

7. Van Rossum T, Ferretti P, Maistrenko OM, Bork P. **Diversity within species: interpreting strains in microbiomes**. *Nat Rev Microbiol.* 2020; **18**(9):491-506.

8. Olm MR, Crits-Christoph A, Bouma-Gregson K, Firek BA, Morowitz MJ, Banfield JF.

inStrain profiles population microdiversity from metagenomic data and sensitively detects shared microbial strains. *Nat Biotechnol.* 2021; **39**(6):727-736.

9. Leimbach A, Hacker J, Dobrindt U. **E. coli as an all-rounder: the thin line between commensalism and pathogenicity**. *Curr Top Microbiol Immunol.* 2013; **358**:3-32.

10. Pierce JV, Bernstein HD. **Genomic diversity of enterotoxigenic strains of Bacteroides fragilis**. *PLoS One.* 2016; **11**(6):e0158171.

11. Welch RA, Burland V, Plunkett G, Redford P, Roesch P, Rasko D, Buckles EL, Liou S-R, Boutin A, Hackett J *et al*. **Extensive mosaic structure revealed by the complete genome sequence of uropathogenic Escherichia coli**. *Proc Natl Acad Sci.* 2002; **99**(26):17020-17024.

12. Sczyrba A, Hofmann P, Belmann P, Koslicki D, Janssen S, Dröge J, Gregor I, Majda S, Fiedler J, Dahms E *et al*. **Critical assessment of metagenome interpretation—a benchmark of metagenomics software**. *Nat Methods.* 2017; **14**(11):1063-1071.

13. Breitwieser FP, Lu J, Salzberg SL. **A review of methods and databases for metagenomic classification and assembly**. *Brief Bioinformatics.* 2019; **20**(4):1125-1136.

14. 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.* 2015; **31**(10):1674-1676.

15. Nurk S, Meleshko D, Korobeynikov A, Pevzner PA. **metaSPAdes: a new versatile metagenomic assembler**. *Genome Res.* 2017; **27**(5):824-834.

16. Peng Y, Leung HC, Yiu S-M, Chin FY. **IDBA-UD: a de novo assembler for single-cell**

and metagenomic sequencing data with highly uneven depth. *Bioinformatics.* 2012; **28**(11):1420-1428.

17. Wu Y-W, Ye Y. **A novel abundance-based algorithm for binning metagenomic sequences using l-tuples**. *J Comput Biol.* 2011; **18**(3):523-534.

18. Wu Y-W, Simmons BA, Singer SW. **MaxBin 2.0: an automated binning algorithm to recover genomes from multiple metagenomic datasets**. *Bioinformatics.* 2016; **32**(4):605-607.

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

20. Bertrand D, Shaw J, Kalathiyappan M, Ng AHQ, Kumar MS, Li C, Dvornicic M, Soldo JP, Koh JY, Tong C *et al.* **Hybrid metagenomic assembly enables high-resolution analysis of resistance determinants and mobile elements in human microbiomes**. *Nat Biotechnol.* 2019; **37**(8):937-944.

21. Chin C-S, Alexander DH, Marks P, Klammer AA, Drake J, Heiner C, Clum A, Copeland A, Huddleston J, Eichler EE *et al.* **Nonhybrid, finished microbial genome assemblies from long-read SMRT sequencing data**. *Nat Methods.* 2013; **10**(6):563.

22. Kolmogorov M, Bickhart DM, Behsaz B, Gurevich A, Rayko M, Shin SB, Kuhn K, Yuan J, Polevikov E, Smith TPL *et al.* **metaFlye: scalable long-read metagenome assembly using repeat graphs**. *Nat Methods.* 2020; **17**(11):1103-1110.

23. DeMaere MZ, Darling AE. **bin3C: exploiting Hi-C sequencing data to accurately resolve metagenome-assembled genomes**. *Genome Biol.* 2019; **20**(1):1-16.

24. Cleary B, Brito IL, Huang K, Gevers D, Shea T, Young S, Alm EJ. **Detection of low-abundance bacterial strains in metagenomic datasets by eigengenome partitioning**. *Nat Biotechnol.* 2015; **33**(10):1053-1060.

25. Peters BA, Kermani BG, Sparks AB, Alferov O, Hong P, Alexeev A, Jiang Y, Dahl F, Tang YT, Haas J *et al.* **Accurate whole-genome sequencing and haplotyping from 10 to 20 human cells**. *Nature.* 2012; **487**(7406):190.

26. Adey A, Kitzman JO, Burton JN, Daza R, Kumar A, Christiansen L, Ronaghi M, Amini S, Gunderson KL, Steemers FJ *et al.* **In vitro, long-range sequence information for de novo genome assembly via transposase contiguity**. *Genome Res.* 2014; **24**(12):2041-2049.

27. Bishara A, Moss EL, Kolmogorov M, Parada AE, Weng Z, Sidow A, Dekas AE, Batzoglou S, Bhatt AS. **High-quality genome sequences of uncultured microbes by assembly of read clouds**. *Nat Biotechnol.* 2018; **36**(11):1067-1075.

28. Wang O, Chin R, Cheng X, Wu M, Mao Q, Tang J, Sun Y, Anderson E, Lam H, Chen D *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 Res.* 2019; **29**(5):798-808.

29. Chen Z, Pham L, Wu T-C, Mo G, Xia Y, Chang PL, Porter D, Phan T, Che H, Tran H *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 Res.* 2020; **30**(6):898-909.

30. Zheng GXY, Lau BT, Schnall-Levin M, Jarosz M, Bell JM, Hindson CM,

Kyriazopoulou-Panagiotopoulou S, Masquelier DA, Merrill L, Terry JM *et al.* **Haplotyping germline and cancer genomes with high-throughput linked-read sequencing**. *Nat Biotechnol.* 2016; **34**(3):303.

31. Danko DC, Meleshko D, Bezdan D, Mason C, Hajirasouliha I. **Minerva: an alignment- and reference-free approach to deconvolve Linked-Reads for metagenomics**. *Genome Res.* 2019; **29**(1):116-124.

32. Xu M, Guo L, Du X, Li L, Peters BA, Deng L, Wang O, Chen F, Wang J, Jiang Z *et al.* **Accurate haplotype-resolved assembly reveals the origin of structural variants for human trios**. *Bioinformatics.* 2021.

33. Bishara A, Liu Y, Weng Z, Kashef-Haghighi D, Newburger DE, West R, Sidow A, Batzoglou S. **Read clouds uncover variation in complex regions of the human genome**. *Genome Res.* 2015; **25**(10):1570-1580.

34. Guo J, Shi C, Chen X, Wang O, Liu P, Yang H, Xu X, Zhang W, Zhu H. **stLFRsv: A Germline Structural Variant Analysis Pipeline Using Co-barcoded Reads**. *Front Genet.* 2021; **12**:222.

35. Weisenfeld NI, Kumar V, Shah P, Church DM, Jaffe DB. **Direct determination of diploid genome sequences**. *Genome Res.* 2017; **27**(5):757-767.

36. Yeo S, Coombe L, Warren RL, Chu J, Birol I. **ARCS: scaffolding genome drafts with linked reads**. *Bioinformatics.* 2017; **34**(5):725-731.

37. Guo L, Xu M, Wang W, Gu S, Zhao X, Chen F, Wang O, Xu X, Seim I, Fan G *et al.* **SLR-superscaffolder: a de novo scaffolding tool for synthetic long reads using a top-to-bottom scheme**. *BMC Bioinformatics.* 2021; **22**(1):1-16.

38. Tolstoganov I, Bankevich A, Chen Z, Pevzner PA. **cloudSPAdes: assembly of synthetic long reads using de Bruijn graphs**. *Bioinformatics.* 2019; **35**(14):i61-i70.

39. Kuleshov V, Snyder MP, Batzoglou S. **Genome assembly from synthetic long read clouds**. *Bioinformatics.* 2016; **32**(12):i216-i224.

40. Wood DE, Lu J, Langmead B. **Improved metagenomic analysis with Kraken 2**. *Genome Biol.* 2019; **20**(1):1-13.

41. Lu J, Breitwieser FP, Thielen P, Salzberg SL. **Bracken: estimating species abundance in metagenomics data**. *PeerJ Comput Sci.* 2017; **3**:e104.

42. Danko DC, Meleshko D, Bezdan D, Mason C, Hajirasouliha I. **Novel Algorithms for the Taxonomic Classification of Metagenomic Linked-Reads**. *bioRxiv.* 2019:549667.

43. Gurevich A, Saveliev V, Vyahhi N, Tesler G. **QUAST: quality assessment tool for genome assemblies**. *Bioinformatics.* 2013; **29**(8):1072-1075.

44. Marks P, Garcia S, Barrio AM, Belhocine K, Bernate J, Bharadwaj R, Bjornson K, Catalanotti C, Delaney J, Fehr A *et al.* **Resolving the full spectrum of human genome variation using Linked-Reads**. *Genome Res.* 2019; **29**(4):635-645.

45. Guo X, Chen F, Gao F, Li L, Liu K, You L, Hua C, Yang F, Liu W, Peng C *et al.* **CNSA: a data repository for archiving omics data**. *Database (oxford).* **2020**.

46. Chen FZ, You LJ, Yang F, Wang LN, Guo XQ, Gao F, Hua C, Tan C, Fang L, Shan RQ. **CNGBdb: China National GeneBank DataBase**. *Heredidas.* 2020; **42**(8):799-809.

47. Parks DH, Imelfort M, Skennerton CT, Hugenholtz P, Tyson GW. **CheckM: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes**. *Genome Res.* 2015; **25**(7):1043-1055.

48. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R, Subgroup GPDP. **The Sequence Alignment/Map format and SAMtools**. *Bioinformatics.* 2009; **25**(16):2078-2079.

49. Chaumeil P-A, Mussig AJ, Hugenholtz P, Parks DH. **GTDB-Tk: a toolkit to classify genomes with the Genome Taxonomy Database**. *Bioinformatics.* 2019; **36**(6):1925-1927.

50. Price MN, Dehal PS, Arkin AP. **FastTree 2 – Approximately Maximum-Likelihood Trees for Large Alignments**. *PLOS ONE.* 2010; **5**(3):e9490.

51. Letunic I, Bork P. **Interactive Tree Of Life (iTOL) v4: recent updates and new developments**. *Nucleic Acids Res.* 2019; **47**(W1):W256-W259.

52. Koren S, Rhie A, Walenz BP, Dilthey AT, Bickhart DM, Kingan SB, Hiendleder S, Williams JL, Smith TP, Phillippy AM. **De novo assembly of haplotype-resolved genomes with trio binning**. *Nat Biotechnol.* 2018; **36**(12):1174-1182.

53. Nicholls SM, Quick JC, Tang S, Loman NJ. **Ultra-deep, long-read nanopore sequencing of mock microbial community standards**. *GigaScience.* 2019; **8**(5).

54. Shin N-R, Whon TW, Bae J-W. **Proteobacteria: microbial signature of dysbiosis in gut microbiota**. *Trends Biotechnol.* 2015; **33**(9):496-503.

55. Johnson KV-A. **Gut microbiome composition and diversity are related to human personality traits**. *Hum Microbiome J.* 2020; **15**:100069.

56. Zeevi D, Korem T, Godneva A, Bar N, Kurilshikov A, Lotan-Pompan M, Weinberger A, Fu J, Wijmenga C, Zhernakova A. **Structural variation in the gut microbiome associates with host health**. *Nature.* 2019; **568**(7750):43-48.

57. Chen L, Wang D, Garmaeva S, Kurilshikov A, Vila AV, Gacesa R, Sinha T, Segal E, Weersma RK, Wijmenga C *et al*. **The long-term genetic stability and individual specificity of the human gut microbiome**. *Cell*. 2021; **184**(9):2302-2315. e2312.