# MegaPath-Nano: Accurate Compositional Analysis and Drug-level Antimicrobial Resistance Detection Software for Oxford Nanopore Long-read Metagenomics


Wui Wang Lui*
*Department of Computer Science*
*The University of Hong Kong*
Hong Kong, China
wwlui@cs.hku.hk

Amy W. S. Leung*
*Department of Computer Science*
*The University of Hong Kong*
Hong Kong, China
wsleung@cs.hku.hk

Henry C. M. Leung*
*Department of Computer Science*
*The University of Hong Kong*
Hong Kong, China
cmleung2@cs.hku.hk

Yan Xin
*Department of Computer Science*
*The University of Hong Kong*
Hong Kong, China
yxin@cs.hku.hk

Jade L. L. Teng
*Department of Microbiology*
*The University of Hong Kong*
Hong Kong, China
llteng@hku.hk

Patrick C. Y. Woo
*Department of Microbiology*
*The University of Hong Kong*
Hong Kong, China
pcywoo@hku.hk

Tak-Wah Lam†
*Department of Computer Science*
*The University of Hong Kong*
Hong Kong, China
twlam@cs.hku.hk

Ruibang Luo†‡
*Department of Computer Science*
*The University of Hong Kong*
Hong Kong, China
rbluo@cs.hku.hk



***Abstract*—Accurate and sensitive taxonomic profiling is essential for any metagenomic analysis for revealing microbial community structure and for potential functional prediction. Antimicrobial resistance (AMR) detection is also a critical task in the clinical diagnosis of infection and antimicrobial therapy. By incorporating Oxford Nanopore Technologies (ONT) sequencing, users benefit from the high-confidence alignment of long reads for taxonomic classification, even among bacteria with similar genomes. Portable ONT devices, such as VolTRAX with MinION, allow short turnaround time for detection and can be used in a lightweight laboratory setting. However, error-prone ONT sequencing reads are still challenging for existing software for accurate taxonomic classification of microbes and detection of AMR down to the drug level.**

**In this paper, we present MegaPath-Nano, the successor to NGS-based MegaPath. It is a high-precision compositional analysis software with drug-level AMR detection for ONT metagenomic sequencing data. MegaPath-Nano performs 1) thorough multi-level filtering against decoy and human reads, while removing noisy alignments, 2) alignment-based taxonomic classification with RefSeq down to strain-level, with an alignment-reassignment algorithm to tackle the challenge of non-unique alignments, based on global alignment distribution, and 3) comprehensive downstream drug-level AMR detection, integrating five AMR databases. In our benchmarks using the Zymo metagenomic datasets, MegaPath-Nano performed better than other existing software for taxonomic classification. We also sequenced five real patient isolates using MinION to benchmark the performance of AMR detection. MegaPath-Nano was the most accurate and provided the most comprehensive output at both the drug and class level of AMR prediction against other state-of-the-art software.**

**MegaPath-Nano is open-source and available at https://github.com/HKU-BAL/MegaPath-Nano.**



***Index Terms*—pathogen detection, antimicrobial resistance prediction, MinION long reads, metagenomics, compositional analysis**


## I. Introduction

The progressive development of short-read Next-Generation Sequencing (NGS) technologies with fewer sequencing errors and more affordable sequencing cost [1] has supported the rapid growth of metagenomic studies. Medical applications of NGS-based metagenomic sequencing for pathogen detection have become more standardized in the past ten years [2, 3, 4]. Accurate and sensitive metagenomic profiling allows 1) taxonomic classification [5], 2) abundance estimation [6], and subsequently, 3) AMR prediction [7].

Since NGS-based metagenomic sequencing is limited by the short length of NGS reads, its resolution for taxonomic classification between microbes at finer levels is low, especially in

*These authors contributed equally to this work


†Correspondence should be addressed to R. L. (rbluo@cs.hku.hk) and T. L. (twlam@cs.hku.hk)



‡R. L. was supported by the ECS (grant number 27204518) of the HKSAR government, and by the URC fund at HKU.

conservative sequence regions [8]. The use of third-generation sequencing Oxford Nanopore Technology (ONT), which generates long reads with an average length of over 10kbp, is a good alternative solution for more sensitive and accurate pathogen detection. With ONT real-time sequencing and base-calling [9], the user also benefits from shorter turnaround time compared to using NGS-based detection. Despite all these advantages, the raw read sequencing error rate of ONT is still high, at about 10% [10]. While most of the existing taxonomic classification or AMR detection software, such as MegaPath [11], Centrifuge [12], Kraken 2 [13], AMR++ [14], and ARGs-OAP [15], were designed for short reads, to the best of our knowledge, there is no software or workflow optimized for ONT metagenomic data that incorporates lower-level taxonomic profiling with comprehensive drug-level AMR detection.

The existing taxonomic classification and pathogen detection software applicable to ONT metagenomic reads can be divided into three main categories: 1) kmer-based read classifier, 2) alignment-based classifier, and 3) Bayesian/EM-based estimator. Kraken 2 with Bracken, is a kmer-based read taxonomic classifier and a Bayesian-based composition estimator. Kraken 2 is memory-efficient and is ultra-fast, but at the expense of giving up useful read-level information. Centrifuge directly applies the Burrows-Wheeler transform [16]. It also runs fast, but it assigns multi-mapped reads to their Lowest Common Ancestor (LCA), which causes a less taxonomy-specific read assignment. Noteworthy, WIMP, which is the proprietary workflow developed by ONT for taxonomic classification, uses Centrifuge as its backend. A recently developed software MetaMaps [17] uses an EM-based approach to estimate microbial compositions with approximate composition-dependent mapping results. The EM algorithm can be resource-thirsty and run much slower than the other methods [18].

Taxonomic profiling allows the detection of potential pathogens in clinical samples. Nevertheless, commonly used AMR detection software does not integrate taxonomic classification results with resistance prediction. Most of these AMR detection tools are either not designed for long-read or cannot detect down to drug level. The NGS-based metagenomics software AMR++ was published together with the accompanying AMR database MEGARes for characterization and quantification of resistance genes. MEGARes comprises representative sequences of AMR genes, reflecting its higher reliability and non-redundant integration of related AMR databases [14]. ONT developed the Antimicrobial Resistance Mapping Application (ARMA) [19] workflow, which aligns ONT reads to CARD (Comprehensive Antibiotic Resistance Database) databases [20] after taxonomic classification by WIMP [21]. CARD, which is the most commonly used database in the field, has well-structured AMR ontologies and constant updates. In addition to AMR gene homolog detection, CARD provides other types of functional annotations for use with other detection methods. The workflow, however, provides users only with the alignment results against CARD for downstream homolog detection and AMR prediction. Another ONT-specific software, ARGpore, utilizes a custom environmental AMR gene database, SARG, which detects AMR only up to the more general class level [22]. Other tools that are not specific to sequencing types, such as ResFinder, identify only a subset of AMR genes contributing to the acquired resistance [23]. AMRfinder [24] uses an NCBI-curated database with HMM and applies a species-specific cut-off for BLAST searches. However, the tool does not provide a taxonomic classification result and therefore is difficult to use it for metagenomic samples. HMM-based software can identify protein sequences with related functions, but with lower sequence identity, so they are useful for remote homologs or novel resistance sequence discovery [25]. There are also databases available for specific types of AMR genes, such as the CBMAR [26] database for beta-lactamase families. Nonetheless, the search function in these databases is rather limited. Depending on their detection methods, there are trade-offs in the 1) availability of genomic contexts, 2) dependency on reference databases, and 3) speed of different tools [25]. In addition, all of these AMR detection tools apply one or two independent AMR gene databases as a reference, which results in different levels of classification resolution and available resistance information. Using a single software or database therefore risks causing incomprehensive prediction output.

Here, we present MegaPath-Nano, comprehensive and accurate software designed and optimized for both metagenomic analysis and AMR detection using ONT long reads. With the ONT-specific optimizations, MegaPath-Nano is the successor to the NGS-based MegaPath, inheriting some of its key features with new functions for AMR detection. It performs 1) data cleansing, 2) taxonomic profiling, and 3) drug-level AMR detection within a single workflow. As a key feature for taxonomic profiling, MegaPath-Nano performs a global-optimization on multiple alignments and reassigns predictably misplaced reads to a single most likely species. As to perform a consistent and comprehensive AMR detection analysis, MegaPath-Nano uses a novel consensus-based approach to detect AMR, incorporating a collection of AMR software and databases. This strategy consolidates the strength of both read- and assembly-based approaches, with (1) high computational efficiency; and (2) utilizes read-level genomic contexts and alignment qualities.

We benchmarked MegaPath-Nano for taxonomic classification using real ONT sequencing data of the ZymoBIOMICS Microbial Community Standard dataset, both with or without human reads. Our results show that MegaPath-Nano has achieved the best performance on all datasets against WIMP, Kraken 2, and MetaMaps. We also sequenced five real patient isolates with known phenotypic AMR screening results. Our results show MegaPath-Nano outperformed ARMA and ARGpore on correct AMR detection in all five isolate datasets. MegaPath-Nano is as well the first publicly available software for both taxonomic classification and drug-level AMR detection using ONT long reads.

## II. METHODS

The complete workflow of MegaPath-Nano is shown in Fig. 1.

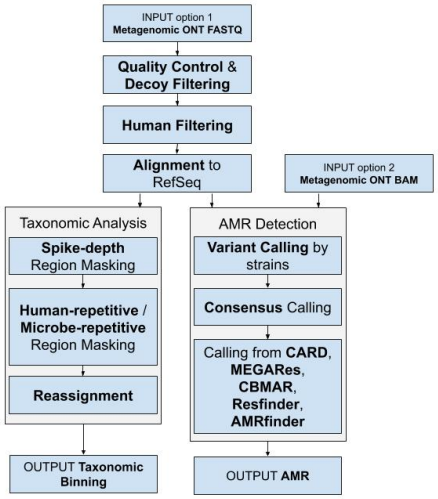


Fig. 1. Complete workflow of MegaPath-Nano

Prior to read alignment, the preprocessing procedure of MegaPath-Nano removes low-quality reads with an average base quality lower than 7, in addition to adapter trimming and read length filtering. The aligner minimap2 [27] is then used in the ONT-read mode to align the reads to the 1) human reference genome database, and 2) a customizable decoy database, including selected plasmid sequences, for further alignment-based data cleansing. The filtering criteria include an alignment score (i.e., BAM auxiliary tag 'AS') over a default threshold value set at 1,000. To specifically exclude certain taxa from the downstream analyses, users can customize the decoy database by adding the reference sequence of the unwanted taxa to the default database. The high confidence reads are then aligned to the >54,000 complete genome assemblies available in the RefSeq [28] database (release 99).

The following sections describe the two main modules of MegaPath-Nano: (A) taxonomic analysis and (B) AMR detection. Note that the AMR detection module has a standalone function, so in addition to the default workflow, users can specify any BAM file as input for AMR detection.

### A. Taxonomic analysis module

*1) Masking alignments in ambiguous regions:* Genomic regions that are difficult to align properly even with long ONT reads, including (1) spike-depth regions [11], (2) human-repetitive regions, and (3) microbe-repetitive regions, are masked to reduce potential alignment errors in the initial three steps. Spike-depth regions are defined as large regions with aligned reads from putative homologs, which are not useful for taxonomic classification. When computing depth for spike filtering, only the alignment with the best alignment score is counted. The spike-depth regions are identified and masked if their depth is greater than the expected maximum depth, $\mu + \alpha \cdot sd$, where $\mu$ is the mean depth, $sd$ is the standard variation, and $\alpha$ is the default at 6 (determined empirically). 2) The human-repetitive region masker masks all regions within the RefSeq references with similarity over 80% in the human genome reference. 3) The microbe-repetitive regions are highly similar regions in high-abundance species that can be aligned to low-abundance species. The similarity cut-off is set to $100\% - 1/8 \times r$, where $r$ is the abundance ratio of a high-abundance species to a low-abundance species. For example, a region in a high-abundance species is masked if it is aligned with a species with eight times lower abundance and over 99% ($100 - 8 \times 1/8$) similarity. These operations ensure that only informative regions in the reference genomes are retained, based on the initial alignments.

*2) Taxonomic binning by alignment reassignment algorithm:* This rationale of the reassignment algorithm was explained in MegaPath. It effectively improves the taxonomic binning specificity and abundance estimation accuracy using Illumina short reads [11]. In MegaPath-Nano, we improved the reassignment algorithm to adapt to the ONT long-read specific properties, such as the variable read length. For reads that are mapped to the repeat-masked reference sequences of multiple species with a high alignment score (AS), if species $s_1$ explains [11] species $s_2$, then the reads common to $s_1$ and $s_2$ are reassigned to $s_1$ from $s_2$. It is defined that $s_1$ explains $s_2$ if $s_1$ weakly explains $s_2$, and no species weakly explains $s_2$. The following conditions determine if $s_1$ weakly explains $s_2$: (1) $Count(s_1) - MCount(s_1, s_2) >= r(s_1)$ and (2) $UCount(s_2) < e \cdot UCount(s_1)$, where $Count(s)$ counts all reads assigned to a species, $s$, $MCount(s)$ is the read count aligned to multiple species in $s$, UCount(s) is the read count aligned uniquely only to $s$, $r$ is an arbitrary ratio of dissimilarity between species, and e is the error rate of sequencing or alignment (default set at 5%). The first condition indicates that $s_1$ has a large proportion of reads that are not samples and that are shared with $s_2$. The second condition signifies that the unique reads of $s_2$ could be caused by a sequencing error or misalignment. This reassignment algorithm is implemented recursively for each species pair that shares reads of multiple high-quality alignments. For those that are uniquely mapped or not reassigned, alignment with the highest AS alignment is selected for further analysis.

### B. AMR detection module

*1) Query sequence error correction with variant calling and consensus generation:* The filtered ONT reads are aligned to the RefSeq database by default for AMR detection in the downstream. In order to tackle the per-read sequencing error in ONT reads, consensus sequences of the aligned reads are generated using the variant calling results. Variant calling with metagenomic data is performed using the bcftools [29] with multi-allelic variant calling mode enabled. While the size of RefSeq is huge, and a typical sample usually covers only a few sequences in RefSeq, it is reasonable to call variants in the covered sequences only. However, that requires a summary

of depth at all sequence positions in RefSeq, which is computationally expensive. Instead, we rely on bedops [30] that can calculate the uncovered regions efficiently. The uncovered regions are masked with 'N' when using the bcftools for consensus. The resulting consensus sequences will be used for downstream AMR detection using three software and five databases.

*2) AMR software and database integration for comprehensive detection and output:* MegaPath-Nano comprehensively integrates major AMR databases and the respective assembly-based AMR detection software, including CARD, ResFinder, AMRfinder, MEGARes, and CBMAR. The consensus sequences generated are aligned to the AMR databases with NCBI BLAST. To search against the protein databases, such as CARD and CBMAR, open reading frame prediction and translation into protein sequences are performed with Prodigal [31]. To avoid miscalling AMR genes or misattributing resistance to non-AMR genes because of an arbitrary BLAST cut-off, a collection of cut-off settings is employed for all tools to reduce variance. The default threshold of identity score ranges from 0.9 to 0.95. The two major models for AMR detection are the homolog model and the variant model [20]. In the homolog model, AMR is detected if consensus sequences match the AMR gene reference sequences with high similarity, using BLAST. In the variant model, AMR is detected with the presence of one or more specific resistance variants. For concatenation of output matrix, only high-confident AMR results with identity score and coverage over 0.9 and 0.6 are retained. The associated AMR genes, identity scores, accession ID, and the number of supports from various databases are reported in a single tabular output for users to evaluate the confidence of each detection result.

*3) Sequencing of five patient isolates with known pathogens and AMR for benchmarking:* Regarding the library preparation for sequencing five phenotypically tested clinical isolates in the AMR detection experiment, approximately 475 ng high molecular weight DNA per sample was used for library preparation using ONT automatic library preparation device VolTRAX V2 following the VSK-VSK002 protocol. The samples were sequenced using ONT MinION sequencer with R9.4.1 flowcells. Depending on the target genome size, each sample was sequenced for 1-2 hr to obtain at least 100X data. The data were basecalled using Guppy v3.3.0 with dna_r9.4.1_450bps_hac configuration.

## III. EXPERIMENTS AND RESULTS

### A. ZymoBIOMICS Microbial Community Standards

The GridION sequencing run on ZymoBIOMICS Microbial Community Standards (Zymo) DNA set 2 [32] with even distribution (8 bacteria, each with 12% genomic DNA; along with two yeasts, each 2%) and log distribution (10 bacterial species with exponentially decreasing genomic DNA proportion) were subsampled into two datasets for benchmarking: (1) 1.2GiB 100% Zymo data, and (2) 10GiB 5% Zymo data, mixed with 95% human NA12878 data obtained from public dataset rel6 [33] to mimic the composition of clinical metagenomic samples.

MegaPath-Nano was benchmarked against What's In My Pot (WIMP), Kraken 2, Bracken and MetaMaps for taxonomic classification. In this experiment, Kraken 2 was benchmarked for taxonomic classification since it provides the assignment information of individual reads, while Bracken was benchmarked for the abundance ranking with the results of Kraken 2. It was reported that the input limit of WIMP was around 1GB, so the data were split into 1GB batches if they exceeded the constraint.

Table I shows that MegaPath-Nano consistently outperformed all other software in precision and sensitivity at the species level on the 10GiB Zymo and human DNA mixture dataset. Overall, MegaPath-Nano showed a 3.9% improvement in recall, on average. While all the software performed relatively well in terms of precision, MegaPath-Nano had the highest precision; one reason for this is the implementation of the reassignment algorithm. The difficulty of taxonomic analysis increases with the size of databases because of the larger number of similar or near-identical regions in metagenomes, such as most of the species (*Listeria monocytogenes*, *Pseudomonas aeruginosa*, *Bacillus subtilis*, etc.) in the Zymo dataset. In metagenomic data, a read having ambiguous alignments to multiple species is common even for long read data and is often misplaced. Furthermore, bacterial strains that are phylogenetically distant can potentially share conserved sequences in their genomes, e.g., ribosomal RNA sequences and conserved single-copy genes [34]. The presence of the high similarity in these conserved regions can cause the error-prone ONT reads to detect or overestimate the abundance of low-abundance microbes in the sample since the ONT reads are derived from high-abundance microbes and assigned to low-abundance or nonexisted microbes. Of note, our reassignment algorithm is unique to MegaPath (for NGS) and MegaPath-Nano (for ONT) for reassigning the reads with multiple alignments to a better-explained species in terms of global read distribution. The effectiveness of the reassignment algorithm can be better illustrated in the log distribution dataset. In the log distribution dataset, *L. monocytogenes* constituted the majority of abundance. However, 0.001% of its reads had primary alignments assigned to the orthologues region of *Enterococcus faecalis*. With the reassignment algorithm, these reads were assigned to *L. monocytogenes* instead, since *L. monocytogenes* explains *E. faecalis*, with the condition that a large number of reads in *L. monocytogenes* are divergent to that in *E. faecalis* and the unique reads supporting *E. faecalis* could stem from a sequencing error or misalignment. Additionally, no species weakly explains *L. monocytogenes* (see Method for more details) The reassignment was also well supported by a reasonably high alignment score to *L. monocytogenes*.

Having a comprehensive database to reference is always critical for classification tasks. MegaPath-Nano utilizes the complete RefSeq database and covers all the reference genomes for the known microbiome. An example was that

TABLE I
BASE-LEVEL PRECISION, SENSITIVITY AND F1-SCORE OF TAXONOMIC ASSIGNMENT TOOLS AT SPECIES LEVEL ON 10GIB MIXTURE (5% ZYMO AND 95% HUMAN METAGENOMIC DATASET)

| | | *MegaPath-Nano* | *WIMP* | *Kraken 2* | *MetaMaps* |
|---|---|---|---|---|---|
| Even | TP | **230M** | 227M | 199M | 221M |
| | FP | **9.64K** | 131K | 264K | 16.9K |
| | FN | **154M** | 187M | 461M | 242M |
| | F1 | **0.96758** | 0.96006 | 0.89577 | 0.94817 |
| | Sensitivity | **0.93724** | 0.92367 | 0.81208 | 0.90152 |
| | Precision | **0.99996** | 0.99942 | 0.99868 | 0.99992 |
| Log | TP | **231M** | 228M | 224M | 227M |
| | FP | **9.64K** | 131K | 264K | 10.7K |
| | FN | **134M** | 162M | 207M | 176M |
| | F1 | **0.97182** | 0.96553 | 0.95520 | 0.96265 |
| | Sensitivity | **0.94523** | 0.93386 | 0.91523 | 0.92802 |
| | Precision | **0.99996** | 0.99943 | 0.99882 | 0.99995 |

True positive (TP) is defined to be the number of Zymo reads assigned to any true species. False positive (FP) is the number of human reads assigned to any of the species. False negative (FN) is the number of Zymo reads that failed to be assigned to the targets, e.g., those filtered, not classified or assigned to human or nonexisting species. True negative (TN) is human reads correctly not assigned to the targets.

in benchmarking the Zymo dataset, the standard 'miniSeq+H' database for MetaMaps does not comprise *Cryptococcus neoformans*, while the standard Kraken 2 database does not contain *C. neoformans* or *Saccharomyces cerevisiae*, although they are common metagenomic species.

Note that MetaMaps by default does not process reads less than 1,000 bp long, so a significant number of Zymo reads are expected to be indiscriminately filtered due to short read length, resulting in a lower number of TP. To preserve the largest number of reads for classification of low abundance taxa, MegaPath-Nano does not enable a read-length filter by default. The value, however, can be updated by user specification. Such advanced settings might require parameter tuning depending on the dataset.

Intensive human filtering is included before data processing by any modules in MegaPath-Nano. It is carried out by aligning query sequences to a human-specific database and filtering out reads that are highly similar to human genome references. It is therefore sensitive to removing most human reads to speeds up the downstream analysis. With intensive human filtering, MegaPath-Nano exhibited the least number of misclassified human reads. Note that a small proportion of human reads were misclassified as bacterial reads consistently by all the software. This illustrates the limitation of filtering in any classification method by default settings when contaminants are extremely similar to the targets. In these cases, we suggest fine-tuning the software parameters if further precision is needed at the expense of sensitivity. Since the same set of human reads was added to the 10GiB mixture in both log and even distributions, it was shown that MegaPath-Nano, WIMP, and Kraken 2 individually have the same set of wrongly assigned human reads in their assignment of both even and log distribution. The set of FP reads of MetaMaps, however, varied from the log and even distributions, which was caused by its mapping algorithms with a composition-dependent prior [17]. In particular, A more substantial abundance of most species in the even distribution dataset might result in the higher likelihood of ambiguous human reads misassigned to the microbe genomes due to the composition-dependent prior. Also, Kraken 2 yielded the highest number of FPs, probably because the kmers of some human and Zymo reads, generated using its default setting, share a higher similarity. In comparison, the similarity between human and Zymo reads is lower. Hence, the kmer-based classification led to a higher number of human kmers misclassified as Zymo than read-based classification.

Table. II illustrates the taxonomic ranking by abundance estimation of the number of reads assigned by each software in the 1.2GiB subset of the 100% Zymo dataset with the log distribution. All the software accurately ranked the top two species with theoretical abundance contributing 98% of the sample. Since *Staphylococcus aureus* has an extremely low theoretical abundance, its expected read count is lower than one in our benchmarking dataset and therefore should be unidentified. Regarding the accuracy of abundance estimation, MegaPath-Nano had the lowest number of nonexisting species in misclassifications due to its reassignment algorithm, so the rankings are closest to the theoretical abundance in all species. Dealing with the complication of multiple-read mapping, Kraken 2 and WIMP employ the classical strategy of assigning a read with multiple mapping to their LCA, although it is linked to the less specific assignment of reads, resulting in their much lower ranking or even missing low-abundance species. Bracken post-processes the assignment from Kraken 2 and reassigns it based on probabilistic estimates of the true composition. Likewise, MetaMaps models the mapping algorithm with a composition-dependent prior. Nonetheless, in our benchmarking result, the abundance ranking suggests that both Bracken and MetaMaps' assignment may be less sensitive to low-abundance species, such as *E. faecalis*. Another straightforward method is to use a read-aligner that can omit reads with multiple alignments. However, this approach causes information loss and therefore less precise alignment results. On the abundance ranking of datasets with even distribution, all the software performed well since its theoretical abundance range is less extreme.

Comparing the overall performance in even and log distributions, it is showed that the log distribution demonstrates higher precision and recall because the minority of species constitutes the majority of genomic DNA, as shown in the theoretical abundance in Table II, so it is more favorable to the benchmarking calculation. The minority species classification performance is not well reflected using an unweighted score. However, precise classification in low-abundance species is sometimes more critical for clinical diagnosis. In fact, with the implementation of the reassignment algorithm in MegaPath-Nano, confidence in the classification of these minority species is guaranteed, even if the number of supporting reads per species is extremely low. All in all, MegaPath-Nano performed the best in taxonomic profiling and read alignment in both

TABLE II
ABUNDANCE RANKING OF EXISTING SPECIES BY TAXONOMIC ASSIGNMENT TOOLS ON 1.2GIB 100% ZYMO METAGENOMIC DATASET WITH THE LOG DISTRIBUTION

| Species | Abundance | *MegaPath-Nano* | *WIMP* | *Bracken* | *MetaMaps* |
|---|---|---|---|---|---|
| 1) *L. monocytogenes* | $8.91\times10^{-1}$ | **1** | **1** | **1** | **1** |
| 2) *P. aeruginosa* | $8.9\times10^{-2}$ | **2** | **2** | **2** | **2** |
| 3) *B. subtilis* | $8.9\times10^{-3}$ | **3** | **3** | 4 | **3** |
| 4) *S. cerevisiae* | $8.9\times10^{-3}$ | **4** | **4** | NA | **4** |
| 5) *S. enterica* | $8.9\times10^{-4}$ | **6** | 7 | **6** | **6** |
| 6) *E. coli* | $8.9\times10^{-4}$ | **7** | 8 | **7** | **7** |
| 7) *L. fermentum* | $8.9\times10^{-5}$ | **10** | 18 | NA | 12 |
| 8) *E. faecalis* | $8.9\times10^{-6}$ | **8** | 27 | 5 | NA |
| 9) *C. neoformans* | $8.9\times10^{-6}$ | **15** | 27 | NA | NA |
| 10) *S. aureus*[1] | $8.9\times10^{-7}$ | NA | **20** | NA | NA |

[1]Expected number of reads assigned is lower than 1.
NA is defined when the species is undetected in the output with the default database. In each species, the ranking result closest to the theoretical abundance ranking is in bold.

tested reference datasets.

### B. AMR detection with real patient isolates

Five real patient samples, including clinical isolates of (1) *Klebsiella pneumoniae*, (2,3) *Escherichia coli* (two sets), (4) *Proteus mirabilis*, and (5) *Proteus* spp. were phenotypically tested against a wide variety of antimicrobial agents by antimicrobial susceptibility testing [35]. MegaPath-Nano provides a list of antimicrobial drugs to which a patient might potentially be resistant, which also include those that might cause moderate resistance, conforming to the conservative practice of drug use. To evaluate AMR detection performance, MegaPath-Nano was compared against ARGpore and ARMA, which are the only software developed for AMR prediction with ONT long reads.

One of the existing software ARGpore delivers results only at the higher class level, whereas MegaPath-Nano can detect a more defined drug level AMR. In order to evaluate our performance against the other software, MegaPath-Nano was therefore benchmarked against ARMA and ARGpore at class level, then further evaluated against the more refined drug-level detection by ARMA. AMR genes detected in any assigned taxon might be caused by AMR by carriage on MGEs, which can lead to the horizontal transfer of AMR genes [25], and therefore not only taxa specific AMR was considered. In the case of a conflicting class derived from phenotypically tested drug resistance, that class is excluded in the benchmarking of class-level AMR detection. For instance, in samples 1, 2, 4, and 5, one of the target class betalactameses included resistance against the drug ceftazidime or ceftriaxone. Since the isolate is resistant to ceftazidime but susceptible to ceftriaxone, benchmarking on the betalactameses class was excluded. Since ARMA provides the read alignment only to AMR genes as the final output, for benchmarking, we further processed the output of AMRA using Bcftools and checked various AMR models based on the CARD database, as ARMA recommended to perform variant calling and consensus generation for confirmation.

Fig. 2 shows that MegaPath-Nano outperformed ARMA by an average of 28% in the number of drug-level correct classifications in all isolates. MegaPath-Nano adopts a conservative approach to aggregate the AMR detection results of similarity search and assembly-based tools with various regularly updated databases. Specifically, the prediction of MegaPath-Nano for drugs in the betalactam class was outstanding, since CBMAR provides supplementary AMR information to drugs in the beta-lactam class, such as Ceftazidime, Ceftriaxone, Cefuroxime, which are not covered by CARD. Moreover, strain-level assemblies are employed as the reference for variant calling and to generate consensus of the covered positions. As mentioned in the methodology, the query sequences are aligned to references of species, which enables species-specific, whole-genome alignment to conserve genomic contexts, e.g., positional information, regulatory sequences, and MGEs for potentially reducing FPs in AMR detection in future development. In contrast, ARMA aligns input sequences to the CARD database with minimap2 to directly estimate AMR genes in the sequences. This approach is useful for detecting AMR in reads that share high similarity to references of AMR genes but lack homology to any species' references. However, the experimental validation results show that the probability of detecting FPs is also increased accordingly.

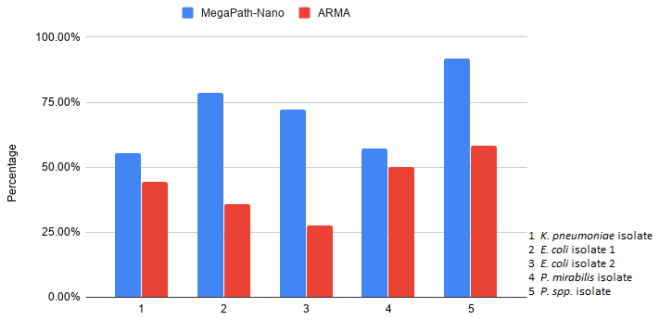


Fig. 2. Drug-level AMR detection accuracy on the five isolates, represented by the percentage of correct classification over the total number of classes or drugs. The absence of AMR indicates that the tool does not output that class or drug. If a database does not contain the information for that class or drug, it is considered misclassified. Note that any AMR detected beyond the susceptibility tested drugs was beyond the scope of this test since they could not be verified. Samples 1, 2, and 4 have 14 drugs each, while samples 3 and 5 have 18 and 12

Regarding class-level AMR detection in Fig. 3, MegaPath-Nano shows a modest improvement in accuracy than ARMA and ARGpore. In particular, its accuracy in sample 3 is higher than that of ARGpore and ARMA by 25% and 13%, respectively. Sample 3 is an *E. coli* isolate, tested with the largest number and variety of drugs[1]. In a homology search of a high-quality consensus ONT query to various databases, MegaPath-Nano ensured the lowest number of false-positives during AMR detection. In our benchmarking results, class-level polymyxin, nitrofuran and vancomycin resistance were often wrongly detected by other tools, which appeared to be

[1]All 18 drugs are listed in the supplementary materials on the GitHub page

over-sensitive. On average, MegaPath-Nano had 28% higher accuracy than ARGpore. Additionally, some of ARGpore's results were misclassified because of the lower level of completeness of its database. The benchmarking results indicate that the more general class-level AMR detection has a considerable discrepancy in accuracy across isolates because the total number of classes differs remarkably among isolates, ranging from only two (sample 5) to eight (sample 3). In our benchmarking experiment, particular classes, such as betalactam, broadly cover a large number of drugs. Some of these classes contain antibiotics of types that are not consistently resistant or susceptible, so the classes are excluded. Although not all of the software could provide AMR prediction results at the refined drug-level, it is suggested that drug-level prediction definitely provides more specific and reliable results than class-level AMR detection.

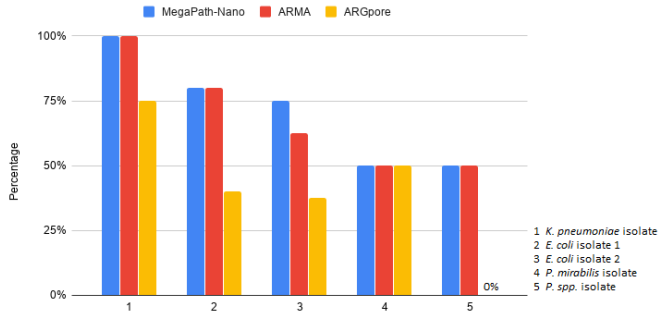


Fig. 3. Class-level AMR detection accuracy on the five isolates. For the results of ARGpore, gene types with any reads aligned are interpreted as resistance to them. The five samples have four, five, eight, four, and three classes, respectively.

### C. Runtime and memory usage

MegaPath-Nano was designed to make use of all available resources in a server to maximize its performance. In our experiments using a server with 512GB of memory, MegaPath-Nano used 24 threads and all RAM, and ran about two hours on a 1.2GiB dataset and four hours on a 10GiB dataset for taxonomic profiling. AMR detection on both datasets took 100GB RAM and one hour. MegaPath-Nano also works in resource-constrained environments by partitioning the whole index into smaller chunks, working on them separately, and aggregating the results at the end. The memory footprint can be reduced to below 64GB, trading off a one-time slowdown. The flexibility of MegaPath-Nano makes it capable of on-site metagenomic analysis and fieldwork using a portable Oxford Nanopore MinION sequencer and limited computational power.

## IV. Conclusions

The ultra-long ONT sequencing technology benefits metagenomic profiling with high alignment specificity. However, its high sequencing error per read remains a hurdle for distinguishing among closely related pathogens in lower taxonomic ranks and for refined drug-level antimicrobial resistance prediction. In this study, we presented MegaPath-Nano, the first publicly available software designed for ONT long reads to carry out 1) accurate alignment-based compositional analysis down to strain-level, and 2) drug-level AMR detection using comprehensively integrated AMR software and databases. We benchmarked against other state-of-the-art software using real sequencing data, and we achieved the best performance in both tasks. In future development, an extended version will be made available with additional optimization for ONT amplicon data. There are also plans to develop a lightweight version of MegaPath-Nano that requires the computational resources of only a laptop, which will be suitable for portable laboratory settings. MegaPath-Nano is, therefore, a well-rounded ONT metagenomic tool for clinical use in practice.

As a contribution to the community for further development, our in-house demo data of five patient isolates are available at http://www.bio8.cs.hku.hk/dataset/MegaPath-Nano/.

## References


[1] J. Hess, T. Kohl, M. Kotrová, K. Roensch, T. Paprotka, V. Mohr, T. Hutzenlaub, M. Brüggemann, R. Zengerle, S. Niemann *et al.*, "Library preparation for next generation sequencing: A review of automation strategies," *Biotechnology Advances*, p. 107537, 2020.

[2] Y. Motro and J. Moran-Gilad, "Next-generation sequencing applications in clinical bacteriology," *Biomolecular detection and quantification*, vol. 14, pp. 1–6, 2017.

[3] R. H. Deurenberg, E. Bathoorn, M. A. Chlebowicz, N. Couto, M. Ferdous, S. García-Cobos, A. M. Kooistra-Smid, E. C. Raangs, S. Rosema, A. C. Veloo *et al.*, "Application of next generation sequencing in clinical microbiology and infection prevention," *Journal of Biotechnology*, vol. 243, pp. 16–24, 2017.

[4] F. Zeeshan and S. Razzak, "Next generation sequencing and its role in clinical microbiology and molecular epidemiology," *Annals of Jinnah Sindh Medical University*, vol. 6, no. 1, pp. 31–32, 2020.

[5] W. Dunne, L. Westblade, and B. Ford, "Next-generation and whole-genome sequencing in the diagnostic clinical microbiology laboratory," *European journal of clinical microbiology & infectious diseases*, vol. 31, no. 8, pp. 1719–1726, 2012.

[6] H. K. Pedersen, V. Gudmundsdottir, H. B. Nielsen, T. Hyotylainen, T. Nielsen, B. A. Jensen, K. Forslund, F. Hildebrand, E. Prifti, G. Falony *et al.*, "Human gut microbes impact host serum metabolome and insulin sensitivity," *Nature*, vol. 535, no. 7612, pp. 376–381, 2016.

[7] B. Li, Y. Yang, L. Ma, F. Ju, F. Guo, J. M. Tiedje, and T. Zhang, "Metagenomic and network analysis reveal wide distribution and co-occurrence of environmental antibiotic resistance genes," *The ISME journal*, vol. 9, no. 11, pp. 2490–2502, 2015.

[8] S. Thankaswamy-Kosalai, P. Sen, and I. Nookaew, "Evaluation and assessment of read-mapping by multiple next-generation sequencing aligners based on genome-wide characteristics," *Genomics*, vol. 109, no. 3-4, pp. 186–191, 2017.

[9] R. R. Wick, L. M. Judd, and K. E. Holt, "Performance of neural network basecalling tools for oxford nanopore sequencing," *Genome biology*, vol. 20, no. 1, p. 129, 2019.

[10] G. A. Logsdon, M. R. Vollger, and E. E. Eichler, "Long-read human genome sequencing and its applications," *Nature Reviews Genetics*, pp. 1–18, 2020.

[11] D. Li, C.-M. Leung, C.-K. Wong, Y. Zhang, W.-C. Law, Y. Xin, R. Luo, H.-F. Ting, and T.-W. Lam, "Megapath: Low-similarity pathogen detection from metagenomic ngs data," in *2018 IEEE 8th International Conference on Computational Advances in Bio and Medical Sciences (ICCABS)*. IEEE, 2018, pp. 1–1.

[12] D. Kim, L. Song, F. P. Breitwieser, and S. L. Salzberg, "Centrifuge: rapid and sensitive classification of metagenomic sequences," *Genome research*, vol. 26, no. 12, pp. 1721–1729, 2016.

[13] D. E. Wood, J. Lu, and B. Langmead, "Improved metagenomic analysis with kraken 2," *Genome biology*, vol. 20, no. 1, p. 257, 2019.

[14] E. Doster, S. M. Lakin, C. J. Dean, C. Wolfe, J. G. Young, C. Boucher, K. E. Belk, N. R. Noyes, and P. S. Morley, "Megares 2.0: a database for classification of antimicrobial drug, biocide and metal resistance determinants in metagenomic sequence data," *Nucleic acids research*, vol. 48, no. D1, pp. D561–D569, 2020.

[15] Y. Xia, A.-D. Li, Y. Deng, X.-T. Jiang, L.-G. Li, and T. Zhang, "Minion nanopore sequencing enables correlation between resistome phenotype and genotype of coliform bacteria in municipal sewage," *Frontiers in microbiology*, vol. 8, p. 2105, 2017.

[16] H. Li and R. Durbin, "Fast and accurate short read alignment with burrows–wheeler transform," *bioinformatics*, vol. 25, no. 14, pp. 1754–1760, 2009.

[17] A. T. Dilthey, C. Jain, S. Koren, and A. M. Phillippy, "Strain-level metagenomic assignment and compositional estimation for long reads with metamaps," *Nature communications*, vol. 10, no. 1, pp. 1–12, 2019.

[18] X. Fan, Y. Yuan, J. S. Liu *et al.*, "The em algorithm and the rise of computational biology," *Statistical Science*, vol. 25, no. 4, pp. 476–491, 2010.

[19] O. N. Technologies, "Real-time detection of antibiotic-resistance genes using oxford nanopore technologies' minion," 2016.

[20] B. P. Alcock, A. R. Raphenya, T. T. Lau, K. K. Tsang, M. Bouchard, A. Edalatmand, W. Huynh, A.-L. V. Nguyen, A. A. Cheng, S. Liu *et al.*, "Card 2020: antibiotic resistome surveillance with the comprehensive antibiotic resistance database," *Nucleic acids research*, vol. 48, no. D1, pp. D517–D525, 2020.

[21] S. Juul, F. Izquierdo, A. Hurst, X. Dai, A. Wright, E. Kulesha, R. Pettett, and D. J. Turner, "What's in my pot? real-time species identification on the minion," *bioRxiv*, p. 030742, 2015.

[22] X. Yin, X.-T. Jiang, B. Chai, L. Li, Y. Yang, J. R. Cole, J. M. Tiedje, and T. Zhang, "Args-oap v2. 0 with an expanded sarg database and hidden markov models for enhancement characterization and quantification of antibiotic resistance genes in environmental metagenomes," *Bioinformatics*, vol. 34, no. 13, pp. 2263–2270, 2018.

[23] E. Zankari, H. Hasman, S. Cosentino, M. Vestergaard, S. Rasmussen, O. Lund, F. M. Aarestrup, and M. V. Larsen, "Identification of acquired antimicrobial resistance genes," *Journal of antimicrobial chemotherapy*, vol. 67, no. 11, pp. 2640–2644, 2012.

[24] M. Feldgarden, V. Brover, D. H. Haft, A. B. Prasad, D. J. Slotta, I. Tolstoy, G. H. Tyson, S. Zhao, C.-H. Hsu, P. F. McDermott *et al.*, "Validating the amrfinder tool and resistance gene database by using antimicrobial resistance genotype-phenotype correlations in a collection of isolates," *Antimicrobial agents and chemotherapy*, vol. 63, no. 11, pp. e00 483–19, 2019.

[25] M. Boolchandani, A. W. D'Souza, and G. Dantas, "Sequencing-based methods and resources to study antimicrobial resistance," *Nature Reviews Genetics*, p. 1, 2019.

[26] A. Srivastava, N. Singhal, M. Goel, J. S. Virdi, and M. Kumar, "Cbmar: a comprehensive $\beta$-lactamase molecular annotation resource," *Database*, vol. 2014, 2014.

[27] H. Li, "Minimap2: pairwise alignment for nucleotide sequences," *Bioinformatics*, vol. 34, no. 18, pp. 3094–3100, 2018.

[28] D. J. Nasko, S. Koren, A. M. Phillippy, and T. J. Treangen, "Refseq database growth influences the accuracy of k-mer-based lowest common ancestor species identification," *Genome biology*, vol. 19, no. 1, pp. 1–10, 2018.

[29] H. Li, "A statistical framework for snp calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data," *Bioinformatics*, vol. 27, no. 21, pp. 2987–2993, 2011.

[30] S. Neph, M. S. Kuehn, A. P. Reynolds, E. Haugen, R. E. Thurman, A. K. Johnson, E. Rynes, M. T. Maurano, J. Vierstra, S. Thomas *et al.*, "Bedops: high-performance genomic feature operations," *Bioinformatics*, vol. 28, no. 14, pp. 1919–1920, 2012.

[31] D. Hyatt, G.-L. Chen, P. F. LoCascio, M. L. Land, F. W. Larimer, and L. J. Hauser, "Prodigal: prokaryotic gene recognition and translation initiation site identification," *BMC bioinformatics*, vol. 11, no. 1, p. 119, 2010.

[32] S. M. Nicholls, J. C. Quick, S. Tang, and N. J. Loman, "Ultra-deep, long-read nanopore sequencing of mock microbial community standards," *Gigascience*, vol. 8, no. 5, p. giz043, 2019.

[33] M. Jain, S. Koren, K. H. Miga, J. Quick, A. C. Rand, T. A. Sasani, J. R. Tyson, A. D. Beggs, A. T. Dilthey, I. T. Fiddes *et al.*, "Nanopore sequencing and assembly of a human genome with ultra-long reads," *Nature biotechnology*, vol. 36, no. 4, pp. 338–345, 2018.

[34] Z. research, "Instruction manual of zymobiomics microbial community standardii (log distribution)," 2018.

[35] L. B. Reller, M. Weinstein, J. H. Jorgensen, and M. J. Ferraro, "Antimicrobial susceptibility testing: a review of general principles and contemporary practices," *Clinical infectious diseases*, vol. 49, no. 11, pp. 1749–1755, 2009.