Sketching and sampling approaches for fast and accurate long read classification

Arun Das[1], Michael C. Schatz[1]

[1]Department of Computer Science, Johns Hopkins University, Baltimore, MD, 21218, USA

Contact: arun.das@jhu.edu, mschatz@cs.jhu.edu

Abstract

In modern sequencing experiments, quickly and accurately identifying the sources of the reads is a crucial need. In metagenomics, where each read comes from one of potentially many members of a community, it can be important to identify the exact species the read is from. In other settings, it is important to distinguish which reads are from the targeted sample and which are from potential contaminants. In both cases, identification of the correct source of a read enables further investigation of relevant reads, while minimizing wasted work. This task is particularly challenging for long reads, which can have a substantial error rate that obscures the origins of each read.

Existing tools for the read classification problem are often alignment or index-based, but such methods can have large time and/or space overheads. In this work, we investigate the effectiveness of several sampling and sketching-based approaches for read classification. In these approaches, a chosen sampling or sketching algorithm is used to generate a reduced representation (a "screen") of potential source genomes for a query readset before reads are streamed in and compared against this screen. Using a query read's similarity to the elements of the screen, the methods predict the source of the read. Such an approach requires limited pre-processing, stores and works with only a subset of the input data, and is able to perform classification with a high degree of accuracy.

The sampling and sketching approaches investigated include uniform sampling, methods based on MinHash and its weighted and order variants, a minimizer-based technique, and a novel clustering-based sketching approach. We demonstrate the effectiveness of these techniques both in identifying the source microbial genomes for reads from a metagenomic long read sequencing experiment, and in distinguishing between long reads from organisms of interest and potential contaminant reads. We then compare these approaches to existing alignment, index and sketching-based tools for read classification, and demonstrate how such a method is a viable alternative for determining the source of query reads. Finally, we present a reference implementation of these approaches at https://github.com/arun96/sketching.

Keywords: Sketching, Sampling, Classification, MinHash, Metagenomics

1. Introduction

Metagenomics has become an increasingly popular area of study over the past two decades, and has enabled us to better understand the diversity, interactions and evolution of microbial communities in a plethora of environments [1]–[3]. Metagenomics has highlighted the problem of being able to quickly and accurately identify the source of a given DNA sequence from all the genomic material in a given sample. This is needed to classify and sort reads for further downstream analysis, and to identify and remove potential contaminants that are present in a sample.

The read classification problem is to identify the source genome of a given input read, usually by comparing the read to a list of potential source genomes and choosing the one with the highest similarity. This comparison may be done naively by comparing the entirety of each read to the entirety of each genome to find the best alignment or through an exhaustive analysis of k-mers present. While these approaches are highly accurate they can incur high computational overheads, which presents an opportunity for lower overhead techniques such as sketching or sampling, especially for long read data.

Sketching is the process of generating an approximate, compact summary of the data (a "sketch"), which retains properties of interest and can be used as a proxy for the original data [4]. Sampling selects a subset of the data, either systematically or randomly, but does not guarantee the preservation of these properties. Each has unique advantages: sketching has been shown to bound error better than sampling [4], [5], while systematic sampling (such as uniform sampling) can provide bounds on the number of samples from specific sections of the original data included in the generated subset. Both sketching and sampling provide simple routes to greatly reduce the size of an input set, while retaining the characteristics and features that identify the set, thus allowing a comparable level of accuracy.

One of the most well-known sketching approaches, and the main one we employ in our work, is MinHash, which was first presented as a method to estimate document similarity using the similarity between their hashed sub-parts [6]. It is now widely used in genomics, such as in Mash [7], which performs fast similarity and distance estimation between two input sequences, and tools such as Mash Screen [8] which uses MinHash to predict which organisms are contained in a mixture. Other tools include MashMap [9], which blends minimizers and MinHash for fast, approximate alignment of DNA sequences, and MHAP [10] to accelerate genome assembly. Beyond MinHash, several related approaches have been proposed, such as bloom filters [11], [12], the HyperLogLog sketch [13], [14], and other sketching approaches to estimate similarity, containment or cardinality [15].

1.1. Approaches to Read Classification

The simplest approach to read classification is to align each query read to all potential source genomes, and using the genome with the best alignment as the predicted source. While the most accurate approach would be exhaustive sequence-to-sequence alignment with dynamic programming, this is impractically slow, so aligners, such as Minimap2 [16] and Winnowmap [17] typically use some form of seed-and-extend that start with exact matches and build out longer regions of high similarity. However, alignment still remains computationally expensive, and offers a level of detail not always necessary in read classification.

A more sophisticated approach is index-based analysis, where a pre-computed index is constructed with sequences that are specific or important to each genome or group of interest. Then each query read is classified by the presence or absence of these pre-identified markers. The foremost examples of this form of read classification are the Kraken [18], [19] set of tools, as well as tools such as CLARK [20] and Centrifuge [21]. While the read classification process in index-based approaches can be extremely fast, there is substantial time and space overhead associated with the construction of the index.

The space, time and computational overhead associated with alignment- and index-based read classification has motivated the need for faster, more accurate, and lower overhead alternatives. Sketching has proven to be a viable solution instead of whole genome comparisons as it provides the level of accuracy required for less demanding tasks such as read classification, while substantially reducing overhead. Examples of this are MashMap [9] and MetaMaps [22], which use approximate similarity instead of exact alignment between regions of two sequences to perform alignment.

In this work, we critically evaluate several sketching and sampling methods that aim to reduce the computational overhead of read classification. We apply sketching, using MinHash- and minimizer-based approaches, as well as uniform sampling, to generate compact, approximate representations of potential source genomes for a given readset. We then classify reads against these representations, and demonstrate that we are able to classify, with a high degree of accuracy, reads from a microbial community and detect contaminants in real and simulated sequencing experiments.

2. Methods

In our methods, we consider several sketching and sampling approaches to generate reduced representations of the source genomes, a "screen" of the genomes. Query reads are then compared against this screen, with the read being classified to the element most similar in the screen **(Figure 1)**.

2.1 Determining sketch and sample size

The single biggest factor in such an approach is the size of the screen, meaning the fraction of the k-mers from the genomes that are stored. The ideal screen size will minimize the input storage requirement while being detailed enough to capture the specificity of each genome. To do this, three main factors must be considered: (1) the size of the genomes; (2) the read length and error rate of the reads we are classifying; and (3) the amount of similarity needed to correctly match a read and its true source genome. We refer to this as the number "target matches" or "shared hashes", which is the number of sketched or sampled k-mers a read and its source genome share. We formalize this using the following formula:

$$Sketch/Sample\ Size = \frac{(\#\ Target\ Matches) \times (Genome\ Size)}{(Read\ Length) \times (1 - Read\ Error\ Rate)^k} \qquad [\mathbf{EQ1}]$$

This formula allows us to sketch and sample at a rate where we expect to retain the target number of k-mers per read length of sequence in the original genome, adjusted for error. We adjust for error by computing the fraction of k-mers we expect to be affected by error at that error rate, and oversampling or oversketching to compensate for this. The resulting sketch or sample has, in expectation, the desired number of error-free hashes stored from each read length of sequence in the genome, and therefore the desired number of error-free shared hashes with a read drawn from the same region. This formula is used to determine the expected number of stored hashes in all our approaches, as the cost of generating similar sized sketches and samples is relatively equal across the approaches.

As shown in **EQ1**, the exact size of the screen depends on the experimental parameters. The compression factor is equal to the targeted number of k-mer matches per read divided by the read length, with an oversampling by a factor of $1/(1 - error)^k$ to correct for errors. This makes these approaches best suited for lower error, longer reads, as these require the fewest number of hashes across the genome, with shorter or higher error reads requiring larger screens. The number of target matches also determines screen size, but as we will see in the results section, high accuracy is possible with small screens that are much smaller than the original data, especially for low error rate and longer reads.

During classification, read k-mers are only compared against the stored k-mers in the screen. This can be done efficiently using hash table-based sets: the space required for representing the reference genomes will typically be only a few percent of the total sequence length, and the number of hash table lookups will also be substantially reduced. Once computed, a screen can be re-used for all future runs, amortizing the cost of screen generation across all future uses. This approach also reduces the cost of updating the set of potential source genomes; instead of rebuilding the whole index, sketches or samples can easily be added or removed, with the rest of the screen left unchanged.

2.2 Overview of sketching and sampling approaches

In this work, we consider several existing sketching and sampling techniques, along with a novel clustering & comparison technique we describe below. A reference implementation of these approaches can be found at https://github.com/arun96/sketching.

Uniform. In this approach, k-mers are uniformly extracted across the genome to reach the desired sample size. The chief benefit of this approach is simplicity, including a guarantee on the maximum distance

between k-mers in our screen, which is generally not guaranteed for alternative approaches. This also guarantees that each read will have a highly predictable amount of overlap with the sampled version of its source genome, though error in these samples can obscure the detection of this overlap. Computationally, uniform sampling is the simplest of the approaches; For a genome of size n, a sample of size s can be generated in $O(n)$ time and stored in $O(s)$ space.

MinHash. This sketching technique enables quickly estimating the similarity between two input sequences by computing the Jaccard coefficient of the selected k-mers extracted from one sequence compared to those selected in a second sequence. There are several widely used methods to generate a MinHash sketch, such as using multiple hash functions or a partitioning of the space of possible k-mers. For our analysis, we use a single hash function, and select the s smallest hash values returned, such as is used in Mash. As hashing is simply a permutation of the input values, this effectively generates a random sampling of s k-mers to be used as the representation of the original genome. In terms of computation, MinHash requires all k-mers to be hashed while maintaining a list of minimal hashes, which can be done using a fixed-size max-heap. A sketch of size s from a genome of size n requires(n) time to hash each k-mer. Then the s smallest hash values can then be identified using a max-heap of size s for a total runtime of $O(n \log s)$.

Weighted MinHash (WMH). This enhances a basic MinHash with weights for each k-mer representing a measure of the k-mer's "importance", with more highly weighted k-mers indicating a greater level of confidence in a match. Weights are typically based on the number of times that an element occurs, or on a predetermined scoring scheme. In our reference implementation, the weight is a measure of "uniqueness"; we compute the weight of a k-mer as the total number of genomes in our screen minus the number of genomes the k-mer is found in. Unique k-mers occurring in a single genome are weighted the highest as these are strong candidates for precisely identifying the source of a read. This is especially useful when considering reads that share the same number of k-mers with multiple potential genomes; having a shared highly weighted k-mer with one of these potential genomes can help determine the source with more accuracy than a random tie break. This approach can be further extended to add a "multiplier" into weighted MinHash, where unique k-mers have their computed weight multiplied by some multiplier M (M=1 in regular weighted MinHash), allowing these highly informative k-mers to play an even larger role in determining the similarity between a read and its genome.

The addition of weight is computationally expensive; for a sketch of size s we must also store $O(s)$ weight values, which effectively doubles the space requirement compared to the basic MinHash approach. There is also the added computation of determining the weights. In the baseline approach mentioned above, we count the number of genomes each k-mer in the sketch is present in; this can take $O(ns)$ time, though this can be minimized by using existing optimized implementations, such as KMC3 [23] or Jellyfish [24].

Order MinHash (OMH). As an alternative to WMH, the consideration of the order of the retained minimal hashes can also filter out spurious matches and prioritize more likely sources for a query sequence. First presented as a method to improve estimation of the edit distance [25], an Order MinHash sketch stores the selected n hashes in ordered sublists of L hashes, in the same order as they occur in the genome, with n/L lists making up the sketch. When two sketches are compared using Order MinHash, the algorithm checks which hashes are shared, along with if the shared hashes are in the correct order relative to each other. This method of comparing two sketches means that two sequences that contain the same k-mers but are rearranged versions of each other will have low similarity scores, while non-ordered MinHash would report high similarity. This approach is also more robust to sequencing errors than selecting a single long k-mer spanning the same distance. As with Weighted MinHash, Order MinHash incurs additional computational overhead. During sketch generation, we must store both the k-mer's hash and its position in the original sequence, in order to construct the ordered sublists. In addition, during comparisons between two sets of hashes, the relative order of any shared hashes must also be considered, meaning simple set comparison is no longer enough.

Minimizer. Minimizers were originally proposed as a sequence compression method [26], but have become popular in genomics due to their ability to succinctly represent large sequences. In the most widely used form of a windowed-minimizer, the algorithm slides a window of size x over the sequence,

and the k-mer with the smallest hash in that window is retained as the minimizer. This is repeated across the entire sequence, and the set of unique minimizers is used as the representation of the full sequence. Unlike MinHash, window-minimizers provide some guarantee on the distance between the retained k-mers in our screen, as this distance is bounded above by twice the window size. For our minimizer-based approach, the window size is computed as the size of the genome divided by the desired number of k-mers per genome, multiplied by a fixed multiplier of two. The reasoning behind this multiplier is quite simple: since the distance between minimizers is uniformly distributed between 0 and the window size, we expect two minimizers per window. Consequently, to find s minimizers for a genome of size n, we simply double the window size. This keeps the generated sketch and sample sizes relatively even across the approaches. Computation of a minimizer sketch of a genome of size n using window size w can be done in $O(nw)$, by choosing the minimum of the w hashes in each of the $O(n)$ windows.

2.3 Clustering sketches

The approaches above create a separate screen for each source genome, where all input reads are compared against all elements of this screen. While this reduces the number of comparisons necessary to classify a read compared to traditional approaches, they still perform a large number of unnecessary comparisons with genomes with low similarity with the query read. To tackle this, we propose a sketch clustering approach to limit the number of comparisons with less-relevant genomes.

To do this, the algorithm first computes a hierarchical clustering of the individual sketches of the input genomes. This groups together similar genomes, whose selected k-mers (and derivative reads) are more likely to be similar. The generated screen is then used to populate the resulting clustering tree: each genome's reduced representation appears in the leaves of the tree, and the reduced representations are combined within internal nodes of the tree, until the root of the tree contains all the elements of the original screen. To limit the overhead of this approach to be comparable to those presented above, the algorithm randomly downsamples the original elements of the screen as the algorithm constructs the tree. This downsampling can be done as a constant factor or by a factor proportional to the height of the tree, depending on the desired total size of the sketch tree. A random downsample could distort set comparison metrics such as Jaccard coefficient estimation, but is less of a concern for this analysis since a single k-mer match is sufficient to explore the children of an internal node. However, extreme downsampling can increase the number of misclassified or unclassified reads as they can remove all shared k-mers.

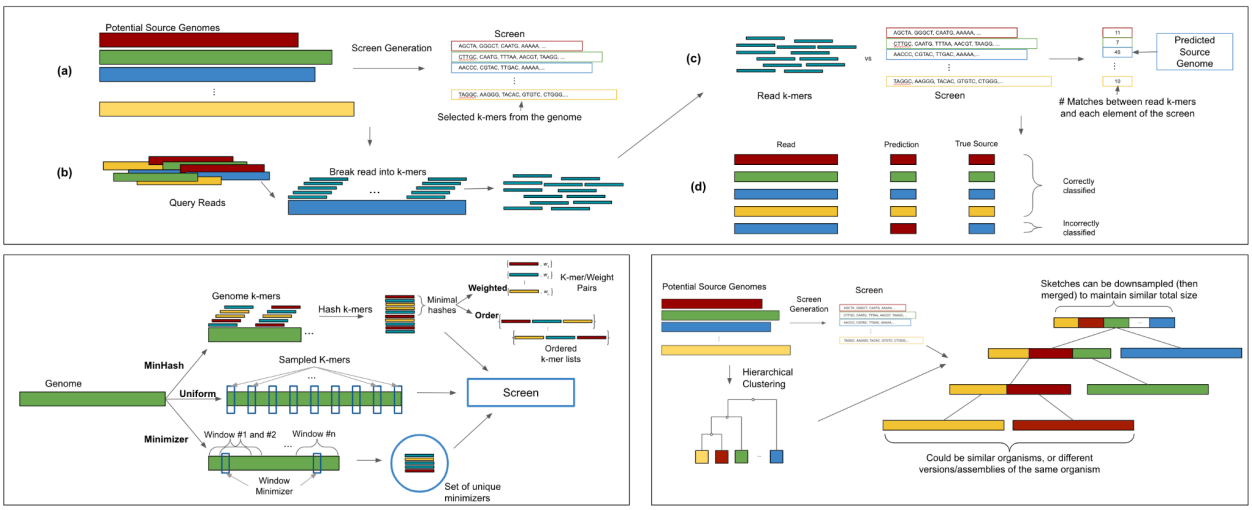


Figure 1. Overview of sketching and sampling methods. (Top) The screen is generated using the desired sketching or sampling approach from potential input genomes, read k-mers are compared against the screen, with the element of the screen most similar to the read predicted as its source. **(Bottom Left)** The different sketching and sampling approaches used to generate a screen. **(Bottom Right)** Sketch clustering: input genomes are clustered, and the generated screen is arranged to match this clustering, with reads compared to the root and then down the tree.

Read classification is then performed by starting at the root, comparing the input read to the stored representations at each of the children of the root, and then descending into the child with which the read is most similar. This process repeats until the algorithm arrives at a leaf, which is the genome predicted to be the source of the read. If, at any point, no child is found with similarity above the required threshold to the query read, the read is left unclassified. This approach quickly prunes genomes with low similarity to the input read, and focuses on the genomes that are likely to be the source of the read, especially genomes that are either from different but similar organisms, or different assemblies of the same genome.

Hierarchical clustering is computationally expensive, requiring $O(n^3)$ time and $\Omega(n^2)$ space to cluster n elements. This is in addition to the cost of sketching genomes to input into the clustering and the cost of the initial screen generation. However, the reduction in the number of comparisons performed per read can compensate for this cost. Given a screen of size x hashes, and the same screen in this clustered form with downsampling factor f, a read will require x comparisons against the original screen but $O(x/f)$ comparisons with the clustered screen. Therefore, with reasonable choice of f, we can reduce the computation per read classified and yield an overall amortized time savings.

3. Results

3.1 Metagenomics Classification

Our main experimental results are based on the widely used Culturable Genome Reference (CGR) community of high quality microbial genomes sequenced from the human gut [27]. From this community, we selected all genomes that were available on RefSeq (as of 12/10/20), giving us 1,310 genomes for our reference database. This community contains several clusters of highly similar genomes that make read classification more difficult (**Table 1**). This difficulty is especially true for approaches that work with reduced representations of the original genomes; unless the differences between these similar genomes are specifically captured, there will be no information available to distinguish between them.

As an example of a simpler community, we also analyze a union of the ZymoBIOMICS Microbial Community Standards (ZYMO) and MBARC-26 [28] reference communities, which, when combined, contain 34 microbial and fungal genomes. This community is much easier to classify within, as the genomes are relatively dissimilar (**Table 1**), and provides a baseline from which to interpret our results.

For our classification analysis, by default we use simulated PacBio HiFi-like reads that are 10Kb long with 1% error rates, with errors uniformly introduced. We simulate 10x coverage of each genome, yielding 4.8M reads for the CGR dataset, and 165K reads for the ZYMO+MBARC-26 dataset. For the classification, we use screen sizes that target 100 shared hashes with each of these reads, or on average one shared hash every 100bp based on **EQ1**; this generates screens that contain approximately 2% of the k-mers present in the original genomes. Results across a range of read lengths, error rates and screen sizes, as well as other experimental parameters, are reported later in this section, and presented in **Figure 2**. Reference implementations and analysis scripts, details on data availability, and scripts to benchmark existing tools can be found at https://github.com/arun96/sketching.

Community	Total Sequence	Number of Organisms	Number of genomes with >X% Mash similarity to another genome in the community					
			X = 5%	X = 25%	X = 50%	X = 70%	X = 90%	X = 95%
CGR	4.85GBp	1310	1218	1189	990	563	315	270
ZYMO + MBARC-26	170MBp	34	2	0	0	0	0	0

Table 1: Overview of the microbial communities.

Classification Experiments

In microbial classification experiments, reads are drawn from a microbial community, and compared against a screen generated from a reference database of known genomes. Under idealized conditions, the database will contain reference genomes from all members of the community, although in practice the community may contain novel species or strains that are not yet characterized leading to poor matches or no matches at all. For simplicity, our simulated reads are drawn from the reference database collection,

and reads are then classified against the screen of all genomes. Accuracy is then measured as the fraction of reads correctly classified as being from the true source genome.

For the human gut microbe community, at our default experimental parameters, we see that all our sketching and sampling approaches achieve approximately 71-75% accuracy (**Figure 2**).We observe that around half of the genomes have classification accuracy over 90%, with the overall accuracy lowered by genomes that have high similarity with other members of the community. For example, we find that of the genomes with less than 50% classification accuracy, 90% have another member of the community with which they are at least 70% similar. This implies that for these genomes, we expect 70% of their reads to be very similar to at least one other genome, greatly increasing the chances of each read being misclassified. This is magnified when considering groups of genomes with >95% similarity, a level of similarity high enough to consider the genomes to be of the same species [29]. Read classification between such genomes, regardless of which approach is used, often becomes random tie-breaking.

Just how disruptive highly similar genomes are to classification accuracy is visible when classifying reads from the simpler ZYMO+MBARC-26. Here, just two of the 34 members have Mash similarity >1.5% with each other, with those two members having a similarity of just 8% (**Table 1**). In our experiments, these two genomes provide the vast majority of misclassified reads, across read lengths and error rates. Overall, with a simpler community like this, any of these approaches achieve >99% classification accuracy, even with much shorter reads and significantly higher error rates.

Effect of experimental parameters on read classification

Read Length. We see increases in performance as read lengths get longer (**Figure 2a**), as we have more opportunities for the screen k-mers to match error-free k-mers in the read. Read length also affects the size of the screen, as longer reads mean smaller screens are necessary to achieve the desired number of shared hashes between a read and its source (**Figure 2c**). Conversely, with shorter reads, the screen sizes must be proportionally larger to maintain the similar levels of accuracy.

Error Rates.We see decreases in accuracy at high error rates (**Figure 2a**), as fewer k-mers remain unaffected by error, accompanied by sharp increases in screen size. With an error rate of 1% (as found in PacBio HiFi reads), we estimate that 81% of 21-mers will remain error free, while at an error rate of 5% (as is found in Oxford Nanopore reads) just 34% of the 21-mers will remain error free. This is even more pronounced at error rates close to 10% (as is found in CLR PacBio data and older Oxford Nanopore reads), where just 10% of the 21-mers can be expected to be unaffected by error. As our approach adjusts screen size to compensate for error rate, this results in extremely large screen sizes to compensate for high error rate (**Figure 2c**).

Target number of shared hashes. The number of target matches determines how densely the reference genomes are represented, and therefore the size of the screen. Low numbers of target matches result in reads being misclassified or unclassified, as there will not be enough detectable similarity with the source genome. However, there is also a plateau in performance as we increase the target number of shared hashes, as some sets of genomes differ only in a small number of k-mers, and a sketching or sampling approach must draw from exactly those places in order to distinguish between them. We see steady increases in performance when increasing the target number of shared hashes up to 3 shared hashes every 200bp, but we see minimal gains beyond this (**Figure 2b**).

K. For k set to at least 20 (**Figure 2d**), we find increasing k can result in minimal increases in performance, yet a larger increase in screen size. This is because longer k-mers have a higher chance of being affected by errors, so larger samples/sketches are necessary to ensure a robust number of error free k-mers remain. This was highly pronounced in our results, e.g. the step up from k=30 to k=50 came with only a 1.3% increase in performance but a 22% larger screen. As we saw similar performance between $20 \leq k \leq 50$, we used $k = 21$ across our other experiments, as it provided the specificity necessary while keeping screen sizes small.

Weight. The addition of weight to traditional MinHash results in a slight increase in performance across read lengths and error rates. This is expected, as the discriminative k-mers now contribute more to the score and help break ties. Including a multiplier has a similar effect and more heavily weighting

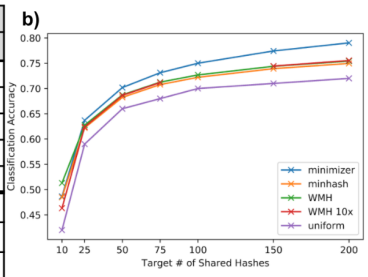
unique k-mers results in correctly breaking even more ties, resulting in another slight increase in performance. The addition of weight saw a 0.5% increase in classification accuracy, with the inclusion of a multiplier of 5 or 10 seeing a further 0.1% increase in performance (**Figure 2e**).

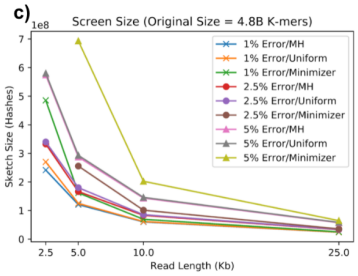
Order. Including order in a MinHash approach has a minimal impact on classification accuracy (**Figure 2e**). Order MinHash was initially proposed as a metric for estimating edit distance, and is most beneficial when determining the similarity between rearranged strings that cannot be distinguished by an unordered MinHash. However, with read classification from this large set of microbial genomes, such rearrangements are not common.

Cluster Downsampling Rate. Using MinHash screens, we compared the accuracy across three approaches to clustering: (1) screens downsampled by a constant factor; (2) screens downsampled based on their height in the sketch tree; and (3) screens that are not downsampled at all. We find that constant factor downsampling approaches, with factors 2 and 4, maintain a good degree of accuracy (71% and 70% respectively, compared to the 72.8% accuracy with MinHash), while keeping the number of comparisons similar to or less than the original MinHash approach. Height-based downsampling approach results in a sharp drop in accuracy (62%), as the screens near the root of the tree are downsampled to the point where discriminative k-mers are lost.

a)

		Read Lengths (Kb)					
		0.5	**2.5**	**5.0**	**10.0**	**25.0**	**50.0**
MinHash	Read Error (%) **0.0**	57.1	65.0	69.0	72.8	77.5	80.8
	1.0	56.6	64.6	68.9	72.8	77.6	80.8
	2.5	55.1	64.1	68.6	72.7	77.8	80.8
	5.0	54.5	62.7	67.8	72.4	77.7	80.7
	10.0	52.5	62.0	66.9	70.8	77.2	80.6
Minimizer	Read Error (%) **0.0**	60.1	70.0	74.0	77.4	82.1	85.8
	1.0	58.7	69.2	72.7	75.6	79.8	83.3
	2.5	57.8	69.0	72.1	74.2	78.0	81.2
	5.0	57.1	68.9	72.3	74.0	76.7	79.8
	10.0	55.6	67.4	71.9	73.6	78.6	80.1
Uniform	Read Error (%) **0.0**	55.3	63.9	68.1	71.1	74.7	77.9
	1.0	54.8	63.5	67.6	70.9	73.7	75.9
	2.5	53.3	63.1	67.1	70.5	72.9	74.8
	5.0	52.8	62.4	66.4	69.5	72.6	73.7
	10.0	52.1	61.9	66.0	68.9	72.1	72.2





d)

K	10	15	20	30	35	40	50
Accuracy (%)	28.5	48.0	72.6	72.9	73.3	73.6	73.9
Sketch Size (10M Hashes)	5.4	5.7	6.0	6.6	7.0	7.3	8.1

e)

	MH	WMH w/ Multiplier					OMH
		1	2	5	10	15	
Accuracy (%)	72.2	72.7	72.7	72.8	72.8	72.8	72.4

Figure 2. Key results across parameters on the CGR dataset. (a) Results across a range of read lengths and error rates, **(b)** the effect on accuracy of increasing the target number of shared hashes, **(c)** The impact read length and error have on sketch size across our approaches, **(d)** The performance of a MinHash approach across a range of k values, and **(e)** the impact of incorporating weights, a range of weight multipliers and order into MinHash.

3.2 Host Contaminant Detection

The goal of the contaminant detection and classification is, for a given read set, to distinguish between reads that come from organisms or sequences of interest, and reads that are from potential contaminants. For our experiments, we considered human reads simulated from GRCh38 [30] and mixed with contaminant reads drawn from a selected microbial community, and classified against a screen containing both the human and microbial genomes. The HiFi-like sequence reads were simulated as above using 10kbp reads at 10x coverage with 1% error. Accuracy is measured as the fraction of human and microbial reads correctly identified as being of interest or as being a contaminant, while also measuring the fraction of contaminant reads that are correctly classified to their source genome.

When using the CGR community as the source of contaminant reads, we find all the sketching and sampling approaches to be successful at distinguishing between microbial and human reads. Across all approaches, >99% of all human reads are correctly distinguished from microbial reads and classified to the chromosome they are drawn from. We also observe that very few microbial reads are misclassified as human, with over 99% correctly identified as being contaminants. This is not unexpected; human and microbial genomes are quite dissimilar, and therefore the sketches or samples of the sequences will also be dissimilar, making read classification successful in nearly all cases. However, it is worth highlighting

that we are able to distinguish between these sequences while storing just 2% of the original k-mers.

After distinguishing between human reads and contaminants, we then attempt to classify the contaminant reads to the exact source genome. The results match what we present in the classification experiments; approximately 75% of the contaminant reads are mapped to the correct genome, with the misclassified reads coming from the genomes discussed in the previous section. We also classify human reads to the chromosome they are drawn from, and are able to do this with >95% accuracy.

3.3 Comparison to existing tools

To evaluate the performance of the sketching and sampling approaches, we also tested several widely used approaches for read classification on the same dataset and experimental settings. As before, accuracy in classification experiments is measured as the number of reads correctly classified as from the microbial genomes they were drawn from, and accuracy in contaminant detection experiments is measured as the number of human and microbial reads identified as human or from any microbial genome respectively.

Alignment-based. To test the effectiveness of alignment-based approaches to read classification, we test **Minimap2** [16] and **Winnowmap** [31]. Minimap2 uses query minimizers as seeds for the alignment, while Winnowmap2 adds a preprocessing step to downweight repetitive minimizers to reduce the chance of them being selected. In both approaches, we align our read sets against the genomes of the selected community, and calculate the predicted source of the read as the sequence to which it is mapped. For microbial classification, we find that both these tools perform slightly better than our MinHash and minimizer based approaches. Compared to an accuracy of 77% and 79% in our MinHash and minimizer approaches with two shared hashes every 100bp, Minimap2 and WinnowMap both achieve an accuracy of 81% (**Table 2**). Both alignment approaches achieve low accuracy on the same genomes that our sketching and sampling approaches struggle on; namely, genomes with high-similarity relatives in the community. These misclassifications are amplified at higher error rates, where the ability to distinguish between similar genomes is reduced. For contaminant detection, both tools are able to correctly distinguish 99.5% of the human and contaminant reads at a 1% sequencing error rate, and over 98% at higher error rates.

Index-based. Kraken2 [19] and **Centrifuge** [21] use a preprocessed index of shared k-mers or compressed genomes respectively to determine the source of a query sequence. Each k-mer in the query sequence is classified to an element in the index, and we determine the source of the query as the element to which a plurality of k-mers are assigned.When using both a pre-built RefSeq database and a custom database built over our test community, we find that genome level identification is difficult between the highly similar members (**Table 2**). We observe large numbers of misclassifications between reads from these similar genomes, as well as classifying many of these reads only to higher taxa, and not to one of the specific genomes. For genomes without similar members in the community, the majority of their reads are correctly classified, giving Kraken2 and Centrifuge an overall classification accuracy of 72% with 1% error reads. At higher error rates, this performance drops sharply, with more reads left unclassified due to a lack of matched k-mers to the generated index. When distinguishing between human and microbial reads, both methods are able to correctly identify >95% of the reads, even at high error rates.

CLARK [20] uses a pre-compiled list of discriminative k-mers for the community it is indexing, and performs classification based on query similarity to this list. While there are still misclassifications and unclassified reads at rates comparable to other tools, CLARK's use of discriminative k-mers slightly reduces the impact of highly similar genomes in the community, allowing it to identify the few differences between them, achieving a classification accuracy of 73.5% with 1% error reads, and making it more resilient against misclassification at higher error rates (**Table 2**). For contaminant detection, CLARK is also able to distinguish over 97% of both human and microbial reads across a range of error rates.

Sketching-based. MashMap [9] computes alignments by estimating k-mer based Jaccard similarity between query sequences with MinHash sketches. We find that MashMap performs worse than classic alignment-based approaches, and similarly to our MinHash approaches, with 74.5% classification on 1% error reads and steady decreases at higher error rates. Alignment boundaries in MashMap are determined

through the Jaccard similarities of sketches. As a result, just as in the MinHash approach, it is susceptible to misclassifications between highly similar genomes. For contaminant detection**,** MashMap is able to distinguish more than 96% of the human and microbial reads, even at higher error rates (**Table 2**).

	Classification Accuracy (%) at Error Rate X			Contaminant Detection: % of Reads identified at Error Rate X					
	X = 1%	X = 5%	X = 10%	X = 1%		X = 5%		X = 10%	
				Human	Contaminant	Human	Contaminant	Human	Contaminant
Minimap2	81.3	77.9	74.1	99.5	99.5	99.2	99.0	98.5	97.9
WinnowMap	81.3	77.9	74.1	99.5	99.5	99.2	99.0	98.5	98.0
Kraken2 *(RefSeq DB)*	72.0	66.0	58.0	99.3	99.2	98.8	98.1	97.2	96.5
Kraken2 *(Custom DB)*	72.2	66.1	58.0	99.3	99.3	98.8	98.1	97.2	96.5
Centrifuge *(RefSeq DB)*	72.2	67.5	62.1	99.3	99.2	98.8	98.4	97.8	97.2
Centrifuge *(Custom DB)*	72.4	67.5	62.1	99.3	99.3	98.8	98.4	97.8	97.2
CLARK	73.5	68.5	65.4	99.5	99.5	99.1	98.9	99.0	98.1
MashMap	74.5	70.8	67.3	99.5	99.4	98.9	98.1	97.7	96.6

Table 2: Performance of existing tools. We find that alignment-based approaches are better at genome-level read classification, with the index- and sketching-based approaches less able to distinguish between highly similar genomes. All tools perform similarly in contaminant detection, with this task less affected by higher error rates.

3.4 Analysis of genuine metagenomics sequencing data

To test the accuracy of our approaches on real sequencing data, we analyzed PacBio HiFi reads from the Human Gut Microbiome Pooled Standards [32] with the CGR community database. For this analysis, we used one vegan and one omnivore dataset, with 1.79M and 1.90M reads respectively of length ~10Kb and median quality of ~Q40. We first aligned these reads using Minimap2, and found 80% of the reads aligned to the CGR community at all, with approximately 54% of the reads having alignment length >5Kb. This threshold drastically reduces the number of classified reads that are mapped to multiple genomes; for these reads, we take the sequence with the longest alignment as the predicted source.

We then classify these reads using our sketching and sampling approaches against a generated screen of the CGR community, built for 10KB, 1% error reads and 100 shared matches per read. We consider a read to be classified if it has at least 5 shared hashes with a genome, and unclassified if it does not share 5 hashes with any genome. With this threshold, approximately 60% of all reads are classified in each of the approaches, with approximately 25% of the classified reads tied between multiple sources (**Table 3**).

We compare these classification results against the alignments generated with Minimap2. Our classification results agree with approximately 60% of the reads classified by Minimap2 with no minimum alignment length, and approximately 80% of those classified with a minimum alignment length of 5Kb (**Table 3**). This increase in consistency is expected, as this threshold limits Minimap2 classification to reads that share a significant amount of sequence with potential source genomes, and are therefore more likely to share a significant amount of similarity with elements in the screen. Without this threshold, some reads are classified based just on small regions of alignment, and any similarity with elements in the screen must come from shared hashes drawn from these short aligned regions; such reads are likely to be misclassified between multiple low scoring genomes, or not classified at all. Investigating the classifications that still do not agree with Minimap2, we see that almost 90% of these reads are instead classified to genomes that are >95% similar to Minimap2's predicted source.

We also classified these reads using Kraken2, with the predicted source of a read being the sequence to which a plurality of its k-mers are assigned. We find that approximately 77% of reads are classified by Kraken2, but approximately 40% of classified reads are only classified to a lowest common ancestor (LCA) instead of a single genome; we count these reads as classified to multiple genomes. We find that Kraken2's classification results agree with approximately 56% of the reads classified by Minimap2 with no alignment threshold, and approximately 82% of the reads classified by Minimap2 with a 5Kb minimum alignment length (**Table 3**). As expected, Kraken2 classified very few reads with <5Kb Minimap2 alignments, leaving those reads unclassified or only classified to a LCA; the majority of

Kraken2's classifications being reads that have significant similarity to a single genome. The remaining reads that do not agree with Minimap2 are either classified only to a LCA, or classified to genomes that are highly similar to Minimap2's predicted source. As with the sketching and sampling approaches, we find that almost 90% of these reads are classified to genomes that are >95% similar genome-wide. The remaining mis-classified reads originate from localized regions of the target genome showing high similarity to other genomes.

		Total number of reads classified	Number of reads classified to multiple genomes	Number of reads with same prediction as Minimap2	
				No Threshold	>5Kb Alignments
Vegan (1.79M Reads)	Minimap2 (No threshold)	1,490,713	485,233	N/A	
	Minimap2 (>5Kb threshold)	966,305	20,365	N/A	
	Kraken2	1,399,341	562,744	827,572	808,633
	MinHash	1,032,396	261,732	890,766	773,055
	Minimizer	1,029,996	262,514	891,388	775,441
	Uniform	1,021,555	264,867	883,465	764,886
Omnivore (1.9M Reads)	Minimap2 (No threshold)	1,530,795	490,501	N/A	
	Minimap2 (>5Kb threshold)	1,036,905	21,550	N/A	
	Kraken2	1,442,671	578,113	855,670	835,888
	MinHash	1,111,102	275,344	915,634	829,802
	Minimizer	1,105,356	278,654	916,998	830,004
	Uniform	1,098,244	281,745	911,554	820,942

Table 3: Performance on real sequencing data. Comparison of the classification of our sketching and sampling approaches against Kraken2 and Minimap2 classifications across two PacBio HiFi Gut Microbiome datasets.

4. Discussion

In this work, we presented and analyzed a range of sketching and sampling approaches for read classification, designed to reduce the space and time overhead for accurate classification across large collections of genomes. Overall, we find sampling and sketching are highly effective compared to index-based approaches, and are within a few percent accuracy of alignment-based approaches. Alignment-based approaches have the advantage that they can assess the entire read sequence, although this increases runtime. Minimizers generally lead to improved accuracy over MinHash-based approaches, chiefly because there is a stronger guarantee on the distance between selected kmers. Among MinHash-based techniques, weighted MinHash enabled modest but measurable improvements while Ordered MinHash enabled minimal performance gains. All approaches correctly distinguished reads from dissimilar genomes but struggled with the classification of reads from highly similar genomes.

Sketching and sampling approaches are able to perform well, however there are still scenarios where these approaches are challenged. The current methods are best suited for longer, low-error reads, and incur a higher footprint and decreased performance when classifying shorter, higher error rate reads. Consequently, a major need for future work is the continued development of sketching and sampling techniques better suited for high error rate environments. This includes the use of gap k-mers [33] to increase error tolerance, or the use of more auxiliary information, such as pre-computed indexes of unique k-mers [34] or augmented MinHash or minimizer-based methods [35], to distinguish similar sequences.

5. Acknowledgements

We would like to thank Benjamin Langmead, Daniel Baker and Bohan Ni for their help. This work was supported in part by National Science Foundation (NSF) grants DBI-1627442, IOS-1732253, and IOS-1758800, National Institutes of Health (NIH) grant U01CA253481, and the Mark Foundation for Cancer Research (19-033-ASP) to M.C.S. This work utilized the computational resources of the Maryland Advanced Research Computing Center (https://www.marcc.jhu.edu/).

References

[1] C. Quince, A. W. Walker, J. T. Simpson, N. J. Loman, and N. Segata, "Shotgun metagenomics, from

sampling to analysis," *Nat. Biotechnol.*, vol. 35, no. 9, pp. 833–844, Sep. 2017.
[2] Handelsman Jo, "Metagenomics: Application of Genomics to Uncultured Microorganisms," *Microbiol. Mol. Biol. Rev.*, vol. 68, no. 4, pp. 669–685, Dec. 2004.
[3] F. P. Breitwieser, J. Lu, and S. L. Salzberg, "A review of methods and databases for metagenomic classification and assembly," *Brief. Bioinform.*, vol. 20, no. 4, pp. 1125–1136, Jul. 2019.
[4] W. P. M. Rowe, "When the levee breaks: a practical guide to sketching algorithms for processing the flood of genomic data," *Genome Biol.*, vol. 20, no. 1, p. 199, Sep. 2019.
[5] G. Cormode, "Data sketching," *Commun. ACM*, vol. 60, no. 9, pp. 48–55, Aug. 2017.
[6] A. Z. Broder, "On the resemblance and containment of documents," in *Proceedings. Compression and Complexity of SEQUENCES 1997 (Cat. No. 97TB100171)*, 1997, pp. 21–29.
[7] B. D. Ondov *et al.*, "Mash: fast genome and metagenome distance estimation using MinHash," *Genome Biol.*, vol. 17, no. 1, p. 132, Jun. 2016.
[8] B. D. Ondov *et al.*, "Mash Screen: high-throughput sequence containment estimation for genome discovery," *Genome Biol.*, vol. 20, no. 1, p. 232, Nov. 2019.
[9] C. Jain, A. Dilthey, S. Koren, S. Aluru, and A. M. Phillippy, "A Fast Approximate Algorithm for Mapping Long Reads to Large Reference Databases," *J. Comput. Biol.*, vol. 25, no. 7, pp. 766–779, Jul. 2018.
[10] K. Berlin, S. Koren, C.-S. Chin, J. P. Drake, J. M. Landolin, and A. M. Phillippy, "Assembling large genomes with single-molecule sequencing and locality-sensitive hashing," *Nat. Biotechnol.*, vol. 33, no. 6, pp. 623–630, Jun. 2015.
[11] B. Solomon and C. Kingsford, "Fast search of thousands of short-read sequencing experiments," *Nat. Biotechnol.*, vol. 34, no. 3, pp. 300–302, Mar. 2016.
[12] C. Sun, R. S. Harris, R. Chikhi, and P. Medvedev, "AllSome Sequence Bloom Trees," *J. Comput. Biol.*, vol. 25, no. 5, pp. 467–479, May 2018.
[13] D. N. Baker and B. Langmead, "Dashing: fast and accurate genomic distances with HyperLogLog," *Genome Biol.*, vol. 20, no. 1, p. 265, Dec. 2019.
[14] F. P. Breitwieser, D. N. Baker, and S. L. Salzberg, "KrakenUniq: confident and fast metagenomics classification using unique k-mer counts," *Genome Biol.*, vol. 19, no. 1, p. 198, Nov. 2018.
[15] G. Marçais, B. Solomon, R. Patro, and C. Kingsford, "Sketching and Sublinear Data Structures in Genomics," *Annual Review of Biomedical Data Science*, Jul. 2019, doi: 10.1146/annurev-biodatasci-072018-021156.
[16] H. Li, "Minimap2: pairwise alignment for nucleotide sequences," *Bioinformatics*, vol. 34, no. 18, pp. 3094–3100, Sep. 2018.
[17] C. Jain, A. Rhie, N. Hansen, S. Koren, and A. M. Phillippy, "A long read mapping method for highly repetitive reference sequences," *bioRxiv*, p. 2020.11.01.363887, Nov. 02, 2020. doi: 10.1101/2020.11.01.363887.
[18] D. E. Wood and S. L. Salzberg, "Kraken: ultrafast metagenomic sequence classification using exact alignments," *Genome Biol.*, vol. 15, no. 3, p. R46, Mar. 2014.
[19] D. E. Wood, J. Lu, and B. Langmead, "Improved metagenomic analysis with Kraken 2," *Genome Biol.*, vol. 20, no. 1, p. 257, Nov. 2019.
[20] R. Ounit, S. Wanamaker, T. J. Close, and S. Lonardi, "CLARK: fast and accurate classification of metagenomic and genomic sequences using discriminative k-mers," *BMC Genomics*, vol. 16, p. 236, Mar. 2015.
[21] D. Kim, L. Song, F. P. Breitwieser, and S. L. Salzberg, "Centrifuge: rapid and sensitive classification of metagenomic sequences," *Genome Res.*, vol. 26, no. 12, pp. 1721–1729, Dec. 2016.
[22] A. T. Dilthey, C. Jain, S. Koren, and A. M. Phillippy, "Strain-level metagenomic assignment and compositional estimation for long reads with MetaMaps," *Nat. Commun.*, vol. 10, no. 1, p. 3066, Jul. 2019.
[23] M. Kokot, M. Dlugosz, and S. Deorowicz, "KMC 3: counting and manipulating k-mer statistics," *Bioinformatics*, vol. 33, no. 17, pp. 2759–2761, Sep. 2017.
[24] G. Marcais and C. Kingsford, "Jellyfish: A fast k-mer counter," *Tutorialis e Manuais*, vol. 1, pp. 1–8,

2012.
[25] G. Marçais, D. DeBlasio, P. Pandey, and C. Kingsford, “Locality-sensitive hashing for the edit distance,” *Bioinformatics*, vol. 35, no. 14, pp. i127–i135, Jul. 2019.
[26] M. Roberts, W. Hayes, B. R. Hunt, S. M. Mount, and J. A. Yorke, “Reducing storage requirements for biological sequence comparison,” *Bioinformatics*, vol. 20, no. 18, pp. 3363–3369, Dec. 2004.
[27] Y. Zou *et al.*, “1,520 reference genomes from cultivated human gut bacteria enable functional microbiome analyses,” *Nat. Biotechnol.*, vol. 37, no. 2, pp. 179–185, Feb. 2019.
[28] E. Singer *et al.*, “Next generation sequencing data of a defined microbial mock community,” *Scientific Data*, vol. 3, no. 1. 2016. doi: 10.1038/sdata.2016.81.
[29] C. Jain, L. M. Rodriguez-R, A. M. Phillippy, K. T. Konstantinidis, and S. Aluru, “High throughput ANI analysis of 90K prokaryotic genomes reveals clear species boundaries,” *Nat. Commun.*, vol. 9, no. 1, p. 5114, Nov. 2018.
[30] V. A. Schneider *et al.*, “Evaluation of GRCh38 and de novo haploid genome assemblies demonstrates the enduring quality of the reference assembly,” *Genome Res.*, vol. 27, no. 5, pp. 849–864, May 2017.
[31] C. Jain *et al.*, “Weighted minimizer sampling improves long read mapping,” *Bioinformatics*, vol. 36, no. Suppl_1, pp. i111–i118, Jul. 2020.
[32] “Data release: Human microbiome samples demonstrate advances in HiFi-enabled metagenomic sequencing,” Aug. 02, 2021. https://www.pacb.com/blog/data-release-human-microbiome-samples-demonstrate-advances-in-hifi-enabled-metagenomic-sequencing/ (accessed Nov. 05, 2021).
[33] M. Ghandi, M. Mohammad-Noori, and M. A. Beer, “Robust k-mer frequency estimation using gapped k-mers,” *J. Math. Biol.*, vol. 69, no. 2, pp. 469–500, Aug. 2014.
[34] K. Zhu *et al.*, “Strain Level Microbial Detection and Quantification with Applications to Single Cell Metagenomics,” *bioRxiv*, p. 2020.06.12.149245, Oct. 04, 2020. doi: 10.1101/2020.06.12.149245.
[35] B. Ekim, B. Berger, and R. Chikhi, “Minimizer-space de Bruijn graphs: Whole-genome assembly of long reads in minutes on a personal computer,” *Cell Syst*, vol. 12, no. 10, pp. 958–968.e6, Oct. 2021.