MTSv: rapid alignment-based taxonomic classification and high-confidence metagenomic analysis

Tara N Furstenau [Corresp., 1], **Tsosie Schneider** [1], **Isaac Shaffer** [1], **Adam J Vazquez** [2], **Jason Sahl** [2], **Viacheslav Fofanov** [1, 2]

[1] School of Informatics, Computing, and Cyber Systems, Northern Arizona University, Flagstaff, Arizona, United States

[2] Pathogen and Microbiome Institute, Northern Arizona University, Flagstaff, Arizona, United States

Corresponding Author: Tara N Furstenau
Email address: tara.furstenau@nau.edu

As the size of reference sequence databases and high-throughput sequencing datasets continue to grow, it is becoming computationally infeasible to use traditional alignment to large genome databases for taxonomic classification of metagenomic reads. Exact matching approaches can rapidly assign taxonomy and summarize the composition of microbial communities, but they sacrifice accuracy and can lead to false positives. Full alignment tools provide higher confidence assignments and can assign sequences from genomes that diverge from reference sequences; however, full alignment tools are computationally intensive. To address this, we designed MTSv specifically for alignment-based taxonomic assignment in metagenomic analysis. This tool implements an FM-index assisted q-gram filter and SIMD accelerated Smith-Waterman algorithm to find alignments. However, unlike traditional aligners, MTSv will not attempt to make additional alignments to a TaxID once an alignment of sufficient quality has been found. This improves efficiency when many reference sequences are available per taxon. MTSv was designed to be flexible and can be modified to run on either memory or processor constrained systems. Although MTSv cannot compete with the speeds of exact k-mer matching approaches, it is reasonably fast and has higher precision than popular exact matching approaches. Because MTSv performs a full alignment it can classify reads even when the genomes share low similarity with reference sequences and provides a tool for high confidence pathogen detection with low off-target assignments to near neighbor species.

MTSv: Rapid alignment-based taxonomic classification and high-confidence metagenomic analysis.

Tara N. Furstenau[1*,] Tsosie Schneider[1], Isaac Shaffer[1], Adam Vazquez[2], Jason Sahl[2], and Viacheslav Y. Fofanov[1,2]

[1] School of Informatics, Computing, and Cyber Systems, Northern Arizona University, Flagstaff, Arizona, USA

[2] Pathogen and Microbiome Institute, Northern Arizona University, Flagstaff, Arizona, USA

Corresponding Author:
Tara N. Furstenau[1]
1295 Knoles Dr, Flagstaff, AZ 86011, USA
Email address: Tara.Furstenau@nau.edu

Abstract

As the size of reference sequence databases and high-throughput sequencing datasets continue to grow, it is becoming computationally infeasible to use traditional alignment to large genome databases for taxonomic classification of metagenomic reads. Exact matching approaches can rapidly assign taxonomy and summarize the composition of microbial communities, but they sacrifice accuracy and can lead to false positives. Full alignment tools provide higher confidence assignments and can assign sequences from genomes that diverge from reference sequences; however, full alignment tools are computationally intensive. To address this, we designed MTSv specifically for alignment-based taxonomic assignment in metagenomic analysis. This tool implements an FM-index assisted q-gram filter and SIMD accelerated Smith-Waterman algorithm to find alignments. However, unlike traditional aligners, MTSv will not attempt to make additional alignments to a TaxID once an alignment of sufficient quality has been found. This improves efficiency when many reference sequences are available per taxon. MTSv was designed to be flexible and can be modified to run on either memory or processor constrained systems. Although MTSv cannot compete with the speeds of exact k-mer matching approaches, it is reasonably fast and has higher precision than popular exact matching approaches. Because MTSv performs a full alignment it can classify reads even when the genomes share low similarity with reference sequences and provides a tool for high confidence pathogen detection with low off-target assignments to near neighbor species.

Introduction

36 Advancements in sequencing technologies over the last few decades have transformed the way
37 that we study complex microbial communities, which in turn has had a large impact on human
38 health, ecology, and industry. In the past, characterization of microbial communities was
39 restricted to sequencing the few individual members that could be successfully isolated and
40 cultured. Fortunately, due to dramatic improvements in the efficiency and cost-effectiveness of
41 sequencing technologies, we can now use shotgun metagenomic sequencing to study the
42 composition of entire microbial communities in a culture-independent manner.

43   Unlike traditional genomic sequencing, which uses DNA isolated from a single organism,
44 metagenomic sequencing is performed on the full complement of DNA captured from a mixture
45 of organisms. Many different methods can be used to characterize metagenomic samples
46 including reference-based and reference-free approaches. Here we focus on reference-based
47 taxonomic classifiers which assign sequencing reads to reference taxa in large sequence
48 databases. Initially, this was accomplished by repurposing existing traditional rapid alignment
49 tools like BLAST (Altschul et al., 1990; Huson et al., 2007; Camacho et al., 2009; Bazinet et al.,
50 2018)), and Bowtie2 (Langmead & Salzberg, 2012; Segata et al., 2012). Traditional read aligners
51 (e.g., Bowtie2) use various heuristics to rapidly search for good (although not necessarily
52 optimal) alignment locations within a single reference genome. Although there may be many
53 potential alignment locations within the genome, aligners will typically perform a prescribed
54 level of search effort and then report only the best alignment found (or in cases of a tie, randomly
55 pick one) (Langmead & Salzberg, 2012). This is appropriate for the application that they were
56 designed for, but it presents a problem when attempting to find alignments against large multi-
57 species reference databases. The problem is that the level of search effort may be expended
58 before all references are checked thereby missing equally good or better alignments to certain
59 taxa. Reference databases already vary widely in the number of sequences per taxa, and this
60 would further bias assignments to taxa that are overrepresented in the database.

61   BLAST and its many variations (Altschul et al., 1990; Morgulis et al., 2008; Camacho et
62 al., 2009) were designed to rapidly search large sequence databases and can report all alignments
63 over a certain quality. This reduces bias by performing a more exhaustive search but requires
64 many more computationally expensive alignments and is too slow to feasibly process large
65 datasets. Similarly, Bowtie2 provides a mode that searches for and reports all valid alignments,
66 but this increases the run time from minutes to days when there are many similar sequences in
67 the database. As databases and sequencing throughput continue to grow, such approaches will
68 become increasingly impractical.

69   Several approaches have been introduced to handle the growing size of reference
70 databases used for metagenomic analysis. Due to the higher levels of conservation of protein
71 compared to DNA, protein-based aligners like DIAMOND (Buchfink, Xie & Huson, 2015;
72 Buchfink, Reuter & Drost, 2021), MMSeqs2 (Steinegger & Söding, 2017; Mirdita et al., 2021),
73 and Kaiju (Menzel, Ng & Krogh, 2016) take advantage of non-redundant protein databases and
74 are orders of magnitude faster than BLAST (Ye et al., 2019). However, protein-based tools tend

75   to classify fewer reads (because the databases contain only coding regions) and they have higher
76   rates of misclassification (Ye et al., 2019). Centrifuge introduced a compression method which
77   reduces the number of redundant reference sequences that must be searched (Kim et al., 2016).
78   However, creating the compressed dataset is computationally intensive and thus difficult to
79   update and customize and users typically rely on versions maintained by the developers. The
80   lossy compression of the database also results in lower sensitivity (Ye et al., 2019). MetaPhlan
81   (Segata et al., 2012) uses Bowtie2 for alignments but restricts alignments to a set of marker
82   genes that are specific to certain taxa. This approach is very quick, but resolution is low and
83   unevenly distributed across taxa (Ye et al., 2019). Metalign (LaPierre et al., 2020) reduces the
84   number of reference genomes by filtering out sequences that are not likely to be in the dataset. It
85   uses a MinHash-based approach (Koslicki & Zabeti, 2019) (like Mash (Ondov et al., 2016) and
86   Sourmash (Brown & Irber, 2016; Pierce et al., 2019)) to identify reference genomes that share a
87   high percent of k-mers with sample sequences and only includes these genomes in the reference
88   set. With this approach, some species are likely to be left out of the filtered database by mistake
89   (e.g., species that diverge from the reference sequences and low abundance organisms) and
90   Metalign only includes a single representative genome per taxon which may reduce assignments.
91   Metalign uses Minimap2 (Li, 2018) to align reads to the filtered database. Minimap2, like other
92   traditional read aligners, was not designed to work with multi-species reference databases and as
93   we show here, it can introduce biases in assignments.
94   Newer methods, based on exact k-mer matching rather than full alignment, have been
95   introduced (e.g., Kraken2 (Wood and Salzberg 2014; Wood, Lu, and Langmead, 2019)) and
96   provide significant speed-ups. However, the heuristics can have large impacts on the accuracy of
97   the results. For example, the selection of a fixed k-mer size requires a fine balance between
98   sensitivity and specificity; longer k-mers will be more specific to a particular taxon but they are
99   less likely to have exact matches due to errors or genetic variation (Lu et al., 2017). These
100 tradeoffs work reasonably well in many but not all cases and can lead to low resolution of certain
101 taxa and/or missed assignments when sequences diverge from the references.
102   Many situations call for higher accuracy and higher confidence in taxonomic assignments
103 and there is a need for a tool that can perform full alignments of metagenomic reads at a
104 reasonable speed. These situations are generally those in which the presence or absence of
105 specific species needs to be established with high confidence. These cases include the detection
106 of organisms known to be associated with clinical syndromes (Chiu and Miller, 2019; Gu,
107 Miller, and Chiu, 2019), public health surveillance efforts (Miller et al., 2013), forensic evidence
108 (Alan and Sarah, 2012; Clarke et al., 2017; Robinson et al., 2021), or in bio threat (Bazinet et al.,
109 2018; Karlsson et al., 2013; Minogue et al., 2019) or biocontamination (De Filippis et al., 2021;
110 Wood et al., 2021) situations. In such cases, a tendency toward false positive assignments can be
111 highly detrimental as independent validation may be needed either through experimentation or
112 often by manually aligning reads to the reference genome of the suspected organism. In such
113 cases, it may be less time consuming to begin with alignment-based taxonomic assignments
114 rather than attempt to validate the results from faster, yet less accurate tools.

115   Here we introduce a new alignment-based taxonomic assignment algorithm called MTSv
116   (https://github.com/FofanovLab/mtsv_tools). Our goal in developing MTSv was to use efficient
117   data structures and algorithms, like those used by traditional read alignment tools (e.g., Bowtie2),
118   in a way that was specifically designed to handle metagenomic data. As described above,
119   traditional read aligners typically report only the top alignments for each read which causes them
120   to miss assignments to multiple taxa in mixed reference databases. However, when forced to
121   report all valid alignments, they are very slow because once they find a valid alignment for a
122   taxon, they continue the exhaustive search across all reference sequences for the same taxon (this
123   is especially slow when there are many similar reference sequences present in the database which
124   is often the case). For taxonomic assignment however, the location and number of potential
125   alignments is not relevant; it is sufficient to know that there is at least one high-quality alignment
126   to a given taxon. With this in mind, we designed our algorithm to bypass further alignment
127   attempts once an alignment of sufficient quality is found resulting in highly accurate assignments
128   with significant speedups. Using both simulated and real datasets, we demonstrate that our full
129   alignment approach (1) offers unbiased assignments and is faster than traditional read aligners
130   (2) drastically reduces false positive assignments compared to metagenomic tools that do not
131   provide full alignments (3) increases the number of assignments for taxa that share low similarity
132   with reference sequences, and (4) provides high confidence detection of pathogens while
133   avoiding off-target assignments to near-neighbor species.

Implementation

134

Building the Metagenomic FM-Index

135

136   MTSv implements a custom metagenomic index (MG-index) based on the FM-index data
137   structure (Ferragina and Manzini, 2000). To build the index, MTSv takes a multi-FASTA file
138   with headers indicating the Taxonomic ID (TaxID) for each reference sequence (Fig. 1). The
139   reference sequences can be sourced from any DNA sequence collection (i.e., GenBank, RefSeq,
140   etc.) and customized to fit the project. Because MTSv was designed to be highly parallelizable,
141   we recommend building multiple indices from smaller chunks of the reference sequences. This
142   helps reduce the memory requirements and allows for faster processing for both index building
143   and assignment. A tool for splitting up the reference files is included with MTSv.
144   During the index build, the sequences in the FASTA file are concatenated while
145   recording the location of sequence boundaries and the TaxID associated with each sequence. A
146   suffix array (Manber and Myers, 1993), Burrows-Wheeler Transform (BWT) (Burrows et al.,
147   1994) and FM-index (Ferragina and Manzini, 2000) are built from the concatenated sequences
148   using the Rust-Bio v0.39.1 package (Köster, 2016). The FM-index with the associated sequence
149   metadata constitutes the MG-index. One MG-index is created per FASTA file, and new indices
150   can be added as the reference collection grows without needing to rebuild any of the existing
151   indices.

Taxonomic Assignment

When classifying metagenomic sequences, MTSv uses the MG-Index to narrow down reference sequence locations that are most likely to contain a good alignment. The binning step begins by extracting overlapping substrings of the same size (seed-size) using a user-provided offset (seed-interval) from each read and its reverse complement (Fig. 2). It then uses the MG-Index to search for exact, un-gapped matches for each seed. The seed matches are sorted by location and grouped into candidate regions using specified windows. The number of hits per candidate is tallied and any candidate that does not meet the user-specified minimum number of seed hits is filtered out.

The filtering step is based on a q-gram algorithm which defines the minimum number of exact k-mer matches (from all $n - k + 1$ overlapping k-mers) that can be expected between an n-length read and a reference sequence with at most e mismatches. In the worst case, where all mismatches are evenly spaced across the alignment, the minimum number of matching k-mers is:

$$m = (n + 1) - k(e + 1) \; (m > 0 \text{ when } \left\lfloor \frac{n}{e + 1} \right\rfloor > k)$$ (Burkhardt et al., 1999; Rasmussen, Stoye,

and Myers, 2006; Reinert et al., 2015; Ukkonen, 1992). If only every l^{th} overlapping k-mer is used, the minimum number of matching k-mers is expected to be $\frac{m}{l}$. Within these parameters, MTSv is highly likely to find all alignments with up to e mismatches but will still find many alignments with more than e mismatches up to the user provided maximum edit distance cutoff.

After the filtering step, the remaining candidate positions are sorted by the number of seed hits in descending order so that the most promising regions are evaluated first. For each candidate region, MTSv extracts the corresponding range from the reference sequence and looks up the TaxID associated with the reference in the MG-index. If the current read has already been successfully aligned to the TaxID associated with the candidate region, no additional alignment is attempted, and the next candidate region is checked. Otherwise, a Single Instruction Multiple Data (SIMD)-accelerated Smith-Waterman alignment (Zhao et al., 2013) is performed between the extracted reference sequence and the query sequence (using a scoring of 1 for matches and -1 for mismatches, gap opening, and gap extension). If the alignment score is sufficiently high, there is one final check to determine if the edit distance is less than or equal to the user-specified edit distance cutoff. If the alignment is considered successful, then no further alignments are attempted for that read against the same TaxID. Skipping all additional alignments to a TaxID avoids many expensive operations and reduces computation time. As the number of reference genomes (and thus the size of a taxonomic bin) continues to rapidly expand, the advantage of this computational speed-up will only become more critical.

Reads are binned against each MG-Index producing separate output files. Each line of the output file contains the read name and a comma separated list of the TaxIDs it aligned to with the edit distance (e.g., "R1:1392=2,1396=3"). The edit distance provided for a TaxID is recorded from the first successful alignment (because no more are attempted afterward) and therefore it is not guaranteed to be the minimum. However, because the alignments start with the most promising candidates first, the edit distance is likely to be near the minimum for each TaxID.

191   Once all output files from the binning step are completed, the files are merged to combine read
192   assignments across multiple MG-Indices.

Materials & Methods

Comparison to existing software

195   We compared the performance of MTSv to two traditional read alignment tools (Bowtie2 v2.4.5
196   (Langmead & Salzberg, 2012) and Minimap2 v2.24 (Li, 2018)) and two commonly used
197   taxonomic assignment tools designed for metagenomic data (Centrifuge v1.0.4 (Kim et al.,
198   2016), and Kraken2 v2.1.2 (Wood, Lu, and Langmead, 2019)). Bowtie2 is a well-established
199   alignment tool that uses an FM-Index assisted seed-and-extend approach similar to MTSv.
200   Minimap2 was introduced more recently and appears to be a replacement for the BWA suite of
201   tools (Li, 2007). It uses a hash table index and a seed-chain-align approach that is faster than
202   Bowtie2 and BWA-MEM for Illumina short read alignments (Li, 2018). Centrifuge was designed
203   for taxonomic classification of metagenomic sequences and uses an FM-Index data structure.
204   However, unlike MTSv which performs a full alignment, Centrifuge only extends seed hits to
205   find the longest exact matching segments and does not account for gaps or mismatches.
206   Centrifuge assigns scores that favor taxa with longer exact matching segments and then assigns
207   the read to the taxon/taxa with the highest score. Kraken2 is a widely used ultrafast metagenomic
208   classification tool that uses an exact k-mer matching approach. Kraken2 builds a database that
209   maps fixed size k-mers from reference sequences to the lowest common ancestor of all
210   organisms that share the k-mer sequence. The k-mers extracted from sequencing reads can then
211   be looked up in a hash table and assigned to the corresponding TaxID. The final assignment is
212   chosen based on a score from a weighted path classification tree (Wood and Salzberg, 2014).
213   Although Centrifuge and Kraken2 (with the help of Bracken (Lu et al., 2017)) can use their
214   taxonomic assignments to estimate species abundance profiles, for comparison purposes, we
215   focused only on their performance on taxonomic assignment.

Building the reference indices

217   To compare the taxonomic assignment tools, reference indices were built from
218   bacterial/archaeal genome assemblies from RefSeq (O'Leary et al., 2016) at the Complete
219   Genome assembly level (121GB) downloaded on October 28th, 2019 (Release 96). TaxIDs were
220   assigned to each sequence using the accession to TaxID mapping file provided with the
221   download (all TaxIDs were labeled at the species level or higher) and there was a total of 4,785
222   unique TaxIDs (4,530 bacteria and 255 archaea). To reduce the memory requirements for
223   building and loading the MG-index, we split the reference file into 13 roughly 10GB chunks and
224   we built individual MG-indices for each chunk. Centrifuge and Kraken2 do not have an option to
225   split up their data structures, so we built them with a single sequence file using default
226   parameters (Kraken2 k-mer length was 35bp). For comparison purposes, the Centrifuge index

227   was built without reference sequence compression and to avoid variation due to masking
228   methods, low complexity regions were not masked for any of the tools.
229   Two additional sets of toy reference indices were built to compare MTSv and the
230   traditional alignment tools. The collection of references was chosen to highlight biases that can
231   be introduced when using traditional alignment tools for taxonomic assignment of metagenomic
232   reads. The first set contained reference sequences of closely related species in the *Bacillus* genus
233   and was dominated by *Bacillus anthracis* genomes (Table 1). The second was dominated by
234   *Staphylococcus aureus* but also contained a *Staphylococcus epidermidis* reference sequence. The
235   accession numbers for the sequences are available in Supplemental Table 1. MTSv, Bowtie2, and
236   Minimap2 indices were built for both reference sequence collections using default parameters.

Parameter settings

238   For Bowtie2, we used parameter settings that were optimized for taxonomic binning of
239   metagenomic reads (Jaillard et al., 2016); namely, a seed length of 20 (-L 20), seed error of 0 (-N
240   0), effort level of 15 consecutive seed extensions (-D 25), and end-to-end mode (--end-to-end –
241   min-score L,-0.6,-0.4). Default values were used for Minimap2. In some cases, the aligners were
242   also run in a way that forced them to find and report all alignments for each read. For Bowtie2
243   this was achieved using the "all" mode (-a) and for Minimap2 we used the "-N" option to retain
244   up to 5,000 alignments. Our goal was to ensure that the search effort extended enough to cover
245   alignments to all references in the index and the limit of 5,000 was chosen because it was much
246   larger than the number of sequences. Bowtie2 also has a -k method to set an upper bound on the
247   number of alignments but in our example, they were effectively the same. When comparing the
248   performance of these tools they were run on the 2.60GHz Intel Skylake Xeon nodes on Northern
249   Arizona University's Monsoon high-performance computing cluster with 120GB of memory
250   using 8 threads each.
251   Default parameters were used for Kraken2 and Centrifuge. Unless otherwise noted, the
252   default parameters were also used for MTSv: the seed size was set to 18bp, the space between
253   subsequent overlapping seeds was 15bp, the minimum number of seed matches to perform a full
254   alignment was set to 1, and the maximum edit distance cutoff was 13% or 20 mismatches for
255   150bp reads. For MTSv, only TaxIDs with the lowest reported edit distance were considered for
256   the final assignment. For performance comparison these tools were also run using 2.60GHz Intel
257   Skylake Xeon nodes using 80GB of memory and 8 threads each.

Performance evaluation using CAMI datasets.

259   To evaluate the performance of MTSv compared to Centrifuge and Kraken2, we used the human
260   microbiome datasets produced for the second Critical Assessment of Metagenome Interpretation
261   (CAMI2) challenge (Meyer et al., 2022). The simulated datasets were modelled on Human
262   Microbiome Project Consortium (Huttenhower et al., 2012) profiles from five different body
263   sites using NCBI RefSeq (O'Leary et al., 2016) complete genome sequences as genome sources
264   (Downloaded on July 8, 2017). We analyzed the short-read data (Illumina HiSeq2000 150bp

265 paired-end reads) from the gastrointestinal tract, oral cavity, airway, skin, and urogenital tract.
266 The ground truth source of each read was provided, and the performance metrics were calculated
267 using Amber v.2.0.2 (Meyer, et al. 2018). For both MTSv and Centrifuge, reads with
268 assignments to multiple taxa were rolled up to the lowest common ancestor using ETE3 v3.1.2
269 (Huerta-Cepas, Serra, and Bork, 2016) whereas Kraken2 does this by default.
270   Amber is a tool that helps standardize comparative assessment of metagenomic binning
271 tools by providing performance metrics evaluated against the ground truth assignments for each
272 read. Amber defines true positives (TP) as sequences that are assigned to the correct taxon at a
273 given taxonomic level and false positives (FP) as those assigned to an incorrect taxon. False
274 negatives (FN) include unassigned sequences and those that are not assigned at the current
275 taxonomic level (e.g., when multiple assignments are classified at a higher common taxonomic
276 level than the one being evaluated). Sample purity (or precision) is the sum of all true positive
277 sequences in the sample divided by the total number of assigned sequences.

278
$$p = \frac{TP}{TP + FP}$$

279 Average purity is the precision calculated for each taxonomic bin averaged across all bins. For
280 average purity small bins have the same weight as larger bins whereas larger bins contribute
281 more to the sample purity value. Sample accuracy was calculated as the sum of true positive
282 sequences divided by the total number of sequences (including those that were unassigned).
283

284
$$a = \frac{TP}{FN + TP + FP}$$

285

286 The average completeness is the sensitivity of each taxonomic bin averaged across all bins.
287 Completeness is calculated as the number of true positive sequences in the bin divided by the
288 true size of the bin. Again, the average completeness weighs all bins equally, regardless of their
289 size.

290
$$c = \frac{TP}{TP + FN}$$

291 Simulating reads

292 We used Mason v2.0.9 (Holtgrewe, 2010) to simulate 150bp Illumina reads using default
293 parameters. To compare the traditional alignment tools using the toy indices, we simulated and
294 combined 1 million reads from a single reference sequence from each species in the index
295 (Supplemental Table 1). When analyzing how divergent genomes impacted taxonomic
296 assignment, we used Mutation-Simulator v2.0.3 (Kühl, 2019) to simulate SNPs, insertions, and
297 deletions in a reference genome for *Faecalibacterium prausnitzii* (GCA_003312465.1) at rates
298 ranging from 0-15% with indels occurring at 1/10 the rate of the SNPs. From these genome
299 sequences we simulated 1 million paired reads using Mason. These reads were then added to the
300 CAMI2 gastrointestinal tract reads which we confirmed did not contain any sequences from *F.*

301   *prausnitzii*. For benchmarking the taxonomic assignment tools, 10 million reads were simulated
302   from the reference sequences in the bacterial RefSeq database.

Pathogen spike-in to soil background

304   Using real sequencing data, we tested the performance of MTSv, Centrifuge, and Kraken2 at
305   detecting pathogen DNA spiked into a complex soil background. We extracted and quantified,
306   via qPCR (Liu et al., 2012), DNA from strains of *Bacillus anthracis, Francisella tularensis*, and
307   *Yersinia pestis* from the Pathogen and Microbiome Institute's culture collection. The pathogen
308   DNA was then spiked into DNA extracted from pooled environmental soil samples collected in
309   Puerto Rico in 2016. We used qPCR to verify that none of the pathogens were present in the soil
310   background prior to the spike-in (for *B. anthracis,* we used primers and protocols targeting a
311   SNP in the *plcR* gene (Easterday et al., 2005), for *F. tularensis* we used the Ft-sp.FTS_0772
312   assay primers and protocols (Öhrman et al., 2021), and for *Y. pestis,* we targeted the *pla* gene
313   using primers and protocols from (Hinnebusch & Schwan, 1993), see Supplemental Methods).
314   DNA from each pathogen was spiked into separate aliquots of the soil background at a 2:98 ratio
315   (roughly 2% of the final mixture). The mixtures and the soil background were sequenced on an
316   Illumina MiSeq platform (V3, 2x300kit). The reads were quality-filtered using Fastp v0.20.122
317   and trimmed to 150bp. The raw reads are available through the NCBI Sequence Read Archive
318   with the following accessions: SAMN31055686 (*B. anthracis* mixture), SAMN31055687 (*Y.*
319   *pestis* mixture), SAMN31055688 (*F. tularensis* mixture), and SAMN31055689 (soil
320   background).

Results

Traditional read alignment tools were not designed or intended for taxonomic assignment against multi-species reference databases.

324   The standard search effort and reporting modes of traditional read alignment tools result in
325   misclassification when used for taxonomic assignment. Bowtie2 and Minimap2 missed
326   thousands of alignments particularly for species that were underrepresented in the reference
327   index (Table 1) or when genomes from closely related species were included. By default,
328   Bowtie2 reports only the best alignment found after a defined level of search effort (choosing
329   one at random if there is a tie) whereas Minimap2 reports up to 5 alternative alignments per read.
330   When the standard search effort and reporting modes were used, many reads aligned to the
331   dominant and/or closely related taxa rather than the correct taxon. One index contained closely
332   related species (>99% identity in some cases) from the *Bacillus* genus (Helgason et al. 2000;
333   Rasko et al., 2005) and was dominated by *Bacillus anthracis* references. Due to the similarity
334   between the genomes, valid alignments for reads of one species were often found within other
335   references. In some cases, the search effort was expended before the alignment to the correct
336   genome was found. The other index included references for *S. aureus* and *S. epidermidis* which

337 share many genes in common but the average nucleotide identity among homologous sequences
338 is low around 75% (Méric et al., 2015). Despite this, many alignments to *S. epidermidis* were
339 missed because *S. aureus* sequences were highly overrepresented (961:1) resulting in biased
340 alignments. Minimap2 performed better than Bowtie2 because it reported some alternative
341 alignments by default. But in either case, when using a multi-species reference index, the level of
342 search effort did not extend to cover all taxa evenly resulting in biases and missed assignments.
343   When Bowtie2 and Minimap2 were forced to find and report all alignments, the number
344 of missed assignments was reduced but the runtime and memory requirements increased (Figure
345 3). Bowtie2 took over 7 hours to align 2 million reads against the *Staphylococcus* index
346 compared to just a couple of minutes in standard mode (~300 times longer). Minimap2 was
347 faster than Bowtie2 although it did not report as many alignments (855M versus 986M), and it
348 failed to find any alignments for over 5,000 reads. Minimap2 took over an hour compared to ~18
349 minutes in standard mode (~4 times longer). Minimap2 also required 111GB of memory when
350 searching for all alignments compared to 7GB in standard mode. MTSv took on average 12
351 minutes to run (~35 times faster than Bowtie2 and 5 times faster than Minimap2) and required
352 9.5GB of memory (only 8% of what Minimap2 required) to achieve the same or better results (in
353 terms of correct taxonomic assignments). MTSv does an exhaustive search of all candidate
354 alignments but it is taxon-aware so it will stop attempting alignments against reference
355 sequences for a TaxID that already has a valid match. Consequently, MTSv was able to skip over
356 1 billion (93%) redundant and computationally expensive alignments while still correctly
357 assigning the *Staphylococcus* reads.

MTSv has high precision and sensitivity and reduces false positive assignments.

360 MTSv had nearly perfect sample purity and much higher average purity compared to Centrifuge
361 and Kraken2 (Figure 4, rows 1 and 2, respectively). This indicates that most of the assignments
362 made by MTSv at each taxonomic level were correct whereas, in every case Centrifuge and
363 Kraken2 had incorrect assignments all the way up to the kingdom level (assigning reads to a
364 false archaea bin as indicated by the average purity of 0.5). High precision is the advantage of
365 using a full alignment approach and this is particularly clear when looking at the number of false
366 positive bins (Fig. 5, middle row). At the species level, Centrifuge and Kraken2 assigned reads to
367 thousands of additional taxa that were not in the sample whereas the highest number of false bins
368 for MTSv was 605 for the airways sample. Additionally, the sizes of the false bins were
369 generally smaller for MTSv (Fig. 5, bottom row).
370   While MTSv has a more conservative approach to making assignments (here the
371 alignment of a 150bp read could only contain up to 20 mismatches compared to Centrifuge
372 which will report a hit if it finds an exact match > 22 bp long) it was still able to assign a large
373 portion of the reads at the species level (Fig. 5, first row). As a result, MTSv maintained high
374 sample accuracy and average completeness (Fig. 4, rows 3 and 4, respectively) by keeping the
375 number of false negatives low. MTSv and Kraken2 had similarly high performance for these

376 metrics while Centrifuge was penalized for having more unassigned reads. The oral sample
377 contained viral sequences that were not included in our bacteria-only reference database, so all
378 the tools had lower average completeness for that sample.

MTSv assigns more reads from genomes that diverge from the reference.

380 The limited number of reference genomes available for taxonomic classification means that
381 sequences in metagenomic samples will often share low similarity with existing database
382 sequences. Divergent sequences contain more gaps and mismatches which will reduce the
383 number of exact k-mer matches. As a result, exact matching approaches tend to miss these
384 assignments completely or have a reduced number of reads assigned to these taxa which can
385 influence downstream abundance estimations.
386   The number of reads assigned accurately by Centrifuge and Kraken2 to *Faecalibacterium*
387 *prausnitzii* dropped more rapidly as the genome diverged from the reference sequence (Fig. 6,
388 left). Even at 95% genome similarity, the number of reads assigned dropped to 91% and 94%,
389 respectively, while MTSv was at 98% using default parameters. At 90% similarity, assigned
390 reads dropped to ~50% for Centrifuge and Kraken2 while MTSv still assigned 74%. The reads
391 were spiked into the CAMI2 gastrointestinal sample to show how the missed assignments
392 affected the proportion of reads assigned to *F. prausnitzii* compared to other species in the
393 sample (Fig. 6, middle). For Centrifuge and Kraken2, the unassigned reads caused *F. prausnitzii*
394 to drop in rank below 3 other species in relative sequence abundance when genome similarity
395 was 90% and below 5 species at 85% (Fig. 6, right). Using MTSv default parameters, *F.*
396 *prausnitzii* dropped below only 3 species at 85% genome similarity.
397   With default parameters, MTSv outperformed Centrifuge and Kraken2 in assigning reads
398 from genomes that diverge from reference sequences. By using more sensitive search parameters
399 (seed size = 16, seed interval = 2) and relaxing the edit distance cutoff to 20%, MTSv performed
400 even better. With these settings, MTSv assigned 84% of *F. prausnitzii* reads and read abundance
401 dropped below only one other species when genome similarity was at 85%.

MTSv provides higher confidence in the detection of pathogens.

403 In microbial forensics and clinical applications, it is important to have a high level of confidence
404 that a detected organism is truly present in the sample. The presence of many false positive
405 assignments can erode this confidence and may require additional time consuming and expensive
406 validation tests. Ideally, adding a single organism to a complex background should not result in
407 many additional false positive assignments. However, although each of the tools was able to
408 detect the correct species in our pathogen spike-in experiments, Centrifuge and Kraken2
409 introduced many additional false positive assignments to near neighbor species.
410   The total number of reads assigned by each tool was low which is expected when
411 performing taxonomic assignments of a complex soil mixture against a relatively limited
412 collection of reference sequences (Table 2). Both Centrifuge and Kraken2 assigned more reads
413 than MTSv but many of the assignments had very low support. The median number of k-mers

414   mapped per read for Kraken2 was 19 (16.4%), 14 (12.1%), and 16 (13.8%) for *B. anthracis, F.*
415   *tularensis,* and *Y. pestis* spike-ins, respectively and the median hit length for Centrifuge was
416   24bp for all three samples. In contrast, all the assignments made by MTSv had 20 or fewer
417   mismatches from the reference sequences. Relaxing the mismatch cutoff to 30 increased the
418   number of assigned reads to 706,182, 680,015, and 604,819, respectively, while still requiring an
419   80% match with the reference sequence.
420   Despite having fewer total reads assigned, MTSv made the most assignments to the
421   spike-in taxa at the species level. For *B. anthracis* and *Y. pestis,* fewer calls could be made at the
422   species level, likely due to high genetic similarity between these and near neighbor species. For
423   example, *Bacillus anthracis* shares >99% average nucleotide identity with its closest relatives in
424   the *Bacillus cereus* group (Helgason et al. 2000; Rasko et al., 2005). When reads were assigned
425   to multiple species in a genus, the assignment was correctly made at the genus level. Unlike the
426   other tools, MTSv was less likely to assign reads to near neighbor species within the same genus
427   and introduced fewer new species assignments.
428   A comparison to the soil background sample (Fig. 7) shows that the additional near
429   neighbor assignments were likely false positives rather than a true signal in the soil background.
430   Because Kraken2 and Centrifuge had many low confidence assignments, we opted to include
431   filtered results for comparison. We used a confidence filter of 0.1 for Kraken2 (Wood, Lu, and
432   Langmead, 2019; Ye et al., 2019) and a score filter of 150 for Centrifuge (Ammer-Herrmenau et
433   al., 2021; Sanderson et al., 2018); neither filter had a large impact on the number of assignments
434   to the spike-in species. The unfiltered Kraken2 results showed large increases (>2-fold) in 13
435   additional taxa in the *Bacillus* genus with a maximum of a 16-fold increase in *B. wiedmannii.*
436   Filtering the low confidence assignments reduced this to 8 taxa with more than a 2-fold increase
437   and maximum 9-fold increase for *B. thuringiensis*. Prior to filtering, Centrifuge showed >2-fold
438   increase in 6 *Bacillus* species but after filtering this increased to 7 species with a maximum of a
439   6-fold increase. Centrifuge appeared to perform better prior to filtering but this was because it
440   had many more assignments to *Bacillus* species in the soil background compared to the other
441   tools and most of them were of very low quality (score < 150, hit length < 32bp). The higher
442   number of (albeit low quality) assignments for *Bacillus* species in the background sample
443   masked the increase due to the spike-in prior to filtering them out. MTSv had only 3 *Bacillus*
444   species with greater than a 2-fold increase with a maximum of 5-fold.

445   ## MTSv is reasonably fast for a full-alignment algorithm and is flexible
446   ## enough to run on lower memory or limited CPU environments.

447   MTSv offers a lot of flexibility when it comes to managing the resources required for taxonomic
448   assignment against large databases. This begins with the index building step. Unlike Centrifuge,
449   Kraken2, and other tools, MTSv allows the indices to be split up into different sized chunks to
450   accommodate resource limitations. For example, because the Centrifuge index must be built all
451   at once and requires a large amount of RAM that scales with the size of the reference genome,
452   we were restricted to using a smaller reference database in this paper to allow direct comparison

453   with Centrifuge. MTSv indices can be built using any reference database or even multiple
454   reference databases because it breaks them into smaller chunks that require less memory to build
455   and run. MTSv can easily combine the assignment results across multiple indices and additional
456   indices can be added without needing to rebuild the entire database.
457   In this paper, we chose to split the reference sequences among 13 different MG-indices
458   for MTSv. Each index required 251GB of memory and an average of 1 hour 23 minutes to build.
459   Alternatively for more memory limited situations, the reference sequences could be split into 66
460   ~2GB files which would each require 49GB of memory and about 18 minutes each to build
461   (Table 3). Both approaches required less memory than Centrifuge's build process which required
462   317GB of RAM. The storage size of the MTSv indices (using the default suffix array and
463   occurrence array sampling intervals of 32 and 64, respectively) was ~3.6 times the size of the
464   original reference sequences. Although this is larger than Kraken2 (~0.3x) and Centrifuge
465   (~0.5x), storage space is arguably the least expensive and most readily available resource. The
466   alternative example in Table 3 (MTSv alt) was built with a higher occurrence array sampling rate
467   of 512 so the files took up less space on disk (~2.3x the original FASTA) but they required
468   slightly longer binning times.
469   Because MTSv performs full alignments, it cannot compete with the speeds of exact
470   matching tools, but it is reasonably fast compared to traditional alignment tools. For taxonomic
471   assignment, the average time per index was about 30 minutes for MTSv and each required about
472   34GB of memory (Table 4). If each of the 13 indices were run simultaneously then the run time
473   would be equal to the time for the longest index (~40 minutes). If each index was run in series, it
474   would take about 6 and a half hours or about 13 times longer than Centrifuge.

Discussion

476   MTSv provides a good middle ground between highly sensitive traditional alignment tools that
477   are too computationally expensive for taxonomic profiling and faster exact matching-based
478   approaches that can be less sensitive and less precise. MTSv was not designed to compete with
479   tools like Kraken2 or Centrifuge, but rather fill a role when higher accuracy and confidence is
480   required from taxonomic assignments and as an alternative to traditional alignment tools used in
481   other metagenomic analysis pipelines (e.g., Metalign (LaPierre et al., 2020)). Here we show that
482   our full alignment approach (1) increases precision and reduces false positive assignments, (2)
483   better captures assignments for taxa that share low similarity with reference sequences, and (3)
484   provides high confidence detection of pathogens while avoiding off-target assignments to near-
485   neighbor species.
486   We compared MTSv to two traditional alignment tools to illustrate the issues that arise
487   from off-label use for taxonomic assignment against multi-species reference databases. Bowtie2
488   is one of the most popular rapid read alignment tools and one of the first to use an FM-index data
489   structure to rapidly look up seeds to extend for full alignments (Langmead & Salzberg, 2012;
490   Alser et al., 2021). Minimap2 is a state-of-the-art hash-based alignment tool that is faster than
491   mainstream short-read aligners with comparable accuracy (Li, 2018; Alser et al., 2021). In our

492 small example, we used a collection of genomes that were dominated by a single species and
493 included species that were closely related. Although this was a toy example to demonstrate a
494 point, it is not far off from reality. Many clinically or economically important species are
495 overrepresented in reference databases and other closely related, yet less relevant species may
496 only have a single complete genome. When using traditional alignment tools, accurate
497 assignment can rely on finding an alignment to a single genome among many before the level of
498 search effort is exhausted. If other valid alignments are found among sequences for closely
499 related taxa, the search effort is more likely to be exhausted before the correct alignment is
500 found. This leads to a bias for assignments to more abundant genomes and noisy results,
501 especially when extended beyond our toy example.
502   One approach to overcome the bias of traditional alignment tools is to increase their
503 search effort to include all valid alignments. Unfortunately, this made the tools very inefficient in
504 terms of speed and, for Minimap2, memory usage, even on a small toy dataset. Attempts to scale
505 traditional read alignment tools to realistic sized metagenomic samples with larger more complex
506 reference databases are sometimes unsuccessful (Jaillard et al., 2016) and increasing the search
507 effort will only make this more impractical. An alternative approach could involve building
508 separate indices for each taxon. Unfortunately, this is also impractical as it would require
509 maintaining and updating over 4,700 separate indices for just the bacterial reference sequence
510 database used here. MTSv was designed specifically to overcome these challenges because it
511 was purpose built to handle taxonomic assignment of metagenomic reads against multi-species
512 reference indices. MTSv does an exhaustive unbiased search for alignments across all reference
513 sequences but it is taxon-aware so once it finds a valid alignment for a taxon, no additional
514 alignments are attempted against any reference for that taxon. This prevents missed assignments
515 while providing large speedups.
516   We compared MTSv to two taxonomic assignment tools, Centrifuge and Kraken2, to
517 demonstrate the situations where MTSv may provide superior performance. Using data with
518 known read assignments, MTSv had high precision and sensitivity at each taxonomic level.
519 Consequently, MTSv reduced the number of false positive species identified by thousands. The
520 false bins introduced with Centrifuge and Kraken2 were larger and therefore are more likely to
521 be treated as a positive signal. Noisy results interfere with true positive signals and make it
522 difficult to detect low abundance taxa with confidence. MTSv reports only high-quality
523 alignments (with >87% similarity with the reference) and reduces the number of false positives
524 thereby increasing overall confidence in the assignment results.
525   Exact match approaches underperform when sequences diverge from reference genomes.
526 Differences between a read and a reference sequence can occur due to sequencing errors, DNA
527 degradation (e.g., ancient DNA (Prys-Jones et al., 2022)), or taxon diversity that is not captured
528 by existing databases. These differences can rapidly decay the potential for exact matches by
529 interrupting stretches of sequence identity and modifying parameters to adjust for this (e.g.,
530 reducing the fixed k-mer size in Kraken2) often sacrifices resolution. By performing a full
531 alignment, MTSv can account for mismatches and assign reads that would be missed by other

532   tools and the alignment sensitivity can be modified easily to capture more divergent sequences.
533   With sensitive parameters, MTSv found 80% of assignments for reads that diverged up to 15%
534   from reference sequences, whereas Kraken2 and Centrifuge found less than 20%. Increasing
535   sensitivity also increases runtime but this can be mitigated by running sensitive mode only on
536   reads that could not be assigned using the default mode.
537   The assignments missed due to sequence divergence are likely to influence downstream
538   species abundance estimates. Both Centrifuge and Kraken2 (with the help of Bracken (Lu et al.,
539   2017) can provide species abundance estimates based on the number of reads assigned to each
540   taxon. Instead of using only species level assignments, Bracken improves abundance estimates
541   by redistributing reads assigned at the genus level or higher among the species level counts.
542   However, it cannot correct for missing or false species-level assignments. We showed that when
543   an organism diverges from reference sequences, Kraken2 and Centrifuge completely miss many
544   assignments which reduces the read count for the species and cannot be corrected using Bracken.
545   For this reason, the abundance estimates are highly dependent on the accuracy of species level
546   assignments. Minor fluctuations in the number of assigned reads will likely have little effect but
547   our results show that divergent genomes caused large shifts in read abundance rankings among
548   species.
549   K-mer based approaches like Kraken2 must strike the right balance between sensitivity
550   and precision when picking a k-mer size. With longer k-mers, fewer exact matches are likely due
551   to sequencing errors or biological variation. However, as k-mer size decreases, there are fewer
552   unique k-mers per taxon and the resolution suffers (Lu et al. 2017; Wood and Salzberg, 2014;
553   Wood, Lu, and Langmead, 2019). This is particularly challenging when trying to detect specific
554   pathogenic species in metagenomic samples when closely related species share many identical k-
555   mers. Low coverage may result in missing the k-mers that differentiate between the taxa and
556   sequencing errors can easily cause false positive assignments. For example, the anthrax
557   pathogen, *Bacillus anthracis*, shares >99% average nucleotide identity with its closest relatives
558   in the *Bacillus cereus* group that includes ubiquitous soil bacteria like *Bacillus cereus* and
559   *Bacillus thuringiensis* (Helgason et al. 2000; Rasko et al., 2005). This high degree of similarity
560   led to the re-analysis of a sensational 2015 study that identified the alarming presence of *B.*
561   *anthracis* in metagenomic samples collected from New York City subways (Ackelsberg et al.
562   2015; Afshinnekoo et al. 2015a, 2015b). The re-analysis revealed that the detection of *B.*
563   *anthracis* was likely an artifact of high levels of sequence overlap with other species. We show
564   here that Kraken2 was susceptible to misclassifying reads within the *Bacillus* genus even after
565   filtering out low confidence assignments. Even tools like KrakenUniq (Breitwieser, Baker &
566   Salzberg, 2018) which were designed specifically to avoid false positives in pathogen detection
567   has been prone to falsely predict the presence of *B. anthracis* when other *Bacillus cereus* Group
568   species are present (Petit III et al., 2018). We are working on developing similar methods that
569   will provide confidence estimates for the true presence of taxa in metagenomic samples, but out
570   of the box, MTSv offered more high-quality assignments in our pathogen detection experiments
571   without creating a large cloud of additional assignments to near neighbor species. This is highly

572 valuable for higher-stakes pathogen detection situations (e.g., analyzing clinical or biothreat
573 samples).
574   Our goal when developing MTSv was to provide an alignment algorithm that was
575 specifically designed for analyzing metagenomic data. Although traditional sequence alignment
576 tools are very efficient, they must perform exhaustive searches to avoid missing alignments using
577 multi-species reference databases which become computationally expensive at scale. In
578 metagenomics, it is less important to identify all the possible alignments and more important to
579 know that an alignment of sufficient quality was found among the reference sequences for a
580 given taxon. MTSv performs a taxon-aware exhaustive search but skips references for TaxIDs
581 where a valid alignment has been found. This allows MTSv to provide full alignment-based
582 taxonomic assignments at speeds that are feasible for metagenomic data. MTSv is more flexible
583 than other existing tools when it comes to modifying the pipeline to work within different
584 computational restrictions. For example, with Centrifuge, the FM-index must be built all at once
585 which requires large amounts of memory that scales with the size of the reference database (we
586 did not have the resources to build anything larger than the bacteria-only complete genomes
587 database). The Centrifuge developers do provide prebuilt indices but that limits control of
588 updates and customization. MTSv indices, on the other hand, can be built in smaller units that
589 suit the user's computing resources and additional reference sequences can be added without
590 needing to rebuild the complete data structure. Although MTSv cannot compete with the
591 assignment speeds of the exact matching approaches, it offers several key advantages while still
592 being faster than other traditional alignment tools.

593 ## Conclusions

594 In conclusion, MTSv provides a novel and relatively fast algorithm for taxonomic classification
595 based on full alignment of nucleotide sequences. It offers greater precision, and confidence in
596 assignments than exact-matching tools which is required in some situations. It offers flexibility
597 to fit different computational needs and can easily expand to accommodate rapidly growing
598 reference databases.

599 ## Acknowledgements

600 The authors are grateful for the programming work provided by Adam Perry, and for the support
601 and guidance of Johannes Köster and other contributors to the Rust-Bio library. The authors also
602 want to acknowledge the support from the Northern Arizona University high-performance
603 computing staff.

604 ## Funding

605 This work has been supported by the Department of Homeland Security Grant [HSHQDC-17-C-
606 B0008/BAA14-003]. Software development, benchmarking, and analysis was performed on

607 Northern Arizona University's Monsoon computing cluster, funded by Arizona's Technology
608 and Research Initiative Fund.

References

610 Ackelsberg J, Rakeman J, Hughes S, Petersen J, Mead P, Schriefer M, Kingry L, Hoffmaster A,
611 Gee JE. 2015. Lack of evidence for plague or anthrax on the New York City subway. *Cell*
612 *Systems* 1:4–5. DOI: 10.1016/j.cels.2015.07.008.

613 Afshinnekoo E, Meydan C, Chowdhury S, Jaroudi D, Boyer C, Bernstein N, Maritz JM, Reeves
614 D, Gandara J, Chhangawala S, Ahsanuddin S, Simmons A, Nessel T, Sundaresh B, Pereira E,
615 Jorgensen E, Kolokotronis S-O, Kirchberger N, Garcia I, Gandara D, Dhanraj S, Nawrin T,
616 Saletore Y, Alexander N, Vijay P, Hénaff EM, Zumbo P, Walsh M, O'Mullan GD, Tighe S,
617 Dudley JT, Dunaif A, Ennis S, O'Halloran E, Magalhaes TR, Boone B, Jones AL, Muth TR,
618 Paolantonio KS, Alter E, Schadt EE, Garbarino J, Prill RJ, Carlton JM, Levy S, Mason CE.
619 2015a. Geospatial resolution of human and bacterial diversity with city-scale metagenomics. *Cell*
620 *Systems* 1:72–87. DOI: 10.1016/j.cels.2015.01.001.

621 Afshinnekoo E, Meydan C, Chowdhury S, Jaroudi D, Boyer C, Bernstein N, Maritz JM, Reeves
622 D, Gandara J, Chhangawala S, Ahsanuddin S, Simmons A, Nessel T, Sundaresh B, Pereira E,
623 Jorgensen E, Kolokotronis S-O, Kirchberger N, Garcia I, Gandara D, Dhanraj S, Nawrin T,
624 Saletore Y, Alexander N, Vijay P, Hénaff EM, Zumbo P, Walsh M, O'Mullan GD, Tighe S,
625 Dudley JT, Dunaif A, Ennis S, O'Halloran E, Magalhaes TR, Boone B, Jones AL, Muth TR,
626 Paolantonio KS, Alter E, Schadt EE, Garbarino J, Prill RJ, Carlton JM, Levy S, Mason CE.
627 2015b. Modern methods for delineating metagenomic complexity. *Cell Systems* 1:6–7. DOI:
628 10.1016/j.cels.2015.07.007.

629 Alan G, Sarah JP. 2012. Microbes as forensic indicators. *Tropical Biomedicine* 29:311–330.

630 Alser M, Rotman J, Deshpande D, Taraszka K, Shi H, Baykal PI, Yang HT, Xue V, Knyazev S,
631 Singer BD, Balliu B, Koslicki D, Skums P, Zelikovsky A, Alkan C, Mutlu O, Mangul S. 2021.
632 Technology dictates algorithms: Recent developments in read alignment. *Genome Biology*
633 22:249. DOI: 10.1186/s13059-021-02443-7.

634 Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. 1990. Basic local alignment search tool.
635 *Journal of Molecular Biology* 215:403–410. DOI: 10.1016/S0022-2836(05)80360-2.

636 Ammer-Herrmenau C, Pfisterer N, van den Berg T, Gavrilova I, Amanzada A, Singh SK, Khalil
637 A, Alili R, Belda E, Clement K, Wahed AA El, Gady EE, Haubrock M, Beißbarth T, Ellenrieder
638 V, Neesse A, Segata N. 2021. Comprehensive wet-bench and bioinformatics workflow for
639 complex microbiota using Oxford Nanopore technologies. *mSystems* 6:e00750-21. DOI:
640 10.1128/mSystems.00750-21.

641 Bazinet AL, Ondov BD, Sommer DD, Ratnayake S. 2018. BLAST-based validation of
642 metagenomic sequence assignments. *PeerJ* 6:e4892–e4892. DOI: 10.7717/peerj.4892.

643 Breitwieser FP, Baker DN, Salzberg SL. 2018. KrakenUniq: confident and fast metagenomics
644 classification using unique k-mer counts. *Genome Biology* 19:198. DOI: 10.1186/s13059-018-
645 1568-0.

646 Breitwieser FP, Lu J, Salzberg SL. 2019. A review of methods and databases for metagenomic
647 classification and assembly. *Briefings in Bioinformatics* 20:1125–1136. DOI:

648   10.1093/bib/bbx120.

649   Brown CT, Irber L. 2016. sourmash: a library for MinHash sketching of DNA. *Journal of Open*
650   *Source Software* 1:27. DOI: 10.21105/joss.00027.

651   Buchfink B, Reuter K, Drost H-G. 2021. Sensitive protein alignments at tree-of-life scale using
652   DIAMOND. *Nature Methods* 18:366–368. DOI: 10.1038/s41592-021-01101-x.

653   Buchfink B, Xie C, Huson DH. 2015. Fast and sensitive protein alignment using DIAMOND.
654   *Nature Methods* 12:59–60. DOI: 10.1038/nmeth.3176.

655   Burkhardt S, Crauser A, Ferragina P, Lenhof H-P, Rivals E, Vingron M. 1999. *Q*-gram based
656   database searching using a suffix array (QUASAR). In: *Proceedings of the Third Annual*
657   *International Conference on Computational Molecular Biology*. RECOMB '99. New York, NY,
658   USA: Association for Computing Machinery, 77–83. DOI: 10.1145/299432.299460.

659   Burrows, M., Wheeler, D., J. 1994. A block-sorting lossless data compression algorithm.
660   Technical Report 124, Digital Equipment Corporation.

661   Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, Madden TL. 2009.
662   BLAST+: architecture and applications. *BMC Bioinformatics* 10:421. DOI: 10.1186/1471-2105-
663   10-421.

664   Chiu CY, Miller SA. 2019. Clinical metagenomics. *Nature Reviews Genetics* 20:341–355. DOI:
665   10.1038/s41576-019-0113-7.

666   Clarke TH, Gomez A, Singh H, Nelson KE, Brinkac LM. 2017. Integrating the microbiome as a
667   resource in the forensics toolkit. *Forensic Science International: Genetics* 30:141–147. DOI:
668   https://doi.org/10.1016/j.fsigen.2017.06.008.

669   Easterday WR, Van Ert MN, Simonson TS, Wagner DM, Kenefic LJ, Allender CJ, Keim P.
670   2005. Use of single nucleotide polymorphisms in the *plcR* gene for specific identification of
671   *Bacillus anthracis*. *Journal of Clinical Microbiology* 43:1995–1997. DOI:
672   10.1128/JCM.43.4.1995-1997.2005.

673   Ferragina P, Manzini G. 2000. Opportunistic data structures with applications. *41st Annual*
674   *Symposium on Foundations of Computer Science, Proceedings*:390–398. DOI:
675   10.1109/SFCS.2000.892127.

676   De Filippis F, Valentino V, Alvarez-Ordóñez A, Cotter PD, Ercolini D. 2021. Environmental
677   microbiome mapping as a strategy to improve quality and safety in the food industry. *Current*
678   *Opinion in Food Science* 38:168–176. DOI: https://doi.org/10.1016/j.cofs.2020.11.012.

679   Gu W, Miller S, Chiu CY. 2019. Clinical metagenomic next-generation sequencing for pathogen
680   detection. *Annual Review of Pathology: Mechanisms of Disease* 14:319–338. DOI:
681   10.1146/annurev-pathmechdis-012418-012751.

682   Helgason E, Okstad OA, Caugant DA, Johansen HA, Fouet A, Mock M, Hegna I, Kolstø AB.
683   2000. *Bacillus anthracis, Bacillus cereus*, and *Bacillus thuringiensis*--one species on the basis of
684   genetic evidence. *Applied and Environmental Microbiology* 66:2627–2630. DOI:
685   10.1128/AEM.66.6.2627-2630.2000.

686   Hinnebusch J, Schwan TG. 1993. New method for plague surveillance using polymerase chain
687   reaction to detect *Yersinia pestis* in fleas. *Journal of Clinical Microbiology* 31:1511–1514. DOI:

688   10.1128/jcm.31.6.1511-1514.1993.

689   Holtgrewe M. 2010. Mason: A read simulator for second generation sequencing data. *Technical*
690   *Report FU Berlin*.

691   Huerta-Cepas J, Serra F, Bork P. 2016. ETE 3: Reconstruction, analysis, and visualization of
692   phylogenomic data. *Molecular Biology and Evolution* 33:1635–1638. DOI:
693   10.1093/molbev/msw046.

694   Huson DH, Auch AF, Qi J, Schuster SC. 2007. MEGAN analysis of metagenomic data. *Genome*
695   *Research* 17:377–386. DOI: 10.1101/gr.5969107.

696   Huttenhower C, Gevers D, Knight R, Abubucker S, Badger JH, Chinwalla AT, Creasy HH, Earl
697   AM, Fitzgerald MG, Fulton RS, Giglio MG, Hallsworth-Pepin K, Lobos EA, Madupu R,
698   Magrini V, Martin JC, Mitreva M, Muzny DM, Sodergren EJ, Versalovic J, Wollam AM,
699   Worley KC, Wortman JR, Young SK, Zeng Q, Aagaard KM, Abolude OO, Allen-Vercoe E, Alm
700   EJ, Alvarado L, Andersen GL, Anderson S, Appelbaum E, Arachchi HM, Armitage G, Arze CA,
701   Ayvaz T, Baker CC, Begg L, Belachew T, Bhonagiri V, Bihan M, Blaser MJ, Bloom T, Bonazzi
702   V, Paul Brooks J, Buck GA, Buhay CJ, Busam DA, Campbell JL, Canon SR, Cantarel BL, Chain
703   PSG, Chen IMA, Chen L, Chhibba S, Chu K, Ciulla DM, Clemente JC, Clifton SW, Conlan S,
704   Crabtree J, Cutting MA, Davidovics NJ, Davis CC, Desantis TZ, Deal C, Delehaunty KD,
705   Dewhirst FE, Deych E, Ding Y, Dooling DJ, Dugan SP, Michael Dunne W, Scott Durkin A,
706   Edgar RC, Erlich RL, Farmer CN, Farrell RM, Faust K, Feldgarden M, Felix VM, Fisher S,
707   Fodor AA, Forney LJ, Foster L, Di Francesco V, Friedman J, Friedrich DC, Fronick CC, Fulton
708   LL, Gao H, Garcia N, Giannoukos G, Giblin C, Giovanni MY, Goldberg JM, Goll J, Gonzalez
709   A, Griggs A, Gujja S, Kinder Haake S, Haas BJ, Hamilton HA, Harris EL, Hepburn TA, Herter
710   B, Hoffmann DE, Holder ME, Howarth C, Huang KH, Huse SM, Izard J, Jansson JK, Jiang H,
711   Jordan C, Joshi V, Katancik JA, Keitel WA, Kelley ST, Kells C, King NB, Knights D, Kong HH,
712   Koren O, Koren S, Kota KC, Kovar CL, Kyrpides NC, La Rosa PS, Lee SL, Lemon KP, Lennon
713   N, Lewis CM, Lewis L, Ley RE, Li K, Liolios K, Liu B, Liu Y, Lo CC, Lozupone CA, Dwayne
714   Lunsford R, Madden T, Mahurkar AA, Mannon PJ, Mardis ER, Markowitz VM, Mavromatis K,
715   McCorrison JM, McDonald D, McEwen J, McGuire AL, McInnes P, Mehta T,
716   Mihindukulasuriya KA, Miller JR, Minx PJ, Newsham I, Nusbaum C, Oglaughlin M, Orvis J,
717   Pagani I, Palaniappan K, Patel SM, Pearson M, Peterson J, Podar M, Pohl C, Pollard KS, Pop M,
718   Priest ME, Proctor LM, Qin X, Raes J, Ravel J, Reid JG, Rho M, Rhodes R, Riehle KP, Rivera
719   MC, Rodriguez-Mueller B, Rogers YH, Ross MC, Russ C, Sanka RK, Sankar P, Fah
720   Sathirapongsasuti J, Schloss JA, Schloss PD, Schmidt TM, Scholz M, Schriml L, Schubert AM,
721   Segata N, Segre JA, Shannon WD, Sharp RR, Sharpton TJ, Shenoy N, Sheth NU, Simone GA,
722   Singh I, Smillie CS, Sobel JD, Sommer DD, Spicer P, Sutton GG, Sykes SM, Tabbaa DG,
723   Thiagarajan M, Tomlinson CM, Torralba M, Treangen TJ, Truty RM, Vishnivetskaya TA,
724   Walker J, Wang L, Wang Z, Ward D V., Warren W, Watson MA, Wellington C, Wetterstrand
725   KA, White JR, Wilczek-Boney K, Wu Y, Wylie KM, Wylie T, Yandava C, Ye L, Ye Y,
726   Yooseph S, Youmans BP, Zhang L, Zhou Y, Zhu Y, Zoloth L, Zucker JD, Birren BW, Gibbs
727   RA, Highlander SK, Methé BA, Nelson KE, Petrosino JF, Weinstock GM, Wilson RK, White O.
728   2012. Structure, function and diversity of the healthy human microbiome. *Nature* 486:207–214.
729   DOI: 10.1038/nature11234.

730   Jaillard M, Tournoud M, Meynier F, Veyrieras J-B. 2016. Optimization of alignment-based
731   methods for taxonomic binning of metagenomics reads. *Bioinformatics (Oxford, England)*

732   32:1779–1787. DOI: 10.1093/bioinformatics/btw040.

733   Karlsson OE, Hansen T, Knutsson R, Löfström C, Granberg F, Berg M. 2013. Metagenomic
734   detection methods in biopreparedness outbreak scenarios. *Biosecurity and Bioterrorism :*
735   *Biodefense Strategy, Practice, and Science* 11 Suppl 1:S146-57. DOI: 10.1089/bsp.2012.0077.

736   Kim D, Song L, Breitwieser FP, Salzberg SL. 2016. Centrifuge: Rapid and sensitive
737   classification of metagenomic sequences. *Genome Research* 26:1721–1729. DOI:
738   10.1101/gr.210641.116.

739   Koslicki D, Zabeti H. 2019. Improving MinHash via the containment index with applications to
740   metagenomic analysis. *Applied Mathematics and Computation* 354:206–215. DOI:
741   https://doi.org/10.1016/j.amc.2019.02.018.

742   Köster J. 2016. Rust-Bio: A fast and safe bioinformatics library. *Bioinformatics* 32:444–446.
743   DOI: 10.1093/bioinformatics/btv573.

744   Kühl M. 2019. Mutation Simulator Github Repository. https://github.com/mkpython3/Mutation-
745   Simulator.

746   Langdon WB. 2015. Performance of genetic programming optimised Bowtie2 on genome
747   comparison and analytic testing (GCAT) benchmarks. *BioData Mining* 8:1. DOI:
748   10.1186/s13040-014-0034-0.

749   Langmead B, Salzberg SL. 2012. Fast gapped-read alignment with Bowtie 2. *Nature Methods*
750   9:357–359. DOI: 10.1038/nmeth.1923.

751   LaPierre N, Alser M, Eskin E, Koslicki D, Mangul S. 2020. Metalign: Efficient alignment-based
752   metagenomic profiling via containment Min Hash. *Genome Biology* 21:242. DOI:
753   10.1186/s13059-020-02159-0.

754   Li H. 2007. BWA Github Repository Readme. https://github.com/lh3/bwa.

755   Li H, Durbin R. 2009. Fast and accurate short read alignment with Burrows-Wheeler transform.
756   *Bioinformatics* 25:1754–1760. DOI: 10.1093/bioinformatics/btp324.

757   Li H. 2018. Minimap2: Pairwise alignment for nucleotide sequences. *Bioinformatics* 34:3094–
758   3100. DOI: 10.1093/bioinformatics/bty191.

759   Liu CM, Aziz M, Kachur S, Hsueh PR, Huang YT, Keim P, Price LB. 2012. BactQuant: an
760   enhanced broad-coverage bacterial quantitative real-time PCR assay. *BMC Microbiology* 12:56.
761   DOI: 10.1186/1471-2180-12-56.

762   Lu J, Breitwieser FP, Thielen P, Salzberg SL. 2017. Bracken: Estimating species abundance in
763   metagenomics data. *PeerJ Computer Science* 3:e104. DOI: 10.7717/peerj-cs.104.

764   Manber U, Myers G. 1993. Suffix arrays - A new method for online string searches. *Siam*
765   *Journal on Computing* 22:935–948. DOI: 10.1137/0222058.

766   Menzel P, Ng KL, Krogh A. 2016. Fast and sensitive taxonomic classification for metagenomics
767   with Kaiju. *Nature Communications* 7:11257. DOI: 10.1038/ncomms11257.

768   Méric G, Miragaia M, de Been M, Yahara K, Pascoe B, Mageiros L, Mikhail J, Harris LG,
769   Wilkinson TS, Rolo J, Lamble S, Bray JE, Jolley KA, Hanage WP, Bowden R, Maiden MCJ,
770   Mack D, de Lencastre H, Feil EJ, Corander J, Sheppard SK. 2015. Ecological overlap and
771   horizontal gene transfer in *Staphylococcus aureus* and *Staphylococcus epidermidis*. *Genome*

772   *Biology and Evolution* 7:1313–1328. DOI: 10.1093/gbe/evv066.

773   Meyer F, Fritz A, Deng Z-L, Koslicki D, Lesker TR, Gurevich A, Robertson G, Alser M,
774   Antipov D, Beghini F, Bertrand D, Brito JJ, Brown CT, Buchmann J, Buluç A, Chen B, Chikhi
775   R, Clausen PTLC, Cristian A, Dabrowski PW, Darling AE, Egan R, Eskin E, Georganas E,
776   Goltsman E, Gray MA, Hansen LH, Hofmeyr S, Huang P, Irber L, Jia H, Jørgensen TS, Kieser
777   SD, Klemetsen T, Kola A, Kolmogorov M, Korobeynikov A, Kwan J, LaPierre N, Lemaitre C,
778   Li C, Limasset A, Malcher-Miranda F, Mangul S, Marcelino VR, Marchet C, Marijon P,
779   Meleshko D, Mende DR, Milanese A, Nagarajan N, Nissen J, Nurk S, Oliker L, Paoli L,
780   Peterlongo P, Piro VC, Porter JS, Rasmussen S, Rees ER, Reinert K, Renard B, Robertsen EM,
781   Rosen GL, Ruscheweyh H-J, Sarwal V, Segata N, Seiler E, Shi L, Sun F, Sunagawa S, Sørensen
782   SJ, Thomas A, Tong C, Trajkovski M, Tremblay J, Uritskiy G, Vicedomini R, Wang Z, Wang Z,
783   Wang Z, Warren A, Willassen NP, Yelick K, You R, Zeller G, Zhao Z, Zhu S, Zhu J, Garrido-
784   Oter R, Gastmeier P, Hacquard S, Häußler S, Khaledi A, Maechler F, Mesny F, Radutoiu S,
785   Schulze-Lefert P, Smit N, Strowig T, Bremges A, Sczyrba A, McHardy AC. 2022. Critical
786   assessment of metagenome interpretation: The second round of challenges. *Nature Methods*
787   19:429–440. DOI: 10.1038/s41592-022-01431-4.

788   Meyer F, Hofmann P, Belmann P, Garrido-Oter R, Fritz A, Sczyrba A, McHardy AC. 2018.
789   AMBER: Assessment of Metagenome BinnERs. *GigaScience* 7. DOI:
790   10.1093/gigascience/giy069.

791   Miller RR, Montoya V, Gardy JL, Patrick DM, Tang P. 2013. Metagenomics for pathogen
792   detection in public health. *Genome Medicine* 5:81. DOI: 10.1186/gm485.

793   Minogue TD, Koehler JW, Stefan CP, Conrad TA. 2019. Next-generation sequencing for
794   biodefense: Biothreat detection, forensics, and the clinic. *Clinical Chemistry* 65:383–392. DOI:
795   10.1373/clinchem.2016.266536.

796   Mirdita M, Steinegger M, Breitwieser F, Söding J, Levy Karin E. 2021. Fast and sensitive
797   taxonomic assignment to metagenomic contigs. *Bioinformatics* 37:3029–3031. DOI:
798   10.1093/bioinformatics/btab184.

799   Morgulis A, Coulouris G, Raytselis Y, Madden TL, Agarwala R, Schäffer AA. 2008. Database
800   indexing for production MegaBLAST searches. *Bioinformatics* 24:1757–1764. DOI:
801   10.1093/bioinformatics/btn322.

802   O'Leary NA, Wright MW, Brister JR, Ciufo S, Haddad D, McVeigh R, Rajput B, Robbertse B,
803   Smith-White B, Ako-Adjei D, Astashyn A, Badretdin A, Bao Y, Blinkova O, Brover V,
804   Chetvernin V, Choi J, Cox E, Ermolaeva O, Farrell CM, Goldfarb T, Gupta T, Haft D, Hatcher
805   E, Hlavina W, Joardar VS, Kodali VK, Li W, Maglott D, Masterson P, McGarvey KM, Murphy
806   MR, O'Neill K, Pujar S, Rangwala SH, Rausch D, Riddick LD, Schoch C, Shkeda A, Storz SS,
807   Sun H, Thibaud-Nissen F, Tolstoy I, Tully RE, Vatsan AR, Wallin C, Webb D, Wu W, Landrum
808   MJ, Kimchi A, Tatusova T, DiCuccio M, Kitts P, Murphy TD, Pruitt KD. 2016. Reference
809   sequence (RefSeq) database at NCBI: Current status, taxonomic expansion, and functional
810   annotation. *Nucleic Acids Research* 44:D733-45. DOI: 10.1093/nar/gkv1189.

811   Öhrman C, Sahl JW, Sjödin A, Uneklint I, Ballard R, Karlsson L, McDonough RF, Sundell D,
812   Soria K, Bäckman S, Chase K, Brindefalk B, Sozhamannan S, Vallesi A, Hägglund E, Ramirez-
813   Paredes JG, Thelaus J, Colquhoun D, Myrtennäs K, Birdsell D, Johansson A, Wagner DM,
814   Forsman M. 2021. Reorganized genomic taxonomy of *Francisellaceae* enables design of robust

815   environmental PCR assays for detection of *Francisella tularensis*. *Microorganisms* 9. DOI:
816   10.3390/microorganisms9010146.

817   Ondov BD, Treangen TJ, Melsted P, Mallonee AB, Bergman NH, Koren S, Phillippy AM. 2016.
818   Mash: Fast genome and metagenome distance estimation using MinHash. *Genome Biology*
819   17:132. DOI: 10.1186/s13059-016-0997-x.

820   Petit III RA, Hogan JM, Ezewudo MN, Joseph SJ, Read TD. 2018. Fine-scale differentiation
821   between *Bacillus anthracis* and *Bacillus cereus* group signatures in metagenome shotgun data.
822   *PeerJ* 6:e5515. DOI: 10.7717/peerj.5515.

823   Pierce NT, Irber L, Reiter T, Brooks P, Brown CT. 2019. Large-scale sequence comparisons
824   with sourmash. *F1000Research* 8:1006. DOI: 10.12688/f1000research.19675.1.

825   Prys-Jones TO, Furstenau TN, Abraham AJ, Shaffer IN, Sobek CJ, Upton JR, Hershauer SN,
826   Wong K, Molina M, Menke S, Mead JI, Ebert CH, Carbone MS, Schuur EAG, Walker FM,
827   Fofanov VY, Doughty CE. 2022. Metagenomic analysis of coprolites from three Late
828   Pleistocene megaherbivores from the southwestern United States. *BioRxiv*:2022.05.06.490351.
829   DOI: 10.1101/2022.05.06.490351.

830   Rasko DA, Altherr MR, Han CS, Ravel J. 2005. Genomics of the *Bacillus cereus* group of
831   organisms. *FEMS Microbiology Reviews* 29:303–329. DOI:
832   https://doi.org/10.1016/j.fmrre.2004.12.005.

833   Rasmussen KR, Stoye J, Myers EW. 2006. Efficient q-gram filters for finding all epsilon-
834   matches over a given length. *Journal of Computational Biology : A journal of Computational*
835   *Molecular Cell Biology* 13:296–308. DOI: 10.1089/cmb.2006.13.296.

836   Reinert K, Langmead B, Weese D, Evers DJ. 2015. Alignment of next-generation sequencing
837   reads. *Annual Review of Genomics and Human Genetics* 16:133–151. DOI: 10.1146/annurev-
838   genom-090413-025358.

839   Robinson JM, Pasternak Z, Mason CE, Elhaik E. 2021. Forensic applications of microbiomics: A
840   review. *Frontiers in Microbiology* 11:608101 DOI: 10.3389/fmicb.2020.608101.

841   Sanderson ND, Street TL, Foster D, Swann J, Atkins BL, Brent AJ, McNally MA, Oakley S,
842   Taylor A, Peto TEA, Crook DW, Eyre DW. 2018. Real-time analysis of nanopore-based
843   metagenomic sequencing from infected orthopaedic devices. *BMC Genomics* 19:714. DOI:
844   10.1186/s12864-018-5094-y.

845   Segata N, Waldron L, Ballarini A, Narasimhan V, Jousson O, Huttenhower C. 2012.
846   Metagenomic microbial community profiling using unique clade-specific marker genes. *Nature*
847   *Methods* 9:811–814. DOI: 10.1038/nmeth.2066.

848   Steinegger M, Söding J. 2017. MMseqs2 enables sensitive protein sequence searching for the
849   analysis of massive data sets. *Nature Biotechnology* 35:1026–1028. DOI: 10.1038/nbt.3988.

850   Ukkonen E. 1992. Approximate string-matching with q-grams and maximal matches.
851   *Theoretical Computer Science* 92:191–211. DOI: https://doi.org/10.1016/0304-3975(92)90143-
852   4.

853   Wood DE, Lu J, Langmead B. 2019. Improved metagenomic analysis with Kraken 2. *Genome*
854   *Biology* 20:257. DOI: 10.1186/s13059-019-1891-0.

855   Wood DE, Salzberg SL. 2014. Kraken: Ultrafast metagenomic sequence classification using
856   exact alignments. *Genome Biology* 15:R46. DOI: 10.1186/gb-2014-15-3-r46.

857   Wood JM, Singh NK, Guan L, Seuylemezian A, Benardini JN, Venkateswaran K. 2021.
858   Performance of multiple metagenomics pipelines in understanding microbial diversity of a low-
859   biomass spacecraft assembly facility. *Frontiers in Microbiology* 12. DOI:
860   10.3389/fmicb.2021.685254.

861   Ye SH, Siddle KJ, Park DJ, Sabeti PC. 2019. Benchmarking metagenomics tools for taxonomic
862   classification. *Cell* 178:779–794. DOI: https://doi.org/10.1016/j.cell.2019.07.010.

863   Zhao M, Lee WP, Garrison EP, Marth GT. 2013. SSW library: an SIMD Smith-Waterman
864   C/C++ library for use in genomic applications. *PLoS One* 8:e82138. DOI:
865   10.1371/journal.pone.0082138.

866

867 **Figure 1. Diagram of MTSv index setup and assignment pipeline.** To build the MG-Index, reference sequences are combined
868 into a reference FASTA file formatted with the TaxID in the header. MTSv breaks up large reference files into smaller chunks with
869 a user defined size and builds an MG-Index from each chunk. For taxonomic assignment, reads are aligned against each of the MG-
870 Indices. The binning outputs from each MG-Index are merged to provide the final taxonomic assignments.

871 **Figure 2. MTSv alignment algorithm.** Reads are broken into seeds of a user-specified size (seed-size) that are separated by a
872 user-specified offset (seed-interval). Using the MG-index, the positions of exact seed hits are found and recorded. Regions that
873 have at least a user-specified minimum number of seed hits (min-seed) are marked as candidate regions. The candidate regions
874 are first sorted (descending) by the number of seed hits to ensure that the most promising alignments are checked first. The
875 candidate region sequence and the query are aligned using a SIMD-accelerated Smith-Waterman algorithm and the alignment is
876 considered successful if the edit distance of the alignment is less than or equal to the user-specified minimum (edit-rate). Once a
877 successful alignment is found for a TaxID, no further candidates will be aligned for that TaxID.

878 **Figure 3. Forcing traditional alignment tools to report all valid alignments reduced bias when using multi-species**
879 **references but it increased computation time and memory.** The plots show the run time (left) and memory usage (right) for
880 each alignment tool. Bowtie2 and Minimap2 were run using the standard mode and "all" mode which forced them to search for
881 and report all alignments. These benchmarks were calculated by running 2 million simulated reads against a toy reference index
882 made up of 961 *Staphylococcus aureus* genomes and 1 *Staphylococcus epidermidis* genome (Table 1) and are the average of 3
883 runs each.

884 **Figure 4. MTSv generally has high purity, accuracy, and sensitivity when assigning reads with known sources.** The CAMI
885 human microbiome datasets (gastrointestinal, airway, oral cavity, skin, and urogenital) were used to compare the taxonomic binning
886 performance of MTSv, Centrifuge, and Kraken2 at each major taxonomic rank. Sample purity (first row) is the sum of all true
887 positive sequences in the sample divided by the total number of sequences assigned. Average purity (second row) is the precision
888 calculated for each taxonomic bin averaged across all bins. Sample accuracy (third row) is the sum of true positive sequences
889 divided by the total number of sequences (including those that were unassigned). The average completeness (fourth row) is the
890 sensitivity of each taxonomic bin averaged across all bins. For both average purity and average completeness, small bins have the
891 same weight as large bins whereas larger bins contribute greater weight in the sample-based metrics.

892 **Figure 5. MTSv classifies a large proportion of reads at the species level and created fewer and smaller false positive bins**
893 **than both Kraken2 and Centrifuge.** The top row shows the percent of the total reads that were classified at each taxonomic rank
894 for each sample. The second row shows the number of false positive bins assigned by each tool and the bottom row is the median
895 size of the false bins ($\log_{10}$ scale).

896 **Figure 6. MTSv assigns more reads from genomes that diverge from reference sequences.** Simulated genomes of
897 *Faecalibacterium prausnitzii* that diverged by 0-15% (100%-85% similarity) from the reference sequence were used to simulate
898 two million reads each that were spiked into the CAMI gastrointestinal dataset. The left panel shows the number of reads from *F.*
899 *prausnitzii* that were assigned correctly (although not necessarily uniquely) by each tool as the genomes diverged. The middle
900 panel is the proportion of reads assigned to *F. prausnitzii* at the species level out of the total number of reads in the CAMI-
901 spikein dataset. The last panel shows how *F. prausnitzii* ranked in read abundance compared to other species in the sample. The
902 grey dashed line in each plot represents the ground truth. MTSv was run using both default and sensitive parameters.

903 **Figure 7. MTSv showed relatively little change in the number of assignments to near-neighbor species in the spike-in**
904 **sample compared to the soil background.** The heat map shows the difference ($\log_2$ ratio of reads per million) in species-level
905 assignments within the *Bacillus* genus between the soil background and the *B. anthracis* spike-in samples using MTSv,
906 Centrifuge, and Kraken2. The heatmap shows only rows where at least one of the columns had a value greater than one indicating
907 an increase in assignments in the spike-in compared to the soil background. A pseudocount of 1 was added to the sample and
908 background counts to avoid undefined ratios.

Standard read alignment approaches provide incorrect and biased assignments when using multi-species references.

We used two different toy reference indices that were each dominated by a single species and simulated 1 million reads from a single reference for each species. Bowtie2 and Minimap2 were run using the standard mode and "all" mode which forced them to search for and report all alignments. The table reports the number (and percent) of missed alignments (reads simulated from a species that did not align back to the same species).
1 **Table 1. Standard read alignment approaches provide incorrect and biased assignments when using multi-species**
2 **references.** We used two different toy reference indices that were each dominated by a single species and simulated 1 million
3 reads from a single reference for each species. Bowtie2 and Minimap2 were run using the standard mode and "all" mode which
4 forced them to search for and report all alignments. The table reports the number (and percent) of missed alignments (reads
5 simulated from a species that did not align back to the same species).

Species	Index Reference Sequences	Number of Missed Assignments				
		MTSv	Bowtie2 v2.4.5		Minimap2 v2.24	
			Standard	All	Standard	All
Bacillus anthracis	116	0 (0%)	254 (0.03%)	0 (0%)	10 (0.001%)	5 (0%)
Bacillus cereus	1	0 (0%)	**190,759 (19%)**	0 (0%)	**12,082 (1.2%)**	105 (0.01%)
Bacillus thuringiensis	1	0 (0%)	**172,807 (17%)**	0 (0%)	**9,225 (1%)**	10 (0.001%)
Bacillus subtilis	1	0 (0%)	316 (0.03%)	0 (0%)	1 (0%)	1 (0%)
Staphylococcus aureus	961	0 (0%)	28 (0%)	0 (0%)	0 (0%)	0 (0%)
Staphylococcus epidermidis	1	0 (0%)	**18,532 (2%)**	0 (0%)	**10,791 (1.2%)**	47 (0%)

MTSv assigned the most reads to the spike-in pathogen at the species level and had fewer assignments to near-neighbor species.

The table shows the results from the *B. anthracis*, *F. tularensis*, and *Y. pestis* spike-in samples using MTSv, Centrifuge, and Kraken2. The table shows the number of reads that were assigned uniquely to the spike-in species, the number of reads that aligned to the spike-in genus, the total reads for both the genus and species columns, the number of reads that were uniquely assigned to a non-spike-in species within the genome (and the number of species), and the total number of assigned reads. The value in parentheses is the percent of total reads. Kraken2 provided assignments at the *Bacillus cereus* Group level, but these were counted as genus assignments.

1 **Table 2. MTSv assigned the most reads to the spike-in pathogen at the species level and had fewer assignments to near-**
2 **neighbor species.** The table shows the results from the *B. anthracis, F. tularensis,* and *Y. pestis* spike-in samples using MTSv,
3 Centrifuge, and Kraken2. The table shows the number of reads that were assigned uniquely to the spike-in species, the number of
4 reads that aligned to the spike-in genus, the total reads for both the genus and species columns, the number of reads that were
5 uniquely assigned to a non-spike-in species within the genome (and the number of species), and the total number of assigned
6 reads. The value in parentheses is the percent of total reads. Kraken2 provided assignments at the *Bacillus cereus* Group level,
7 but these were counted as genus assignments.

	No. Reads	Tool	Reads assigned to spike-in species	Reads assigned to spike-in genus	Combined	Reads assigned to other species within genus (No. Unique Species)	Total Assigned Reads
B. anthracis 2% spike-in	6,353,624	MTSv	54,846 (0.86%)	199,560 (3.1%)	254,406 (4.0%)	687 (51) (0.01%)	384,021 (6.0%)
		Centrifuge	53,443 (0.84%)	195,056 (3.07%)	248,499 (3.91%)	2,443 (80) (0.04%)	1,769,907 (27.9%)
		Kraken2	43,596 (0.69%)	208,798 (3.29%)	252,394 (3.97%)	2,838 (77) (0.04%)	963,480 (15.2%)
F. tularensis 2% spike-in	7,778,056	MTSv	100,492 (1.29%)	1,246 (0.02%)	101,738 (1.31%)	31 (4) (0%)	268,019 (3.4%)
		Centrifuge	98,382 (1.26%)	776 (0.01%)	99,158 (1.27%)	63 (11) (0%)	2,045,753 (26.3%)
		Kraken2	100,104 (1.29%)	1,621 (0.02%)	101,725 (1.31%)	23 (8) (0%)	1,009,429 (12.98%)
Y. pestis 2% spike-in	6,397,098	MTSv	30,966 (0.48%)	112,008 (1.75%)	142,974 (2.2%)	280 (6) (0%)	280,577 (4.6%)
		Centrifuge	29,688 (0.46%)	110,306 (1.72%)	139,994 (2.19%)	420 (14) (0.01%)	1,651,364 (25.8%)
		Kraken2	26,345 (0.41%)	116,199 (1.82%)	142,544 (2.23%)	449 (11) (0.01%)	857,922 (13.4%)

Resources required to build MTSv MG-indices.

The top row of the table shows the resources required to build the MTSv indices used throughout this paper. The second row shows an alternative approach that reduces the amount of memory required for building the indices by splitting the reference sequences into smaller files. The alternative method also used a higher sampling rate of the BWT occurrence array (512 vs. 64) which resulted in a smaller index size. The last two rows show the resources required to build the data structures for Centrifuge and Kraken2.
1 **Table 3. Resources required to build MTSv MG-indices.** The top row of the table shows the resources required to build the
2 MTSv indices used throughout this paper. The second row shows an alternative approach that reduces the amount of memory
3 required for building the indices by splitting the reference sequences into smaller files. The alternative method also used a higher
4 sampling rate of the BWT occurrence array (512 vs. 64) which resulted in a smaller index size. The last two rows show the
5 resources required to build the data structures for Centrifuge and Kraken2.

Tool	CPUs	Wall Clock Time (Avg. H:M:S per job)	Max RAM (Avg. GB per job)	Storage Size on Disk (vs. original FASTA size)
MTSv	1 x 13 jobs	01:23:07 x 13 = 18:00:13	251.0	447.3G (~3.6x)
MTSv (Alt.)	1 x 66 jobs	00:17:44 x 66 = 19:30:24	49.0	298.8G (~2.3x)
Centrifuge v1.0.4	32 x 1 job	13:33:32	316.9	56.5G (~0.5x)
Kraken2 v2.1.2	32 x 1 job	3:55:14	33.5	35.2G (~0.3x)

Resources required to run taxonomic assignments.

The table shows the amount of time and memory requirements to run each tool with 8 threads per job. MTSv was run with 13 different jobs whereas Centrifuge and Kraken2 had only a single job each.

1 **Table 1. Resources required to run taxonomic assignments.** The table shows the amount of time and memory requirements to
2 run each tool with 8 threads per job. MTSv was run with 13 different jobs whereas Centrifuge and Kraken2 had only a single job
3 each.

Tool	CPUs	Wall Clock Time (Avg. H:M:S per job)	Max RAM (Avg. GB per job)
MTSv	8 x 13 jobs	29:51 x 13 = 6:28:07	34
Centrifuge v1.0.4	8 x 1 job	29:59	55
Kraken2 v2.1.2	8 x 1 job	1:11	29

Figure 1

Diagram of MTSv index setup and assignment pipeline.

To build the MG-Index, reference sequences are combined into a reference FASTA file formatted with the TaxID in the header. MTSv breaks up large reference files into smaller chunks with a user defined size and builds an MG-Index from each chunk. For taxonomic assignment, reads are aligned against each of the MG-Indices. The binning outputs from each MG-Index are merged to provide the final taxonomic assignments.

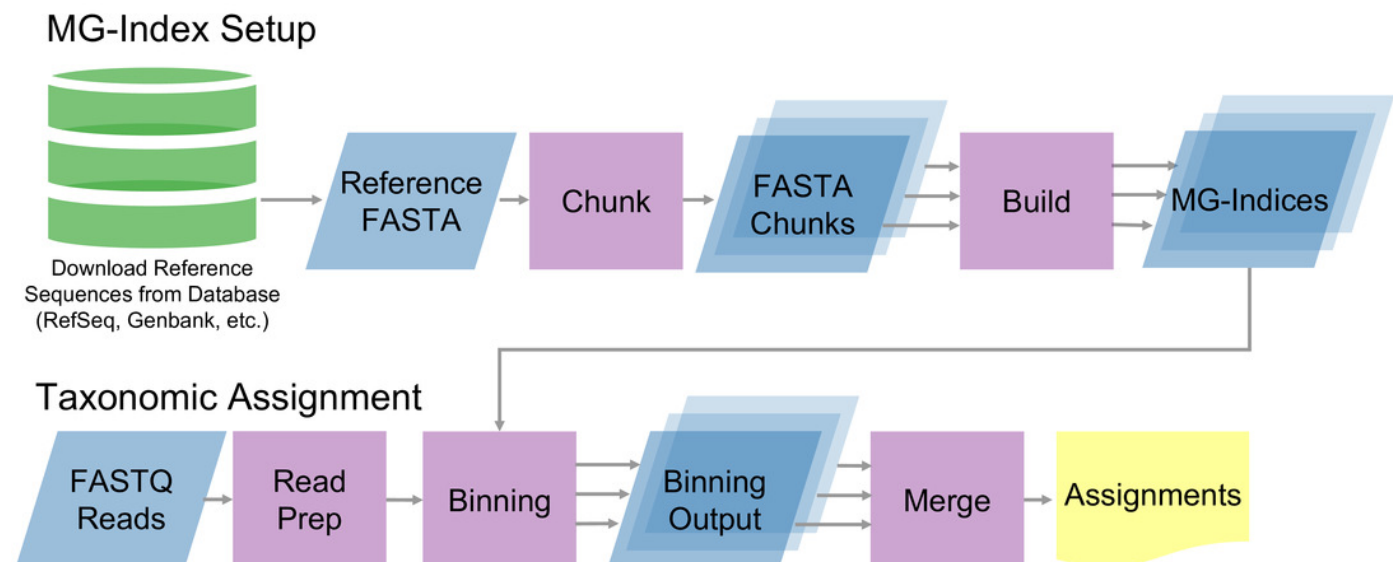



Figure 2

MTSv alignment algorithm.

Reads are broken into seeds of a user-specified size (seed-size) that are separated by a user-specified offset (seed-interval). Using the MG-index, the positions of exact seed hits are found and recorded. Regions that have at least a user-specified minimum number of seed hits (min-seed) are marked as candidate regions. The candidate regions are first sorted (descending) by the number of seed hits to ensure that the most promising alignments are checked first. The candidate region sequence and the query are aligned using a SIMD-accelerated Smith-Waterman algorithm and the alignment is considered successful if the edit distance of the alignment is less than or equal to the user-specified minimum (edit-rate). Once a successful alignment is found for a TaxID, no further candidates will be aligned for that TaxID.

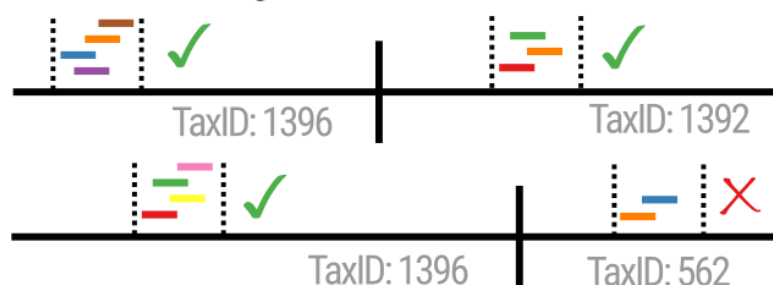
Extract seeds from read

Seed Gap
Seed Size

Look up seed matches in MG-Index

Pick candidate regions with at least 2 matches

TaxID: 1396   TaxID: 1392

TaxID: 1396   TaxID: 562

Sort candidate regions by number of seeds

To check the most promising alignments first

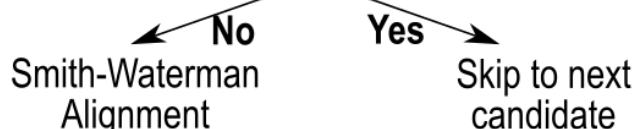
TaxID	Seeds
1396	4
1396	4
1392	3

Alignment

For each candidate:

Has a successful alignment to candidateTaxID been found for this query?

No ← → **Yes**

Smith-Waterman Alignment   Skip to next candidate

Within edit distance tolerance?

No   **Yes**

Hit   **No Hit**

Figure 3

Forcing traditional alignment tools to report all valid alignments reduced bias when using multi-species references but it increased computation time and memory.

The plots show the run time (left) and memory usage (right) for each alignment tool. Bowtie2 and Minimap2 were run using the standard mode and "all" mode which forced them to search for and report all alignments. These benchmarks were calculated by running 2 million simulated reads against a toy reference index made up of 961 *Staphylococcus aureus* genomes and 1 *Staphylococcus epidermidis* genome (Table 1) and are the average of 3 runs each.

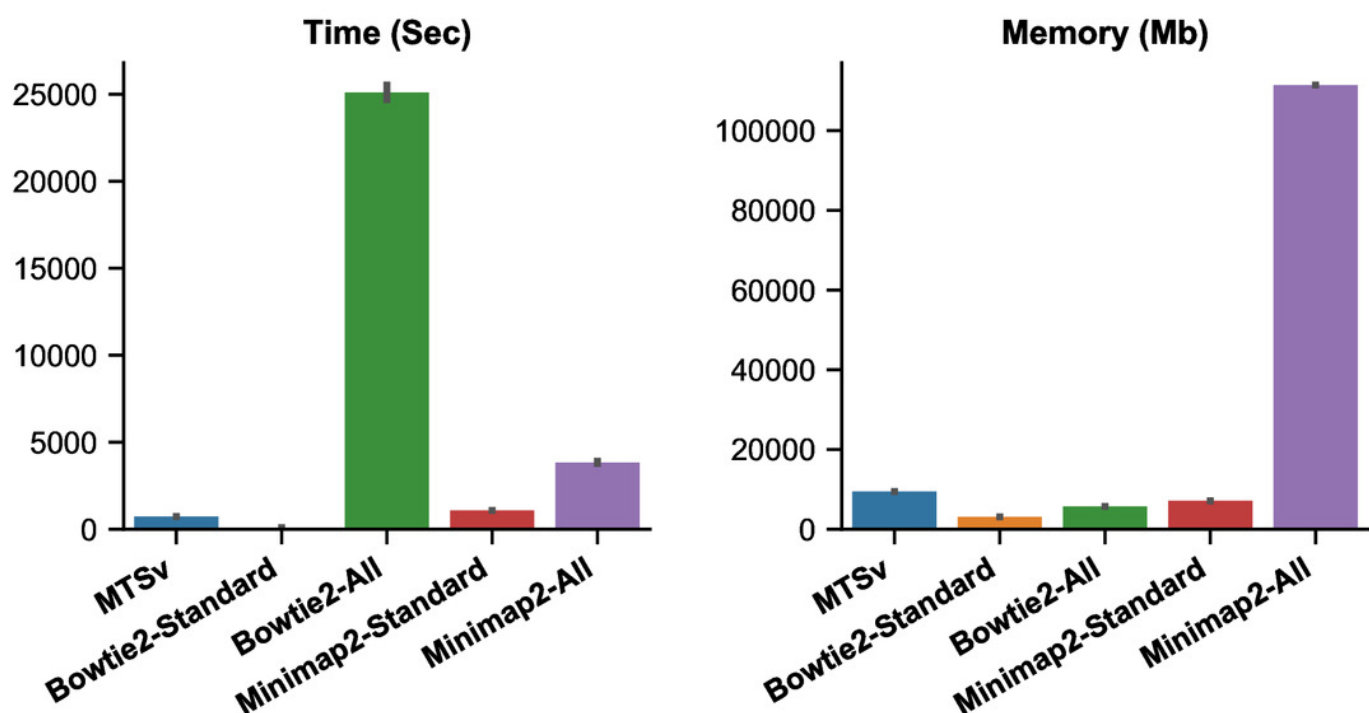



Figure 4

MTSv generally has high purity, accuracy, and sensitivity when assigning reads with known sources.

The CAMI human microbiome datasets (gastrointestinal, airway, oral cavity, skin, and urogenital) were used to compare the taxonomic binning performance of MTSv, Centrifuge, and Kraken2 at each major taxonomic rank. Sample purity (first row) is the sum of all true positive sequences in the sample divided by the total number of sequences assigned. Average purity (second row) is the precision calculated for each taxonomic bin averaged across all bins. Sample accuracy (third row) is the sum of true positive sequences divided by the total number of sequences (including those that were unassigned). The average completeness (fourth row) is the sensitivity of each taxonomic bin averaged across all bins. For both average purity and average completeness, small bins have the same weight as large bins whereas larger bins contribute greater weight in the sample-based metrics.

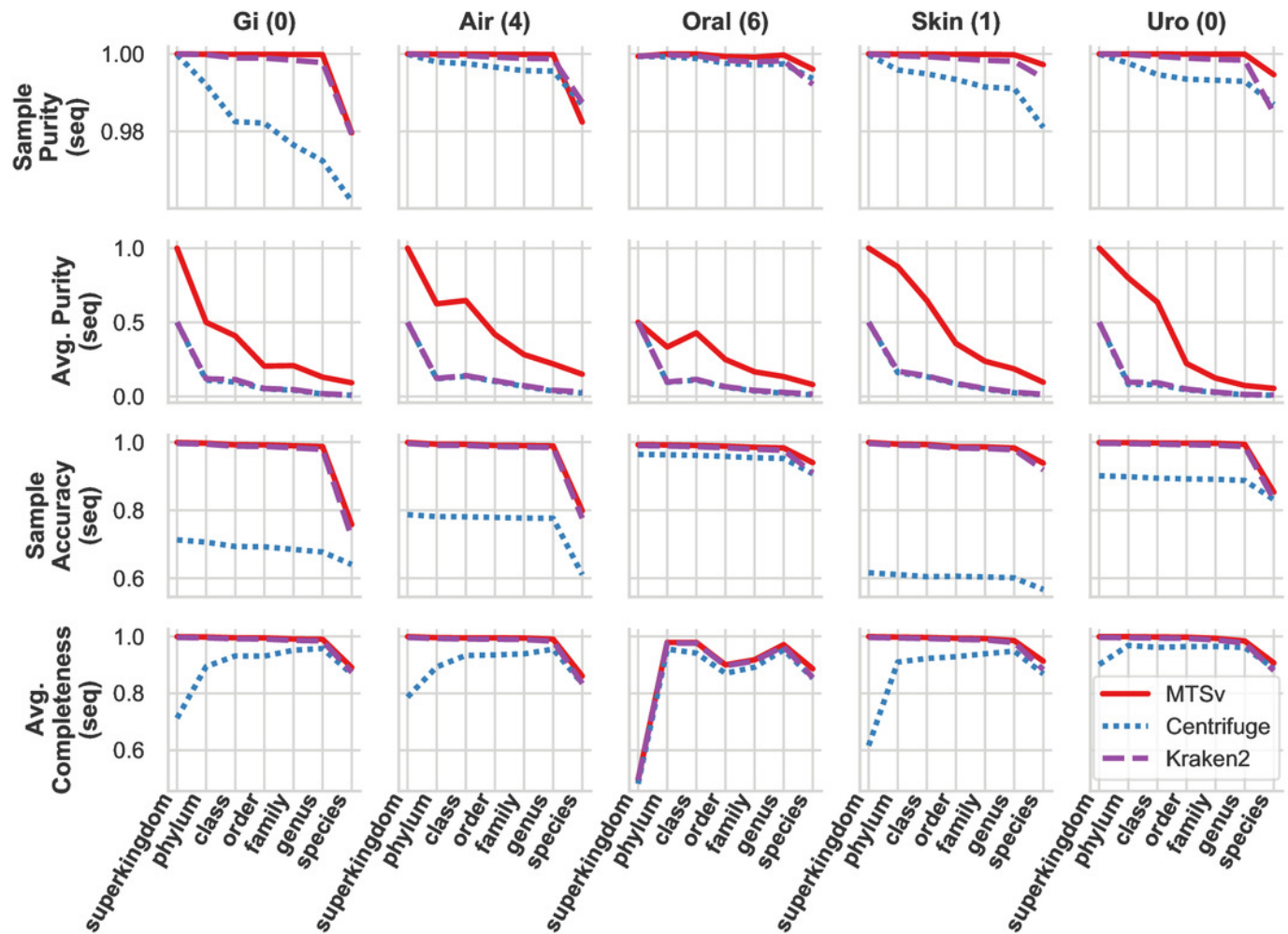



Figure 5

MTSv classifies a large proportion of reads at the species level and created fewer and smaller false positive bins than both Kraken2 and Centrifuge.

The top row shows the percent of the total reads that were classified at each taxonomic rank for each sample. The second row shows the number of false positive bins assigned by each tool and the bottom row is the median size of the false bins ($\log_{10}$ scale).

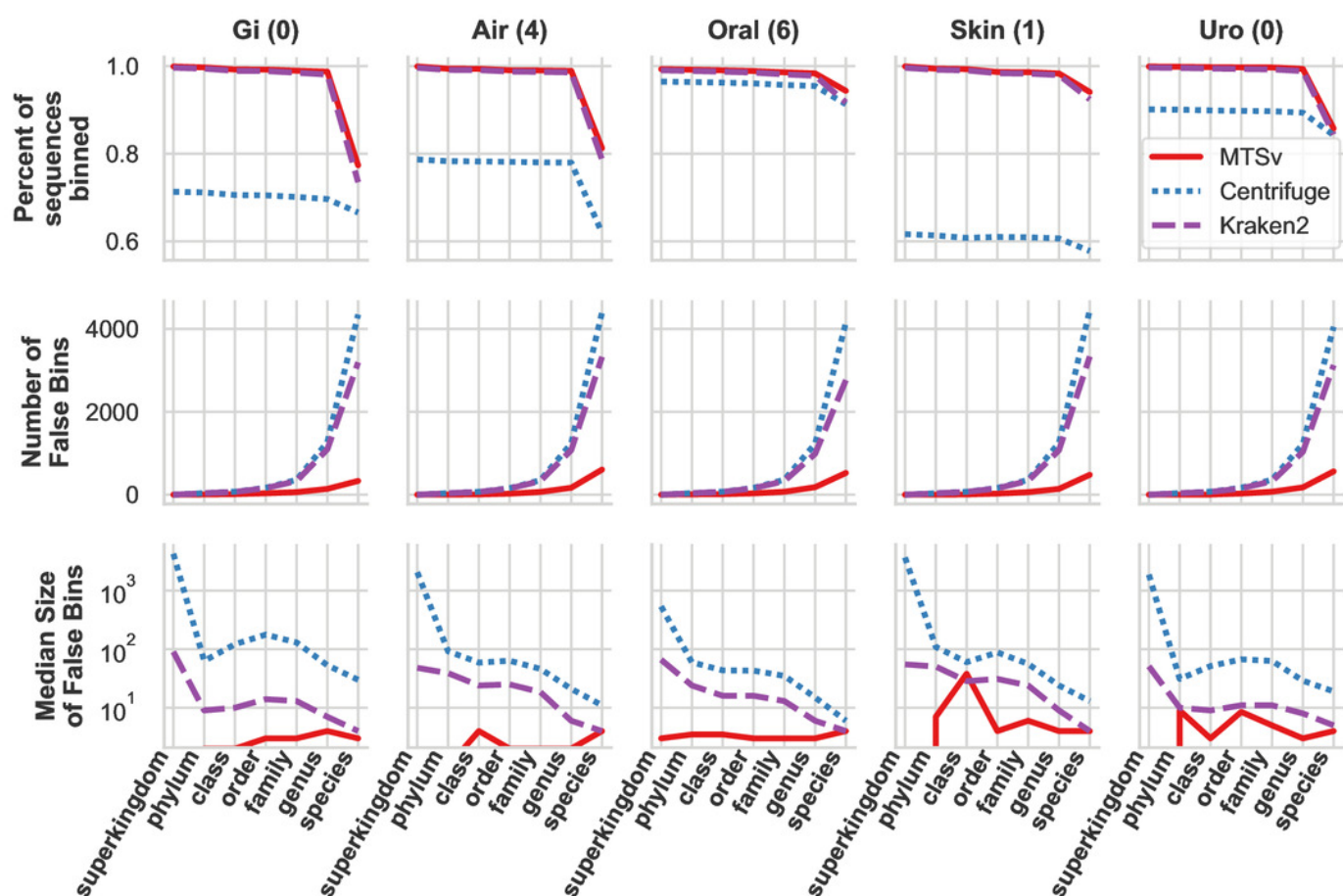



Figure 6

MTSv assigns more reads from genomes that diverge from reference sequences.

Simulated genomes of *Faecalibacterium prausnitzii* that diverged by 0-15% (100%-85% similarity) from the reference sequence were used to simulate two million reads each that were spiked into the CAMI gastrointestinal dataset. The left panel shows the number of reads from *F. prausnitzii* that were assigned correctly (although not necessarily uniquely) by each tool as the genomes diverged. The middle panel is the proportion of reads assigned to *F. prausnitzii* at the species level out of the total number of reads in the CAMI-spikein dataset. The last panel shows how *F. prausnitzii* ranked in read abundance compared to other species in the sample. The grey dashed line in each plot represents the ground truth. MTSv was run using both default and sensitive parameters.

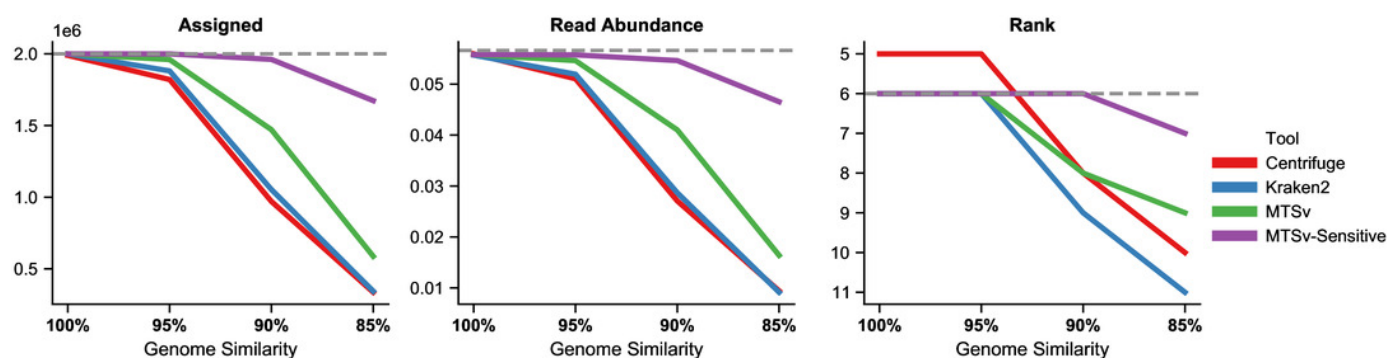



Figure 7

MTSv showed relatively little change in the number of assignments to near-neighbor species in the spike-in sample compared to the soil background.

The heat map shows the difference ($\log_2$ ratio of reads per million) in species-level assignments within the *Bacillus* genus between the soil background and the *B. anthracis* spike-in samples using MTSv, Centrifuge, and Kraken2. The heatmap shows only rows where at least one of the columns had a value greater than one indicating an increase in assignments in the spike-in compared to the soil background. A pseudocount of 1 was added to the sample and background counts to avoid undefined ratios.

	MTSv	Kraken2	Kraken2 Filtered	Centrifuge	Centrifuge Filtered
Bacillus anthracis	12.18	11.41	11.87	11.54	12.00
Bacillus wiedmannii	0.71	4.01	2.00	1.12	1.20
Bacillus thuringiensis	2.32	3.77	3.22	0.96	2.14
Bacillus cereus	1.86	3.77	2.62	1.28	2.13
Bacillus mobilis	0.26	3.56	1.68	1.32	1.32
Bacillus albus	0.82	3.18	1.48	0.47	0.71
Bacillus mycoides	-0.00	2.77	0.53	0.14	0.81
Bacillus sp. ABP14	0.84	2.47	1.45	1.35	1.28
Bacillus paranthracis	0.34	2.39	0.56	0.47	1.02
Bacillus sp. HBCD-sjtu	1.48	2.35	1.18	1.48	1.53
Bacillus cytotoxicus	-0.36	2.07	0.14	-0.19	-0.34
Bacillus pseudomycoides	-0.32	1.71	-0.13	-0.60	-0.36
Bacillus bombysepticus	-0.13	1.36	0.39	0.19	0.39
Bacillus sp. SYJ	0.96	1.14	0.84	1.04	1.53
Bacillus sp. 1NLA3E	0.21	0.66	1.06	0.36	0.49