




# MetaFast: Enabling Fast Metagenomic Classification via Seed Counting and Edit Distance Approximation

**Working Paper****Author(s):**

[Gollwitzer, Arvid](#) ; Alser, Mohammed; Bergtholdt, Joel; [Lindegger, Joël](#) ; Rumpf, Maximilian-David; [Firtina, Can](#) ; Mangul, Serghei; Mutlu, Onur

**Publication date:**

2023-11-03

**Permanent link:**

<https://doi.org/10.3929/ethz-b-000642547>

**Rights / license:**

[Creative Commons Attribution-ShareAlike 4.0 International](#)

**Originally published in:**

arXiv, <https://doi.org/10.48550/ARXIV.2311.02029>

# MetaFast: Enabling Fast Metagenomic Classification via Seed Counting and Edit Distance Approximation

Arvid E. Gollwitzer<sup>2</sup>, Mohammed Alser<sup>1,2</sup>, Joel Bergtholdt<sup>2</sup>, Joël Lindegger<sup>2</sup>,  
Maximilian-David Rumpf<sup>1</sup>, Can Firtina<sup>2</sup>, Serghei Mangul<sup>3</sup>, Onur Mutlu<sup>1,2</sup>

<sup>1</sup>Department of Computer Science, ETH Zürich, 8092 Zürich, Switzerland

<sup>2</sup>Department of Information Technology and Electrical Engineering, ETH Zürich, 8092 Zürich, Switzerland

<sup>3</sup>Department of Clinical Pharmacy, University of Southern California, Los Angeles, CA, 90089, USA

## Abstract

Metagenomics, the study of genome sequences of diverse organisms cohabiting in a shared environment, has experienced significant advancements across various medical and biological fields. Metagenomic analysis is crucial, for instance, in clinical applications such as infectious disease screening and the diagnosis and early detection of diseases such as cancer. A key task in metagenomics is to determine the species present in a sample and their relative abundances. Currently, the field is dominated by either alignment-based tools, which offer high accuracy but are computationally expensive, or alignment-free tools, which are fast but lack the needed accuracy for many applications.

In response to this dichotomy, we introduce *MetaFast*, a tool based on heuristics, to achieve a fundamental improvement in accuracy-runtime tradeoff over existing methods. *MetaFast* delivers accuracy comparable to the alignment-based and highly accurate tool Metalign but with significantly enhanced efficiency. In *MetaFast*, we accelerate memory-frugal reference database indexing and filtering. We further employ heuristics to accelerate read mapping. Our evaluation demonstrates that *MetaFast* achieves a 3.9x speedup over Metalign without compromising accuracy. *MetaFast* is publicly available on: <https://github.com/CMU-SAFARI/MetaFast>.

# Introduction

Metagenomics diverges from traditional genomics by analyzing genome sequences from a diverse community of organisms inhabiting a common environment instead of studying the genome sequences of an individual or a small group of individuals of the same species. This comparison of genome sequences in metagenomic samples against large databases of genomes, known as reference genomes, leads to significant biological findings<sup>1,2</sup>. We outline the four key initial steps in a standard genome sequencing and analysis pipeline: 1) Genomic sequencing data is obtained through sequencing a new sample<sup>3,4</sup>. 2) Basecalling<sup>5</sup> procedures convert raw sequencing data into nucleotides A, C, G, and T in the DNA alphabet. 3) A quality control step<sup>6</sup> removes low-quality subsequences of a read or an entire read sequence. 4) Computational metagenomic analysis lists all taxa in the sample and their corresponding relative abundance levels<sup>7</sup>.

The transformative role of sequencing genomic molecules is stimulating research and development in clinical applications and life sciences. This transformation is due to the successful sequencing of the human genome and the advent of high-throughput sequencing technologies, which have substantially reduced the cost of DNA sequencing. Consequently, bioinformatics has developed many software tools to leverage increasingly large and complex sequencing read sets. These tools have triggered progress in modern biology and become essential to clinical life sciences. For instance, metagenomics has enabled advancements in precision medicine<sup>1,8</sup>, understanding microbial diversity, and the early detection of diseases<sup>9</sup>. All this creates a need for faster and more efficient computational tools.

The increase in available metagenomic HTS datasets<sup>10</sup> has prompted the development of many taxonomic classification and abundance estimation methods, as evidenced by a recent benchmarking study involving a dataset established by the Critical Assessment of Metagenome Interpretation (CAMI) challenge<sup>11</sup>. CAMI covers 20 taxonomic classifiers, including both alignment-based approaches such as GATK<sup>12</sup>, PathSeq<sup>19</sup>, MetaPhlan<sup>13</sup>, and Metalign<sup>14</sup>, and alignment-free approaches such as Kraken2<sup>15</sup>, CLARK<sup>16</sup>, KrakenUniq<sup>17</sup> and Centrifuge<sup>18</sup>. Early approaches for analyzing metagenomic sequencing data were alignment-based and used a reference database. However, the growth of HTS data and reference databases has made read search and alignment based on large databases computationally infeasible. On the other hand, alignment-free tools are less accurate than their alignment-based counterparts.

We consider a representative example detailed in the latest CAMI study: mOTUs<sup>19</sup>, a highly accurate tool, requires four hours to perform taxonomic classification on the same marine (TARA Ocean<sup>20</sup>) dataset we consider in our analysis. This lengthy processing time renders it inapplicable for time-critical or routine procedures. Conversely, Kraken2<sup>21</sup> completes the task in less than six minutes but delivers a very low classification accuracy, with a high false-positive rate. Its F1 score<sup>22</sup> at the species level is only 0.03, an accuracy level that falls short for most medical applications<sup>23</sup>.

Accelerating metagenomic analysis is critical for three reasons:

1. The sequencing and basecalling steps for a sample read set are one-time tasks in most cases, while the reads from a single sequenced sample can be analyzed by multiple studies or at different times in the same study<sup>24</sup>.
2. The analysis throughput is significantly lower than the throughput with which modern sequencing machines generate data. Sequencing throughput is expected to increase even further in the future.<sup>25</sup> As an example, Illumina NovaSeq X Plus systems<sup>26</sup>, considering 10B Flow Cells, generate ~1 Tb of data per run (2 × 50 bp). This amounts to 20 billion reads (paired-end) passing filtering per flow cell. Our analysis, conducted with Metalign<sup>14</sup>, a state-of-the-art metagenomic analysis tool, reveals that data analysis sequenced and basecalled by a high-throughput sequencer<sup>12</sup> in 48 hours lasts 38 days on a high-end server node. Such long analysis times present significant challenges for time-critical metagenomic use cases, including urgent clinical settings and outbreak tracing<sup>27</sup>
3. A single sequencing machine can concurrently sequence numerous samples from diverse sources, such as different patients or environments, thereby achieving remarkable throughput.
4. Furthermore, the extensive computational resources needed for metagenomic analysis make routine screening processes (i.e., for early cancer detection) practically inaccessible to the general public.
5. Sequencing technologies<sup>28</sup> that allow analysis during sequencing<sup>29</sup> further underscore the critical importance of fast metagenomic analysis.

We discuss the role of edit distance approximation methods in read mapping to quickly examine the similarity for every read sequence and potentially matching segments in the reference genome identified during seeding. Traditionally, the mapper performs computationally expensive sequence alignment to determine whether the remaining sequence pairs that pass the filter are similar. We observe that sequence alignment yields data not essential for metagenomic profiling, including the optimal number of edits, their precise locations, and the optimal arrangement of these edits<sup>30,31</sup>. These unnecessary computations waste compute cycles and energy. This necessitates performing the analysis on energy-intensive high-performance computing platforms that incur high costs and are unavailable in remote areas. Thus, many high-throughput applications like disease screening fall short of the possibilities enabled by metagenomic analysis.

We introduce MetaFast, a heuristic computational taxonomic classification approach based on Metalign<sup>14</sup>. Metalign is a state-of-the-art, highly accurate alignment-based tool for metagenomic analysis. We maintain the same level of accuracy as Metalign<sup>14</sup> but, through introducing heuristics, significantly enhance the speed of the analysis pipeline. We make four key contributions to the acceleration of the taxonomic profiling of metagenomic samples:

1. We develop a memory-frugal reference database index structure that enables rapid reference database prefiltering, i.e., containment search. We achieve a twofold reduction in the memory footprint and a 4x speedup over Metalign's reference database filtering procedure.
2. We accelerate the metagenomic read mapping phase by relying on heuristic methods for edit-distance approximation that provide close-to-optimal solutions significantly faster. Our heuristic and alignment-free read mapper delivers a fourfold runtime reduction benchmarked against minimap2.
3. We conduct a rigorous experimental evaluation to examine MetaFast's speed and accuracy. This involves comprehensively benchmarking our application against Metalign<sup>14</sup> to quantify our achieved reduction in execution time and memory footprint.
4. We introduce a novel metagenomic airlifting<sup>32</sup> mechanism that enables swift adaptation of a taxonomic profile following changes to the reference database. Our mechanism allows for adapting the taxonomic profile (subject to a single modification of the reference database) nine times faster than a complete rerun of MetaFast.

We confidently posit that MetaFast will catalyze further advancements in metagenomic research.

# Results

We initially present a brief overview of the MetaFast algorithm. Then, we outline the organization of the reference database and index generation. We perform metagenomic analysis on simulated and real data sets to benchmark MetaFast against Metalign<sup>14</sup>. This evaluation includes accuracy metrics for the benchmarked tools and their computational resource usage.

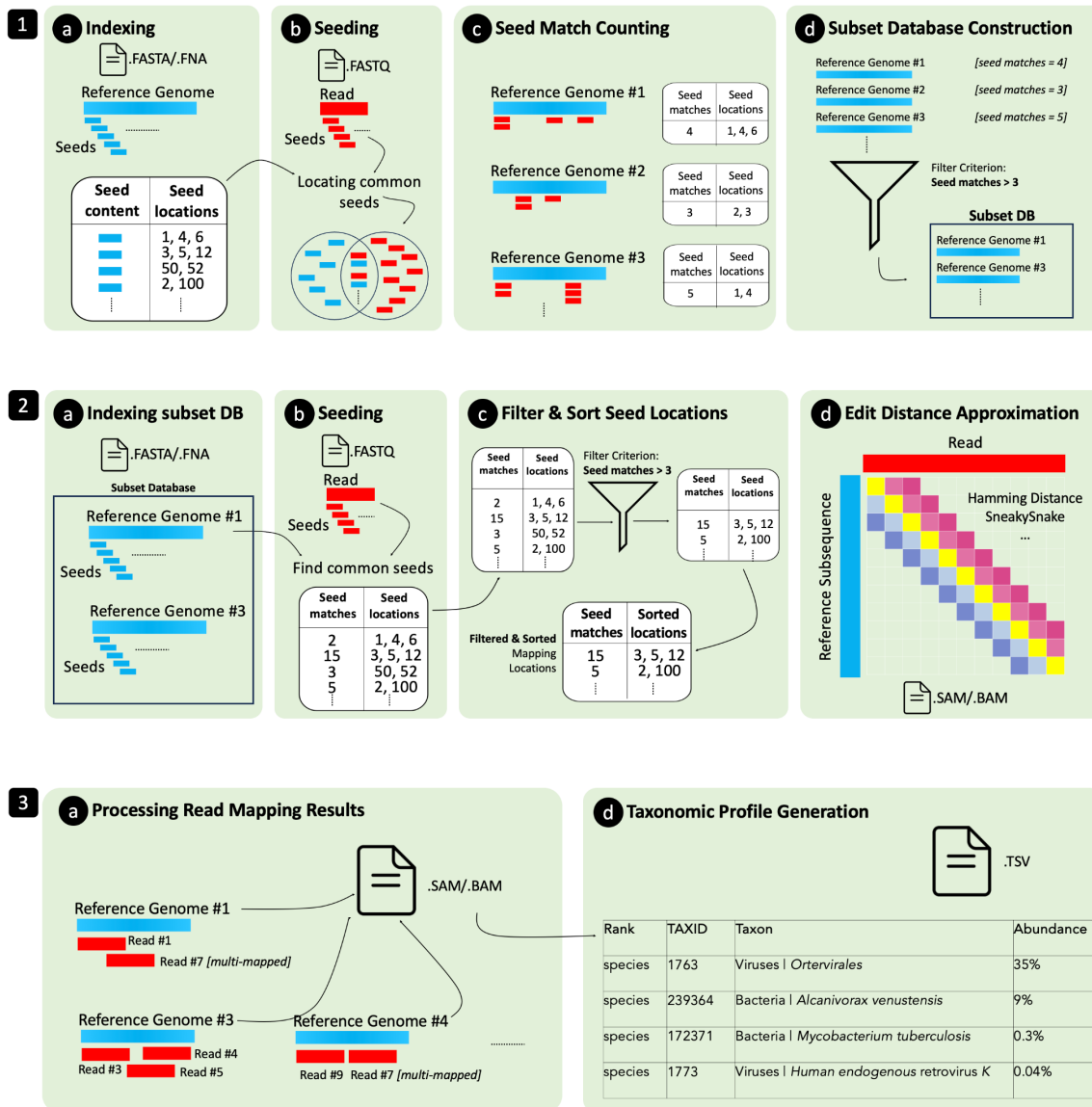
## Methodology Overview

### The MetaFast Pipeline

MetaFast performs metagenomic classification in three stages: 1) Containment search, 2) read mapping, and 3) taxonomic profiling. Each pipeline stage contains one or more sequentially executed *filtering steps*. The first stage (containment search) aims to filter the reference database by generating a much smaller *subset database* from the reference genomes that are likely to be similar to many reads in the metagenomic sample. To quickly quantify the similarity between a set of reference genomes and a set of reads, we perform seeding and count the number of seeds in each reference genome that are also present in the read set. This provides an estimate of how likely it is for each reference genome to be present in the sample. The reference genomes with a number of seed hits above a certain threshold are then included in this subset database to reduce the unnecessary computations for analyzing highly dissimilar reference genomes.

In the second pipeline stage, we perform heuristic read mapping. The goal of our read mapping stage is to filter the metagenomic read set quickly and accurately in two steps. In the first read-filtering step, we examine the mapping locations of all reads and exclude candidate locations that do not achieve a minimum number of seed hits in the subset database. We exclude an entire read sequence if this read does not have at least one mapping location in the subset database with a number of seed hits above the threshold. After this step, we are left with a set of sequence pairs. Each sequence pair consists of a subsequence extracted from a specific mapping location in the subset database and a read sequence. The mapping regions in the subset database are determined based on the locations of seed hits. To accurately quantify the similarities for all the remaining sequence pairs, the second read filtering step uses heuristic algorithms to compute an approximate edit distance for each pair. Sequence pairs with an approximated edit distance above a user-defined cutoff threshold are filtered out from further analysis. We record the associated reference genome, mapping location, and edit distance for all remaining reads.

Based on the read mapping results, we perform taxonomic profiling and relative abundance estimation in the third stage. In this pipeline stage, we perform only one filtering step: we reduce the false positive rate and improve classification accuracy by excluding all organisms with relative abundance estimates below a user-defined cutoff threshold from the final taxonomic profile.



**Figure 1** Overview of the complete MetaFast pipeline.

1) MetaFast first uses Containment Search to build a small subset database. This database filtering stage begins with **a) Indexing**: MetaFast takes sequencing reads and a reference database as input. First, MetaFast collects seeds from all reference genomes in the database. **b) Seeding**: MetaFast extracts seeds from the input read set. **c) Seed Match Counting**: Our memory-frugal methodology quickly estimates the number of seeds in each reference genome that also exist in the reads. **d) Subset Database Construction**: MetaFast selects a significantly smaller subset database comprising reference genomes surpassing a certain seed hit count and builds a subset DB based on these genomes.

2) MetaFast performs Heuristic Read Mapping, beginning with **a) Querying the subset database**: MetaFast loads the seeds from the small subset database created during containment search. **b)** We extract seeds from the read set and determine candidate mapping locations for each read by examining seeds simultaneously contained in the read set and the subset database, i.e., seed matches. **c) Filtering and sorting mapping locations**: In the first filtering step, MetaFast dramatically reduces the workload for later analysis steps by discarding all read candidate mapping locations that do not achieve a minimum number (in our case, three) of seed matches in the subset database. We then sort all remaining mapping locations by their associated seed matches. **d) Edit distance approximation**: MetaFast uses heuristics to approximately compute the edit distance for all remaining reads and mapping locations. We rely on several heuristic methods (like SneakySnake, Hamming Distance, SHD..., etc.) to perform read mapping. All reads with a minimum edit distance exceeding 10% of the read length are discarded.

3) We present the taxonomic profiling results in human-readable form by **a) Streamlined processing of mapping results:** MetaFast streams in the read mapping results as they are generated and computes the absence/presence and relative abundances of taxa on the fly. We consider reads uniquely mapped to a single reference genome, as well as multi-mapped reads. MetaFast examines the number of reads mapped to each reference genome and their respective edit distances to quantify relative abundance levels. **b) Taxonomic profile generation:** We discard all taxa with relative abundance levels below 0.01%. MetaFast then provides a taxonomic profile in the standardized format used by OPAL and CAMI as the final output.

## Containment Search

Containment search measures the similarity between two genomic datasets by calculating the k-mer intersection between their respective k-mer sets. This approach proves useful for key metagenomic applications, such as identifying a small number of relevant organisms potentially present in the metagenomic sample. Containment search commonly involves building a containment index for a set of genomes to determine the intersection of k-mers between some dataset and the index.

Metalign<sup>14</sup> employs KMC3<sup>33</sup> and CMash<sup>34</sup> for selecting relevant reference genomes from a large database. CMash is a state-of-the-art hashing-based approach and uses a k-mer-based ternary search tree<sup>35</sup> (TST) to store variable k-mer sizes. Initially, KMC3 enumerates the k-mers in the reads and intersects these sets with the precomputed k-mers of the reference genomes. The containment MinHash similarity metric, i.e., CMash, then estimates the similarity or containment index between each reference genome and the input sample. The containment index, closely related to the Jaccard index<sup>36</sup>, refers to the share of k-mers in a reference genome that also exists in the reads. Metalign includes all reference genomes above a specific cutoff threshold in a new, reduced database for alignment. We observe that KMC3+CMash generate more than four times the size of the examined reference database as auxiliary data during index construction. Furthermore, during index querying, KMC3 produces auxiliary data of roughly the size of the input read set, and CMash shows large main memory requirements. We propose a memory frugal indexing and efficient seed match counting algorithm to replace both KMC3 and CMash.

## Evaluation Methodology

We initiate our analysis by examining the first two pipeline stages individually. We evaluate the computational resources of each stage. Specifically, we record the runtime and peak resident set size (RSS) for each phase: the first stage performs containment search and reference database filtering, while the second stage performs heuristic read mapping. Finally, we reintegrate the pipeline and analyze the entire system's accuracy. In the final step, we gauge the computational resources again, providing a comprehensive benchmarking strategy against Metalign.

We benchmark our containment search algorithm, directly comparing it with KMC3+CMash. We use 64 compute threads for each tool on a system with an AMD EPYC 7742 64-core processor, 1 TB of main memory, and an SSD<sup>37</sup> with a SATA3 interface. We record the elapsed wall clock time and the main memory footprint in all experiments using the `/usr/bin/time -v` command on Linux. To ensure standardization, we use default parameters for KMC3 and CMash. The runtime measurement



encapsulates both the querying process and the generation of the subset database. Since the reference database filtering procedure employed by Metalign comprises two tools, namely KMC3+CMash, we perform a runtime breakdown to highlight each tool's individual contribution. We record a peak main memory usage of 14 GB for KMC3, independent of the read set. Later, when considering the complete pipeline, we continue our evaluation, focusing on accuracy.

To construct a comprehensive reference database, we incorporate NCBI<sup>38</sup> microbial genome assemblies, encompassing complete and incomplete assemblies from RefSeq<sup>39</sup> and GenBank<sup>40</sup>. The final database comprises 19,807 organisms, amounting to a size of approximately 170 GB.

This large reference database forms the foundation for our genomic analysis. To enhance manageability and efficiency, we divide this database into  $N = 25$  batches. For each batch, we generate an individual index structure (MMI file). This offers numerous advantages, such as parallelized index access, which, as we will show, allows for considerable speedup. Each index (i.e., each MMI file) can be processed individually and independently in a multithreaded fashion. Moreover, our batch-based approach provides an efficient solution for handling updates to reference genomes. If a reference genome is updated, we only need to recreate the index for the affected batch. This localized update approach saves time and computational resources by eliminating the need to process the entire database. We then need to rerun MetaFast only for the affected batch.

We count the seed hits for each reference genome in each batch. The number of compute threads to process an individual batch can be freely chosen. This step quantifies the similarities and differences between the reference genomes and the metagenomic read set. We include all reference genomes that receive a number of seed hits (*seed-hit count*) above an empirically determined threshold in the subset database.

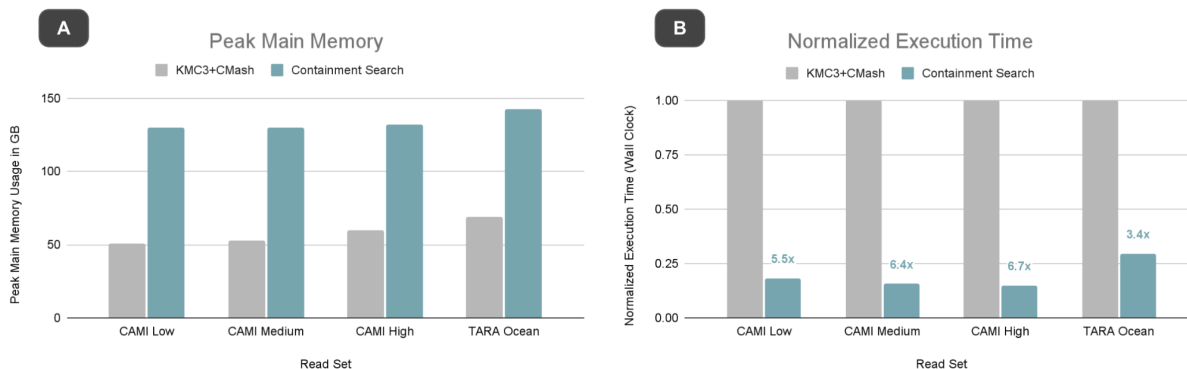
To provide the most comprehensive analysis possible, we examine three datasets from the Critical Assessment of Metagenome Interpretation (CAMI)<sup>11</sup>, one of each diversity class: a low-diversity sample with 99,796,358 reads (RL\_S001insert\_270.fq), medium-diversity with 99,776,814 reads (RM\_S001insert\_270.fq) and a high-diversity community (RH\_S001insert\_270.fq) with 99,811,870 reads. All CAMI datasets have a read length of 150. The CAMI paper<sup>11</sup> further details these communities. Of note, the CAMI communities include many organisms absent from the MetaFast reference database. For instance, for the CAMI high-complexity dataset, MetaFast's database only contains 184 out of the 243 unique species. We further underpin our accuracy and computational resource evaluation with a read set from the TARA Ocean Project<sup>20</sup> (ERR1700889\_1.fastq), accessed February 2023. TARA Ocean reads have a length of 100 base pairs.

We must choose several parameters for our heuristic database filtering, i.e., the containment search approach. As we restrict our considerations to short reads, we rely on minimap2's short read heuristics<sup>41</sup> (k-mer length  $k = 28$ ; window size  $w = 16$  as well as the `-x` and `--sr` presets), which we assume to satisfy our optimality criteria. Since we are only interested in determining the number of seed hits for each

reference genome, we construct all indices using minimap2's `-H` and `--idx-no-seq` options to speed up index construction, reduce storage space, and decrease the main memory footprint later during querying. At the end of our containment search stage, we include a reference genome in the subset database if it shares at least a minimum number of seed hits with the read set.

We aim to find the minimum number of matching seeds that ensures a zero false negative rate. To achieve this, we iterate over several read sets (CAMI low, medium, and high complexity datasets and a TARA Ocean read sample) and incrementally increase the minimum seed hit cutoff value if a false negative occurs. For each read set, we divide all seed hits by the highest observed number of hits for normalization. We test the subset database for false negatives by performing read mapping with minimap2, then generate the taxonomic profile with Metalign's profiling routine and examine the false negative rate (on the species level) using OPAL<sup>42</sup>. Eventually, we make a conservative choice, selecting the lowest possible threshold that still ensures a zero false negative rate across all read sets. In this way, we determine an optimal threshold of 0.0001.

## Seed counting enables fast candidate genome identification



**Figure 2** Memory footprint and execution time analysis of our containment search stage. We benchmark our memory frugal containment search algorithm against the tools KMC3 and CMash employed by Metalign.

**A** Peak main memory usage in GB during the reference database querying and reference genome selection stage.

**B** Normalized execution time (wall clock time) of our containment search stage. We benchmark against KMC3+CMash.

We present the runtime analysis results in *Figure 2 A-B*. We make four key observations:

1. The runtime of KMC3+CMash remains relatively constant for the three CAMI challenge datasets and does not directly scale with the dataset size. The execution time does, however, decrease for the TARA Ocean dataset with a shorter read length of 100.
2. The runtime for our seeding algorithm increases with an increasing number of reads.
3. Our containment search algorithm achieves a 5.5x - 3.4x reduction in runtime.

4. Our containment search algorithm shows a two-fold increase in peak main memory usage. The peak main memory of our seed counting approach remains relatively constant across all read sets. KMC3+CMash, on the other hand, has varying and more unpredictable memory requirements, with an approximately 10 GB increase for the TARA Ocean dataset, as seen in *Figure 2A*.

The relatively constant execution time of KMC3+CMash for CAMI datasets is due to the generally rapid detection of a small overlap of k-mers from the read and reference by KMC3. As a result, even large read sets, which are highly dissimilar to most reference genomes, can be processed swiftly. However, this proves to be practically insignificant, given that the presence of specific organisms in the sample is unpredictable in advance. In contrast, our seeding algorithm iterates over all reads to count seed hits. Thus, its runtime is directly linked to the number of reads in the dataset. We conclude that our containment search algorithm generally provides significant speedup over KMC3 and CMash.

The increased main memory footprint of our containment search methodology is related to the index size (in our case, the set of all MMI files). Given the same underlying set of reference genomes, our indices show approximately twice the storage space requirements compared to the index structures generated by KMC3+CMash. In both cases, the complete index is loaded into the main memory during index querying. We can easily reduce the main memory footprint of our containment search stage by processing batches in the reference database sequentially, which, in turn, leads to an increase in execution time. Based on available hardware resources, our containment search methodology may be, by merely choosing appropriate user arguments, configured to minimize either runtime or main memory usage.

We further analyze the index generation time of our seed-based, memory-frugal indexing methodology, comparing it against KMC+CMash on the same underlying reference data. In the initial step, we allocate 200 threads to each methodology. Our approach requires a five-minute execution time per batch, with a peak memory usage of 150 TB. Owing to the fully parallel execution, given the 200 available compute threads, the entire database is constructed within 15 minutes. On the other hand, KMC3+CMash requires 18 hours and a peak main memory of 300 GB for the same task. We repeat the procedure, this time limiting the thread count to 64 to accommodate realistic scenarios with limited hardware resources. Our method generates the index structure in 1.3 hours with a memory footprint of 90 GB. We were unable to wait for the completion of the index structure generation by KMC3+CMash, which, given the now limited thread count, slowed down dramatically. We terminated the process after 22 hours. In this second attempt, the incomplete KMC3+CMash database generation needed a peak main memory of 700 GB.

## Alignment-free read-mapping

Locating potential subsequences within the reference genome sequence that bears similarity to the read sequence while accommodating differences remains a computationally intensive task. Many researchers aim to address this problem<sup>43</sup> by employing new algorithms, hardware accelerators<sup>44</sup>, and hardware/software codesign<sup>45</sup>.

Minimap2<sup>41</sup>, our baseline, is a state-of-the-art read mapper that effectively maps nearly all existing sequencing read types, including short, ultralong, and accurate long reads. The operation of minimap2 covers four computational steps: 1) indexing, 2) seeding, 3) chaining, and 4) sequence alignment. Initially, minimap2 constructs an index database using minimizer seeds extracted from the reference genomes<sup>31</sup>. In the second step, the minimizer seeds extracted from a read sequence are matched to those extracted from the references. Third, the matching locations are sorted to identify adjacent seeds, which are then used to construct chains of matching seeds. Fourth and finally, a dynamic programming-based algorithm calculates sequence alignment between every two chains of seeds and stores mapping information in a sequence alignment/map (SAM and its compressed representation, BAM) file.

We propose to accelerate minimap2 by refraining from the computationally expensive chaining and DP-based alignment algorithms. We observe that heuristics, in particular the sorting of mapping locations based on seed hits and edit distance approximation algorithms, achieve close-to-optimal results but at a much greater speed. This leads us to our new heuristic and alignment-free read mapper.

## Evaluation Methodology

We achieve read mapping using only heuristic methods. To that end, we perform seeding and evaluate the seed hits for each read. We record each read's mapping location and the corresponding number of seed hits. We exclude all mapping locations with less than three seeds. Subsequently, we sort the mapping locations by their associated seed hits in descending order. We then employ edit distance approximation algorithms to estimate the edit distance of each read for the first three mapping locations with the highest number of seed hits. In the following main filtering step, we exclude all reads that do not achieve an edit distance below a pre-defined threshold (usually 10% - 15% of the read length) for at least one mapping location. For all remaining reads, we record the mapping location and the associated estimated edit distance. We pass this information on to the final profiling stage. We perform this procedure for all seven edit-distance estimation algorithms.

In our selection of heuristic methods, we rely on the first comprehensive overview of edit distance approximation algorithms published from 1993 until 2020<sup>46</sup>. Edit distance approximation algorithms aim to quickly estimate the edit distance between two sequences. The sequences are dissimilar if the edit distance estimate surpasses a user-defined threshold. In genomic studies, sequences with an edit distance less than or equal to a user-defined threshold ( $E$ ) are deemed biologically useful. All surveyed edit distance estimation approaches utilize an edit distance threshold to control rigor. The accuracy of edit distance approximation algorithms significantly depends on the edit distance threshold and the data

analyzed. We employ thresholds of 10%, which we determine to be optimal in most real-world cases and data distributions (refer to supplementary materials). While technically feasible, higher thresholds are rarely biologically useful and thus are rarely observed in real-world applications.

We use one compute thread for each tool on the same system as before, i.e., an AMD EPYC 7742 64-Core Processor, 1 TB of main memory, and an SSD with SATA3 interface<sup>37</sup>.

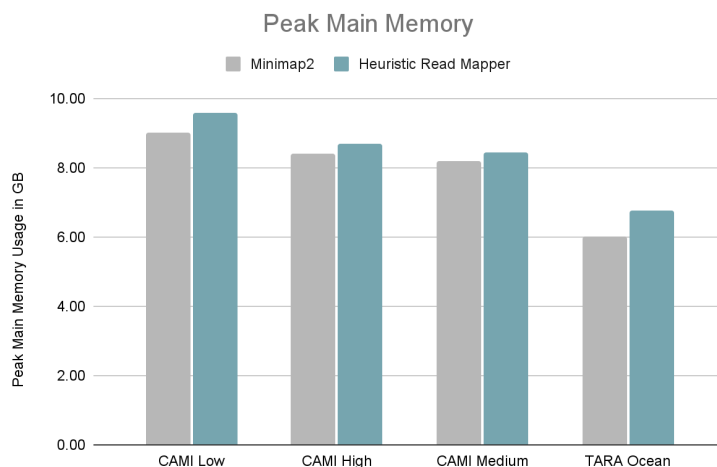
We benchmark our heuristic and alignment-free read mapper against minimap2. We again record the required computational resources, i.e., peak main memory and execution time. To obtain reliable runtime results, we perform each experiment three times and report the average of the observed execution times.

We again base our analysis on the three CAMI datasets, one of each diversity class: RL\_S001insert\_270.fq, RM\_S001insert\_270.fq and RH\_S001insert\_270.fq. We reproduce our analysis with a real-world read set from the TARA Ocean Project (ERR1700889\_1.fastq). This read set was also considered in the latest CAMI challenge.

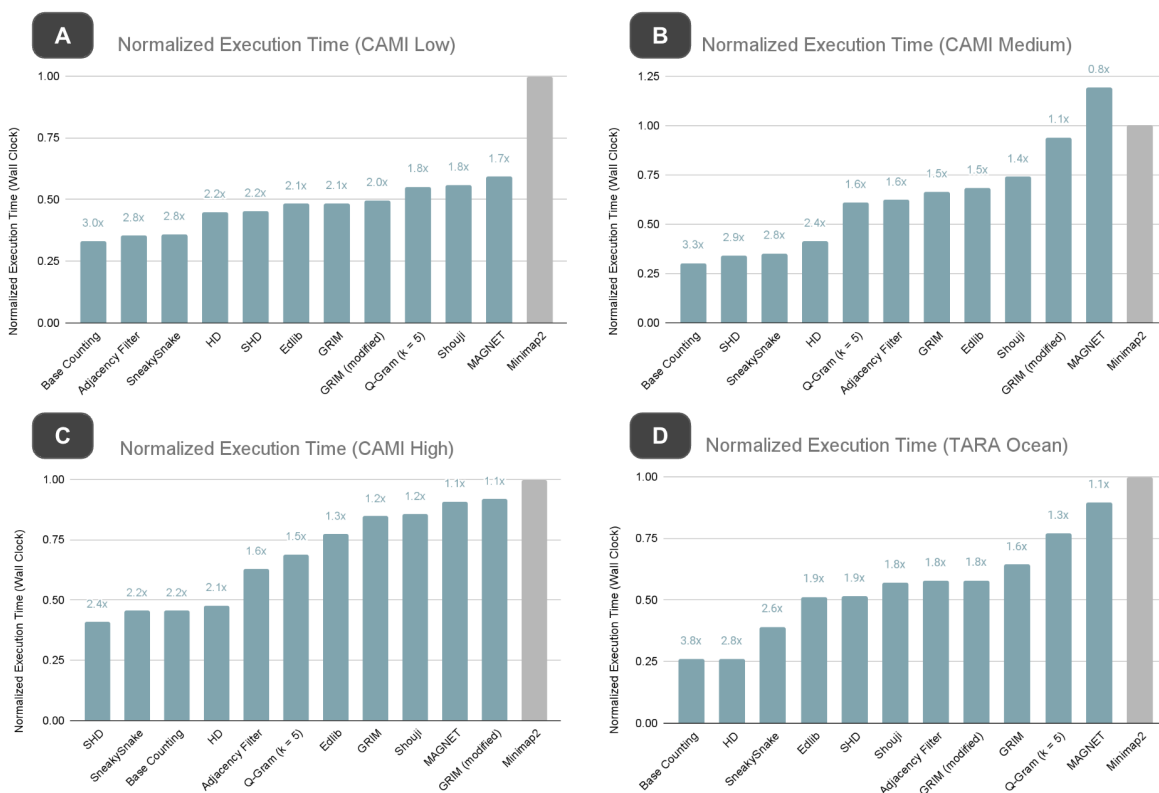
## Seed-based read filtering and edit distance approximation algorithms enable fast read mapping

We present our computational resource evaluation in *Figure 3 and Figure 4 A-D* and make five key observations:

1. Our alignment-free heuristic approach and minimap2 have practically the same main memory requirements.
2. Our alignment-free read mapper has a main memory footprint independent of the edit distance approximation algorithm. The peak main memory usage depends only on the read set and the reference database.
3. Methods such as Base Counting<sup>47</sup>, Adjacency Filter<sup>48</sup>, SneakySnake<sup>49</sup>, HD<sup>50</sup>, and SHD<sup>51</sup> consistently rank among the fastest. On average, these methods achieve a 3.5x speedup over minimap2.
4. In some cases, slower methods such as Magnet<sup>52</sup> yield only minimal speedup against minimap2.
5. If the workload for edit distance approximation algorithms is small, differences in runtime become barely visible, and all edit distance approximation algorithms deliver at least a 3x speedup over minimap2. We observe this effect for the CAMI low dataset.



**Figure 3** Comparison of the main memory footprint of our alignment-free heuristic read mapper against minimap2.



**Figure 4** Runtime analysis of our alignment-free and heuristic read mapper. We report the normalized elapsed wall clock time for each heuristic method employed in the read mapping stage and benchmark against minimap2. Each sub-figure shows the execution for each heuristic method and normalized to the baseline runtime of minimap2.

**A** Normalized execution time evaluated on the CAMI Low dataset. **B** Normalized execution time evaluated on the CAMI Medium dataset. **C** Normalized execution time evaluated on the CAMI High dataset. **D** Normalized execution time evaluated on the TARA Ocean dataset.

In conclusion, seed-based filtering and heuristic edit distance computation generally allow for strong speedup. The linear and some quadratic complexity edit distance approximation algorithms (such as q-gram<sup>53</sup>, GRIM<sup>54</sup>, Base Counting<sup>48</sup>, and SneakySnake<sup>49</sup>) deliver the greatest speedup. The differences in peak main memory usage stem from the index structure (MMI file), which is different for each read set, being loaded into main memory during index querying and seeding. The memory footprint is directly proportional to the size of the index, which in turn depends on the size of the subset database constructed in the containment search stage.

## The complete MetaFast Pipeline: relative abundance estimation and genome identification

Efficient and accurate identification of each microbes presence and relative abundances in an environmental sample directly recovered from its host environment continues to pose a significant challenge<sup>55</sup>. Existing analysis techniques necessitate comparing the genomic composition of the subject sample to a large volume of genomic data and employing computationally intensive algorithms to identify a broad range of microbes<sup>56</sup>. This requirement confines the analysis to high-performance computing platforms, which are typically power-intensive and unavailable in remote areas. There remains a substantial need and room for improvement in existing metagenomic analysis tools<sup>57,58</sup>.

Metalign<sup>14</sup>, a state-of-the-art mapping-based metagenomic analysis tool, utilizes three key steps. Initially, Metalign employs KMC3<sup>33</sup> and CMash<sup>34</sup> to narrow the list of potential candidate organisms in the metagenomic sample. Subsequently, Metalign uses minimap2<sup>41</sup> to map metagenomic reads to the filtered candidate genomes. Finally, Metalign estimates the relative abundances of microbes in the sample by amalgamating information from reads that uniquely map to one genome with those that align to multiple genomes. We propose our accurate and memory-frugal indexing and containment search algorithm and our heuristic read mapper presented earlier as the basis for a novel and fast metagenomic classifier, MetaFast.

### Evaluation Methodology

We combine the previous two pipeline stages, namely containment search and alignment-free read mapping. We then submit the read mapping results, i.e., reads mapped to one or multiple reference genomes and their associated edit distances, to the final stage for taxonomic profile generation. We benchmark MetaFast against Metalign and measure peak main memory and end-to-end execution time.

We evaluate MetaFast's accuracy, considering precision and recall in genome identification and the L1 norm error in relative abundance estimation. We evaluate the accuracy of MetaFast using the CAMI-affiliated analysis software OPAL<sup>42</sup>. Taxonomic profiles generated with Metalign serve as ground truth for our analysis. We restrict our evaluation to the lowest taxonomic rank, i.e., the species level. Both MetaFast and Metalign have the same reference genomes in their respective reference databases. To draw robust conclusions on the accuracy of relative abundance estimation for each heuristic method in the read mapping stage, we compute the average L1 norm error for each edit distance approximation algorithm across all four read sets.

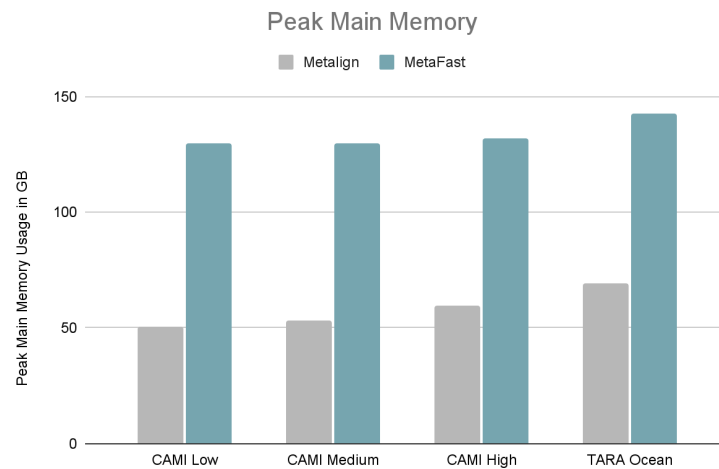


## MetaFast achieves fast and accurate taxonomic classification and abundance estimation

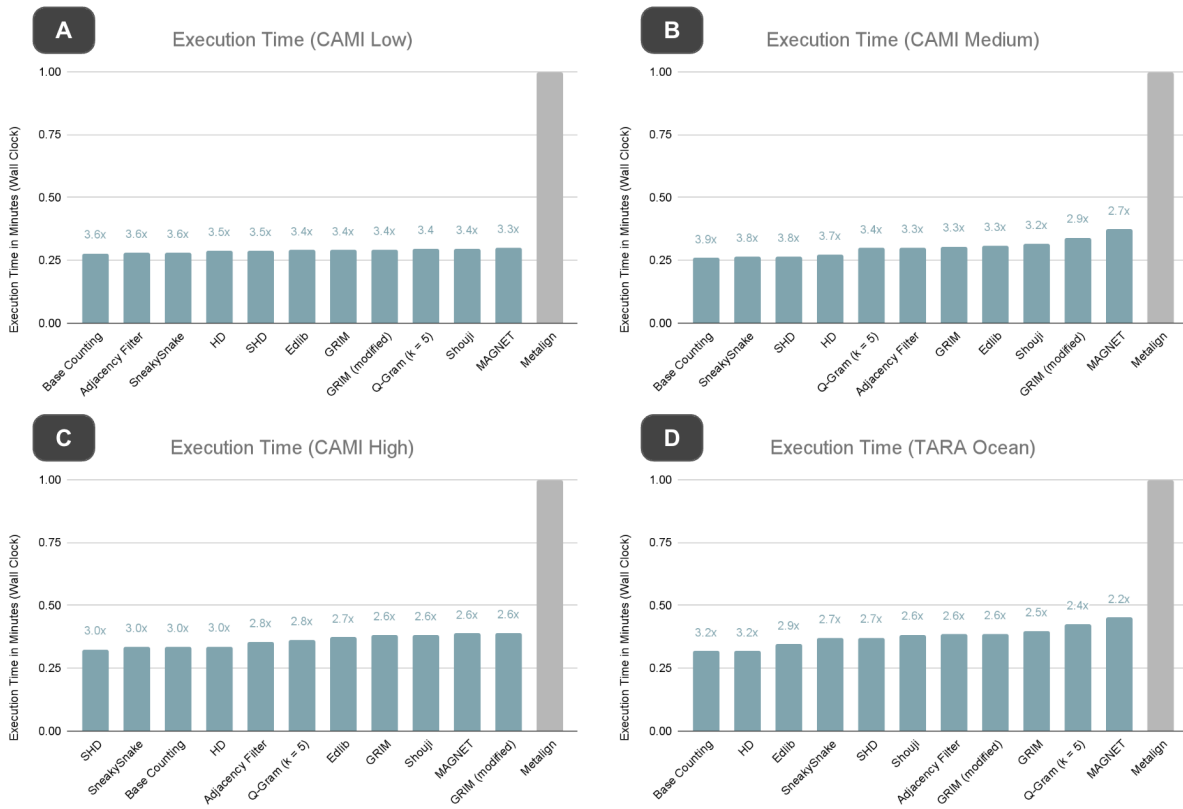
We thoroughly analyze the benefits of metagenomic classification at the species level with MetaFast.

We present our computational resource evaluation in *Figure 5* and *Figure 6 A-D*. We make three key observations:

1. MetaFast achieves a 4.5x speedup over Metalign for the methods SneakySnake, Base Counting, and Hamming Distance.
2. We generally observe an approximately 2-fold increase in peak main memory usage.
3. Hamming Distance, Base Counting, and SneakySnake remain the fastest methods.



**Figure 5** Memory footprint analysis of the complete MetaFast pipeline. We benchmark against Metalign.

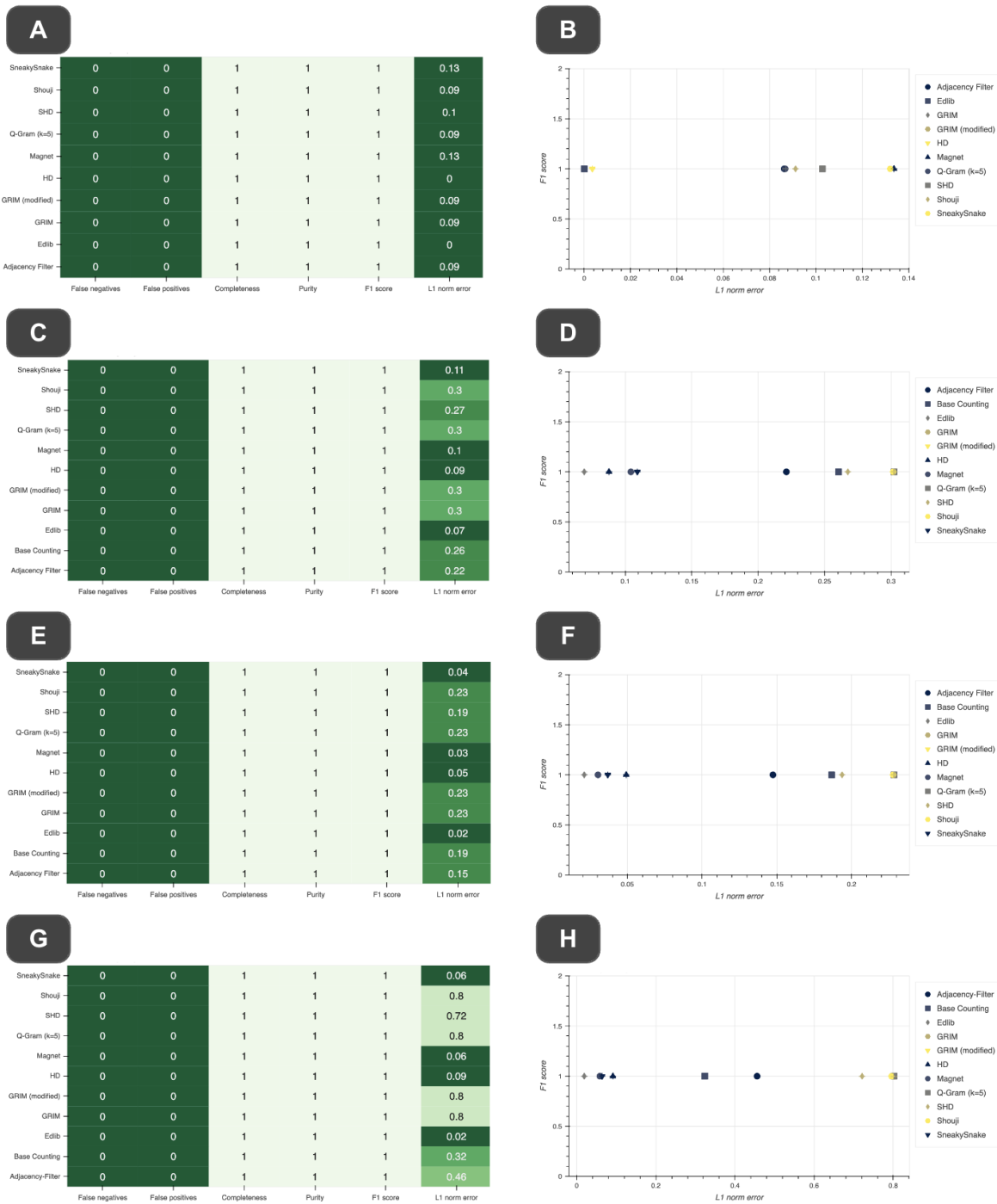


**Figure 6** Runtime analysis of the complete MetaFast pipeline. We report the normalized elapsed wall clock time for each heuristic method employed in the read mapping stage and benchmark against Metalign. Each sub-figure shows the execution for each heuristic method and normalized to the baseline runtime of Metalign. **A** Normalized execution time evaluated on the CAMI Low dataset. **B** Normalized execution time evaluated on the CAMI Medium dataset. **C** Normalized execution time evaluated on the CAMI High dataset. **D** Normalized execution time evaluated on the TARA Ocean dataset.

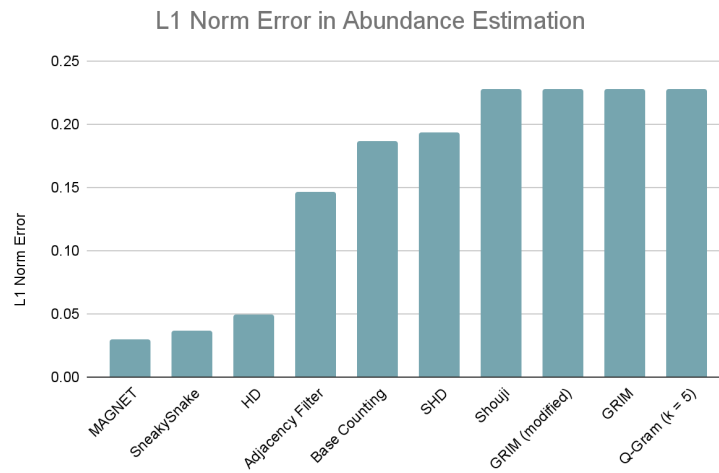
We conclude that MetaFast achieves a significant reduction in execution time for end-to-end metagenomic analyses. Since the metagenomic datasets we consider contain many species absent from MetaFast's database, the subset database constructed at the end of the containment search stage is relatively small (around 500 MB). As a result, the edit distance approximation algorithms employed in the read mapping stage are exposed to a lower workload, and the differences in execution time become less visible.

We present the classification accuracy results in *Figure 7 A-H* and *Figure 8*. We make three key observations:

1. All edit distance approximation algorithms exhibit zero false positive and zero false negative rates, resulting in an F1 score of always one.
2. Heuristic methods in the read mapping stage estimate relative abundance accurately, with minimal errors compared to the precise values computed by Metalign.
3. In the averaged L1 norm error analysis, Magnet<sup>52</sup>, SneakySnake<sup>49</sup>, and Edlib deliver the highest accuracy, i.e., the lowest relative abundance deviations



**Figure 7** Accuracy analysis of the complete MetaFast pipeline. We report the false negative and false positive rates, completeness, purity, and the L1 norm error for each heuristic method employed in the read mapping stage and benchmark against Metalign. We perform all accuracy considerations on the species level. **A** Overview of accuracy metrics for each heuristic method, evaluated on the CAMI Low dataset. **B** F1 score and L1 norm error for each heuristic method, evaluated on the CAMI Low dataset. **C** Overview of accuracy metrics for each heuristic method, evaluated on the CAMI Medium dataset. **D** F1 score and L1 norm error for each heuristic method evaluated on the CAMI Medium dataset. **E** Overview of accuracy metrics for each heuristic method evaluated on the CAMI High dataset. **F** F1 score and L1 norm error evaluated on the CAMI High dataset. **G** Overview of accuracy metrics for the TARA Ocean dataset. **H** F1 score and L1 norm error for the TARA Ocean dataset.



**Figure 8** L1 norm error for each heuristic method employed in the read mapping stage averaged over all four evaluated read sets.

A zero false negative rate in the final classification results again confirms a zero false negative rate in the reference database filtering stage. Upon combining the accuracy analysis with our previous computational resource evaluation, we conclude that SneakySnake offers the best accuracy-runtime tradeoff, while Base Counting stands as the fastest algorithm overall. Magnet is highly accurate in relative abundance estimation but too slow to gain significant speedup over minimap2, and one could simply use an aligner if the read mapping stage is required to deliver such high accuracy. The choice of edit distance approximation algorithm allows optimization for either runtime or accuracy: SneakySnake is the optimal choice for achieving the overall best accuracy-runtime tradeoff; Base Counting may be chosen to optimize for runtime.

In every experiment conducted for taxonomic profiling, we ascertain the correctness of each algorithm, or combination of algorithms, by confirming the presence or absence of each taxon with the correct results generated by Metalign. We consistently observe identical presence or absence of each taxon in all taxonomic profiles for each edit distance approximation algorithm. Consequently, the F1 score invariably equals one. Nevertheless, we notice minor discrepancies between the relative abundance estimates of Metalign (assuming the taxonomy profile of Metalign as the ground truth) and our edit distance approximation algorithms. These differences stem from the varying numbers of mapped reads provided by our algorithm and minimap2. We quantify these differences and represent them as L1 norm error, a measure for deviations in relative abundance.

We conclude that MetaFast correctly and accurately identifies the presence, absence, and relative abundance of taxa in a metagenomic sample.

## Discussion & Future Work

We introduced MetaFast, a computational tool designed to identify and quantify species abundance in a high-throughput sequencing (HTS) sample. MetaFast accelerates the entire computational genome analysis pipeline rather than individual stages. This approach is crucial as focusing on a single stage limits the overall speedup, as Amdahl's Law<sup>60</sup> dictates. We note that improving read mapping performance impacts almost all genomic analyses that use sequencing data. For instance, read mapping constitutes up to 45% of the execution time in cancer genomics studies<sup>61</sup> and 30% in profiling the taxonomy of a multispecies sample<sup>62</sup>. Therefore, our accelerated read mapping stage may also prove useful as a standalone tool.

In MetaFast, we leverage the similarities between indexing and seeding to develop a seeding-based containment search methodology. Indexing, which is off the critical path for most bioinformatics applications, is highly relevant for clinical and medical applications due to frequent modifications to the reference database.

In the reference database filtering, i.e., containment search stage, we observe a 6x speedup over KMC3+CMash when processing one batch, given that we consider reference databases based on the same reference genomes and the same read set. However, the speedup reduces to fourfold when we increase the number of batches and, correspondingly, the number of threads (assuming a thread count equal to or greater than the number of batches, i.e., at least one thread per batch). This slowdown arises from the necessity to combine individually generated results in a multithreaded scenario.

MetaFast's restriction to short reads is a potential limitation. The heuristic methods and empirically determined parameters are optimized for short reads, and extending them to long reads is not entirely trivial. However, there are abundant applications, for example, clinical metagenomics, where short-read sequencing technology prevails<sup>59</sup>.

Sample collection, preparation, and sequencing are key steps in clinical applications<sup>63</sup>. Short-read sequencing technologies, such as Illumina<sup>64</sup>, dominate clinical settings. Furthermore, the equivalent length of all reads stored in the same FASTQ file aids our use case, as most heuristic edit distance approximation methods spend most of their execution time on the longest sequences. The uniformity in length prevents a dramatic increase in runtime that could otherwise occur even in samples that contain only a few long sequences.

A potential optimization to be made in future works is reusing seeds generated from the reads in the containment search stage, later during read mapping. Currently, we perform seeding to locate seeds common to the read set and the full reference database to select a smaller subset database. Then, for read mapping, we repeat this seeding step with the same read set. Instead, we could simply use the seeds from our containment search stage. It still stands to be determined if we may also partially reuse the seed locations we found during containment search. However, since the subset database is very small (approximately 3 MB for CAMI High reads), index construction and seeding are fast. We never observed running times of more than two minutes for this repeated index generation and seeding phase.

It further remains to be examined whether seeding in the read mapping stage significantly impacts runtime for very large read sets.

We anticipate that the foundation established in MetaFast will foster further advancements in metagenomic research. Specifically, we aspire for metagenomic-based screening and early diagnosis methods to become more prevalent and accessible in healthcare, drawing upon tools like MetaFast.

# Methods

The primary objective of MetaFast is a significant reduction in the end-to-end execution time and memory footprint of the indexing, seeding, and read mapping stages in metagenomic analyses. Given two genomic sequences, a reference sequence  $R[0, \dots, r - 1]$  and a query sequence  $Q[0, \dots, q - 1]$ , where  $r \geq q$ , these sequences consist of A, C, G, T in the DNA alphabet  $\{A, C, G, T\}$  in addition to the ambiguous base, N. Our goal is to locate all correct mapping locations of  $Q$  in  $R$  through a methodology that is quick, memory-efficient, and accurate.

## Reference Database Organisation and Index Construction & Querying

To compile a comprehensive reference database, we utilize all NCBI microbial genome assemblies<sup>38</sup>, encompassing both complete and incomplete assemblies from RefSeq and GenBank, as of June 2020. The final database comprises 199,614 organisms, amounting to 170 GB in size (in gzipped form). For index construction, we divide the database into 25 batches, each containing approximately 7 GB of reference data, resulting in a 10 GB index structure (MMI) per batch. We never copy any reference genomes; the 7 GB reference data batch is merely a methodological construct. We employ minimap2-fast to generate each batch's indices (MMIs) and store them in a separate directory. The batches are entirely independent during the index construction phase, with their interplay and possible combination occurring later during database querying. Consequently, indices for all batches can be generated in parallel, with the index generation time for a single batch being approximately five minutes. We refer to the entirety of all indices as our database.

During containment search, we query the database. An individual and independent minimap2-fast instance can process each individual batch, i.e., each MMI file, either in a single or multithreaded fashion. We count the number of seed hits for each batch and then merge the results. Should a thread count lower than the number of batches be chosen in the querying stage, we logically group several batches into clusters. We then process all indices in one cluster sequentially, with one thread. If a reference genome is simultaneously contained in several batches, we compute the number of seed hits for this reference genome for each batch. We then conservatively consider only the highest number of seed hits this reference genome achieves. We include all reference genomes that achieve seed hits above the defined cutoff threshold in the subset database.



## Choice of Edit Distance Approximation Algorithm

There are four main edit distance approximation approaches for genomic sequence comparison. Based on prior benchmarking literature, we identify the methods with the most promising accuracy runtime tradeoff from each methodology.

Name	Year	Methodology	Short/ Long Reads	Native Platform	Language	URL Software
SneakySnake <sup>49</sup>	2019	Pigeonhole	Short/ Long	CPU/ GPU/ FPGA	C/ C++	<a href="https://github.com/CMU-SAFARI/SneakySnake">https://github.com/CMU-SAFARI/SneakySnake</a>
Shouji <sup>65</sup>	2019	Pigeonhole	Short	FPGA	C/ Verilog	<a href="https://github.com/CMU-SAFARI/Shouji">https://github.com/CMU-SAFARI/Shouji</a>
Hamming Distance (HD) <sup>50</sup>	2019	Pigeonhole	Short	CPU		N/A
GRIM-Filter <sup>54</sup>	2018	q-gram	Short	PIM	C	<a href="https://github.com/CMU-SAFARI/GRIM">https://github.com/CMU-SAFARI/GRIM</a>
Magnet <sup>52</sup>	2017	Pigeonhole	Short	CPU	Matlab	<a href="https://github.com/Bilke ntCompGen/MAGNET">https://github.com/Bilke ntCompGen/MAGNET</a>
SHD <sup>51</sup>	2015	Pigeonhole	Short	SIMD	C/ SIMD	<a href="https://github.com/CMU-SAFARI/Shifted-Hamming-Distance">https://github.com/CMU-SAFARI/Shifted-Hamming-Distance</a>
Adjacency Filter <sup>48</sup>	2010	Pigeonhole	Short	CPU	C	<a href="https://github.com/Bilke ntCompGen/mrfast">https://github.com/Bilke ntCompGen/mrfast</a>

Base Counting is the simplest method for comparing genomic sequences, as it merely compares the frequency of individual genomic bases between two sequences, resulting in a time complexity of  $O(n)$ <sup>47</sup>. The q-gram algorithm<sup>53</sup>, seen as an extension of the Base Counting algorithm, compares the abundance of q-long subsequences to obtain a lower-bound estimate of the edit distance with an attractive runtime of  $O(n)$ . Finally, methods including SneakySnake and SHD are subject to the Pigeonhole Principle<sup>66</sup>. The simplest version of this principle is the Hamming Distance algorithm. Methods following the Pigeonhole Principle have a time complexity of  $O(nE)$ , where  $n$  is the sequence length, and  $E$  is the edit distance threshold, given as a percentage of the read length.

## Empirically Determined Parameters

We must choose several parameters for our heuristic database filtering and read mapping approaches. To achieve this, we sweep one parameter while holding all others constant. We repeat this in cyclic rotation to ensure that the choice of one parameter does not preclude the optimal choice of the others. This process continues until we find the optimal combination of parameters. We consider a parameter combination optimal if it leads to the highest accuracy-runtime tradeoff in the end-to-end accuracy and runtime evaluation of the MetaFast pipeline. In our containment search stage, we look for the optimal combination of two parameters:

- The seed length  $k$
- And the minimum number of seed hits a reference genome needs to achieve to be included in the subset database (i.e., the seed-count cutoff value).

We iterate through all possible combinations of these two parameters and find that the a seed length of  $k = 28$  and a seed-count cutoff value of  $0.0001$  lead to the optimal end-to-end accuracy-runtime tradeoff.

For the alignment-free read mapping stage, we again aim to determine the optimal combination of two parameters:

- The minimum number of seeds per mapping location.
- The number of mapping locations with the highest number of seeds to examine.

We know the edit distance threshold to be optimally set to 10%<sup>67</sup>.

Initially, we examine only the single mapping location with the highest number of seeds without requiring a minimum number of seed hits. We iterate over all four read sets as previously described. We profile the mapping results and examine the OPAL<sup>42</sup> report to identify species-level false negatives. If a false negative occurs for any read set, we include the mapping location with the next highest number of seeds. This inclusion continues until the false negatives are resolved.

Conversely, we first allow an unlimited number of mapping locations per read but require a minimum number of seed hits per read and mapping location. We start by considering all mapping locations with at least one seed hit. We incrementally increase the minimum number of seed hits until a false negative occurs in the taxonomic profile. Through this process, we determine the optimal choice to be a minimum number of three seeds per mapping location. All mapping locations not satisfying this criterion are discarded. A read that does not have at least one mapping location with at least three seed hits is filtered out entirely. We then consider the top three mapping locations per read, i.e., the locations with the three highest numbers of seed hits. If a read has fewer than three mapping locations with at least three seed hits, we consider the remaining locations that satisfy the minimum seed hits requirement.

## Taxonomic Profile Generation

In line with Metalign, we adopt the same parameters and choices, presuming their optimality for MetaFast. We observe a runtime of less than two minutes in the profiling stage for all considered read sets, a factor we deem negligible in our analysis.

## Optimization Strategies

We present two distinct optimization strategies for the containment search and read mapping stage, respectively.

### Performing stronger seed-match based filtering

We can choose to increase the threshold of the minimum required number of seeds per mapping location. One increment already reduces the total number of reads for the subsequent edit distance approximation stage. This reduction in workload for the edit distance approximation methods leads to a significant speedup. This parameter can be set as a command-line argument. A high threshold, however, delivering great speedup, may cause false negatives in the final species-level taxonomic profile.

### Multithreaded containment search and parallelized index access

To maximize the parallel processing of our index structure, we initiate a thread for each batch, i.e., each MMI file in the reference database. We allow each file thread to launch  $t$  subthreads, thereby processing each MMI file in a multithreaded fashion with  $t$  parallel compute threads. We are interested in determining the seed-hit count for each reference genome. To achieve this, we keep track of the number of seed hits each reference genome receives through a data structure we call *seedmap*. Each seedmap uses the reference genome's accession number as key and stores the number of seed hits as value.

We aim to efficiently balance the workload among all threads while minimizing waiting times arising from interdependencies of indifferent parallelly active compute threads. Our solution strategy begins by assigning each file thread a vector of seedmaps and a corresponding mutex for each seedmap. The number of seedmaps always equals the number of sub-threads. We do not limit our seed counting process (for one specific MMI file) to a single seedmap, as this would precipitate considerable waiting times for mutex availability.

We aim to ensure an effective balance between the threads and the available seedmaps, thereby avoiding excessive mutex contention. To that end, we hash each accession number and then apply a modulo operation with the number of seedmaps (equivalent to the number of subthreads per MMI file). We increment the value associated with the accession number key in the corresponding seedmap by the seed count after securing the mutex. This relationship may be expressed as

$$\text{seedmap to access} = \text{hash}(\text{accession number}) \% (\text{number of seedmaps}).$$

To construct the subset database, we need to determine the total seed-hit count for each reference genome. We thus merge all seedmaps, once containment mapping is concluded, by summing up the values of each key into a single, ordered seedmap. Given this final seedmap, we can now place all reference genomes with seed-hit counts above the cutoff threshold in the subset database.

## Data & Code Availability

We exclusively relied on publicly available datasets in this paper. The CAMI challenge datasets are available on the [GigaDB](#) website. The prokaryote-isolated Tara Oceans reads used in our study are available on EBI: <https://www.ebi.ac.uk/ena/data/view/PRJEB1787>. The run accessions are ERR598952 and ERR598957.

MetaFast's source code is available on GitHub: <https://github.com/CMU-SAFARI/MetaFast>.

For inquiries or if you wish to collaborate, please contact [arvid.gollwitzer@safari.ethz.ch](mailto:arvid.gollwitzer@safari.ethz.ch).

## Acknowledgments

We thank all members of the SAFARI Research Group for the scholarly environment they provide.

We specifically thank Joel Bergtholdt for his invaluable contributions to this paper while pursuing an undergraduate research project under the mentorship of Arvid E. Gollwitzer.

Arvid E. Gollwitzer especially thanks Dr. Mohammed Alser for his unwavering support and mentorship.

# References

---

1. Dulanto Chiang, A. & Dekker, J. P. From the Pipeline to the Bedside: Advances and Challenges in Clinical Metagenomics. *J. Infect. Dis.* **221**, S331–S340 (2020).
2. Delaney, S. K. *et al.* Toward clinical genomics in everyday medicine: perspectives and recommendations. *Expert Rev. Mol. Diagn.* **16**, 521–532 (2016).
3. Logares, R. *et al.* Environmental microbiology through the lens of high-throughput DNA sequencing: synopsis of current platforms and bioinformatics approaches. *J. Microbiol. Methods* **91**, 106–113 (2012).
4. Koren, S. *et al.* Hybrid error correction and de novo assembly of single-molecule sequencing reads. *Nat. Biotechnol.* **30**, 693–700 (2012).
5. Ledergerber, C. & Dessimoz, C. Base-calling for next-generation sequencing platforms. *Brief. Bioinform.* **12**, 489–497 (2011).
6. Chou, H. H. & Holmes, M. H. DNA sequence quality trimming and vector removal. *Bioinformatics* **17**, 1093–1104 (2001).
7. Sharon, I. *Computational Methods for Metagenomic Analysis*. (Technion - Israel Institute of Technology, Faculty of Computer Science, 2010).
8. Biesecker, L. G. *et al.* The ClinSeq Project: piloting large-scale genome sequencing for research in genomic medicine. *Genome Res.* **19**, 1665–1674 (2009).
9. Lee, Y.-H. & Wong, D. T. Saliva: an emerging biofluid for early detection of diseases. *Am. J. Dent.* **22**, 241–248 (2009).
10. Chen, Y.-C., Wang, C.-T., Lees, D. C. & Wu, L.-W. Higher DNA insert fragment sizes improve mitogenomic assemblies from metagenomic pyrosequencing datasets: an example using Limenitidinae butterflies (Lepidoptera, Nymphalidae). *Mitochondrial DNA Part A* **29**, 840–845 (2018).
11. Meyer, F. *et al.* Critical Assessment of Metagenome Interpretation: the second round of challenges. *Nat. Methods* **19**, 429–440 (2022).
12. Niaré, K., Greenhouse, B. & Bailey, J. A. An optimized GATK4 pipeline for *Plasmodium falciparum*

- whole genome sequencing variant calling and analysis. *Malar. J.* **22**, 207 (2023).
13. Truong, D. T. *et al.* MetaPhlan2 for enhanced metagenomic taxonomic profiling. *Nat. Methods* **12**, 902–903 (2015).
  14. LaPierre, N., Alser, M., Eskin, E., Koslicki, D. & Mangul, S. Metalign: efficient alignment-based metagenomic profiling via containment min hash. *Genome Biol.* **21**, 242 (2020).
  15. Wood, D. E., Lu, J. & Langmead, B. Improved metagenomic analysis with Kraken 2. *Genome Biol.* **20**, 257 (2019).
  16. Ounit, R., Wanamaker, S., Close, T. J. & Lonardi, S. CLARK: fast and accurate classification of metagenomic and genomic sequences using discriminative k-mers. *BMC Genomics* **16**, 236 (2015).
  17. Breitwieser, F. P., Baker, D. N. & Salzberg, S. L. KrakenUniq: confident and fast metagenomics classification using unique k-mer counts. *Genome Biol.* **19**, 198 (2018).
  18. Kim, D., Song, L., Breitwieser, F. P. & Salzberg, S. L. Centrifuge: rapid and sensitive classification of metagenomic sequences. *Genome Res.* **26**, 1721–1729 (2016).
  19. Milanese, A. *et al.* Microbial abundance, activity and population genomic profiling with mOTUs2. *Nat. Commun.* **10**, 1014 (2019).
  20. Sunagawa, S. *et al.* Tara Oceans: towards global ocean ecosystems biology. *Nat. Rev. Microbiol.* **18**, 428–445 (2020).
  21. Lu, J. & Salzberg, S. L. Ultrafast and accurate 16S rRNA microbial community analysis using Kraken 2. *Microbiome* **8**, 124 (2020).
  22. Yacouby, R. & Axman, D. Probabilistic Extension of Precision, Recall, and F1 Score for More Thorough Evaluation of Classification Models. in *Proceedings of the First Workshop on Evaluation and Comparison of NLP Systems* 79–91 (Association for Computational Linguistics, 2020). doi:10.18653/v1/2020.eval4nlp-1.9.
  23. Marian, A. J. Challenges in medical applications of whole exome/genome sequencing discoveries. *Trends Cardiovasc. Med.* **22**, 219–223 (2012).
  24. Loman, N. J. *et al.* High-throughput bacterial genome sequencing: an embarrassment of choice, a world of opportunity. *Nat. Rev. Microbiol.* **10**, 599–606 (2012).
  25. Rodríguez-Ezpeleta, N., Hackenberg, M. & Aransay, A. M. *Bioinformatics for High Throughput*

- Sequencing*. (Springer Science & Business Media, 2011).
26. NovaSeq X specifications.  
<https://emea.illumina.com/systems/sequencing-platforms/novaseq-x-plus/specifications.html>.
  27. Mayer, L. W. Use of plasmid profiles in epidemiologic surveillance of disease outbreaks and in tracing the transmission of antibiotic resistance. *Clin. Microbiol. Rev.* **1**, 228–243 (1988).
  28. Pareek, C. S., Smoczynski, R. & Tretyn, A. Sequencing technologies and genome sequencing. *J. Appl. Genet.* **52**, 413–435 (2011).
  29. Zhang, D. *et al.* Optimized Sequencing Adaptors Enable Rapid and Real-Time Metagenomic Identification of Pathogens during Runtime of Sequencing. *Clin. Chem.* **68**, 826–836 (2022).
  30. Rosenberg, M. S. *Sequence Alignment: Methods, Models, Concepts, and Strategies*. (University of California Press, 2009).
  31. Alser, M. *et al.* From molecules to genomic variations: Accelerating genome analysis via intelligent algorithms and architectures. *Comput. Struct. Biotechnol. J.* **20**, 4579–4599 (2022).
  32. Kim, J. S. *et al.* AirLift: A Fast and Comprehensive Technique for Remapping Alignments between Reference Genomes. *arXiv [q-bio.GN]* (2019).
  33. Kokot, M., Dlugosz, M. & Deorowicz, S. KMC 3: counting and manipulating k-mer statistics. *Bioinformatics* **33**, 2759–2761 (2017).
  34. Liu, S. & Koslicki, D. CMash: fast, multi-resolution estimation of k-mer-based Jaccard and containment indices. *Bioinformatics* **38**, i28–i35 (2022).
  35. Buckingham, L. K-mer based algorithms for biological sequence comparison and search. (Queensland University of Technology, 2022).
  36. Fletcher, S. & Islam, M. Z. Comparing sets of patterns with the Jaccard index. *Australasian Journal of Information* (2018).
  37. Labor Policy. Samsung 870 EVO SATA SSD. *Samsung Semiconductor Global*  
<https://semiconductor.samsung.com/consumer-storage/internal-ssd/870evo/>.
  38. Pruitt, K. D., Tatusova, T., Brown, G. R. & Maglott, D. R. NCBI Reference Sequences (RefSeq): current status, new features and genome annotation policy. *Nucleic Acids Res.* **40**, D130–5 (2012).
  39. *RefSeq Help*. (National Center for Biotechnology Information (US), 2011).



40. *GenBank Overview*.
41. Li, H. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics* **34**, 3094–3100 (2018).
42. Meyer, F. *et al.* Assessing taxonomic metagenome profilers with OPAL. *Genome Biol.* **20**, 51 (2019).
43. Sadasivan, H. *et al.* Accelerating Minimap2 for Accurate Long Read Alignment on GPUs. *J Biotechnol Biomed* **6**, 13–23 (2023).
44. Hardware Accelerator for Minimap-2 Kernel. (2020).
45. Liyanage, K., Gamaarachchi, H., Ragel, R. & Parameswaran, S. Cross Layer Design Using HW/SW Co-Design and HLS to Accelerate Chaining in Genomic Analysis. *IEEE Trans. Comput. Aided Des. Integr. Circuits Syst.* 1–1 (2023) doi:10.1109/TCAD.2023.3236559.
46. Rumpf, M.-D. *et al.* SequenceLab: A Comprehensive Benchmark of Computational Methods for Comparing Genomic Sequences. *arXiv [q-bio.GN]* (2023).
47. Hach, F. *et al.* mrsFAST-Ultra: a compact, SNP-aware mapper for high performance sequencing applications. *Nucleic Acids Res.* **42**, W494–500 (2014).
48. Alser, M. *et al.* GateKeeper: a new hardware architecture for accelerating pre-alignment in DNA short read mapping. *Bioinformatics* **33**, 3355–3363 (2017).
49. Alser, M., Shahroodi, T., Gómez-Luna, J., Alkan, C. & Mutlu, O. SneakySnake: a fast and accurate universal genome pre-alignment filter for CPUs, GPUs and FPGAs. *Bioinformatics* **36**, 5282–5290 (2021).
50. Sun, Y., Aljawad, O., Lei, J. & Liu, A. Genome-scale NCRNA homology search using a Hamming distance-based filtration strategy. *BMC Bioinformatics* **13 Suppl 3**, S12 (2012).
51. Xin, H. *et al.* Shifted Hamming distance: a fast and accurate SIMD-friendly filter to accelerate alignment verification in read mapping. *Bioinformatics* **31**, 1553–1560 (2015).
52. Alser, M., Mutlu, O. & Alkan, C. MAGNET: Understanding and Improving the Accuracy of Genome Pre-Alignment Filtering. *arXiv [q-bio.GN]* (2017).
53. Sutinen, E. & Tarhio, J. On using q-gram locations in approximate string matching. in *Algorithms — ESA '95* 327–340 (Springer Berlin Heidelberg, 1995). doi:10.1007/3-540-60313-1\_153.
54. Kim, J. S. GRIM-Filter. doi:10.1186/s12864-018-4460-0.
55. Hamady, M. & Knight, R. Microbial community profiling for human microbiome projects: Tools,

- techniques, and challenges. *Genome Res.* **19**, 1141–1152 (2009).
56. Sun, Y. *et al.* A large-scale benchmark study of existing algorithms for taxonomy-independent microbial community analysis. *Brief. Bioinform.* **13**, 107–121 (2012).
  57. Prakash, T. & Taylor, T. D. Functional assignment of metagenomic data: challenges and applications. *Brief. Bioinform.* **13**, 711–727 (2012).
  58. Lindgreen, S., Adair, K. L. & Gardner, P. P. An evaluation of the accuracy and speed of metagenome analysis tools. *Sci. Rep.* **6**, 19233 (2016).
  59. Govender, K. N. & Eyre, D. W. Benchmarking taxonomic classifiers with Illumina and Nanopore sequence data for clinical metagenomic diagnostic applications. *Microb Genom* **8**, (2022).
  60. Amdahl, G. M. Computer Architecture and Amdahl's Law. *Computer* **46**, 38–46 (2013).
  61. Dellaire, G., Berman, J. N. & Arceci, R. J. *Cancer Genomics: From Bench to Personalized Medicine*. (Academic Press, 2013).
  62. McCoun, M., Oyebade, A., Estrada-Reyes, Z. M., Pech-Cervantes, A. A. & Ogunade, I. M. Effects of Multi-Species Direct-Fed Microbial Products on Ruminant Metatranscriptome and Carboxyl-Metabolome of Beef Steers. *Animals (Basel)* **11**, (2021).
  63. Weiss, M. M. *et al.* Best practice guidelines for the use of next-generation sequencing applications in genome diagnostics: a national collaborative study of Dutch genome diagnostic laboratories. *Hum. Mutat.* **34**, 1313–1321 (2013).
  64. Adewale, B. A. Will long-read sequencing technologies replace short-read sequencing technologies in the next 10 years? *S. Afr. J. Lab. Clin. Med.* **9**, 1340 (2020).
  65. Alser, M., Hassan, H., Kumar, A., Mutlu, O. & Alkan, C. Shouji: a fast and efficient pre-alignment filter for sequence alignment. *Bioinformatics* **35**, 4255–4263 (2019).
  66. Ajtai, M. The complexity of the pigeonhole principle. *Combinatorica* **14**, 417–433 (1994).
  67. Alser, M. *et al.* Accelerating Genome Analysis: A Primer on an Ongoing Journey. *IEEE Micro* **40**, 65–75 (2020).