Computational and Structural Biotechnology Journal 21 (2023) 444–451

Contents lists available at ScienceDirect

Computational and Structural Biotechnology Journal

journal homepage: www.elsevier.com/locate/csbj

Reference-based read clustering improves the *de novo* genome assembly of microbial strains

Mikang Sim [a,1], Jongin Lee [a,1], Daehong Kwon [a], Daehwan Lee [a], Nayoung Park [a], Suyeon Wy [a], Younhee Ko [b], Jaebum Kim [a,*]

[a] Department of Biomedical Science and Engineering, Konkuk University, Seoul 05029, Republic of Korea
[b] Division of Biomedical Engineering, Hankuk University of Foreign Studies, Gyeonggi-do 17035, Republic of Korea

ARTICLE INFO

Article history:
Received 10 August 2022
Received in revised form 17 December 2022
Accepted 19 December 2022
Available online 21 December 2022

Keywords:
Next-generation sequencing
Read clustering
Reference-based
Microbial genome
Genome assembly

ABSTRACT

Constructing accurate microbial genome assemblies is necessary to understand genetic diversity in microbial genomes and its functional consequences. However, it still remains as a challenging task especially when only short-read sequencing technologies are used. Here, we present a new read-clustering algorithm, called RBRC, for improving *de novo* microbial genome assembly, by accurately estimating read proximity using multiple reference genomes. The performance of RBRC was confirmed by simulation-based evaluation in terms of assembly contiguity and the number of misassemblies, and was successfully applied to existing fungal and bacterial genomes by improving the quality of the assemblies without using additional sequencing data. RBRC is a very useful read-clustering algorithm that can be used (i) for generating high-quality genome assemblies of microbial strains when genome assemblies of related strains are available, and (ii) for upgrading existing microbial genome assemblies when the generation of additional sequencing data, such as long reads, is difficult.

1. Introduction

Using *de novo* genome assembly to concatenate next-generation sequencing (NGS) reads and recover chromosome-level sequences is a difficult task [1]. Although short-read sequence data have been widely used for *de novo* genome assembly, their short length makes the task more difficult [2]. To alleviate this problem, long-read sequencing and assembly technologies have recently been developed [3,4]. However, many microbial genomes have still been sequenced and assembled based on short-read sequences because of their low sequencing error and cost [5–10]. Therefore, numerous studies have researched the effective use of short-read sequences for microbial genome assembly [11–13]. One such method involves clustering read sequences sharing similar features and utilizing the information extracted from the read groups for genome assembly.

Sequence clustering has been extensively used for various types of sequencing data to (i) remove redundant sequences in predicted gene sequences [14–16], (ii) group transcriptome and RNA sequencing [17,18], (iii) detect groups of functionally related protein sequences [19–21], and (iv) group motif sequences [22,23]. These sequence-clustering methods commonly rely on the similarity of sequences, which can be easily found using sequence alignment [24]. This makes the existing clustering methods unsuitable for *de novo* genome assembly, in which the read proximity in a chromosome can play a critical role in the reconstruction of genomes. If read sequences are clustered based on sequence similarity, read sequences that are not physically close can be placed in the same cluster. This is because of the existence of repetitive sequences in many different genome positions, which is especially problematic in short-read sequences. On the other hand, when the sequence similarity-based clustering method is used, read sequences that are physically close in a chromosome cannot be grouped in the same cluster if the level of sequence similarity among them is low. In metagenome studies, sequence clustering is also used to obtain meaningful groups of NGS read sequences from the same species or taxonomic unit using marker genes, tetranucleotide frequencies, and contig or scaffold coverage [25–29]. However, since these approaches utilize various sequence features from different species in

* Corresponding author.
E-mail address: jbkim@konkuk.ac.kr (J. Kim).
[1] These authors contributed equally to this work.

https://doi.org/10.1016/j.csbj.2022.12.032

metagenome samples, it is difficult to directly apply these approaches to the *de novo* assembly of a single genome.

Recent advances in sequencing technologies and assembly algorithms have enabled the accumulation of high-quality genome assemblies for many species, including microbes [30–37]. One of the important features of a microbial genome is that there are multiple strains in the same species [38]. In this case, the high-quality genome assemblies of some strains can be used as valuable resources for assembling the genomes of other strains in the same species. For example, it is possible to ensure that read clusters are especially useful for genome assembly by estimating the proximity of reads in a target strain based on their relative positions in the genome assemblies of other strains of the same microbial species. Then, the read clusters can be used as additional information to produce better genome assemblies. This idea was used to generate the *de novo* genome assembly of a newly sequenced *Arabidopsis thaliana* strain using a single reference strain of the same species [39,40]. The main drawback of these approaches is the use of a single reference genome. Genome rearrangements can occur among microbial strains, even in those of the same species [41]. Therefore, the use of a single reference genome has a limited potential for the accurate assembly of the target strain genome.

Here, we present a new NGS read-clustering algorithm (RBRC) for the *de novo* genome assembly of microbes that utilizes the genome assemblies of multiple closely related microbial strains, called references. Given the NGS paired-end (PE) reads of a target strain and the genome sequences of the references, RBRC constructs syntenic regions among multiple reference genomes and groups the reads that are mapped to the same syntenic regions as members of the same cluster. For unclustered reads, the distance between a pair of reads in the target strain is estimated based on their distance in the reference genome, and this distance is then used to cluster them. Then, the generated clusters are further merged by using the cluster membership of the two end-reads of the PE reads. The performance of our clustering algorithm was evaluated using the simulated NGS reads of a yeast genome. RBRC was also successfully employed to improve the existing fungal and bacterial genome assemblies using the same sequencing data used to generate the original genome assemblies. This clearly shows that this method could also contribute to increasing the quality of the existing genome assemblies of many microbes without requiring additional expensive and time-consuming sequencing data. This clustering algorithm could be very useful for the *de novo* genome assembly of microbes, and furthermore, it will become more useful as more high-quality genome assemblies of various species are accumulated. RBRC and *de novo* assembly programs based on RBRC are available at https://github.com/jkimlab/RBRC.

2. Materials and methods

2.1. Reference-based read-clustering algorithm

This study presents a reference-based read-clustering (RBRC) algorithm for NGS PE reads that utilizes a set of reference genome sequences (Fig. 1). Once whole-genome sequence alignments among reference genomes are generated, syntenic regions among the reference genomes are constructed using the alignments. The reads mapped to the same syntenic regions are then grouped and form a single cluster. These clusters are called synteny-based clusters. For unclustered reads, the distance between two reads in the target strain is estimated by using a weighted sum of their distances in the reference genome sequences. The unclustered reads are then grouped based on those estimated distances, which leads to the generation of distance-based clusters. Finally, the two sets of clusters (synteny-based and distance-based clusters) are merged using the pair information of the PE reads. The details of each step are presented below.

2.2. Pre-processing

For the given NGS PE read sequences of a target strain, low-quality sequences, adapter sequences, and unpaired reads are filtered out using Trimmomatic (version 0.39) [42]. The remaining PE reads are mapped to each given reference genome sequence using BWA-MEM (version 0.7.17) [43], and alignment filtering is performed using SAMtools view (version 1.9) [44] with -f 0x03 -F 0xF00 options. Based on the mapping results, the weight of the reference genomes, defined as the percentage of properly mapped PE reads, is calculated and used to filter out some of the reference genomes (if any) with a low weight (default: 80%). The reference with the highest weight is considered as a leading reference and used as the center in the synteny-based clustering step described in the following subsection.

2.3. Synteny-based clustering

To find conserved genomic regions among all given reference genomes, pairwise whole-genome sequence alignments between the leading reference and each of the other references are generated using lastz (version 1.04.00) [45]; these are then used to create syntenic regions among the reference genomes based on a given resolution by the inferCars program [46]. The constructed syntenic regions are further refined based on the physical read coverage to reduce false positives in the clusters. The physical read coverage is calculated from the mapped reads to the leading reference genome sequence. Specifically, to extract the primary alignment of the properly mapped read pairs, SAMtools (version 1.9) [44] is used to filter alignments with -f 0x03 -F 0xF00 options, and BEDtools (version v2.17.0) [47] is used to calculate the physical read coverage. Syntenic regions are broken at regions with small physical read coverage, and then the regions shorter than the resolution are filtered out. Finally, based on the mapping information of the leading reference, all single reads mapped to the same syntenic regions are grouped into the same cluster.

2.4. Distance-based clustering

All unclustered reads in the previous step are further clustered based on the distances between read pairs in non-syntenic regions in the genomes of multiple reference strains. This clustering step consists of four sub-steps: (i) extracting mapping information of the unclustered reads to all references, (ii) calculating the distance between two reads in each reference genome, (iii) estimating the distance between two reads in a target genome by using their distances in reference genomes, and (iv) clustering the reads based on the estimated distances.

First, for all reference genome sequences, the mapping positions of all unclustered reads are extracted by SAMtools (version 1.9) with the -F 0xF00 option from the read-mapping data generated in the pre-processing step. Next, for a pair of reads mapped to all references as well as the same chromosome in each of the references, the distances between the read pairs in each of the references are calculated. Here, any two reads can make a pair for the distance calculation, regardless of whether they are the two end reads in a PE read or not. For each pair of reads, their distance in the target strain is estimated by using the weighted sum of distances from all references and recorded in a final distance matrix if it is smaller than the cutoff. Here, the weight calculated in the pre-processing step is used as the weight of a reference. Finally, based on estimated distance between unclustered reads, they are clustered using an in-house Perl

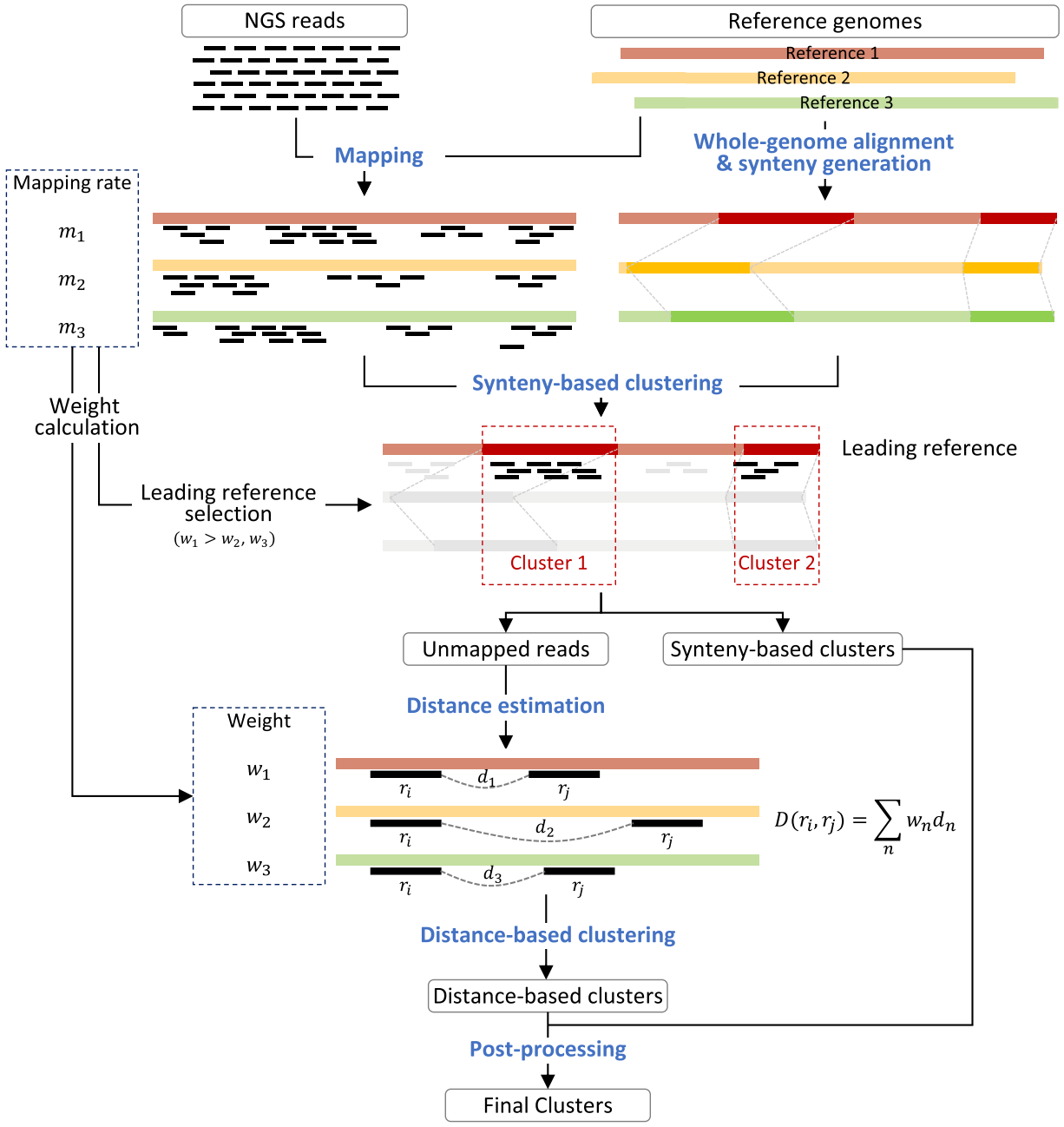

Fig. 1. Overview of reference-based read-clustering (RBRC) algorithm. The reference-based read-clustering algorithm consists of four steps divided by dashed rectangles. (i) In the pre-processing step, low-quality reads are trimmed, and the remaining reads are mapped to reference genomes. The mapping information is used to calculate the weight of references and to choose a leading reference. (ii) In the synteny-based clustering step, syntenies among references are created, and reads that are mapped to the same synteny are assigned to the same cluster. (iii) In the distance-based clustering step, the distance between two unclustered reads in a target strain is estimated based on their distance in the references and used to create additional read clusters. (iv) Finally, using the paired information of reads, the existing clusters are merged in the post-processing step.

script implementing DBSCAN [48], which is a density-based spatial clustering algorithm.

2.5. Post-processing

The read clusters generated in the previous two steps are further merged using the PE read information. First, the number of links between two different clusters is calculated. A link between two different clusters is defined when each end-read of a PE read is assigned to a different cluster. When a pair of clusters has more links than the cutoff, they are merged into a single cluster. Next, if one end of a PE read is not clustered but its pair is in a certain cluster, the unclustered end is assigned to the same cluster as its pair. Finally, if two end-reads of the same PE read are assigned to different clusters, they are discarded from the assigned clusters.

2.6. Simulation of read sequences for a fungal and bacterial genome

To assess the performance of the clustering algorithm, the read sequences of the genomes of one of fungal species (yeast, *S. cerevisiae*) and one of bacterial species (*E. coli*) were simulated. Specifically, complete genome assemblies of five yeast strains (S288C, BY4742, ySR128, ySR127, and KSD-Yc) were downloaded from the NCBI website (Table S1). Using chromosome sequences of the yeast strain S288C, NGS PE reads were simulated using ART [49] with the options of Illumina HiSeq 2500 platform, 101 bp read

length, 500 bp mean fragment length, and four different sequencing coverages (5x, 10x, 30x, and 50x). The other four yeast strains were used as references for clustering.

For simulating the read sequences of the *E. coli* genome, complete genome assemblies of five *E. coli* strains (ST540, ST2747, YD786, MG1655, and O157) were downloaded from the NCBI website (Table S1). The chromosome sequence of the strain ST540 was used to simulate read sequences using ART with the same options including the simulation coverage as in the yeast genome. The other four *E. coli* strains were used as references for clustering.

2.7. Performance assessment of read clustering

The quality of generated clusters was assessed based on entropy [50], which can measure the purity of clustered reads in each cluster. The primary goal of the read clustering is to group reads originated from the same genomic region into the same cluster. Therefore, the quality of a specific read cluster can be measured by examining the number of reads originated from the same genomic region corresponding to the cluster but classified into other clusters. By using the origin of all reads generated during simulation, the corresponding genomic region of a cluster was determined, which was defined as a region from the left-most (the smallest coordinate) read to the right-most (the largest coordinate) read in the cluster. Once the genomic region of a cluster was defined, the cluster membership information of all reads in the genomic region was used to calculate the entropy of the cluster using an in-house Perl script. Entropy for a cluster C was defined as:

$$H(C) = -\sum_{i=1}^{N_c} \frac{n_i}{N_r} \log_2 \frac{n_i}{N_r} \quad (1)$$

where N_r is the total number of reads in the corresponding genomic region of the cluster C, n_i is the number of reads classified into the ith cluster, and N_c is the number of different clusters to which the reads in the genomic region belong. If all reads in the genomic region of the cluster C are classified into the cluster C, then the entropy is 0 because $N_c = 1$ and $N_r = n_i$.

In this assessment, the USEARCH program [14], a clustering program based only on sequence similarity, was additionally used, and its clustering results were compared with the read clusters generated by RBRC. Values of 0.7, 0.8, and 0.9 were used for the similarity parameter of the USEARCH program. In the case of RBRC, used parameter values are shown in Table S2.

2.8. De novo genome assembly using read clusters

To investigate the usability of read clusters generated by RBRC, a *de novo* genome assembly approach was developed using the read clusters constructed by the algorithm with the same parameter values used in the simulation-based assessment. An overview of this approach is illustrated in Fig. S1. For input PE reads and reference genome sequences, read clusters were constructed by RBRC, and contigs were generated for each cluster by the assembly program SPAdes (version 3.15.4) with –careful option [51]. Finally, contig sequences from all clusters were further assembled using SPAdes (version 3.15.4) with the –careful and –nanopore option using all of the qualified input PE reads that passed the filtering step in the clustering algorithm. In this step, SPAdes performed a hybrid assembly by treating the contig sequences as long-read sequences.

The simulated PE datasets described in the previous subsection 2.2 were used to evaluate our assembly approach. The assemblies generated by our approach were compared with those created by the SPAdes program (version 3.15.4) with the –careful option using only the PE reads (not using the read cluster information). For calculating evaluation statistics, QUAST (version 5.0.2) [52] was used with default options and the genome assemblies of the yeast strain S288C and *E. coli* strain ST540 were used as the true assemblies for yeast and *E. coli* respectively.

2.9. Applications of de novo genome assembly

The *de novo* genome assembly approach was employed for the yeast strain Hm-1 (accession number: GCA_003569725.1) [53] which has both NGS PE reads and a scaffold genome assembly. In this application, our assembly approach was used to construct a *de novo* genome assembly of the strain Hm-1 using the NGS PE reads and the reference genomes of the strains ySR128 and ySR127, and the result was compared with the downloaded assembly. The used parameter values for read clustering are described in Table S2.

Additionally, the contig assemblies and corresponding PE reads of two bacteria (*Pseudomonas syringae pv. syringae* strain 2340 and the *Lactobacillus plantarum* strain IYO1511) were downloaded (accession number: GCA_001535725.1 and GCA_011170185.1 respectively) [36,54] and used similarly for the above yeast Hm-1 strain. The strains B728a and CFBP4215 were used as references for strain 2340 and the strains TMW 1.1478 and LB1–2 were used as references for the strain IYO1511. All genome assemblies were obtained from NCBI (accession numbers are listed in Table S1).

The quality of assemblies was assessed by assembly contiguity using N10 to N90 values, and completeness through the scores calculated using BUSCO (version 5.3.2) [55]. In this comparison, short sequences less than 500 bp were ignored, and QUAST was not used because of the absence of a true assembly.

3. Results

3.1. Simulated read-based evaluation of the RBRC clusters

Based on the genome sequences of the yeast strain S288C, a total of 597554, 1195090, 3585270, and 5975450, PE reads were simulated using four different sequencing depths, 5x, 10x, 30x, and 50x, respectively. The simulated reads were then clustered by RBRC using the genome assemblies of four additional yeast strains, BY4742, ySR128, ySR127, and KSD-Yc. In this clustering, a different number of references (from two to four) was used by weighting them based on the rate of properly mapped reads (Materials and methods, Table S3). RBRC produced clusters including more than 99% of the simulated reads in all different settings, and the number of clusters numbers ranged from 135 to 1622 depending on the sequencing depth and the number of used references (Table 1). The sequencing depth was not clearly correlated with the number of clusters, whereas the number of clusters increased as the number of references increased. The simulated read datasets were also clustered by

Table 1
Statistics of clusters generated by RBRC using simulated read datasets of the yeast strain S288C.

Sequencing depth	No. of references	No. of clustered reads	% of clustered reads	No. of final clusters
5x	2	596,576	99.84	508
	3	596,550	99.83	647
	4	594,124	99.43	1,622
10x	2	1,193,862	99.90	135
	3	1,193,460	99.86	148
	4	1,189,390	99.52	357
30x	2	3,582,470	99.92	155
	3	3,581,808	99.90	193
	4	3,573,336	99.67	739
50x	2	5,970,964	99.92	149
	3	5,970,182	99.91	224
	4	5,957,698	99.70	1,033

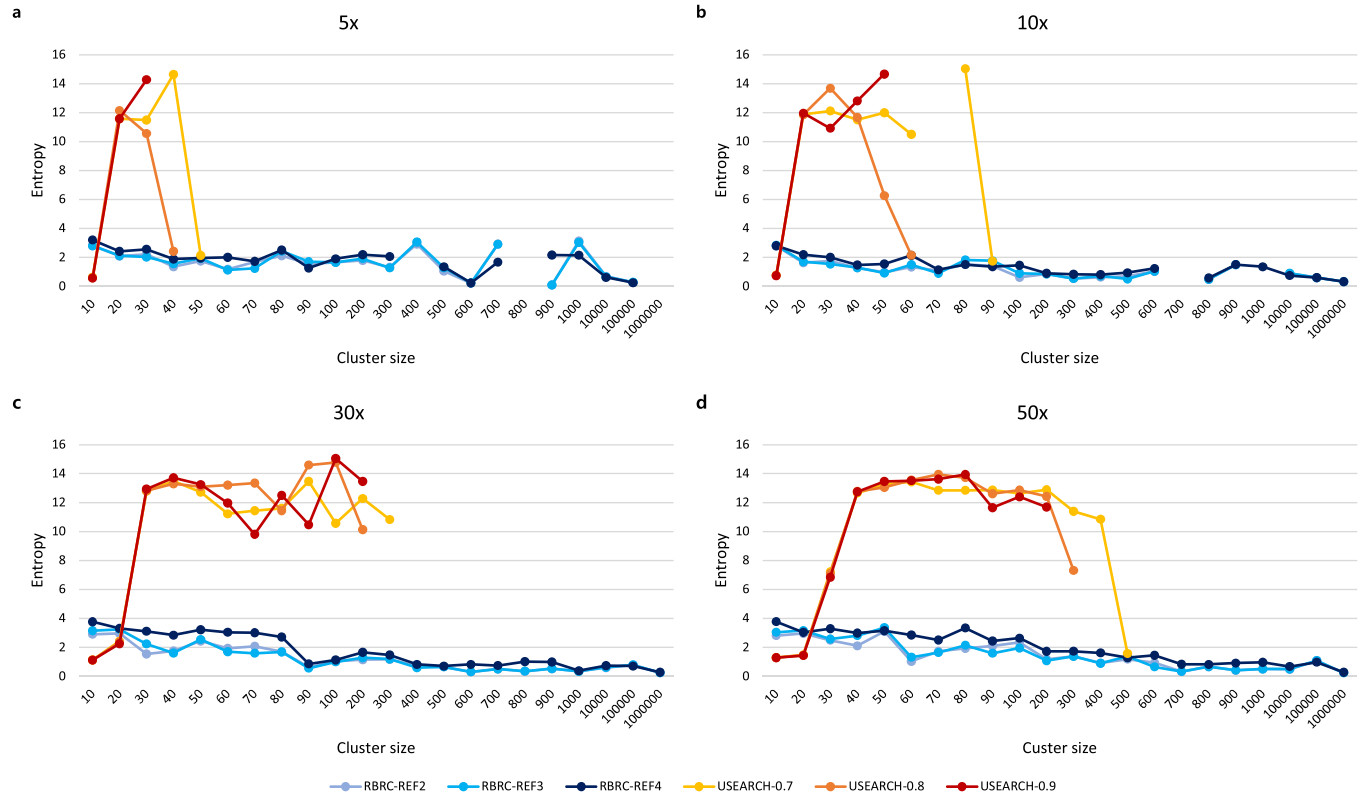

Fig. 2. Evaluation of read clusters generated from the simulated reads of the yeast strain S288C. The entropy of each cluster was calculated, and the average entropy (y-axis) of clusters with a similar size (x-axis) was plotted for four datasets with different sequencing depth, 5x (a), 10x (b), 30x (c), and 50x (d). In the case of RBRC, the number after "REF" represents the number of references used for read clustering in the legend. In the case of USEARCH, the similarity parameter cutoff is displayed as the number after the "-" symbol in the legend.

USEARCH, which only considers the similarity of read sequences (Materials and methods). In this case, a very larger number of clusters (from 160044 to 594187) was created, and a varying number of reads (around from 70% to 100%) were included in those clusters in different settings (Table S4). In addition, a greater number of reads was clustered as the sequencing depth increased.

The primary goal of our clustering is to group reads from the same genomic regions in order to make the clustered reads for local *de novo* genome assembly. Therefore, the quality of generated clusters was assessed by calculating and comparing entropy (Materials and methods). Entropy is a natural measure for assessing the purity of reads, in terms of cluster membership. Good cluster is one that contain all reads from a specific genomic region. One limitation of entropy is the tendency that a smaller cluster is advantageous to have lower entropy. In an extreme case, if there is only one read in a cluster, its entropy is zero. Therefore, for fair comparison, an average entropy of clusters with similar size was calculated and compared. As shown in Fig. 2, the entropy scores of the RBRC clusters were clearly lower than those of USEARCH clusters in all sequencing depths except for clusters with a very small number of reads. In addition, the size of a cluster did not play a critical role in reducing the purity of RBRC clusters, which was improved as a larger number of reads are included in a cluster in all sequencing depths. These results indicate that reference-based read clustering is highly effective in grouping reads generated from similar genomic regions, which cannot be obtained with read sequence similarity alone.

A similar evaluation was also conducted with simulated read sequences from the bacterial genome sequence of the *E. coli* strain ST540 (Materials and methods). The statistics of read clusters by RBRC and USEARCH are available in Tables S5-S7, and the result of quality assessment of clusters, which has a similar pattern as the one of the yeast datasets, are shown in Fig. S2.

3.2. De novo genome assembly-based evaluation of the RBRC clusters

A *de novo* assembly of the yeast strain S288C was constructed (Materials and methods, Fig. S1) with the read clusters generated by RBRC based on two references. The quality of our genome assembly was then compared with another genome assembly generated without the read cluster information (Materials and methods).

For all cases of various sequencing depths, the use of read clusters showed a decrease in the number of contigs (Fig. 3a) with a slight increase in the total contig length (Table S8). The decrease became more apparent when greater sequencing depth was used. For example, in the case of the 50x dataset, the number of contigs was reduced by more than 17% (from 193 to 160). Moreover, it shows that the read cluster information is critical to produce longer contigs (Fig. S3). For example, N50 increased from 177 Kbp to 254 Kbp in the case of the 50x dataset when the read cluster information is incorporated in the assembly step. This gap became more apparent as longer contigs in each assembly were compared (toward N10 values in Figs. 3b and S3).

In addition, assembly contiguity increased remarkably with increasing sequencing depth when the read cluster information was used (Fig. 3b). The improved assembly contiguity affected negatively on the number of misassemblies when 5x read dataset was used (Fig. 3c). However, the effect became almost negligible when a deeper sequencing dataset was used. There was a dramatic increase in the overall assembly quality (NA50) that accounts for both assembly contiguity and the number of misassemblies when the read cluster information was used, especially with the use of a deeper sequencing dataset (Fig. 3d). Similar patterns were also observed for *E. coli* read dataset (Table S9). Overall, these results show that the RBRC read clusters are very useful to improve the quality of the *de*

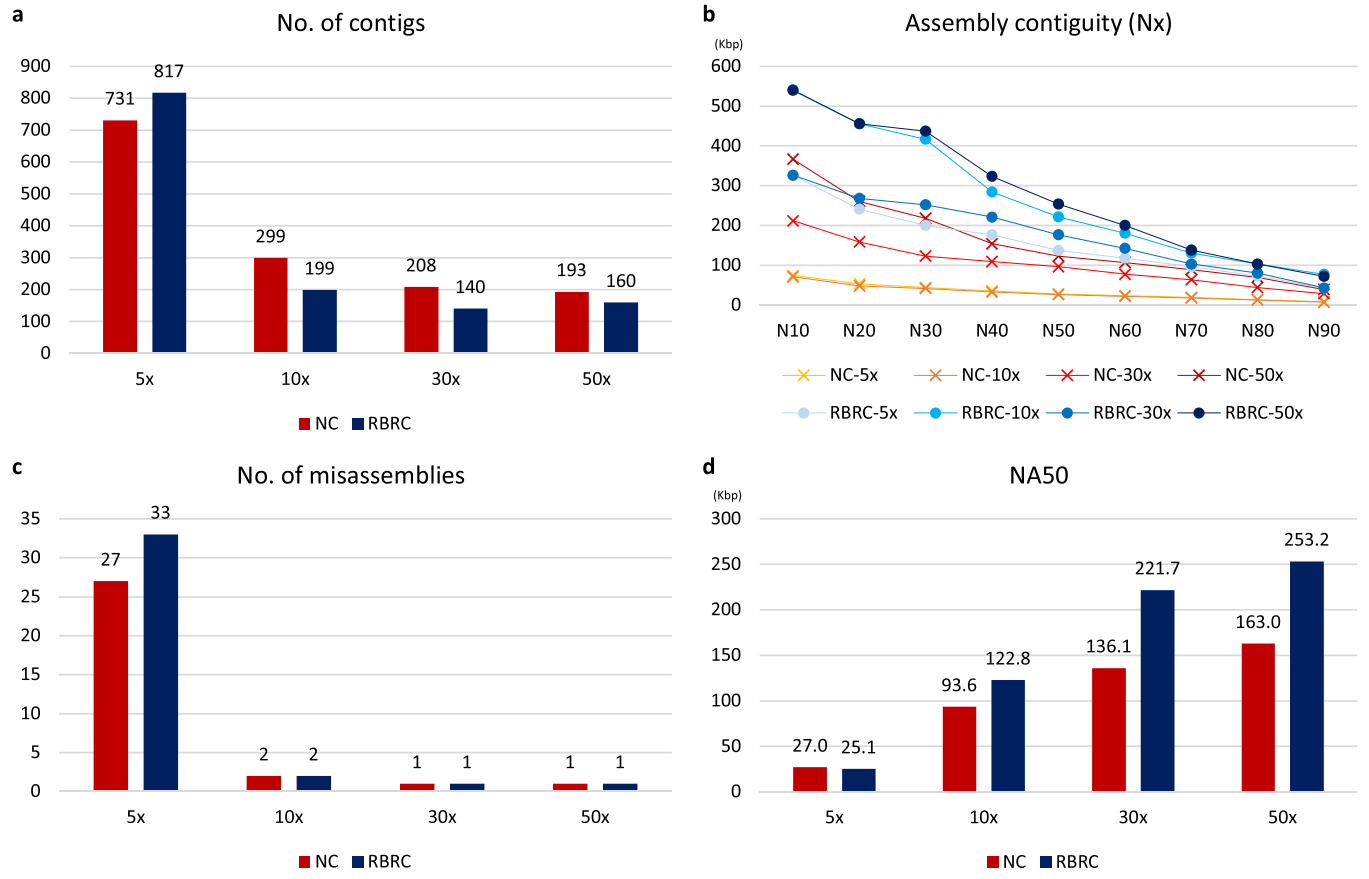

Fig. 3. Evaluation of assemblies constructed from simulated read datasets of the yeast strain S288C with two references. Two sets of contigs were generated (i) using the RBRC algorithm ("RBRC" in legend) and (ii) without using the read cluster information ("NC" in legend); they were then assessed according to the number of contigs (a), assembly contiguity (b), the number of misassemblies (c), and NA50 (d).

novo genome assembly for newly sequenced microbial strain genomes even with low sequencing depth.

3.3. *De novo* genome assembly of additional microbial strains using the RBRC clusters

The RBRC algorithm was applied to *de novo* genome assembly for additional one fungal strain and two bacterial strains with real PE sequencing data (Materials and methods). For the *Saccharomyces cerevisiae* strain Hm-1, the strains ySR128 and ySR127 were used as references (Table S10). Among the total of 4100970 qualified PE reads, 98.47% of reads were grouped into 489 clusters (Table 2). The read clusters were then used to generate the genome assembly of the Hm-1 strain, and its quality was compared with the existing genome assembly of the same strain downloaded from NCBI (Materials and methods). When the RBRC clusters are used, the contiguity and completeness of the assembly were increased compared with the NCBI assembly (Table S11). Although the number of scaffolds increased from 225 to 258, the overall contig lengths including the largest one increased, and total gap size was decreased from 6344 to 2161. In terms of assembly contiguity, the RBRC clusters were very effective to increase the length of assembled sequences in all different length range compared with the NCBI assembly, and the improvement became larger especially when longer sequences were compared (Fig. 4a). Furthermore, such improvement of assembly contiguity was achieved without scarifying the assembly completeness (Fig. 4b). These patterns were also observed when the assembly approach using the RBRC clusters was applied for the bacterial strains, *Pseudomonas syringae pv. syringae* strain 2340 and *Lactobacillus plantarum* strain IYO1511 (Tables S12 and S13).

4. Discussion

We present a novel read-clustering algorithm for microbial strains called RBRC, by integrating the read proximity information from the genome sequences of the strains of the same species. RBRC does not directly use the physical sequence overlap information between two reads. Instead, it estimates and uses the proximity information of two reads based on (i) the organization of the reads in conserved genomic regions among strains of the same species and (ii) their predicted distance, obtained by considering their distance in the genomes of other strains. Proximate reads sequenced from close genomic regions could be grouped using RBRC, even though the reads in the same cluster did not have physical sequence overlap.

Table 2
Clustering statistics for sequencing reads of microbial species.

Statistics	*Saccharomyces cerevisiae* strain Hm-1	*Pseudomonas syringae pv.syringae* strain 2340	*Lactobacillus plantarum strain* IYO1511
No. of total reads	4,100,970	5,664,928	2,668,446
No. of clustered reads	4,038,368	5,076,964	2,510,548
% of clustered reads	98.47	89.62	94.08
No. of clusters	489	991	235

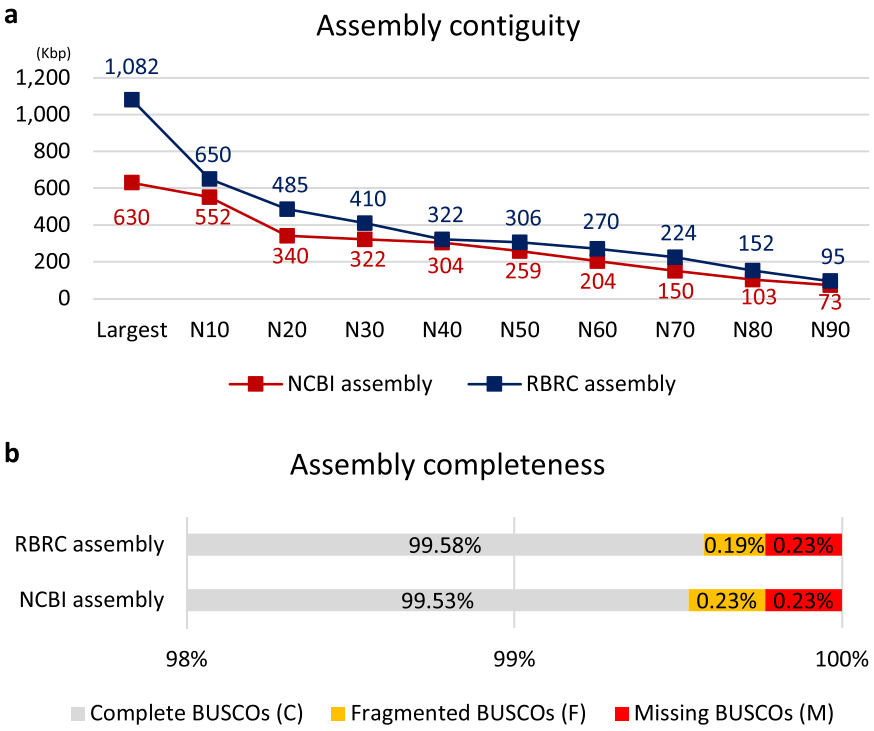

Fig. 4. Evaluation of assemblies of the Saccharomyces cerevisiae strain Hm-1 generated using real sequencing data. Using real sequencing data, contigs were generated by the assembly approach using the RBRC algorithm ("RBRC assembly" in the legend) and compared with the one downloaded from NCBI ("NCBI assembly" in the legend). The quality of the assemblies is shown in terms of (a) assembly contiguity and (b) assembly completeness based on the BUSCO scores.

The quality of predicted read clusters was measured using entropy. Since the entropy has a bias towards small-sized clusters, the predicted read clusters were partitioned into multiple bins based on their size (the number of reads in a cluster), and the entropy of clusters with similar size was compared. As expected, clustering using only sequence similarity failed to produce clusters with large number of reads, which can be localized together into the same genomic region. For example, more than 98% of read clusters generated by USEARCH with similarity parameter 0.7 for 50x yeast dataset consisted of less than 20 reads, and there was no cluster contained more than 500 reads (Fig. 2d). Moreover, their entropy was about three-fold larger than RBRC clusters in most of cluster sizes. Therefore, read clusters generated by the similarity-based clustering method can only provide limited information for genome assembly. On the other hand, the read clusters generated by RBRC show the overall low entropy despite their large size, and they can be more effectively used to construct a more complete and contiguous genome assembly (Figs. 3 and 4).

To maximize the usability of the read clusters generated by RBRC, a genome assembly approach based on RBRC read clusters was developed. This approach is similar with a hybrid genome assembler in the sense that longer read sequences are generated from the given short read sequences and they are used together to construct a genome assembly. However, our approach does not require long read sequences as input because they are generated internally. This is a cost-effective difference of our approach compared with general hybrid genome assemblers. When the assembly approach was applied to the simulated read datasets, it was able to improve the quality of genome assembly. For example, when the 50x yeast dataset was used, longer scaffold sequences were produced without containing misassembled regions compared with genome assemblies generated without using the RBRC read clusters (Fig. 3). The improvement gap in terms of assembly contiguity and misassembly became larger when the sequencing depth of the used datasets became greater. The assembly approach with RBRC read clusters was also successfully applied to *de novo* assembly for fungal and bacterial sequencing reads (Fig. 4, Tables S11-S13). For all sequencing datasets, RBRC showed increased assembly contiguity with high assembly quality scores measured by BUSCO. Our results indicate that RBRC can be used to upgrade many previously sequenced and assembled microbial genomes without additional sequencing data.

5. Conclusions

RBRC is a very useful read-clustering algorithm that can be easily used to generate high-quality genome assemblies of microbial strains when genome assemblies of related strains are available. It can be used not only to create high-quality *de novo* genome assemblies of microbial strains using only short-read sequencing data, but also to upgrade existing genome assemblies constructed based on short reads due to the absence of long reads. The potential contribution of this clustering algorithm will continue to increase as more high-quality genome assemblies of various microbial strains are accumulated.

Funding information

This work was supported by the Konkuk University Researcher Fund in 2021 and the Ministry of Science and ICT of Korea Grant [NRF-2014M3C9A3063544, NRF-2021M3H9A2097134, NRF-2022R1F1A1065159] to JBK, and the Ministry of Science and ICT of Korea Grant [NRF-2020R1F1A1069672] and Hankuk University of Foreign Studies Research Fund to YHK.

CRediT authorship contribution statement

Mikang Sim: Methodology, Software, Validation, Visualization, Formal analysis, Writing – original draft, Writing – review & editing. **Jongin Lee**: Methodology, Software. **Daehong Kwon**: Software. **Daehwan Lee**: Software. **Nayoung Park**: Validation. **Suyeon Wy**:

Validation. **Younhee Ko**: Formal analysis. **Jaebum Kim**: Conceptualization, Methodology, Formal analysis, Writing – original draft, Writing – review & editing.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

Not applicable.

Data Statement

All source codes and example data are available at https://github.com/jkimlab/RBRC.

Consent for publication

Not applicable.

Appendix A. Supporting information

Supplementary data associated with this article can be found in the online version at doi:10.1016/j.csbj.2022.12.032.

References

[1] Myers EW, Sutton GG, Delcher AL, Dew IM, Fasulo DP, et al. A whole-genome assembly of Drosophila. Science 2000;287(5461):2196–204.
[2] Miller JR, Koren S, Sutton G. Assembly algorithms for next-generation sequencing data. Genomics 2010;95(6):315–27.
[3] Rhoads A, Au KF. PacBio sequencing and its applications. Genom Proteom Bioinform 2015;13(5):278–89.
[4] Lu H, Giordano F, Ning Z. Oxford nanopore MinION sequencing and genome assembly. Genom Proteom Bioinform 2016;14(5):265–79.
[5] Wit M., Leng Y., Du Y., Cegielko M., Jabłońska E. et al. (2020) Genome sequence resources for the maize pathogen Fusarium temperatum isolated in Poland. Molecular Plant-Microbe Interactions (ja).
[6] Hamdy AA, Esawy MA, Elattal NA, Amin MA, Ali AE, et al. Complete genome sequence and comparative analysis of two potential probiotics Bacillus subtilis isolated from honey and honeybee microbiomes. J Genet Eng Biotechnol 2020;18(1):1–8.
[7] Ibrahim M, Yar AM, Zaman G, Yan C, Khurshid M, et al. Genome sequence and analysis of Mycobacterium tuberculosis strain SWLPK. J Glob Antimicrob Resist 2018;13:211–3.
[8] Botelho J, Grosso F, Peixe L. Unravelling the genome of a Pseudomonas aeruginosa isolate belonging to the high-risk clone ST235 reveals an integrative conjugative element housing a blaGES-6 carbapenemase. J Antimicrob Chemother 2018;73(1):77–83.
[9] Melo LC, Haenni M, Saras E, Cerdeira L, Moura Q, et al. Genomic characterisation of a multidrug-resistant TEM-52b extended-spectrum β-lactamase-positive Escherichia coli ST219 isolated from a cat in France. J Glob Antimicrob Resist 2019;18:223–4.
[10] Battu L, Ulaganathan K. Whole genome sequencing and identification of host-interactive genes in the rice endophytic Leifsonia sp. ku-ls. Funct Integr Genomics 2020;20(2):237–43.
[11] Page AJ, De Silva N, Hunt M, Quail MA, Parkhill J, et al. Robust high-throughput prokaryote *de novo* assembly and improvement pipeline for Illumina data. Microb Genom 2016;2(8):e000083.
[12] Souvorov A, Agarwala R, Lipman DJ. SKESA: strategic k-mer extension for scrupulous assemblies. Genome Biol 2018;19(1):153.
[13] Al-Okaily AA. HGA: de novo genome assembly method for bacterial genomes using high coverage short sequencing reads. BMC Genomics 2016;17:193.
[14] Edgar RC. Search and clustering orders of magnitude faster than BLAST. Bioinformatics 2010;26(19):2460–1.
[15] Fu L, Niu B, Zhu Z, Wu S, Li W. CD-HIT: accelerated for clustering the next-generation sequencing data. *Bioinformatics*, Article 2012;28(23):3150–2.
[16] Miladi M, Junge A, Costa F, Seemann SE, Havgaard JH, et al. RNAscClust: clustering RNA sequences using structure conservation and graph based motifs. Bioinformatics 2017;33(14):2089–96.
[17] Rao DM, Moler JC, Ozden M, Zhang Y, Liang C, et al. PEACE: parallel environment for assembly and clustering of gene expression. Nucleic Acids Res 2010;38(Web Server issue). W737-742.
[18] Torarinsson E, Havgaard JH, Gorodkin J. Multiple structural alignment and clustering of RNA sequences. Bioinformatics 2007;23(8):926–32.
[19] Hauser M, Mayer CE, Söding J. kClust: fast and sensitive clustering of large protein sequence databases. BMC Bioinformatics 2013;14:248.
[20] Enright AJ, Van Dongen S, Ouzounis CA. An efficient algorithm for large-scale detection of protein families. Nucleic Acids Res 2002;30(7):1575–84.
[21] Nepusz T, Sasidharan R, Paccanaro A. SCPS: a fast implementation of a spectral method for detecting protein families on a genome-wide scale. BMC Bioinformatics 2010;11:120.
[22] Dorr DH, Denton AM. Generalised sequence signatures through symbolic clustering. Int J Data Min Bioinform 2010;4(6):656–74.
[23] Jensen ST, Shen L, Liu JS. Combining phylogenetic motif discovery and motif clustering to predict co-regulated genes. Bioinformatics 2005;21(20):3832–9.
[24] Saito Y, Sato K, Sakakibara Y. Fast and accurate clustering of noncoding RNAs using ensembles of sequence alignments and secondary structures. BMC Bioinformatics 2011;12(Suppl 1):S48.
[25] Wu Y, Tang Y, Tringe S, Simmons B, Singer S. MaxBin: an automated binning method to recover individual genomes from metagenomes using an expectation-maximization algorithm. Microbiome 2014:2.
[26] Wang Z, Lu Y, Sun F, Zhu S. SolidBin: improving metagenome binning with semi-supervised normalized cut. Bioinformatics 2019;35(21):4229–38.
[27] Herath D, Tang SL, Tandon K, Ackland D, Halgamuge SK. CoMet: a workflow using contig coverage and composition for binning a metagenomic sample with high precision. BMC Bioinformatics 2017;18(Suppl 16):571.
[28] Alneberg J, Bjarnason BS, de Bruijn I, Schirmer M, Quick J, et al. Binning metagenomic contigs by coverage and composition. Nat Methods 2014;11(11):1144–6.
[29] Wu YW, Simmons BA, Singer SW. MaxBin 2.0: an automated binning algorithm to recover genomes from multiple metagenomic datasets. Bioinformatics 2016;32(4):605–7.
[30] Metzker ML. Sequencing technologies—the next generation. Nat Rev Genet 2010;11(1):31–46.
[31] Pareek CS, Smoczynski R, Tretyn A. Sequencing technologies and genome sequencing. J Appl Genet 2011;52(4):413–35.
[32] Collins FS, Morgan M, Patrinos A. The Human Genome Project: lessons from large-scale biology. Science 2003;300(5617):286–90.
[33] Koepfli K-P, Paten B, Scientists GKCo, O'Brien SJ. The Genome 10K Project: a way forward. Annu Rev Anim Biosci 2015;3(1):57–111.
[34] Gordon D, Huddleston J, Chaisson MJ, Hill CM, Kronenberg ZN, et al. Long-read sequence assembly of the gorilla genome. Science 2016;352(6281):aae0344.
[35] Pruitt KD, Tatusova T, Brown GR, Maglott DR. NCBI Reference Sequences (RefSeq): current status, new features and genome annotation policy. Nucleic Acids Res 2012;40(D1):D130–5.
[36] Niwa R, Syaputri Y, Horie M, Iwahashi H. Draft Genome Sequence of Lactobacillus plantarum IYO1511, Isolated from Ishizuchi-Kurocha. Microbiol Resour Announcements 2020;9:18.
[37] Palevich N, Palevich FP, Maclean PH, Jauregui R, Altermann E, et al. Whole-Genome Sequencing of Clostridium sp. Strain FP2, Isolated from Spoiled Venison. Microbiol Resour Announcements 2020;9:18.
[38] Dijkshoorn L, Ursing BM, Ursing JB. Strain, clone and species: comments on three basic concepts of bacteriology. J Med Microbiol 2000;49(5):397–401.
[39] Schneeberger K, Ossowski S, Ott F, Klein J, Wang X, et al. Reference-guided assembly of four diverse Arabidopsis thaliana genomes. Proc Natl. Acad Sci U S A 2011;108(25):10249–54.
[40] Lischer HEL, Shimizu KK. Reference-guided de novo assembly approach improves genome reconstruction for related species. BMC Bioinformatics 2017;18(1):474.
[41] Suyama M, Bork P. Evolution of prokaryotic gene order: genome rearrangements in closely related species. Trends Genet 2001;17(1):10–3.
[42] Bolger A, Lohse M, Usadel B. Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics 2014;30(15):2114–20.
[43] Li H, Durbin R. Fast and accurate short read alignment with Burrows–Wheeler transform. Bioinformatics 2009;25(14):1754–60.
[44] Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, et al. The sequence alignment/map format and SAMtools. Bioinformatics 2009;25(16):2078–9.
[45] Harris R.S. Improved pairwise Alignmnet of genomic DNA, 2007.
[46] Ma J, Zhang L, Suh BB, Raney BJ, Burhans RC, et al. Reconstructing contiguous regions of an ancestral genome. Genome Res 2006;16(12):1557–65.
[47] Quinlan AR, Hall IM. BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics 2010;26(6):841–2.
[48] Ester M, Kriegel H-P, Sander J, Xu X. A density-based algorithm for discovering clusters in large spatial databases with noise. Kdd. 1996.
[49] Huang W, Li L, Myers JR, Marth GT. ART: a next-generation sequencing read simulator. Bioinformatics 2011;28(4):593–4.
[50] Shannon CE. A mathematical theory of communication. Bell Syst Tech J 1948;27(3):379–423.
[51] Bankevich A, Nurk S, Antipov D, Gurevich AA, Dvorkin M, et al. SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. J Comput Biol 2012;19(5):455–77.
[52] Gurevich A, Saveliev V, Vyahhi N, Tesler G. QUAST: quality assessment tool for genome assemblies. Bioinformatics 2013;29(8):1072–5.
[53] Takahashi H, Sakagawa E, Sakakibara I, Machida C, Miyaki S, et al. Draft genome sequence of Saccharomyces cerevisiae strain Hm-1, isolated from cotton rose-mallow. Microbiol Resour Announcements 2018;7:13.
[54] Nowell RW, Laue BE, Sharp PM, Green S. Comparative genomics reveals genes significantly associated with woody hosts in the plant pathogen Pseudomonas syringae. Mol Plant Pathol 2016;17(9):1409–24.
[55] Simao F, Waterhouse R, Ioannidis P, Kriventseva E, Zdobnov E. BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. Bioinformatics 2015;31(19):3210–2.