JOURNAL OF COMPUTATIONAL BIOLOGY
Volume 00, Number 00, 2024

Pp. 000–000
DOI: 10.1089/cmb.2024.0663

Imputing Metagenomic Hi-C Contacts Facilitates the Integrative Contig Binning Through Constrained Random Walk with Restart

YUXUAN DU, WENXUAN ZUO, and FENGZHU SUN

ABSTRACT

Metagenomic Hi-C (metaHi-C) has shown remarkable potential for retrieving high-quality metagenome-assembled genomes from complex microbial communities. Nevertheless, existing metaHi-C-based contig binning methods solely rely on Hi-C interactions between contigs, disregarding crucial biological information such as the presence of single-copy marker genes. To overcome this limitation, we introduce ImputeCC, an integrative contig binning tool optimized for metaHi-C datasets. ImputeCC integrates both Hi-C interactions and the discriminative power of single-copy marker genes to group marker-gene-containing contigs into preliminary bins. It also introduces a novel constrained random walk with restart algorithm to enhance Hi-C connectivity among contigs. Comprehensive assessments using both mock and real metaHi-C datasets from diverse environments demonstrate that ImputeCC consistently outperforms other Hi-C-based contig binning tools. A genus-level analysis of the sheep gut microbiota reconstructed by ImputeCC underlines its capability to recover key species from dominant genera and identify previously unknown genera.

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 (Handelsman, 2004; Hugenholtz and Tyson, 2008; Simon and Daniel, 2011; Streit and Schmitz, 2004). 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 (Beitel et al., 2014; Burton

Department of Quantitative and Computational Biology, University of Southern California, Los Angeles, California, USA.

An early version of this paper was published as part of the 2024 Annual International Conference on Research in Computational Molecular Biology (RECOMB).

et al., 2014; Du et al., 2023; Marbouty et al., 2014; 2021; Press et al., 2017; Yaffe and Relman, 2020). 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) (Hugerth et al., 2015) by leveraging Hi-C contacts between contigs (Baudry et al., 2019; DeMaere and Darling, 2019; Du and Sun, 2022; 2023). 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 (Chen et al., 2021; Gounot et al., 2022; Kent et al., 2020; Stalder et al., 2019). Various Hi-C-based contig binning methods have been developed, including HiCBin (Du and Sun, 2022), MetaTOR (Baudry et al., 2019), bin3C (DeMaere and Darling, 2019), and the MetaCC binning module (referred to as MetaCC) (Du and Sun, 2023). 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 (Du and Sun, 2022; Press et al., 2017).

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 (Albertsen et al., 2013), 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. We conducted a comprehensive validation of ImputeCC's performance using a combination of mock and real metaHi-C datasets. In the mock datasets, we demonstrated the effectiveness of our constrained random walk with restart (CRWR) imputation, showcasing its utility and necessity in improving the preclustering of marker-gene-containing contigs. Subsequently, we evaluated ImputeCC's performance against other publicly-available Hi-C-based binning tools using four real metaHi-C datasets sourced from diverse environments, including the human gut (Press et al., 2017), wastewater (Stalder et al., 2019), cow rumen (Bickhart et al., 2019), and sheep gut (Bickhart et al., 2022). ImputeCC's standout performance was particularly evident in the challenging sheep gut environment. In this complex setting, ImputeCC successfully retrieved an impressive total of 408 high-quality and 885 medium-quality MAGs, as assessed by the latest CheckM2 (Chklovski et al., 2023). To the best of our knowledge, this represents the largest number of reference-quality MAGs reported from a single sample. Moreover, we delved into the taxonomic diversity of the captured species in microbial samples by annotating high-quality MAGs generated by various binning methods using GTDB-TK (Chaumeil et al., 2022). ImputeCC consistently demonstrated a significantly broader taxonomic diversity at the species level across all datasets, emphasizing its ability to capture a broader range of microbial taxa. Further downstream ImputeCC's genus-level analysis of the sheep gut microbiota revealed ability of ImputeCC to recover essential species from dominant genera such as *Bacteroides*, showed its potential to detect previously unrecognized genera, and unveiled other high-quality MAGs within the *Alistipes* genus that warrant further experimental investigation to elucidate their characteristics and roles within this ecosystem.

2. METHODS

2.1. Datasets

2.1.1. Mock metaHi-C datasets. The mock community sequencing data were downloaded from the European Nucleotide Archive under project ID PRJEB52977 (Meslier et al., 2022). 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 S1. After filtering the incomplete reference genomes (Supplementary Data S1), we obtained reference genomes of 66 distinct species for the following experiments. The abundances of all species were available from the supplementary data of (Meslier et al., 2022). Since the original dataset lacked Hi-C sequencing reads, we employed sim3C (v0.2) (DeMaere and Darling, 2018) 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.

2.1.2. 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 S2.

The two short-read metaHi-C datasets were derived from the human gut (BioProject: PRJNA413092) (Press et al., 2017) and wastewater (BioProject: PRJNA506462) (Stalder et al., 2019) samples. 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) (Bickhart et al., 2019) and sheep gut (BioProject: PRJNA595610) (Bickhart et al., 2022) samples. 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.

2.2. Data preprocessing

We first conduct essential read cleaning procedures using 'bbduk' from the BBTools suite (v37.25) (Bushnell, 2014) to address issues such as adaptor sequences, low-quality reads, and PCR duplication (Supplementary Data S2). For each metaHi-C dataset, reads from the shotgun library are assembled into longer contigs (Supplementary Data S3). After assembly, processed paired-end Hi-C reads are aligned to these contigs using BWA-MEM (v0.7.17) (Li, 2013) 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 (Du and Sun, 2023) with default parameters to eliminate the systematic biases derived from the number of restriction sites, contig length, and coverage.

2.3. The framework of ImputeCC binning

2.3.1. Detect assembled contigs with single-copy marker genes. Similar to MaxBin (Wu et al., 2014), we identify single-copy marker genes, which are genes typically found as single copies in the majority of genomes (Albertsen et al., 2013) within the assembled contigs. We accomplish this by employing FragGeneScan (Rho et al., 2010) and HMMER (v3.3.2) (Finn et al., 2011) (Supplementary Data S4).

2.3.2. Impute the metagenomic Hi-C contact matrix for contigs containing marker genes. 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 (Du et al., 2022). 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 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 $N^{(0)} = \begin{bmatrix} I_{m\times m} & 0_{m\times(n-m)} \\ 0_{(n-m)\times m} & 0_{(n-m)\times(n-m)} \end{bmatrix}_{n\times n}$, 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

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

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

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

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

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

- $\delta_t < 0.01$,
- 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 (Rao et al., 2014), 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 $P^{'}$.

2.3.3. 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. 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 the 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 (Karp, 1980) 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.

2.3.4. Leiden clustering for all contigs using the information of preliminary bins. We apply the Leiden community detection algorithm (Traag et al., 2019) 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 (Reichardt and Bornholdt, 2006). Notably, the resolution parameter r in the modularity function (Supplementary Data S5) 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) (Parks et al., 2015) 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 Data S6). 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.

2.3.5. 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 Subsection 2.3.4 may exacerbate these mis-binning 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.

2.4. 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 Data S7). Then, we can define 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 ≥90% and contamination ≤5%.

For the real metaHi-C datasets, since the actual genomes are unknown in real samples, we applied CheckM2 (Chklovski et al., 2023) 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 (Chklovski et al., 2023). 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 ≥90% and contamination ≤5%, while MAGs were designated as medium-quality if their completeness ≥50% and contamination ≤10%.

2.5. 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) (Chaumeil et al., 2022) 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) (Ondov et al., 2016) 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.

2.6. Other binners used in benchmarking

All binners used for comparison, i.e., VAMB (v3.0.3) (Nissen et al., 2021), HiCBin (v1.1.0) (Du and Sun, 2022), MetaTOR (v1.1.4) (Baudry et al., 2019), bin3C (v0.1.1) (DeMaere and Darling, 2019), and MetaCC (v1.1.0) (Du and Sun, 2023) were executed with default parameters on all mock and real metaHi-C datasets.

3. RESULTS

3.1. 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. Figure 1 shows the outline of ImputeCC. The core

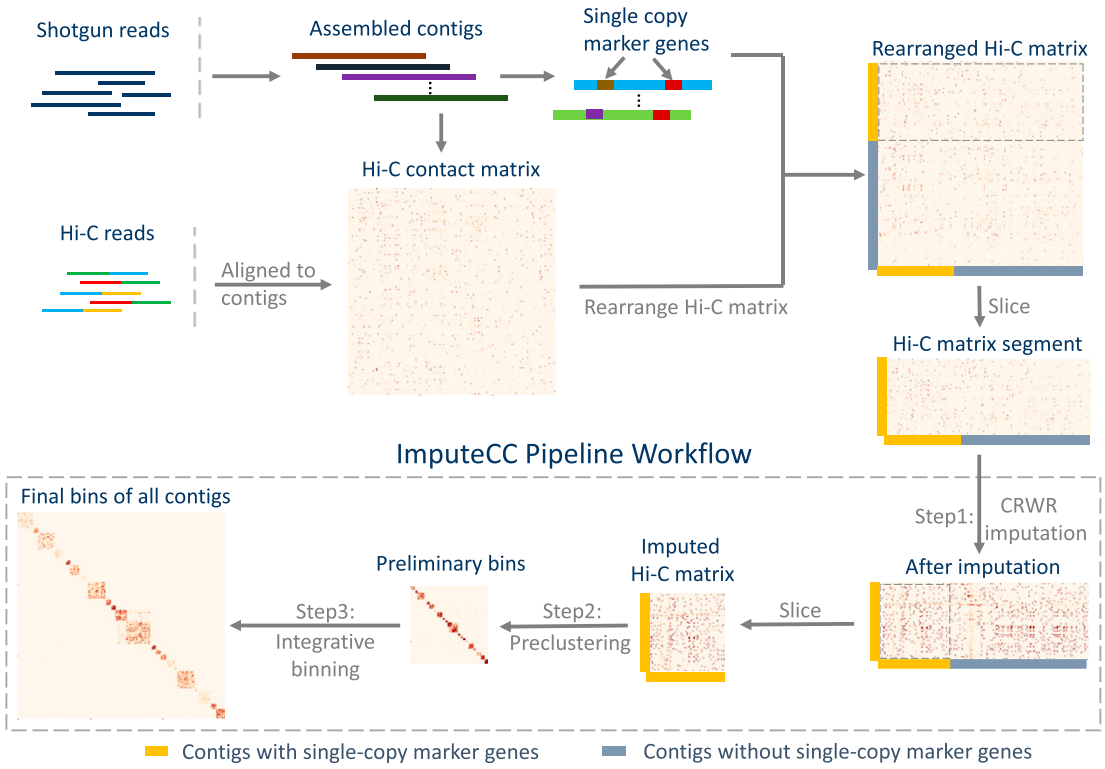

FIG. 1. Overview of the ImputeCC. Given an input of the metagenomic Hi-C contact matrix and contigs containing single-copy marker genes, ImputeCC initiates the imputation of the metaHi-C contact matrix using a new CRWR algorithm, specifically limiting random walks to originate from contigs with marker 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, the Leiden clustering method is applied by ImputeCC to group all assembled contigs, with insights from the preliminary bins guiding the optimization of the binning process. CRWR, constrained random walk with restart; metaHi-C, Metagenomic Hi-C.

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 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 (Traag et al., 2019) to group all assembled contigs, utilizing the information from preliminary bins to optimize the binning process.

3.2. 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. 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 Data S8) for preliminary bins derived from the mock Illumina, Nanopore, and PacBio datasets (see Subsection 2.1.1), resulting in values of 0.976, 0.975, and 0.988, respectively (Fig. 2a). These values indicated that ImputeCC

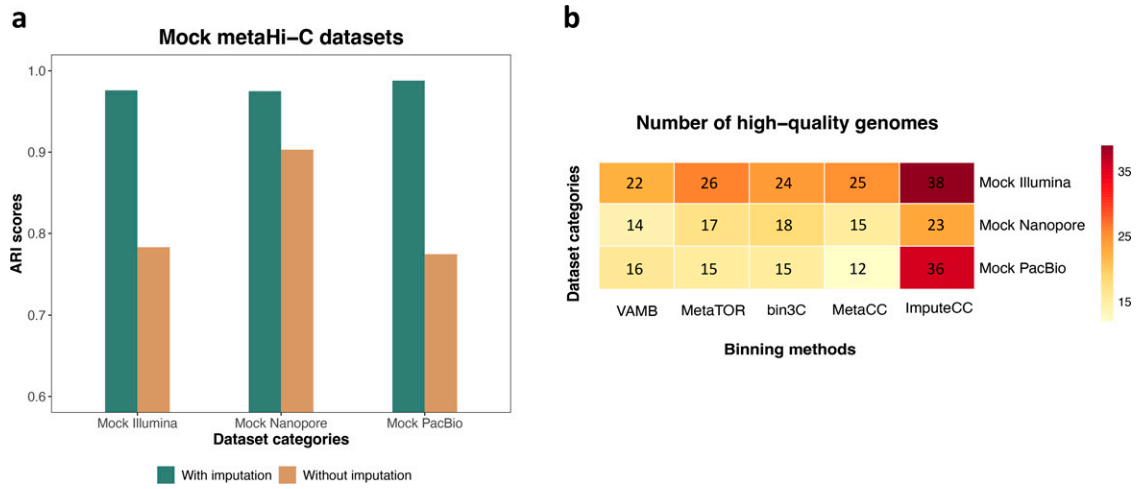

FIG. 2. Benchmarking using the three mock metaHi-C datasets. **(a)** Assessing the quality of preliminary bins using ARI. ImputeCC accurately grouped marker-gene-containing contigs while the CRWR imputation markedly improved the preclustering performance. **(b)** ImputeCC outperformed other binners on all the three mock metaHi-C datasets with respect to the number of retrieved high-quality MAGs (completeness ≥90% and contamination ≤5%). The evaluation criteria of completeness and contamination for MAGs recovered from the mock datasets are detailed in Subsection 2.4. CRWR, constrained random walk with restart; MAG, metagenome-assembled genomes; metaHi-C, Metagenomic Hi-C.

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.

3.3. ImputeCC retrieved the most High-Quality genomes from the mock metaHi-C datasets

We first conducted a comparative evaluation of ImputeCC binning against VAMB (Nissen et al., 2021), MetaTOR (Baudry et al., 2019), bin3C (DeMaere and Darling, 2019), and the MetaCC binning module (referred to as MetaCC) (Du and Sun, 2023) 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 (Du and Sun, 2022) 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 Figure 2b, ImputeCC demonstrated a remarkable ability to reconstruct a markedly larger number of high-quality genomes (completeness ≥90% and contamination ≤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 S3), 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 reads from 10 million to 5 million pairs 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.

3.4. 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 (Chklovski et al., 2023) to evaluate the completeness and contamination of the recovered bins (see Subsection 2.4). In all cases, ImputeCC recovered more high-quality (completeness ≥90% and contamination ≤5%) and medium-quality (completeness ≥50% and contamination ≤10%) bins than the alternatives considered (Fig. 3a and b). Specifically, in the human gut and wastewater short-read metaHi-C datasets, ImputeCC reconstructed 66 and 75 high quality MAGs, outperforming the second-best binner with an increase of 11 (20%) and 10 (15.4%), respectively. For the cow rumen long-read metaHi-C dataset, though bin3C was able to retrieve an equivalent number of high-quality MAGs as ImputeCC, ImputeCC excelled by recovering 90.5% more medium-quality bins. 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. 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.

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 (Chaumeil et al., 2022) (see Subsection 2.5). As shown in Figure 3c, medium-quality bins derived from ImputeCC represented a markedly larger taxonomic diversity at the species level on all datasets.

Finally, we conducted a detailed comparative analysis of the high-quality MAGs retrieved from the sheep gut long-read metaHi-C dataset. We employed Mash (Ondov et al., 2016) 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 2.5). Notably, the majority of high-quality MAGs obtained through other Hi-C-based binning tools were also successfully recovered by ImputeCC (Fig. 4a). 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 (Fig. 4b).

3.5. ImputeCC's Genus-Level analysis unveiled key genera and potential species expansion in the sheep gut microbiota

ImputeCC's genus-level analysis, leveraging its retrieval of 408 high-quality MAGs, has unveiled significant insights into microbial composition of the sheep gut microbiota. Within this complex ecosystem, *Bacteroides* emerges as one of the dominant bacterial genera, well-recognized for its potential influence on the intestinal immune system (Routy et al., 2018; Yatsunenko et al., 2012). ImputeCC's distinctive capabilities stood out as it successfully recovered two critical species from the *Bacteroides* genus, specifically *Bacteroides uniformis* and *Bacteroides vulgatus*, within the sheep gut environment. *B. uniformis* has garnered attention for its reported role in ameliorating immunological dysfunctions and metabolic disorders, often associated with intestinal dysbiosis (Gauffin Cano et al., 2012). In contrast, *B. vulgatus* assumes vital roles in reducing the production of gut microbial lipopolysaccharides and inhibiting atherosclerosis (Yoshida et al., 2018). Notably, among high-quality MAGs, while MetaCC managed to detect the presence of *B. vulgatus*, other binning tools failed to identify the genus *Bacteroides* from the sheep gut dataset. ImputeCC's distinctive capability also emerged as it was the only method that could detect the *Tidjanibacter* genus, a relatively new and less-studied taxonomic group (Xie et al., 2020). This discovery creates opportunities for more research on this genus, offering the potential for exploring its ecological roles within the sheep gut environment. Within the *Rikenellaceae* family, ImputeCC's analysis illuminated the prevalence and diversity of the *Alistipes* genus, which was predominantly found in the gastrointestinal tracts of the healthy human microbiome (Parker et al., 2020; Shkoporov et al., 2015). Specifically, ImputeCC retrieved 17 high-quality MAGs affiliated with *Alistipes*, compared to the 4, 3, and 9 high-quality MAGs recovered by MetaTOR, bin3C, and MetaCC, respectively. Among these 17 MAGs, *Alistipes senegalensi* emerged as a noteworthy species, recognized for its involvement in mannose fermentation (Mishra et al., 2012), suggesting a role of the members from the *Alistipes* genus within the sheep gastrointestinal tract's intricate ecosystem. Furthermore, ImputeCC's analysis unveiled five high-quality MAGs within the *Alistipes* genus that could not be annotated at the species level by GTDB-TK,

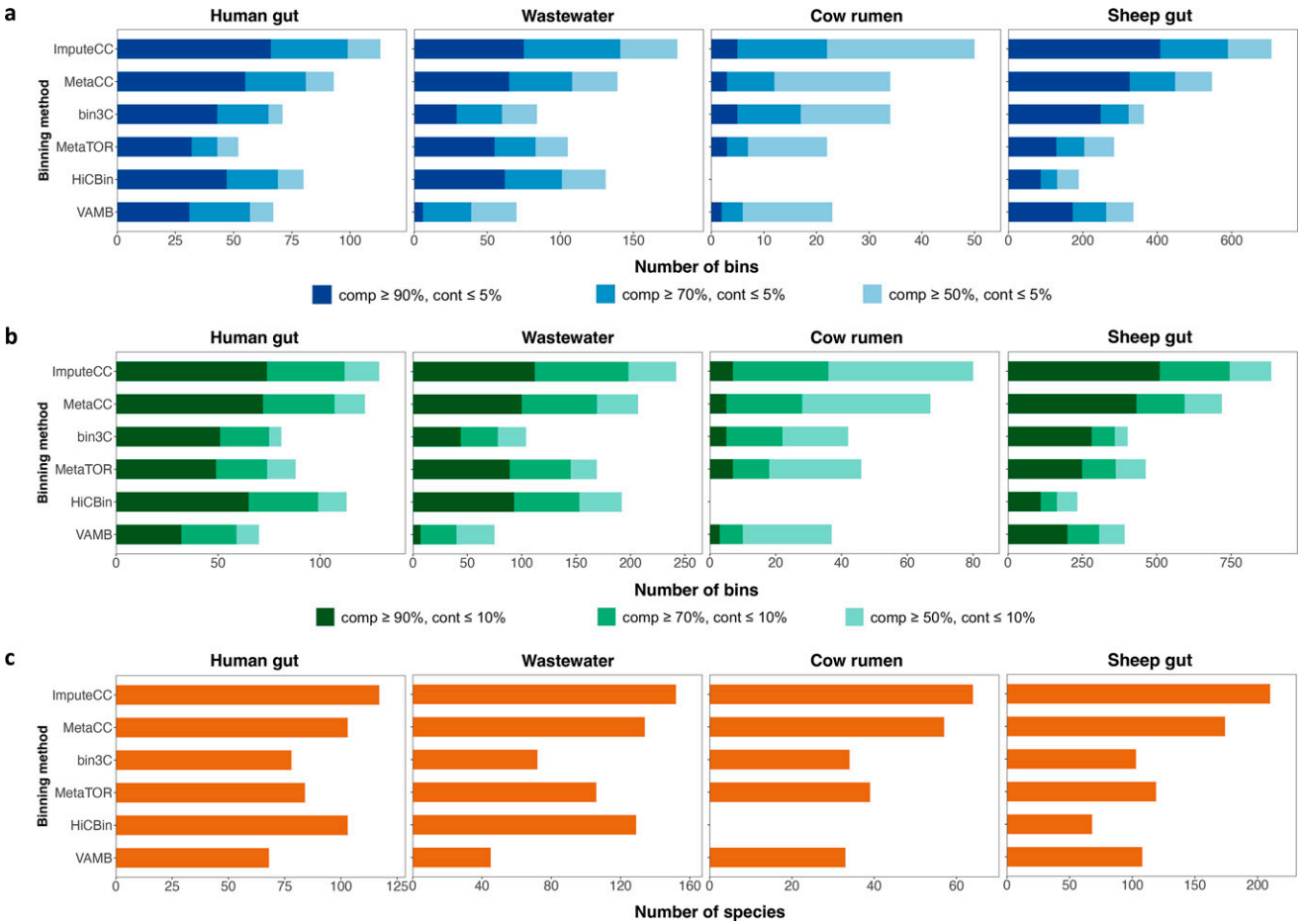

FIG. 3. Benchmarking using the real human gut short-read, wastewater short-read, cow rumen long-read, and sheep gut long-read metaHi-C datasets. **(a)** The number of MAGs with varying completeness (comp) and contamination (cont) ≤5%. ImputeCC consistently outperforms other binning tools, producing a greater number of high-quality bins in all four real metaHi-C datasets. **(b)** The number of MAGs with varying completeness and contamination ≤10%. ImputeCC returned more medium-quality bins when compared to alternative methods for all 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. MAG, metagenome-assembled genomes; metaHi-C, Metagenomic Hi-C.

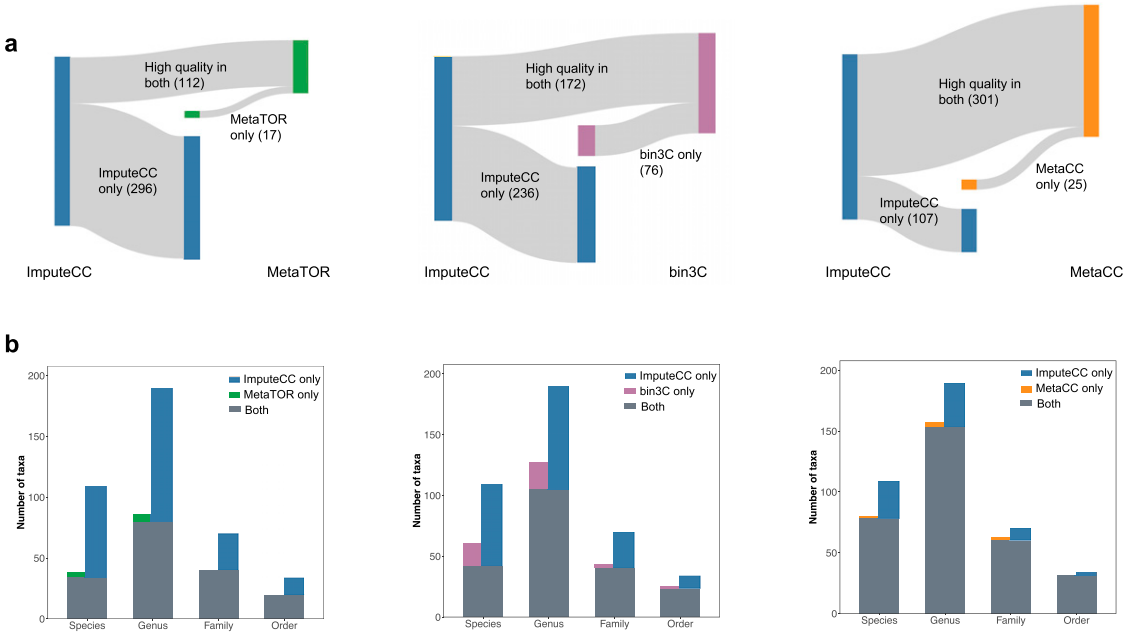

FIG. 4. Comparative analysis of high-quality MAGs retrieved from the sheep gut long-read metaHi-C dataset. **(a)** Comparison of high-quality MAG recovery using ImputeCC and three other Hi-C-based binning tools (MetaTOR, bin3C, and MetaCC), as determined through Mash analysis. ImputeCC successfully retrieved the majority of high-quality MAGs obtained by the alternative Hi-C-based tools, while also surpassing them by reconstructing a significant number of additional high-quality MAGs. **(b)** Annotation analysis of the high-quality MAGs highlighting the enhanced diversity captured by ImputeCC at different taxonomic levels in comparison to its Hi-C-based counterparts, such as MetaTOR, bin3C, and MetaCC. MAG, metagenome-assembled genomes; metaHi-C, Metagenomic Hi-C.

suggesting the potential expansion of species diversity within the *Alistipes* genus. Additional experiments are necessary to gather further data on the phenotypic and physical characteristics of these uncultured members before their definitive identification can be achieved. In conclusion, all these findings underscore the unique efficacy of ImputeCC in advancing our understanding of microbial ecosystems by characterizing the sheep gut microbiota's taxonomic composition and functional potential.

3.6. 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, 204, 25, 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.

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 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 (DeMaere and Darling, 2019; Du and Sun, 2022). 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.

AUTHORS' CONTRIBUTIONS

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 article. Y.D. and W.Z. developed the software. All authors modified and finalized the paper.

AVAILABILITY OF DATA AND MATERIALS

The mock community sequencing data were downloaded from the European Nucleotide Archive under project ID PRJEB52977 (Meslier et al., 2022). The human gut short-read metaHi-C dataset used in this study is available under NCBI BioProject PRJNA413092 (Press et al., 2017). The wastewater short-read metaHi-C dataset is available under NCBI BioProject PRJNA506462 (Stalder et al., 2019). The cow rumen long-read metaHi-C dataset used in this study is available under NCBI BioProject PRJNA507739 (Bickhart et al., 2019). The sheep gut long-read metaHi-C dataset is available under NCBI BioProject PRJNA595610 (Bickhart et al., 2022). The ImputeCC software is freely available at https://github.com/dyxstat/ImputeCC.

CONSENT FOR PUBLICATION

All authors have approved the article for submission.

AUTHOR DISCLOSURE STATEMENT

The authors declare that they have no competing interests.

FUNDING INFORMATION

The research is partially funded by NSF grant EF-2125142.

SUPPLEMENTARY MATERIAL

Supplementary Data S1
Supplementary Data S2
Supplementary Data S3
Supplementary Data S4
Supplementary Data S5
Supplementary Data S6
Supplementary Data S7
Supplementary Data S8
Supplementary Table S1
Supplementary Table S2
Supplementary Table S3

REFERENCES

Albertsen M, Hugenholtz P, Skarshewski A, et al. Genome sequences of rare, uncultured bacteria obtained by differential coverage binning of multiple metagenomes. Nat Biotechnol 2013;31(6):533–538; doi: 10.1038/nbt.2579

Baudry L, Foutel-Rodier T, Thierry A, et al. MetaTOR: A computational pipeline to recover high-quality metagenomic bins from mammalian gut proximity-ligation (meta3C) libraries. Front Genet 2019;10:753; doi: 10.3389/fgene.2019.00753

Beitel CW, Froenicke L, Lang JM, et al. Strain-and plasmid-level deconvolution of a synthetic metagenome by sequencing proximity ligation products. PeerJ 2014;2:e415; doi: 10.7717/peerj.415

Bickhart DM, Kolmogorov M, Tseng E, et al. Generating lineage-resolved, complete metagenome-assembled genomes from complex microbial communities. Nat Biotechnol 2022;40(5):711–719; doi: 10.1038/s41587-021-01130-z

Bickhart DM, Watson M, Koren S, 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 2019;20(1):153; doi: 10.1186/s13059-019-1760-x

Burton JN, Liachko I, Dunham MJ, et al. Species-level deconvolution of metagenome assemblies with Hi-C–based contact probability maps. G3 (Bethesda) 2014;4(7):1339–1346; doi: 10.1534/g3.114.011825

Bushnell B. BBMap: A Fast, Accurate, Splice-Aware Aligner. Technical Report, Lawrence Berkeley National Lab. (LBNL), Berkeley, CA, USA; 2014.

Chaumeil P-A, Mussig AJ, Hugenholtz P, et al. GTDB-Tk v2: Memory friendly classification with the genome taxonomy database. Bioinformatics 2022;38(23):5315–5316; doi: 10.1093/bioinformatics/btac672

Chen Y, Wang Y, Paez-Espino D, et al. Prokaryotic viruses impact functional microorganisms in nutrient removal and carbon cycle in wastewater treatment plants. Nat Commun 2021;12(1):5398; doi: 10.1038/s41467-021-25678-1

Chklovski A, Parks DH, Woodcroft BJ, et al. CheckM2: A rapid, scalable and accurate tool for assessing microbial genome quality using machine learning. Nat Methods 2023;20(8):1203–1212; doi: 10.1038/s41592-023-01940-w

DeMaere MZ, Darling AE. Sim3C: Simulation of Hi-C and Meta3C proximity ligation sequencing technologies. Gigascience 2018;7(2):1–12; doi: 10.1093/gigascience/gix103

DeMaere MZ, Darling AE. bin3C: Exploiting Hi-C sequencing data to accurately resolve metagenome-assembled genomes. Genome Biol 2019;20(1):46; doi: 10.1186/s13059-019-1643-1

Du Y, Fuhrman JA, Sun F. ViralCC retrieves complete viral genomes and virus-host pairs from metagenomic Hi-C data. Nat Commun 2023;14(1):502; doi: 10.1038/s41467-023-35945-y

Du Y, Laperriere SM, Fuhrman J, et al. Normalizing metagenomic Hi-C data and detecting spurious contacts using zero-inflated negative binomial regression. J Comput Biol 2022;29(2):106–120; doi: 10.1089/cmb.2021.0439

Du Y, Sun F. HiCBin: Binning metagenomic contigs and recovering metagenome-assembled genomes using Hi-C contact maps. Genome Biol 2022;23(1):63; doi: 10.1186/s13059-022-02626-w

Du Y, Sun F. MetaCC allows scalable and integrative analyses of both long-read and short-read metagenomic Hi-C data. Nat Commun 2023;14(1):6231; doi: 10.1038/s41467-023-41209-6

Finn RD, Clements J, Eddy SR. HMMER web server: Interactive sequence similarity searching. Nucleic Acids Res 2011;39(Web Server issue):W29–W37; doi: 10.1093/nar/gkr367

Gauffin Cano P, Santacruz A, Moya Á, et al. Bacteroides uniformis CECT 7771 ameliorates metabolic and immunological dysfunction in mice with high-fat-diet induced obesity. PLoS One 2012;7(7):e41079; doi: 10.1371/journal.pone.0041079

Gounot J-S, Chia M, Bertrand D, et al. Genome-centric analysis of short and long read metagenomes reveals uncharacterized microbiome diversity in Southeast Asians. Nat Commun 2022;13(1):6044; doi: 10.1038/s41467-022-33782-z

Handelsman J. Metagenomics: Application of genomics to uncultured microorganisms. Microbiol Mol Biol Rev 2004;68(4):669–685; doi: 10.1128/MMBR.68.4.669-685.2004

Hugenholtz P, Tyson GW. Metagenomics. Nature 2008;455(7212):481–483; doi: 10.1038/455481a

Hugerth LW, Larsson J, Alneberg J, et al. Metagenome-assembled genomes uncover a global brackish microbiome. Genome Biol 2015;16:279; doi: 10.1186/s13059-015-0834-7

Karp RM. An algorithm to solve the m× n assignment problem in expected time O (mn log n). Networks 1980;10(2):143–152; doi: 10.1002/net.3230100205

Kent AG, Vill AC, Shi Q, et al. Widespread transfer of mobile antibiotic resistance genes within individual gut microbiomes revealed through bacterial Hi-C. Nat Commun 2020;11(1):4379; doi: 10.1038/s41467-020-18164-7

Li H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. arXiv 2013; doi: 10.48550/arXiv.1303.3997

Li D, Liu C-M, Luo R, et al. MEGAHIT: An ultra-fast single-node solution for large and complex metagenomics assembly via succinct de Bruijn graph. Bioinformatics 2015;31(10):1674–1676; doi: 10.1093/bioinformatics/btv033

Marbouty M, Cournac A, Flot J-F, et al. Metagenomic chromosome conformation capture (meta3C) unveils the diversity of chromosome organization in microorganisms. Elife 2014;3:e03318; doi: 10.7554/eLife.03318

Marbouty M, Thierry A, Millot GA, et al. MetaHiC phage-bacteria infection network reveals active cycling phages of the healthy human gut. Elife 2021;10:e60608; doi: 10.7554/eLife.60608

Meslier V, Quinquis B, Da Silva K, et al. Benchmarking second and third-generation sequencing platforms for microbial metagenomics. Sci Data 2022;9(1):694; doi: 10.1038/s41597-022-01762-z

Mishra AK, Gimenez G, Lagier J-C, et al. Genome sequence and description of alistipes senegalensis sp. nov. Stand Genomic Sci 2012;6(3):1–16; doi: 10.4056/sigs.2625821

Nissen JN, Johansen J, Allesøe RL, et al. Improved metagenome binning and assembly using deep variational autoencoders. Nat Biotechnol 2021;39(5):555–560; doi: 10.1038/s41587-020-00777-4

Ondov BD, Treangen TJ, Melsted P, et al. Mash: Fast genome and metagenome distance estimation using MinHash. Genome Biol 2016;17(1):132; doi: 10.1186/s13059-016-0997-x

Parker BJ, Wearsch PA, Veloo AC, et al. The genus Alistipes: Gut bacteria with emerging implications to inflammation, cancer, and mental health. Front Immunol 2020;11:906; doi: 10.3389/fimmu.2020.00906

Parks DH, Imelfort M, Skennerton CT, et al. CheckM: Assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. Genome Res 2015;25(7):1043–1055; doi: 10.1101/gr.186072.114

Press MO, Wiser AH, Kronenberg ZN, et al. Hi-C deconvolution of a human gut microbiome yields high-quality draft genomes and reveals plasmid-genome interactions. bioRxiv 2017; doi: 10.1101/198713

Rao SS, Huntley MH, Durand NC, et al. A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping. Cell 2014;159(7):1665–1680; doi: 10.1016/j.cell.2014.11.021

Reichardt J, Bornholdt S. Statistical mechanics of community detection. Phys Rev E Stat Nonlin Soft Matter Phys 2006;74(1 Pt 2):e016110; doi: 10.1103/PhysRevE.74.016110

Rho M, Tang H, Ye Y. FragGeneScan: Predicting genes in short and error-prone reads. Nucleic Acids Res 2010; 38(20):e191; doi: 10.1093/nar/gkq747

Routy B, Gopalakrishnan V, Daillère R, et al. The gut microbiota influences anticancer immunosurveillance and general health. Nat Rev Clin Oncol 2018;15(6):382–396; doi: 10.1038/s41571-018-0006-2

Shkoporov AN, Chaplin AV, Khokhlova EV, et al. Alistipes inops sp. nov. and Coprobacter secundus sp. nov., isolated from human faeces. Int J Syst Evol Microbiol 2015;65(12):4580–4588; doi: 10.1099/ijsem.0.000617

Simon C, Daniel R. Metagenomic analyses: Past and future trends. Appl Environ Microbiol 2011;77(4):1153–1161; doi: 10.1128/AEM.02345-10

Stalder T, Press MO, Sullivan S, et al. Linking the resistome and plasmidome to the microbiome. Isme J 2019;13(10): 2437–2446; doi: 10.1038/s41396-019-0446-4

Streit WR, Schmitz RA. Metagenomics–the key to the uncultured microbes. Curr Opin Microbiol 2004;7(5):492–498; doi: 10.1016/j.mib.2004.08.002

Traag VA, Waltman L, Van Eck NJ. From Louvain to Leiden: Guaranteeing well-connected communities. Sci Rep 2019;9(1):5233; doi: 10.1038/s41598-019-41695-z

Wu Y-W, Tang Y-H, Tringe SG, et al. MaxBin: An automated binning method to recover individual genomes from metagenomes using an expectation-maximization algorithm. Microbiome 2014;2:26; doi: 10.1186/2049-2618-2-26

Wu H, Wang X, Chu M, et al. HCMB: A stable and efficient algorithm for processing the normalization of highly sparse Hi-C contact data. Comput Struct Biotechnol J 2021;19:2637–2645; doi: 10.1016/j.csbj.2021.04.064

Xie Z, Bai Y, Chen G, et al. Modulation of gut homeostasis by exopolysaccharides from Aspergillus cristatus (MK346334), a strain of fungus isolated from Fuzhuan brick tea, contributes to immunomodulatory activity in cyclophosphamide-treated mice. Food Funct 2020;11(12):10397–10412; doi: 10.1039/d0fo02272a

Yaffe E, Relman DA. Tracking microbial evolution in the human gut using Hi-C reveals extensive horizontal gene transfer, persistence and adaptation. Nat Microbiol 2020;5(2):343–353; doi: 10.1038/s41564-019-0625-0

Yatsunenko T, Rey FE, Manary MJ, et al. Human gut microbiome viewed across age and geography. Nature 2012; 486(7402):222–227; doi: 10.1038/nature11053

Yoshida N, Emoto T, Yamashita T, et al. Bacteroides vulgatus and Bacteroides dorei reduce gut microbial lipopolysaccharide production and inhibit atherosclerosis. Circulation 2018;138(22):2486–2498; doi: 10.1161/CIRCULATIONAHA.118.033714

Address correspondence to:
Prof. Fengzhu Sun
Department of Quantitative and Computational Biology
University of Southern California
1050 Childs Way
Los Angeles, CA 90089
USA

E-mail: fsun@usc.edu