Computers in Biology and Medicine 186 (2025) 109743

Contents lists available at ScienceDirect

Computers in Biology and Medicine

journal homepage: www.elsevier.com/locate/compbiomed

Adaptive sequence alignment for metagenomic data analysis

Sami Pietilä [a,1], Tomi Suomi [a,1], Niklas Paulin [a,1], Asta Laiho [a], Yannes S. Sclivagnotis [b], Laura L. Elo [a,c,*]

[a] Turku Bioscience Centre, University of Turku and Åbo Akademi University, FI-20520, Turku, Finland
[b] Orion Corporation, FI-20380, Turku, Finland
[c] Institute of Biomedicine, University of Turku, FI-20520, Turku, Finland

ARTICLE INFO

Keywords:
Metagenomics
Sequence alignment
Sequence assembly
Taxonomic identification

ABSTRACT

With advances in sequencing technologies, the use of high-throughput sequencing to characterize microbial communities is becoming increasingly feasible. However, metagenomic assembly poses computational challenges in reconstructing genes and organisms from complex samples. To address this issue, we introduce a new concept called Adaptive Sequence Alignment (ASA) for analyzing metagenomic DNA sequence data. By iteratively adapting a set of partial alignments of reference sequences to match the sample data, the approach can be applied in multiple scenarios, from taxonomic identification to assembly of target regions of interest. To demonstrate the benefits of ASA, we present two application scenarios and compare the results with state-of-the-art methods conventionally used for the same tasks. In the first, ASA accurately detected microorganisms from a sequenced metagenomic sample with a known composition. The second illustrated the utility of ASA in assembling target genetic regions of the microorganisms. An example implementation of the ASA concept is available at https://github.com/elolab/ASA.

1. Introduction

The emergence of high-throughput sequencing methods has facilitated the development of metagenomics, which involves extracting nucleotide sequences from multiple organisms, such as bacteria, in a particular environment [1–3]. The approach has enabled profiling of entire microbial populations. However, the data analysis poses challenges and requires diverse methods to obtain the necessary microbial profiles [4]. This is particularly true for short read sequencing, which is currently the most common approach in the field. The effectiveness of these methods varies depending on factors such as sequencing length, coverage, quality, and sample complexity, all of which can cause misinterpretations [5]. New techniques are needed to enhance the accuracy and reliability of the results.

Current methodologies for analyzing short shotgun-sequenced reads, and assembling them into larger contiguous regions (contigs), predominantly employ graph-based techniques such as De Bruijn Graphs (DBG), Overlap-Layout-Consensus (OLC) graphs, or greedy methods, as outlined by Ref. [5]. While taxonomic profiling is a prominent application of assembly, alternative profiling techniques that do not require assembly exist. One such technique involves constructing a database of short marker genes sourced from public databases, which enables the identification of species in a given sample [6]. Another prevalent approach, particularly employed in targeted sequencing, utilizes the highly conserved bacterial 16S rRNA gene [2]. Depending on the specific application, high-confidence positive identifications are often crucial. However, it is currently common to encounter a fair number of false positives in reports, which can hinder the attainment of biological research goals [7].

In this work, we introduce a new approach called Adaptive Sequence Alignment (ASA), which iteratively adapts a set of reference sequences to match metagenomic sample content, together with an example implementation (Fig. 1). Unlike other available reference-guided approaches [8–11], ASA is based on iterative partial alignments. To demonstrate the efficacy of ASA, we show its applicability in two distinct research scenarios and compare the results to those of state-of-the-art methods conventionally used for the same tasks. First, we show how ASA can accurately identify the content of a metagenomic sample with low false positive rates, which is crucial, for example, to rule out the presence of potential pathogens in a sample [12]. This is demonstrated

* Corresponding author. Turku Bioscience Centre, University of Turku and Åbo Akademi University, FI-20520, Turku, Finland.
E-mail address: laura.elo@utu.fi (L.L. Elo).
[1] Shared first author.

https://doi.org/10.1016/j.compbiomed.2025.109743
Received 23 May 2024; Received in revised form 20 January 2025; Accepted 21 January 2025
Available online 26 January 2025

through the use of ASA for the assembly of complete 16S rRNA genes, which facilitates the inference of sample taxonomy. (Conventionally, assembling the 16S rRNA gene has posed challenges. Generic *de novo* assembly tools are optimized for this task due to factors such as the high degree of sequence similarity [13].) Second, we demonstrate how ASA can assemble genes and their neighboring regions from a metagenomic sample, which has multiple important applications in biological research [14].

The novelty of ASA lies in its iterative alignment and consensus sequence generation process, which allows for progressive refinement of sequences. This dynamic adaptation of reference sequences through multiple iterations enhances the accuracy of taxonomic identification and gene assembly, setting ASA apart from traditional, alignment-based methods.

2. Materials and methods

2.1. Adaptive sequence alignment (ASA) concept

Our ASA concept is based on an iterative procedure, which first aligns the read sequences against a reference sequence and then forms a consensus sequence to be used as the reference for subsequent alignment iteration.

If the initial reference sequence only partially matches the sample content, the resulting consensus sequence in the alignment iteration typically contains both continuous regions of known nucleotides and regions of unknown nucleotides marked with the letter N (Fig. 1). The unknown nucleotides are found in regions where the reference and the read sequences do not match. With modification to an aligner, it is possible to enable partial alignments where a given proportion of a read sequence aligns to known nucleotides while the rest aligns to Ns. The scoring scheme modification implemented in ASA fills partial references (with regions of Ns) with known nucleotides as reads are aligned at the border of known and unknown nucleotide regions. Repeating the alignment and consensus steps causes the aligner to iteratively build sequences that match progressively better to the sample sequences. Furthermore, sequence assembly beyond the original reference is achieved by appending Ns at the beginning and end of the reference. The uncertainty arising from unknown nucleotides is controlled by limiting the percentage of allowed unknown nucleotides in an alignment of each read. The iterations are stopped when the consensus sequence stops transforming or when the iteration limit (–rounds parameter) is reached. The final consensus sequence reflects the sample content precisely within the limitations posed by read length and coverage. By default, ASA assembles consensus sequences up to the length of a specified reference sequence. If an assembly exceeding the reference length is required, ASA can continuously extend the alignment length in each iteration. In such cases, the length of the alignment is limited either by the iteration limit set by the user or by the availability of read sequences in the dataset, which can be aligned partially, thus extending the sequence further. Additionally, the alignment process will halt if the two consecutive iteration rounds have resulted in an identical consensus sequence, indicating that consensus sequence stops changing. The decision whether to extend the consensus in each round depends on the specific use case. For instance, in taxonomic identification, it may suffice to produce consensus sequences up to the original reference length. However, for assembly purposes, progressive extension can be advantageous, especially in achieving extensive lengths. This facilitates the determination of adjacent sequences to the seed reference sequence.

To implement the ASA concept, a sequence aligner and a tool to build a consensus sequence from the alignment are required. Here, we used a modified version of Bowtie 2 (version 2.2.3) [15] and Samtools (version 0.1.19) [16] for alignments and building of consensus sequences. The Bowtie 2 aligner has a parameter, `--np`, for giving a score penalty for non-A/C/G/T nucleotides. We modified the Bowtie 2 source code by adding parameter `--na` to give a positive score for non-A/C/G/T nucleotides. The new parameter is used in combination with another parameter, `--n-ceil`, which controls the maximum number of permitted non-A/C/G/T characters in an alignment. This allows Bowtie 2 to align reads to unknown nucleotides if the surrounding reference nucleotides support the alignment decision. The default value for `--n_ceil` is 0.3. The iterative ASA workflow was implemented as a Python (version 3) program.

2.2. Benchmark data

To illustrate the applicability of the ASA concept for metagenomic data analysis, we used the public Mock Bacteria Archaea Community (MBARC-26) [17] metagenome benchmark dataset, which is composed of 26 different microorganisms with varying abundance (Table 1). The MBARC-26 Illumina shotgun metagenome sequencing dataset was downloaded from the National Center for Biotechnology Information (NCBI) Sequence Read Archive with accession number SRX1836716, and the complete species information was downloaded as GenBank files.

2.3. Evaluation of taxonomic identification

ASA can be applied to taxonomic identification by first choosing a taxonomic reference database and performing an initial alignment against the references and then choosing top matching references that are iteratively transformed into sequences that match the sample content. The results can be represented in a phylogenetic tree, which can be analyzed manually or with the help of heuristics. To ensure the performance of ASA in general, the number of references in the initial alignment phase passed to the iterative alignment phase should be limited to a selected amount of top matching references (the `--top_refs` parameter in the current ASA implementation). The chosen cutoff is a tradeoff between computing time and confidence that the chosen references generate enough initial alignments to discriminate between different microorganisms in the sample.

We chose a cutoff of 150 references, which yielded 142 consensus sequences covering the expected sample diversity of 26 species. Given the limited number of species (26 in total), numerous consensus sequences were either fully or nearly identical (Fig. 2). To achieve

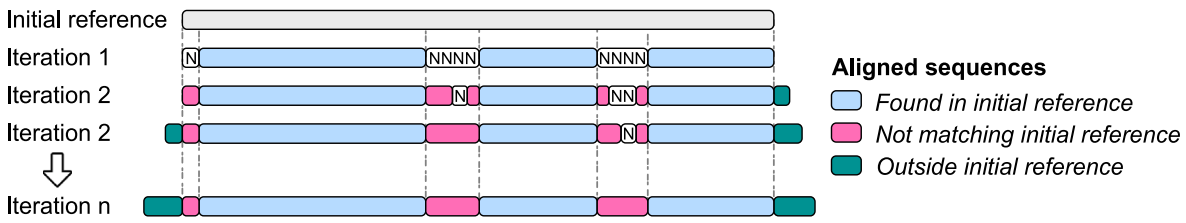

Fig. 1. Illustration of the ASA concept. The ASA concept transforms an initially non-matching reference into a reference that fully matches the sample data. This is achieved by adding unknown nucleotides (N) to the ends of a sequence and encouraging alignments to occur on top of them. The unknown nucleotides are found in regions where the reference and the read sequences do not match. Repeating the alignment and consensus steps enables ASA to iteratively build a contig from the reads.

Table 1
MBARC-26 dataset composition according to read mapping to reference genomes and according to molarity.

Organism	Reference	Molarity	Genome copies per µl	Illumina		Illumina
				% mapped genome	% mapped chromosome	% genome covered
T. roseus	NC_018014.1	4.79E-15	155	2.07	2.13	83.95
C. glutamicum	NC_003450.3	4.91E-16	10	0.30	0.31	99.34
N. dassonvillei	NC_014210.1	2.67E-17	6	0.00	0.00	0.54
O. uli	NC_014363.1	3.40E-14	304	2.26	2.32	100
S. rotundus	NC_014168.1	1.22E-14	149	1.41	1.45	99.96
E. vietnamensis	NC_019904.1	1.26E-15	41	0.62	0.64	99.29
M. silvanus	NC_014212.1	4.38E-14	213	8.56	7.82	99.97
C. perfringens	NC_008261.1	5.20E-16	39	0.42	0.43	99.53
C. thermocellum	NC_009012.1	4.40E-16	15	0.43	0.44	99.11
D. acidiphilus	NC_018068.1	2.68E-14	409	15.11	14.75	99.98
D. meridiei	NC_018515.1	9.89E-15	261	4.61	4.74	99.74
D. gibsoniae	NC_021184.1	2.93E-14	535	6.91	7.11	99.93
S. pyogenes	NC_002737.2	1.53E-15	16	0.43	0.44	99.06
T. composti	NC_019897.1	2.39E-16	7	8.50	8.18	99.82
E. coli	NC_000913.3	3.90E-16	16	0.18	0.19	98.87
F. aurantia	NC_017033.1	2.84E-14	317	3.99	4.11	99.95
H. baltica	NC_012982.1	1.78E-14	400	8.16	8.20	99.99
P. stutzeri	NC_019936.1	1.21E-14	164	1.55	1.56	99.23
S. bongori	NC_015761.1	1.72E-16	31	0.14	0.15	99.31
S. enterica	NC_010067.1	6.69E-16	40	0.52	0.54	99.63
S. smaragdinae	NC_014364.1	2.78E-14	467	11.39	11.72	100
F. pennivorans	NC_017095.1	6.21E-14	672	11.26	11.58	100
C. akajimensis	NC_014008.1	6.85E-15	144	3.41	3.50	99.78
H. ruber	CP003050.1	2.34E-14	614	1.75	1.80	99.92
N. gregoryi	NC_019792.1	3.01E-14	569	2.46	2.53	99.89
N. occultus	NC_019974.1	2.15E-14	933	3.55	3.35	100

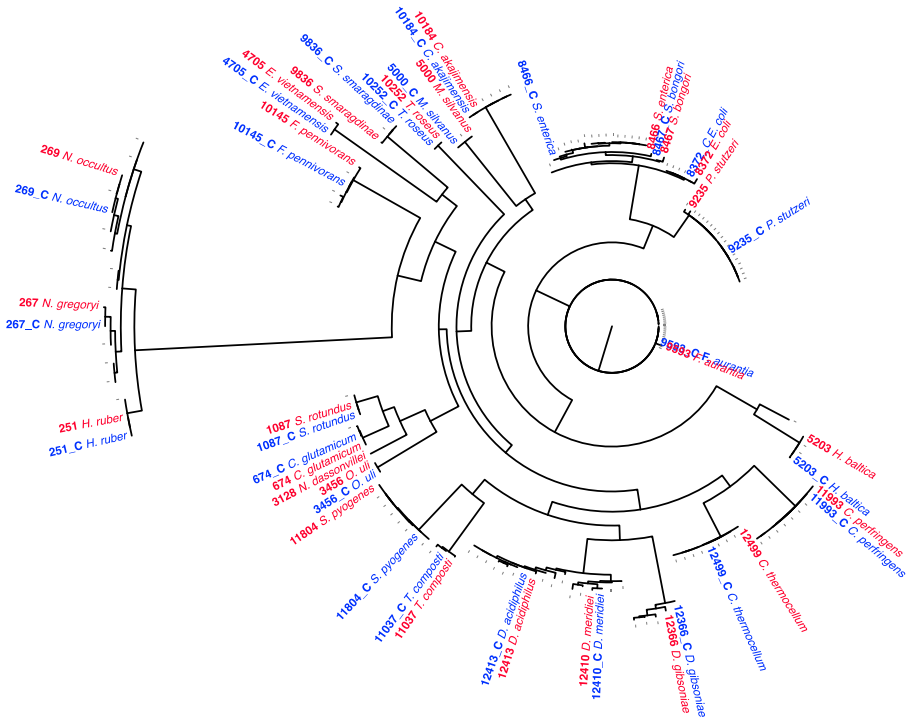

Fig. 2. A phylogenetic tree containing the final consensus sequences and their Ribosomal Database Project reference sequences in the MBARC-26 dataset. Consensus sequences, originating from a MBARC-26 species reference, are marked with a "_C" postfix. Those consensus sequences derived from other references are delineated by gray dashes for clarity. The reference sequences are marked with red color, while blue indicates the closest consensus sequence. (For interpretation of the references to color in this figure legend, the reader is referred to the Web version of this article.)

comprehensive identification, the cutoff value should exceed the anticipated number of species in the sample. Including reference sequences improves the chances of acquiring suitable seed sequences that match the actual sequences present. While having 150 references may be excessive for MBARC-26, it serves well for demonstration purposes. Although an abundance of seed sequences is not detrimental from a methodological perspective, it does impose a computational burden. The maximum number of iteration rounds (the `--rounds` parameter) is an estimate of how many iterations are needed to build a gene region of the required length. Performing 20 iterations was sufficient to discover the 16S rRNA genes that are roughly 1500 nucleotides long. In most instances, the consensus sequence is assembled to the full gene length,

usually without containing Ns. By examining the consensus sequences at each iteration, the assembly rate can be assessed, offering an estimate of the number of rounds necessary. Although the assembly process typically terminates automatically upon achieving full assembly, there are instances where the assembly result may oscillate without progressing. In such cases, the iteration limit ('rounds' parameter) ensures that the algorithm terminates.

To identify the microorganisms in the MBARC-26 benchmark dataset, the whole 16S rRNA genes were assembled from short paired-end 150bp reads using Ribosomal Database Project (RDP) release 11, update 5 [18] as the reference sequence database. With ASA, a maximum of 150 top matching RDP sequences were allowed as candidate references for the iterative processing, and a maximum of 20 alignment iterations were done. A phylogenetic tree containing the final consensus sequences and their RDP reference sequences was produced using FastTree (v2.1) [19] with sequences aligned using ClustalW (v2.1) [20]; we refer to this tree as a *consensus-reference* tree. The taxonomic annotation was determined heuristically by traversing the *consensus-reference* tree backwards from the leaves until the closest reference was found within the threshold of not crossing to another genus. The threshold was determined by calculating the minimum phylogenetic distance between genera in a given family from RDP database sequences.

To assess the performance of the identifications, we calculated the precision (TP/[TP + FP]), recall (TP/[TP + FN]), and F_1-score (2 * [precision * recall]/[precision + recall]), where TP, FP, and FN denote the true positives, false positives, and false negatives, respectively. TP was defined as a correct genus identification based on the closest reference sequence, whereas FP referred to an identification that was not among the 26 microorganisms in the MBARC-26 dataset. FN was defined as a genus present in the MBARC-26 dataset that remained unidentified.

Finally, we compared the ASA identification results with those obtained using four widely used tools for metagenome taxonomic profiling, namely MetaPhlAn4 (v4.1.0) [6,21], mOTUS [22], Kraken2 [23], and MetaPhyler (v1.25) [24], using their default parameters. In addition, we used MEGAHIT (v1.1.3) and metaSPAdes (v3.12.0) to generate *de novo* metagenomic assemblies from the MBARC-26 data, using their default parameters for paired-end assembly. These assemblies were used to evaluate the ability of *de novo* methods to assemble 16S regions, which can be used for taxonomic identification. To target the 16S gene V4 regions in the contigs, we used the Illumina standard primer sequences (GTGCCAGCMGCCGCGGTAA [forward], GGACTACHVGGGTWTCTAAT [reverse, complementary]). The V4 regions were identified directly from the contigs by using a custom Python script (identify-v4-16S-region.py).

2.4. Evaluation of assembling target genetic regions of microorganisms

To assemble target genetic regions of microorganisms, we selected ten species-specific genes from the two most abundant species in the MBARC-26 dataset (*D. acidiphilus* and *S. smaragdinae*) based on their good coverage in the dataset and obtained the corresponding gene sequence from NCBI (identifiers NC_018068.1 and NC_014364.1). The goal was to target genes with a good coverage, thus providing suitable sequences for assembly. From *D. acidiphilus*, we selected the following protein products with identifiers WP_014826593.1, WP_014829047.1, WP_014828152.1, WP_014829474.1, WP_014826879.1, WP_014829310.1, WP_014825176.1, and WP_014828656.1. From *S. smaragdinae*, we selected WP_013253148.1 and WP_013256183.1. We then removed 300 nucleotides from the beginning of these genes and used the truncated genes as seed sequences for ASA to test whether it could reassemble the removed nucleotides using the metagenome sequencing reads.

In addition, we used MEGAHIT (v1.1.3) [25] and metaSPAdes (v3.12.0) [26] to generate *de novo* metagenomic assemblies from the MBARC-26 data, using their default parameters for paired-end assembly, to confirm the content of the removed 300 nucleotides.

3. Results and discussion

We assessed the performance and applicability of the ASA concept for two use cases. First, we applied it for taxonomic identification using the MBARC-26 benchmark data composed of 23 bacterial and 3 archaeal strains (10 phyla, 14 classes). Second, we used it to assemble specific target regions of genes in the dataset. Finally, we compared the performance of ASA with state-of-the-art methods used for these tasks. ASA can offer advantages over alignment-free and Markov model-based algorithms, such as POSSM and iDeLUCS, due to its unique iterative alignment and consensus sequence generation process. This potential is reflected in the comprehensive comparisons with other established methods.

In general, ASA is a highly versatile approach and has multiple plausible applications. The requirement for seed sequences can be described in a very generic manner, and ASA is not restricted to any specific sequence database, such as the RDP. We chose RDP as our reference database due to its relevance and widespread use in similar metagenomic studies, while also acknowledging the RefSeq database as a valuable resource. Alternative databases are also available, such as SILVA, for constructing phylogenetic consensus trees to discern sample composition. The principle of using seed sequences (a database) is that the transformation (i.e., extension) of a known seed sequence to the closest sequence present in the sample uncovers the required information, which varies depending on application. Since the main characteristics of ASA primarily stem from the quality of alignment, it may be influenced by factors like sequencing quality and depth—especially when sequencing depth is so low that the region of interest is not fully covered by read sequences, preventing assembly altogether. However, in the absence of close reference sequences, accuracy may be compromised.

In addition to providing useful use cases on its own, ASA can be used in combination with other methods to complement or validate the results. The sequence assembly is particularly suitable for discovering the surrounding content of a given reference sequence. One could, for example, take the ends of assembled contigs and use ASA to bridge gaps between contigs where the assembly has stopped. The candidate reference sequences provide ASA with insights inaccessible to *de novo* assemblers, potentially enabling ASA to achieve assemblies beyond the reach of *de novo*-based systems especially in scenarios involving highly similar or repeating genetic regions. This warrants further investigation in future research. In addition to its application in metagenomic studies, we expect that the ASA concept can also be employed in genomic data analysis.

3.1. Evaluation of taxonomic identification

First, we assessed the utility of ASA for microbial identification in sequenced metagenomic samples. We used the previously published MBARC-26 benchmark dataset, which contains 26 microorganisms from 23 genera, with varying abundances. The accuracy of ASA was compared with widely used methods for metagenome taxonomic profiling: MetaPhlAn4 [6,21], mOTUS [22], MetaPhyler [24], and Kraken2 [23] (with the maxikraken2_1903_140 GB database).

ASA, mOTUS, MetaPhlAn4, MetaPhyler, and Kraken2 detected 19, 20, 20, 9, and 23 true positives, respectively, out of the known 23 genera present in the sample, while 4, 7, 7, 60, and 1532 false positives were reported, respectively (Fig. 3a). This highlights the ability of the iterative ASA concept to reconstruct sequences with very high accuracy, with a precision of 0.83, recall of 0.83, and F_1 score of 0.83, compared to mOTUS (precision 0.74, recall 0.87, F1 score 0.80), MetaPhlAn4 (precision 0.74, recall 0.87, F_1 score 0.80), MetaPhyler (precision 0.13, recall 0.39, F_1 score 0.20) and Kraken2 (precision 0.02, recall 1, F_1 score 0.03) (Fig. 3b). ASA detected few false positives, which is important, for instance, in ruling out the presence of falsely identified pathogens that might have been found by current state-of-the-art methods.

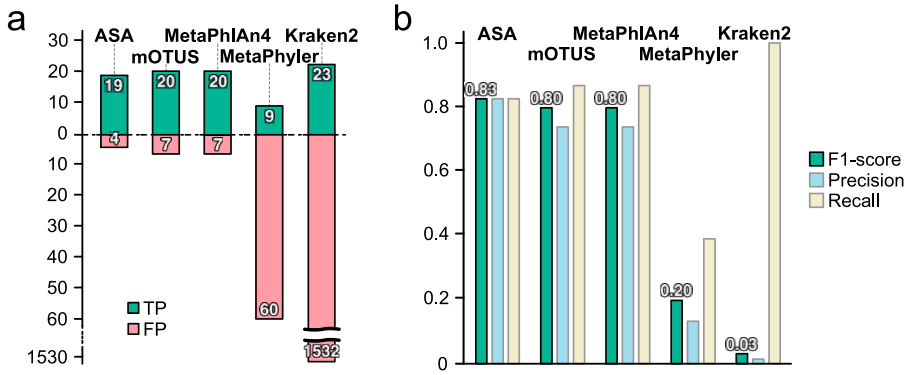

Fig. 3. Performance of ASA compared to mOTUS, MetaPhlAn4, MetaPhyler, and Kraken2 tools to produce correct identifications in the MBARC-26 benchmark dataset. (**a**) Number of true positives (TP) and false positives (FP) produced by each approach. (**b**) Precision, recall, and F_1-score of each approach.

A closer inspection of the ASA results illustrated the benefits of the proposed reference-based concept versus those of contig assembly-based approaches like OLC and DBG. The constructed *consensus-reference* tree of the 142 identified consensus sequences, together with the reference sequences of the known 26 species, showed that the consensus sequences were clustered very closely to 25 out of 26 reference sequences (Fig. 2). Only the *Nocardiopsis dassonvillei* reference did not have any nearby consensus sequences. According to the original publication [17], only a minuscule amount of *N. dassonvillei* sequences were present in the data. These results imply that the reference-based concept is able to distinguish between similar genes as the consensus sequences accurately represent the true sample content.

Of note, besides *N. dassonvillei* (3128), the heuristics failed to detect *Desulfotomaculum gibsoniae* (12366), *Echinicola vietnamensis* (4705), and *Terriglobus roseus* (10252) (Fig. 2). However, there are consensus sequences very close to references 12366, 4705, and 10252, as illustrated in the *consensus-reference tree* (Fig. 2). From the *consensus-reference tree*, it can be determined manually that these genera likely exist in the sample, indicating identifications missed by the automatic heuristic. There is potential to enhance this heuristic strategy to achieve performance levels comparable to manual inspection. This could involve refining the thresholds more precisely to define genus boundaries, which are automatically derived from the RDP database. These thresholds determine the permissible degree of sequence divergence within a genus boundary. Currently, we recommend considering manual inspection of the phylogenetic tree to achieve the highest accuracy when precision is paramount.

One of the main advantages of ASA is that it can assemble the 16S sequences of bacteria than are not precent in the RDP database. For the mock community used, all the microorganisms had a representative in the database. If a community contains a novel species for which a reference does not exist in the database, ASA would still be able to assemble the sequence by using the reference sequence of a closely related species.

Finally, we manually extracted those 25 consensus sequences from the *consensus-reference tree* that had a close-by MBARC-26 reference sequence. We then searched for V4 regions from the consensus sequence, which indicates the presence of 16S gene. The V4 region was found from all the consensus sequences, as expected. Similarly, we conducted a search for V4 regions in the metaSPAdes and MEGAHIT contigs. In contrast to the ASA results, the V4 region was found only in 13 MEGAHIT and 8 metaSPAdes contigs, indicating that the contig assembly may not differentiate well between similar genes.

3.2. Evaluation of assembling target genetic regions of microorganisms

Next, we tested the utility of ASA in assembling specific genetic regions. To do so, we removed nucleotides from the beginning of the ten species-specific genes, selected with good sequencing coverage, from the MBARC-26 dataset and then tested whether we can reassemble the removed nucleotides using the metagenome sequencing reads. We found that ASA was successful in assembling eight out the ten genes, using the default `--conseg_nceil` value of 0.3. Notably, ASA was able to recover the exact nucleotides that were removed. For the other two genes, the alignment initially failed, which led us to hypothesize that the reason was related to homologous genes, which can result in chimeric sequence assemblies. Indeed, with these two genes, a successful assembly was obtained by decreasing the number of allowed non-aligning nucleotides from the default 30 %–10 % (by changing the parameter `--conseq_nceil` from 0.3 to 0.1), making the iterative alignment expand more carefully. Therefore, if the desired assembly is not achieved in particular cases, we recommend adjusting the parameter value from 0.3 to 0.1 until progressive assembly is achieved, while also checking for the presence of chimeric assembly, if possible.

Finally, we compared the ASA results to those obtained by assembling the reads into contigs with the popular MEGAHIT and metaSPAdes tools and searched the contigs for the same ten genes using BWA-MEM [27]. All the tools generated a consistent assembly sequence, confirming the correctness of the assembly produced by ASA.

Although all approaches resulted in a successful assembly, the candidate reference sequences provide information that is not available to tools like MEGAHIT and metaSPAdes. This information may help distinguish between highly similar or repeating genetic areas that are typically very hard to *de novo* assemble correctly.

4. Conclusions

The ASA concept introduces a new type of approach for genomic data processing. It involves an initial identification step according to references, followed by iterative extension and modification of the generated candidate reference sequences through partial alignments until they match the sample content. Such an approach has not been previously explored in other methods. The iterative alignment strategy is particularly well-suited for analyses where accurate sequence alignment is essential for understanding gene expression patterns across different conditions. The flexibility of ASA in handling diverse datasets further enhances its potential in this field, rendering it a reliable tool for comprehensive transcriptomic studies. Rigorous testing has shown the effectiveness of our approach in minimizing false positives and enhancing overall accuracy.

ASA can be described as a hybrid approach that is capable of both taxonomic identification and assembly of target regions of interest. When compared to other microbial identification tools, such as MetaPhlAn4 and MetaPhyler, ASA demonstrated an exceptionally low false positive rate (Fig. 3). Furthermore, the assembled gene sequences generated by ASA were identical to those produced by MEGAHIT and

metaSPAdes. Our evaluation criteria are robust and include comprehensive comparisons with well-established methods, providing a clear and reliable assessment of ASA's performance in taxonomic binning.

In conclusion, this study introduces a new alignment-based approach for genetic sequence assembly and taxonomic profiling, called ASA. It is robust and in principle capable of handling datasets of varying sizes, including MiSeq and HiSeq data. However, empirical assessment with dataset of different sizes was beyond the scope of the current work and remains an area for future research. The ASA method can significantly advance metagenomic analysis by enhancing the accuracy of taxonomic identification and gene assembly. The iterative alignment strategy provides a reliable framework for reconstructing sequences in complex samples, with few false positives and high fidelity. The impact of our findings is substantial, as they contribute to more accurate microbial community profiling and better understanding of microbial functions and interactions. While ASA is presented here as a conceptual framework, we provide an implementation of the algorithm. Notably, the current implementation of the iterative alignment employed by ASA demands substantial computational resources, but in future developments, we anticipate improving the efficiency. In addition, at present, the taxonomic identification procedure yields numerous fully or nearly identical consensus sequences. Specifically, 142 consensus sequences were produced, while the sample contained only 26 species. As a future development, this issue can be addressed by incorporating a sequence de-replication step into the analysis.

To facilitate broader accessibility and collaboration, we have made the ASA implementation available as open source, accessible at https://github.com/elolab/ASA. This resource offers a foundation for further research and applications within the scientific community.

CRediT authorship contribution statement

Sami Pietilä: Writing – review & editing, Writing – original draft, Methodology, Formal analysis, Conceptualization. **Tomi Suomi:** Writing – review & editing, Writing – original draft, Visualization. **Niklas Paulin:** Writing – original draft, Software, Formal analysis. **Asta Laiho:** Writing – review & editing, Supervision. **Yannes S. Sclivagnotis:** Writing – review & editing, Conceptualization. **Laura L. Elo:** Writing – review & editing, Writing – original draft, Supervision, Resources.

Ethics statement

Analysis was conducted using publicly available non-human data, which does not involve private or sensitive information.

Declaration of competing interest

None declared.

Acknowledgements

Prof. Elo reports grants from the European Union's Horizon 2020 research and innovation programme (955321), Research Council of Finland (310561, 314443, 329278, 335434, 335611 and 341342), and Sigrid Juselius Foundation, during the conduct of the study. Our research is also supported by Biocenter Finland, and ELIXIR Finland. The authors wish to thank Olli Uhlgren for Graphical Abstract.

References

[1] A.G. Clooney, F. Fouhy, R.D. Sleator, A. O' Driscoll, C. Stanton, P.D. Cotter, et al., Comparing apples and oranges?: next generation sequencing and its impact on microbiome analysis, PLoS One 11 (2016) e0148028.
[2] A.E. Pérez-Cobas, L. Gomez-Valero, C. Buchrieser, Metagenomic approaches in microbial ecology: an update on whole-genome and marker gene sequencing analyses, Microb. Genom. 6 (2020), https://doi.org/10.1099/mgen.0.000409.
[3] S. Martin, D. Heavens, Y. Lan, S. Horsfield, M.D. Clark, R.M. Leggett, Nanopore adaptive sampling: a tool for enrichment of low abundance species in metagenomic samples, Genome Biol. 23 (2022) 11.
[4] E.R. Mardis, DNA sequencing technologies: 2006–2016, Nat. Protoc. 12 (2017) 213–218.
[5] Q. Chen, C. Lan, L. Zhao, J. Wang, B. Chen, Y.-P.P. Chen, Recent advances in sequence assembly: principles and applications, Brief Funct Genomics 16 (2017) 361–378.
[6] N. Segata, L. Waldron, A. Ballarini, V. Narasimhan, O. Jousson, C. Huttenhower, Metagenomic microbial community profiling using unique clade-specific marker genes, Nat. Methods 9 (2012) 811–814.
[7] Z. Sun, J. Liu, M. Zhang, T. Wang, S. Huang, S.T. Weiss, et al., Removal of false positives in metagenomics-based taxonomy profiling via targeting Type IIB restriction sites, Nat. Commun. 14 (2023) 1–12.
[8] K. Schneeberger, S. Ossowski, F. Ott, J.D. Klein, X. Wang, C. Lanz, et al., Reference-guided assembly of four diverse Arabidopsis thaliana genomes, Proc. Natl. Acad. Sci. U.S.A. 108 (2011) 10249–10254.
[9] E. Bao, T. Jiang, T. Girke, AlignGraph: algorithm for secondary de novo genome assembly guided by closely related references, Bioinformatics 30 (2014) i319–i328.
[10] H.E.L. Lischer, K.K. Shimizu, Reference-guided de novo assembly approach improves genome reconstruction for related species, BMC Bioinf. 18 (2017) 1–12.
[11] I. Rosenboom, T. Scheithauer, F.C. Friedrich, S. Pörtner, L. Hollstein, M.-M. Pust, et al., Wochenende - modular and flexible alignment-based shotgun metagenome analysis, BMC Genom. 23 (2022) 748.
[12] I. Maljkovic Berry, M.C. Melendrez, K.A. Bishop-Lilly, W. Rutvisuttinunt, S. Pollett, E. Talundzic, et al., Next generation sequencing and bioinformatics methodologies for infectious disease research and public health: approaches, applications, and considerations for development of laboratory capacity, J. Infect. Dis. 221 (2019) S292–S307.
[13] C. Yuan, J. Lei, J. Cole, Y. Sun, Reconstructing 16S rRNA genes in metagenomic data, Bioinformatics 31 (2015) i35–i43.
[14] H. Hasegawa, E. Suzuki, S. Maeda, Horizontal plasmid transfer by transformation in Escherichia coli: environmental factors and possible mechanisms, Front. Microbiol. 9 (2018), https://doi.org/10.3389/fmicb.2018.02365.
[15] B. Langmead, S.L. Salzberg, Fast gapped-read alignment with Bowtie 2, Nat. Methods 9 (2012) 357–359.
[16] H. Li, B. Handsaker, A. Wysoker, T. Fennell, J. Ruan, N. Homer, et al., The sequence alignment/map format and SAMtools, Bioinformatics 25 (2009) 2078–2079.
[17] E. Singer, B. Andreopoulos, R.M. Bowers, J. Lee, S. Deshpande, J. Chiniquy, et al., Next generation sequencing data of a defined microbial mock community, Sci. Data 3 (2016) 160081.
[18] J.R. Cole, Q. Wang, J.A. Fish, B. Chai, D.M. McGarrell, Y. Sun, et al., Ribosomal Database Project: data and tools for high throughput rRNA analysis, Nucleic Acids Res. 42 (2014) D633–D642.
[19] M.N. Price, P.S. Dehal, A.P. Arkin, FastTree 2 – approximately maximum-likelihood trees for large alignments, PLoS One 5 (2010) e9490.
[20] M.A. Larkin, G. Blackshields, N.P. Brown, R. Chenna, P.A. McGettigan, H. McWilliam, et al., Clustal W and clustal X version 2.0, Bioinformatics 23 (2007) 2947–2948.
[21] A. Blanco-Míguez, F. Beghini, F. Cumbo, L.J. McIver, K.N. Thompson, M. Zolfo, et al., Extending and improving metagenomic taxonomic profiling with uncharacterized species using MetaPhlAn 4, Nat. Biotechnol. 41 (2023) 1633–1644.
[22] H.-J. Ruscheweyh, A. Milanese, L. Paoli, N. Karcher, Q. Clayssen, M.I. Keller, et al., Cultivation-independent genomes greatly expand taxonomic-profiling capabilities of mOTUs across various environments, Microbiome 10 (2022) 212.
[23] D.E. Wood, J. Lu, B. Langmead, Improved metagenomic analysis with Kraken 2, Genome Biol. 20 (2019) 257.
[24] B. Liu, T. Gibbons, M. Ghodsi, T. Treangen, M. Pop, Accurate and fast estimation of taxonomic profiles from metagenomic shotgun sequences, BMC Genom. 12 (2011) 1–10.
[25] D. Li, C.-M. Liu, R. Luo, K. Sadakane, T.-W. Lam, MEGAHIT: an ultra-fast single-node solution for large and complex metagenomics assembly via succinct de Bruijn graph, Bioinformatics 31 (2015) 1674–1676.
[26] S. Nurk, D. Meleshko, A. Korobeynikov, P.A. Pevzner, metaSPAdes: a new versatile metagenomic assembler, Genome Res. 27 (2017) 824–834.
[27] H. Li, R. Durbin, Fast and accurate short read alignment with Burrows–Wheeler transform, Bioinformatics 25 (2009) 1754–1760.