Ariadne: Barcoded Linked-Read Deconvolution Using de Bruijn Graphs

Lauren Mak[1,2], Dmitry Meleshko[1,2], David C. Danko[1,2], Waris N. Barakzai[3], Natan Belchikov[4], and Iman Hajirasouliha[2,5,*]

[1] Tri-Institutional Computational Biology & Medicine Program, Weill Cornell Medicine of Cornell University, NY, USA
[2] Institute for Computational Biomedicine, Department of Physiology and Biophysics, Weill Cornell Medicine of Cornell University, NY, USA
[3] Department of Computer Science, New York University, NY, USA
[4] Physiology, Biophysics & Systems Biology Program, Weill Cornell Medicine of Cornell University, NY, USA
[5] Englander Institute for Precision Medicine, The Meyer Cancer Center, Weill Cornell Medicine, NY, USA
* imh2003@med.cornell.edu

Abstract. *De novo* assemblies are critical for capturing the genetic composition of complex samples. Linked-read sequencing techniques such as 10x Genomics' Linked-Reads, UST's TELL-Seq, Loop Genomics' LoopSeq, and BGI's Long Fragment Read combines 3′ barcoding with standard short-read sequencing to expand the range of linkage resolution from hundreds to tens of thousands of base-pairs. The application of linked-read sequencing to genome assembly has demonstrated that barcoding-based technologies balance the tradeoffs between long-range linkage, per-base coverage, and costs. Linked-reads come with their own challenges, chief among them the association of multiple long fragments with the same 3′ barcode. The lack of a unique correspondence between a long fragment and a barcode, in conjunction with low sequencing depth, confounds the assignment of linkage between short-reads.
Results: We introduce Ariadne, a novel linked-read deconvolution algorithm based on assembly graphs, that can be used to extract single-species read-sets from a large linked-read dataset. Ariadne deconvolution of linked-read clouds increases the proportion of read clouds containing only reads from a single fragment by up to 37.5-fold. Using these enhanced read clouds in *de novo* assembly significantly improves assembly contiguity and the size of the largest aligned blocks in comparison to the non-deconvolved read clouds. Integrating barcode deconvolution tools, such as Ariadne, into the post-processing pipeline for linked-read technologies increases the quality of *de novo* assembly for complex populations, such as microbiomes. Ariadne is intuitive, computationally efficient, and scalable to other large-scale linked-read problems, such as human genome phasing.

Availability: The source code is available on GitHub: https://github.com/lauren-mak/Ariadne

Keywords: Linked-read sequencing · Assembly graphs · Metagenomics · Barcode deconvolution

1 Introduction

Next generation sequencing technologies underpin the large-scale genetic analyses of complex genomic mixtures. However, reconstruction of multiple genomes or identification of structural variations is more computationally demanding than assembling a single haploid genome. Though standard Illumina short-read sequencing is the most popular platform for genome-wide characterizations due to its sequencing depth to cost ratio, the length of Illumina short-reads limits the linkage information that can be extracted from a single sequencing insert. To address this, researchers are increasingly using long-read sequencing technologies such as those from Pacific Biosciences or Oxford Nanopore Technologies to approximate single-cell resolution from complex mixtures [1]. However, with significantly higher per-base error rates and higher requirements for input DNA than short-read sequencing, combined with higher costs and lower throughput, long-read approaches are generally impractical for mixtures of genomes or structural variants of variable frequency.

1.1 Overview of Linked-Read Technology

Alternative approaches that rely on short-read sequencing while still generating long-range linkage information include Hi-C and sparse synthetic long-read (SSLR) technologies [21]. While 10x Genomics' linked reads

has been discontinued after nearly half a decade of widespread usage, a variety of other barcode-based SSLR methods such as TELL-Seq, LoopSeq, and BGI's Long Fragment Reads have been commercialized recently with the promise of (near-)single-molecule resolution along with simplified library preparation procedures and compatibility with standard Illumina sequencing machines. For detailed explanations of each method's capabilities and library preparation procedures, we direct the reader to [5,29,28,25]. Though the applications of 10x are well-known, this is the first head-to-head comparison of multiple linked-read technologies to profile their strengths and weaknesses on metagenomics datasets of similar complexity.

We refer to the set of reads that share a $3'$ barcode as a read cloud. The aptly named linked-read technology balances coverage with cost, and is ideally applicable to the problems of metagenomic assembly and structural variant detection [3,24,12].

1.2 Applications of Linked-Reads

Metagenomics Many taxonomic lineages identified in large-scale studies are not represented in reference sequence databases, and are not associated with isolated cultures [11]. However, metagenomic assembly that relies on existing reference genomes will inherently bias the genomic reconstruction of a mixed population, generally towards the most common and already well-characterized species within the sample [1]. To avoid database bias, the accurate assembly of long contiguous blocks of sequence information is critical for accurate and representative taxonomic classification of the species in a microbial population [1,3,8,13]. Several *de novo* assembly tools are available depending on the properties of the dataset. Whereas cloudSPAdes was designed as an all-purpose SSLR assembler, Athena and Supernova were designed for metagenomics and human data respectively [3,24,26].

Human Genome Phasing and Structural Variant Identification The reconstruction of haplotypes (phasing) and detection of structural variation are analogous to the *de novo* assembly of complex populations, but for individual organisms. Previous work combining short- and linked-read technology has demonstrated the feasibility of using linked-read sequencing to achieve medium-length contiguity resolution [18]. Further applications illustrate the capability of linked-reads to resolve human haplotypes and structural variants [29,16,17].

1.3 Challenges Posed by Linked-Read Sequencing

Despite the additional linkage information offered by linked-read sequencing, there are new computational challenges involved in applying barcoded reads to *de novo* assembly. Because metagenomic samples are intrinsically multiplexed samples of multiple species, long-range linkage information is confounded by the multiplicity of fragments assigned to $3'$ barcodes. Existing systems employ on the order of $10^{6\text{-}7}$ $3'$ barcodes [16]. In previous studies with the 10x Genomics system, it was observed that there were 2–20 long fragments of DNA per $3'$ barcode [7]. The larger the barcoded read cloud, the more likely that reads tended to originate from multiple fragments. Our analyses suggest that at least 97% of read clouds are composed of $\geq$2 fragments, with the exception of a LoopSeq dataset. In the absence of additional information about the sample, it is difficult to distinguish the genomic origin of a random assortment of reads with the same barcode.

Furthermore, each fragment of DNA is only fractionally covered by reads (typically 10-20%). Because overall coverage is reduced, linked-reads provide long-range information at the expense of short-range blocks of contiguous sequence. Though it is not possible to complete *de novo* assembly using only reads with the same barcode, this limitation is compatible with phasing large-scale structural variation [9,23].

1.4 The Barcode Deconvolution Problem

The barcode deconvolution problem, previously described in [22,7], is defined as the assignment of each read with a given $3'$ barcode to a subgroup such that every read in the subgroup originates from the same fragment or contiguous genomic segment. A solution to the barcode deconvolution problem for a set of read clouds would be a map from each read cloud to a function which solves the barcode deconvolution problem for that read cloud. Reads with the same $3'$ barcode that are highly likely to have originated from the same

metagenomic fragment are more likely to co-occur in one another's sequence space within the assembly graph than reads from different fragments.

Though the multiplicity of genomic fragments to $3'$ barcodes has been problematic since the inception of linked-read sequencing, the barcode deconvolution problem has only been addressed recently by two computational tools: EMA [22] and Minerva [7]. The EMA approach augments read alignment for barcoded reads based on alignment to a reference sequence. In the process of generating probabilistic alignments, reads from a single read cloud are sub-grouped into sets of reads that map close to each other. EMA is particularly applicable for highly repetitive regions where a barcoded read can align to multiple locations within the genome [22]. However, the EMA approach relies on the user to supply reference sequence(s) as *a priori* information about the input sample. We also demonstrate that EMA (as of the date of publication) is unable to recognize $> 99.9\%$ of linked-read barcodes, even with barcode whitelists tailored for each dataset. While the species composition of popular sampling sites such as the human gut are well-characterized, such reference-based methods are not designed for microbial samples from under-studied environments such as urban landscapes [6]. Minerva does not require a reference genome, using instead k-mer similarities between read clouds to approximately solve the barcode deconvolution problem for metagenomic samples [7]. However, Minerva is memory-inefficient and requires extensive parameter optimization to deconvolve all of the read clouds in a dataset.

In this paper, we present Ariadne as an advanced approach to tackle barcode deconvolution. Instead of using read alignments to reference sequences, which are unknown for poorly characterized environmental microbial samples [6], or relying on computationally expensive string-based graphs, Ariadne leverages the linkage information encoded in the full de Bruijn-based assembly graph generated by a *de novo* assembly tool such as cloudSPAdes [24] to generate up to 37.5-fold more read clouds containing only reads from a single fragment, improve the summed NA50 by up to 500 kbp, and maintain a proportional rate of misassembly relative to *de novo* assembly without prior barcode deconvolution. Searching through the pre-made assembly graph for reads in the neighboring sequence space makes the search for co-occurring reads computationally tractable, generalizable, and scalable to large datasets.

2 Methods

2.1 Algorithm Overview

We have developed Ariadne, an algorithm that approximately solves the barcode deconvolution problem for linked-read sequencing datasets. Ariadne deconvolves read clouds by positioning each read on a de Bruijn-based assembly graph, and grouping reads within a read cloud that are located on nearby edges of the assembly graph. The grouped reads are termed enhanced read clouds. Currently, Ariadne is implemented as a module of cloudSPAdes version 3.12.0.

The Usage of Barcode Information in De Novo Assembly The following is a summary of cloudSPAdes' usage of barcodes to identify contigs from assembly graphs. For a more detailed description, see the Materials and Methods section of [24]. cloudSPAdes constructs and iteratively simplifies a de Bruijn assembly graph using sequencing reads and several sizes of k-mers (by default, 21, 33, and 55 base-pairs). The goal is to recover genomic cycles through the assembly graph that correspond to whole chromosomes. Due to read error, incomplete sequencing coverage, limited sequencing depth, and biological phenomena such as repetitive regions, de novo assemblers employ a number of heuristics to find the optimal unbranching paths through the graph [3,24,26]. In the case of linked-reads and cloudSPAdes, barcodes provide one mechanism of identifying edges that are likely part of the same genomic cycle [24]. If read i carrying barcode b aligns to edge i, then edge i is said to be associated with barcode b. The barcode similarity between two edges is the proportion of associated barcodes in common. By using barcode similarities to identify edges that likely originated from the same genomic region, cloudSPAdes connects edges within the assembly graph to approximate chromosomes.

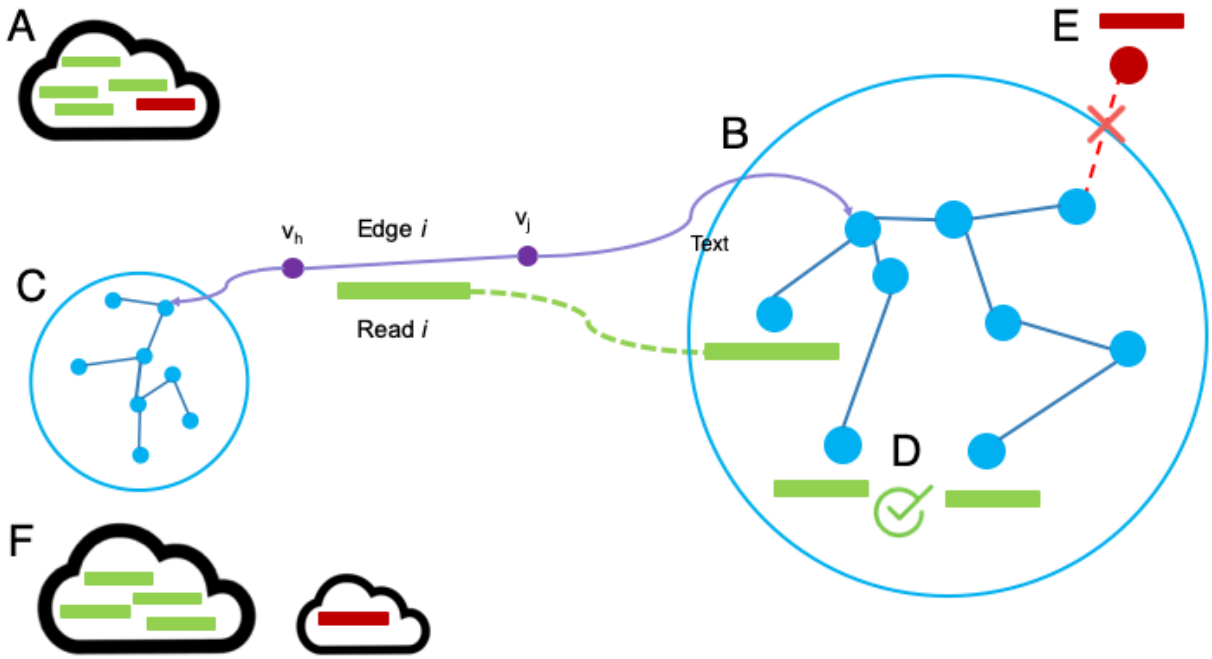

Fig. 1. Graphical description of the Ariadne deconvolution process. **(A)** Reads with the same 3′ barcode are in a read cloud. Green and red reads originate from different fragments. Vertices and edges included in the central read's **(B)** forward and **(C)** reverse Dijkstra are highlighted in blue. These vertices and edges comprise read i's search-distance-limited connected subgraph within the whole assembly graph. **(D)** Reads aligning to edges in this connected subgraph are added to read i's connected set. The paired read of the central read i is indicated with a dashed line. **(E)** Reads originating from different fragments likely coincide with non-included vertices. **(F)** Connected read-sets with at least one intersection (i.e.: one read in common) are output together as an enhanced read cloud.

Due to fragment-to-barcode multiplicity, barcoded metagenomic read clouds are likely comprised of reads from a few species. Since these reads are all carrying the same barcode b, the long edges associated with barcode b may be erroneously connected to form contigs with genomic material from multiple species. Chimeric contigs are hazardous for downstream analyses, since as the binning of contigs into metagenomic-assembled genomes [20,4].

The Assembly Graph Conception of a Fragment Searching through the cloudSPAdes-generated assembly graph for reads in the neighboring sequence space is equivalent to approximating the sequence content of a genomic fragment given the entirety of the sequence content in the input read dataset. By necessity, reads originating from the same fragment must i) have the same 3′ barcode and ii) be no more further apart than the total length of the fragment. The user can specify the search distance according to *a priori* knowledge about the median size of genomic fragments generated in the first step of linked-read sequencing. This process is diagrammed in Figure 1. Reads can be mapped to a de Bruijn graph. A read (Figure 1A green and red bars) is said to coincide with an edge of the assembly graph (Figure 1 purple lines) if the read's sequence aligns to the edge's sequence.The read can also be said to coincide with the vertices bordering the edge it coincides with. For example, in Figure 1, read i coincides with edge i, and vertices v_h and v_j.

The Dijkstra Graph Conception of a Fragment The Dijkstra graphs (Figure 1C and D) comprising the nearby assembly graph for each read represent the potential sequence space that the read-originating fragment occupies on that assembly graph.

A Dijkstra graph contains the shortest paths from source node to all vertices in the given graph. The Dijkstra graphs for a read are comprised of the vertices bordering assembly graph edges that are reachable

within the maximal search distance. The maximum search distance is a user-provided parameter limiting the size of the Dijkstra graph, or the search space to be considered in the assembly graph. The maximum search distance reflects the user's *a priori* knowledge of the size of a fragment.

2.2 The Ariadne Algorithm

Ariadne requires a prior step to locate reads within the (meta)genome of the sample. Where EMA requires an alignment step using the bwa read mapper to generate initial mappings for its barcoded reads [22,15], Ariadne uses the raw assembly graph generated by any of the SPAdes family of *de novo* assemblers to identify the locations of reads relative to one another in the sample's sequence space. Though Ariadne is intended to be a standalone tool, in the future, it will be possible to integrate Ariadne as an intermediary step in the *de novo* assembly procedure to improve the assembly graph *in situ*.

The assembly graph is first generated by applying cloudSPAdes to the raw linked-read metagenomics reads as described in [19]. For a more detailed explanation of the process of forming the assembly graph, we refer the reader to [2,19]. Subsequently, the Ariadne deconvolution algorithm is applied to the raw reads, with the cloudSPAdes-generated assembly graph supplying potential long-range linkage connections derived from the dataset.

Step 1: Extract the assembly graph from a cloudSPAdes run. A simple representation of an assembly graph is depicted in (Figure 1B). In the cloudSPAdes assembly procedure, the assembly graph is obtained from the raw assembly graph by condensing each non-branching path into a single edge and closing gaps, removing loops, bulges, and redundant contigs. In the process of constructing the assembly graph, cloudSPAdes also maps each read to the assembly graph, generating the mapping path of a read, or the edges in the assembly graph either partially or fully spanned by the read (Figure 1 read/edge i, vertices h and j). The mapping path P_i, or finite walk, of read i is comprised of the set of edges e from graph G that read i covers (Equation 1).

$$P_i = \{e_1, e_2, \ldots, e_n\} \tag{1}$$

The mapping path can equivalently be described as a vertex sequence V_i, which is composed of the set of vertices that border each of the edges in P_i (Equation 2).

$$V_i = \{v_1, v_2, \ldots, v_{n+1} : \phi(e_i) = v_i, v_{i+1}\} \tag{2}$$

Importantly, the vertices and edges in the assembly graph are unique numeric indices replacing their sequence content. Instead of having to compare the sequences comprising the assembly graph edges, reads sharing edges can be identified by the indices in their mapping paths. As such, for the purposes of barcode deconvolution, reads do not need to be re-mapped along the assembly graph. Instead, the index-based mapping paths and vertex sequences of each read are used to locate the read. This represents a considerable speed-up over the Minerva procedure, which relies on hashing string-based sequence comparisons between reads and read clouds. Steps 2 and 3 are trivially parallelized such that the deconvolution procedure processes as many read clouds as there are threads available.

Step 2: Generating connected read-sets for each read *i*. If the number of reads in the read cloud is greater than the user-set size cutoff, the read cloud (i.e.: tagged with the same $3'$ barcode) is loaded into memory (Figure 1A). For each read i, the following steps are conducted to identify other reads with the same $3'$ barcode that potentially originated from the same fragment.

Step 2A: Generating forward and reverse Dijkstra graphs. In Figure 1, read i aligns to edge i, which is bordered by vertices v_h and v_j. Assuming that read i is oriented in the $5' \rightarrow 3'$ direction, vertex v_j is represents the $3'$-most sequence that read i is contiguous with. By finding edges reachable within search distance d_max of v_h and v_j, we will be able to all reads with the same barcode that are likely to originate from the same fragment. Reads that are oriented in the $3' \rightarrow 5'$ direction can be reverse-complemented to apply this same procedure.

To facilitate this search, a forward Dijkstra graph (Figure 1C) is constructed starting at the end-vertex of the edge that read i coincides with, v_j. The forward Dijkstra graph $D_{f,i}$ from v_j comprises the set of

vertices v_k, which are all vertices in the assembly graph reachable within the maximum search distance d_{max} from v_j (Equation 3).

$$D_{f,i} = \{v_k : min(distance(v_j, v_k)) \leq d_{max}\} \tag{3}$$

The process of constructing the Dijkstra graph, which represents the nearby sequence space of read i, is as follows. The tentative distances to all downstream vertices in the assembly graph are set to the maximum search distance d_{max}. From vertex v_j, the tentative distances to these vertices are calculated from lengths of the edges connecting the vertices. This value is compared to the current distance value, and the smaller of the two is assigned as the actual distance. The vertex with the smallest actual distance from the current node is selected as the next node from which to find minimal paths, and this process is repeated. This process concludes for any path from v_j when the smallest tentative distance to vertices v_k is the maximum search distance.

Correspondingly, a reverse Dijkstra graph is constructed with the goal node set as v_h, the start-vertex of the edge that the start of the read coincides with (Figure 1D). For the reverse Dijkstra $D_{r,i}$, v_h is treated as the terminal node of Dijkstra graph construction, and the set of distances in $D_{r,i}$ is comprised of the vertices v_g such that the minimal edge-length distance between v_h and v_g is less than the maximum search distance.

Step 2B: Identifying other reads potentially originating from the same (meta)genomic fragment. Due to long-range linkage, reads from the same genomic fragment should occur in both the nearby assembly graph and each others' Dijkstra graphs. As such, all other reads j in the same read cloud are compared to read i and its forward and reverse Dijkstra graphs. For every read j in the read cloud, if one of the following conditions holds, the read j is added to the connected read-set of read i. If any vertex in the vertex sequence of j, V_j (or its reverse complement sequence) is found in the forward or reverse Dijkstra graphs (Figure 1E). This process is conducted via binary search of sorted vertex names in the forward and reverse Dijkstra graphs, eliminating the need to test for sequence string equivalency as in Minerva. Note that edge lengths can range from hundreds to tens of thousands of base-pairs in length. If read i aligned to an edge that was 10 kilo-base-pairs long, many reads from the same originating fragment may in fact align to the same edge as read i. Ariadne thus deconvolves read clouds by dynamically identifying the maximal sets of connected reads within each cloud.

Else, if neither of the above is true, the read is likely to have originated from a different fragment than read i (Figure 1F) despite being tagged with the same $3'$ barcode, and thus no connecting component is built.

Step 3: Output maximal sets of connected reads, or enhanced read clouds. Each enhanced read cloud is the set of reads from the same read cloud that are part of the same search-distance-limited connected subgraph, as identified in step 2B. These enhanced read clouds are output as solutions to the barcode deconvolution problem (Figure 1G). Each original read cloud is either subdivided into two or more enhanced read clouds, or kept intact if all of the reads were found within the search distance d_{max} of each other. To avoid a preponderance of trivial-sized enhanced read clouds, if there is an enhanced read cloud generated that is composed solely of a pair of reads, the two reads are instead added to the smallest enhanced read cloud generated from the same original cloud.

2.3 Mathematical Justification of Fragment Dissimilarity

We provide a short summary of the mathematical model that justifies the modelling of fragments as search-distance limited connected subgraphs within a de Bruijn assembly graph. The details of the mathematical model for drawing fragments of DNA from a linked-read sequencing sample are fully described in [7], the article describing Minerva, the previous iteration of Ariadne. In brief, random fragments of genomic sequence on the scale of kilo-base-pairs from a large number of species with genomes that are many mega-base-pairs in length are unlikely to be similar in terms of sequence to overlap on the assembly graph, even with k-mers as small as 22 bp. For metagenomic datasets, at least 95% of read clouds consist of fragments that do not overlap in this model. This is important because it means that overlapping fragments, and thus sprurious connections, are uncommon and will not hinder deconvolution in most read clouds. The algorithm is theoretically capable of uniquely deconvolving 95% of read clouds by searching the assembly graph surrounding the read for

connecting reads. The only drawback of this model is that it does not account for the fact that individual fragments may have similar sequences, such as those resulting from repetitive regions or highly conserved genes.

3 Results

3.1 Benchmarking Datasets

We used four datasets, one of which was used in [7], and another in [24] to explore the effects of barcode deconvolution on read cloud composition and assembly quality. Two of the datasets, MOCK5 LoopSeq MOCK20 TELL-Seq, are being analysed for the first time in this work. We refer readers to Supp. Table 1 for a detailed description of the synthetic microbial communities and the technical preparation of the MOCK5 10x sequencing dataset. The MOCK5 10x dataset consists of 97,491,080 reads, 91,101,472 of which are barcoded (93.4%). The species comprising the community are *Escherichia coli, Enterobacter cloacae, Micrococcus luteus, Pseudomonas fluorescens, and Staphylococcus epidermidis*. The MOCK5 LoopSeq dataset consists of 75,107,814 reads, all of which are barcoded. The species comprising the community are *Escherichia coli, Porphyromonas gingivalis, Pseudomonas aeruginosa, Rhodobacter sphaeroides, and Streptococcus mutans*. While both comprised of 5 species, the MOCK5 10x and MOCK5 LoopSeq datasets share only one species—*Escherichia coli*—and thus cannot be considered as a comparison of 10x and LoopSeq technology. The original MOCK20 10x dataset is 332,575,310 reads large, but for efficiency a random subset of 100 million reads was used in the subsequent analyses. Of those 100 million reads, 94,151,528 were barcoded (94.2%). The species comprising the community are *Streptococcus mutans, Porphyromonas gingivalis, Staphylococcus epidermidis, Escherichia coli, Rhodobacter sphaeroides, Bacillus cereus, Pseudomonas aeruginosa*, Streptococcus agalactiae, Clostridium beijerinckii, Staphylococcus aureus, Acinetobacter baumannii, Neisseria meningitidis, Propionibacterium acnes, Helicobacter pylori, Lactobacillus gasseri, Bacteroides vulgatus, Deinococcus radiodurans, Actinomyces odontolyticus, Bifidobacterium adolescentis, and Enterococcus faecalis*. Similarly, the original MOCK20 TELL-Seq dataset is 210,463,786 reads large, but subsetted to 100 million reads. All of the reads were barcoded, and the community composition was the same. See Supp. Table 1 for a tabularized summary and download information.

3.2 Gold-Standard ('Reference') Cloud Deconvolution

The use of mock communities allowed us to infer the genomic fragment of origin of the barcoded reads. These gold-standard fragment assignments served as a benchmark to compare read clouds and assemblies with and without deconvolution. We mapped the reads to the reference sequences of the species that were known to form the mock communities using Bowtie2 [14] (version version 2.3.4.1). Using the read mapping positions along the reference genome, we further subsetted the reads into gold-standard read clouds such that the left- and right-most starting positions of reads in the same cloud were no further than 200 kbp apart, in case multiple fragments from the same genome were present in the same read cloud. We termed this method 'reference deconvolution' as it represents the maximum likelihood inference of the genomic fragments that originated reads tagged with the same $3'$ barcode.

3.3 Selection of Search Distances

Prior to conducting *de novo* metagenomics assembly, [24] estimated several dataset parameters to accurately model metagenomic fragments as paths through the assembly graph. To estimate the average fragment length, a method termed single linkage clustering was used to partition reads into clusters corresponding to alignments of likely fragments to the known reference genomes of the species comprising the sample. For example, the mean fragment size of the MOCK5 10x dataset was estimated to be 39,139 base-pairs. This estimated fragment length is analogous to the maximum search distance. Though unlikely, if the fragment size distribution is multimodal, the maximum search distance will model long-range linkage poorly at the fragment level, and by extension reduce deconvolution quality. For a complete analysis of the estimated fragment length, fragments per $3'$ barcode, and overall coverage of MOCK5 10x, see the Supplementary Materials of [24] that also used the same data. Several search distances smaller than the estimated fragment

length—5, 10, 15, and 20 kilo-base-pairs (kbp)—were selected for this study. The Results section focuses on Ariadne performance with a search distance d_{max} =5 kbp, which provides the best balance of deconvolution accuracy, assembly quality, and computational efficiency.

3.4 Ariadne Generates a Large Number of High-Quality Enhanced Read Clouds

Ariadne generated enhanced read clouds, or subgroups of original read clouds, that largely corresponded to individual fragments of DNA. We measured the quality of each enhanced read cloud using two metrics: Shannon entropy index $H = \sum p_i \log p_i$ and purity $P = \max(\vec{p})$ where p_i indicates the proportion of reads in an (enhanced) read cloud that originates from the same most prevalent 200 kbp region in a single reference sequence. Prior to these quality checks, we excluded reads from standard and enhanced read groups of size 2 or smaller (i.e.: consisting of a single read-pair or smaller), which are trivially pure.

Table 1. Read cloud summary statistics. For Ariadne deconvolution, we used a search distance of 5 kbp. The 3-5th columns contain the average and standard deviations of read cloud statistics. Prop. Under, Comp., and Over refer to the proportion of total original or deconvolved read clouds that were over- or under-deconvolved, or completely and exactly comprised of all of the reads from a single inferred genomic fragment.

Dataset	**Deconv. Method**	**Avg. Purity**	**Avg. Entropy**	**Avg. Size**	**Num. Clouds**	**Prop. Under**	**Prop. Comp.**	**Prop. Over**
MOCK5 10x	None	0.53±0.16	1.61±0.52	63.51 ±51.64	1 425 430	0.98	0.02	0
MOCK5 10x	Reference	0.99±0.05	0.03±0.14	18.47 ±19.26	4 733 136	0.06	0.91	0.03
MOCK5 10x	Ariadne	0.89±0.22	0.37±0.7	12.18 ±15.26	4 753 473	0.25	0.04	0.71
MOCK5 LoopSeq	None	0.93±0.11	0.31±0.31	269.18 ±356.67	276 665	0.68	0.32	0
MOCK5 LoopSeq	Reference	0.98±0.07	0.06±0.21	116.19 ±252.59	637 012	0.14	0.82	0.04
MOCK5 LoopSeq	Ariadne	0.88±0.21	0.4±0.7	101.44 ±225.71	734 143	0.35	0.14	0.51
MOCK20 10x	None	0.38±0.15	2.28±0.62	186.49 ±129.32	503 205	0.97	0.03	0
MOCK20 10x	Reference	0.99±0.06	0.05±0.18	29.31 ±29.26	3 146 784	0.13	0.83	0.04
MOCK20 10x	Ariadne	0.94±0.18	0.22±0.64	13.56 ±22.69	6 920 196	0.13	0.05	0.83
MOCK20 TELL-Seq	None	0.39±0.15	2.37±0.59	160.45 ±110.51	623 026	0.98	0.02	0
MOCK20 TELL-Seq	Reference	1±0.02	0±0.05	23.83 ±27.65	4 067 621	0.01	0.99	0
MOCK20 TELL-Seq	Ariadne	0.87±0.24	0.47±0.89	21.96 ±30.74	4 074 321	0.25	0.04	0.71

Without deconvolution, nearly 100% of the reads are in mixed-origin read clouds, or clouds that are comprised of reads that have likely originated from different linked-read fragments (Table 1). Since the goal of linked reads is to approximate the linkage range of long fragments, having mixed-origin clouds as a result of fragment-to-barcode multiplicity confounds downstream applications such as taxonomic classification. Ariadne deconvolution reduces the proportion of multi-origin read clouds by at least 2-fold in the MOCK5 LoopSeq dataset, up to 7.5-fold in the MOCK20 10x dataset (Table 1 column 7). Furthermore, Ariadne at least doubles the proportion of completely deconvolved read clouds, which represent the entirety of the reads from a single inferred genomic fragment, except in the case of the MOCK5 LoopSeq dataset (Table 1 column

8). In comparison, under-deconvolved clouds are clouds comprised of reads that originate from multiple fragments, whereas over-deconvolved clouds are comprised of single-origin reads that are a subset of all of the reads from that fragment. The proportion of single-origin reads, the sum of complete and over-deconvolved read clouds, has increased between 2- and 37.5-fold. Furthermore, read cloud entropy has decreased between 5- to 10-fold, except in the case of MOCK5 LoopSeq. In decreasing the read cloud size and increasing read cloud purity, Ariadne deconvolution has also reduced the number of fragments originating the reads in a cloud (Table 1). If any one cloud is sampled at random, enhanced read clouds will on average be comprised of 1/5 of the species in the original cloud.

Without deconvolution, there is a large spread of P at each read cloud size. Larger clouds are more likely to contain reads of multiple species origins and thus lower purity. The exception to this trend is the MOCK5 LoopSeq dataset (Figure 2 top row). Since Ariadne deconvolution generally decreases this trend (except in the case of LoopSeq), it is unlikely that our results have been inflated by a large number of small and trivially pure clouds. Furthermore, reference-based deconvolution generates a similar number of clouds of a similar size to Ariadne, indicating that search-distance based subgrouping models linked-read fragment boundaries sufficiently.

While the 10x and MOCK20 TELL-Seq size-to-purity distributions are similar, the MOCK5 LoopSeq distribution peaks at $P = 0.96$, which explains why Ariadne deconvolution has minimal effect on read cloud quality. For the rest of the datasets, with respect to size, the relative purity of Ariadne-enhanced read clouds is significantly larger than that of the original read clouds (Figure 2 bottom row). The ideal deconvolution based on reference-mapped positions is shown in the Figure 2 middle row, where there are extremely few clouds with $P < 1$. Even with deconvolution, there are still large read clouds (≥ 100 bp) in the deconvolved set, indicating that Ariadne is capable of maintaining the integrity of existing single-origin read clouds through a limited search of the assembly graph (Figure 2). Overall, applying Ariadne deconvolution to linked-read datasets generates enhanced read clouds that closely resemble the size-to-purity distribution of the reference deconvolution, thereby approximating the ideal deconvolution without *a priori* information about the microbial composition of the originating data.

We have additionally demonstrated the effect of increased search distance on overall dataset quality metrics (Supp. Table 2). The number of read clouds and the average entropy increase as the search distance increases, with minimal increases in the read cloud purity along with the proportion of single-origin (complete and over-deconvolved read clouds). Given with the large increases in computational runtime, only results with the deconvolution search distance of 5 kbp are included going forward in the main text. Similar results- smaller and on average fewer-origin read clouds- can be observed with the Shannon entropy measure (Supp. Fig. 1).

3.5 Enhanced Read Clouds Improve Metagenomic Assembly

Since Ariadne was able to deconvolve the original linked-read dataset into high-quality and non-trivial enhanced read clouds, we applied Ariadne to the full 97M-read MOCK5 10x dataset, the full 75M-read MOCK5 LoopSeq dataset, and 100M randomly sampled reads from each of the MOCK20 10x and MOCK20 TELL-Seq datasets to generate enhanced read clouds as input to cloudSPAdes [24] in metagenomics mode. While both comprised of 5 species, the only species common to the MOCK5 10x and MOCK5 LoopSeq datasets is *Escherichia coli*. As such, while they are comparable in terms of input community complexity, they cannot be treated as a comparison of linked-read technology with respect to de novo assembly. The assembly quality of the resulting scaffolds with reference and Ariadne deconvolution was compared to that without prior deconvolution. cloudSPAdes generates *de novo* assemblies a wide variety of sequencing data into contigs and scaffolds, and has been benchmarked on the MOCK5 10x and MOCK20 10x datasets previously [24]. As such, we have used similar metaQUAST metrics to evaluate and compare the quality of the assemblies.

The largest improvements are in the overall assembly contiguity and the largest alignment, demonstrating that enhanced read clouds with increased fragment specificity generate higher-quality assemblies (Figure 3). NA50 is another measure of assembly contiguity, reporting the length of the aligned block such that using longer or equal-length contigs produces half of the bases of the assembly. *De novo* assemblies generated from reference- or Ariadne-deconvolved read clouds are significantly more contiguous than those generated from the original, non-deconvolved read clouds (Figure 3). To calculate the overall performance improvements

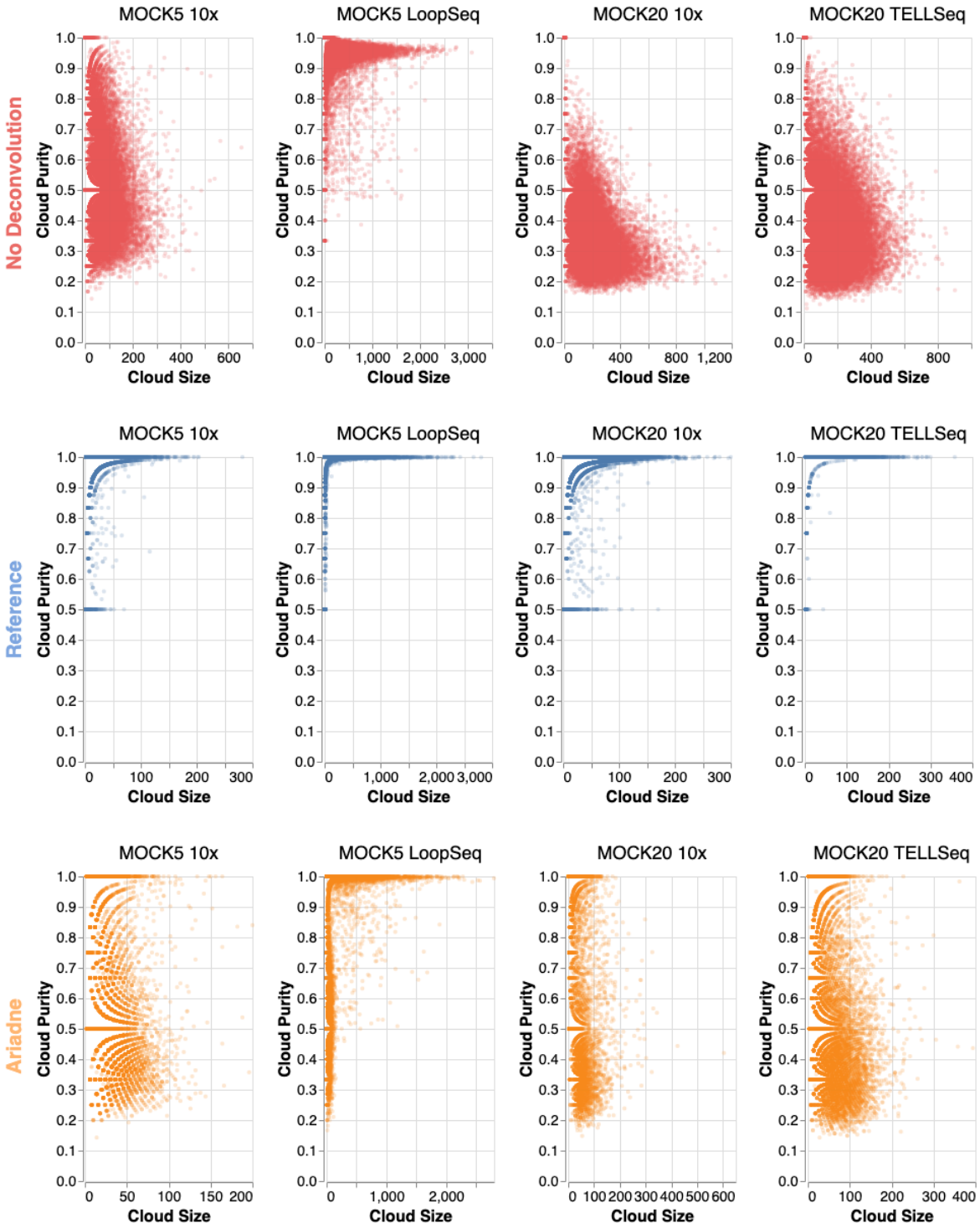

Fig. 2. The size-weighted purity of linked-read read clouds increases after applying deconvolution methods. All graphs were generated from 40,000 randomly sampled clouds from the dataset. **Top row**: No deconvolution. **Middle row**: Reference deconvolution based on read alignment to species and then grouping reads in 200 kbp regions. **Bottom row**: Ariadne deconvolution with a search distance of 5 kbp and a minimum cloud size cutoff of 6.

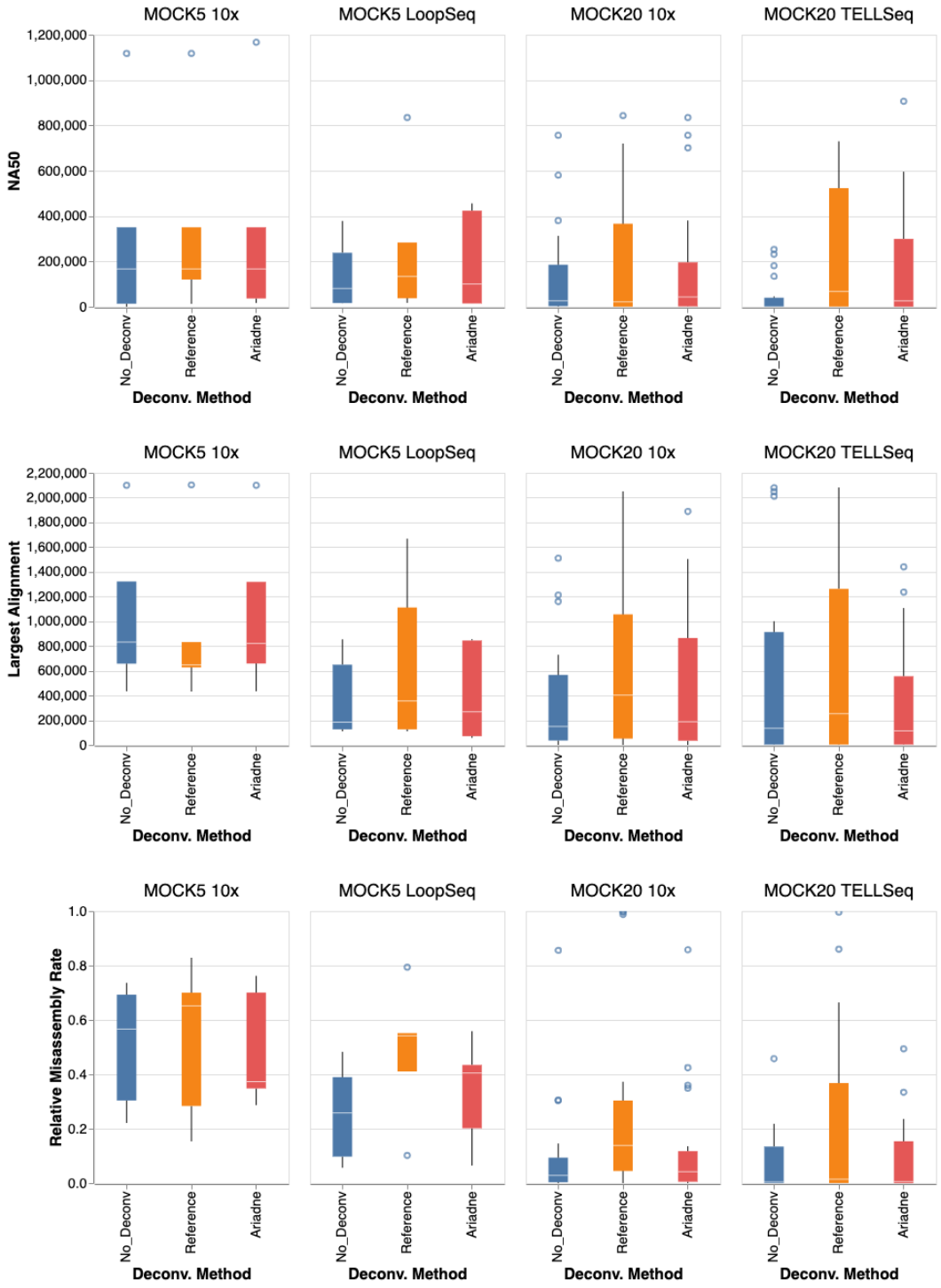

Fig. 3. Read cloud deconvolution improves metagenomic assembly compared to raw linked-read data. We compared assemblies built from raw linked reads (no deconvolution) to assemblies built from reads deconvolved using two methods: reference deconvolution, which maps reads to reference genomes, and deconvolution using Ariadne. **Top Row**: The NA50 of assemblies for each species in each sample between deconvolved reads and raw barcoded reads. Larger numbers indicate better performance. **Middle Row**: The largest alignment of assemblies for each species in each sample between deconvolved reads and raw barcoded reads. Larger numbers indicate better performance. **Bottom row**: The proportion of misassembled bases p_{miss} is the number of bases in misassembled contigs over the total number of assembled bases.

when assembling each dataset, the assembly statistic for each species was summed and the relative difference between reference- or Ariadne-enhanced scaffolds and no-deconvolution scaffolds calculated. For NA50 and largest alignment, this means the no-deconvolution summed NA50 were subtracted from that of the reference- or Ariadne-enhanced scaffolds. For the rate of misassembled bases, the number of misassembled bases was divided by the total assembly length. Then, the reference- or Ariadne-enhanced scaled misassembly rates were divided by that of the no-deconvolution scaffolds.

While the differences between the deconvolution methods and no deconvolution were minimal in the MOCK5 10x dataset (Figure 3 first column), there were significantly positive differences in NA50 and largest alignment in all other datasets. While Ariadne generally produces shorter NA50 and largest alignments than the ideal genomic deconvolution, Ariadne assemblies significantly outperform no-deconvolution assemblies in terms of contiguity statistics (Figure 3 top and middle). However, in the case of the two MOCK5 datasets, the Ariadne assembly NA50 gains were larger than the reference deconvolution assembly gains, suggesting that the most ideal read cloud partition does not always generate the most contiguous assemblies (Figure 3 top first and second panels).

In terms of misassembly rate, Ariadne assemblies largely match no-deconvolution assemblies and significantly outperform reference deconvolution. With the MOCK20 10x dataset, the third quartile of misassembly rate by reference deconvolution is 8-fold larger than that of no deconvolution, whereas Ariadne assemblies contained up to 2-fold more at maximum (Figure 3 bottom third panel). Ariadne underperforms assembly without deconvolution in terms of largest alignment and misassembly rate in MOCK20 TELL-Seq (Figure 3 bottom last panel). However, both Ariadne- and reference-enhanced assembly obtain NA50 for 4 species that the no-deconvolution assembly failed to capture, as well as some substantial differences between Ariadne and reference-based deconvolution with respect to no deconvolution (Table 2). In other datasets, there may be species that are more easily reconstructed with read cloud deconvolution that cloudSPAdes would not otherwise find long, contiguous reference subpaths for through the assembly graph. Outlier values of NA50, largest alignment, and misassembly rates were omitted from Figure 3 for visual clarity were all from the reference deconvolution scaffolds and can be found in Supp. Table 4. There were no major changes in the fraction of reference bases that were reconstructed, the total aligned length, the mean number of mismatches, and the number of contigs.

Table 2. Both Ariadne and reference deconvolution increase the summed NA50 of *de novo* assembled metagenomes. The rightmost column shows the difference between the summed NA50 of assemblies obtained from deconvolved reads and the summed NA50 of non-deconvolved assembly. The asterisk (*) indicates species that had no associated NA50 in the non-deconvolved assembly.

Dataset	**Species**	**Deconv. Method**	**Difference in NA50 (bp)**
MOCK5 10x	Enterobacter cloacae*	Ariadne	16 655
MOCK20 TELL-Seq	Streptococcus agalactiae*	Ariadne	464 957
MOCK20 TELL-Seq	Staphylococcus epidermidis*	Ariadne	489 177
MOCK20 TELL-Seq	Streptococcus mutans*	Ariadne	906 518
MOCK20 TELL-Seq	Staphylococcus aureus*	Ariadne	106 598
MOCK5 10x	Enterobacter cloacae*	Reference	13 628
MOCK20 10x	Rhodobacter sphaeroides	Reference	-580 531
MOCK20 10x	Staphylococcus epidermidis	Reference	639 974
MOCK20 TELL-Seq	Staphylococcus epidermidis*	Reference	1 503 747
MOCK20 TELL-Seq	Streptococcus agalactiae*	Reference	2 047 925
MOCK20 TELL-Seq	Streptococcus mutans*	Reference	2 012 488
MOCK20 TELL-Seq	Staphylococcus aureus*	Reference	134 663

3.6 Runtime and Performance

The time complexity of the read cloud deconvolution process, which consists of Dijkstra graph construction and binary searches through that graph, is to $O(E \log V)$, where E = number of edges and V = number of

vertices in the forward or reverse Dijkstra graphs. The number of vertices searched for connectivity (step 2b) depends on technical properties intrinsic to the dataset—such as the number of reads, the read coverage, error rate, number of repeats in the metagenome, and number of chromosomes—and the user-set search distance parameter. In all cases, Ariadne consumes approximately as much memory as the initial assembly graph generation process, with the exceptions of the MOCK5 datasets (Supp. Table 3). When the search distance is 5 kbp or less for the 10x datasets, Ariadne's runtime is approximately similar to that of the entire de novo assembly process (Supp. Table 1). At search distances larger than 5 kbp as well as for non-10x datasets, Ariadne's runtime becomes significantly larger than the time needed to generate the assembly graph, likely due to the increased number of nodes considered per deconvolution process and the increased graph connectivity (Supp. Table 3). For instance, when comparing the two MOCK20 datasets, Ariadne deconvolution with a search distance of 5 kbp takes nearly 5 times as long to complete for MOCK20 TELL-Seq, despite the number of reads being identical and *de novo* assembly taking nearly the same amount of time.

3.7 EMA, Longranger and Lariat, and Minerva

EMA takes linked-read barcodes into account to align linked reads to a reference sequence(s) using a latent variable model, thereby serving as an alternative to Bowtie2-based read alignment to generate gold-standard read cloud deconvolution. However, the first step of the EMA pipeline (last downloaded: May 7, 2021), `ema count`, which counts barcodes and partitions the original FastQ file into a number of deconvolution bins, is unable to recognize most to all of the barcodes, even when custom whitelists with the exact list of barcodes in the dataset are directly provided (Supp. Table 5). In one case, `ema count` failed to detect any barcodes in the MOCK5 LoopSeq dataset altogether. In the best-case scenario with the MOCK20 10x dataset, `ema count` recognized 199,488 barcoded reads. The core deconvolution module in the pipeline, `ema align`, would have been able to deconvolve at maximum 0.002% of a 94-million read dataset (Supp. Table 5). To deconvolve reads that are not recognized as barcoded, the EMA pipeline uses the same procedure as our reference deconvolution pipeline, with bwa as the read aligner instead of Bowtie2. Due to the paucity of aligned reads and the subsequent lack of deconvolution, EMA-enhanced read clouds were not featured in this analysis. For ease of use, we decided to use our own reference-based deconvolution pipeline, which automates all of the read alignment, read-subgrouping, and FastQ generation steps with a single submission command. Similar barcode recognition issues were encountered with the Longranger align pipeline and the Lariat aligner it was based on. Similar to EMA, Lariat incorporates barcode information to align linked reads to a reference sequence(s). However, the available FastQ files for all four of the datasets were not comprised of raw Illumina BCL files or the raw output of `longranger mkfastq`. As such, it was not possible to apply Longranger align or Lariat to the four datasets. Minerva was similarly unable to deconvolve a sufficient number of read clouds (0.002% at best with the MOCK5 LoopSeq dataset), and it too was not featured in this full analysis. However, both Ariadne and Minerva completed a deconvolution test-run of 20-million reads from the MOCK5 10x dataset. In summary, while Minerva generated nearly 100% single-species read clouds, it was only able to deconvolve 4% of the reads in total (Supp. Figs. 2 and 3, Table 3). Performance-wise, the Minerva runs were 4 times as long and consumed 3 times as much memory. The results are explained in greater detail in the Supplementary Materials (pg. 6 and 7).

4 Discussion

We have developed Ariadne, a novel linked-read deconvolution algorithm based on assembly graphs that addresses the barcode deconvolution problem. Ariadne represents a complete usage of the linkage information in linked-read technology. Ariadne deconvolution generates enhanced read clouds that are up to 37.5-fold more likely to be single-origin, which will improve downstream applications that depend on approximating the long-range linkage information from a single species, such as taxonomic classification. *De novo* assemblies of mock metagenome communities are significantly more contiguous with Ariadne-enhanced read clouds, without the outsized increase in misassembly rate as observed with the ideal deconvolution strategy. In terms of the dataset-specific results, increasing the number of species did not change the shape of the size-to-purity distribution greatly, reflecting the inter-technology similarities in barcode-to-fragment multiplicity.

As such, Ariadne is capable of improving assembly results and read cloud composition across all of the linked-read strategies, and unlike EMA or the Longranger pipeline, easily facilitates re-analyses of existing linked-read datasets for higher-resolution *de novo* assembly and taxonomic classification without pre-existing knowledge of the originating species. Ariadne deconvolution has the largest impact when the input microbial community is large and complex, which is ideal for environmental microbial samples with minimal prior characterization [6].

There have been other recent developments in the linked-read space, intended to extract as much linkage information from barcoded reads as possible in order to maximize recovery of input genomic information. Instead of tackling the barcode deconvolution problem, Guo et al. [10] and Weng et al. [27] innovate downstream of the *de novo* assembly problem. By using linked-reads in combination with graph- or k-mer-based methods, SLRsuperscaffolder and IterCluster attempt to generate longer and higher-quality assemblies. Either of these, paired with the largely single-origin read clouds generated by Ariadne, could potentially improve the NA50 and average alignment size generated by cloudSPAdes.

Though Ariadne relies on cloudSPAdes parameters to generate the assembly graph (e.g., iterative k-mer sizes), the program by itself only has two: search distance and size cutoff. The maximum search distance determines the maximum path length of the Dijkstra graphs surrounding the focal read. Since each read is modelled as the center of a genomic fragment, the search distance can be thought of as the width of the fragment. As such, it should be set as the mean estimated fragment length, as determined using other means such as *a priori* knowledge of shearing duration and intensity or mapping same-barcoded reads to known reference genomes. While the estimated mean length of a metagenomic fragment according to [24] is approximately 40 kbp, the most balanced results were obtained with a significantly shorter search distance. Of the distances tried in this analysis, to balance the highest-quality assembly results, single-origin read clouds, and computational efficiency, we recommend that the user set a search distance 5 kbp or shorter to generate their own enhanced read clouds. If the user has additional information that their linked-read data is comprised of larger read clouds that are nearly single-origin, such as in the MOCK5 LoopSeq dataset, then larger size cutoffs should be tried. Similarly, if the goal is to generate as much alignable genetic material as possible without as much concern for synteny or the relative spacing of genes along the genome, then larger search distances can be tried. Reference-based deconvolution, which directly approximates genomic fragments by mapping reads to the the known reference sequence composition of the sample, is better for large contiguous aligned blocks. However, the high misassembly rate is detrimental to the overall integrity of the assembly and should be carefully applied even in cases where the species composition of the sample is known. For some linked-read technologies, the average genomic fragment size must be estimated prior to library preparation to ensure the correct balance of DNA and reagent molarity [5,29,28]. This average fragment size serves as a useful upper bound of an appropriate search distance.

cloudSPAdes only uses barcode information to find subpaths of the genomic cycle after the entire assembly graph has been made. While the current version of Ariadne requires an initial run of cloudSPAdes to generate the assembly graph required for the deconvolution process, future versions will incorporate Ariadne and cloudSPAdes' scaffold generation module separately from the rest of cloudSPAdes, thus eliminating the need for and the resource consumption of an additional run of cloudSPAdes with the Ariadne-enhanced linked-reads. In its current form, Ariadne is implemented as a module of cloudSPAdes to directly ingest the generated assembly graph and generate enhanced read clouds. Further algorithmic improvements to maximize the correspondence between $3'$ barcodes and fragments may bring about significant improvements in assembly quality.

Acknowledgements

We would like to thank Indira Wu and Tuval Ben-Yehezkel at Loop Genomics for providing the MOCK5 LoopSeq dataset, as well as Yu Xia and Tom Chen at Universal Sequencing for providing the MOCK20 TELL-Seq dataset.

Funding

This work was supported by an NIGMS Maximizing Investigators' Research Award (MIRA)[grant number R35 GM138152-01 to I.H.]. LM, DM and DCD were also supported by the Tri-Institutional Training Pro-

gram in Computational Biology and Medicine (CBM) funded by the NIH grant 1T32GM083937. WNB was supported by the Tri-Institutional Computational Biology Summer Program.

References

1. Ayling, M., Clark, M.D., Leggett, R.M.: New approaches for metagenome assembly with short reads. Briefings in Bioinformatics **00**(January), 1–11 (2019). https://doi.org/10.1093/bib/bbz020
2. Bankevich, A., Nurk, S., Antipov, D., Gurevich, A.A., Dvorkin, M., Kulikov, A.S., Lesin, V.M., Nikolenko, S.I., Pham, S., Prjibelski, A.D., Pyshkin, A.V., Sirotkin, A.V., Vyahhi, N., Tesler, G., Alekseyev, M.A., Pevzner, P.A.: SPAdes: A new genome assembly algorithm and its applications to single-cell sequencing. Journal of Computational Biology **19**(5), 455–477 (2012). https://doi.org/10.1089/cmb.2012.0021
3. Bishara, A., Moss, E.L., Kolmogorov, M., Parada, A.E., Weng, Z., Sidow, A., Dekas, A.E., Batzoglou, S., Bhatt, A.S.: High-quality genome sequences of uncultured microbes by assembly of read clouds. Nature Biotechnology **36**(11), 1067–1080 (2018). https://doi.org/10.1038/nbt.4266
4. Brown, C.L., Keenum, I.M., Dai, D., Zhang, L., Vikesland, P.J., Pruden, A.: Critical evaluation of short, long, and hybrid assembly for contextual analysis of antibiotic resistance genes in complex environmental metagenomes. Scientific Reports **11**(1) (2021). https://doi.org/10.1038/s41598-021-83081-8
5. Chen, Z., Pham, L., Wu, T.C., Mo, G., Xia, Y., Chang, P., Porter, D., Phan, T., Che, H., Tran, H., Bansal, V., Shaffer, J., Belda-Ferre, P., Humphrey, G., Knight, R., Pevzner, P., Pham, S., Wang, Y., Lei, M.: Ultra-low input single tube linked-read library method enables short-read ngs systems to generate highly accurate and economical long-range sequencing information for de novo genome assembly and haplotype phasing. bioRxiv p. 852947 (01 2019)
6. Danko, D., Bezdan, D., Afshinnekoo, E., Ahsanuddin, S., Bhattacharya, C., Butler, D.J., Chng, K.R., De Filippis, F., Hecht, J., Kahles, A., Karasikov, M., Kyrpides, N.C., Leung, M.H., Meleshko, D., Mustafa, H., Mutai, B., Neches, R.Y., Ng, A., Nieto-Caballero, M., Nikolayeva, O., Nikolayeva, T., Png, E., Sanchez, J.L., Shaaban, H., Sierra, M.A., Tong, X., Young, B., Alicea, J., Bhattacharyya, M., Blekhman, R., Castro-Nallar, E., Cañas, A.M., Chatziefthimiou, A.D., Crawford, R.W., Deng, Y., Desnues, C., Dias-Neto, E., Donnellan, D., Dybwad, M., Elhaik, E., Ercolini, D., Frolova, A., Graf, A.B., Green, D.C., Hajirasouliha, I., Hernandez, M., Iraola, G., Jang, S., Jones, A., Kelly, F.J., Knights, K., Łabaj, P.P., Lee, P.K., Shawn, L., Ljungdahl, P., Lyons, A., Mason-Buck, G., McGrath, K., Mongodin, E.F., Moraes, M.O., Nagarajan, N., Noushmehr, H., Oliveira, M., Ossowski, S., Osuolale, O.O., Özcan, O., Paez-Espino, D., Rascovan, N., Richard, H., Rätsch, G., Schriml, L.M., Semmler, T., Sezerman, O.U., Shi, L., Song, L.H., Suzuki, H., Court, D.S., Thomas, D., Tighe, S.W., Udekwu, K.I., Ugalde, J.A., Valentine, B., Vassilev, D.I., Vayndorf, E., Velavan, T.P., Zambrano, M.M., Zhu, J., Zhu, S., Mason, C.E.: Global Genetic Cartography of Urban Metagenomes and Anti-Microbial Resistance. bioRxiv p. 724526 (aug 2019). https://doi.org/10.1101/724526, https://doi.org/10.1101/724526
7. Danko, D.C., Meleshko, D., Bezdan, D., Mason, C., Hajirasouliha, I.: Minerva: an alignment- and reference-free approach to deconvolve Linked-Reads for metagenomics. Genome research **29**(1), 116–124 (2019). https://doi.org/10.1101/gr.235499.118, http://www.ncbi.nlm.nih.gov/pubmed/30523036
8. Georganas, E., Egan, R., Hofmeyr, S., Goltsman, E., Arndt, B., Tritt, A., Buluc, A., Oliker, L., Yelick, K.: Extreme scale de novo metagenome assembly. Proceedings - International Conference for High Performance Computing, Networking, Storage, and Analysis, SC 2018 pp. 122–134 (2019). https://doi.org/10.1109/SC.2018.00013
9. Greer, S.U., Nadauld, L.D., Lau, B.T., Chen, J., Wood-Bouwens, C., Ford, J.M., Kuo, C.J., Ji, H.P.: Linked read sequencing resolves complex genomic rearrangements in gastric cancer metastases. Genome Medicine **9**(1), 1–17 (2017). https://doi.org/10.1186/s13073-017-0447-8
10. Guo, L., Xu, M., Wang, W., Gu, S., Zhao, X., Chen, F., Wang, O., Xu, X., Fan, G., Deng, L., Liu, X.: SLR-superscaffolder: A de novo scaffolding tool for synthetic long reads using a top-to-bottom scheme. bioRxiv (2019). https://doi.org/10.1101/762385
11. Hug, L.A., Baker, B.J., Anantharaman, K., Brown, C.T., Probst, A.J., Castelle, C.J., Butterfield, C.N., Hernsdorf, A.W., Amano, Y., Ise, K., Suzuki, Y., Dudek, N., Relman, D.A., Finstad, K.M., Amundson, R., Thomas, B.C., Banfield, J.F.: A new view of the tree of life. Nature Microbiology **1**(5), 1–6 (2016). https://doi.org/10.1038/nmicrobiol.2016.48
12. Karaoğlanoğlu, F., Ricketts, C., Ebren, E., Rasekh, M.E., Hajirasouliha, I., Alkan, C.: VALOR2: characterization of large-scale structural variants using linked-reads. Genome Biology **21**(1) (2020). https://doi.org/10.1186/s13059-020-01975-8
13. Koren, S., Harhay, G.P., Smith, T.P., Bono, J.L., Harhay, D.M., Mcvey, S.D., Radune, D., Bergman, N.H., Phillippy, A.M.: Reducing assembly complexity of microbial genomes with single-molecule sequencing. Genome Biology **14**(9) (2013). https://doi.org/10.1186/gb-2013-14-9-r101
14. Langmead, B., Salzberg, S.L.: Fast gapped-read alignment with Bowtie 2. Nature Methods **9**(4), 357–359 (2012). https://doi.org/10.1038/nmeth.1923

15. Li, H., Durbin, R.: Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics **25**(14), 1754–1760 (2009). https://doi.org/10.1093/bioinformatics/btp324
16. Marks, P., Garcia, S., Barrio, A.M., Belhocine, K., Bernate, J., Bharadwaj, R., Bjornson, K., Catalanotti, C., Delaney, J., Fehr, A., Fiddes, I.T., Galvin, B., Heaton, H., Herschleb, J., Hindson, C., Holt, E., Jabara, C.B., Jett, S., Keivanfar, N., Kyriazopoulou-Panagiotopoulou, S., Lek, M., Lin, B., Lowe, A., Mahamdallie, S., Maheshwari, S., Makarewicz, T., Marshall, J., Meschi, F., O'Keefe, C.J., Ordonez, H., Patel, P., Price, A., Royall, A., Ruark, E., Seal, S., Schnall-Levin, M., Shah, P., Stafford, D., Williams, S., Wu, I., Xu, A.W., Rahman, N., MacArthur, D., Church, D.M.: Resolving the full spectrum of human genome variation using Linked-Reads. Genome Research **29**(4), 635–645 (2019). https://doi.org/10.1101/gr.234443.118
17. Meleshko, D., Marks, P., Williams, S., Hajirasouliha, I.: Detection and assembly of novel sequence insertions using linked-read technology. bioRxiv p. 551028 (2019)
18. Mostovoy, Y., Levy-Sakin, M., Lam, J., Lam, E.T., Hastie, A.R., Marks, P., Lee, J., Chu, C., Lin, C., Dzakula, Z., Cao, H., Schlebusch, S.A., Giorda, K., Schnall-Levin, M., Wall, J.D., Kwok, P.Y.: A hybrid approach for de novo human genome sequence assembly and phasing. Nature Methods **13**(7), 587–590 (2016). https://doi.org/10.1038/nmeth.3865
19. Nurk, S., Meleshko, D., Korobeynikov, A., Pevzner, P.A.: MetaSPAdes: A new versatile metagenomic assembler. Genome Research **27**(5), 824–834 (2017). https://doi.org/10.1101/gr.213959.116
20. Sangwan, N., Xia, F., Gilbert, J.A.: Recovering complete and draft population genomes from metagenome datasets. Microbiome **4**(1) (2016). https://doi.org/10.1186/s40168-016-0154-5
21. Sedlazeck, F.J., Lee, H., Darby, C.A., Schatz, M.C.: Piercing the dark matter: Bioinformatics of long-range sequencing and mapping. Nature Reviews Genetics **19**(6), 329–346 (2018). https://doi.org/10.1038/s41576-018-0003-4, http://dx.doi.org/10.1038/s41576-018-0003-4
22. Shajii, A., Numanagić, I., Whelan, C., Berger, B.: Statistical Binning for Barcoded Reads Improves Downstream Analyses. Cell Systems **7**(2), 219–226.e5 (2018). https://doi.org/10.1016/j.cels.2018.07.005
23. Spies, N., Weng, Z., Bishara, A., McDaniel, J., Catoe, D., Zook, J.M., Salit, M., West, R.B., Batzoglou, S., Sidow, A.: Genome-wide reconstruction of complex structural variants using read clouds. Nature Methods **14**(9), 915–920 (2017). https://doi.org/10.1038/nmeth.4366
24. Tolstoganov, I., Bankevich, A., Chen, Z., Pevzner, P.A.: CloudSPAdes: Assembly of synthetic long reads using de Bruijn graphs. Bioinformatics **35**(14), i61–i70 (2019). https://doi.org/10.1093/bioinformatics/btz349
25. Wang, O., Chin, R., Cheng, X., Yan Wu, M.K., Mao, Q., Tang, J., Sun, Y., Anderson, E., Lam, H.K., Chen, D., Zhou, Y., Wang, L., Fan, F., Zou, Y., Xie, Y., Zhang, R.Y., Drmanac, S., Nguyen, D., Xu, C., Villarosa, C., Gablenz, S., Barua, N., Nguyen, S., Tian, W., Liu, J.S., Wang, J., Liu, X., Qi, X., Chen, A., Wang, H., Dong, Y., Zhang, W., Alexeev, A., Yang, H., Wang, J., Kristiansen, K., Xu, X., Drmanac, R., Peters, B.A.: Efficient and unique cobarcoding of second-generation sequencing reads from long DNA molecules enabling cost-effective and accurate sequencing, haplotyping, and de novo assembly. Genome Research **29**(5), 798–808 (2019). https://doi.org/10.1101/gr.245126.118
26. Weisenfeld, N.I., Kumar, V., Shah, P., Church, D.M., Jaffe, D.B.: Direct determination of diploid genome sequences. Genome Research **27**(5), 757–767 (2017). https://doi.org/10.1101/gr.214874.116
27. Weng, J., Chen, T., Xie, Y., Xu, X., Zhang, G., Peters, B.A., Drmanac, R.: IterCluster: a barcode clustering algorithm for long fragment read analysis. PeerJ **8**, e8431 (2020). https://doi.org/10.7717/peerj.8431
28. Wu, I., Kim, H.S., Ben-Yehezkel, T.: A Single-Molecule Long-Read Survey of Human Transcriptomes using LoopSeq Synthetic Long Read Sequencing. bioRxiv pp. 1–14 (2019). https://doi.org/10.1101/532135
29. Zheng, G.X., Lau, B.T., Schnall-Levin, M., Jarosz, M., Bell, J.M., Hindson, C.M., Kyriazopoulou-Panagiotopoulou, S., Masquelier, D.A., Merrill, L., Terry, J.M., Mudivarti, P.A., Wyatt, P.W., Bharadwaj, R., Makarewicz, A.J., Li, Y., Belgrader, P., Price, A.D., Lowe, A.J., Marks, P., Vurens, G.M., Hardenbol, P., Montesclaros, L., Luo, M., Greenfield, L., Wong, A., Birch, D.E., Short, S.W., Bjornson, K.P., Patel, P., Hopmans, E.S., Wood, C., Kaur, S., Lockwood, G.K., Stafford, D., Delaney, J.P., Wu, I., Ordonez, H.S., Grimes, S.M., Greer, S., Lee, J.Y., Belhocine, K., Giorda, K.M., Heaton, W.H., McDermott, G.P., Bent, Z.W., Meschi, F., Kondov, N.O., Wilson, R., Bernate, J.A., Gauby, S., Kindwall, A., Bermejo, C., Fehr, A.N., Chan, A., Saxonov, S., Ness, K.D., Hindson, B.J., Ji, H.P.: Haplotyping germline and cancer genomes with high-throughput linked-read sequencing. Nature Biotechnology **34**(3), 303–311 (2016). https://doi.org/10.1038/nbt.3432, http://dx.doi.org/10.1038/nbt.3432