doi: 10.1093/nar/gkz017

Amino acid based *de Bruijn* graph algorithm for identifying complete coding genes from metagenomic and metatranscriptomic short reads

Jiemeng Liu[1,2,†], Qichao Lian[1,†], Yamao Chen[1] and Ji Qi[1,*]

[1] State key Laboratory of Genetic Engineering, Institute of Plant Biology, School of Life Sciences, Fudan University, Shanghai 200433, China and [2] The T-Life Research Center, Fudan University, Shanghai 200433, China

Received November 27, 2018; Revised December 19, 2018; Editorial Decision January 06, 2019; Accepted January 08, 2019

ABSTRACT

Metagenomic studies, greatly promoted by the fast development of next-generation sequencing (NGS) technologies, uncover complex structures of microbial communities and their interactions with environment. As the majority of microbes lack information of genome sequences, it is essential to assemble prokaryotic genomes ab initio aiming to retrieve complete coding genes from various metabolic pathways. The complex nature of microbial composition and the burden of handling a vast amount of metagenomic data, bring great challenges to the development of effective and efficient bioinformatic tools. Here we present a protein assembler (MetaPA), based on *de Bruijn* graph searching on oligopeptide spaces and can be applied on both metagenomic and metatranscriptomic sequencing data. When public homologous protein sequences are involved to guide the assembling procedures, MetaPA assembles 85% of total proteins in complete sequences with high precision of 83% on real high-throughput sequencing datasets. Application of MetaPA on metatranscriptomic data successfully identifies the majority of actively transcribed genes validated in related studies. The results suggest that MetaPA has a good potential in both metagenomic and metatranscriptomic studies to characterize the composition and abundance of microbiota.

INTRODUCTION

Metagenomics, which treats all the genetic materials directly extracted from environmental microbial samples as a whole for research (1), has become a significant methodology to study uncultured microbe in various habitats including human body, oceans, soil, etc. (2–5). Studies on metagenomics lead to the knowledge to thousands of previously unknown microbes and their interactions with habitats, thus promoting further understanding on evolutionary history of life (6) and bringing out the application to medicine and industry (7–9). As an essential step in most metagenomic sequencing projects, identifying coding sequences from vast amount of next-generation sequencing (NGS) reads, severely affects the profiling of both taxonomic and functional composition. However, the intrinsic complexity of microbiome brings great challenges to the development of efficient and effective bioinformatic algorithms for short NGS reads assembly, when facing diversified similarities among various genome sequences.

To our knowledge, there are currently two alternative approaches to predict genes from raw metagenomic reads. One type of approaches, e.g. MetaVelvet (10), IDBA-UD (11), MEGAHIT (12) and metaSPAdes (13), assemble short NGS reads into longer contigs, from which genes are identified by utilizing the third-party gene finders (14). Considering the species abundance variation represented by the uneven coverage of different prokaryotic genomes, these methods make great improvement compared with those designed for the assembly of single genome, e. g. SOAPdenovo2 (15). However, these nucleotide based assembly approaches may fail to obtain complete coding sequences as synonymous polymorphisms are widely distributed among populations, species or strains. As a consequence, their final output sequences might be fragmentary, especially on metagenomic samples with complex microbial community structures.

The other type of algorithms, e.g. SFA-SPA (16,17) and inGAP-CDG (18), directly deals with peptides or translated nucleotide sequences via prediction of open reading frames (ORFs) from short raw reads, aiming to reconstruct complete protein sequences by assembling short peptide sequences. As protein coding genes occupy the majority of the genomes of microbes (19,20), and are highly conserved across strains within the same species, these methods advance in obtaining 'pan-proteome' in given metagenomic samples compared with the nucleotide based assem-

[*] To whom correspondence should be addressed. Tel: +86 21 31246533; Email: qij@fudan.edu.cn
[†] The authors wish it to be known that, in their opinion, the first two authors should be regarded as joint First Authors.

blers. However, false ORF-prediction, arisen from the short length of metagenomic reads, unavoidably leads to the generation of many false gene sequences. Furthermore, current available peptide assemblers are designed by utilizing oligopeptides with fixed-length (*k*-mers), thus are difficult to handle a mixture of homologous sequences with various similarities.

In this study, we present a novel protein assembler, MetaPA, to predict coding sequences from the high-throughput metagenomic data by optimizing ORF detection and peptide assembly simultaneously. MetaPA adopts a *de Bruijn* graph based strategy and depends on multiple graphs constructed by k-mers with different lengths (up to 24 amino acids), whose iteration benefits the correction of sequencing error and the reduction of false ORF prediction, and leads to simpler *de Bruijn* graphs, which yield more completely assembled proteins. In addition, published microbial protein sequences, if available, can be introduced to guide the assembly of metagenomic sequences. Tested by sequencing datasets of either synthetic or real microbial communities, MetaPA succeeded in detecting more complete protein sequences with higher accuracy compared with other approaches. A case study of metatranscriptomic samples showed that the majority of raw reads could be mapped onto the proteins assembled by MetaPA, benefiting accurate estimation of both taxonomic abundance and functional gene expressions.

MATERIALS AND METHODS

To obtain complete proteins of multiple microorganisms from high-throughput metagenomic data, MetaPA performs the assembly of short reads on the amino acid level based on two considerations. First, due to the complex nature of microbial communities, the assembly of individual prokaryotic genes greatly reduces the challenges than that of complete genome sequences, without losing too much information since the majority of genomic regions of microbes are coding sequences. Second, the presence of variations in coding regions or intergenic regions among different prokaryotic strains often leads to difficulties in assembling single-species sequences, while the conservation of related homologs in amino acid levels decreases the mathematical complexity of assembly, e.g. the popular used *de Bruijn* graphs adopted in this study. In addition, MetaPA gradually increases the lengths of *k*-mers to improve the assembly quality. Organism-specific proteins and nearly identical ones among species/strains are assembled from graphs with shorter *k*-mers, while proteins with lower similarities are possibly resolved into individual proteins when using longer *k*-mers. Briefly, we employ a multiple-step strategy to process an assembly job (Supplementary Figure S1): (i) predict ORFs from short nucleotide reads and translate them to protein segments; (ii) construct *de Bruijn* graphs in the space of oligopeptides, where a node denotes a *k*-mer and an edge represents a (*k*+1)-mer to connect two overlapping *k*-mers; (iii) simplify the graph and decompose it into subgraphs (denoted by connected components), from which the longest/shortest paths are called to evaluate confidence of associated *k*-mers; (iv) for each short read, evaluate ORF candidates according to summarized confidence score of *k*-mers, then repeat the steps of (i–iv) using longer *k*-mers; (v) in the procedure of calling paths from sub-graphs, read sequences and paired-end information are utilized to make decisions when meeting forks or crosses. If not determined, reference protein sequences are adopted as templates to guide the procedure of path searching; (vi) remove redundantly assembled proteins and false sequences predicted from intergenic regions.

Prediction of ORFs from raw reads

MetaPA adopts the universal DNA codon table to translate each query nucleotide sequence into coding sequences by using a six-frame translation strategy similar to OrfPredictor (21). Each of the six ORF candidates is evaluated by considering the presence/absence of start/stop codons: an ORF is directly translated when it has no stop codon; if only one stop codon is observed, the longer translated segment, either the one before the stop codon, or the one from the start codon (following the stop codon) to the end of the reads, is adopted; ORFs with two or more stop codons are ignored. After the six ORF candidates have been evaluated, only the one who has the longest ORF fragment is considered as the correct translation. Meanwhile, ORFs with translated protein segments shorter than 20 a.a. are also ignored from further analyses. In case that more than one ORF pass the filter, one of them is randomly chosen in this round of ORF prediction. These ORF candidates are then further evaluated by comparing with assembled proteins (see section ”Recalling of ORFs from raw reads under the guidance of de Bruijn graph”) resolved from reconstructed *de Bruijn* graphs to correct false ORFs called previously and potential sequencing errors in raw nucleotide reads.

Construction of *de Bruijn* graphs

Oligopeptides (k-mers) are collected from putative ORFs of query nucleotide reads, to construct a *de Bruijn* graph, which is represented by a hash map with k-mers as nodes. A tip is removed if shorter than twice of the length of k-mer with read coverage lower than 4×; a bubble is merged when two subpaths of the bubble have identical length with a single amino acid difference, e.g. exhibiting sequence similarity of $(2k - 1)/2k$ (Supplementary Figure S4). MetaPA further utilizes information of average insertion length of the paired-end (PE) sequencing library to predict a potential path connecting two ends of a read. Translated peptides of PE reads are mapped to the graph by using a strategy of spectral alignment (22). To estimate average insert size (the average distance of two ends of reads, denoted by l) and its standard variation (denoted by δ), the minimal distance (number of *k*-mers from N- to C-terminal on the graph) between the two translated products from the same PE read is collected for counting distributions of insert sizes. When l and δ are estimated, distance between two ends of each PE read in the graph is re-evaluated, a read is considered as ‘normally mapped’ if its insert size falls into range of $l \pm 3\delta$. DNA libraries usually adopt longer insert size than length of read sequence, leaving a gap between two ends. When calculating reads coverage on a path, MetaPA fills the gap between two ends of a normally mapped PE read by putting an extra coverage to the uncovered nodes in the gap. This procedure benefits the assembling of longer proteins while slightly increases the complexity of related graphs.

Decomposition of graphs and identification of protein sequences

An initial *de Bruijn* graph constructed from whole metagenomic data consists of multiple connected components. For each connected component, MetaPA traverses each nonbranching node (both in-degree and out-degree < 2) in descending order of reads coverage, and considers the given node as a seed to extend on both directions (defined as a subpath) until reaching to a fork or a cross (Supplementary Figure S2A). The decision of which alternative path to choose is made according to the following rules (Supplementary Figures S4–S6): (i) Amino acid sequences translated from raw reads are selected if contain the *k*-mer adjacent to the fork. The peptides are then 'aligned' to each of the alternative paths to examine if connections of multiple continuous *k*-mers are supported. Paths not covered by translated sequences are ignored from subsequent analyses. (ii) In addition, links of paired ends provide examination by using longer range connections of *k*-mers (up to the insert size *l*) than those of sequences from one end. Paths are chosen when one or more paired-end links fully span the k-mer junction in a fork/cross. (iii) Finally, redundant paths are removed from the collections (Supplementary Figures S4–S6).

The path search progress of MetaPA is benefitted by using public protein sequences from organisms with whole genome information, of which the number is sharply increasing in recent years, to guide the assembly procedure (Supplementary Figures S1 and S2B). When reference proteins are available as templets, subgraphs sharing 60% *k*-mers or more with a reference protein are selected. MetaPA then employs a dynamic programming strategy to find a path having the best alignment against the reference protein for each pair of start/end nodes. These alignments are further refined by using a Smith–Waterman algorithm, and the path with the optimal score is chosen for further evaluation even when lacking supports from sequences or paired-end links from reads as described in the above procedure.

Recalling of ORFs from raw reads under the guidance of *de Bruijn* graph

To further improve the performance of MetaPA, all the six ORF candidates of each raw read are compared with the *de Bruijn* graph obtained in the previous round to correct false ORF predictions or potential sequencing errors of raw reads. When protein sequences are recalled from the refined prediction of ORFs, longer k-mers are adopted to construct a *de Bruijn* graph for distinguishing heterozygous proteins with higher resolution. The pipeline of the strategy is shown in Supplementary Figure S3.

First, we download 2463 bacterial genomes from National Center for Biotechnology Information (NCBI), and use inGAP-sv to simulate 2 × 100 bp paired-end short reads with coverage of 30×. For each species, a *de Bruijn* graph is built and all alternative paths are outputted and compared with reference proteins. We find the majority of predicted proteins (62%) with length longer than 300 a.a. are well matched to reference proteins. On the contrary, 80% of assembled proteins shorter than 150 a.a. lack such supports (Supplementary Figure S8). Accordingly, we classify proteins called from the previous round of *de Bruijn* graph into three classes based on their lengths: proteins of high confidence (300 a.a. or longer), those of low confidence (from 150 a.a. to 300 a.a.) and unreliable ones (150 a.a. or shorter). Furthermore, we also find that many long false ORFs are possibly predicted from high GC content regions (60% or higher, Supplementary Figure S8), and are often associated with unusual amino acid abundance compared with true proteins (Supplementary Figure S9). MetaPA thus classifies those predicted proteins as unreliable ones when they own high abundance of Arginine (15% or more), Serine (14% or more), Histidine (7.5% or more) or Glycine (14% or more) according to the comparison results on the simulated short reads (Supplementary Figure S9). Next, each k-mer is labeled as high/low confident or unreliable ones depending on their appearance on the predicted proteins with different confidence levels in the previous round of assembly, ambiguous k-mers associated with non-unique labels are ignored from further analysis.

Second, six-frame translations are performed on each raw read to produce putative proteins for evaluation under the guidance of the subgraphs of different confident levels. In details, each frame candidate is compared with all subgraphs by using a strategy of spectral alignment (22), and the numbers of high/low confident and unreliable *k*-mers are counted. Among the six potential frames, the ORF candidate having the most high-confident *k*-mers is chosen for graph construction; the low-confident *k*-mer numbers will be an additional criterion when no judgement can be made by the high-confident *k*-mers, so as to the unreliable *k*-mers. Furthermore, single amino acid inconsistency in assembly graphs caused by sequencing errors is corrected by using a similar algorithm as described in (22), which is designed for nucleotide sequences.

At last, the updated ORFs are used to construct a new graph based on (*k*+1)-mers to refine protein assembly. The procedures are repeated until the length of k-mers reaches to a maximum value defined by users.

Calculation of CDS sequences and refining protein callings

Raw reads are aligned to assembled proteins by using a fast alignment tool, Diamond (23), only those reads showing high similarities (fully aligned, up to 2 mismatches and no gap allowed) with the targeted proteins are adopted for further analysis. Nucleotide sequence of each protein is called by calculating consensus sequences upon mapped reads, on which each nucleotide is evaluated according to a Bayesian model (24) for detection of potential polymorphisms. For each CDS, MetaPA employs a similar strategy as inGAP-sv (25) does to find break points, where lack coverage of paired-end reads with normal distance and strands. A CDS is split into fragments to avoid chimeric protein callings when meeting two requirements: it exhibits one or more break points; a pair of neighbouring spited fragments have different targets of the reference protein database. Proteins with similarity threshold of 95% are considered as redundant sequences and are removed by using CD-HIT (26). CDS sequences are further compared to remove assembled products from antisense strand of coding genes by using CD-HIT-EST (26) with options of '-G 0 -c 0.95 -aS 0.9'. In addition, MetaPA applies a model of supporting vector machine (SVM) trained by intergenic regions of the 2463 species from NCBI, to recognize pseudo proteins assembled by translated sequences of non-coding short reads.

Datasets of real sequences

Dataset 1 (a synthetic sequencing data). To compare the performance of MetaPA with that of other methods upon real sequencing data of microbial communities (27), a dataset sequenced on Illumina platform (with accession number as SRR606249, 11.1 Gb) is downloaded from the NCBI Short Reads Archive. The raw reads are trimmed to remove adaptor sequences and low-quality reads by using Trimmomatic 0.36 (28) with the parameter 'LEADING:3 TRAILING:3 SLIDINGWINDOW:4:15 MINLEN:50'. These filtered reads are then mapped onto the reference database including 64 complete prokaryotic genomes by using BWA (29) to profile species compositions under the synthetic conditions. Those reads lacking matches on the reference genomes are ignored from the analyses. Performance of MetaPA and other metagenomic assemblers upon these short reads are evaluated by comparing the assembled proteins/contigs with the 64 reference proteomes. Another 5872 prokaryote proteomes from NCBI, excluding the 64 proteomes, are downloaded to guide the assembly procedure of MetaPA.

Dataset 2 (two metagenomic datasets of human stool samples). Real metagenomic datasets are adopted to compare the performance of MetaPA with other methods on proteome assembly. Two datasets of human stool samples (sample SRS078176 and SRS022524) are downloaded from the NCBI Short Reads Archive. Both of them are sequenced on Illumina platform and produce 2 × 95 bp and 2 × 100 bp paired-end reads, respectively. Sample SRS078176, consisting of 10.1 Gb sequencing data, is expected to exhibit higher sequencing depth per species than sample SRS022524 containing only 2.3 Gb sequencing data, since they owe comparable microbial community complexities. The raw reads are trimmed under the same procedure as in Dataset 1. The HMP project releases a list of 2307 reference genome assemblies (30–32), which are adopted for evaluation of performance of all the methods. We download 5936 prokaryote genomes from NCBI to provide an independent database for guiding the assembly procedure of MetaPA.

Dataset 3 (a metagenomic and a metatranscriptomic datasets of the same stool sample). Compared with metagenomic data, which profile structure of microbial communities, metatranscriptomic data provide snapshot of their transcription activity. To validate the utilization of MetaPA for assembling proteins from either DNA-seq or RNA-seq sequences, we download a metagenomic dataset (SRS302292, SRS302293, SRS302298 and SRS302307) and a metatranscriptomic dataset (SRS302300, SRS302306, SRS302315 and SRS302319) from NCBI in a study of human stool samples by Giannoukos *et al*. (33). These two datasets are sequenced from the same stool sample, constituting data of 5.0 Gb and 7.9 Gb, respectively, providing a cross validation for examining the performance of these methods. The raw reads are trimmed under the same procedure as in the above datasets. The assembly analyses of MetaPA are guide by using 5936 prokaryote proteomes adopted in Dataset 2.

RESULTS

We compared the performance between MetaPA and six other assemblers, among which SFA-SPA (17) is also an amino acid based assembler, and all the other five are nucleotide based approaches, namely IDBA-UD (11), MEGAHIT(12), metaSPAdes (13), MetaVelvet (10) and SOAPdenovo2 (15). SOAPdenovo2, on the other hand, is an assembler designed for single genome and has been widely applied on many metagenomic studies like MetaHIT (34) and HMP (30–32). Contigs obtained by IDBA-UD, MEGAHIT, metaSPAdes, MetaVelvet, and SOAPdenovo2 were given to FragGeneScan (14) for further prediction of ORFs and proteins to facilitate the comparisons among different approaches. The length of *k*-mers were selected as 18 a.a. for MetaPA when dealing with metagenomic datasets and 14 a.a. for metatransriptomic data. *k*-mers of 51 nt were applied for both MetaVelvet and SOAPdenovo2 and as 6 a.a. for SFA-SPA, while parameters for the other approaches were set as default.

Performance comparisons on the benchmark data (Dataset 1)

Shakya *et al*. (27) built a synthetic prokaryotic community, owning 16 organisms of Archaea and 48 members of Bacteria. As this data covered most phyla with published whole genome information, it was adopted as a benchmark for the quantitative comparisons among the seven methods. Briefly, application of MetaPA on this dataset yielded a total of 177 764 sequences composing of 58.8 Mega amino acids [Maa], with an average length of 331 a.a. As a comparison, the other six methods obtained 178 383–555 551 sequences in total (44.2–63.0 Maa, with an average length of 113–289 a.a.) (Figure 1, Supplementary Figure S10). It showed MetaPA outperforms the other methods on obtaining longer proteins when yielding similar amount of total assembly products. In detail, 147 857 sequences by MetaPA (83.2% of total, Figure 1B and Supplementary Figure S10B) were full-length proteins (matched 90% or longer region of reference proteins), which covered 85.9% of the total reference proteins (Figure 1A, Supplementary Figure S10A). MetaPA showed the highest recall and precision rates (85.9% and 83.2%), followed by metaSPAdes (81.5% and 78.5%), SOAP (67.5% and 59.5%), MetaVelvet (58.1% and 67.8%), SFA-SPA (51.6% and 35.1%), MEGAHIT (26.1% and 16.4%), IDBA_UD (25.4% and 16.2%). In addition, only a few reference proteins were assembled as fragmented sequences by MetaPA, e.g. 5.2% proteins were 70–90% complete, 1.6% were 50–70% complete, only 0.8% were 30–50% complete (Figure 1A), illustrating a better performance of MetaPA than those of the other methods. It

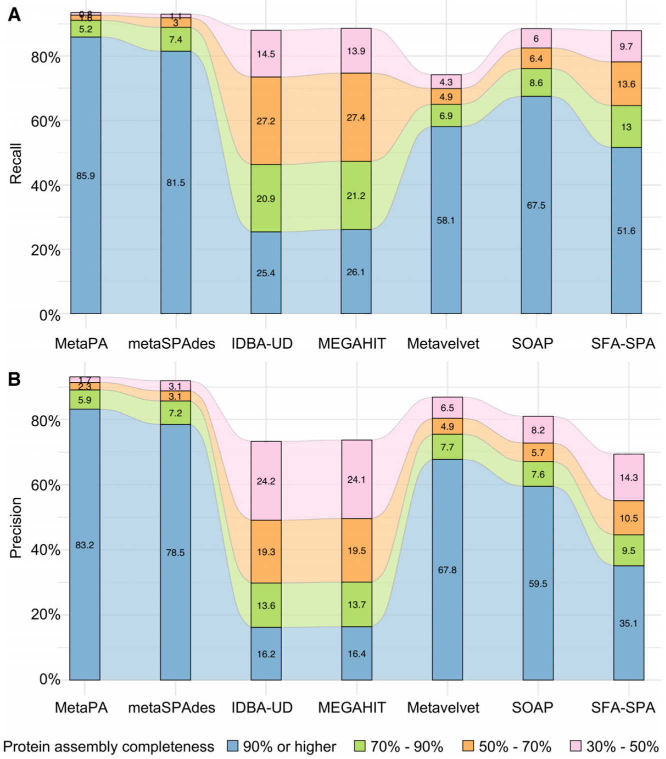

Figure 1. Performance of the seven approaches on a sequencing dataset of microbial synthetic communities. (**A**) Recall and (**B**) precision values of assembled sequences are displayed, where sequences are classified into four levels according to their completeness compared with reference proteins. An assembled sequence is considered to be matched to a reference protein when over 70% of the query sequence is aligned with identity higher than 80%.

is worth noting that most metagenome assemblers, both amino acid based and nucleotide based (except IDBA_UD and MEGAHIT), only produced hundreds of chimeric sequences (see definition in Supplemental Figure S7), representing 0.5% or lower proportion of the total assembled proteins (Supplementary Figure S11A). Considering redundant sequences, MetaPA outputted slightly higher numbers (6697) than other methods (313–5,608 proteins) (Supplementary Figure S11B), suggesting some nodes in related subgraphs are repeatedly adopted in MetaPA. Nevertheless, MetaPA performed well as other methods, except SFA-SPA, in yielding small amount of sequences unaligned to reference proteins (Supplementary Figure S11C).

The 64 prokaryotes in the synthetic community exhibited wide range of abundance, from 9.4× for *Burkholderia xenovorans* to 310.0× for *Nanoarchaeum equitans*. To provide detailed comparison of the seven methods for assembling proteins upon various amounts of short reads from different abundant species, the 64 organisms are classified into three groups: 9 species with 100× or higher coverage, 45 species with abundance from 20× to 100× and 10 organisms under 20×. As shown in Figure 2A and Supplementary Table S1, MetaPA successfully assembled most amounts of prokaryotic proteins (with completeness of 50% or higher) in all groups, i.e. 90.1% at the high abundance group, 90.4% at the median and 88.1% at the low abundance group. While

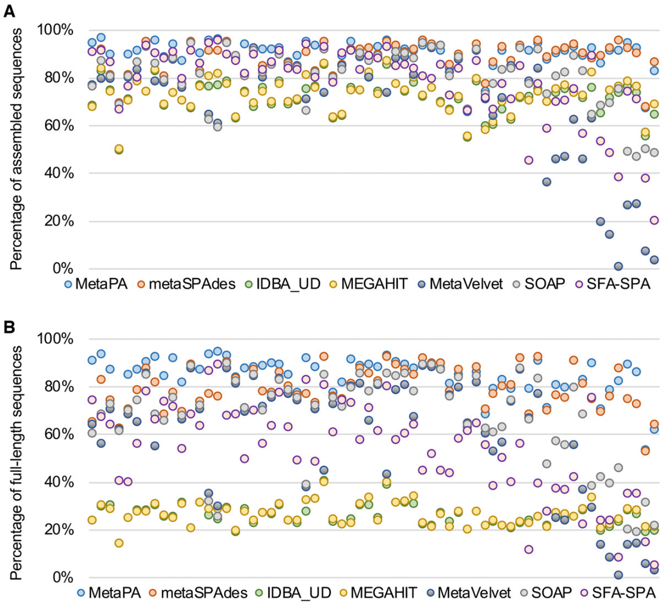

Figure 2. Evaluation on sequence fragmentation of the output proteins of the seven approaches in the synthetic metagenomic dataset. (**A**) Percentage of assembled sequences with completeness of at least 50% or (**B**) full-length recalled sequences (completeness of 90% or higher) for each of the 64 prokaryotic genomes in the synthetic metagenomic dataset.

the performances of metaSPAdes, MEGAHIT, IDBA_UD and SOAP were also insensitive to species abundance, from 72.4–83.6% (high abundance) to 72.2–88.7% (median abundance) and to 64.5–88.7% in the low abundance group. On the contrary, only 27.0% and 56.4% proteins were successfully assembled by MetaVelvet and SFA-SPA, respectively, upon bacterial genomes with coverage of under 20×. When comparing effectiveness of these methods on obtaining full-length proteins, MetaPA could still recall 85.5%, 85.2% and 78.2% of reference proteins for species with high, median and low abundance, respectively, outperform the other methods with recall of 25.6–74.2%, 25.8–80.6% and 17.6–73.9% (Figure 2B and Supplementary Table S2).

Performance evaluation on two human gut metagenomic samples (Dataset 2)

Real metagenomics samples often have relatively complex community structures by including microbes with highly variable abundance and/or from a wide spectrum of taxonomy. To evaluate the performance of the seven approaches on real metagenomic datasets, we applied them on two human gut metagenomic NGS datasets released by the HMP project (30–32). The two microbiomes show similar community complexities containing 118 and 92 bacteria, with 100 or more genes per species, as detected in the assembly of the two datasets by HMP, respectively, but vary in data size. The predicted coding genes by each approach were aligned to the reference genomes of HMP for classification and evaluation. As shown in Table 1, the application of these methods on the deeper sequenced dataset SRS078176 (10.1 Gb) showed most of algorithms yielded similar amount of product size (42.7–59.0 Maa in total) except SFA-SPA (32.5 Maa). Moreover, MetaPA obtained 181 925 sequences with an average length of 287 a.a. illustrating a better performance on the assembly completeness than other methods, which exhibited 288 053–340 725 sequences with an average length of 100–173 a.a. Especially, 56 031 sequences yielded by MetaPA (30.8% of total) were almost of full-length, qualified by reference proteins in the HMP database, more than those (17 592 to 51 360 sequences) of the other six methods (Table 1). Consistently, MetaPA outperformed the other methods by providing fewer fragmented sequences at various levels of incompleteness (Table 1). We then compare the performance of these algorithms on the more challenging dataset SRS022524, which owns only 2.3 Gb data. MetaPA yielded similar amount of product size (amino acids) with the longest average protein length (299 a.a.), providing further evidence for the higher effectiveness of MetaPA on detecting more completed protein sequences compared with the other approaches.

A case study of MetaPA on a sample with both metagenomic and metatranscriptomic sequencing data (Dataset 3)

To test the performance of MetaPA in dealing with various types of microbiome data, based on DNA-seq or RNA-seq, we adopted a pair of metagenome (5.0 Gb) and metatranscriptome (7.9 Gb) datasets from one human stool sample sequenced by Giannoukos *et al*. (33). Consistent with previous study, the majority of metagenomic reads (88.5%) from the microbial community came from 19 bacterial species belonging to three phyla: Firmicutes (13 species), Bacteroidetes (5 species), and Actinobacteria (1 species). Consistently, the 19 species also dominated in the metatranscriptome dataset (94.6% of total reads) but exhibited higher degree of abundance variation, thus providing a cross validation to measure the effectiveness of the seven methods upon DNA or RNA based studies. As a result, MetaPA outperformed other methods in similar patterns as shown in the other two datasets (Figure 3 and Table 2).

35 694 440 of 39 248 382 metagenomic reads were assembled by MetaPA, yielding 63 977 proteins with an average length of 315 a.a. A total of 41 471 proteins (64.8%) were classified into the 19 species from the three phyla of Firmicutes, Bacteroidetes and Actinobacteria (Figure 3, Figure 4A and Table 2), accounting for over 78.7% of the assembled reads, consistent with the observation by Giannoukos *et al*. Further examination of the assembled proteins for each species illustrated that the majority of reference proteins (72.1%) were recalled (30% or longer regions were covered) although their sequencing depth fluctuated from 7× (*Bacteroides* sp. 3-1-23) to 484X (*Prevotella copri* DSM 18205). Moreover, 52.7% proteins of *Ruminococcus obeum* were recalled even when its sequencing depth was as low as 4×, while 19.9% proteins were assembled for *Clostridium nexile* with depth of 0.7× (Figure 4A).

We then applied MetaPA on the metatranscriptome dataset, within which 59 527 614 of 62 883 662 short reads (94.7%) were utilized to produce 29,726 proteins (Table 2). Consistently, the majority of assembled proteins (20 475, 68.9%) accounting for 80.3% of assembled reads, came from the 19 species detected in the metagenomic data, exhibiting similar abundance variation pattern as reported by Giannoukos *et al*.: *Prevotella copri* DSM 18205, as the most

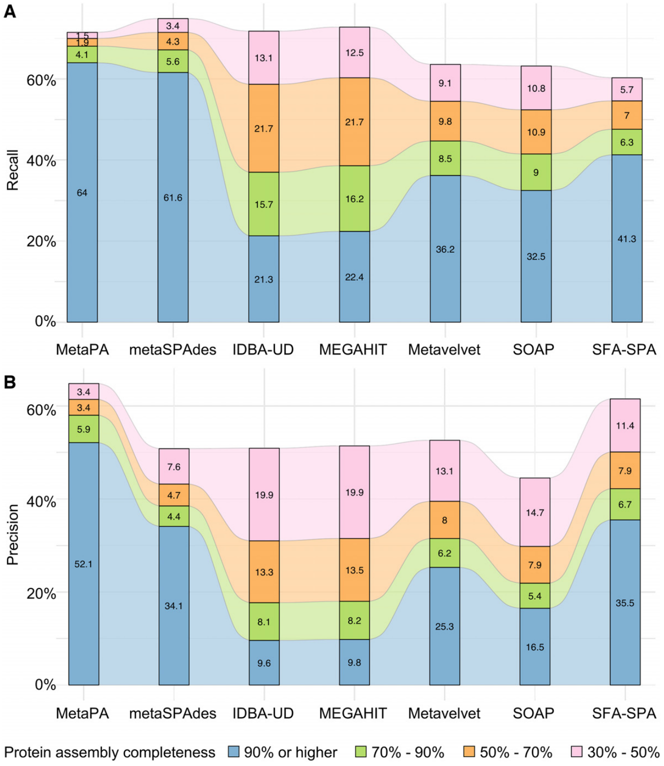

Figure 3. Performance of the seven approaches on real metagenomic sequencing data of a human stool sample. (A) Recall and (B) precision values are displayed in different levels of completeness.

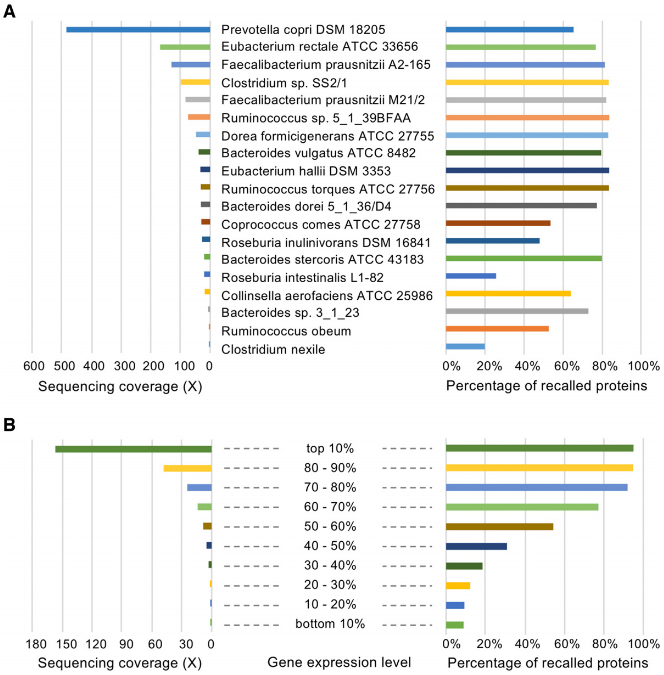

Figure 4. Analyses of a paired metagenomic and metatranscriptomic datasets of a human stool sample by MetaPA. (**A**) Display of sequencing coverages (the left panel) and recall rates (the right panel) for the top 19 abundant species in the metagenomic dataset. (**B**) Sequencing coverage (the left panel) and recall rates (the right panel) for genes with different expression levels in the metatranscriptomic dataset. Protein sequences assembled by MetaPA with completeness of 30% or higher are taken into account.

Table 1. Summary of performance of seven approaches on two metagenomic datasets of human stool samples. All the assembled sequences from the seven approaches were then blast against the HMP reference genomes

Approach	MetaPA	metaSPAdes	IDBA_UD	MEGAHIT	MetaVelvet	SOAP	SFA-SPA
Sample			SRS078176 (10.1 Gb)				
Total assembled size (a.a.)	52 216 405	59 025 179	48 410 556	47 462 147	53 378 828	42 697 585	32 539 419
Total fragments	181 925	340 725	294 972	288 053	410 314	305 629	326 102
Mean length	287.0	173.2	164.1	164.8	130.1	139.7	99.8
Full-length fragments[a]	56 031	51 360	19 231	19 174	27 947	31 646	17 592
70∼90%[b]	8 521	8 159	16 291	16 087	10 544	6 225	5 036
50∼70%[c]	6 185	11 496	24 704	24 871	15 789	8 810	6 179
30∼50%[d]	7 351	23 313	38 419	38 163	28 836	16 370	9 659
0–30%[e]	15 481	54 156	41 506	41 756	57 459	38 453	15 309
no hit	88 356	192 241	154 821	148 002	269 739	204 125	272 327
Time	21h21m	9h55m	5h43m	2h31m	6h1m	1h51m	29h40m
Sample			SRS022524 (2.3 Gb)				
Total assembled size (a.a.)	36 287 032	38 872 956	33 029 099	33 964 537	37 752 147	27 874 834	14 266 759
Total fragments	121 331	226 220	199 447	208 275	335 125	251 574	145 741
Mean length	299.1	171.8	165.6	163.1	112.7	110.8	97.9
Full-length fragments[a]	41 467	35 707	14 129	14 559	18 060	14 779	9 182
70∼90%[b]	8 821	6 296	11 837	12 266	7 541	5 853	3 257
50∼70%[c]	5 843	7 894	18 144	19 026	11 472	8 910	4 086
30∼50%[d]	6 301	13 645	27 417	29 057	21 860	17 630	6 693
0–30%[e]	9 557	27 768	29 172	32 594	46 576	40 712	12 785
no hit	49 342	13 4910	98 748	100 773	229 616	163 690	109 738
Time	6h	4h	2h30m	1h34m	2h	58m	11h4m
Threads	8	8	8	8	1	8	8

[a] Refers to assembled sequences covering 90% or more of amino acids on targeted reference proteins.

[a-e] An assembled sequence is considered to be matched to a reference protein in HMP when over 70% of the query sequence is aligned with identity higher than 80%.

Table 2. Summary of performance of seven approaches on a pair of metagenomic and metatranscriptome samples. All the assembled sequences from the seven approaches were then blast against a database with 19 bacteria genomes

Approach	MetaPA	metaSPAdes	IDBA_UD*	MEGAHIT	MetaVelvet	SOAP	SFA-SPA
Sample		SRS302292, SRS302293, SRS302298, SRS302307 (5.0 Gb)					
Total assembled size (a.a.)	20 152 734	21 056 101	18 875 766	20 141 499	15 995 998	17 856 096	12 865 390
Total fragments	63 977	101 754	114 227	120 939	110 834	152 258	116 283
Mean length	315.0	206.9	165.2	166.5	144.3	117.3	110.6
Full-length fragments[a]	33 320	29 435	9 441	10 152	16 497	13 815	15 068
70–90%[b]	3 773	3 790	7 924	8 492	4 073	4 543	2 851
50–70%[c]	2 207	4 036	13 123	14 060	5 195	6 584	3 336
30–50%[d]	2 171	6 539	19 592	20 706	8 533	12 319	4 851
0–30%[e]	2 651	10 961	19 327	19 783	14 628	25 307	6 541
no hit	19 855	46 993	44 820	47 746	61 908	89 690	83 636
Time	7h57m	3h19m	5h18m	1h11m	1h18m	2h14m	10h43m
Sample		SRS302300, SRS302306, SRS302315, SRS302319 (7.9 Gb)					
Total assembled size (a.a.)	8 404 524	7 817 382	7 669 930	7 162 149	7 705 853	5 586 267	5 472 030
Total fragments	29 726	42 874	54 947	42 112	82 977	57 395	70 095
Mean length	282.7	182.3	139. 6	170.1	92.9	97.3	78.1
Full-length fragments[a]	10 929	9 468	3 301	4 306	2 993	2 427	4 179
70–90%[b]	3 883	2 200	2 957	3 200	1 854	1 476	1 059
50–70%[c]	2 760	2 660	5 070	5 169	2 923	2 317	1 364
30–50%[d]	2 903	4 777	9 584	8 675	6 050	5 049	2 283
0–30%[e]	3 175	7 899	15 691	8 575	13 704	11 550	3 957
no hit	6 076	15 870	18 344	12 187	55 453	34 576	57 253
Time	26h51m	3h3m	2h48m	52m	2h13m	52m	28h18m
Threads	8	8	8	8	1	8	8

[a] refers to assembled sequences covering 90% or more of amino acids on targeted reference proteins.

[a-e] an assembled sequence is considered to be matched to a reference protein in 19 bacteria genomes when over 70% of the query sequence is aligned with identity higher than 80%.

[*] IDBA-UD and IDBA_Trans were used for predicting sequences from metagenomic and metatranscriptome dataset, respectively.

transcriptionally active species, occupied 53.9% of total assembled reads, followed by *Bacteroides vulgatus* ATCC 8482 (11.7%) and *Bacteroides dorei* 5_1_36/D4 (11.5%). Considering variation in gene expression, the assembled protein sequences were classified into ten categories according to their sequencing coverage (Figure 4B). As a result, the majority of proteins (77.2–95.1%) were recalled for the genes with expression level of top 40%, i.e. with sequencing coverage higher than 14×. The possibility of protein assembly descended along with the decrease of their abundance from 54.4% (8×) to 30.9% (5×). We also observed that the genes in the last two categories had average sequencing depth lower than 0.8×, leading to insufficient coverage of reads on proteins and fragmented assembling results.

DISCUSSION

In this study, we described MetaPA, a method for the assembly and prediction of protein sequences from metagenomic or metatranscriptomic short reads. As the presences of stop codons lead to the decomposition of *de Bruijn* graph into smaller subgraphs, each of which denotes a sequence cluster representing a few homologous protein sequences, the complex tasks of assembling whole microbial genomes were simplified into resolving of single proteins from mess of multiple homolog sequences. Thus, this strategy makes MetaPA outperform these nucleotide based assemblers discussed in this study. In addition, MetaPA is capable of utilizing longer *k*-mers (12–24 a.a.) compared with another amino acid based method SFA-SPA, to further simplify the decomposition procedure of graph thus enhance the effectiveness on calling of complete protein sequences. Actually, the recall and precision of MetaPA on real metagenomic datasets were slightly improved along with the increasing of *k*-mers lengths and reached the optimum values with the length of *k*-mer as 18 a.a. (Supplemental Figure S12), which occupied approximately half of 100 bp reads. In addition, the percentage of chimeric sequences produced by MetaPA decreases when applying longer *k*-mers and reaches to a plateau value at 18 a.a. (0.75%). These evaluations suggest a choice of 18-mer for MetaPA is most effective and efficient when applying upon metagenomic dataset with similar sequencing amount to these adopted in this study. Adoption of shorter *k*-mers, e.g. with length of 14 a.a., benefits protein assembly upon low abundant organisms by sacrificing computation time. Recent studies (35,36) suggest the challenge can be partially overcome by applying strategies of flow cytometry and single-cell sequencing, when dealing with microbiome with complex community structure. If samples exhibiting tens or hundreds of organisms are extracted, MetaPA is capable of uncovering most of proteins from these mini-metagenomes associated with relatively lower variation on species abundance.

It is worth noting that MetaPA consists of both functions of assembling and ORF prediction, while a third-party software is necessary for the other five nucleotide based assemblers for gene calling. Nevertheless, MetaPA consumes similar computational time with them, although MetaPA attempts to iterate construction of *de Bruijn* graphs by utilizing various lengths of *k*-mers to improve its performance. MetaPA requires up to 50 Gb memory, larger than those of other methods but still affordable on many servers, for assembling these metagenomic sequencing data in this study. Still, a more efficient way of MetaPA when dealing with long k-mers calls for further study. All analyses were done on a mini-server with 32 CPUs and 256 Gb memory.

DATA AVAILABILITY

The source code of MetaPA are available at https://sourceforge.net/projects/metapa/.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

Computational analyses were completed using the biological supercomputing server of Computing Center of Beijing Institutes of Life Science.

FUNDING

National Natural Science Foundation of China [31770244 and 31570224 to J.Q.]; State Key Laboratory of Genetic Engineering at Fudan University.

Conflict of interest statement. None declared.

REFERENCES

1. Amann,R.I., Ludwig,W. and Schleifer,K.H. (1995) Phylogenetic identification and in-situ detection of individual microbial-cells without cultivation. *Microbiol. Rev.*, **59**, 143–169.
2. Venter,J.C., Remington,K., Heidelberg,J.F., Halpern,A.L., Rusch,D., Eisen,J.A., Wu,D., Paulsen,I., Nelson,K.E., Nelson,W. *et al.* (2004) Environmental genome shotgun sequencing of the Sargasso Sea. *Science*, **304**, 66–74.
3. Gill,S.R., Pop,M., Deboy,R.T., Eckburg,P.B., Turnbaugh,P.J., Samuel,B.S., Gordon,J.I., Relman,D.A., Fraser-Liggett,C.M. and Nelson,K.E. (2006) Metagenomic analysis of the human distal gut microbiome. *Science*, **312**, 1355–1359.
4. Leininger,S., Urich,T., Schloter,M., Schwark,L., Qi,J., Nicol,G.W., Prosser,J.I., Schuster,S.C. and Schleper,C. (2006) Archaea predominate among ammonia-oxidizing prokaryotes in soils. *Nature*, **442**, 806–809.
5. Arumugam,M., Raes,J., Pelletier,E., Le Paslier,D., Yamada,T., Mende,D.R., Fernandes,G.R., Tap,J., Bruls,T., Batto,J.M. *et al.* (2011) Enterotypes of the human gut microbiome. *Nature* **473**, 174–180.
6. Spang,A., Saw,J.H., Jorgensen,S.L., Zaremba-Niedzwiedzka,K., Martijn,J., Lind,A.E., van Eijk,R., Schleper,C., Guy,L. and Ettema,T.J.G. (2015) Complex archaea that bridge the gap between prokaryotes and eukaryotes. *Nature*, **521**, 173–179.
7. Li,L.L., McCorkle,S.R., Monchy,S., Taghavi,S. and van der Lelie,D. (2009) Bioprospecting metagenomes: glycosyl hydrolases for converting biomass. *Biotechnol. Biofuels*, **2**, 10.
8. Raes,J., Letunic,I., Yamada,T., Jensen,L.J. and Bork,P. (2011) Toward molecular trait-based ecology through integration of biogeochemical, geographical and metagenomic data. *Mol. Syst. Biol.*, **7**, 473.
9. Qin,J., Li,Y., Cai,Z., Li,S., Zhu,J., Zhang,F., Liang,S., Zhang,W., Guan,Y., Shen,D. *et al.* (2012) A metagenome-wide association study of gut microbiota in type 2 diabetes. *Nature*, **490**, 55–60.
10. Namiki,T., Hachiya,T., Tanaka,H. and Sakakibara,Y. (2012) MetaVelvet: an extension of Velvet assembler to de novo metagenome assembly from short sequence reads. *Nucleic Acids Res.*, **40**, e155.
11. Peng,Y., Leung,H.C., Yiu,S.M. and Chin,F.Y. (2012) IDBA-UD: a de novo assembler for single-cell and metagenomic sequencing data with highly uneven depth. *Bioinformatics*, **28**, 1420–1428.

12. Li,D., Liu,C.M., Luo,R., Sadakane,K. and Lam,T.W. (2015) MEGAHIT: an ultra-fast single-node solution for large and complex metagenomics assembly via succinct de Bruijn graph. *Bioinformatics*, **31**, 1674–1676.
13. Nurk,S., Meleshko,D., Korobeynikov,A. and Pevzner,P.A. (2017) metaSPAdes: a new versatile metagenomic assembler. *Genome Res.*, **27**, 824–834.
14. Rho,M., Tang,H. and Ye,Y. (2010) FragGeneScan: predicting genes in short and error-prone reads. *Nucleic Acids Res.*, **38**, e191.
15. Luo,R., Liu,B., Xie,Y., Li,Z., Huang,W., Yuan,J., He,G., Chen,Y., Pan,Q., Liu,Y. *et al.* (2012) SOAPdenovo2: an empirically improved memory-efficient short-read de novo assembler. *Gigascience*, **1**, 18.
16. Yang,Y. and Yooseph,S. (2013) SPA: a short peptide assembler for metagenomic data. *Nucleic Acids Res.*, **41**, e91.
17. Yang,Y., Zhong,C. and Yooseph,S. (2015) SFA-SPA: a suffix array based short peptide assembler for metagenomic data. *Bioinformatics*, **31**, 1833–1835.
18. Peng,G., Ji,P. and Zhao,F. (2016) A novel codon-based de Bruijn graph algorithm for gene construction from unassembled transcriptomes. *Genome Biol.*, **17**, 232.
19. Bernardi,G. and Bernardi,G. (1986) Compositional constraints and genome evolution. *J Mol. Evol.*, **24**, 1–11.
20. Koonin,E.V. and Wolf,Y.I. (2008) Genomics of bacteria and archaea: the emerging dynamic view of the prokaryotic world. *Nucleic Acids Res.*, **36**, 6688–6719.
21. Min,X.J., Butler,G., Storms,R. and Tsang,A. (2005) OrfPredictor: predicting protein-coding regions in EST-derived sequences. *Nucleic Acids Res.*, **33**, W677–W680.
22. Pevzner,P.A., Tang,H. and Waterman,M.S. (2001) An Eulerian path approach to DNA fragment assembly. *Proc. Natl. Acad. Sci. U.S.A.*, **98**, 9748–9753.
23. Buchfink,B., Xie,C. and Huson,D.H. (2015) Fast and sensitive protein alignment using DIAMOND. *Nat. Methods*, **12**, 59–60.
24. Qi,J., Zhao,F., Buboltz,A. and Schuster,S.C. (2010) inGAP: an integrated next-generation genome analysis pipeline. *Bioinformatics*, **26**, 127–129.
25. Qi,J. and Zhao,F. (2011) inGAP-sv: a novel scheme to identify and visualize structural variation from paired end mapping data. *Nucleic Acids Res.*, **39**, W567–W575.
26. Fu,L., Niu,B., Zhu,Z., Wu,S. and Li,W. (2012) CD-HIT: accelerated for clustering the next-generation sequencing data. *Bioinformatics*, **28**, 3150–3152.
27. Shakya,M., Quince,C., Campbell,J.H., Yang,Z.M.K., Schadt,C.W. and Podar,M. (2013) Comparative metagenomic and rRNA microbial diversity characterization using archaeal and bacterial synthetic communities. *Environ. Microbiol.*, **15**, 1882–1899.
28. Bolger,A.M., Lohse,M. and Usadel,B. (2014) Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics*, **30**, 2114–2120.
29. Li,H. and Durbin,R. (2009) Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics*, **25**, 1754–1760.
30. Chain,P.S.G., Grafham,D.V., Fulton,R.S., FitzGerald,M.G., Hostetler,J., Muzny,D., Ali,J., Birren,B., Bruce,D.C., Buhay,C. *et al.* (2009) Genome project standards in a new era of sequencing. *Science*, **326**, 236–237.
31. Human Microbiome Jumpstart Reference Strains Consortium, Nelson,K.E., Weinstock,G.M., Highlander,S.K., Worley,K.C., Creasy,H.H., Wortman,J.R., Rusch,D.B., Mitreva,M., Sodergren,E. *et al.* (2010) A catalog of reference genomes from the human microbiome. *Science*, **328**, 994–999.
32. Ribeiro,F.J., Przybylski,D., Yin,S., Sharpe,T., Gnerre,S., Abouelleil,A., Berlin,A.M., Montmayeur,A., Shea,T.P., Walker,B.J. *et al.* (2012) Finished bacterial genomes from shotgun sequence data. *Genome Res.*, **22**, 2270–2277.
33. Giannoukos,G., Ciulla,D.M., Huang,K., Haas,B.J., Izard,J., Levin,J.Z., Livny,J., Earl,A.M., Gevers,D., Ward,D.V. *et al.* (2012) Efficient and robust RNA-seq process for cultured bacteria and complex community transcriptomes. *Genome Biol.*, **13**, R23.
34. Qin,J.J., Li,R.Q., Raes,J., Arumugam,M., Burgdorf,K.S., Manichanh,C., Nielsen,T., Pons,N., Levenez,F., Yamada,T. *et al.* (2010) A human gut microbial gene catalogue established by metagenomic sequencing. *Nature*, **464**, 59–65.
35. Ji,P., Zhang,Y., Wang,J. and Zhao,F. (2017) MetaSort untangles metagenome assembly by reducing microbial community complexity. *Nat. Commun.*, **8**, 14306.
36. Parks,D.H., Rinke,C., Chuvochina,M., Chaumeil,P.A., Woodcroft,B.J., Evans,P.N., Hugenholtz,P. and Tyson,G.W. (2017) Recovery of nearly 8,000 metagenome-assembled genomes substantially expands the tree of life. *Nat. Microbiol*, **2**, 1533–1542.