Bioinformatics
doi.10.1093/bioinformatics/xxxxxx
Advance Access Publication Date: Day Month Year
Original papers

Original paper

HiFine: integrating Hi-c-based and shotgun-based methods to reFine binning of metagenomic contigs

Yuxuan Du[1] **and Fengzhu Sun**[1,*]

[1] Department of Quantitative and Computational Biology, University of Southern California, USA.

[*] To whom correspondence should be addressed.

Associate Editor: XXXXXXX

Received on XXXXX; revised on XXXXX; accepted on XXXXX

Abstract

Motivation: Metagenomic binning aims to retrieve microbial genomes directly from ecosystems by clustering metagenomic contigs assembled from short reads into draft genomic bins. Traditional shotgun-based binning methods depend on the contigs' composition and abundance profiles and are impaired by the paucity of enough samples to construct reliable co-abundance profiles. When applied to a single sample, shotgun-based binning methods struggle to distinguish closely related species only using composition information. As an alternative binning approach, Hi-C-based binning employs metagenomic Hi-C technique to measure the proximity contacts between metagenomic fragments. However, spurious inter-species Hi-C contacts inevitably generated by incorrect ligations of DNA fragments between species link the contigs from varying genomes, weakening the purity of final draft genomic bins. Therefore, it is imperative to develop a binning pipeline to overcome the shortcomings of both types of binning methods on a single sample.

Results: We develop HiFine, a novel binning pipeline to refine the binning results of metagenomic contigs by integrating both Hi-C-based and shotgun-based binning tools. HiFine designs a strategy of fragmentation for the original bin sets derived from the Hi-C-based and shotgun-based binning methods, which considerably increases the purity of initial bins, followed by merging fragmented bins and recruiting unbinned contigs. We demonstrate that HiFine significantly improves the existing binning results of both types of binning methods and achieves better performance in constructing species genomes on publicly available datasets. To the best of our knowledge, HiFine is the first pipeline to integrate different types of tools for the binning of metagenomic contigs.

Availability: HiFine is available at https://github.com/dyxstat/HiFine.

Contact: fsun@usc.edu

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

Metagenomics is a field that characterizes the diversity of species from microbial samples without the cultivation or isolation of microorganisms (Handelsman, 2004). Metagenomic studies reveal the complex community structures and establish interactions between microbial organisms (Hugenholtz and Tyson, 2008). High throughput metagenomic shotgun sequencing technologies directly capture genomic fragments from various environments and generate a tremendous number of short reads (Albertsen *et al.*, 2013). These shotgun reads can be either clustered into groups to reduce the size of metagenomic datasets (Luo *et al.*, 2019; Balvert *et al.*, 2021) or assembled into longer contigs, which are usually a portion of the full-length genomes (Li *et al.*, 2015; Nurk *et al.*, 2017). To retrieve the complete genomes present in microbial ecosystems, assembled contigs are grouped into bins that represent draft genomes of different species. This grouping process, termed binning, is the foundation of the downstream taxonomic profiling and functional analysis.

Traditional shotgun-based binning methods make use of the contigs' compositions and/or abundance profiles (Alneberg *et al.*, 2014; Wu *et al.*, 2016; Lu *et al.*, 2017; Kang *et al.*, 2019). Compositions of contigs usually refer to GC-content and oligonucleotide frequencies and the shotgun-based binning tools assume that contigs from the same genome share similar compositions (Chatterji *et al.*, 2008; Yang *et al.*, 2009). Besides, it has been shown that coverage profiles of metagenomic contigs from the same genome are highly correlated across multiple samples (Nielsen *et al.*, 2014). Although some shotgun-based binning pipelines have achieved good retrieval performance combining the information of the compositions and abundance, effective co-abundance profiles cannot be constructed if there are not enough samples due to the cost of sequencing or the limited ability to collect samples, which is common in clinical studies, for instance. When applied to a single sample, the shotgun-based binning methods can merely rely on the composition information to group contigs and struggle to distinguish closely related species with similar genomic compositions.

Hi-C-based binning is an alternative binning approach designed for a single sample based on the high-throughput chromosome conformation capture (Hi-C) experiments (Lieberman-Aiden *et al.*, 2009). Metagenomic Hi-C is a genomic proximity ligation technique generating millions of paired-end reads linking metagenomic squences in close three-dimensional distance within cells (Burton *et al.*, 2014; Beitel *et al.*, 2014). Therefore, the number of Hi-C read pairs connecting two assembled contigs is significantly related to the probability that contigs belong to the same genome, resulting in multiple binning pipelines making use of Hi-C interactions to group contigs (Press *et al.*, 2017; DeMaere and Darling, 2019; Du and Sun, 2022). In the Hi-C-based binning analysis, paired-end Hi-C sequencing reads derived from the same community of the shotgun library are mapped to the assembled contigs to construct raw contact maps between contigs. Raw Hi-C contact maps are then normalized to remove the high experimental biases (Du *et al.*, 2022). Finally, contigs are grouped using normalized contact maps to obtain draft genomic bins. The potential of Hi-C to deconvolute metagenomes and separate closely related genomes has been demonstrated on synthetic and real microbial communities (DeMaere and Darling, 2019; Du and Sun, 2022). However, spurious inter-species contacts inevitably generated from the ligation of DNA fragments between species link contigs from various genomes. Hence, contigs from different species may be incorrectly grouped into highly complete bins as contamination, weakening the purity of the final draft genomes (Stalder *et al.*, 2019; Du *et al.*, 2022).

To tackle the shortcomings of both shotgun-based and Hi-C-based binning methods on a single sample, we put forward HiFine, a novel single-sample binning pipeline to refine the binning results of metagenomic contigs by integrating existing Hi-C-based and shotgun-based binning tools. HiFine is a generic approach for integrating Hi-C based with shotgun based binning methods and consists of three steps. In the first step, HiFine designs a strategy of fragmentation by selecting out the intersections as fragmented bins between two bin sets constructed by a Hi-C-based binning method and a shotgun-based binning tool. Theoretically, contigs within fragmented bins are more likely to come from the same genomes as they have been grouped together according to both criteria of proximity contacts and composition similarity. Hence, our fragmentation approach can greatly increase the purity of bins, which is also demonstrated by our experimental results in Subsection 4.2. Considering that some genomes can only be detected by Hi-C-based or shotgun-based binning methods and thus are not included in the fragmented bins, HiFine adds original bins that remain relatively complete after removing shared contigs into the set of fragmented bins. In the second and third steps, HiFine merges the fragmented bins that potentially belong to the same species and recruits contigs that are not contained in the fragmented bins by reusing the normalized Hi-C contact maps. To the best of knowledge, HiFine is the first pipeline to integrate different types of binning tools on a single sample and is able to significantly improve the existing binning results of both types of binning methods.

2 Methods

2.1 Obtain the initial binning results from different types of binning methods

As HiFine aims at refining binning of metagenomic contigs by integrating both Hi-C-based and shotgun-based tools, we first need to generate two initial binning sets, where one binning set is constructed by a Hi-C-based binning method and the other set is derived by a shotgun-based binning pipeline.

2.2 Pipeline of the HiFine refinement method

2.2.1 Step1: Construct fragmented bins

We design a fragmentation algorithm to obtain a set of fragmented bins. The pseudo-code workflow of the algorithm to generate fragmented bins can be found in Supplementary Materials Algorithm S1. Specifically, the output bin sets generated by a Hi-C-based binning method and a shotgun-based binning tool are used as inputs in the first step. Since HiFine is developed to refine bins derived from the same set of assembled contigs, intersections between two bin sets were carried out by contig indices assigned by the assembly software as groups of shared contigs, which are then extracted and removed from the original bins as fragmented bins. Since contigs within the same fragmented bins have been clustered together by both the shotgun-based binning method in terms of composition similarity and the Hi-C-based binning method in terms of proximity contacts, they are more likely to come from the same genome. Therefore, we expect that the strategy of fragmentation can improve the purity of each fragmented bin, which has also been demonstrated in Subsection 4.2.

Moreover, Hi-C-based and shotgun-based binning sometimes identify different genomes due to different abilities to ascertain the same pool of genomes (Stalder *et al.*, 2019). In other word, a few genomes are only detected by Hi-C-based or shotgun-based binning methods and thus cannot be included into the fragmented bins by figuring out the groups of shared contigs. Bins containing such kind of genomes always remain relatively complete after we remove shared contigs from the original bins and we refer to these bins as *remaining complete bins*. Discarding remaining complete bins may lead to the loss of some detected genomes. Therefore, if the total length of retained contigs in one bin is larger than 80% of the original bin size and above a lower bound restriction (default, 500 kbp), we regard this bin as the remaining complete bin and then add this bin into fragmented bins.

After the first step, we can obtain fragmented bins from two sets of draft bins constructed by one Hi-C-based binning method and one shotgun-based binning tool. The fragmented bins come from two sources: sets of intersected contigs from the two types of binning approaches and remaining complete bins.

Despite the advantage of our strategy of fragmentation, taking the intersection between two sets of bins generates some small genomic bins, where multiple fragmented bins may belong to the same genome. Moreover, some contigs are included in either set of bins but are not contained in the fragmented bins after the strategy of fragmentation. Therefore, we design the second and third steps to solve these two problems, respectively.

2.2.2 Step2: Merge fragmented bins

To solve the problem of small genomic bins, we merge the fragmented bins that potentially belong to the same species in the second step.

We employ the Hi-C contact maps normalized by HiCzin (Du *et al.*, 2022), a state-of-the-art normalization method designed for metagenomic Hi-C contact maps, to merge the fragmented bins. HiCzin applied a zero-inflated negative binomial regression framework to remove potential experimental biases, including the number of enzymatic restriction sites on contigs, contig length and coverage. The specific steps to generate the normalized Hi-C contact maps can be found in Supplementary Materials Section 1. Let M and F_k denote the normalized Hi-C contact maps and the k-th fragmented bin, respectively, and let M_{c_1,c_2} represent the normalized Hi-C contacts between contigs c_1 and c_2. Noticeably, M_{c_1,c_2} reflects the proximity between two contigs. The larger M_{c_1,c_2} is, the closer c_1 and c_2 tend to be, indicating that contigs c_1 and c_2 are more likely to belong to the same genome. To measure the similarity between two fragmented bins, we design a modularity-like bin-to-bin similarity score S using the normalized Hi-C contact maps M as

$$S_{F_i,F_j} = \frac{\sum_{c_1 \in F_i, c_2 \in F_j} M_{c_1,c_2}}{\#F_i \times \#F_j}, \tag{1}$$

where S_{F_i,F_j} represents the similarity between the bins F_i and F_j, c_1 and c_2 denote the contigs in the bins and $\#F_i$ and $\#F_j$ are the number of contigs in bins F_i and F_j.

The similarity score S_{F_i,F_j} reflects average Hi-C contacts between two fragmented bins. High similarity score indicates close relationship with respect to the proximity contacts. In a special case when F_j is equal to F_i, the similarity score S_{F_i,F_i} becomes the average Hi-C contacts within fragmented bin, i.e,

$$S_{F_i,F_i} = \frac{\sum_{c_1,c_2 \in F_i;\, c_1 \neq c_2} M_{c_1,c_2}}{(\#F_i)^2}. \tag{2}$$

In fact, S_{F_i,F_i} can also be regarded as the bin-to-bin similarity between the fragmented bin F_i and its copy. If similarity score between the fragmented bin F_j and F_i is slightly smaller or even larger than the score between the fragmented bin F_i and its copy, we consider that fragmented bin F_j is a closely-related bin to F_i and has the potential to merge with F_i.

Based on the aforementioned discussions, we put forward an algorithm to merge the fragmented bins. Let $\mathcal{F}$ denote the set of all fragmented bins. Algorithm 1 describes the pseudo-codes of the merging process where the function $\overline{\textit{UpdateMI}}$ is presented in Supplementary Materials Algorithm S2. To render the merging strategy less tedious and more scalable, we assume that larger fragmented bins are more stable in the merging procedure and we always attempt to merge smaller bins to larger ones. Under this assumption, we sort and handle the fragmented bins in descending order of the bin size. For each sorted fragmented bin F_i ($i \in [1, |\mathcal{F}|]$), we define that a fragmented bin F_j is *closely-related* to F_i if

$$S_{F_i,F_j} \geq \alpha \times S_{F_i,F_i}, \tag{3}$$

where j is larger than i. Since the set of fragmented bins $\mathcal{F}$ has been sorted in descending order according to their size, the bin size of F_j is smaller than that of F_i. In this way, all closely-related bins of F_i can be identified. Noticeably, α (default, 0.6) serves as a merging coefficient and the merging standard becomes stricter with the increase of α. Initially, all fragmented bins are not merged. During the merging steps, we merge all closely-related bins of F_1 to F_1 in the first step. Then in the i-th step, we deal with the fragmented bin F_i. On the one hand, if the bin F_i has not yet been merged, we then check the current merging status of all its closely-related bins as follows:

- If all closely-related bins of F_i are not merged, we then merge those bins into F_i.
- If one or more closely-related bins have been merged to a bin F_k, we then merge the bin F_i and the rest of closely-related bins to F_k.
- If closely-related bins are merged to different bins, then the bin F_i is regarded as an ambiguous bin and discarded.

On the other hand, if the bin F_i has already been merged to one bin, denoted by F_k, situations become more complicated:

Algorithm 1 Merge fragmented bins.

```
Input: The set of fragmented bins, $\mathcal{F}$; The normalized Hi-C contact
    matrix, $M$; Merging coefficient, $\alpha$ (default, 0.6);
Output: Merging index, $MI$;
Sort the fragmented bins within $\mathcal{F}$ by bin size in descending order;
Initialize the merging index $MI$ as an zero vector with length $|\mathcal{F}|$;
for each $i \in [1, |\mathcal{F}|]$ do
    Initialize an empty list $L_{F_i}$ and $I_{F_i}$;
    Compute $S_{F_i,F_i}$ as Eq. (2);
    for each $j \in [i+1, |\mathcal{F}|]$ do
        Compute $S_{F_i,F_j}$ as Eq. (1);
        if $S_{F_i,F_j} \geq \alpha \times S_{F_i,F_i}$ then
            if $MI[j] \neq 0$ and $MI[j] \notin I_{F_i}$ then
                $I_{F_i} = I_{F_i} \cup \{I[j]\}$;
            else if $MI[j] = 0$ then
                $L_{F_i} = L_{F_i} \cup \{j\}$;
            end if
        end if
    end for
    $MI = \overline{\textit{UpdateMI}}(MI, L_{F_i}, I_{F_i})$;
end for
return $MI$;
```

- If all closely-related bins of F_i are not merged, we then merge those bins into F_k.
- If one or more closely-related bins have also been merged to F_k, we then merge the rest of unmerged closely-related bins to F_k.
- If closely-related bins are merged to one bin that is different from F_k or merged to more than one bin, the bin F_i is then regarded as an ambiguous bin and removed from F_k.

The pseudo-codes in Supplementary Materials Algorithm S2 follow the aforementioned rules. Algorithm 1 outputs the merging index MI where the k-th fragmented bin will be merged to the $MI[k]$-th bin. Therefore, fragmented bins with the same index will be merged together. Moreover, those bins with merging index 0 are ambiguous bins and are discarded. Finally, we can obtain a set of merged bins, denoted by $\mathcal{B}$.

2.2.3 Step3: Recruit unbinned contigs into merged bins

The third step is designed to recruit contigs that are not included in the merged bins. We reutilize the normalized Hi-C contact maps M to assign the unbinned contigs into merged bins and hence we only consider unbinned contigs showing in the contact maps.

We first define the contig-to-bin association A between a contig c and a merged bin $B \in \mathcal{B}$ as

$$A_{c,B} = \frac{\sum_{c_1 \in B} M_{c,c_1}}{\#B}, \tag{4}$$

where c_1 denotes the contigs in the merged bin B, M_{c,c_1} is the normalized contacts between contigs c and c_1 and $\#B$ represents the number of contigs in the bin B.

Then, for each unbinned contig c, we compute contig-to-bin association from the contig to each of the merged bin in $\mathcal{B}$. We identify the bin with the highest association as the potential bin to recruit the contig c. However, there exist some mistakes if we recruit all unbinned contigs to the merged bins according to the contig-to-bin association as some contigs do not belong to any of the merged bins. To solve this problem, a discarding procedure is introduced. Specifically, assume that the k-th merged bin $B_k \in \mathcal{B}$ is the potential bin of the contig c. We compute the bin-to-bin

Algorithm 2 Recruit unbinned contigs into merged bins.

```
Input: The set of merged bins, $\mathcal{B}$; The normalized Hi-C contact maps, $M$; Recruiting coefficient, $\beta$ (default, 0.3);
Output: Unbinned contig set, $C$; Recruiting index, $RI$;
Construct unbinned contig set $C$;
Initialize the recruiting index $R$ as an zero vector with length $|C|$;
for each $B \in \mathcal{B}$ do
    Compute the bin-to-bin similarity score $S_{B,B}$ as Eq. 2
end for
for each $i \in [1, |C|]$ do
    for each $B \in \mathcal{B}$ do
        Compute the contig-to-bin association $A_{C_i,B}$ as Eq. 4
    end for
    $j = \underset{k}{\operatorname{argmax}} A_{C_i,B_k}$
    if $A_{C_i,B_j} \geq \beta \times S_{B_j,B_j}$ then
        $RI[i] = j$
    end if
end for
    return $C, RI$;
```

similarity score S_{B_k,B_k} as Eq. (2) and compare A_{c,B_k} and S_{B_k,B_k}. If

$$A_{c,B_k} \geq \beta \times S_{B_k,B_k}, \tag{5}$$

where β (default 0.3) denotes the recruiting coefficient. Noticeably, with the increase of β, the contig recruiting adheres to stricter criteria, leading to higher recruiting accuracy but fewer recruited contigs. The default value of β can balance the 'trade-off' between the recruiting quality and the total bin size as shown in Subsection 4.3. We assign contig c to the merged bin B_k. Otherwise, we discard this contig because the contig-to-bin association is not strong enough to ensure the recruitment.

Algorithm 2 gives the pseudo-code of the recruiting step. The input of the algorithm is the set of merged bins obtained from the second step, the normalized Hi-C contact maps, and the value of recruiting coefficient. Algorithm 2 then outputs the set of the unbinned contigs C and the recruiting index RI, where the k-th contig in the set C will be assigned to the $RI[k]$-th merged bin in $\mathcal{B}$. Contigs with recruiting index 0 fail to pass the recruiting criteria (5) and are discarded. Therefore, after the recruiting step, we can obtain the final refinement bins, denoted by $\mathcal{R}$.

In conclusion, HiFine employs three steps to refine the Hi-C-based and shotgun-based binning results. It first designs a fragmentation algorithm by selecting out groups of shared contigs as fragmented bins between two bin sets constructed by a Hi-C-based binning method and a shotgun-based binning tool. Bins that remain relatively complete after removing shared contigs are also added into the set of fragmented bins. In the second and third steps, HiFine merges the fragmented bins that potentially come from the same species and recruits unbinned contigs by reusing the normalized Hi-C contact maps.

3 Experiments

3.1 Datasets

We validated our method on three publicly available datasets.

3.1.1 The synthetic metagenomic yeast datasets

The synthetic metagenomic yeast sample was composed of 16 yeast strains from 13 yeast species (NCBI accession: SRR1263009 and SRR1262938) (Burton *et al.*, 2014). Shotgun libraries were constructed by the Nextera DNA Sample Preparation Kit (Illumina) and included 85.7 million read pairs at 101 bp per read. Hi-C libraries were prepared using HindIII and NcoI restriction endonuclease (NEB) and then sequenced by the HiSeq and MiSeq Illumina platforms. Raw Hi-C dataset contained 81 million read pairs at 100 bp per read. As contigs' identities and species in this sample are known, we can validate the ability to retrieve species genomes and the accuracy of contig recruitment for HiFine.

3.1.2 The inoculated beer datasets

The inoculated beer sample was generated from the top of a wine barrel containing the spontaneously inoculated beer (NCBI accession: SRR5890763, SRR5890764) (Smukowski Heil *et al.*, 2018). Shotgun libraries were prepared using the Nextera Kit (Illumina). Hi-C libraries were prepared with HindIII and NcoI restriction enzymes and then sequenced on the NextSeq 500 Illumina platform. After sequencing, 27.5 million and 28.1 million of read pairs at 100 bp per read were produced for the shotgun and Hi-C libraries, respectively. This sample is used for evaluating the performance of HiFine on the low-complexity real sample.

3.1.3 The human gut datasets

The human gut datasets were derived from a fecal sample of a human subject (NCBI accession: SRR6131122, SRR6131123, and SRR6131124) (Press *et al.*, 2017). Two restriction enzymes MluCI and Sau3AI (New England Biolabs) were used to construct two different Hi-C libraries. The shotgun and Hi-C libraries were sequenced on the Illumina HiSeqX platform at 151 bp. The raw shotgun libraries consisted of 250,884,672 read pairs and the sequencing of two Hi-C libraries produced 41,733,770 read pairs (Sau3AI library) and 48,798,091 read pairs (MluCI library), respectively. This sample is used for evaluating the performance of HiFine on the high-complexity real sample.

3.2 Deriving the initial sets of binning results

In this paper, we selected two state-of-the-art Hi-C-based binning tools HiCBin (v1.0.0) (Du and Sun, 2022) and bin3C (v0.1.1) (DeMaere and Darling, 2019), and three popular shotgun-based binning methods MetaBAT2 (v2.12.1) (Kang *et al.*, 2019), MaxBin2 (v2.2.4) (Wu *et al.*, 2016), and VAMB (v3.0.3) (Nissen *et al.*, 2021). Therefore, we can obtain six combinations as inputs of HiFine on the three datasets, i.e., HiCBin+MetaBAT2, HiCBin+MaxBin2, HiCBin+VAMB, bin3C+MetaBAT2, bin3C+MaxBin2, and bin3C+VAMB. HiFine was validated for all six cases on the three datasets. We presented the results of combining HiCBin and MetaBAT2 by HiFine in the following section in the main text and the results of other combinations were shown in Supplementary Materials Figures S2-S6.

HiCBin (v1.0.0) (Du and Sun, 2022) and MetaBAT2 (v2.12.1) (Kang *et al.*, 2019) were exploited to generate the initial binning results. HiCBin is a state-of-the-art metagenomic Hi-C-based binning pipeline. After the normalization of raw Hi-C contact maps, the Leiden algorithm (Traag *et al.*, 2019) combined with the Reichardt and Bornholdt's Potts model (Reichardt and Bornholdt, 2006) was utilized to cluster contigs based on the normalized Hi-C contacts. HiCBin outperformed all other Hi-C-based binning pipelines on both synthetic and real microbial samples (Du and Sun, 2022). MataBAT2 was one of the most popular shotgun-based binning tools and achieved one of the best binning performance in the CAMI Challenge Dataset (Sczyrba *et al.*, 2017). MetaBAT2 computed the composition scores by integrating both normalized tetra-nucleotide frequency scores and abundance scores, followed by contig clustering using a modified label propagation algorithm.

Specifically, after the pre-processing of raw shotgun and Hi-C libraries, short reads were assembled into contigs using MEGAHIT (v1.2.9) (Table 1) and Hi-C read pairs were aligned to the assembled contigs (see Supplementary Materials Section 2). Then, we employed HiCBin and MetaBAT2 to derive two initial binning sets. HiCBin was run with

Table 1. Contigs assembled by MEGAHIT for three datasets.

Dataset	Contig num	N50	Total length
Synthetic yeast	6,566	63,305	126,030,343
Beer	4,418	57,561	67,968,628
Human gut	105,267	14,166	530,969,816

Note: Contig num represents the number of contigs assembled by MEGAHIT. N50 is defined by the length of the shortest contig where contigs with longer and equal length cover at least 50% of the assembly.

parameters '-min-signal 2 -min-binsize 100000' and MetaBAT2 was executed with default parameters.

3.3 Evaluation metrics

To evaluate HiFine, we identified the ground-truth on the species level of assembled contigs in all three datasets. For the metagenomic yeast datasets, as the reference genomes of all strains within the sample are known, we downloaded all reference genomes of these 16 yeast strains and then aligned contigs to reference genomes at the species level by BLASTn (Ye *et al.*, 2006) (see Supplementary Materials Section 3). For the real beer and human gut samples, TAXAassign (v0.4) (https://github.com/umerijaz/TAXAassign) was utilized to label the contigs by searching in the NCBI nucleotide database with the 95% identity percentage. After labeling contigs, three common clustering metrics including F-score, Adjusted Rand Index (ARI), and Normalized Mutual Information (NMI) were used to evaluate the binning results. To demonstrate our strategy of fragmentation, we apply a homogeneity metric to measure the extent of clusters containing only data points of a single class (Rosenberg and Hirschberg, 2007). The homogeneity metric examines the conditional entropy of the class distribution given the proposed clustering and is equal to one when each cluster contains only members of a single class. Definitions of all metrics are shown in Supplementary Materials Section 4.

4 Results

4.1 Performance of HiFine on three datasets

- For the synthetic yeast datasets (Figure 1a), HiCBin and MetaBAT2 achieved 0.908 and 0.608 in terms of F-score, 0.894 and 0.480 in terms of ARI, and 0.895 and 0.712 in terms of NMI, respectively. All three binning metrics were improved to 0.963, 0.958, and 0.959 by HiFine. In the third step of HiFine, 1221 contigs were recuited into the bins. Among these contigs, 1187 contigs (97.2%) were assigned correctly, demonstrating the high accuracy of our recruiting algorithm. Moreover, 12 out of 13 species were detected in exactly 12 bins by HiFine with high purity in each bin. The exception was *P. pastoris*, which had a very low fraction of read coverage as reported in the original paper ((Burton *et al.*, 2014)). In comparison, HiCBin detected 12 species except *P. pastoris* in 14 bins, where two draft genomes belonged to the same species *S. cerevisiae* and one bin was highly contaminated. As for MetaBAT2, 11 species except *P. pastoris* and *S. kudriavzevii* could be detected in 13 bins, where two bins came from *A. gossypii*, two belonged to *L. kluyveri*, and two other bins were highly contaminated. Consequently, HiFine could perform better in constructing species genomes.
- Figure 1b shows the results for the inoculated beer datasets. MetaBAT2 and HiCBin showed very high scores on this low-complexity sample. The F-scores of MetaBAT2 and HiCBin were 0.945 and 0.991, respectively. The F-score was further improved to 0.998 by HiFine. The values of ARI of MetaBAT2 and HiCBin were 0.924 and 0.989, respectively. HiFine further increased this score to 0.997. As for NMI, MetaBAT2 and HiCBin got 0.898 and 0.974, respectively. In comparison, the NMI was improved to 0.995 by HiFine. Therefore, HiFine could improve the binning quality on the low-complexity real sample.
- The results of the human gut datasets were shown in Figure 1c. HiCBin and MetaBAT2 obtained 0.788 and 0.678 in terms of F-score, respectively, which was further improved to 0.931 by HiFine. The values of ARI of HiCBin and MetaBAT2 were 0.768 and 0.662, respectively while the ARI score of HiFine was further increased to 0.926. The HiFine improved the NMI from 0.791 and 0.854 achieved by HiCBin and MetaBAT2, respectively to 0.911. Hence, we validated the significant improvement of HiFine on the highly complicated microbial community.

Moreover, for all three datasets, the total bin size of HiFine with default parameter was close to the total size of HiCBin and were much larger than that of MetaBAT2 (Supplementary Materials Table S2). This indicated that HiFine could achieve better binning quality and at the same time group most of the contigs.

4.2 The fragmentation strategy improves the purity of bins

As shown in Table 2, the homogeneity scores of the fragmented bins constructed by the first step of HiFine were larger than the scores of both sets of bins generated by MetaBAT2 and HiCBin for all three datasets. As the homogeneity metric reflected the purity of the bins, this finding demonstrated that contigs from the same fragmented bin are more likely to come from the same genome and proved the rationality of our strategy to select out shared contigs.

Table 2. Homogeneity scores of bins from MetaBAT2, HiCBin, and fragmented bins in HiFine for three datasets.

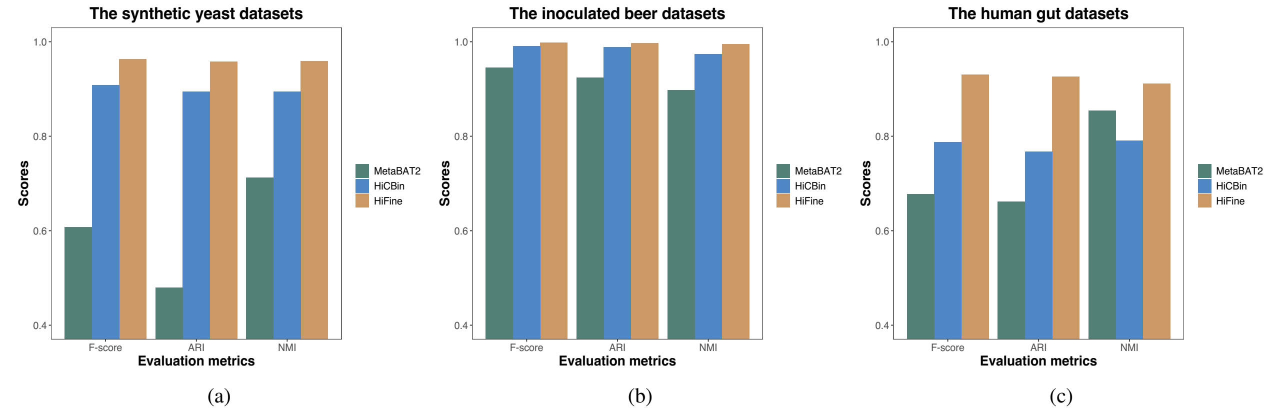
Dataset	MetaBAT2	HiCBin	Fragmented bins in HiFine
Synthetic Yeast	0.627	0.909	0.970
Beer	0.962	0.978	0.999
Human gut	0.907	0.877	0.959

4.3 Evaluating the impacts of merging and recruiting coefficients on the performance of HiFine

We study the impacts of two coefficient parameters (i.e., the merging coefficient α in the second step and the recruiting coefficient β in the third step) on the performance of HiFine.

For the merging coefficient α, the binning results were not affected by α on the low-complexity beer sample (Figure 2a). On the synthetic yeast datasets, the three clustering metrics (i.e., F-score, ARI and NMI) of HiFine were all stable when the merging coefficient α was no more than the default value (0.6) and slightly decreased when α was larger than 0.6 (Figure 2c). On the human gut datasets, the clustering metrics first increased and then decreased with the increase of α and reached the local maximum when α was equal to the default value 0.6 (Figure 2e).

For the recruiting coefficient β, the clustering metrics were still stable on the low-complexity beer sample (Figure 2b). On both the synthetic yeast (Figure 2d) and the human gut (Figure 2f) datasets, the three clustering metrics increased with the increase of β when β was no more than the default value (0.3). However, the growth significantly slowed down when β was larger than 0.3. The NMI even decreased from 0.911 to 0.909 when β increased from 0.3 to 0.4. Meanwhile, the proportion of the total bin size within the total length of the assembled contigs decreased with the increase of β. This can be explained by the fact that the smaller β leads to looser recruiting criteria, resulting in more recruited contigs but relatively



(a)

(b)

(c)

Fig. 1. HiFine outperforms MetaBAT2 and HiCBin based on (a) the synthetic yeast datasets, (b) the inoculated beer datasets, and (c) the human gut datasets using evaluation metrics of F-score, ARI and NMI.

lower recruitment accuracy. Therefore, we selected the default value of β as 0.3 to balance the 'trade-off' between the binning quality and the total bin size.

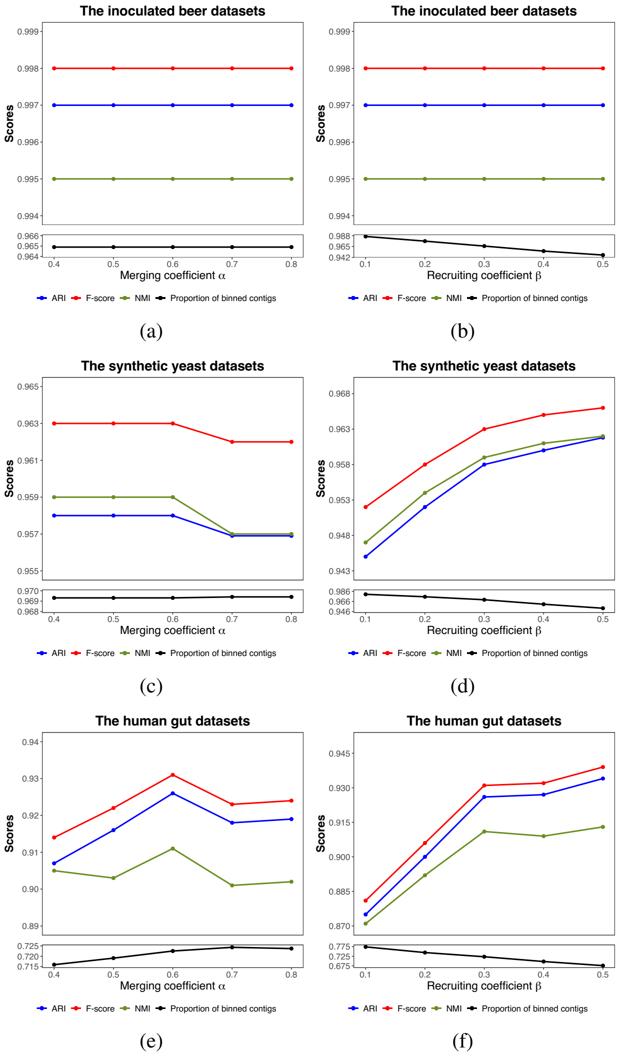
In conclusion, the choice of different merging coefficients and the recruiting coefficients did not have a large impact on the final binning results of HiFine and our default values could yield relatively satisfying and consistent binning performance on all three samples.

4.4 Comparing HiFine with assembly-graph based refinement pipeline GraphBin

In the read assembly process, contigs are constructed from the underlying assembly graph providing valuable connected relationship between contigs (Pevzner *et al.*, 2001; Simpson and Durbin, 2012). Therefore, assembly graph can be utilized in the binning refinement. GraphBin is a typical pipeline making use of the assembly graph to improve the binning results of existing tools (Mallawaarachchi *et al.*, 2020). To evaluate the refinement performance of the assembly graph on a single sample, we applied GraphBin in the MEGAHIT version with default parameters on all three datasets and compared the binning results of GraphBin to that of the initial bins (Supplementary Materials Figure S7). We found that the GraphBin could not improve the binning quality of the initial bins on our three datasets. In comparison, the binning scores were significantly increased by HiFine. From our observations, one potential reason was that the assembly graphs were extremely sparse for all three datasets, which might be ascribed to the relatively small shotgun library derived from only one sample. Specifically, there were only 4,740, 14,140, 416,808 edges in three assembly graphs corresponding to 6,566, 4,418, and 105,267 assembled contigs, respectively. Therefore, the assembly graphs could not provide enough information to refine the initial bins on a single sample.

5 Conclusion and Discussion

In this paper, we developed a binning method, HiFine, to integrate the Hi-C-based and shotgun-based binning tools on a single sample. We put forward the strategy of fragmentation by selecting out the intersected contigs between sets of bins from different types of binning tools. Experimental results confirmed the effectiveness of our strategy of fragmentation in the first step to improve the purity of bins and validated the high accuracy of the recruiting process of unbinned contigs in the third step. We also demonstrated that HiFine achieved a significant improvement on the existing binning results of both types of binning methods and performed



(a)

(b)

(c)

(d)

(e)

(f)

Fig. 2. Impacts of the (a) merging coefficient and (b) recruiting coefficient on the inoculated beer sample, (c) merging coefficient and (d) recruiting coefficient on the synthetic yeast sample, and (e) merging coefficient and (f) recruiting coefficient on the human gut sample.

better in retrieving species genomes. The high-purity genomes generated by HiFine can significantly facilitate the downstream analyses, such as tracking horizontal gene transfer and probing virus-host interactions.

However, our method has its own limitations. We observe that HiFine might not have a significant improvement if one of the existing binning method only grouped a considerably smaller number of contigs than the other type of binning method. Intuitively, if this situation happens, the set size of intersected contigs will be extremely small compared to the larger initial bin set. Then the fragmented bin set will be dominated by the relatively complete bins from the initial bin set. In other word, the fragmented bin set does not have a significant difference with the larger initial bin set. Normally, the Hi-C-based binning method outperforms the shotgun-based method when applied to the single sample. Hence, this problem might happen to the shotgun-based binning tool.

In the future research, there still exists a lot of work to do on a single sample. The idea of HiFine can be regarded as integrating the composition and proximity contact information together to refine the binning quality. Hence, one natural question is whether we can further improve the binning performance by incorporating more information from different sources, such as gene prediction information and DNA methylation.

Acknowledgements

YD and FS conceived the ideas and designed the study. YD implemented the methods, carried out the computational analyses, and drafted the manuscript. FS modified and finalized the paper.

Funding

The research is partially funded by NIH grants (R01GM120624 and R01GM131407) and NSF EF-2125142. YD is supported by the Viterbi Fellowship.

References

Albertsen, M., Hugenholtz, P., Skarshewski, A., Nielsen, K. L., Tyson, G. W., and Nielsen, P. H. (2013). Genome sequences of rare, uncultured bacteria obtained by differential coverage binning of multiple metagenomes. *Nature biotechnology*, **31**(6), 533–538.

Alneberg, J., Bjarnason, B. S., De Bruijn, I., Schirmer, M., Quick, J., Ijaz, U. Z., Lahti, L., Loman, N. J., Andersson, A. F., and Quince, C. (2014). Binning metagenomic contigs by coverage and composition. *Nature methods*, **11**(11), 1144–1146.

Balvert, M., Luo, X., Hauptfeld, E., Schönhuth, A., and Dutilh, B. E. (2021). OGRE: Overlap Graph-based metagenomic Read clustEring. *Bioinformatics*, **37**(7), 905–912.

Beitel, C. W., Froenicke, L., Lang, J. M., Korf, I. F., Michelmore, R. W., Eisen, J. A., and Darling, A. E. (2014). Strain-and plasmid-level deconvolution of a synthetic metagenome by sequencing proximity ligation products. *PeerJ*, **2**, e415.

Burton, J. N., Liachko, I., Dunham, M. J., and Shendure, J. (2014). Species-level deconvolution of metagenome assemblies with Hi-C–based contact probability maps. *G3: Genes, Genomes, Genetics*, **4**(7), 1339–1346.

Chatterji, S., Yamazaki, I., Bai, Z., and Eisen, J. A. (2008). CompostBin: A DNA composition-based algorithm for binning environmental shotgun reads. In *Annual International Conference on Research in Computational Molecular Biology*, pages 17–28. Springer.

DeMaere, M. Z. and Darling, A. E. (2019). bin3C: exploiting Hi-C sequencing data to accurately resolve metagenome-assembled genomes. *Genome biology*, **20**(46), 1–16.

Du, Y. and Sun, F. (2022). HiCBin: binning metagenomic contigs and recovering metagenome-assembled genomes using Hi-C contact maps. *Genome biology*, **23**(63), 1–21.

Du, Y., Laperriere, S. M., Fuhrman, J., and Sun, F. (2022). Normalizing metagenomic Hi-C data and detecting spurious contacts using zero-inflated negative binomial regression. *Journal of Computational Biology*, **29**(2), 106–120.

Handelsman, J. (2004). Metagenomics: application of genomics to uncultured microorganisms. *Microbiology and molecular biology reviews*, **68**(4), 669–685.

Hugenholtz, P. and Tyson, G. W. (2008). Metagenomics. *Nature*, **455**(7212), 481–483.

Kang, D. D., Li, F., Kirton, E., Thomas, A., Egan, R., An, H., and Wang, Z. (2019). MetaBAT 2: an adaptive binning algorithm for robust and efficient genome reconstruction from metagenome assemblies. *PeerJ*, **7**, e7359.

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**(10), 1674–1676.

Lieberman-Aiden, E., Van Berkum, N. L., Williams, L., Imakaev, M., Ragoczy, T., Telling, A., Amit, I., Lajoie, B. R., Sabo, P. J., Dorschner, M. O., *et al.* (2009). Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science*, **326**(5950), 289–293.

Lu, Y. Y., Chen, T., Fuhrman, J. A., and Sun, F. (2017). COCACOLA: binning metagenomic contigs using sequence COmposition, read CoverAge, CO-alignment and paired-end read LinkAge. *Bioinformatics*, **33**(6), 791–798.

Luo, Y., Yu, Y. W., Zeng, J., Berger, B., and Peng, J. (2019). Metagenomic binning through low-density hashing. *Bioinformatics*, **35**(2), 219–226.

Mallawaarachchi, V., Wickramarachchi, A., and Lin, Y. (2020). GraphBin: refined binning of metagenomic contigs using assembly graphs. *Bioinformatics*, **36**(11), 3307–3313.

Nielsen, H. B., Almeida, M., Juncker, A. S., Rasmussen, S., Li, J., Sunagawa, S., Plichta, D. R., Gautier, L., Pedersen, A. G., Le Chatelier, E., *et al.* (2014). Identification and assembly of genomes and genetic elements in complex metagenomic samples without using reference genomes. *Nature biotechnology*, **32**(8), 822–828.

Nissen, J. N., Johansen, J., Allesøe, R. L., Sønderby, C. K., Armenteros, J. J. A., Grønbech, C. H., Jensen, L. J., Nielsen, H. B., Petersen, T. N., Winther, O., *et al.* (2021). Improved metagenome binning and assembly using deep variational autoencoders. *Nature biotechnology*, **39**(5), 555–560.

Nurk, S., Meleshko, D., Korobeynikov, A., and Pevzner, P. A. (2017). metaSPAdes: a new versatile metagenomic assembler. *Genome research*, **27**(5), 824–834.

Pevzner, P. A., Tang, H., and Waterman, M. S. (2001). An Eulerian path approach to DNA fragment assembly. *Proceedings of the national academy of sciences*, **98**(17), 9748–9753.

Press, M. O., Wiser, A. H., Kronenberg, Z. N., Langford, K. W., Shakya, M., Lo, C.-C., Mueller, K. A., Sullivan, S. T., Chain, P. S., and Liachko, I. (2017). Hi-C deconvolution of a human gut microbiome yields high-quality draft genomes and reveals plasmid-genome interactions. *biorxiv*, **198713**.

Reichardt, J. and Bornholdt, S. (2006). Statistical mechanics of community detection. *Physical review E*, **74**(1), 016110.

Rosenberg, A. and Hirschberg, J. (2007). V-measure: A conditional entropy-based external cluster evaluation measure. In *Proceedings of the 2007 joint conference on empirical methods in natural language processing and computational natural language learning (EMNLP-CoNLL)*, pages 410–420.

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

Simpson, J. T. and Durbin, R. (2012). Efficient de novo assembly of large genomes using compressed data structures. *Genome research*, **22**(3), 549–556.

Smukowski Heil, C., Burton, J. N., Liachko, I., Friedrich, A., Hanson, N. A., Morris, C. L., Schacherer, J., Shendure, J., Thomas, J. H., and Dunham, M. J. (2018). Identification of a novel interspecific hybrid yeast from a metagenomic spontaneously inoculated beer sample using Hi-C. *Yeast*, **35**(1), 71–84.

Stalder, T., Press, M. O., Sullivan, S., Liachko, I., and Top, E. M. (2019). Linking the resistome and plasmidome to the microbiome. *The ISME journal*, **13**(10), 2437–2446.

Traag, V. A., Waltman, L., and Van Eck, N. J. (2019). From Louvain to Leiden: guaranteeing well-connected communities. *Scientific reports*, **9**(5233), 1–12.

Wu, Y.-W., Simmons, B. A., and Singer, S. W. (2016). MaxBin 2.0: an automated binning algorithm to recover genomes from multiple metagenomic datasets. *Bioinformatics*, **32**(4), 605–607.

Yang, B., Peng, Y., Leung, H. C., Yiu, S.-M., Chen, J.-C., and Chin, F. Y. (2009). Unsupervised binning of environmental genomic fragments based on an error robust selection of l-mers. In *Proceedings of the third international workshop on Data and text mining in bioinformatics*, pages 3–10.

Ye, J., McGinnis, S., and Madden, T. L. (2006). BLAST: improvements for better sequence analysis. *Nucleic acids research*, **34**(suppl_2), W6–W9.