Development of a Novel Metagenomic Classification Method

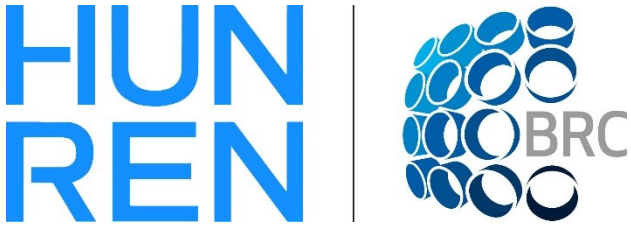
PhD thesis

Bertalan Vilmos Takács

Supervisor:

Dr. Lajos Haracska

DOCTORAL SCHOOL OF BIOLOGY



FACULTY OF SCIENCE AND INFORMATICS

UNIVERSITY OF SZEGED

INSTITUE OF GENETICS

HUN-REN BIOLOGICAL RESEARCH CENTRE

2025

I. Table of Contents

I. Table of Contents 2
II. List of figures 5
III. List of abbreviation 6
IV. Introduction 7
1. An overview of the microbiome: ecology, diversity and host interactions 7
1.1. Composition and functions of the gut microbiome 8
1.1.1. Metabolic role of the gut microbiota 9
1.1.2. Host health modulating effect 9
1.1.3. The microbiota-gut-brain axis 10
1.2. Oral 10
1.3. Vaginal microbiome 11
1.4. Skin microbiome 12
2. Identifying microbiome members 12
2.1. Culture-based classification of bacteria 13
2.2. Culture-independent classification of bacteria 13
2.2.1. DNA-based classification methods 14
2.2.2. Next-Generation Sequencing (NGS) 14
2.2.3. 16S sequencing 15
2.2.4. Shotgun sequencing 16
3. Metagenomic classification of bacteria 17
3.1. Types of classification algorithms 18
3.2. DNA sequence alignment 19
3.3. Main challenges in classification 21
4. Standards in metagenomics 22
4.1. *In vivo* standard communities 22
4.2. *In vitro* standards 23
4.3. *In silico* standards 23
5. Clinical relevance of metagenomics 24

V. Aims 25
VI. Materials and Methods 26
1. Creation of NABAS+ (Novel Alignment-based Biome Analysis Software +) 26
2. Creating an in-house *in silico* dataset 26
3. Collecting data from the CAMI II challenge 28
4. CAMI II sample generation 29
5. Collecting and analysing data from deeply sequenced microbial community standards 29
6. Collecting and analysing real-world metagenomic sequencing data 31
7. Collecting and analysing clinical data 32
8. Selecting and setting up classifiers for the initial gut metagenome analysis 32
9. Selecting and setting up reference classifiers for NABAS+ benchmarking 33
10. Running the classifiers 33
11. Statistical analysis 34
12. Software used and code availability 35
VII. Results 36
1. Comparing the performance of nine commonly used metagenomic classifiers 36
2. Introducing a novel metagenomic classifier, NABAS+ 38
3. Testing NABAS+ on in-house generated datasets 41
4. Benchmarking NABAS+ and the low group on the CAMI II gastrooral *in silico* data 46
5. Examining and re-creating an outlier CAMI II sample 48
6. Testing classifiers on the regenerated CAMI II sample19 48
7. Testing classifier performance on Zymo standards 49
8. Demonstrating NABAS+'s utility on a real-world clinical dataset 50
VIII. Discussion 52
IX. Conclusions 56
X. Acknowledgements 57
XI. Bibliography 58
XII. List of Publications 70
Other publications 70

Conference talks 71
Posters 72
XIII. Magyar nyelvű összefoglaló 73
XIV. Summary in English 77

II. List of figures

Figure 1: Most notable taxa of four microbial communities of the human body. (A) and (B) important functions of the gut microbiome. Adapted from Hou et al., 2022 8
Figure 2.: Examples of methods of bacterial identification 12
Figure 3: Comparison of the 16S and shotgun sequencing workflows. 15
Figure 4: Broad view on basic local alignment and the seed-and-extend alignment 20
Figure 5: Number of found species and Jaccard-distance results of 9 metagenomic classifiers 36
Figure 6: Simplified workflow of NABAS+ 38
Figure 7: Number of species identified by each classifier across the in-house datasets 42
Figure 8: Number of identified species in the samples per read number 43
Figure 9: F1 scores of tested classifiers on the in-house dataset 44
Figure 10: Performance comparison on the CAMI II dataset 47
Figure 11.: Classification performance on the (A) Zymo CSI and (B) Zymo CSII datasets 49

III. List of abbreviation

ANI	Average Nucleotide Identity
ASV	Amplicon Sequence Variant
BWA	Burrows-Wheeler Alignment
CAMI	Critical Assessment of Metagenome Interpretation
CSI	Community Standard I
CSII	Community Standard II
FDR	False Discovery Ratio
IMViC	Indole Test, Methyl Red Test, Voges-Proskauer Test, Citrate Utilization Test
LCA	Latest Common Ancestor
MAG	Metagenome-assembled genome
MEM	Maximum Exact Matches
MetaPhlAn	Metagenomic Phylogenetic Analysis
MLST	Multilocus Sequence Typing
NABAS+	Novel Alignment-based Biome Analysis Software +
NCBI	National Center for Biotechnology Information
NGS	Next-generation Sequencing
OTU	Operational Taxonomic Unit
PCR	Polymerase Chain Reaction
qPCR	Quantitative Polymerase Chain Reaction
rRNA	Ribosomal Ribonucleic Acid
SCFA	Short Chain Fatty Acids
SNP	Single Nucleotid Polymorphism

IV. Introduction

1. An overview of the microbiome: ecology, diversity and host interactions

We live in a world of microorganisms. There are very few truly sterile, habitable environments in nature (Cockell, 2021, Michán-Doña et al., 2024). Aside from a few extreme inhospitable environments (Dragone et al., 2021, Payler et al., 2019), every corner of our surroundings is inhabited by microorganisms: bacteria, archaea, viruses, fungi and other eukaryotes. These organisms are so ubiquitous that humans carry them even to previously uninhabitable spaces (Salido et al., 2025). The community of microscopic organisms living in a specific environment is referred to as microbiome (Berg et al., 2020). Each microbiome is shaped by their environment, and in turn shapes that environment. This is true not only in abiotic environmental settings, but also within and on the surfaces of multicellular organisms, all of which host diverse microbial communities (Bordenstein & Theis, 2015). These microorganisms influence the health and development of the host, on an individual but also on an evolutionary scale (Zilber-Rosenberg & Rosenberg, 2008).

Beyond host interactions, microbes also engage with one another through numerous direct and indirect interactions, such as competing for the same resources, developing biofilm, anti-microbial peptides, and quorum sensing (Coyte et al., 2015). Consequently, a microbiome can be viewed as an ecological community rather than a collection of independent species. As such, diversity metrics like alpha-diversity (measuring the number and distribution of species in a sample), beta-diversity (measuring the different abundance of species between communities), species richness (measuring the number of species present in a community), and evenness (measuring how equally the different species are distributed) are commonly applied (Galloway-Peña & Hanson, 2020).

In this dissertation, the term microbiome refers specifically to the human microbiome—the community of microorganisms residing within and on the surface of the human body. Although the human body also serves as host for numerous viruses, fungi, and other eukaryotic microorganisms, the focus here is limited to the bacterial component of these

communities. These bacteria are high in number and rich in diversity (Huttenhower et al., 2012a) and influence the host body in several ways, both beneficial and detrimental (Clemente et al., 2012). The most extensively studied human microbiome is that of the gut, but other important bacterial communities also inhabit the vagina, the oral cavity, and the skin (Figure 1).

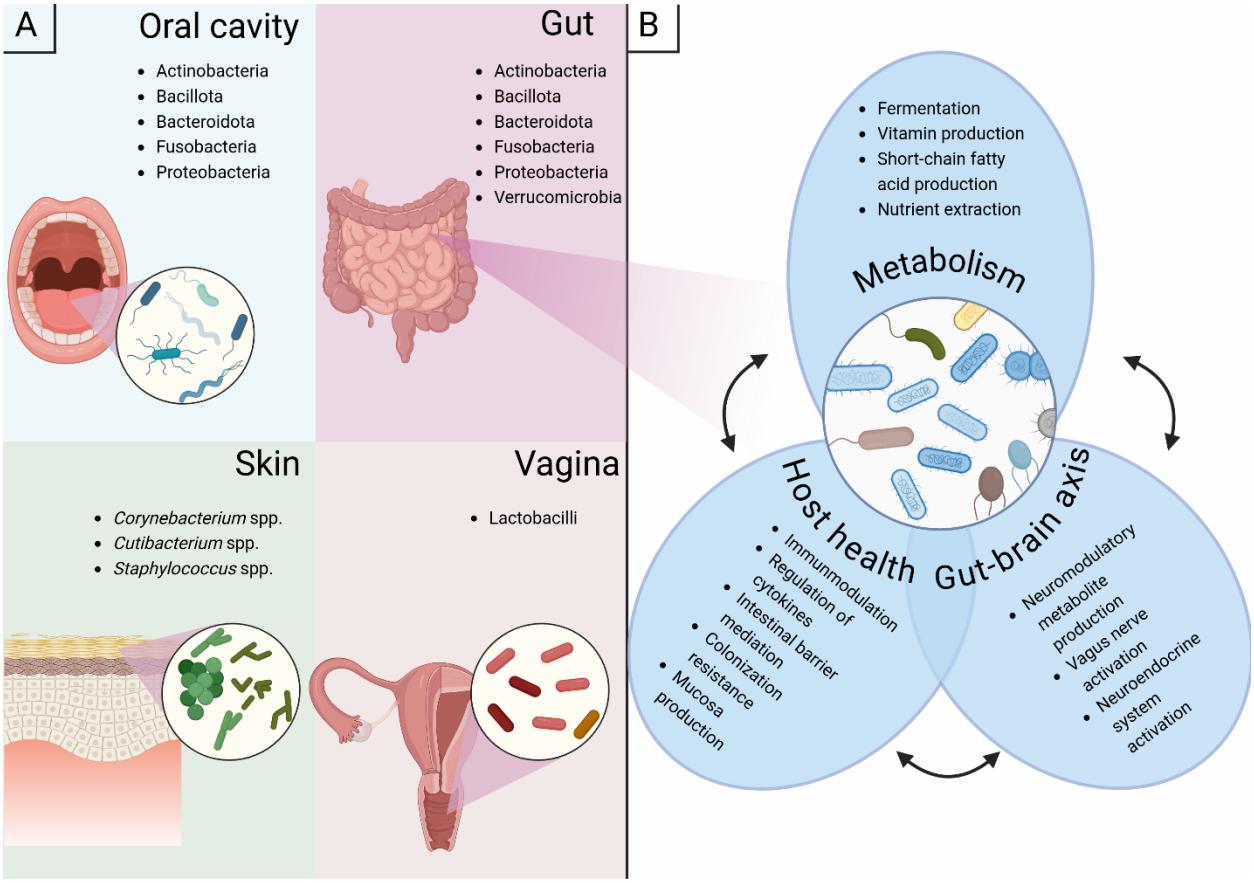


Figure 1: Most notable taxa of four microbial communities of the human body. (A) and (B) important functions of the gut microbiome. Adapted from Hou et al., 2022

1.1. Composition and functions of the gut microbiome

The most bacteria-rich environment inside the human body is the gut, particularly the large intestine. In this region, the number of bacterial cells is estimated to be roughly comparable to, or slightly greater than, the number of human cells of the organ (Sender et al., 2016). This richness is combined with a high diversity; there are on average 200-1000 species in the healthy gut (Turnbaugh et al., 2007). The gut microbiome also shows high levels of specificity to the host: microbial “fingerprints” have been shown to work as

an accurate means of identifying individuals (Franzosa et al., 2015). Bacteria in the gut mainly belong to 5 phyla: Bacillota (formerly Firmicutes), Bacteroidota (formerly Bacteroidetes), Proteobacteria, Actinobacteria and Verrucomicrobia (Hou et al., 2022). Additionally, the gut is populated by archaea, most notably the methanogens *Methanobrevibacter smithii* and *Methanosphaera stadtmanae* (Gaci et al., 2014).

In the following section I give a brief introduction of 3 important properties of the gut microbiota, nonetheless this list is not exhaustive, and the listed functions are not mutually exclusive.

1.1.1. Metabolic role of the gut microbiota

Gut bacteria produce enzymes for the digestion of proteins, carbohydrates and fatty acids and are able to digest complex carbohydrates that would otherwise be hard to digest or indigestible for the host (Flint et al., 2012). Through the digestion of these carbohydrates, microbial members can produce short-chain fatty acids (SCFA), such as acetate and butyrate, which play important roles in immunomodulation and the regulation of the epithelial barrier (Mann et al., 2024)

Moreover, some genera of gut bacteria (e.g. *Bacteroides*, *Bifidobacterium*, and *Enterococcus)* can contribute to the production of beneficial nutrients and vitamins (Morowitz et al., 2011). The most notable examples are the vitamin group B (such as biotin (B7), folate (B9), and cobalamin (B12)) and vitamin K (Tarracchini et al., 2024). It has been estimated that up to half of the daily Vitamin K requirement is provided by gut bacteria (Morowitz et al., 2011).

The metabolic properties of the gut microbiome do not always align with our intentions since certain drug compounds can be partly or completely digested by members of the gut microbiome, thus influencing drug pharmacodynamics and medical treatment (Tsunoda et al., 2021). Furthermore, certain bacterial metabolic byproducts can lead to health conditions, such as bloating or diarrhea (Sachdev & Pimentel, 2013).

1.1.2. Host health modulating effect

The gut microbiota acts as a shield in protecting the host from exogenous and potentially pathogenic microorganisms through a process termed as “colonization resistance” (Hou

et al., 2022). The gut microbiome also contributes to the modulation of the barrier function of intestinal cells (Takiishi et al., 2017) and to the regulation of the intestinal mucus barrier (Paone & Cani, 2020).

On the other hand, bacteria in the gastrointestinal tract can lead directly or indirectly to diseases. Notable examples include *Helicobacter pylori* causing chronic gastritis (Robin Warren & Marshall, 1983) and different strains of *Escherichia coli* facilitating inflammatory bowel disease (Mirsepasi-Lauridsen et al., 2019). The gut microbiome can also serve as a host for opportunistic pathogens (Dey & Ray Chaudhuri, 2023) and a reservoir for antibiotic resistance genes (Anthony et al., 2020).

1.1.3. The microbiota-gut-brain axis

A particularly curious effect of the gut microbiome is on the nervous system. The gut has a large number of nerve cells and is sometimes referred to as the "second brain" (Gershon, 1999). There is evidence that the gut microbiome is in a bidirectional communication with the central nervous system through these nerve cells, known as the "microbiota-gut-brain axis" (Loh et al., 2024). The microbiome is capable of producing or influencing the production of several neurotransmitters, such as serotonin (Yano et al., 2015), dopamine (Wang et al., 2021), and gamma-aminobutyric acid (Strandwitz et al., 2018). The microbiome is also able to regulate microglial maturation and cell death through SCFAs (Huang et al., 2023) and thus is theorized to have a role in different neurodegenerative diseases, such as Alzheimer's disease (Dodiya et al., 2021) and Parkinson's disease (Sampson et al., 2016).

1.2. Oral microbiome

The oral microbiome is the second largest microbial community in the human body, and its composition shows a high degree of overlap with the gut microbiome on a higher taxonomic level (phylum, class, family). The most notable oral bacterial phyla are Bacillota (formerly Firmicutes), Proteobacteria, Bacteroidota (formerly Bacteroidetes), Actinobacteria and Fusobacteria (Hou et al., 2022). The oral cavity can be further divided into different distinct microbial habitats, such as the tongue, tooth surfaces, and buccal mucosa, each with their own distinct microbial composition (Baker et al., 2023). While the

core members of the healthy oral microbiome are stable, there are rare taxa and strain variations enough to distinguish one person from another (Arumugam et al., 2025).

The disruption of a healthy oral flora can lead to diseases, such as dental caries and periodontal disease (Baker et al., 2023). Dysbiosis of the oral microbiome also shows association with systemic diseases, including cancer and rheumatoid arthritis (Kumar, 2013).

1.3. Vaginal microbiome

Probably the second most researched bacterial community in the human body, the vaginal microbiome also plays a significant role in disease development and prevention.

The healthy vaginal microbiome is dominated by a few species, mainly from the *Lactobacillus* and related genera (Amabebe & Anumba, 2018). These bacteria produce lactic acid (hence the name), lowering the pH of the vaginal microenvironment, and supposedly inhibiting the growth of other, less beneficial bacteria (Amabebe & Anumba, 2018). Thus, similarly to the gut, the vaginal microbiota also has an important function in colonization resistance (Mei & Li, 2022). Lactic acid may also modulate the immune response of the host (Chee et al., 2020).

There are cases when the vaginal microbiota does not consist primarily of lactobacilli but is instead dominated by facultative or obligate anaerobes such as *Gardnerella vaginalis*, *Prevotella* spp., *Mobiluncus* spp., *Ureaplasma urealyticum*, and *Mycoplasma hominis*. The presence of these bacteria is associated with higher pH levels and a condition called bacterial vaginosis (Abou Chacra et al., 2022). The presence of these bacteria has also been associated with the acquisition of sexually transmitted infections and spontaneous preterm birth (Ding et al., 2021).

1.4. Skin microbiome

As the skin is constantly in contact with the outside environment, its microbial composition is heavily influenced by environmental factors, such as heat, moisture, and outside pathogens, (Baker et al., 2023). An interesting property of the skin is that depending on its physiological characteristics (whether it's oily, moist, or dry) different sites of the human body harbor different bacteria (Costello et al., 2009). The healthy skin microbiome is composed mainly of the *Cutibacterium* spp. (formerly *Propionibacter* spp.), *Staphylococcus* spp. and *Corynebacterium* spp., bacterial species along with fungi from the *Malassezia* genus (Byrd et al., 2018). Members of the skin microbiome have been connected to skin issues, such as acne (*Cutibacterium acnes,* in Dréno et al., 2018) and atopic dermatitis (*Staphylococcus aureus,* in Kim et al., 2019).

2. Identifying microbiome members

As the microbiome can affect human health in several different ways, it is important for the members of a given community to be identified correctly. The misidentification of bacteria and especially of pathogens in a clinical setting can lead to misdiagnosis and unnecessary treatment of patients. Currently, there are several methods for the identification of bacteria from microbial samples (Figure 2), each with their own strengths and weaknesses.

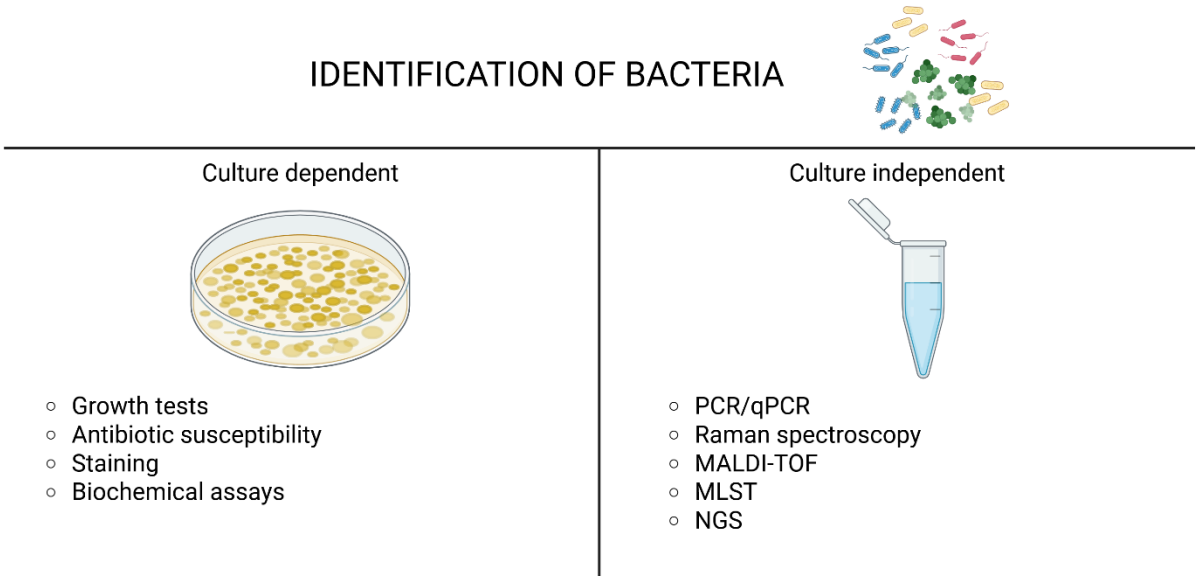


Figure 2.: Examples of methods of bacterial identification

2.1. Culture-based classification of bacteria

Classical microbiology methods are the original way to identify species from a microbiome sample. In these, most commonly, the sample is spread on an agar plate, and bacteria are grown under controlled conditions. The identity of the bacterial species is then decided based on phenotypic tests (e.g. color and morphology of the colony, growth on specific media) alongside staining and antibiotic susceptibility tests (Giuliano et al., 2019). The probably most well-known staining technique is Gram-staining, which targets the prokaryotic cell wall and divides bacteria into two non-disjunct groups: Gram-positives and Gram-negatives (Bartholomew & Mittwer, 1952). Metabolic capabilities of the studied bacteria, measured by biochemical assays, can also serve as the basis of identification (Altheide, 2019). For example, the IMViC, which was developed for the differentiation of *Enterobacteria*, is based on the detection of 4 distinct metabolic processes namely indole production, acid production, acetylmethylcarbinol (acetoin) production, and citrate utilization (Powers & Latt, 1977).

These methods are still considered the “gold standard” for bacterial identification and are required when describing a novel bacterial species: an isolated and pure culture of the “type strain” of the new species is necessary for acceptance by the International Code of Nomenclature of Prokaryotes (Parker et al., 2019). Nonetheless, despite their accuracy, culture-based methods have several shortcomings as they are expensive, slow (Goelzer & Fromion, 2011), low throughput, and the detection of a specific species often requires a dedicated test. Additionally, these methods may not provide a representative picture of the ratio of bacteria in a sample, as different bacteria can grow better or worse in laboratory conditions than in their natural environment (Steensels et al., 2019). Although culturing methods are constantly improving, there is still a substantial number of gut bacteria that can’t be cultured (Liu et al., 2021).

2.2. Culture-independent classification of bacteria

With the expansion of understanding of the molecular processes of prokaryotic cells, novel laboratory methods are continuously developed for the culture-independent classification of bacteria. These methods target structural, metabolic and other features of these microorganisms and often are done in parallel with culturing. These methods

include Raman spectroscopy (Krynicka et al., 2025), matrix assisted laser desorption ionization-time of flight mass spectrometry (MALDI-TOF MS) (Singhal et al., 2015), among others. However, the current most popular prokaryotic classification methods focus on DNA.

2.2.1. DNA-based classification methods

Nowadays, there is a vast selection of classification methods for differentiating prokaryotes based on their DNA content. They are common in that they focus on parts of the bacterial genomes which are characteristic of a given bacterium and can be used to differentiate it from the rest. These DNA-based methods include polymerase chain reaction (PCR) and quantitative PCR (qPCR) based-methods (Kralik & Ricchi, 2017), multi-locus sequence typing (MLST) (Maiden, 2006), plasmid profile analysis, and next-generation sequencing methods (Adzitey et al., 2012). The latter is also the focus of this dissertation.

2.2.2. Next-Generation Sequencing (NGS)

Although DNA sequencing has been possible since the 1970s, the high-throughput sequencing methods brought a breakthrough in the last 30 years (reviewed in Kumar et al., 2019). These methods made it possible to find the sequence of large amounts of DNA at the same time.

In brief, the most widely adopted NGS technology, originally developed by Solexa and subsequently commercialized by Illumina, is based on sequencing by synthesis (reviewed in Hu et al., 2015). Sequencing by synthesis works as follows: template DNA sequences are anchored to the surface of a sequencing chip and are extended through polymerase reactions. In each cycle, fluorescently marked termination nucleotides are added to the mix. The incorporation of these nucleotides results in a fluorescent emission of a specific wavelength that can be detected by a camera. This camera takes a high-resolution image of every cycle, and at the end of the sequencing runs the generated images are “translated into” nucleotides through a process called base-calling (Cacho et al., 2016). This method made it possible to determine the sequence of thousands of DNA templates in the same run.

The high throughput capability of this method proved to be especially useful in the case where the DNA content of a sample doesn't come from a single organism, but from a community of tens or hundreds of species The term metagenomics itself, which refers to the study of the genetic content of an environmental sample, has been around the same time as these high-throughput methods (Handelsman et al., 1998). High-throughput sequencing in the last 25 years has brought several breakthroughs to the field.

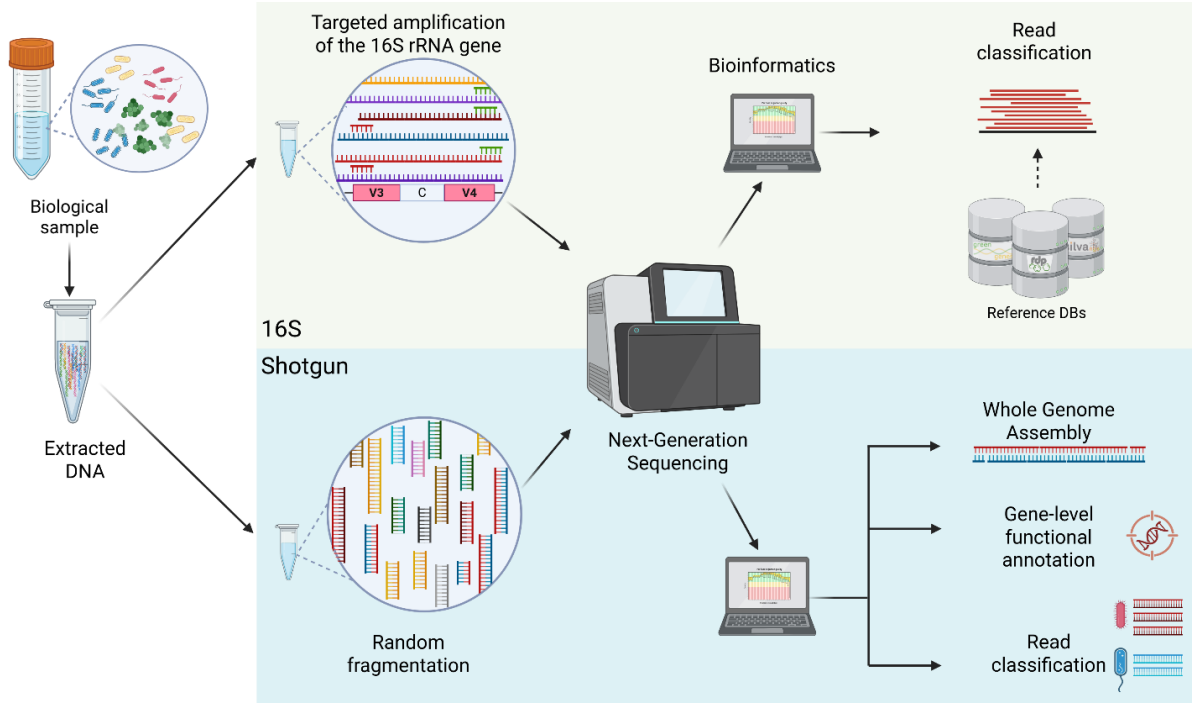


Figure 3: Comparison of the 16S and shotgun sequencing workflows. V3,V4: variable regions of the 16S rRNA gene, C: conserved region

2.2.3. 16S sequencing

The more traditional sequence-based identification method of bacterial species is the so-called 16S sequencing (Janda & Abbott, 2007). This method is based on the 16S rRNA gene, which can be found in every bacterial genome. This gene is particularly useful for microbial classification, due to its conserved regions, which are very similar among all bacteria, as well as variable regions, that show species- or even sub-species-level specificity. There are PCR-based methods targeting specific SNPs of this gene, but NGS-based 16S classification is a far more popular method (Figure 3). Finding the sequence of this gene is often enough to classify bacteria in metagenomic samples, and it offers a

quick and cheap solution to the problem. However, 16S sequencing-based identification has several weaknesses:

- It doesn't provide any additional genomic information other than the sequence of variable regions of this one specific gene, and such can't be used for metagenomics
- Bacteria can have multiple copies of the same 16S rRNA gene and these copies can differ in their sequence (Lin et al., 2022). Moreover, classification can give a different result based on the variable region targeted by the sequencing, leading to inaccuracies (Zhang et al., 2024).

Because of these limitations, 16S sequencing is no longer regarded as a true metagenomic approach, since it targets only a single marker gene rather than the full genomic content of a community. For this reason, it has largely been replaced by shotgun metagenomics, which captures the complete DNA pool of all organisms in a sample. However, there are still specific cases, for example, the sample is highly contaminated by the eukaryotic host's DNA, or genus-level classification is sufficient, when 16S is commonly used (Durazzi et al., 2021). Additionally, sequencing the whole 16S gene using long read sequencing technologies, such as Nanopore (Aja-Macaya et al., 2025) and PacBio (Buetas et al., 2024) can partially overcome these limitations and improve taxonomic classification.

16S classification algorithms apply roughly the same basic principles: they order the sequences into operational taxonomic units (OTUs) based on sequence similarity, then compare these to a reference database (such as SILVA or Greengenes). A more novel method is treating each sequence as amplicon sequence variant (ASV) identifying them separately, then merging abundance data on species level. This can give a more fine-grained view on the composition of the bacterial community (Marizzoni et al., 2020).

2.2.4. Shotgun sequencing

Modern microbial analysis increasingly relies on metagenomic methods, which aim to analyze not just one gene per bacterial taxon, but the whole genetic content of each species in the sample. Shotgun sequencing is named after the DNA processing technique applied: extracted DNA is randomly fragmented into short DNA fragments that collectively

represent the entire genomes of all organisms in the sample, similarly to the scatter pattern of a shotgun blast (Anderson, 1981).

Shotgun sequencing can be applied in more diverse ways compared to 16S (Figure 3):

- Creating metagenome-assembled genomes (MAGs):
  MAGs are an important tool to study microbial diversity and characterizing new, often unculturable, organisms. The creation of MAGs is done by assembling reads into longer sequences (contigs and then scaffolds) and then grouping the results: creating groups of scaffolds based on short sequence similarities. Each bin corresponds to a putative individual bacterial genome. Assembly and binning are always followed by quality assessment, where the completeness of the genomes and the presence of contaminants is tested (Setubal, 2021).
- Gene identification and functional profiling:
  This provides information on how bacteria function in a community and how they interact with the host. Additionally, it can be used to identify antibiotic-resistance genes, which is critical for clinical treatment (Boolchandani et al., 2019). Functional annotation typically involves two steps: gene prediction, where the potentially coding sequences are identified, and annotation, where predicted proteins are compared with protein families in databases and annotated functionally.
- Taxonomic classification:
  Shotgun sequencing can be used for more accurate classification and can provide sub-species level information for certain bacteria. The following section details the workings of such classification algorithms.

Overall, all methods analyzing metagenomic data are strongly influenced by the quality of the sample, sequencing depth, and the bioinformatic methods applied.

3. Metagenomic classification of bacteria

To make sense of metagenomic data and to find species corresponding to DNA sequences (the classification process), bioinformatics methods are needed. As the available metagenomic data is increasing exponentially, the number of classification strategies grows as well. These classification algorithms apply various methods to identify

bacterial taxa in the samples.

3.1. Types of classification algorithms

There are several approaches to identify bacteria in metagenomic samples. The most popular metagenomic classifiers apply some kind of alignment algorithm in their workflow.

While the algorithmic principles applied by metagenomic classifiers are diverse, they can be broadly divided into 3 categories:

1. *k-mer based classifiers*: These classifiers process the reference genomes by splitting it into short sequences (k-mers) and then building a taxonomic tree of the sequences, trying to find the taxon in which the k-mer appears (latest common ancestor, LCA). The classification of the samples is done similarly: splitting the reads into k-mers and attempting to place these k-mers on the LCA tree. The most notable examples are Kraken (Wood & Salzberg, 2014), Kraken2 (Wood et al., 2019). and Bracken (Lu et al., 2017), which is an extension of Kraken. K-mer based classification methods tend to be fast but less specific, compared to other methods (Garrido-Sanz et al., 2022).
2. *Markergene-based classifiers*: These classifiers process the reference genomes by taxon-specific sequences. During classification, the algorithm compares the reads to this taxon-specific database, assigns the recognized sequences to taxon and discards the rest. The most notable example of marker gene-based classification is MetaPhlAn (Beghini et al., 2021). This method tends to produce good sensitivity and specificity, with medium running speed (Ye et al., 2019).
3. *Genomic alignment-based classifiers*: These methods tend to be computation- and time-intensive. They rely on building an indexed reference database of complete bacterial genomes and applying a local alignment algorithm to match the reads to the reference sequences. GOTTCHA (Freitas et al., 2015) is a notable example of alignment-based classification.

There are, of course, classifiers that don't fit these categories. For example, Kaiju (Menzel

et al., 2016) and DIAMOND-Megan (Bağcı et al., 2021) which utilize a DNA-to-protein sequence-based search strategy. These tools are becoming less popular due to their high computational and time requirements and low specificity.

3.2. DNA sequence alignment

Sequence alignment is the most important step of comparing biological sequences. Simply put, alignment is the process when two strings (character sequences) are arranged in a way that shows the highest-scoring similarity between the two. In genomics alignment is typically done against a reference, which can be a gene, chromosome or whole genome of an organism. The completeness and accuracy of this reference highly influences the quality of the alignment.

As the size of the reference grows, alignment becomes progressively less trivial. The need for alignment algorithms that can simplify the process or at least reduce the time and computational resources needed is growing. Many metagenomic classifiers rely on some kind of alignment or alignment-like strategy to compare reads or other query sequences to reference genomes, and they apply various methods for saving time and computational resources (as detailed in the next section). The method of alignment, where only a subsequence of the query sequence is matched to a subsequence of the reference, is called local alignment. A classic example of a local alignment algorithm is the Smith-Waterman algorithm (Smith & Waterman, 1981), which finds the matching substrings between the reference and the query by filling out a scoring matrix: a matrix that gives scores for matches and mismatches, and penalties for gaps (insertions or deletions). The algorithm applies dynamic programming to maximize the number of matching nucleotides while minimizing mismatches and indels. Although the Smith-Waterman Algorithm is too computationally intensive for large-scale genomics, it serves as the foundation for numerous modern alignment algorithms (Daily, 2016, Delcher et al., 2002). In the following, as an example, I will introduce one such algorithm, the Burrows-Wheeler Aligner (BWA) (Li & Durbin, 2009) and its Maximum Exact Matches version (BWA-MEM) (Li, 2013), which is a robust, frequently used alignment algorithm in bioinformatics and underlies the software described in this dissertation.

Indexing the reference is the initial step of BWA. During this step the reference is

transformed into an indexed suffix array via Burrows-Wheeler transformation (Burrows & Wheeler, 1994) and FM-indexing (Ferragina & Manzini, 2000). This transformation enables fast and computationally efficient substring search. In the original version of BWA reads are represented by fixed-length seeds, which are mapped to the reference via exact matches in the FM-index (Figure 4). These seeds are then extended and scored with a modified version of the Smith-Waterman algorithm and chained together into candidate alignments, with the highest-scoring chains reported as the final local alignments.

BWA-MEM, the variant used in my dissertation, is optimized for longer reads. Instead of fixed-length seeds, it first scans reads for minimizers, which serve as representative of the query, and then uses the FM-index to identify maximal exact matches (MEMs). Among these, it selects supermaximal exact matches (SMEMs) — matches that cannot be extended in either direction without mismatch and are not contained within larger matches. These SMEMs act as seeds for chaining and local extension, as in the classic algorithm. This adaptive seeding strategy makes BWA-MEM faster, more efficient, and more suitable for aligning long or noisy reads than the original BWA.

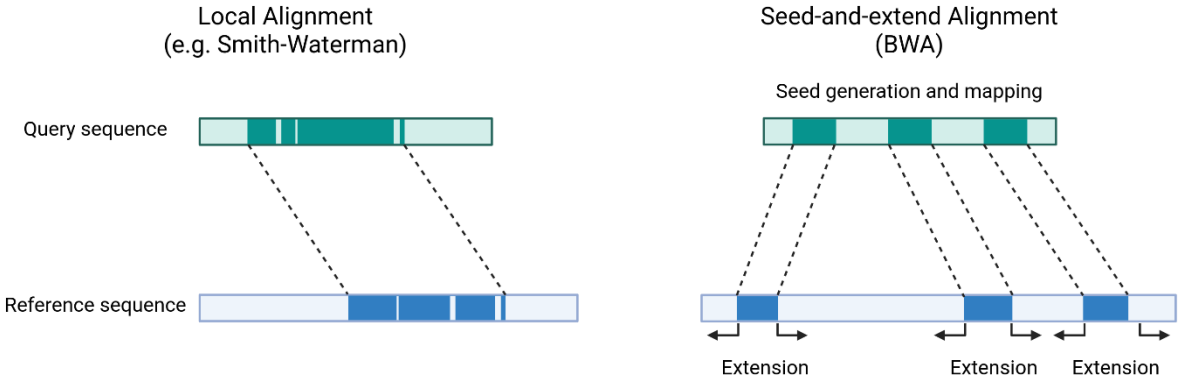


Figure 4: Broad view on basic local alignment and the seed-and-extend alignment
Adapted from Florescu & Ahmed, 2022

3.3. Main challenges in classification

There are several problems in metagenomic classification, that all classification algorithms must deal with:

- The quality and size of the reference database
  There is a large amount of bacterial sequencing data available, but it varies in quality and species specificity. E.g. for extensively studied species, such as *E. coli*, there are several, deeply sequenced strains available. Lesser-known or recently discovered species usually have one such reference. Simply comparing every sequencing read to every available bacterial reference sequence would be an exorbitant computational task.
- The quickly changing bacterial taxonomic landscape
  As the field of bacterial genomics is quickly evolving, new species are constantly discovered, old ones are reassigned, and existing clades are renamed. Metagenomic classifiers need to be constantly updated and need to have a reliable source of taxonomic information.
- Similarity between bacterial species and evolutionally conserved regions
  Defining and dividing between bacterial species is far from trivial (Doolittle, 2012). The current scientific is that bacteria of the same species have at least 95% average nucleotide identity (ANI) in their genomes, which corresponds to ~99% similarity in the 16S rRNA gene (Konstantinidis et al., 2006). Additionally, because of the methods of horizontal gene transfer, bacteria that aren't closely related can carry similar sequences in their genome. These conserved regions can make classification "noisy" and can increase the number of false positives (Paul et al., 2020).
- Making classification computationally efficient
  Classifiers apply different methods to filter and index the reference database and the metagenomic sample, to reduce computational costs.
- Making classification as accurate as possible
  Misclassification can lead to several problems, especially if the metagenomic analysis is performed for the purpose of clinical pathogen identification.

For example, as mentioned, Kraken2 divides up reference genomes to k-mers (short, 20–30 bp long sequences) and selects common sequences between bacteria on different taxonomic levels, thus making searching the reference database very efficient (k-mer search can be 1–3 magnitudes faster than full-sequence alignment (Bokulich, 2025)).

MetaPhlAn reduces the reference genome size by selecting sequences that are strictly specific to one bacterium. This allows for fast classification, although the number of classified reads will depend on the quality and the diversity of the sample.

In general, Kraken2 tends to overestimate the number of species in a sample, while MetaPhlAn generally underestimates. The accuracy of these two classifiers can be influenced by factors such as the read depth and the size of the reference genome (Sun et al., 2021). In general, there is no perfect method, only one that is suitable for the specific problem. To compare the performance of the growing number of classifiers, benchmarking against “gold standard” samples is needed.

4. Standards in metagenomics

For the proper benchmarking of the performance of metagenomic classifiers, it is necessary to have a ground truth of bacterial communities with known composition. This allows for the fair comparison of classifiers and can reveal their weaknesses, strengths, and specific uses of different tools. Several such standards are available for metagenomics; they differ in their composition, complexity, and use, but can be grouped into three general categories: *in vivo*, *in vitro* and *in silico* standards.

4.1. *In vivo* standard communities

There are several, well-described microbial communities inoculated in living organisms, such as minimal microbiomes sustained in gnotobiotic mice (e.g. Altered Schaedler Flora (Brand et al., 2015)). These “synthetic communities” are often results of a reductionist approach toward understanding the microbiota: they contain only the “key” species of a microbial community, and they don’t attempt to model real-world diversity (Raghu et al., 2024). These *in vivo* standards can provide a close representation of a real-world microbiome. These standard communities may, however, be contaminated or colonized

by unknown bacteria, making their maintenance time-consuming and costly (Basic & Bleich, 2019).

4.2. *In vitro* standards

In vitro standards are created by mixing either the cells or isolated DNA of well-known and frequently studied bacterial species (Morgan et al., 2010). These standards are frequently used in laboratory practice as controls and validation of metagenomics. Because these samples aren't taken from the real-world, they show less resemblance to actual, real microbiomes. However, this can be also the strength of these standards: make it possible mixing of bacteria together that wouldn't be able to co-exist in nature. As such, *in vitro* standards allow testing on a broader range than *in vivo* ones.

4.3. *In silico* standards

Generating *in silico* metagenomic standards has several advantages over real-world sequenced DNA samples. The number and identity of genomic sequences used for the generation, the number, length, insert size and quality of reads can be strictly controlled: we can be sure where each read in the sample comes from.

There are several methods to make sure that the generated reads resemble real-world samples as close as possible. Modeling the distribution of quality scores of the reads can take several directions, from simply giving each base the same quality score to the same value to different average error rates and sequencer-specific error profiles. Random sequencing errors, insertions and deletions can also be modeled (Fritz et al., 2019).

As an example, Critical Assessment of Metagenome Interpretation (CAMI) is one of the most well-known examples of these standardization efforts, with two rounds of challenges so far (Sczyrba et al., 2017, Meyer et al., 2022). The purpose of these challenges is to provide gold standard data for classifiers as well as to compare the existing algorithms on different tasks, such as classifying very diverse samples, identifying a novel bacterial

strain in a medical context, or assembling and taxonomically classifying the genome of previously unknown bacterial species.

5. Clinical relevance of metagenomics

Correctly identifying bacteria in human samples is a critical task in clinical analysis. Currently, this is most often performed by culturing methods introduced at the beginning of the chapter. Because some pathogens are difficult or impossible to culture in laboratory settings, and because broad-spectrum antibiotics are often applied before identification, these traditional methods can lack in sensitivity and can result in needless complications for the patient (reviewed in Gu et al., 2019). Metagenomics could mean a solution for these problems, providing a non-targeted, high-throughput, sensitive method for pathogen identification. For clinical use, any new detection method needs to be strictly validated (Kan et al., 2024). In the case of metagenomics, this can be achieved by using the previously mentioned *in vitro* and *in silico* standards, or alternatively, using real-world clinical data, tested by both culturing and metagenomics-based methods (Angel et al., 2025).

As previously mentioned, metagenomic classification faces several challenges: In my doctoral work, I devised and benchmarked a novel metagenomic classification algorithm, named Novel Alignment-based Biome Analysis Software + (NABAS+) to address these challenges.

V. Aims

As I detailed in the introduction, metagenomics has revolutionized our ability to characterize complex microbial communities, yet the accurate classification of bacterial species within these datasets remains a challenge. To fully harness the potential of metagenomics for routine diagnostic use, there is a clear need for classification tools that combine high taxonomic specificity with robustness and computational efficiency. The validation of such tools requires benchmarking against datasets of precisely known composition that model realistic microbial communities. In this doctoral work, my aim was to develop, optimize, and validate a novel metagenomic classifier capable of accurate species-level identification, and to demonstrate its applicability to clinical pathogen detection.

To achieve these goals, I set out to:

- Develop a novel metagenomic classifier, NABAS+ that minimizes false-positive identification, leveraging high-quality reference genomes and the BWA-MEM algorithm
- Establish comprehensive benchmarking datasets by generating i*n silico* metagenomic samples as well as collecting *in silico* and *in vitro* standards for objective performance evaluation.
- Benchmark NABAS+ along with other classifiers and evaluate its performance using various metrics, such as F1 score, precision and recall.
- Demonstrate the applicability of the NABAS+ in clinical metagenomics by assessing its performance in correctly detecting pathogenic species in a real-world sample

VI. Materials and Methods

1. Creation of NABAS+ (Novel Alignment-based Biome Analysis Software +)

To address the problems highlighted in the Introduction, we aimed to create a novel metagenomic classifier, which we named NABAS+ (Novel Alignment-based Biome Analysis Software +). We used an underutilized alignment algorithm in metagenomic classification, BWA-MEM. The algorithmic workings of NABAS+ are detailed in the Results section.

We chose the Java programming language to implement our algorithm, for its speed, large available codebase and multithreading capabilities. This allows us to integrate our classifier into larger software environments and analysis pipelines. Furthermore, to facilitate the use of NABAS+ by the broader scientific community, we created a standalone version that can be run from the command line, along with test datasets and a reference database.

2. Creating an in-house *in silico* dataset

To evaluate the performance of our classifier, first we built a dataset of *in silico* metagenomic samples, generated in-house. With these samples our aim was to study the accuracy of NABAS+ on samples modeling real-world environments, with different read depth, as well as to compare the performance of our tool to the current industry-leader classifiers.

Our first set of test samples were generated using an in-house *in silico* NGS read generating algorithm, from 6 mock microbial communities, five of which were retrieved from the work of Ounit and Lonardi (Ounit & Lonardi, 2016). The composition of these communities was the following:

“Buc12”: This community contains 12 microbial species found in the buccal microbiota, as reported in (Franzosa et al., 2015) and (Huttenhower et al., 2012b) including

Haemophilus influenzae, *H. parainfluenzae*, *Neisseria subflava*, and *Veillonella dispar* as well as eight species from the *Streptococcus* genus.

“CParMed48”: For this community, 48 species were selected from Proteobacteria, Acidobacteria, Bacteroidota, Actinobacteria, and Planctomycetes phyla, based on (Reese et al., 2016) reporting the most common bacteria in city parks and medians in Manhattan.

“Gut20”: This community consisted of 20 species commonly found in the human gut, described by (Kuleshov et al., 2016), from the *Streptococcus*, *Listeria*, and *Lactobacillus* genera among others.

“Hous31”: This community contains 31 species typically found in Western homes, as described in (Ruiz-Calderon et al., 2016). These species belong to the Streptococcaceae, Lactobacillaceae, Pseudomonadaceae, Intrasporangiaceae, and Rhodobacteraceae families.

“Hous21”: This community is composed of 21 species from the dominant organisms found in the bathroom and kitchen, reported in (Adams et al., 2015), namely, *Propionibacterium acnes* and the *Corynebacterium*, *Streptococcus*, and *Acinetobacter* genera.

Additionally, we created a “Custom 100” community by randomly selecting 100 species from a 500 species list containing the most common bacteria of the human gut microbiota.

Relative abundances of the species in the communities were uniformly distributed. For each bacterium, we collected the latest available reference genome from RefSeq. Our algorithm utilized these genomic sequences in the following manner: picking a random start point in the genome and copy the sequence for a set number of bases to create a mock read.

We created an in-house *in silico* sample generating algorithm, to model real-life Illumina runs. Our algorithm works with a collection of fasta-formatted reference genomes, picks random starting points with a set length of insert size and adds random insertions, deletions and “sequencing errors” (bases different from the reference) with a set frequency.

Reads were generated to model Illumina 2*151 bp reads, with a minimum insert length of 50 and expected length of 250 bp. The read lengths followed Gaussian distribution with a standard deviation of 50 bp. At each position of every generated read, there was a 0.4% possibility of mismatch, 0.25% of 1-base-long deletion, and 0.15% of 1-base-long insertion. Insertions and deletions were allowed to happen concurrently, making longer indels possible. The Phred quality scores were set as consistent "A", coding the Q score 32, which corresponds to the error rate of $6.3*10^{-4}$ (Ewing et al., 1998). In real-world NGS-analysis a read with an average Q score above 30 is considered good quality (Ewing & Green, 1998).

Our reference data contained 212 species overall. From each community, we generated 6 samples with different read numbers: 5×10^5, 10^6, 2×10^6, 5×10^6, and 10^7 reads.

3. Collecting data from the CAMI II challenge

The gastrooral subset of second CAMI Toy Human Microbiome Project dataset (DOI: 10.4126/FRL01-006425518) was downloaded from the author's website (https://cami-challenge.org/datasets/), using the provided camiClient.jar, along with the provided NCBI RefSeq version and NCBI taxonomy. Reads were de-interleaved and given Casava 1.8-style headers before analysis, using a custom script, FixFastqHeaders.jar.

Because of the relative outdatedness of the reference database versions of the CAMI2 data (datasets were generated in January 2019), older, corresponding database versions were utilized for each classifier. We aimed to use database versions published in the same timeframe as CAMI2 datasets.

- MetaPhlAn3: 'mpa_v31_CHOCOPhlAn_201901'
- Kraken2: 'minikraken2_v2_8GB_201 904_UPDATE';
- GOTTCHA: 'GOTTCHA_BACTERIA_c4937_k24_u30_xHUMAN3x.species'

In the case of NABAS+, we built the reference database using genomes labelled as 'representative' or 'reference' from RefSeq (as of 8 January 2019, shared by the CAMI II authors), using a custom script.

4. CAMI II sample generation

The *in silico* sequencing sample was regenerated using the abundance file and settings provided in the CAMI II challenge with CAMISIM (Fritz et al., 2019) version 1.3. To get a more modern representation of each species in the sample, we collected the most recent ‘reference’ or ‘representative’ genome corresponding to the species from NCBI Refseq. After creation, the created sample was treated the same way as the rest of the CAMI2 samples.

5. Collecting and analysing data from deeply sequenced microbial community standards

To demonstrate the accuracy of our tool on *in silico* real-world shotgun sequencing as well, we used deep sequenced ZymoBIOMICS Microbial Community Standards (Nicholls et al., 2019).

These community standards consist of 8 bacterial and 2 fungal species, which are common members of the human gut microbiota. Community Standard I (CSI) contains these species in equal distribution, while Community Standard II (CSII) has species in exponential distribution. For our benchmark, we only considered the bacterial species in both communities.

Species	Relative abundance (%)
Bacillus subtilis	12.5
Enterococcus faecalis	12.5
Escherichia coli	12.5
Limosilactobacillus fermentum	12.5
Pseudomonas aeruginosa	12.5
Salmonella enterica	12.5
Staphylococcus aureus	12.5

Table 1 Bacterial composition of the CSI dataset

Species	Relative abundance (%)
Bacillus subtilis	0.89
Enterococcus faecalis	0.00089
Escherichia coli	0.089
Limosilactobacillus fermentum	0.0089
Listeria monocytogenes	89.1
Pseudomonas aeruginosa	8.9
Salmonella enterica	0.089
Staphylococcus aureus	0.000089

Table 2 Bacterial composition of the CSII dataset

Illumina sequencing results of the two Zymo datasets were retrieved from the ENA archive. The datasets contained 8.8 million (2*151 bp, MiSeq) and 47.8 million read pairs (2×101 bp, HiSeq) of the CSI and CSII samples, respectively. Quality control of the sequencing data was performed with Trimmomatic (Bolger et al., 2014), with default parameters, with an average minimum quality of 20 and a minimum sequence length of 75 bp. TruSeq Y adapters were removed from reads using Cutadapt (Martin, 2011).

MultiQC (Ewels et al., 2016) was used to assess quality after trimming, using the default command.

Additionally, we set up classifiers for the analysis of the Zymo data with more modern database versions, where it was available:

Kraken2: Kraken2 standard bacterial database

MetaPhlAn3: “mpa_vOct22_CHOCOPhlAnSGB_202212”

NABAS+: RefSeq, 2022

In the case of GOTTCHA, because its database is updated infrequently, we used the same reference database as for the analysis of the CAMI II samples. Since not all databases contained the updated taxonomies for *Bacillus subtilis* (Dunlap et al., 2020) and *Lactobacillus fermentum* (Zheng et al., 2020), these taxa were treated as “groups” to reconcile the differences in taxon names.

6. Collecting and analysing real-world metagenomic sequencing data

For the initial testing of classifiers, we used human metagenomic data provided by DeltaBio 2000 Ltd. These datasets originated from stool samples of healthy humans. By using the microbiome analysis service DeltaBio 2000 Ltd, patients agree to the anonymized use of their samples for research purposes.

Stool samples were collected, and DNA was isolated using the QIAamp PowerFecal Pro DNA Kit. Next-generation sequencing libraries were prepared using Illumina Nextera XT DNA Library Preparation Kit (FC-131-1096 Illumina) according to the manufacturer's

instructions. For quality control, libraries were run on a BioAnalyzer2100 instrument using High Sensitivity DNA Kit (5067-4626 Agilent). Fragment libraries were sequenced on an Illumina NextSeq500 instrument with 2*150 bp chemistry (20024904 Illumina). Quality control and trimming was done using Trimmomatic, Cutadapt and MultiQC, according to the parameters described previously.

7. Collecting and analysing clinical data

We aimed to test NABAS+ on samples that come from real-world clinical settings but still contain species whose presence has been verified with laboratory methods other than metagenomics. For this we collected a dataset of 20 samples, coming from a study describing 330 samples with verified pathogenic content (Angel et al., 2025). These samples have been tested with PCR-based and MCS (molecular, culture and sensitivity) assays for common pathogens, e.g. *Salmonella* and *Campylobacter spp*.. Samples were collected as paired-end FastQ files, from ENA, from the accession PRJNA1156595.). Quality control and trimming was done using Trimmomatic, Cutadapt and MultiQC, according to the parameters described previously.

8. Selecting and setting up classifiers for the initial gut metagenome analysis

For the initial comparison of human microbiome samples, we set up 8 different metagenomic classifiers, chosen based on popularity and diversity of classification algorithms applied. The classifiers were run with their respective default databases and standard commands when applicable, on a desktop computer equipped with 64 GB RAM and 12 processor threads, on a Linux operating system. For this initial analysis, we compared unfiltered results.

For Kaiju, we tested both the freely available webserver *(Kaiju Web Server - Submit Jo*b, n.d.) and the desktop version of the software were tested. This allowed us to run Kaiju with the NCBI nr database, which we could not process locally due to computational limitations. During the publishing process of these results, this webserver has become unavailable. Classifiers and databases used:

- Bracken/Kraken2: Kraken2 standard bacterial database
- Centrifuge: ‘p_compressed_2018_4_15’
- CLARK: ‘bacteria’
- DIAMOND: “NCBI nr”
- GOTTCHA: ‘GOTTCHA_BACTERIA_c4937_k24_u30_xHUMAN3x.species’
- Kaiju (local): ‘RefSeq’
- Kaiju-webserver: ‘NCBI nr’
- MetaPhlAn3: ‘mpa_v31_CHOCOPhlAn_201 901’

9. Selecting and setting up reference classifiers for NABAS+ benchmarking

For the initial test on our in-house in silico dataset, we tested NABAS+ along the 8 classifiers listed in the previous section. For further evaluation against the CAMI 2 and Zymo datasets, we benchmarked our tool along three reference classifiers, namely MetaPhlAn3, Kraken, and GOTTCHA. These tools showed the highest similarity to our own in the initial testing, applied diverse approaches in metagenomic classification, showed popularity of the metagenomic community, as well as good performance in other benchmarking studies. Additionally, by selecting GOTTCHA, we could compare our classifier to another BWA-based algorithm.

10. Running the classifiers

Classifiers were run with default commands when applicable. In the case of Kraken2, an additional threshold was set to provide a fair comparison (as unfiltered Kraken2 is known to produce a lot of false positives): we only accepted species supported by at least 100 fragments. MetaPhlAn3 and GOTTCHA results were not filtered post-classification.

In the GOTTCHA analysis, the ‘--minQ 0’ parameter was used for CAMI2 samples 4, 8, 10, 13, and 15, to avoid the ‘0% of reads passing filters’ exception. Without this parameter, GOTTCHA was unable to identify any species from these samples.

Classifiers were run with default settings where applicable, on a desktop computer equipped with 64 GB RAM and 12 processor threads, on a Linux operating system.

11. Statistical analysis

To test the accuracy of our classifier, we utilized several commonly used statistical metrics. Classifier outputs were collected and compared to the reference datasets. Results were considered only at the species level; non-bacterial and non-archaeal hits were removed, and percentages were recalculated only for the remaining species.

Precision or positive predicted value is used to calculate the ratio of true positives (in our case of species correctly classified) based on the following formula:

$$\frac{TP}{TP + \mathrm{FP}}$$

Recall, also known as true positive rate (TPR) measures the percentage of actual positive samples that were correctly identified by the classifier, and it was calculated as follows:

$$\frac{TP}{TP + FN}$$

F1-Score(Chinchor, 1992): This score is the harmonic mean of Precision and Recall, calculated as:

$$\frac{2 * TP}{2 * TP + FP + FN}$$

False Discovery Rate (Benjamini & Hochberg, 1995) was calculated as:

$$\frac{FP}{TP + FP + TP + TN}$$

Where TP is the number of true-positive species, FP is the number of false positives and FN is the number of false negatives.

Additionally, the following diversity metrics were also calculated:

Jaccard-distance (Jaccard, 1912):

$$1 - \frac{TP}{\text{Number of species in either ground truth or classification result}} * 100$$

Bray-Curtis distance (Bray & Curtis, 1957):

$$1 - \frac{\text{The sum of the lesser abundance values for TP species}}{2 * \text{TP} + \text{FP} + \text{FN}} * 100$$

12. Software used and code availability

Development, comparisons and calculations were performed in Java, graphs were generated with Python3.10 using the seaborn package and in R (Version 4.2.1.), using the ggplot2 (version 3.3.6.) and ggbreak (version 0.1.1.) packages.

Figure 1-4 and 6 were generated with BioRender.

All the generated code, including a stand-alone version of NABAS+ is available on GitHub at the following repository: https://github.com/TakacsBertalan/NABAS_paper_scripts

VII. Results

1. Comparing the performance of nine commonly used metagenomic classifiers

There are several freely available metagenomic classifiers in the metagenomics community. Although there are less- and more popular ones, currently there is not a universally accepted best classifier in the community. To study the similarity metagenomic classifiers show in their results, we tested nine of such tools on five real-world shotgun sequenced stool samples. We found that these software showed highly discordant results testing on the same samples. There was a large disagreement in the number and identity of classified species (Figure 5).

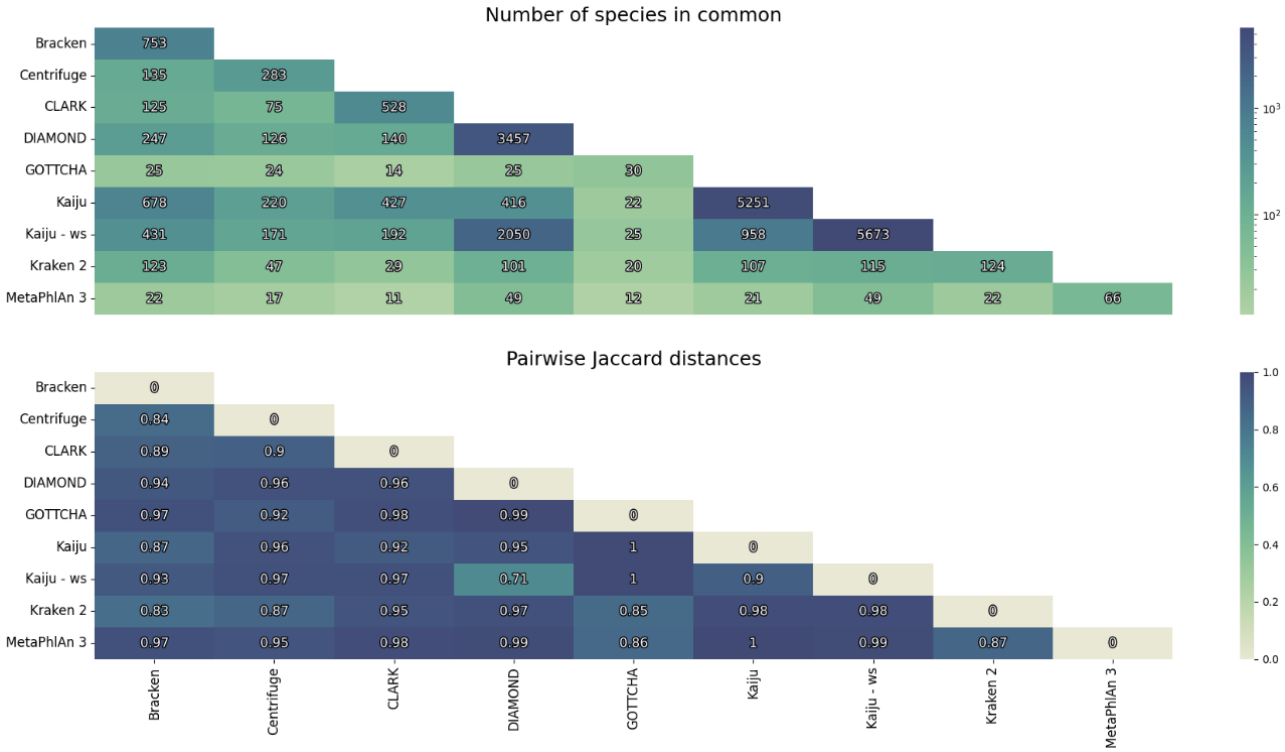


Figure 5: Number of found species and Jaccard-distance results of 9 metagenomic classifiers

Kaiju-ws: Kaiju-webserver

(Takács et al., 2025)

Based on the number of common species, we found the biggest agreement between Kaiju and Kaiju-webserver, which is not surprising as the latter is an implementation of the former. Both local and webserver versions of Kaiju also showed high concordance with the DIAMOND results. This is probably due to the same nucleotide-to-protein search strategy applied by both classifiers. We also found a large concordance between Kraken2 and Bracken, which is again to be expected, as they are developed by the same research group and work on the same base principles.

When calculating Jaccard-distance, which takes into account not only species identity but also the number of total species identified, we could observe the biggest similarity between DIAMOND and Kaiju-ws. Along with similar classification principles applied, this could be caused by the reference database used by both classifiers, ncbi-nr. We also observed a higher concordance between Kraken2-Bracken and Kaiju-Kaiju-ws. Interestingly, we also found larger similarities between Kraken2, GOTTCHA and MetaPhlAn3, the latter two classifiers identified the lowest number of total species.

To demonstrate that these differences are not solely due to the varying analytical depth of the classifiers, we also calculated the number of species identified by all tools in common. On average, only five species per sample were shared across all classifiers, revealing a substantial discrepancy between the consensus set and the total number of species reported individually.

It’s important to point out that this preliminary experiment was done using the default settings for all classifiers and fine-tuning the tools would bring these results closer to each other. At the same time, this fine-tuning needs to be experiment- and classifier-specific as there are no “gold standard” threshold values or “best settings” provided for each classifier.

To address these challenges, we developed NABAS+, a metagenomic classifier designed to deliver consistent, high-confidence species identification by minimizing false positives and relying on high-quality reference genomes. The following section outlines the main design principles and workflow of NABAS+.

2. Introducing a novel metagenomic classifier, NABAS+

We aimed to create a reliable, accurate metagenomic classification algorithm. Figure 6 shows a brief overview of the main steps of NABAS+.

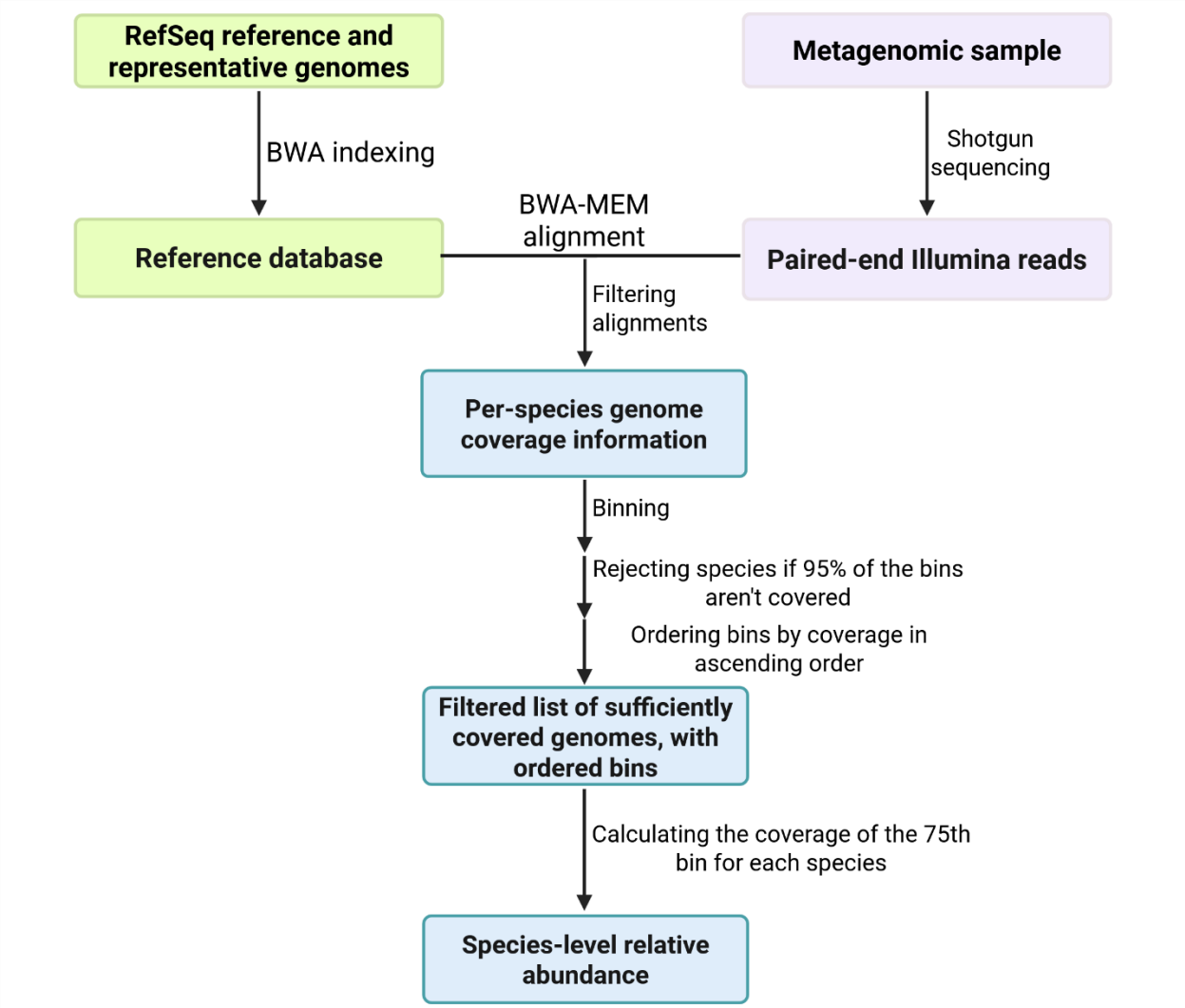


Figure 6: Simplified workflow of NABAS+
(Takács et al., 2025)

The first part of the analysis is database creation. This needs to be done once initially and every time the reference sequences are updated (e.g. a new species is added). To get the most reliable classification results, we used genomes RefSeq database, selecting a representative genome for each bacterium. Genomes flagged as “representative” are generally considered the current highest quality representation of the genome in the given

species in RefSeq. By picking a representative for each species, we aimed to ensure high-quality, unambiguous classification results. Additionally, this reduced the size of the reference database considerably: RefSeq contains over 440000 genome assemblies and 22,082 of them are flagged as “representative”, as of September 2025 *(*NCBI Insights, 2025).

To make the indexing and the alignment process more efficient, we processed the databases in chunks, each containing ~ $8*10^9$ bases. This makes updating our database simpler as we don’t need to replace all genomes for an update: it’s enough to replace the respective chunk. Furthermore, this makes it easy to add any new genome by simply creating and indexing a new chunk.

Chunking the database also reduces the RAM requirements of our tool considerably. We estimate that indexing of one such chunk requires approximately 16 GBs of RAM, while indexing the whole database at the same time would require 140+ GBs and that the RAM cost of the BWA-MEM alignment is reduced similarly. Estimations were based on the original BWA paper (Li & Durbin, 2009).

The second stage is metagenomic classification: NABAS+ processes paired-end Illumina reads (2*151 bp). In the first step of the analysis, the reads are aligned against our reference database using BWA-MEM. BWA-MEM works by scanning the reference for maximum exact matches, creating ‘seeds’, from which alignment can be extended. Alignments are then scored with the Smith-Waterman algorithm. This, however, makes the mapping more extensive compared to other popular metagenomic classifiers. By contrast, MetaPhlAn aligns the query sequences against a database of clade-specific sequences using Bowtie2, which utilizes the Burrows-Wheeler transformation, similarly to BWA. While Kraken2 processes reference sequences as *k*-mers (usually with a length of 31 bases), builds a taxonomic tree of the latest common ancestor (LCA), and maps these query sequences to this tree.

This extensive alignment is the most time-consuming step of running NABAS+ and takes advantage of the multi-threading capability of BWA-MEM and the split database. According to our estimates, based on the work of Hanussek et al. (2021), using 12 CPU threads accelerates alignment five- to eightfold, with further improvements possible

through additional parallelization. Samples are aligned to each database chunk individually, and the SAM outputs are converted into BAM format using samtools. The BAM files are then merged, and low-quality alignments are filtered out, based on the CIGAR string of each aligned read, which is a string representation of the alignment. By default, we allow a maximum of 10 edit distances (including mismatches, indels, and softclips) between the read and the reference; above that the alignment gets rejected. From the header of the BAM files, we collect the covered genomes and reject those that do not have a minimum of 10 aligned reads. The set cut-off values for edit distances and the minimum number of aligned reads were determined through extensive empirical testing and can be modified by the user.

To assign taxonomy to the species corresponding to the genome, we build a hash map containing the names and corresponding taxonomic node, using the *nodes.dmp* and *names.dmp* of the NCBI Taxonomy dump package. This is done only once per run, after the alignment step.

For each remaining genome, we calculate an actual genome coverage (% of the genome covered by the reads) and a hypothetical genome coverage (total read length of sequences aligning to the genome/length of the genome) value. If hypothetical coverage/actual genome coverage is larger than 3.5, we reject the genome. This filters out genomes with disproportionate coverage as well as strongly over-represented regions. Through empirical testing, 3.5 has been found to be the optimal threshold value, lowering it decreased specificity in general.

Subsequently, we divide each genome into 100 equal-sized bins. To access genome length information fast, we read the *.ann* file of each database chunk and store it in a hash map. For each genome, we count the bins that have at least one read. The species corresponding to the genome is considered present if 95 of the 100 bins have at least one read. For these genomes, we order the bins in an ascending order and count the number of reads in the 75th bin so that over-represented, possibly non-species-specific regions are avoided. During the filtering steps, the results are collected in an Excel file

showing the taxonomy of the identified species per bin, average coverage values, and relative abundances. This Excel file is the final report of the analysis.

Because the previously mentioned microbial samples were taken from real-world gut microbial communities, we can't be exactly sure about their exact microbial composition without further laboratory tests. To properly test our novel algorithm and to compare its performance to other classifiers, we needed datasets with exact known compositions.

3. Testing NABAS+ on in-house generated datasets

To evaluate the accuracy and robustness of NABAS+, we first tested its performance on six *in silico* mock metagenomic communities generated using our in-house read simulation pipeline.

We created 6 such communities, comprising 212 bacterial species in total, modeling both environmental and human-associated microbiomes. The experiment was intended to evaluate the performance and robustness of NABAS+ under idealized conditions in which the exact community composition was known, providing a controlled baseline for later comparisons with more complex datasets.

The following figure (Figure 7) shows the number of species identified by each classifier for each sample. Samples from the same community with different read numbers were averaged.

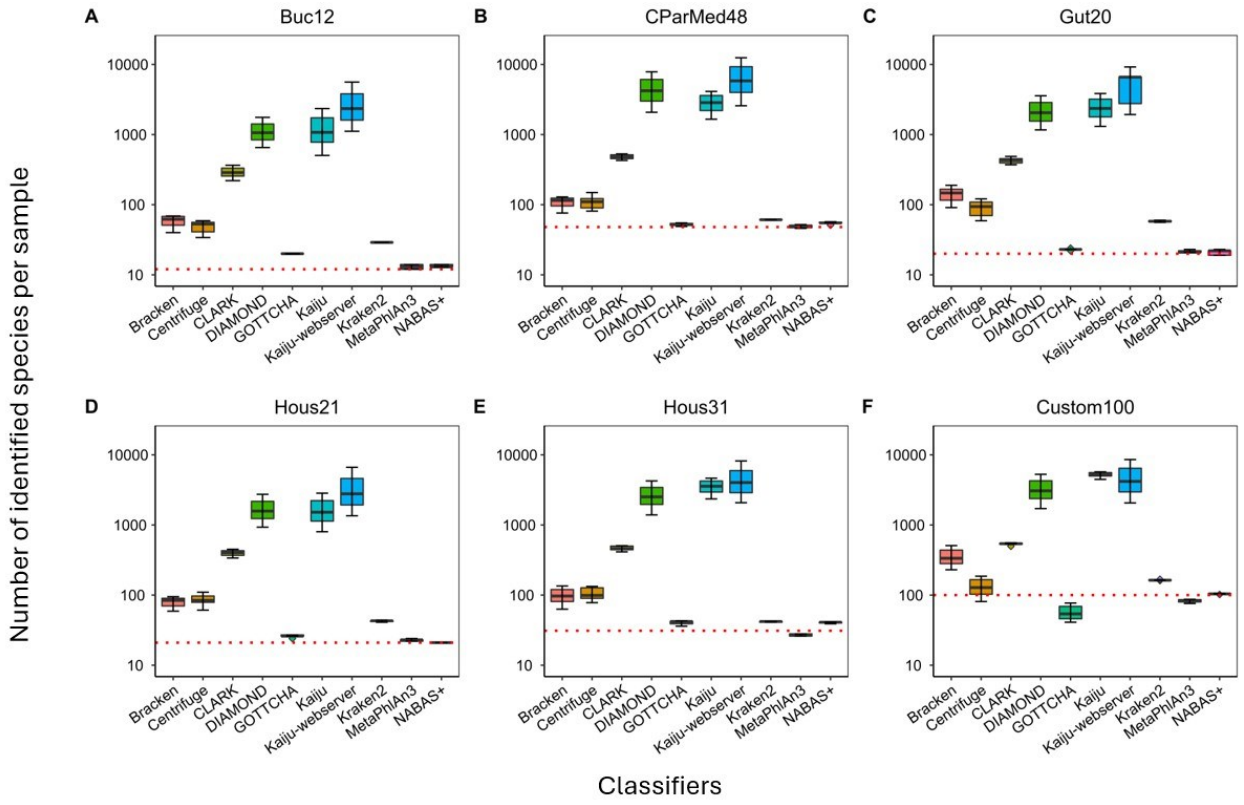


Figure 7: Number of species identified by each classifier across the in-house datasets

Dotted red line: actual number of species present in each community

(Takács et al., unpublished)

On this dataset we could show that NABAS+ accurately classified the right number of species in each sample. The composition and "origin" of the dataset did not seem to influence the classification accuracy of our tool: it produced correct results on communities modelling urban environments (CParMed48) as well as ones modelling the human oral and gastrointestinal microbiome (Buc12, Gut20 and Custom100).

Similar to the results shown in Section 1, there was a considerable variation in the number of species detected by the different classifiers. Classifiers based on similar algorithmic principles, such as DIAMOND and Kaiju, produced correspondingly similar outputs. In contrast, the tools whose results most closely matched those of NABAS+—Kraken2, MetaPhlAn3, and GOTTCHA—are built on distinct underlying methodologies.

Finding the optimal read number for a sample is an important question of metagenomics. A “too shallow” sequencing can lead to losing low-abundance species (Pereira-Marques et al., 2019) and sequencing “too deeply” could lead to the multiplication of classification errors and an increase in false positives.

By creating a different visualization for the same experiment, we can get a better picture on how differing read numbers (ranging from $5*10^5$ to $1*10^7$) affect the classification accuracy of the classifiers and NABAS+. Figure 8 shows the number of species found in the samples in the 6 mock communities. As earlier, the red dotted line indicates the actual number of species in the samples.

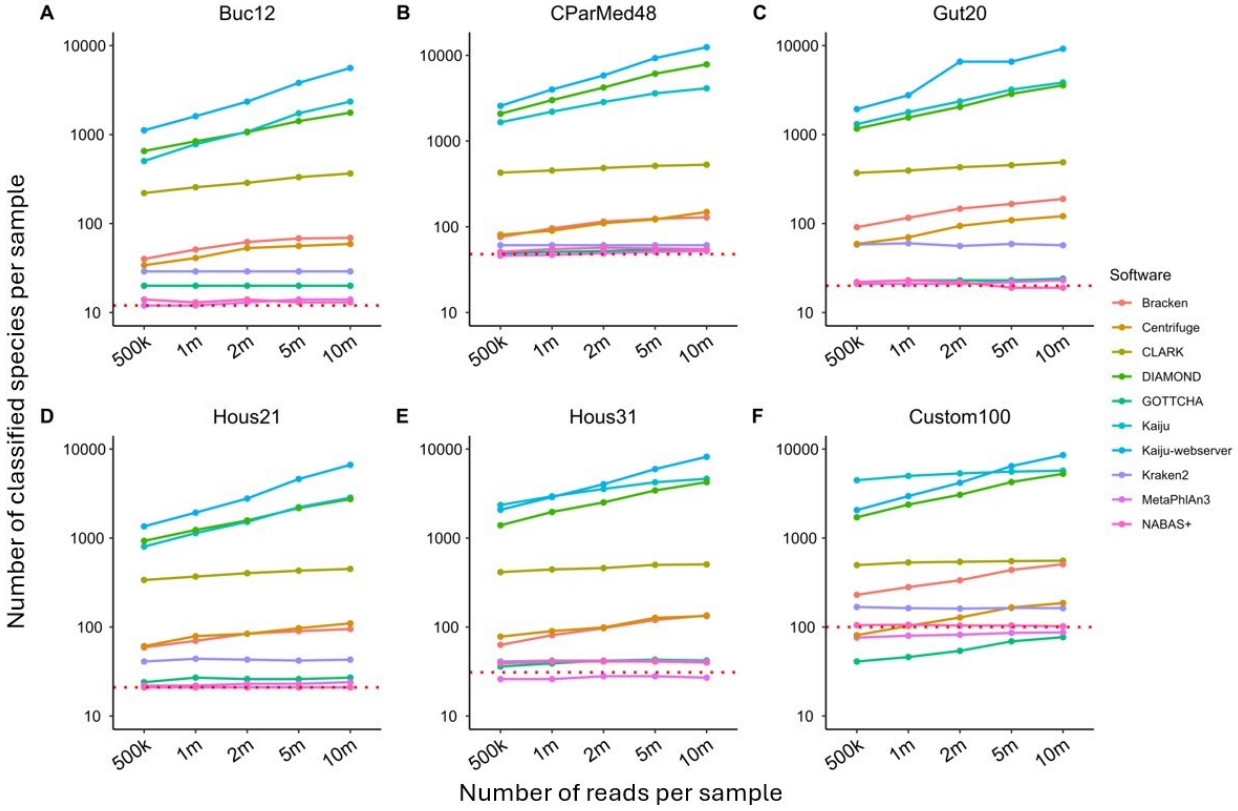


Figure 8: Number of identified species in the samples per read number

(Takács et al., unpublished)

NABAS+ maintained stable accuracy across all sequencing depths tested. The number of detected species remained close to the true value even at the highest coverage levels, indicating that our tool is largely insensitive to changes in sequencing depth. This

contrasts with several classifiers that exhibited inflated species counts as coverage increased.

The classifiers whose results most closely resembled those of NABAS+ again exhibited similar behavior. Their detected species counts were largely unaffected by changes in sequencing depth. A consistent pattern was observed: MetaPhlAn3 tended to slightly underestimate the number of species, whereas Kraken2 tended to overestimate it.

To provide an integrated performance measure, we next evaluated the F1 score for each classifier, combining precision and recall to capture both sensitivity and specificity.

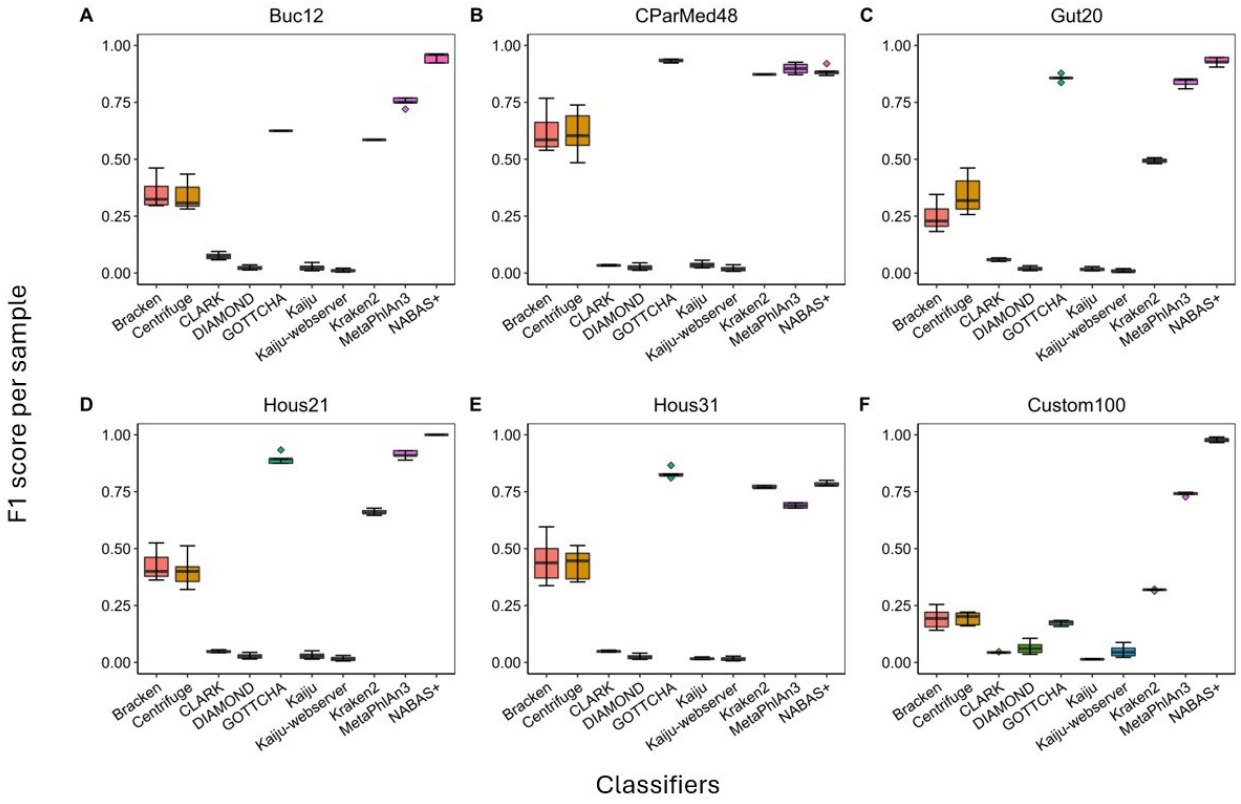


Figure 9: F1 scores of tested classifiers on the in-house dataset

(Takács et al., unpublished)

NABAS+ achieved uniformly high F1 scores across all six mock communities, confirming its balanced precision and recall. Its performance remained stable on all datasets, including those modeling human and environmental microbiomes. While some classifiers

(e.g., Kraken2) showed reduced precision on specific datasets such as Custom100 and Gut20, NABAS+ maintained consistent accuracy and specificity throughout.

These results confirm that NABAS+ can accurately identify bacterial species in controlled *in silico* samples and that sequencing depth has little effect on its classification performance. Its reliability across communities of different origin indicates that the tool generalizes well beyond the human gut microbiome.

Even though these results are promising for our classifier, there are important caveats we must point out: since our samples were created from RefSeq reference genomes, with uniform distribution, they resembled an ideal metagenomic sample, rather than a realistic one. In the real world, metagenomic datasets tend to be “noisier” often containing sequences of low-abundant or lesser-known species. Additionally, since the NABAS+ reference database was also created based on RefSeq, simply relying on this dataset to verify our classifier would carry the danger of overestimating the accuracy of our software. Moreover, we benchmarked NABAS+ against classifiers run with default settings: the caveats detailed in the previous section apply here as well.

To further assess the robustness of NABAS+ under more realistic yet controlled conditions, we next benchmarked it using the CAMI II gastrooral dataset, a set of *in silico* samples designed to emulate human microbiome composition.

4. Benchmarking NABAS+ and 3 other classifiers on the CAMI II gastrooral *in silico* data

Building on the findings from our in-house datasets, we extended our analysis to the CAMI II gastrooral dataset to test NABAS+ under more realistic, yet still well-defined conditions.

Based on the results of the previous section, we decided to continue the benchmark against the 3 classifiers that showed the most similar performance to NABAS+ in the previous experiments: Kraken2, Metaphlan3 and GOTTCHA. All 3 (and especially Kraken2 and Metaphlan3) have a high number of citations and are broadly used by the scientific community.

After running the classifiers as described in the ‘Materials and methods’ section, we utilized the following metrics to evaluate their performance: F1 Score, Precision, and Recall.

We found that MetaPhlAn3, GOTTCHA, and NABAS+ produced the highest F1 scores (Fig. 10A). As it was previously observed by other studies, Kraken2 tended to produce a high number of false positives, leading to a lower precision score (Fig. 10B), despite the applied threshold mentioned in the ‘Materials and methods’ section. It is also important to point out that Kraken2 often outperformed the other tools in Recall (Fig. 10C), indicating that it was able to identify more positive true species in the samples.

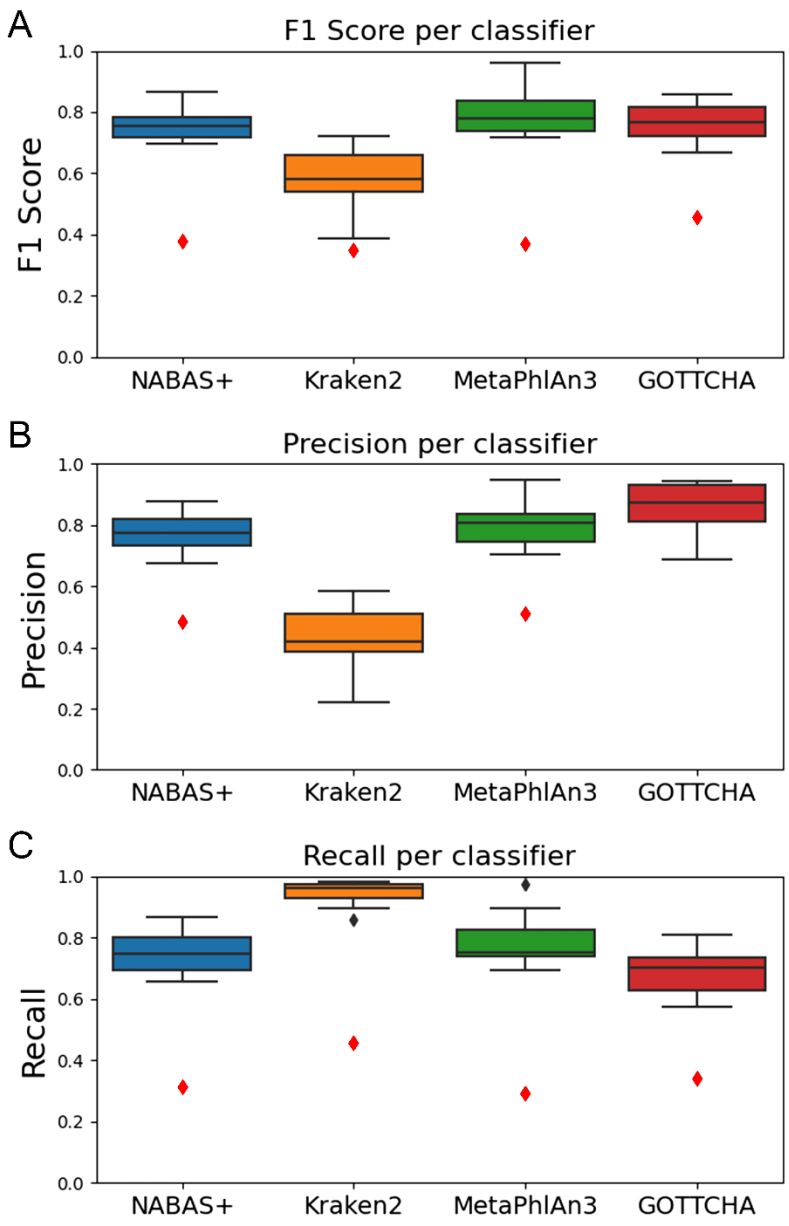


Figure 10: Performance comparison on the CAMI II dataset

The outlying sample19 is marked with red

(Takács et al., 2025)

5. Examining and re-creating an outlier CAMI II sample

In the case of one sample of the CAMI II “human gastrooral” dataset (sample19), we observed that all our tested classifiers produced low classification accuracy (Figure 10). The examination of this sample showed that the reference genomes it was created from were often from the first half of the 2010s, or were of low confidence, while the reference database of the classifiers likely had more recent, higher quality references for the same species. We re-created this sample with CAMISIM to see if the low classification accuracy was due to this discrepancy between genome versions. We theorized that re-creating this sample with current genomes will improve classification accuracy. For this, we utilized the same CAMISIM settings as the authors of the original CAMI II dataset but replaced the reference genomes with the latest RefSeq reference version.

6. Testing classifiers on the regenerated CAMI II sample19

Sample name	Classifier	Precision	Recall	F1 Score
sample19-new	GOTTCHA	0.912	0.646	0.756
sample19-old	GOTTCHA	0.688	0.344	0.458
sample19-new	Kraken2	0.431	0.969	0.596
sample19-old	Kraken2	0.386	0.458	0.419
sample19-new	MetaPhlAn3	0.778	0.729	0.753
sample19-old	MetaPhlAn3	0.509	0.292	0.371
sample19-new	NABAS+	0.719	0.719	0.719
sample19-old	NABAS+	0.484	0.313	0.380

Table 3 Performance of the classifier on the original and newly generated CAMI II sample19
(Takács et al., 2025)

After running classification on this sample, we observed an increase across all the classification metrics for all the examined classifiers. This indicates that classification performance depends not only on algorithmic design but also on the quality and currency of the reference database.

7. Testing classifier performance on Zymo standards

So far, we have presented NABAS+'s accuracy on *in silico* generated datasets. To take testing a step further, we wanted to measure its accuracy in real-world NGS samples as well. We used two Illumina sequencing runs of this community, both encompassing the same eight bacterial species and two fungal species. One sample (Zymo CSI) contained the bacterial species in equal abundance and the fungal ones in small percentage and was sequenced for 8.8 million reads, whereas the other had the species with logarithmically distributed abundances and contained 47.8 million reads (Zymo CSII).On the Zymo CSI dataset, all classifiers performed well (Figure 11A), with all species accurately identified and relative abundances close to ground truth; GOTTCHA, Kraken2, and MetaPhlAn3 found a significant percentage of false-positive species, compared to NABAS+ that found none. Comparing Bray-Curtis distances showed the same (Table 4): NABAS+ produced the lowest distance from the original composition of both Zymo sets, suggesting that it was able to give the most accurate classification out of the classifiers studied.

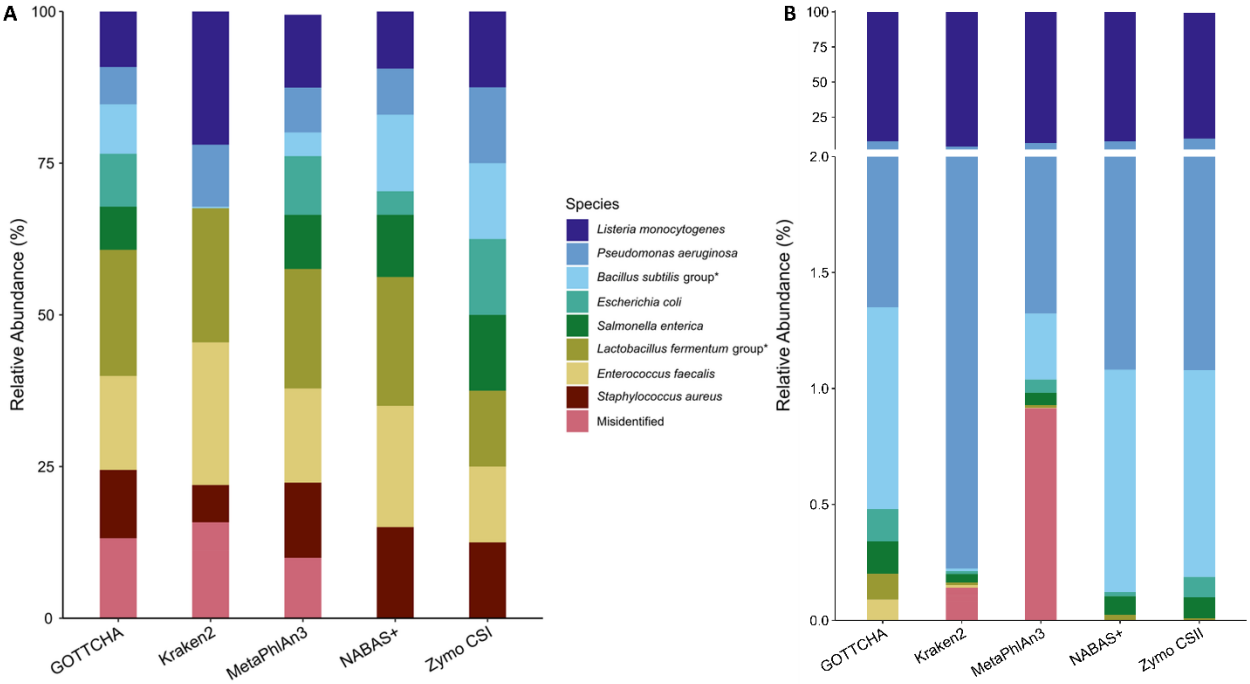


Figure 11.: Classification performance on the (A) Zymo CSI and (B) Zymo CSII datasets (Takács et al., 2025)

Classifier	Zymo CSI	Zymo CSII
GOTTCHA	0.622	0.515
MetaPhlAn3	0.604	0.526
Kraken2	0.73	0.534
NABAS+	0.594	0.515

Table 4. Bray-Curtis distances between the classifier results and the original compositions of the Zymo datasets
(Takács et al., 2025)

Zymo CSII data showed similar results across all four classifiers (Figure 11B). The logarithmic nature of species abundances may explain why there was little variance in the Bray-Curtis distances between the classifiers. Kraken2 and MetaPhlAn3 both identified some species falsely, and though GOTTCHA showed high sensitivity, NABAS+ was still the classifier that produced the closest value to the ground truth.

The analysis of the Zymo standards served as a crucial link between controlled in silico experiments and real-world sequencing data. Across both Zymo datasets, NABAS+ consistently achieved the most accurate reconstruction of microbial community composition, yielding the lowest Bray–Curtis distance to the known ground truth and producing no false positives. Notably, it successfully detected *Enterococcus faecalis* at a relative abundance of only 0.00089%, demonstrating its good sensitivity. These results, obtained from genuine Illumina sequencing runs rather than simulated reads, confirm that NABAS+ performs reliably under experimental conditions and is well suited for subsequent application to clinical stool samples.

8. Demonstrating NABAS+'s utility on a real-world clinical dataset

As mentioned in the Introduction, the use of metagenomics in a clinical setting is not considered standard practice currently. Because we developed NABAS+ mainly for the analysis of the human gut microbiome, it is crucial that if we identify pathogens in the sample, we can be sure about their identity. To demonstrate if NABAS+ was fit for this purpose, we obtained a dataset from Angel et al., 2025 and analyzed a random subset of the samples using our own method.

Eleven of these samples were confirmed as positive for either *Salmonella enterica*, *Aeromonas veroni*, or *Campylobacter jejuni*, while the other nine tested negative for the same species (Table 5).

Number of samples = 20	Positive NABAS+ result	Negative NABAS+ result
Positive laboratory test result	11	0
Negative laboratory test result	0	9

Table 5. Confusion matrix of the pathogen identification using NABAS+
Laboratory testing was either culturing or PCR-based, as described by Angel et al. (Takács et al., 2025)

Our results show that NABAS+ was able to accurately identify the correct pathogen in the infected samples and did not report any pathogenic species in the negative samples. This illustrates the potential applicability of NABAS+ in a clinical setting. Taken together, our findings highlight NABAS+ as a robust and reliable metagenomic classifier that performs on par with, and in several cases surpasses, currently available tools. Its ability to combine precision and robustness underscores its potential for integration into clinical microbiome diagnostics and broader microbiological research.

VIII. Discussion

This dissertation highlights the importance of the human microbiome and demonstrates that the challenges of metagenomic classification remain far from solved. Although multiple classifiers are available, they often produce highly discordant results when tested on real metagenomic samples. These discrepancies concern both the number and the identity of the species classified. As metagenomics is increasingly introduced into clinical diagnostics, reliable and accurate classification methods are urgently needed.

In our work, we showed that nine of the most commonly used classifiers display considerable disagreement on real-world stool metagenomic samples. This means that analysing a sample with multiple tools does not necessarily increase accuracy; rather, it may introduce additional uncertainty. In research use, post-classification filtering is often applied to classification results, to filter out false positive hits, but the thresholds of this are often decided on a case-by-case basis.

These challenges motivated us to develop our own classifier, NABAS+, designed to minimize false positives by relying on reliable genome assemblies. Unlike most tools, NABAS+ was developed with the aim of avoiding the need for post-classification filtering, thereby displaying a robust performance across different sample origins.

Our results demonstrate that NABAS+ performs well both on *in silico* datasets and real-world metagenomic standards, and that it can also be successfully applied to clinical patient data. On our in-house *in silico* dataset, we observed that NABAS+ consistently performed well in detecting the correct number of species and showing a high F1 score. Changing the read numbers did not seem to affect NABAS+, it showed robust performance on samples ranging from $5*10^5$ to $1*10^7$.

We further benchmarked NABAS+ against the other three classifiers of similar performance—Kraken2, MetaPhlAn3, and GOTTCHA—using standardized microbial community datasets. On CAMI II data, MetaPhlAn3 performed best, while NABAS+ and GOTTCHA followed closely. Kraken2 showed weaker performance, largely due to high false positive rates. This illustrates the good performance of our tool on *in silico* datasets

not generated from RefSeq reference genomes, with more complex species distribution and sequencing error profile than our own.
On Zymo standards, NABAS+ showed good performance, producing no false positives and yielding the lowest Bray-Curtis distance from the ground truth. Importantly, NABAS+ was also able to correctly identify pathogens in clinical stool samples, underscoring its utility in real-world diagnostic applications. Beyond the scope of this dissertation, NABAS+ has already been applied successfully in a clinical study of Crohn's disease (Bacsur et al., 2024), further demonstrating its real-world applicability.

Overall, our findings show that NABAS+, a BWA MEM-based alignment classifier, can produce comparable results, and in some cases performs better than the most popular metagenomic tools. Our results indicate alignment-based classification methods are capable of showing good performance when paired with curated databases, despite their underutilization in recent years. Although alignment-based approaches are dismissed by some because of their high computational demands, NABAS+ mitigates this limitation through a carefully filtered reference database. By including only reliable reference or representative assemblies—one genome per species—the total number of reference genomes is reduced approximately twentyfold. In addition, dividing the database into smaller segments further decreases computational requirements, lowering the memory needed for indexing and alignment by roughly a factor of eight, according to our estimates. Moreover, unlike other classifiers that depend on clade-specific or marker-based databases (where creating a new reference database can be computationally intensive and thus not happen regularly), this split means NABAS+ databases can be easily rebuilt and updated with new genome versions using simple BWA indexing. We observed that in the case of GOTTCHA, where some samples had to be ran with lowered thresholds to produce any results. We believe that this is partly due to the outdatedness of GOTTCHA's reference database.

Nevertheless, we identified limitations in all classifiers when analysing CAMI II sample 19, which was generated from older, less reliable genome assemblies. Recreating the sample with modern assemblies markedly improved performance, suggesting that certain CAMI II datasets may no longer be suitable benchmarks without modernization. This issue is

likely not unique to CAMI II and may extend to other *in silico* benchmarking resources, reflecting the rapid evolution of metagenomics. We also identified discrepancies between classifier performance and the quality of widely used benchmarking datasets. Given the rapid pace of developments in bacterial taxonomy and genome sequencing, our results suggest that in silico benchmarking datasets should be updated regularly to remain relevant for classifier evaluation.

Despite its strengths, NABAS+ has certain limitations. Because it relies on a curated database, it is not suitable for analysing samples dominated by unknown species or for the discovery of novel taxa. NABAS+ was designed for species-level metagenomic classification, primarily in the context of the human gut microbiome, and has not been optimized for non-human or highly diverse environments (e.g., soil, wastewater). Even though NABAS+ performed well on the in-house generated samples that model such environments (e.g. CParMed48, Hous31), we suggest the fine-tuning of parameters and database composition before applying it to environmental samples.

Even though NABAS+ showed robust performance across a broad range of read numbers, it is important to point out that these findings were limited to controlled datasets with uniform read quality and distribution. To establish broader conclusions about minimum read requirements of in real-world applications, further studies are needed.

Strain level classification is an important task in metagenomics, especially if we intend to use these methods in clinical practice, as different strains of the same bacterium can have vastly different effects on human health. Testing our software's capabilities to distinguish between different strains or subtypes was beyond the scope of this dissertation. However, we think that proper database customization could enable this in future.

While this dissertation focused on bacterial communities, NABAS+ also showed promise for virome analysis, and a eukaryotic reference database is under development. Additionally, long-read metagenomics has been gaining traction in the recent years, and we think that, with parameter optimization, NABAS+ could likely be adapted for long-read sequencing data. Similar alignment-based strategies are already employed in that field (Li et al., 2021) and there are already classifiers utilizing BWA-MEM (Curry et al., 2022),

suggesting that NABAS+ could be fine-tuned or extended work in a similar manner. Collectively, these directions highlight the versatility of NABAS+ and its potential to evolve into a comprehensive framework for metagenomic classification across diverse sequencing platforms and organism groups.

IX. Conclusions

In the described work, I demonstrated the discordance between different, commonly used metagenomic classifiers and introduced a novel classification tool, NABAS+. NABAS+ is based on the alignment algorithm BWA-MEM and has a reference database containing reliable bacterial genomes from the RefSeq database. NABAS+ was written in Java and is freely available as a stand-alone software.

I created custom *in silico* datasets to test this software and collected other benchmarking datasets, including data from the CAMI II challenge, deeply sequenced microbial mock communities from Zymo and real-life stool samples containing pathogens, collected in a hospital environment.

Benchmarking NABAS+ against the mentioned metagenomic classifiers on both in silico and real-world samples showed comparable performance to the most commonly used and most accurate classifiers (GOTTCHA, Kraken2, MetaPhlAn3). NABAS+ was particularly suitable for minimizing the number of false positives and produced the highest similarity to the deeply sequenced Zymo standards.

We were also able to demonstrate the clinical applicability of NABAS+ on a real-world dataset, where it was able to find the correct pathogens in the infected samples, while it did not classify false positives in the non-infected samples.

Overall, we could demonstrate that NABAS+ is a reliable tool not only for research but also for clinical settings, contributing to improved accuracy and reproducibility in metagenomic studies and its ability may benefit the broader metagenomic community.

X. Acknowledgements

I would like to thank my supervisor, Lajos Haracska for his work and support.

I am thankful for the help of the members of the HUN-REN BRC Carcinogenesis and Mutagenesis Laboratory, as well as members of the HUN-REN BRC Genetics Institute for their support. Special thanks go to my fellow PhD students, Alexandria Qorri, Emese Pekker and Valentin Varga who helped me a lot during my PhD years.

I am grateful to my opponents for their work and contribution.

I would also like to thank Gábor Jaksa, Lajos Pintér, Zoltán Gyuris, all current and former employees of Deltabio 2000 and European Life Technologies, including but not limited to Bence Széplaki, Bernadett Csányi and Zsófia Szabados-Tóth.

I would like to thank Fulbright Hungary for funding my semester as a Visiting Student Researcher in the USA. I am thankful for the opportunity to work at the Knight lab at the University of California, San Diego. Special thanks to Rob Knight, Caitlyn Guccione, Kristina Chan, Kylie Chan, and Tiffany Zhang. I am grateful for the Hungarian Fulbright Chapter for their help, especially Anna Becsei and Károly Jókay.

I am grateful for the support and patience of my family, my girlfriend, and friends; without their help, this work couldn't have been completed.

This work received funding from the National Research, Development, and Innovation Office (GINOP-2.3.2–15-2016–00020, GINOP-2.3.2-15-2016-00024, GINOP-2.3.2–15-2016–00026, and RRF-2.3.1-21-2022-00015). Project no. RRF-2.3.1-21-2022-00015 has been implemented with the support provided by the European Union. This work was supported by the European Union's Horizon 2020 research and innovation program under grant agreement no. 739593.

XI. Bibliography

Abou Chacra, L., Fenollar, F., & Diop, K. (2022). Bacterial vaginosis: What do we currently know? *Frontiers in Cellular and Infection Microbiology*, *11*. https://doi.org/10.3389/FCIMB.2021.672429

Adams, R. I., Bateman, A. C., Bik, H. M., & Meadow, J. F. (2015). Microbiota of the indoor environment: a meta-analysis. *Microbiome*, *3*(1), 49. https://doi.org/10.1186/S40168-015-0108-3

Adzitey, F., Huda, N., & Rahmat Ali, G. R. (2012). Molecular techniques for detecting and typing of bacteria, advantages and application to foodborne pathogens isolated from ducks. *3 Biotech*, *3*(2), 97. https://doi.org/10.1007/S13205-012-0074-4

Aja-Macaya, P., Conde-Pérez, K., Trigo-Tasende, N., Buetas, E., Nasser-Ali, M., Nión, P., Rumbo-Feal, S., Ladra, S., Bou, G., Mira, Á., Vallejo, J. A., & Poza, M. (2025). Nanopore full length 16S rRNA gene sequencing increases species resolution in bacterial biomarker discovery. *Scientific Reports 2025 15:1*, *15*(1), 1–12. https://doi.org/10.1038/s41598-025-10999-8

Altheide, S. T. (2019). Biochemical and culture-based approaches to identification in the diagnostic microbiology laboratory. *American Society for Clinical Laboratory Science*, *32*(4), 166–175. https://doi.org/10.29074/ASCLS.2019001875

Amabebe, E., & Anumba, D. O. C. (2018). The vaginal microenvironment: The physiologic role of Lactobacilli. *Frontiers in Medicine*, *5*(JUN), 389042. https://doi.org/10.3389/FMED.2018.00181

An Updated Bacterial and Archaeal Reference Genome Collection is Available! - *NCBI Insights*. (2025, September 2). https://ncbiinsights.ncbi.nlm.nih.gov/2025/09/02/bacterial-and-archaeal-reference-genome-collection/

Anderson, S. (1981). Shotgun DNA sequencing using cloned DNase I-generated fragments. *Nucleic Acids Research*, *9*(13), 3015. https://doi.org/10.1093/NAR/9.13.3015

Angel, N. Z., Sullivan, M. J., Alsheikh-Hussain, A., Fang, L., MacDonald, S., Pribyl, A., Wills, B., Tyson, G. W., Hugenholtz, P., Parks, D. H., Griffin, P., & Wood, D. L. A. (2025). Metagenomics: a new frontier for routine pathology testing of gastrointestinal pathogens. *Gut Pathogens*, *17*(1), 1–17. https://doi.org/10.1186/S13099-024-00673-1

Anthony, W. E., Burnham, C. A. D., Dantas, G., & Kwon, J. H. (2020). The gut microbiome as a teservoir for sntimicrobial tesistance. *The Journal of Infectious Diseases*, *223*(Suppl 3), S209. https://doi.org/10.1093/INFDIS/JIAA497

Arumugam, V., Nagaraj, V., Muthanandam, S., Arumugam, S., & Ramasamy, D. (2025). Human oral microbiome as forensic biomarkers for individual identification: A

systematic review. *Indian Journal of Microbiology Research*, *11*(4), 230–242. https://doi.org/10.18231/J.IJMR.2024.042

Bağcı, C., Patz, S., & Huson, D. H. (2021). DIAMOND+MEGAN: Fast and easy taxonomic and functional analysis of short and long microbiome sequences. *Current Protocols*, *1*(3), e59. https://doi.org/10.1002/CPZ1.59

Baker, J. L., Mark Welch, J. L., Kauffman, K. M., McLean, J. S., & He, X. (2023). The oral microbiome: diversity, biogeography and human health. *Nature Reviews Microbiology 2023 22:2*, *22*(2), 89–104. https://doi.org/10.1038/s41579-023-00963-6

Bartholomew, J. W., & Mittwer, T. (1952). The Gram stain. *Bacteriological Reviews*, *16*(1), 1. https://doi.org/10.1128/BR.16.1.1-29.1952

Basic, M., & Bleich, A. (2019). Gnotobiotics: Past, present and future. *Laboratory Animals*, *53*(3), 232–243. https://doi.org/10.1177/0023677219836715

Beghini, F., McIver, L. J., Blanco-Míguez, A., Dubois, L., Asnicar, F., Maharjan, S., Mailyan, A., Manghi, P., Scholz, M., Thomas, A. M., Valles-Colomer, M., Weingart, G., Zhang, Y., Zolfo, M., Huttenhower, C., Franzosa, E. A., & Segata, N. (2021). Integrating taxonomic, functional, and strain-level profiling of diverse microbial communities with bobakery 3. *ELife*, *10*. https://doi.org/10.7554/ELIFE.65088

Benjamini, Y., & Hochberg, Y. (1995). Controlling the False Discovery Rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society: Series B (Methodological)*, *57*(1), 289–300. https://doi.org/10.1111/J.2517-6161.1995.TB02031.X

Berg, G., Rybakova, D., Fischer, D., Cernava, T., Vergès, M. C. C., Charles, T., Chen, X., Cocolin, L., Eversole, K., Corral, G. H., Kazou, M., Kinkel, L., Lange, L., Lima, N., Loy, A., Macklin, J. A., Maguin, E., Mauchline, T., McClure, R., … Schloter, M. (2020). Microbiome definition re-visited: old concepts and new challenges. *Microbiome*, *8*(1), 1–22. https://doi.org/10.1186/S40168-020-00875-0

Bokulich, N. A. (2025). Integrating sequence composition information into microbial diversity analyses with k-mer frequency counting. *MSystems*, *10*(3). https://doi.org/10.1128/MSYSTEMS.01550-24

Bolger, A. M., Lohse, M., & Usadel, B. (2014). Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics*, *30*(15), 2114–2120. https://doi.org/10.1093/BIOINFORMATICS/BTU170

Boolchandani, M., D'Souza, A. W., & Dantas, G. (2019). Sequencing-based methods and resources to study antimicrobial resistance. *Nature Reviews Genetics 2019 20:6*, *20*(6), 356–370. https://doi.org/10.1038/s41576-019-0108-4

Bordenstein, S. R., & Theis, K. R. (2015). Host biology in light of the microbiome: ten principles of holobionts and hologenomes. *PLOS Biology*, *13*(8), e1002226. https://doi.org/10.1371/JOURNAL.PBIO.1002226

Brand, M. W., Wannemuehler, M. J., Phillips, G. J., Proctor, A., Overstreet, A. M.,

Jergens, A. E., Orcutt, R. P., & Fox, J. G. (2015). The altered Schaedler flora: continued applications of a defined murine microbial community. *ILAR Journal*, *56*(2), 169. https://doi.org/10.1093/ILAR/ILV012

Bray, J. R., & Curtis, J. T. (1957). An ordination of the upland forest communities of southern Wisconsin. *Ecological Monographs*, *27*(4), 325–349. https://doi.org/10.2307/1942268

Buetas, E., Jordán-López, M., López-Roldán, A., D'Auria, G., Martínez-Priego, L., De Marco, G., Carda-Diéguez, M., & Mira, A. (2024). Full-length 16S rRNA gene sequencing by PacBio improves taxonomic resolution in human microbiome samples. *BMC Genomics*, *25*(1), 1–13. https://doi.org/10.1186/S12864-024-10213-5

Burrows, M., & Wheeler, D. J. (1994). *A block-sorting sossless data compressionaAlgorithm*.

Byrd, A. L., Belkaid, Y., & Segre, J. A. (2018). The human skin microbiome. *Nature Reviews Microbiology 2018 16:3*, *16*(3), 143–155. https://doi.org/10.1038/nrmicro.2017.157

Cacho, A., Smirnova, E., Huzurbazar, S., & Cui, X. (2016). A comparison of base-calling algorithms for Illumina sequencing technology. *Briefings in Bioinformatics*, *17*(5), 786–795. https://doi.org/10.1093/BIB/BBV088

Chee, W. J. Y., Chew, S. Y., & Than, L. T. L. (2020). Vaginal microbiota and the potential of Lactobacillus derivatives in maintaining vaginal health. *Microbial Cell Factories 2020 19:1*, *19*(1), 1–24. https://doi.org/10.1186/S12934-020-01464-4

Chinchor, N. (1992). *MUC-4 Evaluation Metrics*. https://aclanthology.org/M92-1002

Clemente, J. C., Ursell, L. K., Parfrey, L. W., & Knight, R. (2012). The impact of the gut microbiota on human health: An integrative view. In *Cell*. https://doi.org/10.1016/j.cell.2012.01.035

Cockell, C. S. (2021). Are microorganisms everywhere they can be? *Environmental Microbiology*, *23*(11), 6355–6363. https://doi.org/10.1111/1462-2920.15825

Costello, E. K., Lauber, C. L., Hamady, M., Fierer, N., Gordon, J. I., & Knight, R. (2009). Bacterial community variation in human body habitats across space and time. *Science*, *326*(5960), 1694–1697. https://doi.org/10.1126/SCIENCE.1177486

Coyte, K. Z., Schluter, J., & Foster, K. R. (2015). The ecology of the microbiome: Networks, competition, and stability. *Science*, *350*(6261), 663–666. https://doi.org/10.1126/SCIENCE.AAD2602

Curry, K. D., Wang, Q., Nute, M. G., Tyshaieva, A., Reeves, E., Soriano, S., Wu, Q., Graeber, E., Finzer, P., Mendling, W., Savidge, T., Villapol, S., Dilthey, A., & Treangen, T. J. (2022). Emu: Species-level microbial community profiling for full-length Nanopore 16S reads. *Nature Methods*, *19*(7), 845. https://doi.org/10.1038/S41592-022-01520-4

Daily, J. (2016). Parasail: SIMD C library for global, semi-global, and local pairwise sequence alignments. *BMC Bioinformatics*, *17*(1), 1–11. https://doi.org/10.1186/S12859-016-0930-Z

Delcher, A. L., Phillippy, A., Carlton, J., & Salzberg, S. L. (2002). Fast algorithms for large-scale genome alignment and comparison. *Nucleic Acids Research*, *30*(11), 2478. https://doi.org/10.1093/NAR/30.11.2478

Dey, P., & Ray Chaudhuri, S. (2023). The opportunistic nature of gut commensal microbiota. *Critical Reviews in Microbiology*, *49*(6), 739–763. https://doi.org/10.1080/1040841X.2022.2133987

Ding, C., Yu, Y., & Zhou, Q. (2021). Bacterial Vaginosis: Effects on reproduction and its therapeutics. *Journal of Gynecology Obstetrics and Human Reproduction*, *50*(9), 102174. https://doi.org/10.1016/J.JOGOH.2021.102174

Dodiya, H. B., Lutz, H. L., Weigle, I. Q., Patel, P., Michalkiewicz, J., Roman-Santiago, C. J., Zhang, C. M., Liang, Y., Srinath, A., Zhang, X., Xia, J., Olszewski, M., Zhang, X., Schipma, M. J., Chang, E. B., Tanzi, R. E., Gilbert, J. A., & Sisodia, S. S. (2021). Gut microbiota-driven brain Aβ amyloidosis in mice requires microglia. *Journal of Experimental Medicine*, *219*(1). https://doi.org/10.1084/JEM.20200895/212889

Doolittle, W. F. (2012). Population Genomics: How bacterial species form and why they don't exist. *Current Biology*, *22*(11), R451–R453. https://doi.org/10.1016/J.CUB.2012.04.034

Dragone, N. B., Diaz, M. A., Hogg, I. D., Lyons, W. B., Jackson, W. A., Wall, D. H., Adams, B. J., & Fierer, N. (2021). Exploring the boundaries of microbial habitability in soil. *Journal of Geophysical Research: Biogeosciences*, *126*(6), e2020JG006052. https://doi.org/10.1029/2020JG006052

Dréno, B., Pécastaings, S., Corvec, S., Veraldi, S., Khammari, A., & Roques, C. (2018). *Cutibacterium acnes* (*Propionibacterium acne*s) and acne vulgaris: a brief look at the latest updates. *Journal of the European Academy of Dermatology and Venereology*, *32*, 5–14. https://doi.org/10.1111/JDV.15043

Dunlap, C. A., Bowman, M. J., & Zeigler, D. R. (2020). Promotion of *Bacillus subtilis* subsp. *inaquosorum*, *Bacillus subtilis* subsp. *spizizenii* and *Bacillus subtilis* subsp. *stercoris* to species status. *Antonie van Leeuwenhoek, International Journal of General and Molecular Microbiology*, *113*(1), 1–12. https://doi.org/10.1007/S10482-019-01354-9

Durazzi, F., Sala, C., Castellani, G., Manfreda, G., Remondini, D., & De Cesare, A. (2021). Comparison between 16S rRNA and shotgun sequencing data for the taxonomic characterization of the gut microbiota. *Scientific Reports*, *11*(1), 3030. https://doi.org/10.1038/S41598-021-82726-Y

Ewing, B., & Green, P. (1998). Base-calling of automated sequencer traces using phred. II. Error probabilities. *Genome Research*, *8*(3), 186–194. https://doi.org/10.1101/gr.8.3.186

Ewing, B., Hillier, L. D., Wendl, M. C., & Green, P. (1998). Base- calling of automated sequencer traces using Phred. I. accuracy assessment. *Genome Research*, *8*(3), 175–185. https://doi.org/10.1101/GR.8.3.175

Ferragina, P., & Manzini, G. (2000). Opportunistic data structures with applications. *Annual Symposium on Foundations of Computer Science - Proceedings*, 390–398. https://doi.org/10.1109/sfcs.2000.892127

Flint, H. J., Scott, K. P., Duncan, S. H., Louis, P., & Forano, E. (2012). Microbial degradation of complex carbohydrates in the gut. *Gut Microbes*, *3*(4), 289. https://doi.org/10.4161/GMIC.19897

Florescu, S., & Ahmed, N. (2022). *GPU acceleration of BWA-MEM DNA sequence alignment*. https://repository.tudelft.nl/record/uuid:4dd99ea2-6955-4e39-8e40-4198da4667f4

Franzosa, E. A., Huang, K., Meadow, J. F., Gevers, D., Lemon, K. P., Bohannan, B. J. M., & Huttenhower, C. (2015). Identifying personal microbiomes using metagenomic codes. *Proceedings of the National Academy of Sciences of the United States of America*, *112*(22), E2930–E2938. https://doi.org/10.1073/PNAS.1423854112

Freitas, T. A. K., Li, P. E., Scholz, M. B., & Chain, P. S. G. (2015). Accurate read-based metagenome characterization using a hierarchicalsuite of unique signatures. *Nucleic Acids Research*, *43*(10), e69. https://doi.org/10.1093/NAR/GKV180

Fritz, A., Hofmann, P., Majda, S., Dahms, E., Dröge, J., Fiedler, J., Lesker, T. R., Belmann, P., Demaere, M. Z., Darling, A. E., Sczyrba, A., Bremges, A., & McHardy, A. C. (2019). CAMISIM: Simulating metagenomes and microbial communities. *Microbiome*, *7*(1), 1–12. https://doi.org/10.1186/S40168-019-0633-6

Gaci, N., Borrel, G., Tottey, W., O'Toole, P. W., & Brugère, J. F. (2014). Archaea and the human gut: New beginning of an old story. *World Journal of Gastroenterology : WJG*, *20*(43), 16062. https://doi.org/10.3748/WJG.V20.I43.16062

Galloway-Peña, J., & Hanson, B. (2020). Tools for Analysis of the Microbiome. *Digestive Diseases and Sciences*, *65*(3), 674. https://doi.org/10.1007/S10620-020-06091-Y

Garrido-Sanz, L., Senar, M. À., & Piñol, J. (2022). Drastic reduction of false positive species in samples of insects by intersecting the default output of two popular metagenomic classifiers. *PLOS ONE*, *17*(10). https://doi.org/10.1371/JOURNAL.PONE.0275790

Gershon, M. D. (1999). The enteric nervous system: A second brain. *Hospital Practice*, *34*(7), 31–52. https://doi.org/10.3810/HP.1999.07.153

Giuliano, C., Patel, C. R., & Kale-Pradhan, P. B. (2019). A guide to bacterial culture identification and results interpretation. *Pharmacy and Therapeutics*, *44*(4), 192. https://pmc.ncbi.nlm.nih.gov/articles/PMC6428495/

Gu, W., Miller, S., & Chiu, C. Y. (2019). Clinical metagenomic Next-Generation

Sequencing for pathogen detection. *Annual Review of Pathology: Mechanisms of Disease*. https://doi.org/10.1146/annurev-pathmechdis-012418-012751

Handelsman, J., Rondon, M. R., Brady, S. F., Clardy, J., & Goodman, R. M. (1998). Molecular biological access to the chemistry of unknown soil microbes: a new frontier for natural products. *Chemistry & Biology*, *5*(10). https://doi.org/10.1016/S1074-5521(98)90108-9

Hou, K., Wu, Z. X., Chen, X. Y., Wang, J. Q., Zhang, D., Xiao, C., Zhu, D., Koya, J. B., Wei, L., Li, J., & Chen, Z. S. (2022). Microbiota in health and diseases. *Signal Transduction and Targeted Therapy 2022 7:1*, *7*(1), 1–28. https://doi.org/10.1038/s41392-022-00974-4

Hu, Z., Cheng, L., & Wang, H. (2015). The Illumina-Solexa sequencing protocol for bacterial genomes. *Methods in Molecular Biology (Clifton, N.J.)*, *1231*, 91–97. https://doi.org/10.1007/978-1-4939-1720-4_6

Huang, Y., Wu, J., Zhang, H., Li, Y., Wen, L., Tan, X., Cheng, K., Liu, Y., Pu, J., Liu, L., Wang, H., Li, W., Perry, S. W., Wong, M. L., Licinio, J., Zheng, P., & Xie, P. (2023). The gut microbiome modulates the transformation of microglial subtypes. *Molecular Psychiatry 2023 28:4*, *28*(4), 1611–1621. https://doi.org/10.1038/s41380-023-02017-y

Huttenhower, C., Gevers, D., Knight, R., Abubucker, S., Badger, J. H., Chinwalla, A. T., Creasy, H. H., Earl, A. M., Fitzgerald, M. G., Fulton, R. S., Giglio, M. G., Hallsworth-Pepin, K., Lobos, E. A., Madupu, R., Magrini, V., Martin, J. C., Mitreva, M., Muzny, D. M., Sodergren, E. J., White, O. (2012a). Structure, function and diversity of the healthy human microbiome. *Nature 2012 486:7402*, *486*(7402), 207–214. https://doi.org/10.1038/nature11234

Jaccard, P. (1912). T he distribution of the flora in the Alpine zone. *New Phytologist*, *11*(2), 37–50. https://doi.org/10.1111/J.1469-8137.1912.TB05611.X

Janda, J. M., & Abbott, S. L. (2007). 16S rRNA gene sequencing for bacterial identification in the diagnostic laboratory: Pluses, perils, and pitfalls. *Journal of Clinical Microbiology*, *45*(9), 2761. https://doi.org/10.1128/JCM.01228-07

Kaiju web server - Submit Job. (n.d.). Retrieved November 2, 2022, from https://kaiju.binf.ku.dk/server

Kan, C. M., Tsang, H. F., Pei, X. M., Ng, S. S. M., Yim, A. K. Y., Yu, A. C. S., & Wong, S. C. C. (2024). Enhancing clinical utility: Utilization of international standards and guidelines for metagenomic sequencing in infectious disease diagnosis. *International Journal of Molecular Sciences 2024, Vol. 25, Page 3333*, *25*(6), 3333. https://doi.org/10.3390/IJMS25063333

Kim, J., Kim, B. E., Ahn, K., & Leung, D. Y. M. (2019). Interactions between atopic dermatitis and *Staphylococcus aureus* infection: Clinical implications. *Allergy, Asthma & Immunology Research*, *11*(5), 593. https://doi.org/10.4168/AAIR.2019.11.5.593

Konstantinidis, K. T., Ramette, A., & Tiedje, J. M. (2006). The bacterial species definition in the genomic era. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *361*(1475), 1929. https://doi.org/10.1098/RSTB.2006.1920

Kralik, P., & Ricchi, M. (2017). A basic guide to real time PCR in microbial diagnostics: Definitions, parameters, and everything. *Frontiers in Microbiology*, *8*(FEB), 239909. https://doi.org/10.3389/FMICB.2017.00108

Krynicka, P., Koulaouzidis, G., Skonieczna-Żydecka, K., Marlicz, W., & Koulaouzidis, A. (2025). Application of Raman spectroscopy in non-invasive analysis of the gut microbiota and its impact on gastrointestinal health. *Diagnostics 2025, Vol. 15, Page 292*, *15*(3), 292. https://doi.org/10.3390/DIAGNOSTICS15030292

Kuleshov, V., Jiang, C., Zhou, W., Jahanbani, F., Batzoglou, S., & Snyder, M. (2016). Synthetic long read sequencing reveals the composition and intraspecies diversity of the human microbiome. *Nature Biotechnology*, *34*(1), 64. https://doi.org/10.1038/NBT.3416

Kumar, K. R., Cowley, M. J., & Davis, R. L. (2019). Next-Generation Sequencing and emerging technologies. *Seminars in Thrombosis and Hemostasis*, *45*(7), 661–673. https://doi.org/10.1055/S-0039-1688446/ID/JR02634-15/BIB

Kumar, P. S. (2013). Oral microbiota and systemic disease. *Anaerobe*, *24*, 90–93. https://doi.org/10.1016/J.ANAEROBE.2013.09.010

Li, G., Liu, Y., Li, D., Liu, B., Li, J., Hu, Y., & Wang, Y. (2021). Fast and accurate classification of meta-genomics long reads with deSAMBA. *Frontiers in Cell and Developmental Biology*, *9*, 643645. https://doi.org/10.3389/FCELL.2021.643645

Li, H. (2013). *Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM*. https://arxiv.org/abs/1303.3997v2

Li, H., & Durbin, R. (2009). Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinformatics*, *25*(14), 1754–1760. https://doi.org/10.1093/BIOINFORMATICS/BTP324

Lin, J.-N., Lai, C.-H., Lin, S.-Y., Lee, C.-C., Lee, N.-Y., Liu, P.-Y., Yang, C.-H., & Huang, Y.-H. (2022). Effect of of intragenomic sequence heterogeneity among multiple 16S rRNA genes on species identification of *Elizabethkingia*. *Microbiology Spectrum*, *10*(5). https://doi.org/10.1128/SPECTRUM.01338-22

Liu, C., Du, M. X., Abuduaini, R., Yu, H. Y., Li, D. H., Wang, Y. J., Zhou, N., Jiang, M. Z., Niu, P. X., Han, S. S., Chen, H. H., Shi, W. Y., Wu, L., Xin, Y. H., Ma, J., Zhou, Y., Jiang, C. Y., Liu, H. W., & Liu, S. J. (2021). Enlightening the taxonomy darkness of human gut microbiomes with a cultured biobank. *Microbiome*, *9*(1), 1–29. https://doi.org/10.1186/S40168-021-01064-3

Loh, J. S., Mak, W. Q., Tan, L. K. S., Ng, C. X., Chan, H. H., Yeow, S. H., Foo, J. B., Ong, Y. S., How, C. W., & Khaw, K. Y. (2024). Microbiota–gut–brain axis and its therapeutic applications in neurodegenerative diseases. *Signal Transduction and Targeted Therapy 2024 9:1*, *9*(1), 1–53. https://doi.org/10.1038/s41392-024-01743-

1

Lu, J., Breitwieser, F. P., Thielen, P., & Salzberg, S. L. (2017). Bracken: Estimating species abundance in metagenomics data. *PeerJ Computer Science*, *2017*(1), e104. https://doi.org/10.7717/PEERJ-CS.104

Maiden, M. C. J. (2006). Multilocus sequence typing of bacteria. *Annual Review of Microbiology*, *60*, 561–588. https://doi.org/10.1146/ANNUREV.MICRO.59.030804.121325

Mann, E. R., Lam, Y. K., & Uhlig, H. H. (2024). Short-chain fatty acids: linking diet, the microbiome and immunity. *Nature Reviews Immunology 2024 24:8*, *24*(8), 577–595. https://doi.org/10.1038/s41577-024-01014-8

Marizzoni, M., Gurry, T., Provasi, S., Greub, G., Lopizzo, N., Ribaldi, F., Festari, C., Mazzelli, M., Mombelli, E., Salvatore, M., Mirabelli, P., Franzese, M., Soricelli, A., Frisoni, G. B., & Cattaneo, A. (2020). C omparison of bioinformatics pipelines and operating systems for the analyses of 16S rRNA gene amplicon sequences in Human Fecal Samples. *Frontiers in Microbiology*, *11*, 506925. https://doi.org/10.3389/FMICB.2020.01262

Martin, M. (2011). Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet.Journal*, *17*(1), 10–12. https://doi.org/10.14806/EJ.17.1.200

Mei, Z., & Li, D. (2022). The role of probiotics in vaginal health. *Frontiers in Cellular and Infection Microbiology*, *12*, 963868. https://doi.org/10.3389/FCIMB.2022.963868

Menzel, P., Ng, K. L., & Krogh, A. (2016). Fast and sensitive taxonomic classification for metagenomics with Kaiju. *Nature Communications 2016 7:1*, *7*(1), 1–9. https://doi.org/10.1038/ncomms11257

Meyer, F., Fritz, A., Deng, Z. L., Koslicki, D., Lesker, T. R., Gurevich, A., Robertson, G., Alser, M., Antipov, D., Beghini, F., Bertrand, D., Brito, J. J., Brown, C. T., Buchmann, J., Buluç, A., Chen, B., Chikhi, R., Clausen, P. T. L. C., Cristian, A., … McHardy, A. C. (2022). Critical Assessment of Metagenome Interpretation: the second round of challenges. *Nature Methods 2022 19:4*, *19*(4), 429–440. https://doi.org/10.1038/s41592-022-01431-4

Michán-Doña, A., Vázquez-Borrego, M. C., & Michán, C. (2024). Are there any completely sterile organs or tissues in the human body? Is there any sacred place? *Microbial Biotechnology*, *17*(3), e14442. https://doi.org/10.1111/1751-7915.14442

Mirsepasi-Lauridsen, H. C., Vallance, B. A., Krogfelt, K. A., & Petersen, A. M. (2019). *Escherichia coli* pathobionts associated with inflammatory bowel disease. *Clinical Microbiology Reviews*, *32*(2). https://doi.org/10.1128/CMR.00060-18

Morgan, J. L., Darling, A. E., & Eisen, J. A. (2010). Metagenomic sequencing of an in vitro-simulated microbial community. *PLOS ONE*, *5*(4), e10209. https://doi.org/10.1371/JOURNAL.PONE.0010209

Morowitz, M. J., Carlisle, E. M., & Alverdy, J. C. (2011). Contributions of intestinal

bacteria to nutrition and metabolism in the critically ill. *The Surgical Clinics of North America*, *91*(4), 771. https://doi.org/10.1016/J.SUC.2011.05.001

Nicholls, S. M., Quick, J. C., Tang, S., & Loman, N. J. (2019). Ultra-deep, long-read nanopore sequencing of mock microbial community standards. *GigaScience*, *8*(5). https://doi.org/10.1093/GIGASCIENCE/GIZ043

Ounit, R., & Lonardi, S. (2016). Higher classification sensitivity of short metagenomic reads with CLARK-S. *Bioinformatics*, *32*(24), 3823–3825. https://doi.org/10.1093/BIOINFORMATICS/BTW542

Paone, P., & Cani, P. D. (2020). Mucus barrier, mucins and gut microbiota: the expected slimy partners? *Gut*, *69*(12), 2232. https://doi.org/10.1136/GUTJNL-2020-322260

Parker, C. T., Tindall, B. J., & Garrity, G. M. (2019). International code of nomenclature of Prokaryotes. *International Journal of Systematic and Evolutionary Microbiology*, *69*(1), S1. https://doi.org/10.1099/IJSEM.0.000778

Paul, B., Kavia Raj, K., Murali, T. S., & Satyamoorthy, K. (2020). Species-specific genomic sequences for classification of bacteria. *Computers in Biology and Medicine*, *123*, 103874. https://doi.org/10.1016/J.COMPBIOMED.2020.103874

Payler, S. J., Biddle, J. F., Lollar, B. S., Fox-Powell, M. G., Edwards, T., Ngwenya, B. T., Paling, S. M., & Cockell, C. S. (2019). An ionic limit to life in the deep subsurface. *Frontiers in Microbiology*, *10*(MAR). https://doi.org/10.3389/FMICB.2019.00426

Pereira-Marques, J., Hout, A., Ferreira, R. M., Weber, M., Pinto-Ribeiro, I., Van Doorn, L. J., Knetsch, C. W., & Figueiredo, C. (2019). Impact of host DNA and sequencing depth on the taxonomic resolution of whole metagenome sequencing for microbiome analysis. *Frontiers in Microbiology*, *10*(JUN), 454372. https://doi.org/10.3389/FMICB.2019.01277

Powers, E. M., & Latt, T. G. (1977). Simplified 48-hour IMVic test: an agar plate method. *Applied and Environmental Microbiology*, *34*(3), 274–279. https://doi.org/10.1128/AEM.34.3.274-279.1977

Raghu, A. K., Palanikumar, I., & Raman, K. (2024). Designing function-specific minimal microbiomes from large microbial communities. *Npj Systems Biology and Applications 2024 10:1*, *10*(1), 1–9. https://doi.org/10.1038/s41540-024-00373-1

Reese, A. T., Savage, A., Youngsteadt, E., Mcguire, K. L., Koling, A., Watkins, O., Frank, S. D., & Dunn, R. R. (2016). Urban stress is associated with variation in microbial species composition-but not richness-in Manhattan. *The ISME Journal*, *10*(3), 751–760. https://doi.org/10.1038/ISMEJ.2015.152

Robin Warren, J., & Marshall, B. (1983). Unidentified curved bacilli on gastric epithelium in active chronic gastritis. *The Lancet*, *321*(8336), 1273–1275. https://doi.org/10.1016/S0140-6736(83)92719-8

Ruiz-Calderon, J. F., Cavallin, H., Song, S. J., Novoselac, A., Pericchi, L. R.,

Hernandez, J. N., Rios, R., Branch, O. H., Pereira, H., Paulino, L. C., Blaser, M. J., Knight, R., & Dominguez-Bello, M. G. (2016). Microbiology: Walls talk: Microbial biogeography of homes spanning urbanization. *Science Advances*, *2*(2). https://doi.org/10.1126/SCIADV.1501061

Sachdev, A. H., & Pimentel, M. (2013). Gastrointestinal bacterial overgrowth: pathogenesis and clinical significance. *Therapeutic Advances in Chronic Disease*, *4*(5), 223. https://doi.org/10.1177/2040622313496126

Salido, R. A., Zhao, H. N., McDonald, D., Mannochio-Russo, H., Zuffa, S., Oles, R. E., Aron, A. T., El Abiead, Y., Farmer, S., González, A., Martino, C., Mohanty, I., Parker, C. W., Patel, L., Portal Gomes, P. W., Schmid, R., Schwartz, T., Zhu, J., Barratt, M. R., … Knight, R. (2025). The International Space Station has a unique and extreme microbial and chemical environment driven by use patterns. *Cell*, *188*(7), 2022-2041.e23. https://doi.org/10.1016/J.CELL.2025.01.039

Sampson, T. R., Debelius, J. W., Thron, T., Janssen, S., Shastri, G. G., Ilhan, Z. E., Challis, C., Schretter, C. E., Rocha, S., Gradinaru, V., Chesselet, M. F., Keshavarzian, A., Shannon, K. M., Krajmalnik-Brown, R., Wittung-Stafshede, P., Knight, R., & Mazmanian, S. K. (2016). Gut microbiota regulate motor deficits and neuroinflammation in a model of Parkinson's disease. *Cell*, *167*(6), 1469-1480.e12. https://doi.org/10.1016/J.CELL.2016.11.018

Sczyrba, A., Hofmann, P., Belmann, P., Koslicki, D., Janssen, S., Dröge, J., Gregor, I., Majda, S., Fiedler, J., Dahms, E., Bremges, A., Fritz, A., Garrido-Oter, R., Jørgensen, T. S., Shapiro, N., Blood, P. D., Gurevich, A., Bai, Y., Turaev, D., … McHardy, A. C. (2017). Critical Assessment of Metagenome Interpretation—a benchmark of metagenomics software. *Nature Methods 2017 14:11*, *14*(11), 1063–1071. https://doi.org/10.1038/nmeth.4458

Sender, R., Fuchs, S., & Milo, R. (2016). Revised estimates for the number of human and bacteria cells in the body. *PLoS Biology*. https://doi.org/10.1371/journal.pbio.1002533

Setubal, J. C. (2021). Metagenome-assembled genomes: concepts, analogies, and challenges. *Biophysical Reviews*, *13*(6), 905. https://doi.org/10.1007/S12551-021-00865-Y

Singhal, N., Kumar, M., Kanaujia, P. K., & Virdi, J. S. (2015). MALDI-TOF mass spectrometry: an emerging technology for microbial identification and diagnosis. *Frontiers in Microbiology*, *6*(AUG), 791. https://doi.org/10.3389/FMICB.2015.00791

Smith, T. F., & Waterman, M. S. (1981). Identification of common molecular subsequences. *Journal of Molecular Biology*, *147*(1), 195–197. https://doi.org/10.1016/0022-2836(81)90087-5

Steensels, J., Gallone, B., Voordeckers, K., & Verstrepen, K. J. (2019). Domestication of industrial microbes. *Current Biology*, *29*(10), R381–R393. https://doi.org/10.1016/J.CUB.2019.04.025

Strandwitz, P., Kim, K. H., Terekhova, D., Liu, J. K., Sharma, A., Levering, J.,

McDonald, D., Dietrich, D., Ramadhar, T. R., Lekbua, A., Mroue, N., Liston, C., Stewart, E. J., Dubin, M. J., Zengler, K., Knight, R., Gilbert, J. A., Clardy, J., & Lewis, K. (2018). GABA-modulating bacteria of the human gut microbiota. *Nature Microbiology 2018 4:3*, *4*(3), 396–403. https://doi.org/10.1038/s41564-018-0307-3

Sun, Z., Huang, S., Zhang, M., Zhu, Q., Haiminen, N., Carrieri, A. P., Vázquez-Baeza, Y., Parida, L., Kim, H. C., Knight, R., & Liu, Y. Y. (2021). Challenges in benchmarking metagenomic profilers. *Nature Methods 2021 18:6*, *18*(6), 618–626. https://doi.org/10.1038/s41592-021-01141-3

Takiishi, T., Fenero, C. I. M., & Câmara, N. O. S. (2017). Intestinal barrier and gut microbiota: Shaping our immune responses throughout life. *Tissue Barriers*, *5*(4), e1373208. https://doi.org/10.1080/21688370.2017.1373208

Tarracchini, C., Lugli, G. A., Mancabelli, L., van Sinderen, D., Turroni, F., Ventura, M., & Milani, C. (2024). Exploring the vitamin biosynthesis landscape of the human gut microbiota. *MSystems*, *9*(10), e00929-24. https://doi.org/10.1128/msystems.00929-24

Tsunoda, S. M., Gonzales, C., Jarmusch, A. K., Momper, J. D., & Ma, J. D. (2021). Contribution of the gut microbiome to drug disposition, pharmacokinetic and pharmacodynamic variability. *Clinical Pharmacokinetics 2021 60:8*, *60*(8), 971–984. https://doi.org/10.1007/S40262-021-01032-Y

Turnbaugh, P. J., Ley, R. E., Hamady, M., Fraser-Liggett, C. M., Knight, R., & Gordon, J. I. (2007). The Human Microbiome Project. *Nature 2007 449:7164*, *449*(7164), 804–810. https://doi.org/10.1038/nature06244

Wang, Y., Tong, Q., Ma, S. R., Zhao, Z. X., Pan, L. Bin, Cong, L., Han, P., Peng, R., Yu, H., Lin, Y., Gao, T. Le, Shou, J. W., Li, X. Y., Zhang, X. F., Zhang, Z. W., Fu, J., Wen, B. Y., Yu, J. B., Cao, X., & Jiang, J. D. (2021). Oral berberine improves brain dopa/dopamine levels to ameliorate Parkinson's disease by regulating gut microbiota. *Signal Transduction and Targeted Therapy*, *6*(1). https://doi.org/10.1038/S41392-020-00456-5

Wood, D. E., Lu, J., & Langmead, B. (2019). Improved metagenomic analysis with Kraken 2. *Genome Biology*, *20*(1), 1–13. https://doi.org/10.1186/S13059-019-1891-0

Wood, D. E., & Salzberg, S. L. (2014). Kraken: Ultrafast metagenomic sequence classification using exact alignments. *Genome Biology*, *15*(3), 1–12. https://doi.org/10.1186/GB-2014-15-3-R46

Yano, J. M., Yu, K., Donaldson, G. P., Shastri, G. G., Ann, P., Ma, L., Nagler, C. R., Ismagilov, R. F., Mazmanian, S. K., & Hsiao, E. Y. (2015). Indigenous bacteria from the gut microbiota regulate host serotonin biosynthesis. *Cell*, *161*(2), 264–276. https://doi.org/10.1016/J.CELL.2015.02.047

Ye, S. H., Siddle, K. J., Park, D. J., & Sabeti, P. C. (2019). Benchmarking metagenomics tools for taxonomic xlassification. *Cell*, *178*(4), 779–794. https://doi.org/10.1016/J.CELL.2019.07.010

Zhang, H., Wang, X., Chen, A., Li, S., Tao, R., Chen, K., Huang, P., Li, L., Huang, J., Li, C., & Zhang, S. (2024). Comparison of the full-length sequence and sub-regions of 16S rRNA gene for skin microbiome profiling. *MSystems*, *9*(7). https://doi.org/10.1128/MSYSTEMS.00399-24

Zheng, J., Wittouck, S., Salvetti, E., Franz, C. M. A. P., Harris, H. M. B., Mattarelli, P., O'toole, P. W., Pot, B., Vandamme, P., Walter, J., Watanabe, K., Wuyts, S., Felis, G. E., Gänzle, M. G., & Lebeer, S. (2020). A taxonomic note on the genus *Lactobacillus*: Description of 23 novel genera, emended description of the genus *Lactobacillus beijerinck* 1901, and union of Lactobacillaceae and Leuconostocaceae. *International Journal of Systematic and Evolutionary Microbiology*, *70*(4), 2782–2858. https://doi.org/10.1099/IJSEM.0.004107

Zilber-Rosenberg, I., & Rosenberg, E. (2008). Role of microorganisms in the evolution of animals and plants: the hologenome theory of evolution. *FEMS Microbiology Reviews*, *32*(5), 723–735. https://doi.org/10.1111/J.1574-6976.2008.00123.X

XII. List of Publications

MTMT ID: 10074085

Publications used in this thesis

1. *Advancing metagenomic classification with NABAS+: a novel alignment-based* approach
   Bertalan Takács, Gábor Jaksa, Erda Qorri, Zoltán Gyuris, Lajos Pintér and Lajos Haracska

   NAR Genomics and Bioinformatics, Volume 7, Issue 3, September 2025

   Impact Factor: **5.0**

2. *Shotgun Analysis of Gut Microbiota with Body Composition and Lipid Characteristics in Crohn's Disease*
   Péter Bacsur, Tamás Resál, Bernadett Farkas, Boldizsár Jójárt, Zoltán Gyuris, Gábor Jaksa, Lajos Pintér, Bertalan Takács, Sára Pál, Attila Gácser, Kata Judit Szántó, Mariann Rutka, Renáta Bor, Anna Fábián, Klaudia Farkas, József Maléth, Zoltán Szepes, Tamás Molnár and Anita Bálint

   Biomedicines 2024, *12*(9), 2100;

   Impact factor: **3.9**

Other publications

1. *Prolonged activity of the transposase helper may raise safety concerns during DNA transposon-based gene therapy*
   Gergely Imre, Bertalan Takács, Erik Czipa, Andrea Bakné Drubi, Gábor Jaksa, Dóra Latinovics, Andrea Nagy, Réka Karkas, Liza Hudoba, Bálint Márk Vásárhelyi, Gabriella Pankotai-Bodó, András Blastyák, Zoltán Hegedűs, Péter Germán, Balázs Bálint, Khaldoon Sadiq Ahmed Abdullah, Anna Georgina Kopasz, Anita Kovács, László G Nagy, Farkas Sükösd, Lajos Pintér, Thomas Rülicke, Endre Barta, István Nagy, Lajos Haracska and Lajos Mátés

Molecular Therapy Methods & Clinical Development, Volume 29, 145 – 159

Impact Factor: 4.7

2. *A Comprehensive Evaluation of the Performance of Prediction Algorithms on Clinically Relevant Missense Variants*
   Erda Qorri, Bertalan Takács, Alexandra Gráf, Márton Zsolt Enyedi, Lajos Pintér, Ernő Kiss and Lajos Haracska

   International Journal of Molecular Sciences. 2022; 23(14):7946

   Impact Factor: 4.9

3. *Distinct Gut Microbiota Profiles in Unruptured and Ruptured Intracranial Aneurysms: Focus on Butyrate-Producing Bacteria*
   Péter Csécsei, Bertalan Takács, Lídia Pasitka, Réka Várnai, Zoltán Péterfi, Brigitta Orbán, Máté Czabajszki, Csaba Oláh and Attila Schwarcz

   Journal of Clinical Medicine. 2025; 14(10):3488

   Impact Factor: 3.0

Conference talks

1. *Comparative Analysis of the Microbial Composition of Different Tumor Tissues*
   Talk at the Microbiota and Cancer Immunity Conference, Taipei, Taiwan, 2024
2. *The devil is in the details: Eliminating false metagenomic classification with a novel algorithm*
   Talk at the 61st Annual Meeting of the Hungarian Society of Laboratory Diagnostics, Budapest, Hungary, 2023
3. *Understanding the effects of COVID-19 on the microbiome using bioinformatics and machine learning*
   Talk at the 5th National Conference of Young Biotechnologists, Gödöllő, Hungary, 2022

Posters

1. *The devil in the details: Eliminating false metagenomic classification with a novel algorithm*
   Poster at the Hungarian Molecular Life Sciences 2023 Conference, Eger, Hungary, 2023
2. *Understanding the effects of COVID-19 on the microbiome using bioinformatics and machine learning*
   Poster at the 25th Spring Wind Conference,Pécs, Hungary, 2022
3. *Impacts of COVID-19 on the Microbiome: A Bioinformatics and Machine Learning Study*
   Poster at the 21st European Conference on Computational Biology, Sitges, Spain, 2022

XIII. Magyar nyelvű összefoglaló

Az emberi mikrobiom az emberi szervezettel együtt élő mikroorganizmusok összességét jelenti. Ezen organizmusok száma magasabb, mint a testet felépítő sejtek teljes száma és jelentős hatást gyakorolnak a szervezetre: pozitív és negatív módon is képesek befolyásolni annak egészségét. A bél, szájüreg, hüvely és a bőr mind saját mikrobiális közösséggel rendelkeznek, amelynek összetétele, diverzitása szoros összefüggésben áll különféle élettani és kóros folyamatokkal. A mikrobiom vizsgálata ezért az elmúlt két évtizedben az orvosbiológiai kutatások központi területévé vált. E fejlődést elősegítette a nagy áteresztőképességű molekuláris biológiai módszerek megjelenése és elterjedése is. Ilyen módszerek az ezen munkában tárgyalt újgenerációs szekvenálási módszerek is, mint a 16S- és "shotgun"-szekvenálás, amelyek lehetővé teszik egy mintában jelen levő teljes mikrobiális közösség egyidejű meghatározását. A nagy mennyiségű adat ugyanakkor az elemzésére alkalmas bioinformatikai technológiák fejlődését is szükségessé tette. Az elmúlt években számos olyan szoftver született, amelyeket a mikrobiom összetételének azonosítására fejlesztettek. Ezek pontossága, referencia-adatbázisa, számítási igénye jelentős eltérést mutat egymástól. Az azonosítás pontossága különösen fontos akkor, ha a metagenomikai módszereket klinikai környezetben, például patogén fajok azonosítására kívánjuk alkalmazni. Ehhez olyan algoritmusokra van szükség, amelyek a lehető legnagyobb pontossággal képesek meghatározni a minták összetételét, minimalizálják a hamis pozitív találatok számát, emellett gyorsak és skálázhatóak is.

Doktori munkám során az alábbi célokat tűztem ki:

- Egy új metagenomikai azonosító algoritmus fejlesztése
- Az algoritmus teljesítményének teszteléséhez szükséges adatsorok előállítása és összegyűjtése
- Összehasonlító tesztelés
- Az algoritmus hatékonyságának bemutatása valós klinikai adatokon

Kilenc mikrobiális azonosításra alkalmas szoftvert választottunk ki az összehasonlításhoz, ezek a következők voltak: Bracken, Centrifuge, CLARK, DIAMOND, GOTTCHA, Kaiju, Kaiju-ws, Kraken2, MetaPhlAn3. A válogatás során célunk az volt, hogy olyan szoftvereket válasszunk, amelyek népszerűek a tudományos közösségben és működésük egymástól jelentősen eltérő algoritmikus megoldásokat implementál. A kilenc szoftvert "shotgun" módszerrel szekvenált emberi széklet mikrobiom mintán teszteltük. Összehasonlítottuk a szoftverek által azonosított baktériumok számát és identitását, illetve azt, hogy egymáshoz mennyire hasonlító eredményt hoztak.

Saját azonosító algoritmusunk, a Novel Alignment-based Biome Analysis Software + fejlesztése során a fő szempontok a pontosság és a hamis pozitívok kiküszöbölése volt. Az algoritmust Java nyelven implementáltuk, kihasználva annak skálázhatóságát és gyorsaságát, programunk így integrálható egy nagyobb szoftverbe. Emellett elkészítettük a szoftver egy önállóan futtatható "stand-alone" verzióját is.

Az tesztelést a következő adatsorokon végeztük:

- Saját fejlesztésű *in silico* adatsorok, összesen 30 minta. A minták különböző emberi szervezetből (pl. bél, szájüreg) és a környezetből (pl. városi park, aszfalt) gyűjtött mikrobiális közösségek összetételét modellezi, összesen 212 baktériumfajból, $5*10^5$-től $1*10^7$ terjedő leolvasási mélységekben.
- Critical Assessment of Metagenome Interpretation II (CAMI) "toy human gastrooral" adatsor, 20 minta. Szabadon hozzáférhető *in silico* minták. Az emberi gasztroorális traktus mikrobiális összetételét modellező, metagenomikai kutatásokban rendszeresen használt adatsor
- Zymo Community Standard I és II (CSI és CSII) Illumina szekvenált adatai, 2 minta. Kereskedelmi forgalomban elérhető standardok, 10, humán mikrobiomban is gyakran előforduló baktérium DNS-ének keverékét tartalmazzák, pontosan ismert arányban.
- Klinikai minták, 20 minta. Shotgun-szekvenált emberi bél-mikrobiom minták, melyek valós kórházi körülmények között voltak gyűjtve és patogén-tartalmukat

laboratóriumi módszerek segítségével is vizsgálták. Tizenegy minta hordozott valamilyen patogént (*Aeromonas*, *Campylobacter* vagy *Salmonella* fajokat),

A kilenc vizsgált program összehasonlítása kimutatta, hogy ugyanazon mikrobiom minták elemzése jelentősen eltérő eredményeket ad, mind a fajok száma, mind azok egyezése tekintetében, szoftvertől függően. Az azonosított fajok identitásában alacsony volt az átfedés.

A NABAS+ teljesítményét több szempontból vizsgáltuk:

- *In silico* adatokon a NABAS+ magas pontossággal határozta meg a fajok számát és identitását. Teljesítménye összemérhető volt tudományos közösségben leggyakrabban használt és tesztünk által is legpontosabbnak ítélt algoritmusokhoz (Kraken2, MetaPhlAn3, GOTTCHA). Az összehasonlítás további részét a NABAS+ mellett ezzel a három szoftverrel végezük.
- A CAMI II adatsorokon a NABAS+ következetesen jó teljesítményt nyújtott, a vezető módszerekkel összehasonlítható eredményekkel. Egy minta (sample19) újragenerálása tovább javította az azonosítás pontosságát.
- A Zymo standardokon a NABAS+ kiemelkedően szerepelt: pontosan detektálta az alacsony abundanciájú fajokat is, és ez volt az egyetlen szoftver, amely nem azonosított hamis pozitívokat. A Bray-Curtis távolság alapján a NABAS+ eredményei álltak legközelebb a valós összetételhez.

Klinikai mintákon a NABAS+ pontosan azonosította a fertőzött mintákban jelen levő patogéneket, és nem jelzett hamis pozitív találatot egészséges mintákban.

Összefoglalva: kutatásom során egy új metagenomikai azonosító szoftvert, a NABAS+-t fejlesztettem ki és validáltam. A NABAS+ teljesítménye a legjobb jelenlegi algoritmusokkal összemérhető, az azonosítás során minimalizálja a hamis pozitívok számát. Ez különösen nagy jelentőséggel bír a klinikai alkalmazásokban, ahol a téves azonosítás súlyos következményekkel járhat.

Eredményeim alapján a NABAS+ nemcsak kutatási, hanem klinikai környezetben is megbízható eszközként alkalmazható, hozzájárulva a metagenomikai vizsgálatok

pontosságához és reprodukálhatóságához. A szoftver fejlesztése és publikusan elérhetővé tétele elősegítheti a tudományos közösség szélesebb körű felhasználását is.

XIV. Summary in English

The human microbiome refers to the community of microorganisms living in association with the human body. Their number exceeds that of human cells, and they exert significant influence on the host, they are able to affect health positively and negatively. Distinct microbial communities inhabit the gut, oral cavity, skin, and vagina, whose composition and diversity are closely associated with a wide range of physiological and pathological processes. Consequently, the study of the microbiome has become a focus of biomedical research over the past two decades. This development has been greatly facilitated by the emergence and widespread use of high-throughput molecular biology techniques. Among these, next-generation sequencing methods such as 16S rRNA and shotgun metagenomics have enabled the simultaneous characterization of entire microbial communities within a sample. The large volume of data generated by these techniques, however, has also raised the need for the advancement of bioinformatic approaches. In recent years, numerous software have been developed for the classification of metagenomic data, but their accuracy, reference databases, and computational requirements vary considerably.

Accuracy is particularly critical when metagenomic methods are applied in clinical contexts, for instance in the identification of pathogenic species. For such applications, algorithms must provide highly accurate results, minimize false positives, and at the same time be fast and scalable.

During my doctoral work, the objectives were the following:

- To develop a novel metagenomic classification algorithm
- To generate and collect datasets required for algorithm validation
- To conduct comparative benchmarking against existing methods
- To demonstrate the algorithm’s performance on real clinical data

We selected nine metagenomic classifiers to include in our comparative analysis: Bracken, Centrifuge, CLARK, DIAMOND, GOTTCHA, Kaiju, Kaiju-ws, Kraken2, and

MetaPhlAn3. The selection criteria for these software were their popularity in the scientific community and the diversity of algorithmic approaches they implement. The nine tools were tested on shotgun-sequenced human stool microbiome samples, and their outputs were compared in terms of the number and identity of detected bacterial species, as well as the degree of similarity among results.

Our own classification algorithm, Novel Alignment-based Biome Analysis Software+ (NABAS+), was developed with an emphasis on accuracy and the elimination of false positives. The software was implemented in Java, exploiting its speed and scalability, which also ensures integration with other bioinformatic pipelines. In addition, we created a stand-alone command-line version to facilitate its independent use.

The performance of NABAS+ was evaluated using the following datasets:

- “In-house” *in silico* datasets (30 samples): These simulated microbial communities modeled samples from various human body sites (e.g., gut, oral cavity) and environmental sources (e.g., urban park, asphalt). In total, 212 bacterial species were included, with sequencing depths ranging from 5×10^5 to 1×10^7 reads.
- Critical Assessment of Metagenome Interpretation II (CAMI II) “toy human gastrooral” dataset (20 samples): A freely available *in silico* dataset that models the microbial composition of the human gastrooral tract and is widely used in metagenomic research.
- Zymo Community Standards I and II (CSI and CSII, 2 samples): Commercially available reference standards consisting of DNA from 10 bacterial species commonly found in the human microbiome, in precisely defined ratios.
- Clinical samples (20 samples): Shotgun-sequenced human gut microbiome samples collected under real hospital conditions, in which pathogen content had also been confirmed using laboratory diagnostic methods. Eleven samples

contained pathogenic species (*Aeromonas, Campylobacter* or *Salmonella*), while nine were pathogen-free.

The nine classifiers produced strongly discordant results on the real-world metagenomic samples. The number and identity of detected taxa differed widely between classifiers, with only limited overlap in species identification.

The performance of NABAS+ was evaluated as follows:

- *In silico* datasets: NABAS+ accurately determined both the number and identity of species, achieving performance comparable to the reliable classifiers identified in our study (Kraken2, MetaPhlAn3, and GOTTCHA). Subsequent comparisons were therefore restricted to these three leading tools.
- CAMI II dataset: NABAS+ consistently produced results comparable to those of the best-performing methods. Re-generation and -analysis of one outlier sample (sample19) further improved classification accuracy.
- Zymo standards: NABAS+ performed exceptionally well, accurately detecting low-abundance species and uniquely avoiding false positive identifications. Based on Bray-Curtis dissimilarity, NABAS+ produced results closest to the known reference composition.
- Clinical samples: NABAS+ successfully identified pathogens present in infected samples and did not report any false positives in healthy samples.

In conclusion, in this study, I developed and validated a novel metagenomic classification tool, NABAS+. Its performance is comparable to that of the most widely used state-of-the-art algorithms, while uniquely minimizing false positives during classification. This property is of particular importance in clinical applications, where misidentification of pathogens can have serious consequences.

Our results demonstrate that NABAS+ is a reliable tool not only for research but also for clinical settings, contributing to improved accuracy and reproducibility in metagenomic studies. The development and public availability of this software may further support its widespread adoption within the scientific community.