ImputeCC enhances integrative Hi-C-based metagenomic binning through constrained random-walk-based imputation

Yuxuan Du[1][0000−0002−0568−3838], Wenxuan Zuo[1], and Fengzhu Sun[1*]

Department of Quantitative and Computational Biology, University of Southern California, USA
{yuxuandu,wzuo,fsun}@usc.edu

Abstract. Metagenomic Hi-C (metaHi-C) enables the recognition of relationships between contigs in terms of their physical proximity within the same cell, facilitating the reconstruction of high-quality metagenome-assembled genomes (MAGs) from complex microbial communities. However, current Hi-C-based contig binning methods solely depend on Hi-C interactions between contigs to group them, ignoring invaluable biological information, including the presence of single-copy marker genes. Here, we introduce ImputeCC, an integrative contig binning tool tailored for metaHi-C datasets. ImputeCC integrates Hi-C interactions with the inherent discriminative power of single-copy marker genes, initially clustering them as preliminary bins, and develops a new constrained random walk with restart (CRWR) algorithm to improve Hi-C connectivity among these contigs. Extensive evaluations on mock and real metaHi-C datasets from diverse environments, including the human gut, wastewater, cow rumen, and sheep gut, demonstrate that ImputeCC consistently outperforms other Hi-C-based contig binning tools. ImputeCC's genus-level analysis of the sheep gut microbiota further reveals its ability and potential to recover essential species from dominant genera such as *Bacteroides*, detect previously unrecognized genera, and shed light on the characteristics and functional roles of genera such as *Alistipes* within the sheep gut ecosystem.
Availability: ImputeCC is implemented in Python and available at `https://github.com/dyxstat/ImputeCC`. The Supplementary Information is available at `https://doi.org/10.5281/zenodo.10776604`.

Keywords: Metagenomic Hi-C · Integrative Contig Binning · MetaHi-C Contact Map Imputation · Constrained Random Walk With Restart

1 Introduction

Metagenomics is revolutionizing microbial ecology by enabling the exploration of complex microbial communities in diverse environments without the need for traditional microbial isolation or cultivation [17,18]. The recent combination of Hi-C sequencing with whole metagenomic shotgun sequencing leads to the development of the metagenomic Hi-C (metaHi-C) technique, which has provided novel perspectives on species diversity and the interactions among microorganisms within a single microbial sample [5,12,22,27]. In metaHi-C experiments, shotgun sequencing extracts genomic fragments from a microbial sample, while Hi-C sequencing conducted on the same microbial sample generates DNA-DNA proximity ligations within the same cells, resulting in millions of paired-end Hi-C short reads. These fragmented shotgun reads are assembled into longer contigs, forming the basis for aligning paired-end Hi-C reads. MetaHi-C contacts, representing the number of Hi-C read pairs linking contig pairs, reveal contig relationships based on physical proximity within the microbial community. Depending on whether the shotgun libraries in metaHi-C experiments are constructed using second-generation or third-generation sequencing technologies, metaHi-C experiments can be classified into either short-read or long-read metaHi-C datasets, respectively. Considering contigs originating from the same genome exhibit enriched Hi-C contact frequencies relative to those derived from distinct genomes, the process of Hi-C-based binning emerges and aims at grouping fragmented contigs into metagenome-assembled genomes (MAGs) [19] by leveraging Hi-C contacts between contigs [2,11,14]. The resulting MAG collections serve as fundamental prerequisites for downstream analyses, such as the elucidation of the metabolic potentials and functional roles of diverse microorganisms, as well as the exploration of virus-host interactions [8,35]. Various Hi-C-based contig binning methods have been developed, including HiCBin [14], MetaTOR [2], bin3C [11], and the MetaCC binning module (referred to as MetaCC) [15]. Compared to conventional shotgun-based binning tools reliant on sequence composition and contig coverage for contig clustering, Hi-C-based binning methods demonstrate their superior ability in MAG recovery using only one single sample [14,27].

However, existing Hi-C-based binning methods rely solely on Hi-C interactions for contig grouping, overlooking valuable biological information encapsulated within single-copy marker genes. These genes, present as single copies in the vast majority of genomes [1], hold the great potential to discriminate between contigs originating from distinct species when shared among them. This omission underscores a critical gap in current approaches, leaving ample room for enhancement and improved analyses. In response, we introduce ImputeCC, an integrative binning tool designed for metaHi-C datasets. ImputeCC manages to harness the comprehensive insights offered by both Hi-C interactions and single-copy marker genes to optimize the contig binning process. To thoroughly assess the effectiveness of ImputeCC, we conduct simulations for both short-read and long-read metaHi-C datasets. Subsequently, we demonstrate ImputeCC's performance against other publicly-available Hi-C-based binning tools using a diverse set of real short-read and long-read metaHi-C datasets including the human gut short-read [27], wastewater short-read [32], cow rumen long-read [4], and sheep gut long-read [3] metaHi-C datasets. ImputeCC's superior performance is particularly evident in the challenging sheep gut environment, where ImputeCC successfully retrieves an impressive total of 408 high-quality and 885 medium-quality MAGs, as assessed by the latest CheckM2 [9]. To the best of our knowledge, this represents the largest number of reference-quality MAGs reported from a single microbial sample. Furthermore, ImputeCC's genus-level analyses of the sheep gut microbiota reveal ability of ImputeCC to recover essential species from dominant genera and showed its potential to detect previously unrecognized genera.

2 Results

2.1 Overview of ImputeCC

ImputeCC is an integrative Hi-C-based binner that leverages the combined power of Hi-C interactions and single-copy marker genes in the contig binning process. Fig. 1 shows the outline of ImputeCC. The core concept of ImputeCC involves the preclustering of marker-gene-containing contigs guided by two fundamental principles: I) Contigs sharing the same single-copy marker gene originate from distinct species with high probability; II) Contigs without overlapping single-copy marker genes are likely from the same genome when connected by robust Hi-C signals. To address the challenge that marker-gene-containing contigs from the same genome may not be effectively linked by Hi-C contacts due to the locality characteristics of proximity ligations, we design a new constrained random walk with restart (CRWR) algorithm to impute the metaHi-C contact matrix before preclustering, with all random walks limited to start from marker-gene-containing contigs. Subsequently, by leveraging the imputed Hi-C matrix in conjunction with the aforementioned principles, ImputeCC can accurately precluster contigs with single-copy marker genes, establishing them as preliminary bins. Finally, the tool applies Leiden clustering [33] to group all assembled contigs, utilizing the information from preliminary bins to optimize the binning process.

2.2 ImputeCC achieved accurate preclustering for contigs containing single-copy marker genes

Since ImputeCC relies on the information provided by preliminary bins for final contig clustering, the quality of these preliminary bins, as established during the preclustering step, holds a pivotal role in affecting the final binning results of ImputeCC. Mock metaHi-C datasets were created by combining simulated Hi-C reads with real shotgun sequencing data from a manually curated microbial community (see Subsection 3.1). The shotgun data were obtained from the Illumina HiSeq 3000, ONT MinION R9, and PacBio Sequel II platforms. These datasets, named 'mock Illumina', 'mock Nanopore', and 'mock PacBio', each comprised a combination of simulated Hi-C reads and real shotgun reads corresponding to the specific sequencing platform. Since the ground truth of all contigs from the mock metaHi-C datasets were known, we could leverage the mock datasets to assess the quality of the preclustering of preliminary bins. Specifically, we calculated the Adjusted Rand Index (ARI) clustering evaluation metric (Supplementary Note 1) for preliminary bins derived from the mock Illumina, Nanopore, and

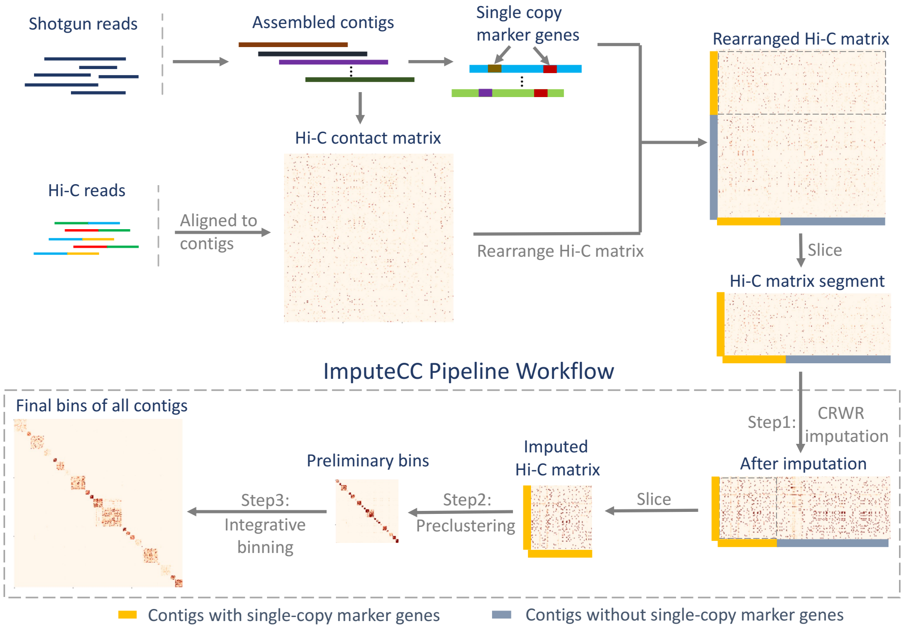

Fig. 1 Overview of the ImputeCC. Given an input of the metagenomic Hi-C contact matrix and contigs containing single-copy marker genes, ImputeCC initiates the imputation of the metaHi-C contact matrix using a new constrained random walk with restart (CRWR) algorithm, specifically limiting random walks to originate from contigs with marker genes. Subsequently, ImputeCC segregates and retains the imputed contact matrix exclusively for marker-gene-containing contigs, using it in conjunction with the characteristics of single-copy marker genes to effectively precluster these contigs as preliminary bins. Finally, ImputeCC applies the Leiden clustering method to group all assembled contigs, with insights from the preliminary bins guiding the optimization of the binning process.

PacBio datasets, resulting in values of 0.976, 0.975, and 0.988, respectively (Fig. 2a). These values indicated that ImputeCC could accomplish precise preclustering for contigs with single-copy marker genes. Furthermore, we performed preclustering directly using NormCC-normalized Hi-C contacts, omitting the imputation step. In this context, the ARI values for preliminary bins derived from the three mock datasets were decreased to 0.783, 0.903, and 0.775, respectively (Fig. 2a), underscoring the significant enhancement in the construction of preliminary bins achieved through our CRWR imputation.

2.3 ImputeCC retrieved the most high-quality genomes from the mock metaHi-C datasets

We first conducted a comparative evaluation of ImputeCC binning against VAMB [24], MetaTOR [2], bin3C [11], and the MetaCC binning module (referred to as

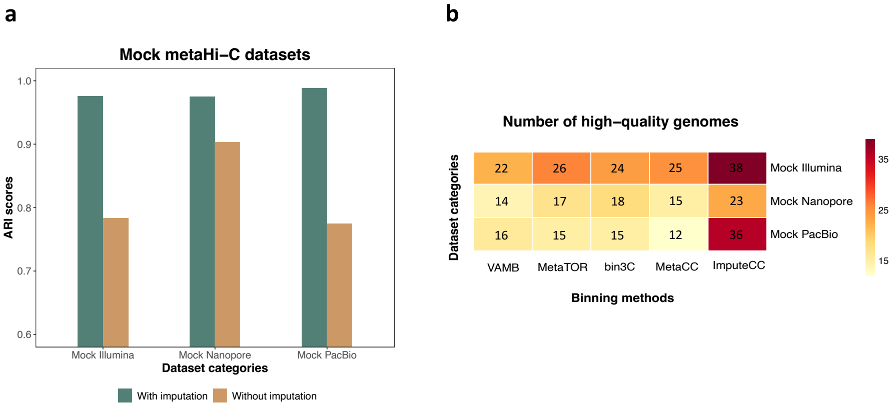

Fig. 2 Benchmarking using the three mock metaHi-C datasets. (a) Assessing the quality of preliminary bins using ARI. ImputeCC accurately grouped marker-gene-containing contigs while the CRWR imputation markedly improved the preclustering performance. (b) ImputeCC outperformed other binners on all the three mock metaHi-C datasets with respect to the number of retrieved high-quality MAGs (completeness $\geq$ 90% and contamination $\leq$ 5%). The evaluation criteria of completeness and contamination for MAGs recovered from the mock datasets are detailed in Subsection 3.4.

MetaCC) [15] using the three mock metaHi-C datasets. In addition to VAMB, a popular shotgun-based binning tool that utilizes sequence composition and coverage information, three other tools in consideration are Hi-C-based. It is important to note that another publicly available Hi-C-based binner HiCBin [14] was excluded from the benchmarking study on the mock datasets due to its inability to converge when applied to the mock Nanopore and PacBio datasets. As shown in Fig. 2b, ImputeCC demonstrated a remarkable ability to reconstruct a markedly larger number of high-quality genomes (completeness $\geq$ 90% and contamination $\leq$ 5%) across all the three mock datasets. Specifically, ImputeCC outperformed the second-highest result by 46.2%, 27.8%, and 125% in terms of high-quality genome reconstruction for the mock Illumina, Nanopore, and PacBio datasets, respectively. Notably, the number of mapped Hi-C read pairs for the mock Nanopore dataset was considerably lower in comparison to the mock Illumina and PacBio datasets (Supplementary Table 1), which can be attributed to the relatively higher error rate associated with Nanopore R9 long reads. This disparity in read mapping could be one of the contributing factors for ImputeCC retrieving a comparatively lower number of high-quality genomes from the mock Nanopore dataset. Finally, we evaluated ImputeCC's stability against Hi-C sequencing depth by downsampling the Hi-C read pairs from 10 million to 5 million in the mock datasets. The recovery of high-quality MAGs slightly declined from 38 to 36 in the Illumina dataset and from 23 to 21 in the Nanopore dataset, while the PacBio dataset consistently yielded 36 MAGs.

These results highlighted ImputeCC's resilience to reduced Hi-C read counts, ensuring its reliable performance in the mock metaHi-C datasets.

2.4 ImputeCC markedly outperformed existing binners on real metaHi-C datasets

To validate ImputeCC on real metaHi-C data, we applied it to two short-read and two long-read metaHi-C datasets from four different environments: human gut, wastewater, cow rumen, and sheep gut. Here, we compared ImputeCC to all four publicly-available Hi-C-based binners, namely HiCBin, MetaTOR, bin3C, and MetaCC, in addition to VAMB. Given the absence of reference genomes in real-world datasets, we utilized the CheckM2 [9] to evaluate the completeness and contamination of the recovered bins (see Subsection 3.4). The results from the two long-read metaHi-C datasets are presented in Fig. 3, while those from the two short-read metaHi-C datasets can be found in Supplementary Fig. 1. In all cases, ImputeCC recovered more high-quality (completeness $\geq$ 90% and contamination $\leq$ 5%) and medium-quality (completeness $\geq$ 50% and contamination $\leq$ 10%) bins than the alternatives considered. Notably, the sheep gut long-read metaHi-C dataset, owing to its high complexity, posed a greater challenge. ImputeCC binning retrieved 408 high-quality MAGs, markedly outperforming VAMB, HiCBin, MetaTOR, bin3C, and MetaCC with an increase of 235 (135.8%), 321 (369%), 279 (216.3%), 160 (64.5%), and 82 (25.2%), respectively (Fig. 3a). ImputeCC was also able to recover 125.8%, 279.8%, 91.1%, 120.1% and 23.1% more medium-quality bins than VAMB, HiCBin, MetaTOR, bin3C, and MetaCC, respectively (Fig. 3b).

Moreover, we explored the capability of different binners to capture the species diversity in microbial samples by annotating all medium-quality and high-quality bins generated by different binners on all real metaHi-C datasets using GTDB-TK [7] (see Subsection 3.5). As shown in Fig. 3c and Supplementary Fig. 1c, medium-quality bins derived from ImputeCC represented a markedly larger taxonomic diversity at the species level on all datasets. We further conducted a detailed comparative analysis of the high-quality MAGs retrieved from the sheep gut long-read metaHi-C dataset. We employed Mash [25] to identify cases where ImputeCC binning and three other Hi-C-based binning tools (MetaTOR, bin3C, and MetaCC) retrieved identical high-quality MAGs on the sheep gut long-read metaHi-C dataset (see Subsection 3.5). Notably, the majority of high-quality MAGs obtained through other Hi-C-based binning tools were also successfully recovered by ImputeCC (Supplementary Fig. 2a). In contrast, ImputeCC binning went beyond by reconstructing a substantial number of high-quality MAGs that remained inaccessible to the other binning tools. Further annotation analyses of the high-quality MAGs demonstrated ImputeCC recovered more distinct taxa at various taxonomic levels compared to Hi-C-based alternatives, including bin3C, MetaTOR, and MetaCC (Supplementary Fig. 2b).

Finally, ImputeCC's analysis at the genus level, leveraging its recovered high-quality MAGs, has unveiled significant insights into microbial composition of the sheep gut microbiota (Supplementary Note 2). Within this complex ecosystem, ImputeCC highlighted the dominance of the *Bacteroides* genus, known for influencing intestinal immunity [31,36], and uniquely detected critical species within it, such as *Bacteroides uniformis* and *Bacteroides vulgatus*. It was also the only tool to uncover the *Tidjanibacter* genus and extensively characterized the *Alistipes* genus, revealing species with potential roles in the sheep gut ecosystem and suggesting a broader species diversity. These capabilities demonstrate ImputeCC's unparalleled contribution to elucidating the sheep gut's microbial composition and its functional significance.

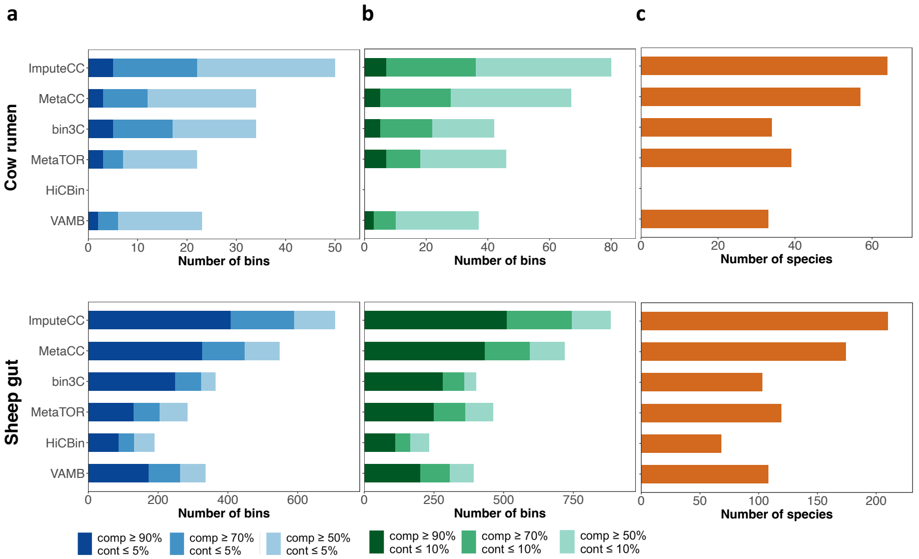

Fig. 3 Benchmarking using the real cow rumen and sheep gut long-read metaHi-C datasets. (a) The number of MAGs with varying completeness (comp) and contamination (cont) $\leq$ 5%. ImputeCC consistently outperforms other binning tools, producing a greater number of high-quality bins in both long-read metaHi-C datasets. (b) The number of MAGs with varying completeness and contamination $\leq$ 10%. ImputeCC returned more medium-quality bins when compared to alternative methods for both datasets. (c) Comparative analysis of the taxonomic diversity at the species level within medium-quality bins obtained by different binning tools. ImputeCC's binning approach stands out by capturing the broadest range of microbial species in medium-quality MAGs.

2.5 Running time analysis of the ImputeCC

On an Intel Xeon Processor E5-2665 with a clock speed of 2.40 GHz and 50 GB of allocated memory, the ImputeCC pipeline spent 64 min, 204 min, 25 min, and 2,115 min on the human gut short-read, wastewater short-read, cow rumen long-read, and sheep gut long-read metaHi-C datasets, respectively.

3 Materials and Methods

3.1 Datasets

Mock metaHi-C datasets. The mock community sequencing data were downloaded from the European Nucleotide Archive under project ID PRJEB52977 [23]. The mock community comprises 71 strains representing 69 distinct species and underwent comprehensive sequencing using the Illumina HiSeq 3000, ONT MinION R9, and PacBio Sequel II platforms, generating three different shotgun libraries. The specific accession numbers and sizes of these three shotgun libraries are shown in Supplementary Table 2. After filtering the incomplete reference genomes (Supplementary Note 3), we obtained reference genomes of 66 distinct species for the following experiments. The abundances of all species were available from the supplementary data of [23]. Since the original dataset lacked Hi-C sequencing reads, we employed sim3C (v0.2) [10] to simulate metagenomic Hi-C reads based on the 66 reference genomes and their known abundances in the mock community, utilizing parameters '-n 10000000 -l 150 -e MluCI -e Sau3AI -m hic –insert-sd 20 –insert-mean 350 –insert-min 150 –linear –simple-reads'. Subsequently, we combined the same simulated Hi-C library with the three shotgun libraries, respectively, to construct three mock metaHi-C datasets. These mock Hi-C datasets were named according to the shotgun library incorporated in the mock dataset, resulting in the 'mock Illumina,' 'mock PacBio,' and 'mock Nanopore' metaHi-C datasets. Each mock dataset comprised real shotgun reads sequenced from a known mock community, along with simulated Hi-C reads.

Real metaHi-C datasets. Four publicly-available real metaHi-C datasets were utilized in this study, comprising two short-read metaHi-C datasets and two long-read metaHi-C datasets. The specific sizes of the raw datasets are detailed in Supplementary Table 3.

The two short-read metaHi-C datasets were derived from the human gut (BioProject: PRJNA413092) [27] and wastewater (BioProject: PRJNA506462) [32] samples, respectively. Each short-read metaHi-C dataset consisted of both shotgun and Hi-C libraries originating from the same sample source. The construction of Hi-C sequencing libraries involved the use of restriction endonucleases Sau3AI and MluCI. Sequencing of both the shotgun and Hi-C libraries was carried out on Illumina platforms, producing 150-base pair reads. The two long-read metaHi-C datasets were obtained from cow rumen (BioProject: PRJNA507739) [4] and sheep gut (BioProject: PRJNA595610) [3] samples, respectively. The cow rumen long-read metaHi-C dataset comprised uncorrected PacBio long-read libraries and Hi-C libraries. The error-prone PacBio long reads were generated using both the PacBio RSII and PacBio Sequel platforms. Hi-C libraries for this dataset were prepared using the Sau3AI and MluCI restriction enzymes and subsequently sequenced on an Illumina HiSeq 2000, producing 80-base pair reads. The sheep gut long-read metaHi-C dataset consisted of PacBio circular consensus sequencing (CCS) long-read libraries and Hi-C sequencing libraries. The PacBio CCS long reads, characterized by high accuracy with average Q scores exceeding 20, were referred to as HiFi reads. Distinct Hi-C libraries for the sheep gut long-read metaHi-C dataset were generated using the Sau3AI and MluCI restriction enzymes and sequenced at a length of 150 base pairs.

3.2 Data preprocessing

We first conduct essential read cleaning procedures using 'bbduk' from the BBTools suite (v37.25) [6] to address issues such as adaptor sequences, low-quality reads, and PCR duplication (Supplementary Note 4). For each metaHi-C dataset, reads from the shotgun library are assembled into longer contigs (Supplementary Note 5). After assembly, processed paired-end Hi-C reads are aligned to these contigs using BWA-MEM (v0.7.17) [21] with the '-5SP' parameter to prioritize the alignment with the lowest read coordinate as the primary alignment. Subsequent alignment filtering steps include the removal of unmapped reads, secondary and supplementary alignments, and alignments with low quality (nucleotide match length < 30 or mapping score < 30). We count Hi-C read pairs aligned to two contigs as raw Hi-C contacts between contigs and those contigs with fewer than two Hi-C contacts are excluded. Raw Hi-C contacts are normalized by NormCC [15] with default parameters to eliminate the systematic biases derived from the number of restriction sites, contig length, and coverage.

3.3 The framework of ImputeCC binning

Detect assembled contigs with single-copy marker genes. Similar to [34], we identify single-copy marker genes, which are genes typically found as single copies in the majority of genomes [1] within the assembled contigs. We accomplish this by employing FragGeneScan [30] and HMMER (v3.3.2) [16] (Supplementary Note 6).

Impute the metagenomic Hi-C contact matrix for contigs containing marker genes. According to the second principle of preclustering outlined in Subsection 2.1, the effective preclustering of contigs with single-copy marker genes partially depends on the expectation that marker-gene-containing contigs can be reliably linked through robust Hi-C interactions if they come from the same genome. However, this expectation encounters a practical limitation attributed to the localized characteristics of proximity ligations, which implies that even when two contigs share the same genomic origin, they may fail to establish Hi-C contacts if they are not in close spatial proximity within the cell, thereby contributing to the sparsity of the metagenomic Hi-C contact matrix [13]. To facilitate improved connections among marker-gene-containing contigs originating from the same genome through Hi-C interactions, we design a metagenomic

Hi-C contact matrix imputation method. This involves employing a constrained random walk with restart (CRWR) technique to amplify the within-cell Hi-C signals specially for marker-gene-containing contigs. Specifically, we define m and n as the number of contigs containing single-copy marker genes and the total number of assembled contigs, respectively. Let H denote the NormCC-normlized Hi-C contact matrix, where the entry H_{ij} represents the normalized Hi-C contacts between contig i and j. We first set all diagonal entries of H as zero and reorganize the matrix H by moving the contigs containing marker genes to the first m rows and m columns consistently and denote the reorganized matrix as H'. Then, the reorganized matrix H' is further normalized by its row sum and let M denote the matrix after the row-sum normalization, i.e.,

$$M_{ij} = \frac{H'_{ij}}{\sum_k H'_{ik}}. \tag{1}$$

We use $N^{(t)}$ to represent the matrix after the t-th iteration of random walk with restart and limit that all random walks can only start from the contigs with marker genes. Mathematically, the random walk starts from the initial matrix $N^{(0)} = \left[\begin{array}{c:c} I_{m\times m} & 0_{m\times(n-m)} \\ \hdashline 0_{(n-m)\times m} & 0_{(n-m)\times(n-m)} \end{array}\right]_{n\times n}$, and $N^{(t)}$ is computed recursively by the following:

$$N^{(t)} = (1-p)\cdot N^{(t-1)}\cdot M + p\cdot T, \tag{2}$$

where $T = N^{(0)}$ denotes the restarting matrix, and p (default, 0.5) serves as the restarting probability used to maintain a balance between the influence of global and local network structures. Notably, since the last $n-m$ rows of all iteration matrices N are kept to be zero, the formula (2) can be simplified by omitting the last $n-m$ rows of N and T. As a result, the new RWR can be represented as

$$\begin{aligned} \tilde{N}^{(0)} &= \tilde{T} = [I_{m\times m}|0_{m\times(n-m)}]_{m\times n}, \\ \tilde{N}^{(t)} &= (1-p)\cdot\tilde{N}^{(t-1)}\cdot M + p\cdot\tilde{T}. \end{aligned} \tag{3}$$

To avoid the imputed matrix becoming too dense, we only retain the largest τ percent (default, 20) of non-zero entries in $\tilde{N}^{(t)}$ after each iteration, i.e.,

$$\tilde{N}^{(t)} = \tilde{N}^{(t)} \circ \mathbf{1}_{\{\tilde{N}^{(t)} > C_t^\tau\}}, \tag{4}$$

where C_t^τ is a $(100-\tau)$-th percentile of all non-zero entries in $\tilde{N}^{(t)}$; $\mathbf{1}$ represents an indicator matrix and $\mathbf{1}_{ij} = 1$ only if $\tilde{N}^{(t)}_{ij} > C_t^\tau$; $\circ$ denotes the mathematical operator of element-wise matrix multiplication.

Let $\delta_t = ||\tilde{N}^{(t)} - \tilde{N}^{(t-1)}||_2$. The iteration ends if either of the following two conditions is satisfied:

- δ_t<0.01,
- Early stop if $\delta_t - \delta_{t-1} < 0.001$ for a consecutive five times.

Let $\hat{N}$ denote the final matrix output from the imputation. Then the first m columns of $\hat{N}$, denoted by $P_{m\times m}$, can exactly represent the imputed Hi-C matrix for contigs with marker genes. Finally, we transform the matrix P to a symmetric matrix P' and further normalize P' to eliminate the contigs' coverage biases derived from the imputation using the Square Root Vanilla Coverage (sqrtVC) method [28], i.e.,

$$\begin{aligned} P' &= P + P^T, \\ Q &= D^{-\frac{1}{2}} P' D^{-\frac{1}{2}}, \end{aligned} \tag{5}$$

where D is a diagonal matrix where each elements D_{ii} is the sum of the i-th row of P'.

Precluster contigs with marker genes as preliminary bins. Leveraging the imputed Hi-C matrix Q as well as the characteristics of single-copy marker genes, we would like to accurately precluster contigs with marker genes as preliminary bins following the two principles outlined in Subsection 2.1. Specifically, we first sort all categories of detected marker genes by the number of contigs containing the marker genes. If several marker genes correspond to the same number of contigs, they are further sorted by the gene length. Then, we use a greedy strategy to iteratively construct the preliminary bins as follows:

- Initialization: choose all contigs from the first marker gene and initialize preliminary bin set, denoted by $\mathcal{B}$, with each bin containing one contig.
- Iteration: in the k-th iteration, we select all contigs containing the k-th marker gene and only handle contigs that have not been assigned to any preliminary bins in $\mathcal{B}$. Let $\mathcal{C}$ denote the set of contigs to be processed in the iteration. We then define the contig-to-bin Hi-C similarity between a contig $c \in \mathcal{C}$ and a bin $B \in \mathcal{B}$ as:

  $$S_{c,B} = \frac{\sum_{c_1 \in B} Q_{c,c_1}}{\#B} \tag{6}$$

  where c_1 denotes the contigs in the preliminary bin B, Q_{c,c_1} is the imputed Hi-C contacts between contigs c and c_1 and $\#B$ represents the number of contigs in B. In this way, we can construct a undirected bipartite graph, where the top nodes are contigs from the set $\mathcal{C}$ and the bottom nodes are preliminary bins from the set $\mathcal{B}$. The weighted edges between top nodes and bottom nodes represent the contig-to-bin Hi-C similarity. To assign the contigs to preliminary bins, we leverage the Karp's algorithm [20] to find a maximum-weight matching between contigs and preliminary bins. For each contig in the set $\mathcal{C}$ with a matching preliminary bin, if the contig-to-bin Hi-C similarity is above the median of non-zero entries in the imputed matrix Q, we attribute the contig to its matching preliminary bin; otherwise, the contig will be discarded. Finally, we add all unmatched contigs to $\mathcal{B}$ as new preliminary bins, with each new bin containing one unmatched contig.
- Repeat the iteration step until all marker genes are processed.

Leiden clustering for all contigs using the information of preliminary bins. We apply the Leiden community detection algorithm [33] to the NormCC-normalized Hi-C contact matrix H to cluster all assembled contigs, using the preliminary bin set as an initial framework. The Leiden algorithm iteratively merges and refines communities to maximize modularity, a metric that quantifies the partitioning quality. To incorporate preliminary bin information, we initialize contig memberships based on preliminary bins, ensuring that contigs from the same preliminary bin are placed within the same community, while contigs not associated with any preliminary bins are initially assigned to individual communities. Throughout the Leiden iterations, these assignments for contigs from preliminary bins remain fixed. Consequently, contigs from the same preliminary bin coalesce into the same cluster, while those from different preliminary bins form distinct clusters after the Leiden clustering.

Moreover, since the Leiden algorithm is modularity-based, we select a flexible modularity function based on the Reichardt and Bornholdt's Potts model [29]. Notably, the resolution parameter r in the modularity function (Supplementary Note 7) is a hyper-parameter that determines the relative importance assigned to the configuration null part compared to the links within the communities. To ascertain the optimal resolution parameter, we conduct parallel executions of the Leiden algorithm using various resolution values and automatically select the most favorable outcome. Specifically, we identify lineage-specific genes, which act as indicators of genome quality, through the application of the CheckM (v1.1.3) [26] function 'checkm analyze'. Consequently, for any given contig bin, we employ the same evaluation strategy as CheckM to efficiently estimate its precision and recall (Supplementary Note 8). Subsequently, for each resolution parameter value, we count the number of genomic bins with precision exceeding 95% and recall surpassing 90%, 70%, and 50%, respectively. Finally, we automatically select the resolution value that maximizes the sum of three count numbers as the optimal choice.

Integrative strategy to obtain the final bins. It is essential to acknowledge that the preliminary bins may not be entirely accurate. This can occur, for instance, in cases where genome coverage is insufficient or marker genes are fragmented into several pieces. Furthermore, our clustering strategy in previous steps may exacerbate these mis-binnings arising from the preliminary bin assignments. Consequently, it is still meaningful to apply the Leiden algorithm to cluster contigs independently, without relying on the preliminary bin information. The selection of the resolution parameter follows the same methodology as previously described. We denote the resulting bin sets as $\mathcal{F}_{\text{pre}}$ and $\mathcal{F}_{\text{null}}$ for the Leiden clustering with and without preliminary bin information, respectively. We then implement an iterative greedy strategy to integrate these two bin sets. Specifically, in each iteration of this integrative procedure, we assess the quality of all existing MAGs from $\mathcal{F}_{\text{pre}}$ and $\mathcal{F}_{\text{null}}$ using the metric:

$$\text{Recall} - 2 \times (100 - \text{Precision}). \tag{7}$$

The MAG displaying the highest estimated quality across both bin sets is selected for further consideration. In situations where two or more MAGs exhibit identical estimated quality scores, ties are resolved by selecting the MAG with the greatest N50 statistic and bin size. Following the selection of a MAG, it is moved from the corresponding bin set to the final bin set, and any contigs belonging to the selected MAG are also removed from the other bin set, if present. This iterative procedure continues until the highest quality MAG identified falls below 10. Finally, we can obtain the final bin set through the integration.

3.4 Evaluating the quality of recovered MAGs from the mock and real metaHi-C datasets

For the mock metaHi-C datasets, where all species within the mock microbial community were known, the species identity of the assembled contigs could be determined (Supplementary Note 9). Then, we can define the the completeness and contamination of each MAG recovered from the mock datasets. Specifically, for each MAG, we segregated the lengths of contigs according to their respective reference genomes and attributed the MAG to the reference genome with the largest cumulative contig length, denoted as $L(q)$. The length of the corresponding reference genome was denoted as $L(r)$, and the total length of the MAG was referred to as $L(v)$. The completeness of a MAG was quantified as $\frac{L(q)}{L(r)}$, while the contamination of a MAG was defined as $\frac{L(v)-L(q)}{L(v)}$. Finally, we classified high-quality genomes obtained from the mock datasets as those MAGs with completeness $\geq 90\%$ and contamination $\leq 5\%$.

For the real metaHi-C datasets, since the actual genomes are unknown in real samples, we applied CheckM2 [9] to evaluate the completeness and contamination of retrieved MAGs. CheckM2 is an advanced machine learning-based method for assessing the quality of draft genomic bins, offering improved accuracy and computational speed compared to existing tools [9]. Based on the CheckM2 assessments of completeness and contamination, we categorized the resolved MAGs from real metaHi-C datasets as high-quality if their completeness $\geq 90\%$ and contamination $\leq 5\%$, while MAGs were designated as medium-quality if their completeness $\geq 50\%$ and contamination $\leq 10\%$.

3.5 MAG analyses on real metaHi-C datasets

To assess the capacity of various binning methods in capturing taxonomic diversity within real metaHi-C datasets, we performed taxonomic annotation on all high-quality and medium-quality bins using GTDB-TK (v2.1.0, Release: R207 v2) [7] with the function 'classify_wf' to extract the taxonomic information of the MAGs recovered by different binning methods.

Furthermore, to identify overlapping high-quality bins retrieved from the sheep gut long-read metaHi-C dataset between ImputeCC binning and other Hi-C-based binning approaches, we utilized Mash (v2.2) [25] with 10,000 sketches per bin to calculate the Mash distance between high-quality bins from different bin sets. Bins with a Mash distance below 0.01 were considered MAGs originating from the same genome.

3.6 Other binners used in benchmarking

All binners used for comparison, i.e., VAMB (v3.0.3) [24], HiCBin (v1.1.0) [14], MetaTOR (v1.1.4) [2], bin3C (v0.1.1) [11], and MetaCC (v1.1.0) [15] were executed with default parameters on all mock and real metaHi-C datasets.

4 Discussions

In this work, we developed ImputeCC, an integrative Hi-C-based contig binning methods. ImputeCC combines Hi-C interactions with the intrinsic discriminative potential of single-copy marker genes by preclustering marker-gene-containing contigs as preliminary bins. To enhance the Hi-C connectivity of marker-gene-containing contigs, ImputeCC introduces a constrained random walk with restart (CRWR) approach to impute the metaHi-C contact matrix. Finally, ImputeCC employs Leiden clustering to group all assembled contigs, optimizing the binning process by leveraging information from the preliminary bins. Evaluations of ImputeCC using a wide range of diverse mock/real metaHi-C datasets have demonstrated its effectiveness for retrieving reference-quality MAGs and shown its potential to unravel the structure of microbial ecosystems and their resident microorganisms. Notably, we utilized CheckM2 in assessing the binning performance for the four real metaHi-C datasets. Although CheckM2 represents the most advanced software for evaluating bin quality in real metagenomic samples, it is essential to delve further into the accuracy of this machine-learning-based validation method in reflecting the true completeness and contamination levels of the recovered MAGs. Moreover, previous research has established the efficacy of Hi-C-based binning over shotgun-based approaches [11,14]. Accordingly, our benchmarking analyses focus on Hi-C-based methods, comparing ImputeCC with similar tools and including VAMB as a reference shotgun-based method.

ImputeCC offers several promising avenues for expansion. For instance, when dealing with large MAGs characterized by high abundances, there is potential in imputing normalized Hi-C contacts for contigs within these MAGs to facilitate the scaffolding process. Moreover, exploring imputation methods that consider additional information, such as the sequence composition of contigs, could yield improved imputation results.

Acknowledgments. Y.D. and F.S. conceived the ideas and designed the study. Y.D. implemented the methods, carried out the computational analyses, and drafted the manuscript. Y.D. and W.Z. developed the software. All authors modified and finalized the paper. The research is partially funded by NSF grant EF-2125142. The authors declare no competing interests.

References

1. Albertsen, M., Hugenholtz, P., Skarshewski, A., Nielsen, K.L., Tyson, G.W., Nielsen, P.H.: Genome sequences of rare, uncultured bacteria obtained by differential coverage binning of multiple metagenomes. Nat Biotechnol **31**(6), 533–538 (2013)
2. Baudry, L., Foutel-Rodier, T., Thierry, A., Koszul, R., Marbouty, M.: MetaTOR: a computational pipeline to recover high-quality metagenomic bins from mammalian gut proximity-ligation (me) libraries. Front Genet **10**, 753 (2019)
3. Bickhart, D.M., Kolmogorov, M., Tseng, E., Portik, D.M., Korobeynikov, A., Tolstoganov, I., Uritskiy, G., Liachko, I., Sullivan, S.T., Shin, S.B., et al.: Generating lineage-resolved, complete metagenome-assembled genomes from complex microbial communities. Nat Biotechnol **40**(5), 711–719 (2022)
4. Bickhart, D.M., Watson, M., Koren, S., Panke-Buisse, K., Cersosimo, L.M., Press, M.O., Van Tassell, C.P., Van Kessel, J.A.S., Haley, B.J., Kim, S.W., et al.: Assignment of virus and antimicrobial resistance genes to microbial hosts in a complex microbial community by combined long-read assembly and proximity ligation. Genome Biol **20**, 153 (2019)
5. Burton, J.N., Liachko, I., Dunham, M.J., Shendure, J.: Species-level deconvolution of metagenome assemblies with Hi-C–based contact probability maps. G3 (Bethesda) **4**(7), 1339–1346 (2014)
6. Bushnell, B.: BBMap: a fast, accurate, splice-aware aligner. Tech. rep., Lawrence Berkeley National Lab.(LBNL), Berkeley, CA (United States) (2014)
7. Chaumeil, P.A., Mussig, A.J., Hugenholtz, P., Parks, D.H.: GTDB-Tk v2: memory friendly classification with the genome taxonomy database. Bioinformatics **38**(23), 5315–5316 (2022)
8. Chen, Y., Wang, Y., Paez-Espino, D., Polz, M.F., Zhang, T.: Prokaryotic viruses impact functional microorganisms in nutrient removal and carbon cycle in wastewater treatment plants. Nat Commun **12**, 5398 (2021)
9. Chklovski, A., Parks, D.H., Woodcroft, B.J., Tyson, G.W.: CheckM2: a rapid, scalable and accurate tool for assessing microbial genome quality using machine learning. Nat Methods **20**, 1203–1212 (2023)
10. DeMaere, M.Z., Darling, A.E.: Sim3C: simulation of Hi-C and Meta3C proximity ligation sequencing technologies. GigaScience **7**(2), gix103 (2018)
11. DeMaere, M.Z., Darling, A.E.: bin3C: exploiting Hi-C sequencing data to accurately resolve metagenome-assembled genomes. Genome Biol **20**, 46 (2019)
12. Du, Y., Fuhrman, J.A., Sun, F.: ViralCC retrieves complete viral genomes and virus-host pairs from metagenomic Hi-C data. Nat Commun **14**, 502 (2023)
13. Du, Y., Laperriere, S.M., Fuhrman, J., Sun, F.: Normalizing Metagenomic Hi-C Data and Detecting Spurious Contacts Using Zero-Inflated Negative Binomial Regression. J Comput Biol **29**, 106–120 (2022)
14. Du, Y., Sun, F.: HiCBin: binning metagenomic contigs and recovering metagenome-assembled genomes using Hi-C contact maps. Genome Biol **23**, 63 (2022)
15. Du, Y., Sun, F.: MetaCC allows scalable and integrative analyses of both long-read and short-read metagenomic Hi-C data. Nat Commun **14**, 6231 (2023)
16. Finn, R.D., Clements, J., Eddy, S.R.: HMMER web server: interactive sequence similarity searching. Nucl Acids Res **39**(suppl_2), W29–W37 (2011)
17. Handelsman, J.: Metagenomics: application of genomics to uncultured microorganisms. Microbiol Mol Biol Rev **68**(4), 669–685 (2004)
18. Hugenholtz, P., Tyson, G.W.: Metagenomics. Nature **455**(7212), 481–483 (2008)

19. Hugerth, L.W., Larsson, J., Alneberg, J., Lindh, M.V., Legrand, C., Pinhassi, J., Andersson, A.F.: Metagenome-assembled genomes uncover a global brackish microbiome. Genome Biol **16**, 279 (2015)
20. Karp, R.M.: An algorithm to solve the m× n assignment problem in expected time O (mn log n). Networks **10**(2), 143–152 (1980)
21. Li, H.: Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. arXiv (2013). https://doi.org/10.48550/arXiv.1303.3997
22. Marbouty, M., Cournac, A., Flot, J.F., Marie-Nelly, H., Mozziconacci, J., Koszul, R.: Metagenomic chromosome conformation capture (meta3C) unveils the diversity of chromosome organization in microorganisms. eLife **3**, e03318 (2014)
23. Meslier, V., Quinquis, B., Da Silva, K., Plaza Oñate, F., Pons, N., Roume, H., Podar, M., Almeida, M.: Benchmarking second and third-generation sequencing platforms for microbial metagenomics. Sci Data **9**(1), 694 (2022)
24. 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.: Improved metagenome binning and assembly using deep variational autoencoders. Nat Biotechnol **39**, 555–560 (2021)
25. Ondov, B.D., Treangen, T.J., Melsted, P., Mallonee, A.B., Bergman, N.H., Koren, S., Phillippy, A.M.: Mash: fast genome and metagenome distance estimation using MinHash. Genome Biol **17**, 132 (2016)
26. Parks, D.H., Imelfort, M., Skennerton, C.T., Hugenholtz, P., Tyson, G.W.: CheckM: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. Genome Res **25**(7), 1043–1055 (2015)
27. 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., Liachko, I.: Hi-C deconvolution of a human gut microbiome yields high-quality draft genomes and reveals plasmid-genome interactions. bioRxiv (2017). https://doi.org/10.1101/198713
28. Rao, S.S., Huntley, M.H., Durand, N.C., Stamenova, E.K., Bochkov, I.D., Robinson, J.T., Sanborn, A.L., Machol, I., Omer, A.D., Lander, E.S., et al.: A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping. Cell **159**(7), 1665–1680 (2014)
29. Reichardt, J., Bornholdt, S.: Statistical mechanics of community detection. Phys Rev E **74**(1), 016110 (2006)
30. Rho, M., Tang, H., Ye, Y.: FragGeneScan: predicting genes in short and error-prone reads. Nucl Acids Res **38**(20), e191–e191 (2010)
31. Routy, B., Gopalakrishnan, V., Daillère, R., Zitvogel, L., Wargo, J.A., Kroemer, G.: The gut microbiota influences anticancer immunosurveillance and general health. Nat Rev Clin Oncol **15**, 382–396 (2018)
32. Stalder, T., Press, M.O., Sullivan, S., Liachko, I., Top, E.M.: Linking the resistome and plasmidome to the microbiome. ISME J **13**(10), 2437–2446 (2019)
33. Traag, V.A., Waltman, L., Van Eck, N.J.: From Louvain to Leiden: guaranteeing well-connected communities. Sci Rep **9**, 5233 (2019)
34. Wu, Y.W., Tang, Y.H., Tringe, S.G., Simmons, B.A., Singer, S.W.: MaxBin: an automated binning method to recover individual genomes from metagenomes using an expectation-maximization algorithm. Microbiome **2**(26) (2014)
35. Yaffe, E., Relman, D.A.: Tracking microbial evolution in the human gut using Hi-C reveals extensive horizontal gene transfer, persistence and adaptation. Nat Microbiol **5**(2), 343–353 (2020)
36. Yatsunenko, T., Rey, F.E., Manary, M.J., Trehan, I., Dominguez-Bello, M.G., Contreras, M., Magris, M., Hidalgo, G., Baldassano, R.N., Anokhin, A.P., et al.: Human gut microbiome viewed across age and geography. Nature **486**, 222–227 (2012)