Minerva: An Alignment and Reference Free Approach to Deconvolve Linked-Reads for Metagenomics

David C. Danko[1,2], Dmitry Meleshko[1,2], Daniela Bezdan[2], Christopher Mason[2,3], and Iman Hajirasouliha[2,4,*]

[1]Tri-Institutional Computational Biology & Medicine Program, Cornell University, NY, USA
[2]Institute for Computational Biomedicine, Department of Physiology and Biophysics, Weill Cornell Medicine of Cornell University, NY, USA
[3]The Feil Family Brain and Mind Research Institute, Weill Cornell Medicine, NY, USA
[4]Englander Institute for Precision Medicine, The Meyer Cancer Center, Weill Cornell Medicine, NY, USA
[*]Corresponding author

Abstract

Emerging linked-read technologies (aka read-cloud or barcoded short-reads) have revived interest in standard short-read technology as a viable way to understand large scale structure in genomes and metagenomes. Linked-read technologies, such as the 10X Chromium system, use a microfluidic system and a set of specially designed 3' barcodes (aka UIDs) to tag short DNA reads which were originally sourced from the same long fragment of DNA; subsequently these specially barcoded reads are sequenced on standard short read platforms. This approach results in interesting compromises. Each long fragment of DNA is covered only sparsely by short reads, no information about the relative ordering of reads from the same fragment is preserved, and typically each 3' barcode matches reads from 5-20 long fragments of DNA. However, the cost per base to sequence is far lower than single molecule long read sequencing systems, far less input DNA is required, and the error rate is that of standard short-reads.

Linked-reads represent a new set of algorithmic challenges. In this paper we formally describe one particular issue common to all applications of linked-read technology: the deconvolution of reads with a single barcode into clusters that correspond to a single long fragment of DNA. We introduce **Minerva**, A graph-based algorithm which approximately solves the barcode deconvolution problem for metagenomic data (where reference genomes may be incomplete or unavailable). Additionally, we demonstrate that deconvolved barcoded reads significantly improve downstream results by improving the specificity of taxonomic assignments, and by improving the ability of topic models to identify clusters of related sequences.

1 Introduction

1.1 Linked-Read Sequencing

Recently, long-read sequencing technologies (e.g. PacBio, Oxford Nanopore) have become commercially available. These techniques promise the ability to improve de novo assembly, particularly in metagenomics [5, 9]. While these technologies offer much longer reads than standard short-read sequencing, their base-pair error rates are substantially higher than short reads (10-15% error vs. 0.3%). More important, long-read technologies have substantially higher costs, lower throughput, and require large amounts of DNA, or PCR amplification, as input. Currently, this makes long-reads impractical for large-scale screening of whole genome or metagenome samples and most low-input clinical samples.

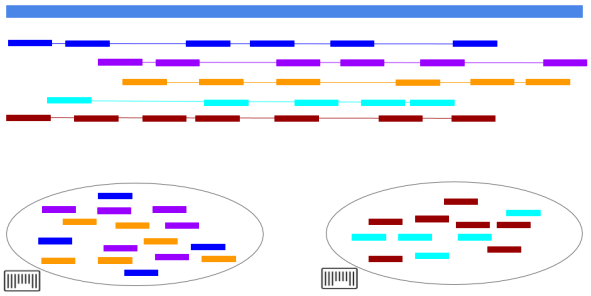

Figure 1: Cartoon representing linked-read sequencing, Different colors represent reads from different fragments that can be mixed together and tagged with the same barcode.

As an alternative, low-cost, and low-input (∼1ng) DNA library preparation techniques using microfluidic molecular barcoding methods have recently emerged (e.g. Moleculo/Illumina, 10X Genomics) that address these shortcomings. With these new technologies, input DNA is sheared into long fragments of tens to hundreds of kbp before a barcode is ligated to short reads from the fragments such that short reads from the same fragment share the same barcode (N.B. this barcode is in addition to any barcode used for sample deconvolution). Finally, the short reads are sequenced using industry standard sequencing technologies (e.g. Illumina Hiseq) . This process is commonly referred to as, linked-read sequencing. Linked-reads offer additional *long-range* information over standard short-reads that are present on the same DNA fragment (i.e. phased) and has been used recently to phase large-scale somatic structural variations [7, 17] and to improve metagenomic assembly [13].

With respect to sequence coverage, linked-read sequencing technologies consist of two key parameters: C_f, the average physical coverage of the target genome with long fragments, and C_r, the average coverage of long fragments with standard short-reads. The overall sequence coverage can be, thus, approximated as $C = C_f \times C_r$. In the case of the 10X Genomics system, which is the technique we used to generate our metagenomics linked-read data, C_r is kept very light (e.g. $0.1 - 0.2\times$; covering each long fragment only fractionally), while the total coverage C is in the typical range of standard short-read experiments.

State of the art linked-read sequencing systems group reads drawn from several long fragments under the same barcode. This is done to improve throughput given a finite number of barcodes; in practice there are roughly 5-20 long fragments represented per barcode. This can complicate downstream applications like metagenomic assembly and structural variant detection. Since many fragments are represented per barcode, a single barcode is insufficient evidence that its representative sequences occur near one another in the absence of a strong prior (e.g. an assembled genome).

Each long fragment of DNA is only fractionally covered by short reads (typically 10-20%). Thus, reads from the same fragment do not typically overlap (this is not the case with certain older technologies and types of sample preparation). As such it is not possible to do de novo assembly using only reads with the same barcode. However, linked-reads still provide significantly more information about the proximity of short reads than standard short read sequencing. Given a large amount of background sequence the co-occurrence of particular sequences in multiple barcodes provides strong guarantees that the co-occurring sequences were drawn from the same underlying fragments of DNA, provided the sequences are not extremely common.

1.2 Applications of Linked-Reads in Metagenomics

Linked-reads carry several potential benefits for metagenomic research compared to standard short-reads. Linked-reads can be used to improve taxonomic classification of reads, improve the assembly of microbial genomes, identify horizontally transferred sequences, quantify the genetic structure of low-abundance organisms and catalog intra-sample genetic structural variants. In the near term many algorithms for analyzing short read sequences can be used on linked-read data without modification.

Compared to long-read sequencing, linked-reads can be used to sequence samples far more deeply for the same amount of money and can accept much smaller amounts of input DNA. Even at the same read

depth, linked-reads may be more useful for studying low abundance organisms because linked-reads span a much longer stretch of a genome for the number of bases sequenced (i.e. very high physical coverage).

In this paper, we address the barcode deconvolution problem, a fundamental problem of using linked-reads for metagenomics. We show that several applications of linked reads can improve on short reads and that specifically addressing the barcode deconvolution problem may improve downstream results for multiple applications.

1.3 The Barcode Deconvolution Problem

We formally define the barcode deconvolution problem. As input we are given a set of n reads. Each read has the same barcode and an unobserved class that represents the fragment that the read was drawn from. For a given barcode with n reads drawn from f fragments we have $\vec{R} = \langle r_1, \cdots, r_n \rangle$ where r_i represents the unobserved class for read i, and $\vec{F}$, the set of possible fragments $[1, f]$.

A solution to the barcode deconvolution problem would be a function

$$D : \vec{R} \mapsto \vec{F}$$

such that

$$D(i) = D(j), \forall_j r_i = r_j, \forall_i \in [1, n]$$

When a reference genome is available and structural variants are not expected the barcode deconvolution problem is relatively trivial. All individual reads with the same barcode can be mapped to the reference genome using any good read alignment method. Reads from the same fragment will tend to be clustered near one another on the reference genome and, unless the genome in question is small, there is little chance that reads from different fragments will be proximal. However, if a reference genome is not available, is relatively small, or structural variation is present read mapping may not provide a good solution to the barcode deconvolution problem. All of these conditions are common in metagenomics.

1.4 Summary

We have developed a novel method, **Minerva**, that explicitly uses barcode co-occurrence information to cluster reads with the same barcode into groups of reads that were drawn from a single long fragment. Our approach draws heavily from the field of topic modeling in Natural Language Processing (NLP) which studies methods to find groups of co-occurring words in text. In particular, we introduce a procedure 2.2 that achieves great specificity and performance when certain features of linked-reads are explicitly encoded as well as a probabilistic generative model justifying key assumptions of our procedure 2.1. We report our negative results, models that we tested but performed poorly 2.5. Finally, we demonstrate how our techniques can be effectively applied to real metagenomic linked-read data 4.1 and improve downstream analysis 4.2, 4.3.

2 Barcode Deconvolution Using a Graphical Model

We developed a graph based algorithm to subdivide reads with the same barcode into groups that better corresponded to a single molecule. The core intuition behind this approach was that reads from the same fragment would tend to overlap with similar sets of reads that had different barcodes. Though each set of reads with the same barcode can contain reads from several fragments (5-20), it is unlikely that two sets of reads with two different barcodes would contain reads from multiple separate overlapping fragments. If multiple reads with the same barcode overlapped with reads that had another barcode it would suggest that the original reads came from the same fragment.

2.1 Mathematical Justification of the Model

We have developed a simple and conservative generative model to justify our intuition that reads from the same fragment will overlap with similar barcodes. This model is similar to empirical results and can be used to inform the parameters used for deconvolution.

We model each microbial genome G_i in a metagenome G as a discrete collection of fragments F_i that do not overlap. The probability of selecting a given fragment $F_{i,j}$ from a given microbial genome G_i is given by a uniform distribution. We model each microbial genome as a set of N_g discrete fragments

that do not overlap and model the probability of selecting a given genome conservatively as a Geometric distribution.

$$P(F = F_{i,j}|G_i) = \frac{1}{|F_i|}$$

$$|F_i| = N_g \forall i \in 1 : |G|$$

$$P(G = G_i) = \frac{1}{2^i}$$

This gives us the probability of drawing a single fragment as well as the probability of drawing the same fragment twice

$$P(F_{i,j}) = \frac{1}{G_f * 2^i}$$

$$P(F_{w,x} = F_{y,z}|G_w = G_y) = \frac{1}{N_g}, P(G_w = G_y) = \lim_{|G|\to\infty} \frac{1}{2}$$

$$p_f = P(F_{w,x} = F_{y,z}) = \frac{1}{2N_g}$$

We model each barcode as a selection of N_b fragments with the simplification that a fragment is never drawn twice even though the probability of drawing a given fragment is not altered between draws. We can model the probability of two barcodes sharing exactly k fragments.

$$X(k) = P(|B_i \cap B_j| = k) = \binom{N_b}{k} * p_f^k * (1 - p_f)^{N_b - k}$$

$$P(|B_i \cap B_j| > 1) = \sum_{i=2}^{N_b} X(i) = 1 - X(0) - X(1)$$

If we choose reasonable, conservative, values for all parameters $N_b = 5, N_g = 100, |G| = 10$

$$p_f = \frac{1}{200}$$

$$X(1) \approx 0.0245$$

$$X(\geq 2) = 1 - X(0) - X(1) \approx 0.000248$$

$$\frac{X(1)}{1 - X(0) - X(1)} \approx 100$$

We find that it is about 100 times more likely to have exactly one overlapping fragment than multiple overlapping fragments with the same barcode, even with conservative parameters. However, this model does not account for the fact that individual fragments may have similar sequences, this is a major source of noise for **Minerva**. However, we can use the parameters of this model to justify removing any overlaps that occur far more often than expected.

We can further inform our models parameters by accounting for the light coverage of each long fragment. On average, each fragment in a dataset is only fractionally covered at a rate of C_r (with a read depth of 1). While the precise coverage might vary between fragments this parameter can be used to estimate the size of overlaps between fragments and their expected sparsity. Two long fragments could be expected to overlap at C_r^2 points in their overlap. By measuring actual overlap compared to the number of reads not overlapping it should be possible to estimate the size of the overlap. In cases where fragments overlap much more frequently than C_r^2 it can be inferred that the sequence present is repetitive or well conserved.

These facts are used in **Minerva** to filter connections between repetitive regions, restrict overlaps to regions of a certain length, and to heuristically filter comparisons between barcodes unlikely to have significant overlap. This carries practical performance benefits and reduces errors.

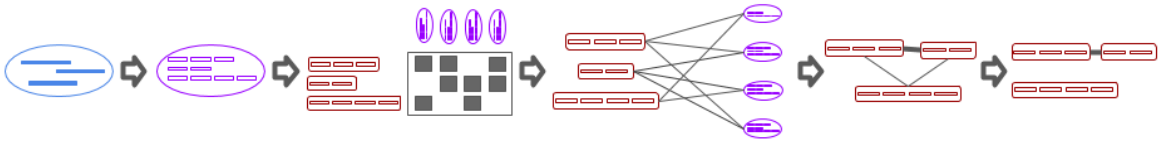

Figure 2: From left 1) Fragments are generated 2) Fragments are sequenced and tagged 3) Reads in a given barcode are aligned to other barcodes 4) A bipartite graph between reads and barcodes is constructed 5) A graph between reads that co-occur with barcode is constructed 6) Reads are clustered into groups

2.2 A Graphical Model of Barcode Deconvolution

We have developed a graphical model that effectively deconvolves the reads within a given barcode. In essence, the model works by constructing a bipartite graph between all reads with a particular barcode and all other barcodes. Reads have an edge to a barcode if they are found to contain a kmer that is specific to exactly one read in the foreign barcode. Once the bipartite graph is constructed barcodes and reads with too many or too few edges (by user supplied parameters) are removed. The filtered bipartite graph is used to construct an adjacency matrix between reads and the matrix is clustered into groups of reads.

Initially each barcode in a given dataset is parsed into a sketch of minimizing kmers (see section 2.4, figure 2 part 2). Global counts for kmers are retained, kmers that occur exactly once or 10 times more than average are discarded. Singleton kmers cannot possibly overlap between barcodes and overly common kmers tend to create false positives, this is analogous to removing stop words in Natural Language Processing applications. A map of kmers to reads is retained for every barcode.

After parsing, each barcode is compared to every other barcode (figure 2 parts 3 and 4). Comparisons between barcodes that share too many kmers are discarded as these likely represent low complexity or evolutionarily conserved regions as opposed to real overlaps. Comparisons between barcodes that share too few kmers are also rejected. The intersection of the kmer sets between both barcodes is calculated and edges are added between every read with an intersecting kmer and a node representing the foreign barcode. If the degree of the node representing the barcode is too high it is discarded. All cutoffs are determined by user supplied thresholds.

Each bipartite graph representing the reads with a given barcode is given a final round of filtering where reads that matched too many barcodes are removed based on a user supplied threshold. The filtered bipartite graph is converted to an adjacency matrix of reads where the similarity between reads is equivalent to the number of barcodes both reads mutually overlapped with (figure 2 part 5). This adjacency matrix is converted to a binary matrix by setting all values below a user supplied threshold to zero and all remaining values to one (figure 2 part 6).

All connected components in the binary matrix are found. Connected components consisting of single reads are discarded, the remaining components define clusters. This process is analogous to DBSCAN [4] for graphs.

Depending on the parameters and dataset used the clustering step may produce clusters that are large and imprecise or small and incomplete; this is partly a consequence of using a small number of discrete components to define connections. To mitigate this issue we have introduced two enhancements to the clustering step that may optionally be used. We have based these enhancements on the two step procedure used for community recovery by [2] and [14].

We term the first enhancement read-rescue; in some instances it is useful to rescue components that consist of single reads by connecting them to clusters that they unambiguously share connections with. To do this all single reads are checked for connections to any read in a cluster. If the single read is connected to enough reads (by a user supplied parameter) in no more than one cluster the read is added to that cluster. We term the second enhancement edge-cracking; when clusters are too large it is useful to break certain edges that bridge two individually well-connected components. To do this all edges in the adjacency matrix are considered. If the two reads attached to an edge do not share enough neighboring reads (by a user supplied parameter) the edge is removed from the adjacency graph.

2.3 Information Theory Bounds on Barcode Deconvolution

We note that the barcode deconvolution on the graphical model we have described in 2.2 is analagous to the cluster recovery problem [6] in Information Theory. In particular, barcodes that overlap with reads being deconvolved provide linkage information between reads. We have used this linkage information to construct a graph between the reads being deconvolved with the expectation that reads from the same

fragment will have a better chance of being linked than reads from different fragments. Formally we say that two reads are connected with probability p if they are from the same fragment and probability q if they are from different fragments. Termed differently p is the true positive rate while q is the false positive rate.

If we make a simplifying assumption that all fragments in our barcode produced equal numbers of reads we can use the formula determined by [8] to determine the minimum connectivity of linking barcodes necessary to deconvolve our reads. We define the number of reads per fragment as $\frac{N_r}{N_b}$, where N_r is the total number of reads being deconvolved and N_b is the number of fragments in the given barcode.

For the community recovery problem [8] have provided a lower bound on the size of graph that can be accurately clustered given values of p and q, regardless of the algorithm used. If a graph is smaller than this threshold it is unlikely that it will be possible to distinguish clusters from spurious edges. This boundary requires us to assume that all fragments with the same barcode produced equal numbers of reads. We define the number of reads per fragment as $\frac{N_r}{N_b}$, where N_r is the total number of reads being deconvolved and N_b is the number of fragments in the given barcode. With this definition we can apply the following inequality to barcode deconvolution:

$$\frac{N_r}{\log N_r}(\sqrt{p} - \sqrt{q})^2 > N_b$$

If a barcode graph does not meet this inequality it is unlikely that it will be possible to accurately reconstruct all clusters, This can be used to estimate the minimum number of reads and maximum error rates which can lead to effective barcode deconvolutions. In principle, this inequality should apply to all barocde deconvolution algorithms that can be formulated as a graph. However different algorithms may have different values of p and q. We also note that the above formula is based on asymptotic behavior for graphs with thousands of nodes. Typical deconvolution graphs have less than 100 nodes.

2.4 Minimum Sparse Hashing

Minerva frequently tests whether pairs of reads overlap. Many solutions to finding overlaps between reads exist, such as: sequence clustering algorithms, sequence aligners, and kmer matching. These techniques typically make trade offs between overall performance and error rates. Since Minerva is meant to be relatively fast and can tolerate some errors we elected to use a minimal sparse hash of kmers to match read pairs. This technique was originally developed independently for biological sequence search and natural language document search [10, 16] (in natural language search the technique is referred to as winnowing). While the original application of this technique in biology defined minimization as the lexicographic minimum of a set of sequences we use a uniform random hash function to determine the minimal sequence in a set. This is a common practical enhancement and the mathematics of its use are recently detailed by [15].

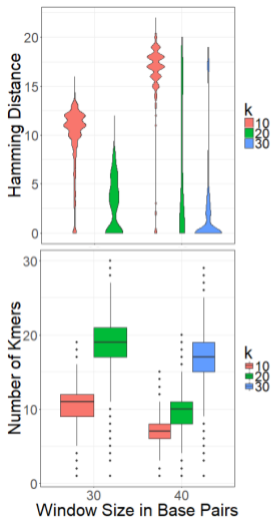

Figure 3: Top, Hamming distance between windows in reads that share kmers, using various parameters. Bottom, number of representative kmers per read

Minimum sparse hashing for sequences takes three parameters, a length k, a window size w, and a hash function h. Given a set K of $n, n \geq w$ kmers, the min-sparse hash computes the hash h of each kmer then selects the kmer with the smallest hash from each consecutive set of w kmers in K. The final set of minimizers is the unique set of kmers generated W, since each consecutive window shares $w - 1$ kmers there is a good chance that each window shares the same minimum with its predecessor. Formally $W = \{\min(h(k) \forall k \in K_{(i,i+w)}) \forall i \in 0 : (n - w)\}$. This algorithm guarantees that any pair of reads with an exact overlap of at least $w + k - 1$ bases will share at least one minimum sparse kmer while drastically reducing the number of kmers which must be stored in memory (figure 3). In certain implementations minimum sparse kmers may also improve performance by allowing a kmer that can be stored in a single 64 bit cell ($k \leq 32$) of memory to represent a longer sequence.

Minimum sparse kmers are prone to false positives when presented with similar, but not identical, runs of w bases in read pairs. We measured this phenomenon by comparing all kmers of length w from pairs of reads that share a minimum sparse hash. Figure 3 shows the minimum hamming distance for windows of length w between reads that share a minsparse hash. When k is larger the average hamming

distance is smaller though outliers persist. Small values of k produce many distant false positives. In practice we found that false positives could be controlled through downstream techniques. Raising k from 20 to 30 ($w = 40$) improved accuracy and precision but not overwhelmingly so ($k = 10$ performed poorly in all cases).

The mathematics that underly minimum sparse hashing may also be used to efficiently approximate the overlap between sets, another important operation for Minerva. We did not use this technique in our current implementation of Minerva but plan to explore this for later versions.

2.5 Failed Models

While developing this model we tested several variants that performed poorly. In this section we report these models as negative results, parameters and outcomes may be found in the supplement. We define three distinct components of our model:

1. Barcode reduction, a technique to generate a matrix with rows representing reads and columns related to barcodes
2. Cell weighting, a technique to reweight the cells of our reduced matrix
3. Clustering, alternate techniques to cluster our reduced, reweighted matrix.

We tried several techniques to reduce our matrix (our final technique was to build a bipartite graph that was converted into an adjacency matrix) including: Principal Component Analysis, Latent Dirichlet Allocation (treating each read as a document of barcodes), the jaccard, dice and kulsinski distance matrices between reads.

For most reductions we tried to reweight our matrices by normalizing row and column sums. We also tried term frequency - inverse document frequency and just inverse document frequency. For our final model we did not use any reweighting scheme.

To cluster our reduced and reweighted matrix we tried xmeans, kmedoids, hierarchical clustering with various linkages and Markov clustering. Hierarchical clustering yielded reasonable performance but dbscan was faster, had more intuitive parameters, and had slightly better performance.

Additionally, we tried to use LDA globally (treating each barcode as a document of kmers) to discover sets of kmers that uniquely identified fragments. This technique provided surprisingly good barcode deconvolution (and automatically defined globally clusterings of barcodes) but was extremely computationally costly and did not yield results as good as our final model.

3 Primary Data Sets

We tested Minerva using two primary real data sets from two microbial mock communities. The first community (Dataset 1) contained five bacterial species: *E. coli, Enterobacter cloacae, Micrococcus luteus, Pseudomonas antarctica, and Staph. epidermis.* The second community (Dataset 2) contained 8 bacterial species and 2 fungi: *Bacillus subtilis, Cryptococcus neoformans, Enterococcus faecalis, E. coli, Lactobacillus fermentum, Listeria monocytogenes, Psuedomonas aeruginosa, Sachharomyces cerevisiae, Salmonella enterica, and Staphylococcus aureus.*

Roughly 1ng of high molecular weight DNA was extracted from each sample. The HMW DNA was processed using a 10X Chromium instrument and we prepared sequencing libraries. Each library was sequenced on an Illumina Hiseq with 2x150 paired-end reads. Roughly 20M reads were generated for each sample, for testing we selected 10M reads from each while ensuring that we only selected complete

Table 1: Dataset Properties

	Dataset 1	Dataset 2
Number of Read Pairs	$10 * 10^6$	$10 * 10^6$
Number of Species	5	10
Mean Barcode Richness	2.74	5.79
Mean Barcode Size	7.399	7.515
Barcode N50 Size	11	11
Barcode N90 Size	4	4

barcodes. Both samples showed some evidence of human contamination, reads that mapped to the human genome were not removed from the samples (but were not used to generate statistics on barcode purity) since some amount of human DNA is typical in metagenomic samples.

In both samples reads were evenly distributed over 3M barcodes. We used *LongRanger Basic* [1] to attach barcodes to reads and perform error correction on barcodes. Both samples has a similar number of reads per barcode. Sample 2 had more species represented in each barcode on average, though not necessarily more fragments since fragments can originate from the same genome. Statistics about the datasets are summarized in Table 1. Both datasets are available through the 10X Metagenomics Consortium.

4 Experimental Results

4.1 Minerva Deconvolves Barcodes

Minerva was able to identify subgroups in barcodes that largely corresponded to individual fragments of DNA. We term these subgroups 'Enhanced Barcodes'. We measured the quality of each enhanced barcode using two metrics: shannon diversity index $H = \sum p_i \log p_i$ and purity $P = \max(\vec{p})$ where p_i indicates the proportion of an enhanced barcode that belongs to each fragment. These values are shown in figure 4 as compared to barcodes that were not enhanced. In general, Minerva produced a large number of perfect ($P = 1, H = 0$) enhanced barcodes. Note that the results in figure 4 were produced without using read-rescue or edge-cracking.

We also tested whether the quality of the enhanced barcodes changed with the number of reads in the barcode. We found a small inverse relationship between barcode size and purity but established that our previous results were not being inflated by a large number of very small enhanced barcodes (enhanced barcodes of size 1 are always excluded from results).

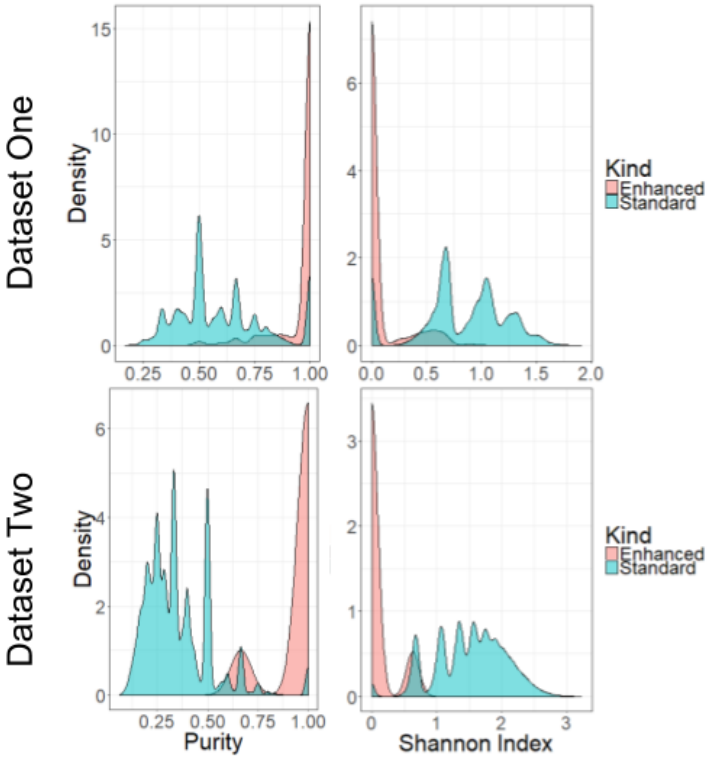

Figure 4: Clockwise from top left 1) Purity in dataset one for enhanced and standard barcodes 2) Shannon index in dataset one for enhanced and standard barcodes 3) Shannon index in dataset two for enhanced and standard barcodes 4) Purity in dataset two for enhanced and

In testing we found three parameters that seemed to have the most effect of Minerva's performance. The number of links required between reads to form a cluster (*eps*), the kmer size within each MinSparse hash (K), and the maximum allowed frequency of each read (*maxk*). In supplementary figure 1 we show how these parameters affect Minerva's performance under three different metrics: mean enhanced barcode purity, mean enhanced barcode size, and total reads clustered. Large, pure, and complete clusters being the ideal. We found that Minerva's parameters could be used to tune performance between very large and very pure enhanced barcodes depending on the downstream application.

4.2 Enhanced Barcodes can be Clustered into Meaningful Groups Using Topic Models

After deconvolving barcodes into enhanced barcodes it is useful to group enhanced barcodes that likely came from the same genome. This is essentially a clustering problem. Initially we explored graphical approaches similar to our algorithm for read cloud deconvolution. These algorithms relied on the assumption that elements being clustered would have small numbers of distinguishing elements and a relatively high a priori probability of originating from the same cluster. When dealing with individual barcodes these assumptions proved reasonable; faced with the complexity of a full dataset these assumptions became inaccurate and graph based algorithms performed poorly.

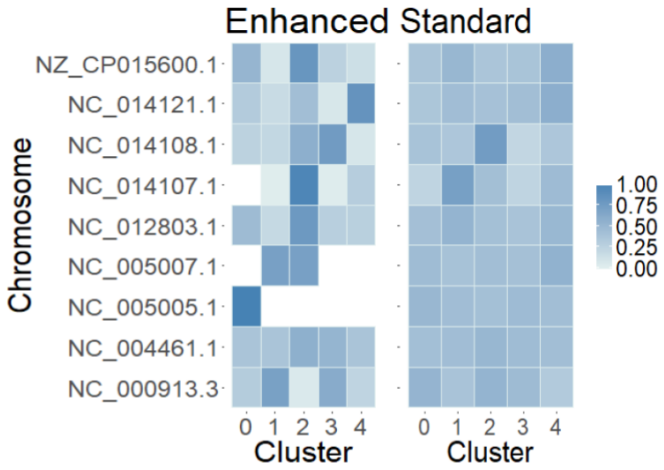

Figure 5: Abundance of different chromosomes across clusters as assigned by LDA. Enhanced barcodes dramatically improve LDA's ability to distinguish structure.

With relatively little structure in the data that could be known a priori we turned to topic modeling algorithms to discover implicit genetic structures in our data. Latent Dirichlet Allocation is a classic model in Natural Language Processing [3]. LDA is a generative model that assumes data was created using a certain, well defined, stochastic process. Unlike our graph based algorithm this model had a large number of parameters that could be adjusted until a model that fit the data well was found. Training the model consists of finding parameters that make it more likely that the observed data would be generated using the given stochastic process; typically this is done with Gibbs sampling.

Typically LDA is used to analyse corpora of natural language. Natural language corpora are organized into documents (e.g. emails or book chapters) that consist of words. The base version of LDA does not consider what order words in a document occur, just how often each word occurs in a given document; this is referred to as a bag-of-words model. Formally, documents are modeled as a sparse vector over a large vocabulary of words where entries represent the number of times a word occurs in the document. LDA maps documents from a high dimensional word-space to a lower dimensional topic-space. In NLP topics typically have intuitive interpretations as thematically consistent units. A key advantage of LDA is that it can classify words based on context (i.e. a river bank vs. a financial bank), this may be useful for classifying conserved motifs.

We used LDA to project enhanced and standard barcodes into a lower dimensional space. We treated each barcode as a document containing minimum sparse kmers as words. We removed kmers that occured far more often than average in a process similar to removing stop-words in NLP. We ran LDA with hyper-parameter optimization on our barcode-documents and clustered to obtain a topic cluster for each barcode using the implementation LDA in Mallet [11]. Using X-Means we clustered the barcode vectors into discrete groups.

Without enhanced barcodes LDA essentially cannot distinguish any structure, with enhanced barcodes LDA can generate clusters that are less diverse. This could be used to improve assemblies by clustering similar reads and reducing spurious connections.

4.3 Minerva Improves Short Read Taxonomic Assignment

We used Minerva to improve the specificity of short read taxonomic assignments obtained from Kraken, a popular pseudo-alignment based tool [18]. We selected Kraken because it was found to have good precision but relatively poor recall in a study by McIntyre et al. [12]. We observed that if reads from a single linked fragment were classified using Kraken or similar pseudo-aligners many of the reads could not be assigned to low taxonomic ranks, however since the reads must have come from the same underlying fragment of DNA they must all have the same taxonomic rank. This fact could be used to rescue reads without a specific annotation. Since Minerva approximates pure read groups we used it to rescue taxonomic assignments used by Kraken. Since Minerva's enhanced barcodes can contain impurities we restricted annotation-rescue to cases where the given taxonomic assignment was not in direct conflict with the lowest level taxonomic assignment present. (i.e. a read assigned by Kraken to Bacteroidetes would never be promoted to E. Coli). Additionally we required that the lowest taxonomic rank be

supported by at least two annotated reads, in principle reads could actually be demoted.

Using this technique we were able to rescue a large number of reads from unspecific taxonomic assignments. This is shown in figure 6 for E. Coli. All observed taxonomic assignments were correct after promotion. Without enhanced reads many annotations cannot be rescued or are incorrect.

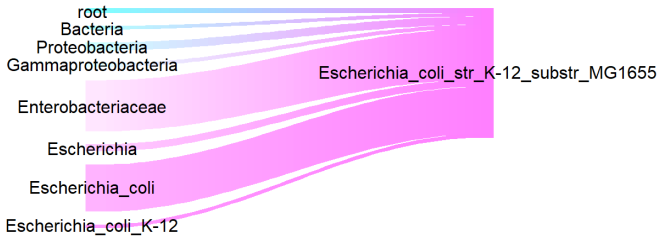

Figure 6: Using enhanced barcodes we can promote the taxonomic assignment of reads. Width of each frond is proportional to the number of reads promoted from a specific rank.

5 Conclusion

We have introduced **Minerva** a graph based algorithm to provide a partial solution to the barcode deconvolution problem. By design Minerva provides conservative solutions to barcode deconvolution for metagenomics and uses essentially no information (except kmer overlaps) about the sequences being clustered. As such Minerva is a relatively pure demonstration of how information can be extracted from linked-reads. With some modification the algorithms underlying Minerva may even be useful for detecting structural variations and other genetic structures in the human genome.

However, the current version of Minerva could be enhanced by leveraging a number of practical sequence features: such as known taxonomic assignment, GC content, tetramer frequency, or motifs. These have been shown to be good indicators of lineage in metagenomics and could be easily incorporated to improve Minerva's clusterings. In particular, taxonomic assignments could be incorporated into Minerva to evaluate barcode deconvolution, since there is no a-priori reason to think reads with a known taxonomic classification would be deconvolved more effectively than reads that could not be classified.

The current version of Minerva provides reasonable performance but can still up to 100 cpu-hours to deconvolve some large-scale datasets. A large performance bottleneck is Minerva's routine to calculate the size of an intersection between two sets which is naive and exact. Jain et al. [9] has shown that bloom filters can be effectively used to speed up the calculation of set intersection in biology with acceptable errors. Future versions of Minerva will employ similar techniques to improve performance. Minerva uses the same parameters to process every barcode, however the nature of linked-read sequencing provides a rich source of information that could be used to optimize model parameters for deconvolving individual barcodes. This would require a more thorough mathematical model of linked reads which we leave to a future work. Similarly, external sequence annotation could be incorporated as a practical approach to setting parameters for individual barcodes though it is unlikely that such a technique would generalize to non-microbial applications.

Of particular interest to us is the possibility of using Minerva to directly improve downstream applications. For simple applications Minerva may be used with a single set of parameters to produce a deconvolution that meets certain requirements. For applications built to take advantage of barcode deconvolution Minerva could be run with multiple parameters to produce increasingly strict tiers of enhancement. This may be particularly important for de-bruijn graph assembly. DBG assembly typically relies on effectively trimming and finding paths through a de-bruijn graph. Multiple tiers of linkage between reads could be used to inform trimming or path finding programs about likely paths and spurious connections. This could likely be modeled either as an information theory or probabilistic approach depending on the situation and assembler.

Overall we believe that Minerva is an important step towards building techniques designed to take advantage of linked-reads. Linked-reads have the potential to dramatically improve detection of large genetic structures without dramatically increasing sequencing costs and while taking advantage of existing techniques to process short reads.

6 Availability

The version of Minerva used to write this paper is freely available at `https://github.com/dcdanko/minerva_barcode_deconvolution`. The fresh data sets used to evaluate Minerva is available upon request.

7 Acknowledgement

We acknowledge Victoria Popic and her team at Illumina Inc. for fruitful discussions on the problem and her help in benchmarking. We sincerely thank 10X Genomics Inc. for providing us valuable fresh sequencing data sets used in this study for evaluations. We also thank Stephen Williams of 10X Genomics Inc. for coordinating the 10X Metagenomics consortium in which our team participates. DCD and DM were supported by the Tri-Institutional Training Program in Computational Biology and Medicine (CBM) funded by an NIH T32 grant.

References

[1] 10x genomics long ranger pipelines: https://support.10xgenomics.com/genome-exome/software/pipelines/latest/advanced/other-pipelines.

[2] Emmanuel Abbe, Afonso S. Bandeira, and Georgina Hall. Exact recovery in the stochastic block model. *IEEE Transactions on Information Theory*, 62(1):471–487, 2016.

[3] David M Blei, Andrew Y Ng, and Michael I Jordan. Latent Dirichlet Allocation. *Journal of Machine Learning Research*, 3:993–1022, 2003.

[4] Martin Ester, Hans-Peter Kriegel, Jorg Sander, and Xiaowei Xu. A density-based algorithm for discovering clusters in large spatial databases with noise. In *International Conference on Knowledge Discovery and Information Retrieval*, pages 226–231, 1996.

[5] J. A. Frank, Y. Pan, A. Tooming-Klunderud, V. G. H. Eijsink, A. C. McHardy, A. J. Nederbragt, and P. B. Pope. Improved metagenome assemblies and taxonomic binning using long-read circular consensus sequence data. 6:25373 EP –, 05 2016.

[6] M. Girvan and M. E. J. Newman. Community structure in social and biological networks. *Proceedings of the National Academy of Sciences*, 99(12):7821–7826, 2002.

[7] Stephanie U. Greer, Lincoln D. Nadauld, Billy T. Lau, Jiamin Chen, Christina Wood-Bouwens, James M. Ford, Calvin J. Kuo, and Hanlee P. Ji. Linked read sequencing resolves complex genomic rearrangements in gastric cancer metastases. *Genome Medicine*, 9(1):57, 2017.

[8] Bruce Hajek, Yihong Wu, and Jiaming Xu. Achieving Exact Cluster Recovery Threshold via Semidefinite Programming: Extensions. In *IEEE Transactions on Information Theory*, volume 62, pages 5918–5937, 2016.

[9] Miten Jain, Sergey Koren, Josh Quick, Arthur C Rand, Thomas A Sasani, John R Tyson, Andrew D Beggs, Alexander T Dilthey, Ian T Fiddes, Sunir Malla, Hannah Marriott, Karen H Miga, Tom Nieto, Justin O'Grady, Hugh E Olsen, Brent S Pedersen, Arang Rhie, Hollian Richardson, Aaron Quinlan, Terrance P Snutch, Louise Tee, Benedict Paten, Adam M. Phillippy, Jared T Simpson, Nicholas James Loman, and Matthew Loose. Nanopore sequencing and assembly of a human genome with ultra-long reads. *bioRxiv*, 01 2017.

[10] Guillaume Marçais, David Pellow, Daniel Bork, Yaron Orenstein, Ron Shamir, and Carl Kingsford. Improving the performance of minimizers and winnowing schemes. In *Bioinformatics*, volume 33, pages i110–i117, 2017.

[11] Andrew Kachites McCallum. MALLET: A Machine Learning for Language Toolkit, 2002.

[12] Alexa McIntyre, Rachid Ounit, Ebrahim Afshinnekoo, Robert Prill, Elizabeth Henaff, Noah Alexander, Sam Minot, David Danko, Jonathan Foox, Sofia Ahsanuddin, Scott Tighe, Nur A Hasan, Poorani Subramanian, Kelly Moffat, Shawn Levy, Stefano Lonardi, Nick Greenfield, Rita Colwell, Gail Rosen, and Christopher E Mason. Comprehensive Benchmarking and Ensemble Approaches for Metagenomic Classifiers. *bioRxiv*, jun 2017.

[13] Eli Moss, Alex Bishara, Ekaterina Tkachenko, Joyce B Kang, Tessa M Andermann, Christina Wood, Christine Handy, Hanlee Ji, Serafim Batzoglou, and Ami S Bhatt. De novo assembly of microbial genomes from human gut metagenomes using barcoded short read sequences. *bioRxiv*, apr 2017.

[14] Elchanan Mossel, Joe Neeman, and Allan Sly. Consistency Thresholds for Binary Symmetric Block Models. *arXiv preprint arXiv:1407.1591*, 2014.

[15] Yaron Orenstein, David Pellow, Guillaume Marçais, Ron Shamir, and Carl Kingsford. Compact universal k-mer hitting sets. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, volume 9838, pages 257–268, 2016.

[16] Saul Schleimer, Daniel S. Wilkerson, and Alex Aiken. Winnowing: Local Algorithms for Document Fingerprinting. *Proceedings of the 2003 ACM SIGMOD international conference on on Management of data - SIGMOD '03*, pages 76–85, 2003.

[17] Noah Spies, Ziming Weng, Alex Bishara, Jennifer McDaniel, David Catoe, Justin M Zook, Marc Salit, Robert B West, Serafim Batzoglou, and Arend Sidow. Genome-wide reconstruction of complex structural variants using read clouds. *Nat Meth*, advance online publication:–, 07 2017.

[18] Derrick E Wood and Steven L Salzberg. Kraken: ultrafast metagenomic sequence classification using exact alignments. *Genome Biology*, 15(3):R46, 2014.