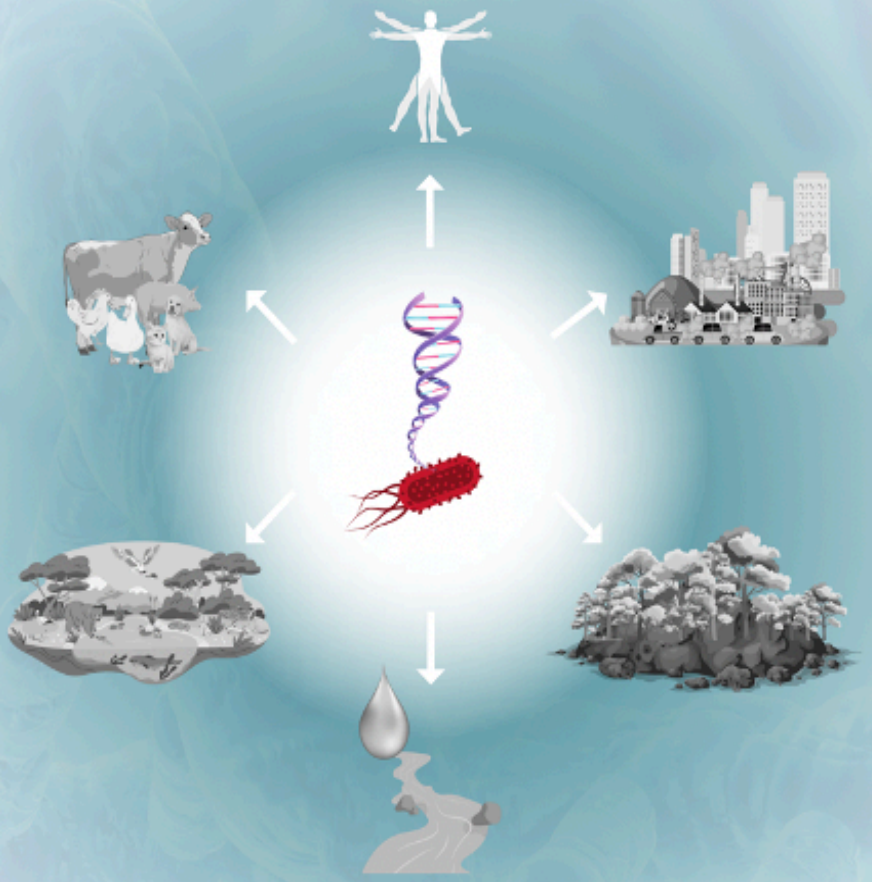
Université de Liège
Faculté de Médecine vétérinaire
Département des Sciences des denrées alimentaires (DDA)
Secteur Microbiologie



Détermination des sources de contamination bactériennes sur base de l'étude de leurs séquences ribosomiques

Sources tracking of bacterial contamination based on ribosomal sequence analysis

Hiba JABRI
Thèse présentée en vue de l'obtention du grade de Doctorat en Sciences Vétérinaires
Année académique 2024-2025

UNIVERSITE DE LIEGE

FACULTE DE MEDECINE VETERINAIRE

DEPARTEMENT DES SCIENCES DES DENREES ALIMENTAIRES

SECTEUR DE MICROBIOLOGIE

DÉTERMINATION DES SOURCES DE CONTAMINATION BACTÉRIENNE SUR LA BASE DE L'ÉTUDE DE LEURS SÉQUENCES RIBOSOMIQUES

SOURCES TRACKING OF BACTERIAL CONTAMINATION BASED ON RIBOSOMAL SEQUENCE ANALYSIS

Hiba JABRI

THESE PRESENTEE EN VUE DE L'OBTENTION DU GRADE DE

DOCTEUR EN SCIENCES VETERINAIRES

ANNEE ACADEMIQUE 2024-2025

Preface

This thesis represents the culmination of my research efforts, drawing on a range of methodologies and tools to present findings as clearly and effectively as possible. To ensure the clarity, coherence, and precision of the written content, I utilized AI-based writing assistance tools during the drafting and revision stages. These tools supported grammar refinement, style enhancement, and content structuring, allowing me to focus on the scientific rigor and originality of the work. The final manuscript, however, remains my own, reflecting my ideas, research findings, and critical analyses.

Table of Contents

Doctoral Committee

President

Coraline RADERMECKER [1]

Jury members

Prof. Stéphane CHAILLOU [2]
Prof. Djamel DRIDER [3]
Prof. Annick WILMOTTE [4]
Dr. Jean-François CABARAUX [5]
Dr. Damien THIRY [6]
Dr. Henry-Michel CAUCHIE [7]

Supervisors

Dr. Bernard TAMINIAU [8]
Prof. Georges DAUBE [8]

Committee members

Pr. Denis BAURAIN[9]
Pr. Véronique DELCENSERIE[8]

[1] Faculty of Veterinary Medicine, Department of Functional Sciences (DSF), GIGA I3 - Immunophysiology, University of Liège (Belgium).

[2] Research Director at INRAE, MICALIS Institute, Food Microbial Ecology team (France).

[3] Charles Viollette Institute. Polytech Lille (office C315). IUT A (office 3A36). Postal address: Prof. Djamel Drider. ICV-Polytech-Lille office C315. Cité Scientifique. 59655 Villeneuve d'Ascq (France).

[4] Faculty of Science, Department of Life Sciences, Diversity and Molecular Ecology of Cyanobacteria, Faculty of Science, Department of Life Sciences, Integrative Biological Sciences (InBioS), University of Liège (Belgium).

[5] Faculty of Veterinary Medicine, Department of Veterinary Resource Management (DRA), Ecology of Health and Animal Production, Faculty of Veterinary Medicine, University of Liège (Belgium).

[6] Faculty of Veterinary Medicine, Department of Infectious and Parasitic Diseases (DMI), Veterinary Bacteriology and Animal Bacterial Diseases, FARAH: Veterinary Public Health (Belgium).

[7] Faculty of Science, Department of Environmental Science and Management, Arlon Environment Campus (Belgium, Luxembourg).

[8] Laboratory of Food Microbiology, Fundamental and Applied Research for Animals and Health Center (FARAH), University of Liège (Belgium).

[9]Eukaryotic Phylogenomics, InBioS-PhytoSYSTEMS, University of Liège (Belgium).

Acronyms

ARB: ARBitrary; a Software environment for sequence data form SILVA database
BB: Bacteria Biotope
BFI: Bacterial Fingerprint Identification
BioNLP: Biomedical Natural Language Processing
DDBJ: DNA Data Bank of Japan
DGGE: Denaturing Gradient Gel Electrophoresis
DNA: Deoxyribonucleic Acid
EBI: European Bioinformatics Institute
ENVO: Environment Ontology
FIB: Fecal Indicator Bacteria
FORENSIC: FORest Enteric Source Identification
GTDB: Genome Taxonomy Database
HTS: High-Throughput Sequencing
INSDC: International Nucleotide Sequence Database Collaboration
LDMs: Library-Dependent Methods
LH-PCR: Length Heterogeneity Polymerase Chain Reaction
LIMs: Library-Independent Methods
LSU: Large Subunit
MIGS/MIMS: Minimum Information about a Genomic/Metagenomic Sequence
MIMARKS: Minimum Information about a Marker Gene Sequence
MST: Microbial Source Tracking
NCBI: National Center for Biotechnology Information
NGS: Next Generation Sequencing
OTT: Open Tree of life Taxonomy
OWL: Ontology Web Language
PCR: Polymerase Chain Reaction
qPCR: Quantitative Polymerase Chain Reaction
RDP: Ribosomal Database Project
RNA: Ribonucleic Acid
SeqEnv: Sequence Environment Software
SILVA: Database from Latin SILVA, forest, http://www.arb-silva.de
SRA: Sequence Read Archive
SQL and MySQL: Structured Query Language
SSU: Small subunit
SMRT: Single Molecule Real Time
STENSL: Microbial Source Tracking with ENvironment SeLection
TGS: Third Generation Sequencing
T-RFLP: Terminal Restriction Fragment Length Polymorphism
WGS: Whole Genome Sequencing
XML: Extensible Markup Language
MAR: Antibiotic Resistance
MAG: Metagenome-Assembled Genomes

Figure List

List of tables

Abstract- Résumé

Abstract

Bacteria are ubiquitous, colonizing a wide range of habitats and biotopes, and their diversity is shaped by varying environmental conditions and levels of competition. This results in complex and diverse bacterial communities across different ecosystems. Molecular biology has significantly advanced our ability to track bacterial sources, enabling precise identification of contamination origins through various techniques. Among these, Microbial Source Tracking (MST) and Bacterial Fingerprint Identification (BFI) are commonly used, alongside other traditional approaches.

However, these source-tracking methods face challenges, such as the influence of complex environmental conditions, inadequate sampling, and variability between techniques, which can limit their effectiveness. To overcome these limitations, alternative methods are needed. This thesis employs a targeted gene using the metagenomics study, to enhance bacterial source tracking. By focusing on 16S rDNA genes, this method provides a detailed analysis of bacterial communities, offering a more refined profile of contamination sources.

The research led to the development of two innovative tools: the Metagenomic Source Tool (MGST) and the Sequence Tracking tool (ST). These tools were specifically designed to pinpoint the origins of bacterial contamination in various matrices. To support these tools, a specialized database named “BiotopeBac-DB” was established, tailored for diverse environmental applications. The MGST was used to validate this database, while the ST was applied in studies targeting contamination sources in distinct settings, such as veterinary restrooms and bathing waters.

This work aims to enrich metagenetic analysis by incorporating information on the origins of contamination, with the goal of enhancing the understanding and tracking of bacterial sources in various environments.

Résumé

Les bactéries sont omniprésentes et colonisent une grande variété d'habitats et de biotopes, leur diversité étant influencée par les conditions environnementales et les niveaux de compétition. Cela génère des communautés bactériennes complexes et diversifiées dans différents écosystèmes. La biologie moléculaire a considérablement amélioré notre capacité à suivre l'origine des contaminations bactériennes, permettant une identification précise des sources grâce à diverses techniques, telles que le Microbial Source Tracking (MST) et l'identification d'empreintes bactériennes (BFI), ainsi que d'autres approches classiques.

Cependant, ces méthodes de traçabilité des sources rencontrent des limites, notamment en raison des conditions environnementales complexes, d'un échantillonnage insuffisant et de la variabilité entre les techniques, ce qui peut affecter leur efficacité. Pour surmonter ces obstacles, d'autres méthodes sont nécessaires. Cette thèse propose l'utilisation d'une méthode ciblée basée sur l'étude de la métagénomique, pour améliorer le suivi des sources bactériennes. En ciblant les gènes l'ARN ribosomal 16S, cette méthode offre une analyse détaillée des communautés bactériennes, fournissant un profil plus précis des sources de contamination.

Les recherches ont conduit au développement de deux outils innovants : le Metagenomic Source Tool (MGST) et le Sequence Tracking tool (ST), spécialement conçus pour identifier l'origine des matrices contaminées. Pour soutenir ces outils, une base de données spécialisée nommée 'BiotopeBac' a été créée et adaptée à diverses applications environnementales. Le MGST a été utilisé pour valider cette base de données, tandis que le ST a été appliqué dans des études ciblant les sources de contamination dans des environnements distincts, tels que les sanitaires de cliniques vétérinaires et les eaux de baignade.

Ces travaux visent à enrichir l'analyse métagénétique en intégrant des informations sur l'origine des contaminations, afin de mieux comprendre et améliorer le suivi des sources bactériennes dans différents environnements.

Chapter I

General Introduction

I.1 Microorganisms and their diversity

Prokaryotic microorganisms represent the most extensive source of genetic diversity on Earth, existing long before more complex life forms like protists, animals, and plants. They dominate the biosphere, particularly in the Earth's crust, and contribute significantly to the planet's biomass. These microorganisms are essential for maintaining ecological balance, as they drive biogeochemical cycles, recycle nutrients, and break down both natural and human-made waste. Their activities are critical for the health and stability of ecosystems.

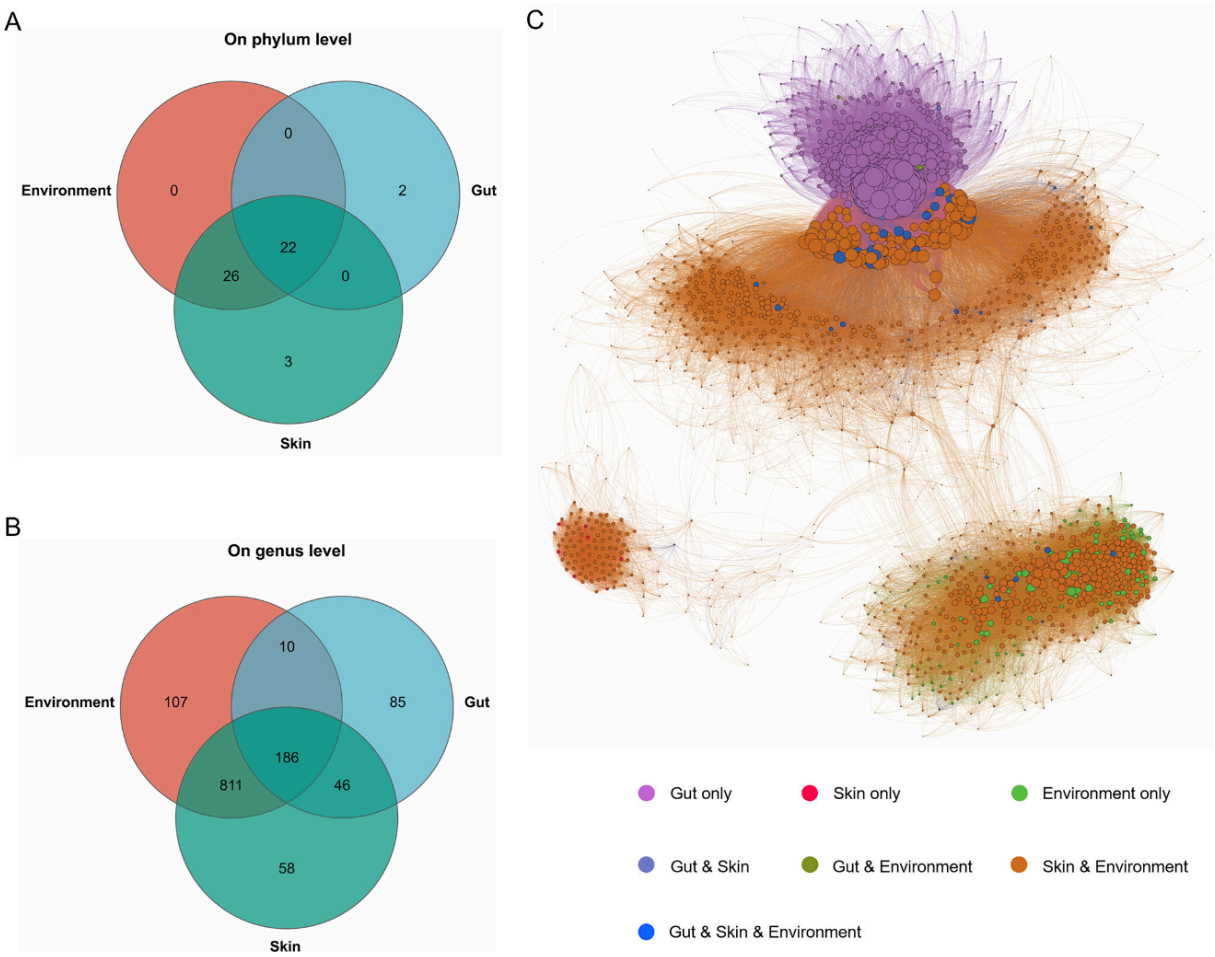
In addition to their ecological roles, microorganisms are a rich source of bioactive compounds, enzymes, polymers, and biotechnological tools. While much attention has been given to their role in human diseases, their broader contributions to industries such as food production and pharmaceuticals are equally important. For example, microorganisms are involved in processes like wine fermentation and cheese ripening, highlighting their integral role in daily life. Studying these organisms is key to unlocking their untapped potential for scientific and industrial applications.

Understanding the diverse ecological niches where bacteria thrive and their roles within these environments is crucial for advancing both science and industry. Research in this area has provided valuable insights into bacterial distribution across various habitats, including soil, water, and host-associated environments. For instance, studies on soil microbiomes have revealed the complex relationships between bacteria, plant health, and nutrient cycling (Philippot et al., 2013; Fierer, 2017). Similarly, research in aquatic ecosystems has explored bacterial diversity and their roles in nutrient cycling and pollutant degradation (Newton et al., 2011; Kirchman, 2016). Investigations into host-associated microbiota, such as the human microbiome, have also demonstrated the importance of bacterial communities in maintaining host health and influencing physiological processes (Human Microbiome Project Consortium, 2012; Sender et al., 2016). These studies underscore the need for a comprehensive understanding of bacterial functions in specific environments, which is essential for applications in environmental conservation and biotechnology.

In medical and health research, high-throughput sequencing has revolutionized our ability to study bacterial sources and improve source tracking. This technology has enabled the exploration of bacterial communities in diverse environments, from the densely populated

gastrointestinal tract (Scheithauer et al., 2016; Tropini et al., 2017) to low-biomass sites like the lower respiratory tract (Dickson et al., 2014; Dickson et al., 2017; Yatera et al., 2018) and the uterus (Collado et al., 2016; Perez-Muñoz et al., 2017). The development of high-throughput sequencing techniques builds on earlier advancements, such as the pioneering work of Sogin et al. (2005), who were among the first to use operational taxonomic units (OTUs) without relying on cloning techniques, significantly advancing the field of microbial diversity research. However, debates continue regarding the sterility of certain niches, such as the placental microbiome (Pelzer et al., 2017; Kuperman et al., 2020), and the challenges of studying low-biomass samples (Drengenes et al., 2019; Dahlberg et al., 2019). These discussions highlight the complexities of microbial diversity research and the ongoing challenges in defining the boundaries of microbial ecosystems.

Figure 1, adapted from Zhong et al. (2023), provides a fascinating look into the bacterial communities of white rhinoceroses, revealing intricate interactions across different ecological niches. The Venn diagrams in Fig. 1A and B show significant overlap in bacterial phyla and genera among the gut, skin, and environmental samples, indicating a complex network of interactions. Shared phyla include Firmicutes, Actinobacteriota, Chloroflexi, Bacteroidota, and Proteobacteria, reflecting a rich diversity across these niches. This aligns with the concept of microbial niches in the human body, where distinct microbial communities inhabit different anatomical sites (Lloyd-Price et al., 2016). Additionally, the co-occurrence network pattern in Fig. 1C highlights the interconnectedness of bacterial species within the gut, skin, and environment, suggesting that these communities mutually influence each other. This finding enhances our understanding of microbial ecology and the complexity of microbial landscapes in different niches.



Source: figure from Zhong et al., 2023

Figure 1. Interactions among bacterial communities in different niches of white rhinoceroses. *(A) Venn diagram based on the phylum level of bacterial communities in the gut, skin, and environment of white rhinoceroses. (B) Venn diagram based on the genus level of bacterial communities in the gut, skin, and environment of the white rhinoceroses. (C) The co-occurrence network pattern depicts the interactions among bacterial communities from the gut, skin, and environment. High clustering coefficients indicate strong connectivity within each niche, while distinct clustering regions suggest intricate microbial relationships across the different ecological niches.*

Despite significant progress, there remains a substantial gap in our understanding of prokaryotic diversity, highlighting the need for further research and innovation. As highlighted by Le Guyader (2008), exploring the full potential of these microorganisms requires deeper investigation. High-throughput sequencing of the 16S ribosomal RNA (rRNA) gene has become a key tool for studying microbial communities in various human-associated environments. Studies by Hong et al. (2016), Botterel et al. (2018), and Wang et al. (2017) have demonstrated the power of this approach, helping to address critical gaps in our knowledge.

Building on these advancements, Microbial Source Tracking (MST) has become an essential method for identifying the sources of microbial contamination in different environments. MST uses molecular and genetic techniques to trace microbes back to their origins, whether human, animal, or environmental. By analyzing unique microbial markers and genetic signatures, MST enables precise identification of contamination sources, which is crucial for environmental monitoring, public health, and water quality management. In the following section, we will explore the methodologies of MST and its applications in environmental and clinical microbiology, emphasizing its role in protecting ecosystems and advancing scientific understanding.

I.2 Detection of Sources for Eukaryotes, Bacteria, Parasites, and Viruses

Identifying the sources of contamination by eukaryotes, bacteria, parasites, and viruses is crucial for environmental microbiology, public health, and ecosystem management. Over time, various techniques have been developed to detect and trace these microorganisms back to their origins. These methods range from traditional microbiological approaches to advanced molecular and bioinformatic tools, each tailored to the unique characteristics of the organisms being studied.

For eukaryotes, such as pathogenic fungi and protists, environmental DNA (eDNA) analysis has become a key method. By isolating and amplifying DNA from water, soil, or sediment samples, eDNA techniques allow researchers to detect specific eukaryotic organisms without the need for culturing. Metagenomic sequencing takes this further by identifying entire eukaryotic communities, helping trace sources of organisms like *Cryptosporidium* in contaminated drinking water (Xiao et al., 2000). Additionally, fluorescence *in situ* hybridization (FISH) has been used to visualize specific eukaryotic cells in environmental samples, providing spatial and source information (Amann et al., 1995).

For bacteria, detection methods have shifted from traditional culture-based techniques to advanced molecular approaches. Culture-independent methods, such as 16S rDNA gene sequencing, enable the identification of bacterial taxa in complex environmental samples with high precision (Woese & Fox, 1977). However, 16S rDNA sequencing has limitations, including insufficient resolution to distinguish closely related species or strains, reliance on incomplete reference databases that may misclassify novel taxa, and amplification biases that skew community representation (Jovel et al., 2016). These constraints reduce its utility in pinpointing exact contamination sources, as many host-specific markers (e.g., human vs. animal Bacteroides variants) require strain-level discrimination. To address this, targeted methods like quantitative PCR (qPCR) and multiplex PCR are employed to detect specific bacterial markers linked to fecal contamination, such as host-associated *Bacteroides spp*. sequences, which bypass 16S ambiguities (Bernhard & Field, 2000). While 16S rDNA is a cornerstone for differentiating prokaryotes (bacteria and archaea) from eukaryotes, it cannot determine microbial functional traits. This limitation arises because the 16S rRNA gene encodes structural ribosomal RNA rather than metabolic or functional genes. Additionally, the presence of multiple 16S rRNA gene copies per bacterial cell can introduce quantification biases,

complicating efforts to correlate taxonomic identity with ecological roles. These approaches are applied in scenarios like tracing *Escherichia coli* in agricultural runoff or identifying bacterial pathogens in recreational waters to reduce public health risks.

For parasites, molecular techniques like nested PCR and loop-mediated isothermal amplification (LAMP) are highly effective for detecting DNA from organisms such as *Giardia lamblia* and *Cryptosporidium parvum* in water samples (Fayer et al., 2000). These methods are especially useful in low-concentration situations where traditional microscopy or immunoassays may fail. Immunological techniques, such as enzyme-linked immunosorbent assays (ELISA), are also used to detect parasitic antigens in environmental and clinical samples, enabling rapid identification of contamination sources during outbreaks. For example, combining LAMP and ELISA has been used to monitor municipal water supplies in developing regions, reducing waterborne parasitic infections (Notomi et al., 2000).

Viruses present unique challenges due to their small size and diversity, but advances in molecular biology have provided effective detection tools. Reverse transcription-PCR (RT-PCR) is widely used to detect RNA viruses, such as noroviruses and hepatitis A virus, in environmental water samples (Kitajima et al., 2020). Virome analysis, a subset of metagenomics, allows researchers to characterize entire viral communities, helping identify the origin of specific viral strains during outbreaks. For instance, virome sequencing has been used to trace enteric viruses in urban wastewater, distinguishing human-derived contamination from animal sources (Bibby & Peccia, 2013). Digital droplet PCR (ddPCR) further enhances viral detection by offering high sensitivity in quantifying low-abundance viral pathogens.

These techniques are often combined with isotopic and chemical fingerprinting to provide a more complete understanding of contamination pathways. For example, stable isotope analysis paired with microbial detection has been used to trace agricultural pollutants, including microbial contaminants, back to specific farms (Kendall et al., 1998). Applications of these methods span water quality monitoring, epidemiological investigations, and ecosystem conservation. In coastal environments, they have been critical in identifying sources of harmful algal blooms, while in urban settings, they help pinpoint the origin of microbial contamination in stormwater systems.

Overall, the detection of eukaryotes, bacteria, parasites, and viruses has been greatly advanced by molecular and bioinformatic innovations. These tools provide critical insights into microbial

community dynamics and their interactions with human and environmental systems, enabling targeted interventions and sustainable management practices.

As the detection of microbial contaminants has advanced, the need to pinpoint their specific sources has become increasingly important. Microbial Source Tracking (MST) addresses this challenge by focusing on identifying the origins of microbial contamination, particularly fecal pollution, in environmental systems. With the help of molecular and bioinformatic tools, MST allows researchers to distinguish between human, animal, and environmental sources of microbes, providing critical information for public health interventions and environmental management. This approach has become indispensable in addressing contamination issues, as it not only identifies the presence of harmful microorganisms but also traces their pathways, enabling more effective and targeted mitigation strategies.

MST was initially developed to identify the sources of fecal contamination in environmental waters, such as rivers, lakes, and coastal areas. MST uses DNA-based techniques, including polymerase chain reaction (PCR) and quantitative PCR (qPCR), to detect and quantify genetic markers that are specific to certain sources. These markers, often found in bacterial genera like Bacteroides, help trace contamination back to its origin (Field et al., 2007). Over time, MST has been adapted to track microbial contamination in other environments, such as soils, sediments, and animal microbiomes (Shanks et al., 2010; Harwood et al., 2013). This versatility has made MST a valuable tool not only for water quality monitoring but also for broader ecological and environmental studies.

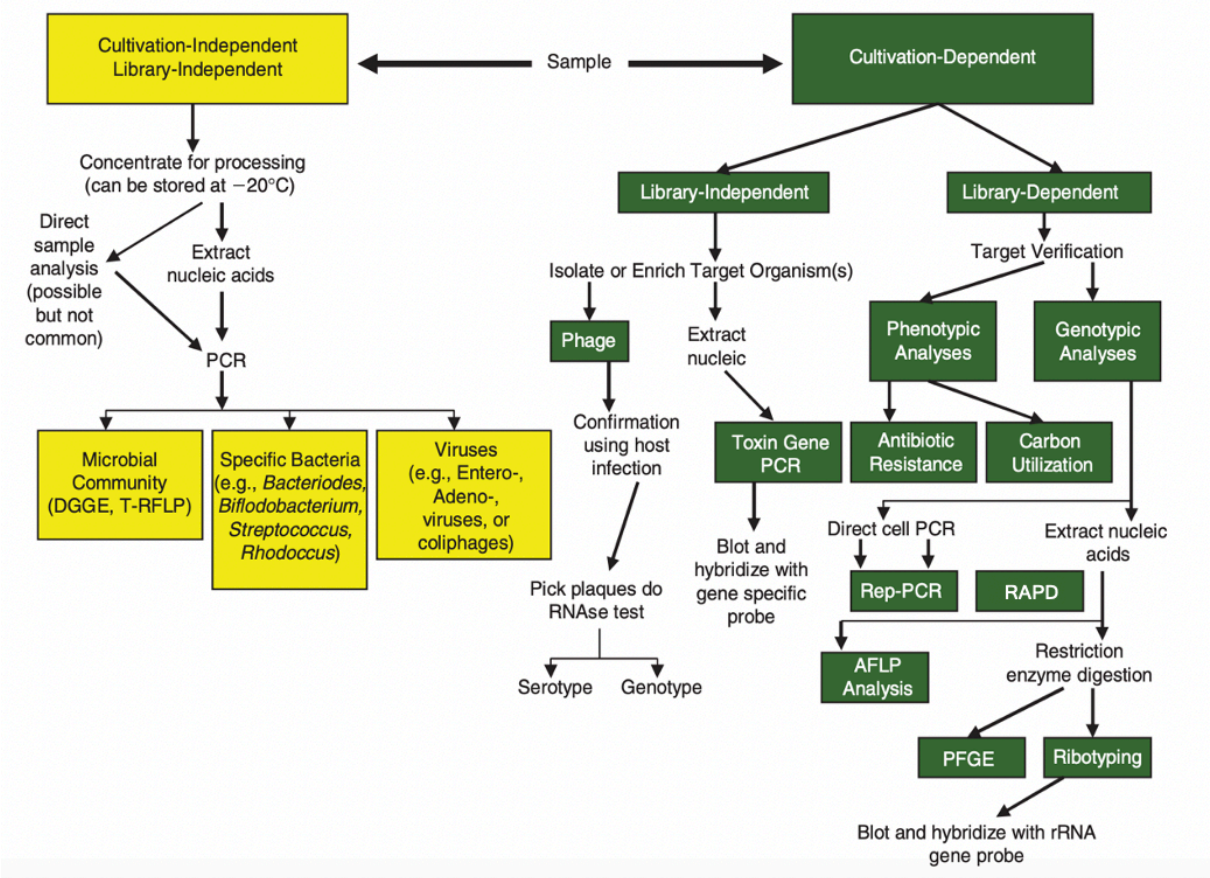
MST methods are generally divided into two categories: Library-Dependent Methods (LDMs) and Library-Independent Methods (LIMs). LDMs involve creating a database of fecal bacteria from known sources and comparing environmental samples to this database to identify contamination sources. LIMs, on the other hand, do not require a reference library and instead use genetic or chemical markers specific to certain miro-organismes to directly identify contamination sources from environmental samples (Harwood et al., 2014).

One key focus of environmental microbiology is source tracking, which can be used to identify the origin of microbial fecal contamination among others. Once the source is identified, control measures can be implemented to address the contamination. Various MST methods have been developed to distinguish between human and animal sources and to trace specific discharges into surface and groundwater.

Library-Dependent Methods (LDMs) rely on comparing environmental isolates to a database of known sources. For example, phenotypic methods might analyze how bacteria like E. coli or enterococci respond to antibiotics or carbon sources, matching these patterns to those in a reference library. While LDMs are quantitative, sensitive, and reproducible, they are time-consuming, require extensive databases, and can have higher false-positive rates compared to LIMs (Fong & Lipp, 2005).

Library-Independent Methods (LIMs) do not require a reference library. Instead, they use host-specific genetic markers to identify contamination sources directly. For example, genotypic methods like PCR can detect markers in bacteria such as *Bacteroides* or *Bifidobacterium* to differentiate between human and animal sources (Bernhard & Field, 2000; Bernhard et al., 2003). Other LIMs include using enterotoxin genes in *E. coli* to distinguish between domestic animal sources (Khatib et al., 2002) or enterococcal surface protein (ESP) in *Enterococcus faecium* to differentiate human and animal contamination (Scott et al., 2005). Additionally, male-specific coliphages and viral pathogens like hepatitis A virus can be used to trace contamination sources (Fong & Lipp, 2005; Stewart-Pullaro et al., 2006).

LIMs are faster and more accurate in distinguishing human from non-human sources. However, they require expensive equipment and are limited by our understanding of how well host-specific markers persist in the environment. Figure 2 provides a flowchart summarizing the different approaches used in microbial source tracking.



Source: figure from Charles P.GERBA :Indicator Microorganisms 2015

Figure 2. Illustrative diagram depicting various microbial source tracking (MST) methods. *The methods include PCR (Polymerase Chain Reaction), Rep-PCR (Repetitive-Intergenic DNA Sequence PCR), DGGE (Denaturing Gradient Gel Electrophoresis), T-RFLP (Terminal-Restriction Fragment Length Polymorphism), PFGE (Pulsed Field Gel Electrophoresis), RAPD (Randomly Amplified Polymorphic DNA), and AFLP (Amplified Fragment Length Polymorphism).*

Figure 2 presents a visual overview of microbial source tracking (MST) methodologies, highlighting their applications and technical differences. The figure organizes common molecular techniques such as PCR, DGGE, PFGE, RAPD, and AFLP into two categories: Library-Dependent Methods (LDMs), which rely on reference databases, and Library-Independent Methods (LIMs), which use host-specific markers. By systematically comparing these approaches, Figure 2 clarifies how MST tools identify sources of fecal contamination, connecting molecular biology principles with environmental monitoring objectives. For example, LDMs (e.g., PFGE) are shown alongside LIMs (e.g., host-specific PCR), demonstrating how these methods complement each other in source attribution. This visual

representation is particularly useful for readers who may not be specialists in the field, providing a clear comparison of evolving MST practices, such as the transition from labor-intensive fingerprinting techniques (e.g., DGGE) to faster, marker-based assays. Furthermore, the figure underscores the advancements in MST technologies, illustrating how improvements in genetic analysis have increased precision in public health and forensic applications. In summary, Figure 2 enhances the text by offering a structured and accessible depiction of MST workflows, aiding in the understanding of complex methodologies.

MST has been widely used for various purposes, including regulatory compliance, pollution remediation, and risk assessment. It plays a critical role in public health and water quality management by enabling the identification of fecal contamination sources (Scott et al., 2005). MST employs a range of molecular and bioinformatic techniques to analyze microbial communities and distinguish between different contamination sources. These methods take advantage of the genetic diversity of microbes, often targeting specific genetic markers unique to particular sources.

The 2002 review "Microbial Source Tracking: Current Methodology and Future Directions" by Scott et al. highlights the limitations of traditional fecal pollution monitoring, which relied on general microbial indicators like *E. coli, Enterococcus spp.,* and *Clostridium perfringens*. While these indicators effectively signaled fecal contamination, they could not distinguish between human and animal sources a critical gap for targeted remediation. For example, *E. coli* abundance in water confirmed fecal pollution but provided no insight into whether contamination originated from livestock, wildlife, or sewage. The authors emphasize how this ambiguity hindered public health responses, as risks differ vastly between sources (e.g., human feces may carry pathogens like Norovirus, whereas livestock waste might prioritize zoonotic agents) (Scott et al., 2002).

Scott et al. then introduces emerging MST methodologies designed to overcome these shortcomings, such as library-dependent approaches (e.g., ribotyping, PFGE) that compare environmental isolates to fecal source libraries, and library-independent techniques (e.g., host-specific PCR markers like *Bacteroides* HF183 for humans). Scott et al. argue that MST's ability to discriminate sources revolutionizes water quality management by linking contamination to specific origins, enabling precise interventions. The study also foresees future advancements in

genetic marker discovery and standardization, which have since materialized with next-generation sequencing and qPCR-based assays (Scott et al., 2002).

Traditional fecal pollution monitoring relied on general microbial indicators like *E. coli* and *Enterococcus spp*., which lacked specificity for identifying contamination sources. Scott et al. (2002) outline the evolution of MST methodologies to address this gap, emphasizing the rationale of exploiting host-adapted microbial traits. Genetic methods assume that bacterial lineages develop unique genomic fingerprints due to host-specific adaptations (e.g., nutrient availability, pH), enabling differentiation of human versus animal sources. Phenotypic approaches, such as antibiotic resistance (MAR) or biochemical profiles, similarly target traits shaped by host environments. Early microbiological methods, like the fecal *coliform/streptococcus* ratio, were abandoned due to variability, while *Bifidobacterium spp*. obligate anaerobes abundant in human feces emerged as promising human-specific indicators. Despite advantages (e.g., inability to replicate post-excretion, signaling recent contamination), *Bifidobacterium* faced challenges like rapid environmental die-off and detection sensitivity. These limitations underscored the need for advanced molecular techniques, paving the way for modern library-dependent (e.g., ribotyping) and library-independent (e.g., host-specific PCR) MST frameworks.

Recent innovations, such as metagenomics and STENSL (Microbial Source Tracking with ENvironment SeLection) by Raza et al. (2021) and An et al. (2022), have expanded the scope of MST. Contributions from Bowen et al. (2023), who developed a comprehensive reference library for MST in the mid-Atlantic United States, and Vanderzalm et al. (2023), who applied MST to identify fecal pollution in coral reef lagoons, demonstrate the method's adaptability across diverse environments. Frank et al. (2024) revealed marine sediments as a reservoir for *Escherichia coli* using signature-based and novel amplicon sequencing approaches. Nam et al. (2023) evaluated MST markers on food-contact surfaces, providing insights into contamination sources in school cafeterias, while Vadde et al. (2022) characterized fecal pollution in karst aquifers, showcasing MST's versatility in different hydrological settings.

This research integrates key elements of MST, including source identifiers, detection methods, and analytical frameworks. Source identifiers, such as genotypes or phenotypes of E. coli, F-specific RNA coliphages, and host-specific markers in *Bacteroides/Prevotella*, are combined with detection methods like ribotyping and length heterogeneity PCR (LH-PCR). Analytical

frameworks then link water sample results to specific fecal sources. This thesis aims to contribute to the evolving field of source tracking by extending its focus beyond traditional fecal indicator bacteria (FIB) to include all bacteria sampled in the study area.

To provide further context, the next section will explore the evolution of bacterial identification techniques. From Pasteur's early microscopy to modern PCR and molecular methods, we will trace the development of tools that have significantly advanced our understanding of microbial life. This historical perspective will highlight the transformative impact of technological progress on bacterial identification and the ongoing pursuit of more accurate and comprehensive methods.

I. 3 Evolutionary Shifts in Bacterial Identification: From Pasteur's Microscopy to PCR's Molecular Insights

Beginning in the 1870s, Louis Pasteur conducted important experiments to identify bacteria based on their shape, biochemical properties, and physiological traits (Pasteur, 1876). At the time, bacteria were observed under microscopes, with or without staining, and their shapes, motility, and cell wall structures were studied (Koch, 1884). This work laid the foundation for classical microbiology and led to the development of bacterial identification methods that relied on phenotypic traits, such as growth patterns on culture media containing specific nutrients or metabolic substrates (Buchanan, 1918). These methods became reliable tools for identifying known pathogenic bacteria in medical settings (Koch, 1882; Bergey et al., 1923).

However, these traditional approaches have limitations when applied to complex environments like soil, sediments, freshwater, and seawater (Amann et al., 1995). In these natural habitats, most microorganisms cannot be cultured in the lab, making it difficult to study their diversity and interactions (Staley & Konopka, 1985). Many bacteria exist in a dormant state or interact with other microorganisms, further complicating efforts to fully understand their roles in these ecosystems (Lewis, 2007; Lennon & Jones, 2011). As a result, studying microbial diversity in such environments remains a major challenge (Rappé & Giovannoni, 2003).

The field of molecular biology began to address these challenges with the discovery of the double-helix structure of DNA by Watson and Crick in 1953, which revealed the complementary nature of DNA strands and hinted at a potential copying mechanism (Watson & Crick, 1953). This foundational discovery paved the way for understanding DNA replication and repair (Meselson & Stahl, 1958). In 1957, Arthur Kornberg identified the first DNA polymerase, an enzyme essential for DNA synthesis, though it required a primer and could only copy DNA in one direction (Kornberg, 1957). By the 1970s, researchers like Gobind Khorana and Kjell Kleppe explored DNA repair synthesis and envisioned a two-primer system for DNA replication, laying the groundwork for future advancements in DNA amplification (Khorana et al., 1971; Kleppe et al., 1971).

The advent of DNA sequencing in 1977, pioneered by Frederick Sanger, further advanced molecular biology by enabling the reading of DNA sequences using DNA polymerase, primers, and nucleotide precursors. By the 1980s, all the components necessary for DNA amplification were in place. However, it was not until 1983 that Kary Mullis, while working at Cetus

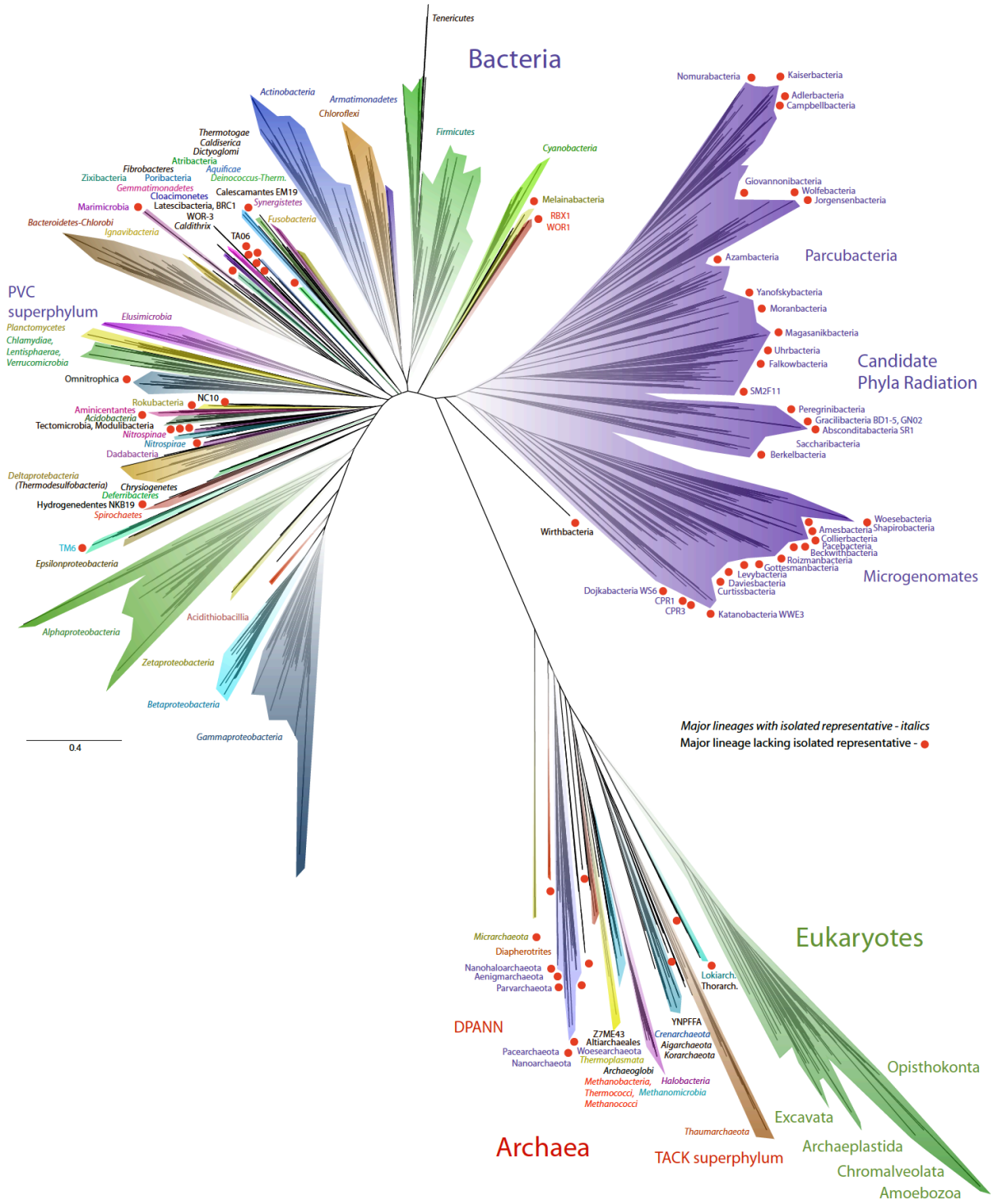
Corporation, combined these elements to develop the Polymerase Chain Reaction (PCR). Mullis's innovation involved using two primers and repeated cycles of DNA polymerase activity to exponentially amplify specific DNA segments, revolutionizing the study of genetics and microbial diversity.

In the 1990s, PCR became a cornerstone of molecular biology, allowing scientists to amplify and study bacterial DNA with greater facility. This breakthrough opened new possibilities for identifying bacteria, detecting their genomes, and studying genetic mutations. It also provided a powerful tool for exploring microbial diversity, particularly for non-cultivable bacteria. PCR's utility was further enhanced by techniques like Southern blotting, which confirmed the amplification of specific DNA segments and validated the accuracy of PCR products (Mullis et al., 1986).

Today, PCR remains a critical tool in microbial research, enabling the study of complex environments and unculturable microorganisms. Its development marked a turning point in microbiology, bridging the gap between classical phenotypic methods and modern molecular techniques, and continues to drive advancements in microbial ecology, biotechnology, and medical research.

Using PCR and other DNA amplification techniques, researchers discovered that mutations in bacterial genomes or specific genes accumulate over generations (Dauga et al., 2005). This finding has been crucial for phylogenetic studies, which aim to reconstruct the evolutionary relationships between organisms. Unlike traditional methods that focus on physical similarities and differences, phylogenetic analysis of DNA sequences provides a deeper understanding of microbial evolution. By calculating the evolutionary distances between generations, scientists can construct phylogenetic trees, where branch lengths represent mutation rates (Figure 3). This

approach offers a more comprehensive view of microbial history and relationships.



Source: figure from Hug et al., 2016

Figure 3. An innovative portrayal of the phylogenetic tree unfolds, featuring 92 named bacterial phyla, 26 archaeal phyla, and the five eukaryotic supergroups. *This tree is a result of genomic sampling of previously unexamined environments and includes over 1,000 uncultivated and little-known organisms, along with known sequences. The color-coded branches help distinguish between major groups, and the labels indicate specific subgroups or species. The red dots on certain branches mark "Major lineages with isolated representative*

taxa," highlighting the areas where more research and discovery are needed. The prominent lineages are perceptibly differentiated by distinctive colors and identified with italicized, well-established lineage names. Lineages lacking isolated representatives are conspicuously marked with non-italicized names and accentuated by red dots, underscoring their unique status. This representation affords a comprehensive panorama of the diversity and interrelationships among these microbial groups.

The study of microbial diversity has been transformed by modern discovery strategies, which have built on the advancements in phylogenetic classification, surpassing traditional phenotypic methods. The rise of molecular biology has opened new doors to understanding non-cultivable bacteria, revealing their hidden roles and complexities. Among these strategies, Denaturing Gradient Gel Electrophoresis (DGGE) (Muyzer et al., 1993) has proven invaluable, separating DNA fragments based on their sequences to provide a snapshot of microbial diversity in a sample. Similarly, terminal restriction fragment length polymorphism (T-RFLP) (Liu et al., 1997) complements this by analyzing DNA fragments after digestion with restriction enzymes, offering detailed insights into microbial community composition.

Quantitative Polymerase Chain Reaction (qPCR) (Bustin, 2000) has also emerged as a powerful tool, enabling rapid and sensitive detection and quantification of specific microbial taxa, shedding light on their abundance across different environments. Cloning techniques, particularly the creation and analysis of clone libraries, capture the genetic diversity within microbial communities. Meanwhile, high-throughput sequencing technologies have open new opportunities in microbial ecology by enabling metagenomic analyses, which explore entire microbial communities in unprecedented detail.

The integration of these technologies has not only improved our ability to characterize microbial diversity but also deepened our understanding of the functional roles of microbial taxa in complex ecosystems. Techniques like DGGE, T-RFLP, and amplicon sequencing are central to microbial community fingerprinting, which analyzes the composition and diversity of bacterial communities across various environments. These methods have revealed unique microbial fingerprints, significantly advancing our understanding of bacterial ecology and its applications in environmental science, agriculture, medicine, and biotechnology.

For example, in environmental science, these techniques are used to study bacterial communities in polluted soils, identifying species that can degrade pollutants and aiding in

bioremediation efforts. In agriculture, amplicon sequencing helps characterize soil microbiomes, identifying beneficial bacteria that improve soil fertility and crop health by fixing nitrogen or suppressing plant pathogens. In medicine, T-RFLP and amplicon sequencing are used to study changes in the gut microbiome of patients with chronic diseases like ulcerative colitis or irritable bowel syndrome, guiding the development of personalized treatments. In biotechnology, DGGE and amplicon sequencing are employed to discover bacteria with probiotic properties or those capable of producing industrial enzymes, leading to innovations in food and pharmaceutical industries.

Together, these approaches are driving a revolution in microbial ecology, providing a more nuanced and holistic understanding of the microbial world. As we continue to explore microbial ecosystems, it is essential to expand our analytical tools. This brings us to advanced methodologies like whole-genome sequencing (WGS), metagenetics, and metagenomics, which offer a more comprehensive analysis of microbial communities beyond traditional source tracking. In the following section, we will examine how these methods integrate to enhance our understanding of microbial diversity and function across different environments.

I.4 Metagenetics, metagenomics and whole-genome sequencing

Metagenetics, metagenomics, and whole-genome sequencing are three distinct approaches used to study microbial communities, each with its own methods, technologies, and applications.

The advent of Next-Generation Sequencing (NGS), marked by transformative technologies like Illumina (Metzker, 2010) and 454 pyrosequencing (Margulies et al., 2005), has enable the development of these strategies. These technologies have overcome the limitations of traditional culture-based methods, uncovering the vast diversity of non-cultivable microorganisms that were previously inaccessible.

Unlike traditional sequencing methods, NGS platforms can process millions of DNA fragments simultaneously, enabling the rapid and cost-effective generation of large amounts of sequence data. This high-throughput capability has been particularly transformative in metagenomic studies, allowing researchers to identify rare and novel microbial taxa and explore microbial communities at a genomic scale. Figure 4 illustrates the evolution of sequencing technologies and their applications over time.

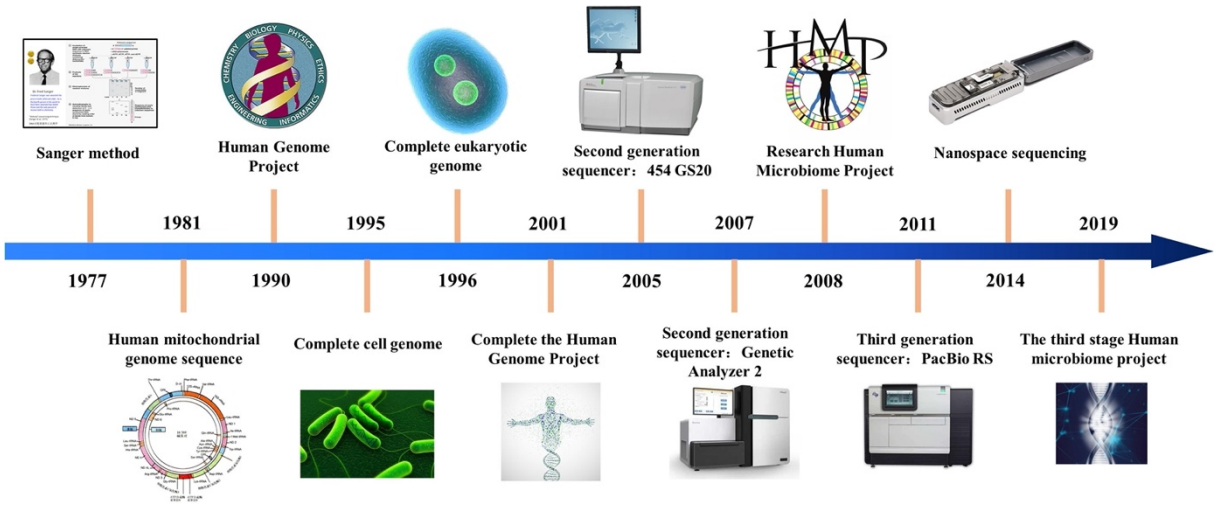
Metagenetics focuses on the targeted sequencing of conserved marker genes, such as the 16S rRNA gene for bacteria and archaea and the 18S rRNA gene for fungi. As shown by Woese and Fox (Woese, 1977), the ribosomal small subunit is a conserved marker, suited for phylogenetic analysis and for taxonomic classification. Its sequencing enables researchers to identify both known and novel microbial taxa.

In contrast, metagenomics bypasses marker genes and sequences all the genetic material in a sample. This provides a comprehensive view of the genomic content of microbial communities, enabling deeper insights into microbial diversity and function. Figure 5 illustrates the workflows for both targeted metagenetics and broader metagenomics approaches, highlighting the experimental and bioinformatics steps involved.

Metagenome-Assembled Genomes (MAGs) represent a powerful approach to studying microbial communities by reconstructing individual genomes directly from metagenomic data. Unlike traditional methods that rely on isolating and culturing individual species, MAGs allow researchers to piece together genomes from complex environmental samples, even for unculturable microorganisms. This is achieved through advanced computational techniques that

assemble short DNA reads into longer, contiguous sequences, which are then binned into individual genomes based on sequence similarity and coverage. MAGs provide valuable insights into the genetic potential, metabolic pathways, and ecological roles of microbial species within a community. They have become a cornerstone of modern metagenomics, enabling the discovery of novel species, functional genes, and microbial interactions in diverse environments, from soil and oceans to the human gut. However, the quality of MAGs depends on factors such as sequencing depth, assembly algorithms, and binning strategies, which can influence the completeness and accuracy of the reconstructed genomes.

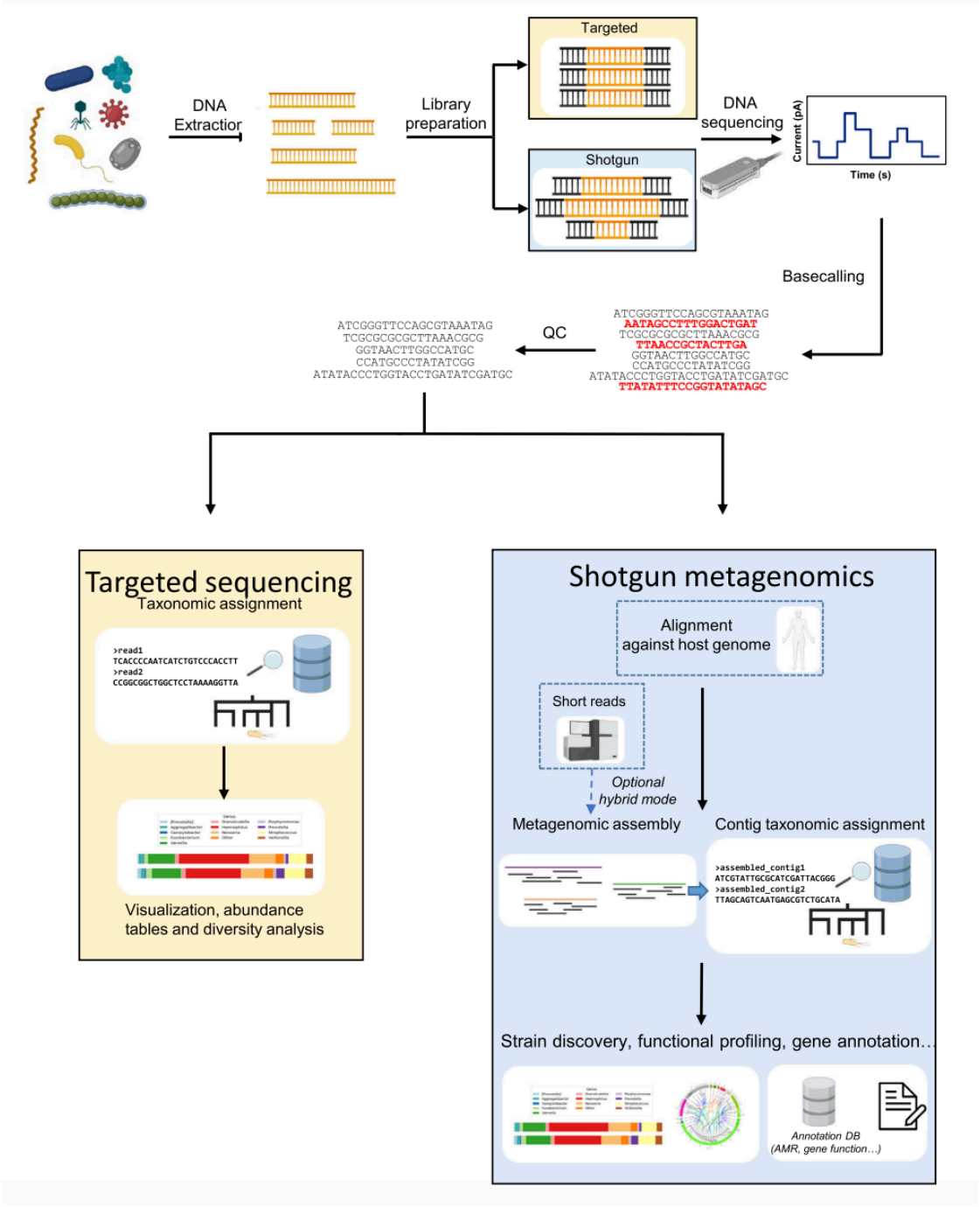
Whole-Genome Sequencing (WGS) involves sequencing the entire DNA of an organism's genome. It is used for in-depth studies of individual species, pathogen genomics, and genetic research, providing the most comprehensive genetic information available. WGS, along with high-throughput targeted gene sequencing, has greatly advanced our understanding of microbial genomes.



Source: figure from Yang, Aimin et al.,2020

Figure 4. Evolution of DNA sequencing technologies over time. *This timeline illustrates the progression of DNA sequencing technologies from the introduction of Sanger sequencing to the modern era of High-Throughput Sequencing (HTS). Each key milestone is represented, including the transformative impact of Next Generation Sequencing (NGS) techniques like Illumina and 454 pyrosequencing. The graph highlights the increasing efficiency, throughput, and cost-effectiveness of sequencing methods, showcasing their pivotal role in advancing our understanding of microbial communities and ecosystems.*

The most used marker genes in metagenetics are the 16S rRNA for bacteria and archaea and the 18S rRNA, ITS2, or 28S rRNA for fungi and eukaryotes (Prosser, 2015; Schoch et al., 2012). While viruses make up about 80% of the microbiota, they lack specific marker genes, making them harder to study using these methods (Paez-Espino et al., 2016; Roux et al., 2015). The 16S rRNA gene, approximately 1,500 base pairs long, contains nine hypervariable regions (V1-V9) interspersed with highly conserved regions (Clarridge, 2004). These conserved regions can be targeted with universal primers for amplification, while the variable regions allow researchers to distinguish between different microbial taxa (Hugenholtz et al., 1998; Head et al., 1998; Baker et al., 2003; Chakravorty et al., 2007). A sequence similarity cut-off of 98.65% in the 16S rRNA gene is used to differentiate between distinct species, making it a highly sensitive and standard tool for microbial identification and classification (Bharti et al., 2019; Yarza et al., 2014; Lamoril et al., 2008).



Source: figure from Ciuffreda et al., 2021

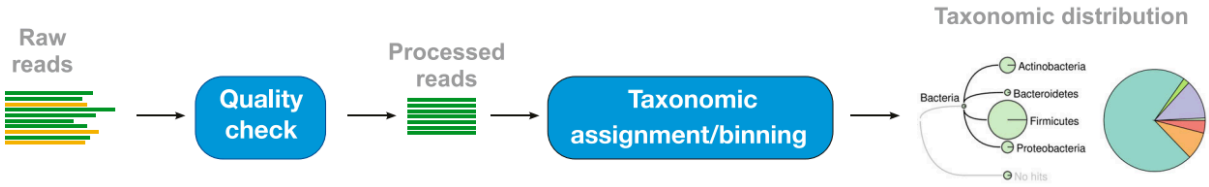
Figure 5. General workflow for generation and analysis of data for targeted and metagenomics studies. *This figure encapsulates the workflow for targeted and metagenomics studies, beginning with sample collection and DNA extraction, followed by library preparation and high-throughput sequencing. Data processing ensures quality control and assembly, while bioinformatics analysis involves taxonomic classification against databases like SILVA or GreenGenes. Functional annotation predicts gene functions using resources like KEGG or CARD, leading to data interpretation that explores microbial diversity and environmental correlations. The workflow culminates in visualization and reporting, summarizing the findings for scientific communication*

Next-generation sequencing (NGS) technologies, such as Illumina, often target short segments of the 16S rRNA gene, focusing on one or more of its nine hypervariable regions (V1-V9). The V1-V3 and V3-V4 regions are particularly popular in sequencing studies due to their role in accurate taxonomic identification and profiling of microbial communities. This approach offers several advantages: lower error rates, reduced complexity, easier alignment with reference genomes, and less ambiguity in identifying taxa. Additionally, short-read sequencing is cost-effective, enabling high-throughput analysis with less computational power, making it ideal for large-scale studies and community profiling.

Long-read sequencing, such as nanopore sequencing, enables the amplification and analysis of the entire 16S rRNA gene, providing a more complete representation of microbial taxa in a sample. PacBio and nanopore sequencing are two widely used long-read sequencing technologies, each with unique strengths and applications. PacBio, developed by Pacific Biosciences, uses Single Molecule Real-Time (SMRT) sequencing to produce long, high-quality, and highly accurate reads. Nanopore sequencing, from Oxford Nanopore Technologies, offers real-time sequencing with read lengths extending into megabases. It is particularly useful when rapid sequencing, portability, or extremely long reads are needed.

Although nanopore sequencing has a higher error rate compared to other technologies, it offers full-length 16S rRNA gene sequencing, improving species-level resolution and the detection of rare taxa (Benítez-Páez et al., 2016; Shin et al., 2016; Nygaard et al., 2020; Rodríguez-Pérez et al., 2020). Tools like NanoCLUST (Rodríguez-Pérez et al., 2020) and other data-polishing algorithms help mitigate errors, producing highly accurate consensus sequences. This advancement allows for precise species-level differentiation and classification of microbial communities, significantly enhancing the resolution and accuracy of metagenetics in microbiology studies.

A typical bioinformatic analysis involves grouping sequencing reads into taxonomic units and calculating the microbial composition of samples. Figure 6 illustrates this workflow, highlighting the steps from sequencing to taxonomic profiling.



Source:figure from Huson DH et al., 2016, page 6

Figure 6. Basic taxonomic binning workflow. *This flowchart shows the steps from raw reads to taxonomic assignment with quality checks.*

For microbial community studies, especially those focusing on the 16S rRNA gene, short-read sequencing is often used to target key variable regions that help identify different species. This approach allows researchers to analyze community diversity and study microbial ecosystems effectively. Researchers can access 16S rRNA gene sequences in general databases like GENBANK (https://www.ncbi.nlm.nih.gov/) and specialized databases like the Ribosomal Database Project (RDP) (https://rdp.cme.msu.edu/), as well as others listed in Table 1. These databases, including the comprehensive and regularly updated ARB-SILVA (https://arb-silva.de/), provide quality-checked ribosomal RNA sequences that support accurate taxonomic classification and enhance our understanding of microbial diversity.

Table 1. Summary of the most widely used reference databases for metagenetics and metagenomics for bacterial analysis.

Database	*Description*	*Area of Use*
GENBANK	A comprehensive repository of sequence data from all domains of life.	General Research
MG-RAST	A metagenomics analysis server that allows functional comparison of microbial communities.	Ecosystems, Human Microbiome Studies
MetaHIT	Focuses on human intestinal microbiota, providing a platform for metagenomic studies.	Human Health, Disease Research
Ensembl Genomes	Offers genome-scale data from a variety of organisms.	Comparative Genomics, Evolutionary Studies

SILVA	Provides quality-checked ribosomal RNA sequences for accurate taxonomic classification.	Biodiversity Monitoring, Microbial Ecology
GreenGenes	A 16S rRNA gene database providing curated taxonomy for bacterial and archaeal sequences.	Microbial Diversity Studies, Taxonomic Classification
GTDB	A genome-based taxonomy database providing standardized bacterial and archaeal taxonomy based on genome sequences.	Microbial Diversity Studies, Taxonomic Classification

Bioinformatic processing is essential for obtaining accurate and reliable results from 16S rDNA amplicon sequencing. The goal is to achieve total sensitivity (detecting all 16S rRNA sequences and taxa) and total specificity (avoiding false positives), while also accurately representing the relative abundances of microbial communities in their natural environments. However, current methods often fall short due to imperfect recall (missing some sequences or taxa) or imperfect precision (detecting false sequences or taxa). These issues stem from various workflow-related flaws, such as biases in sample preparation (e.g., DNA extraction, PCR, library preparation), suboptimal experimental design (e.g., choice of amplicons and primers), errors introduced during sequencing, and limitations in bioinformatic analysis strategies (Kozich et al., 2013; Wesolowska-Andersen et al., 2014; Almeida et al., 2018). Addressing these challenges is critical to improving the accuracy and reliability of 16S rRNA amplicon sequencing and advancing our understanding of microbial communities in diverse environments.

As microbiology continues to evolve, the rise of high-throughput sequencing (HTS) has revolutionized the study of microbial communities. However, these advancements also bring new challenges. The massive volume of data generated by HTS technologies, such as Illumina, PacBio, and Oxford Nanopore, demands robust computational tools and storage solutions. Additionally, processing, analyzing, and interpreting these large datasets present significant bioinformatic hurdles, including issues with data quality, error rates, and the difficulty of assembling short reads. In the next section, we will explore these complexities and discuss the need for improved tools to efficiently manage and analyze large-scale genomic data.

I.5 The rise of high-throughput sequencing data and its challenges

The rise of high-throughput sequencing (HTS) platforms, including Next-Generation Sequencing (NGS) and Third-Generation Sequencing (TGS), has generated vast amounts of data, now stored in large public databases. Key resources like GenBank provide access to nucleotide sequences and their protein translations, while the Sequence Read Archive (SRA) serves as a repository for DNA sequencing data, particularly short reads (less than 1,000 base pairs) produced by HTS. These databases are part of the International Nucleotide Sequence Database Collaboration (INSDC), a partnership between the NCBI, the European Bioinformatics Institute (EBI), and the DNA Data Bank of Japan (DDBJ). While these resources have grown exponentially, managing, processing, and tracing interconnected data remains a significant challenge.

In microbial taxonomy, 16S rRNA sequencing has proven effective for bacterial identification. However, alternative approaches like the Genome Taxonomy Database (GTDB) (Parks et al., 2021) offer a broader genomic perspective. Unlike 16S rRNA sequencing, which focuses on a single genetic marker, GTDB uses whole-genome sequences to infer phylogenetic relationships, providing a more detailed understanding of microbial evolution and taxonomy. While 16S rRNA sequencing is widely used for its simplicity, GTDB complements it by offering higher resolution for evolutionary studies. Integrating these approaches can lead to a more comprehensive understanding of microbial diversity and its ecological implications.

The rapid growth of sequencing data also presents challenges for accurately identifying and classifying bacteria in electronic databases. One major issue is that sequences in public databases are often annotated by the submitting authors without rigorous validation. For example, a 16S rDNA sequence might be labeled as a "bacterial environmental sequence" without precise taxonomic identification. Additionally, there is no standardized peer review process for taxonomic annotations, increasing the risk of errors. This makes it essential to handle these databases carefully to ensure reliable results.

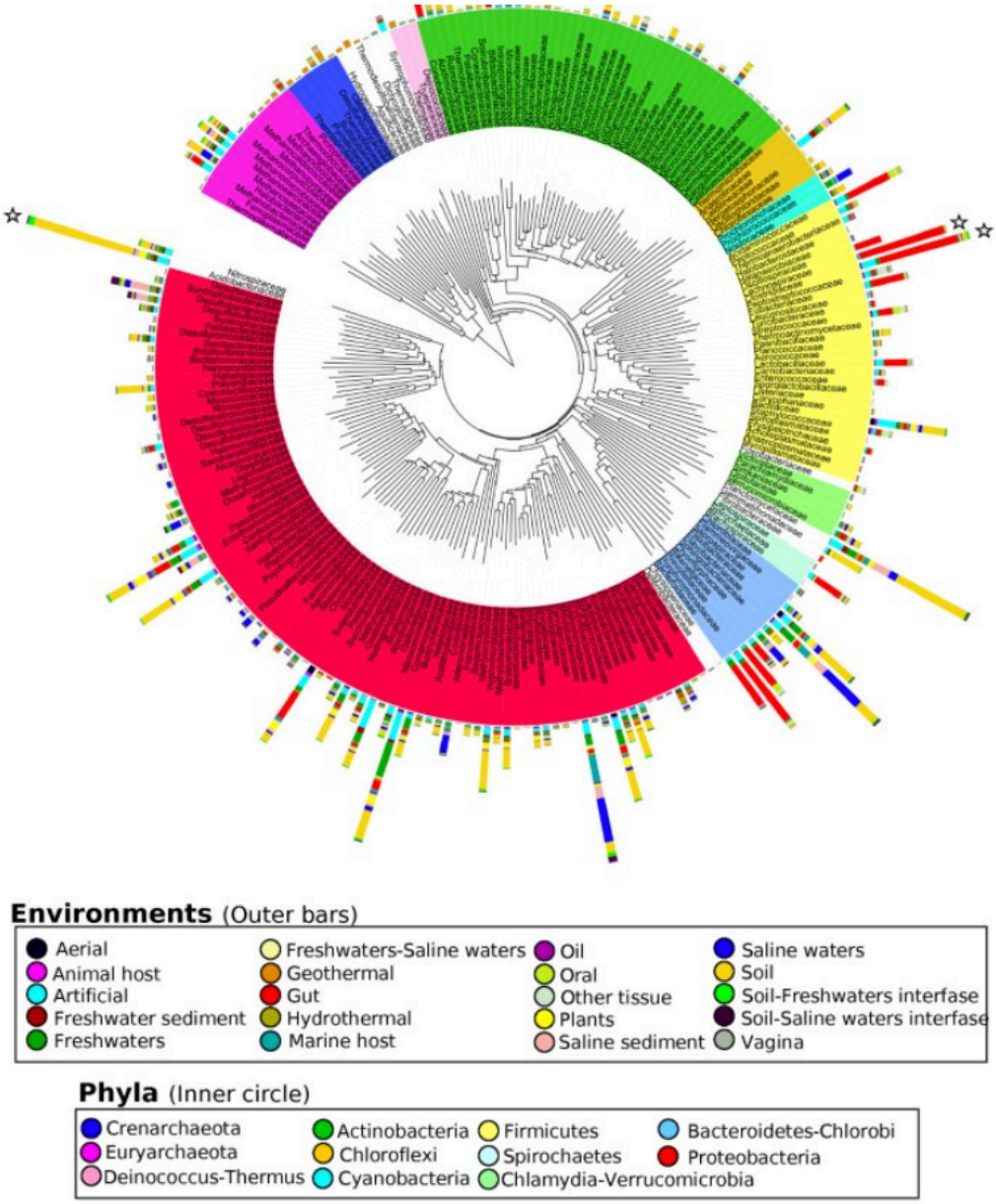
Another challenge is the lack of a uniform protocol for taxonomic identification across scientific journals, which can lead to mislabeling and incorrect attributions. Poor sequence quality, such as chimeric sequences or low-quality reads, further complicates accurate classification. These inaccuracies can propagate errors in metagenomic studies that depend on

precise taxonomy. Despite efforts to improve annotation quality, manual curation remains labor-intensive and slow, limiting progress.

Furthermore, the growing volume of publications on bacterial biotopes and the difficulty of extracting information from text-based formats create barriers to understanding the relationships between bacteria and their environments. The absence of tools or databases to parse this textual data leaves a significant gap in our knowledge of bacterial distribution across different habitats. This information is critical for fields like human and animal health, agriculture, food production, and environmental science. Therefore, selecting appropriate sequencing data is crucial for studying bacterial groups and their habitats to bridge this gap.

In this context, this thesis builds on the foundational work of Tamames et al. (2010) (Figure 7) to expand our understanding of prokaryotic taxa distribution. By exploring microbial communities in various ecological niches, this research aims to contribute to a broader understanding of ecosystem dynamics and their implications.

Choosing the right sequencing datasets is a crucial step in microbial research, as it directly impacts the accuracy and scope of the analysis. With the immense diversity of microbial species and the limitations of current sampling efforts, researchers must carefully evaluate which sequencing approach best suits their study goals. Key factors to consider include read length, sequencing depth, and taxonomic resolution, as these influence the dataset's effectiveness. The next section will discuss strategies for selecting appropriate sequencing datasets, ensuring they align with research objectives and provide meaningful insights into microbial diversity.



Source: figure from Tamames et al. 2010

Figure 7. **The distribution of Taxonomic Families in Various Environments.** *This figure, extracted from the seminal work of Tamames and colleagues in 2010 on "Environmental distribution of prokaryotic taxa," presents a comprehensive exploration of the distribution of individual taxonomic families across different environment types. The inner circle shows a phylogenetic tree, based on the assumption that one representative sequence has been taken from each family. It was arbitrarily rooted in the split between bacteria and archaea, and families are colored with the corresponding phyla. Families containing at least 10 observations have been treated. The outer circle contains the number of times each family has been observed in a sample from a particular environment, marked by the bars. The stars marked on the bars have been reduced to one third of their original size for clarity.*

I.6 Choosing the right sequencing datasets and limitation

The immense diversity of microbial species makes cataloging them a complex task (Whitman et al., 1998). Many species are often missed in sampling efforts, leading to biases and an incomplete understanding of microbial community structures (Green et al., 2006). Additionally, our knowledge of the factors influencing bacterial distribution across environments remains limited. For example, entire bacterial groups are often broadly categorized as "marine" or "terrestrial," which may oversimplify their ecological roles. This raises questions about whether such categorizations reflect true environmental adaptations or simply gaps in our knowledge. To address this, standardized descriptions of habitats are essential. Initiatives like the Genomic Standards Consortium's MIGS/MIMS (Minimum Information about a Genomic/Metagenomic Sequence) specifications aim to improve environmental metadata in research (Field et al., 2008).

Despite these challenges, modern research has made significant strides in exploring microbial diversity across ecosystems. By focusing on well-sampled environments, researchers can better study microbial taxa and overcome some of these limitations. In this thesis, we address the challenge of identifying bacteria labeled as "unknown" or "unclassified" at the species level. Such classifications are common in environmental samples, where databases like NCBI often lack detailed taxonomic information, while resources like SILVA and RDP provide more refined classifications (Hinchliff et al., 2015). To bridge this gap, we aim to develop a framework for identifying and tagging these "unknown" species by analyzing their distribution across different environments and linking them to ecological factors like soil type or water quality.

Beyond SILVA, other datasets like the Genome Taxonomy Database (GTDB) and NCBI have become valuable resources for studying microbial taxonomy. Studies by Smith et al. (2014) and Parks et al. (2018) have used these datasets to explore bacterial and archaeal diversity, providing genome-based taxonomies that complement traditional approaches (Parks et al., 2021). Integrating insights from these diverse datasets enhances our ability to understand microbial distribution patterns across environments.

Since beginning our thesis in 2017, we have focused on the ARB-SILVA database, a key resource for rRNA alignment and quality control (Pruesse et al., 2007; Quast et al., 2013). We

use SILVA version 132, released in 2017, which contains over 6 million aligned sequences—a significant increase from the 461,823 sequences in version 91 (2007). This version includes high-quality, nearly complete sequences, filtered to exclude chimeras and ensure data integrity. Our choice of the SILVA SSU Parc dataset reflects our commitment to precision in microbial community analysis.

Our research methodology involves sampling, extracting bacterial DNA, targeting the 16S rDNA, and analyzing microbiota using the SILVA SSU Parc dataset as a reference. This dataset serves as the foundation for developing a tracking system to identify microbial sources, whether from hosts or environments. This approach represents a shift toward more refined and comprehensive methods for studying microbial communities.

Building on the SILVA datasets, the next phase of our research focuses on understanding microbial habitat interactions. By leveraging advanced classification systems that capture the physicochemical properties of habitats, we aim to deepen our insights into microbial ecology. This will help refine ecological models and inform strategies in environmental and applied microbiology.

I.7 Exploiting microbial habitat information

Understanding microbial habitats requires a detailed and adaptable classification system that meets the needs of modern microbial ecology research. Such a system must balance breadth and depth, covering the wide diversity of microbial environments while distinguishing their unique physicochemical properties. This depth is essential for uncovering the complex relationships between microorganisms and their surroundings. Additionally, the system should be dynamic, allowing for regular updates, easy maintenance, and automated use to keep pace with new discoveries and technological advancements.

A well-designed classification system should align with the goals of microbial biodiversity studies. It must be user-friendly for microbiologists, grouping similar environments to enable meaningful analysis. By organizing habitats into logical clusters, researchers can better understand the shared characteristics and ecological dynamics of specific environments.

In comparative analysis, metadata plays a crucial role. Metadata provides context for microbial samples, including details like geographical location, sampling time, procedures, and environmental conditions. This information is vital for interpreting microbial community composition and characteristics, especially in metagenomics. While structured text descriptions are useful, they often fall short of capturing the full richness of metadata. Therefore, improving metadata standards is essential for unlocking its potential. A well-structured metadata framework supports cross-study comparisons, meta-analyses, and predictive models, enhancing our understanding of microbial ecology and its broader applications. As technology advances, metadata standards must evolve to remain relevant and useful.

Building on the importance of metadata, the next challenge is standardizing terminology and sample annotations. Consistent and accurate metadata is critical for organizing, retrieving, and comparing information about microbial habitats. Harmonizing content, syntax, and terminology will improve data interoperability and support the growing need for reliable, reusable data in microbial ecology. This standardization lays the foundation for future advancements in the field.

I.8 Ontology Work: Classification of Terms and Relationships

Ontologies are formal, structured frameworks that represent knowledge in a specific domain by defining concepts (terms), their properties, and the relationships between them (Ashburner et al., 2000; Smith et al., 2007). In microbial ecology, ontologies standardize and organize metadata, enabling consistent annotation, data integration, and interoperability across studies. Key components of ontologies include terms (e.g., "soil," "human gut"), relationships (e.g., "part_of," "located_in"), hierarchies (e.g., "environment" > "aquatic environment"), and attributes (e.g., pH, temperature) (Buttigieg et al., 2013a). Creating an ontology involves domain analysis, term definition, relationship establishment, hierarchy construction, validation, and integration with tools and databases (Jensen et al., 2012).

To address the need for consistent and detailed metadata, the microbial and molecular ecology communities have developed standards like MIMARKS (Minimum Information about a MARKer gene Sequence) (Field et al., 2008). MIMARKS provides guidelines for annotating microbial samples with comprehensive metadata, ensuring standardized content, syntax, and terminology. This standardization is crucial for organizing and retrieving information about sample origins. Additionally, harmonizing reference keywords for bacterial habitats supports the creation of databases and tools that use sequence alignment to infer sample sources. This aligns with the FAIR data principles, which emphasize making data findable, accessible, interoperable, and reusable.

One well-known ontology is the Environment Ontology (ENVO) (Buttigieg et al., 2013a), which provides a comprehensive framework for describing environmental biotopes, such as "soil," "marine environment," and "freshwater ecosystem." Similarly, OntoBiotope (Karadeniz & Özgür, 2015b) focuses on microbial habitats related to food and environmental samples.

Text mining tools, such as BioNLP (Robert Bossy, 2019) and Seqenv (Sinclair et al., 2016a), leverage these ontologies to extract and structure biotope information from scientific publications. These tools use machine learning and ontological dictionaries to identify bacterial habitats, including those related to food (Chaix et al., 2019). However, they face limitations when handling uncultivable bacteria, which make up over 90% of bacterial diversity. Similarly, Seqenv relies on raw text from publications, which can lead to mislabeled sequences due to a lack of standardized biotope terms. Another tool, FORENSIC, offers a practical solution for identifying fecal sources in water quality testing by using a library of human and animal

reference signatures, eliminating the need for users to create their own databases (Roguet et al., 2020).

Despite their utility, existing ontologies like ENVO and OntoBiotope primarily focus on environmental biotopes, leaving a gap in understanding bacteria associated with host organisms.

However, text mining approaches face challenges such as spelling errors, homonyms, and inconsistencies in terminology, which are common in biological contexts (Jensen et al., 2012). These limitations exclude uncultured microorganisms and can lead to confusion. Despite these challenges, our work aims to improve data quality and research outcomes by standardizing metadata and developing reliable tools for inferring bacterial habitats.

I.9 Bayesian Probability: Philosophy and Application in Microbial Source Tracking

Bayesian probability is a statistical framework that interprets probability as a measure of belief or confidence in an event, updated as new evidence becomes available (for an introduction on bayeseian network in computational biology, see Needham, et al 2007). Unlike classical ("frequentist") probability, which defines probability as the long-run frequency of events, Bayesian probability incorporates prior knowledge (a "prior") and updates this belief with observed data to produce a posterior probability. This approach is formalized in Bayes' theorem:

$$P(H|D) = \frac{P(D|H).P(H)}{\mathrm{P(D)}}$$

where P(H|D)P(H|D) is the posterior probability of hypothesis H given data D, P(D|H) is the likelihood of observing D under H, P(H)is the prior probability of H, and P(D) is the probability of the data.

Bayesian methods emphasize subjectivity in probability, acknowledging that prior knowledge and assumptions influence conclusions. This contrasts with frequentist approaches, which rely solely on observed data and avoid incorporating prior beliefs. In microbial source tracking, Bayesian frameworks are particularly powerful because they allow researchers to integrate existing ecological knowledge (e.g., known host-specific markers or environmental distributions) with new sequencing data to objectively infer contamination sources. By iteratively updating probabilities as more data is collected, Bayesian models reduce bias and adapt to evolving evidence, making them ideal for complex, noisy biological systems.

Bayesian probability underpins many microbial source tracking tools, such as SourceTracker (Knights et al., 2011), which estimates the proportional contribution of different sources (e.g., human, animal, soil) to a microbial community. For example, if a water sample contains Bacteroides sequences associated with humans, the Bayesian model combines prior knowledge of Bacteroides prevalence in human feces with observed sequence abundances to calculate the posterior probability that the contamination originated from humans. This approach quantifies uncertainty, providing confidence intervals for source attributions a critical feature for environmental monitoring and forensic investigations. Bayesian models also address challenges

like incomplete reference databases or low-biomass samples by probabilistically assigning sequences to the most likely sources, even when exact matches are absent. This flexibility is invaluable in studying uncultured bacteria or novel environments, where traditional methods may fail.

While Bayesian methods dominate source tracking, other probabilistic frameworks are also used:

1. **Frequentist Probability:** Focuses on maximum likelihood estimation (MLE) without incorporating prior knowledge. Tools like FEAST (Shenhav et al., 2019) use MLE to estimate source contributions but lack Bayesian adaptability to update hypotheses with new data.
2. Machine Learning (ML) Models:
    - **Random Forest:** A non-Bayesian ML algorithm that constructs decision trees to classify sources based on sequence features (e.g., marker gene abundances). While powerful for pattern recognition, it treats probabilities as fixed frequencies and does not integrate prior ecological knowledge.
    - **Support Vector Machines (SVM):** Maps sequences to high-dimensional spaces to classify sources but offers limited interpretability compared to Bayesian posterior probabilities.
3. **Hidden Markov Models (HMMs):** Used in tools like PhyloSource (Martiny et al., 2006) to model transitions between microbial states across environments but require predefined transition probabilities, unlike Bayesian adaptive priors.

Bayesian methods excel in microbial source tracking by quantifying uncertainty (e.g., confidence intervals for contamination sources) and handling incomplete data. They probabilistically assign sequences to likely sources even without exact database matches, aiding studies of uncultured bacteria. In contrast, frequentist tools like FEAST use maximum likelihood estimation but lack Bayesian adaptability. Machine learning models (e.g., Random Forest, SVM) classify sources based on patterns but treat probabilities as fixed frequencies, ignoring prior ecological knowledge. Hidden Markov Models (e.g., PhyloSource) require predefined transition probabilities, unlike Bayesian adaptive priors. Hybrid approaches, such as combining Bayesian inference with machine learning, could enhance predictive power while retaining interpretability. Table 2 below summarizes the

key strengths and limitations of Bayesian frameworks in microbial source tracking, balancing their adaptability and objectivity against inherent challenges.

Table 2. Advantages and Challenges of Bayesian Approaches in Microbial Source Tracking

Advantages	*Limitations*
Objectivity: Explicit priors reduce bias in ambiguous data interpretation.	**Prior Sensitivity:** Incorrect priors (e.g., outdated host-specific markers) skew results.
Uncertainty Quantification: Posterior probabilities provide confidence metrics for predictions.	**Computational Complexity**: Resource-intensive for high-dimensional datasets.
Adaptability: Priors update with new data refining hypotheses as databases expand.	

I.10 Source Tracking Tools: advantages and disadvantages

Advances in bioinformatics have led to the development of numerous tools for microbial source tracking, each designed to address specific challenges in analyzing microbial communities and studying host-microbiome interactions. While widely used tools like SourceTracker and MetaPhlAn remain central to the field, newer frameworks such as SeqEnv, BioNLP, FORENSIC, and GenBank-hosted datasets have expanded the scope of microbial source identification and ecological analysis.

SeqEnv matches microbial sequences to environmental parameters using reference databases, linking microorganisms to their ecological niches. This approach is particularly useful for environmental microbiology but is limited by the quality and completeness of the metadata in reference databases (Sinigalliano et al., 2019).

BioNLP uses natural language processing (NLP) to analyze metagenomic data, extracting annotations and mapping microbial communities to their sources. While effective for synthesizing large datasets, its reliance on annotated literature and structured ontologies limits its use in novel or poorly characterized environments (Fang et al., 2017).

FORENSIC is a specialized tool for microbial source attribution, using machine learning and statistical modeling to improve prediction accuracy. It is highly effective in forensic microbiology, such as criminal investigations and environmental contamination tracing. However, its need for extensive training data and computational resources can make it less accessible to some researchers (Clancy et al., 2021).

GenBank, hosted by the NCBI, is a key resource for studying host-microbiome interactions. A recent study by Ramanan et al. (2022) demonstrated its ability to correlate host-microbiome data across diverse environments through comparative genomic and metadata analyses. However, the accuracy of its analyses depends on the quality and completeness of the submitted sequences and metadata.

Together, these tools (summarized in Table 3) form a robust and evolving framework for microbial source tracking. They enable applications in public health, ecology, and forensic research. When combined with traditional techniques like PCR and metagenomic sequencing, these bioinformatics tools provide deep insights into microbial community structures and their origins, advancing our understanding of microbial ecology and host-microbiome dynamics.

Table 3. Summary of the most bioinformatic tool tracking the source of bacteria with their advantages and disadvantages.

Tool	*Advantages*	*Disadvantages*	*Type of Data Used*	*Citation*
SourceTracker	- Estimates proportional contributions from various sources - Broad applicability	- Requires a diverse and comprehensive reference database - Struggles with rare microorganisms	16S rRNA, metagenomic data	Knights et al., 2011
MetaPhlAn	- High taxonomic resolution (species/genus level) - Marker-based detection	- Limited ability to identify novel microbes - Does not directly quantify environmental or host contributions	Marker genes, metagenomic data	Segata et al., 2012
FEAST	- Computationally efficient - Works with minimal training data	- Poor performance in resolving highly similar microbial communities	16S rRNA, metagenomic data	Shenhav et al., 2019
PhyloSource	- Integrates phylogenetic data - Robust even with limited datasets	- High computational demands - Less user-friendly than alternatives	Phylogenetic markers, metagenomic data	Martiny et al., 2006
Kraken2/Bracken	- Fast and accurate classification - Scalable for large datasets	- Dependent on completeness of reference genomes - Limited in poorly studied environments	Whole-genome sequencing, metagenomic data	Wood et al., 2019; Lu et al., 2017
mlST	- Data-driven insights using machine learning - Learns patterns for improved accuracy	- Requires extensive training data - Complex implementation and analysis	Genomic, metagenomic, and metadata	Ramazzotti et al., 2019
SeqEnv	- Links sequences to environmental metadata - Useful for niche-specific tracking	- Limited by the resolution and completeness of environmental databases	Environmental DNA, metagenomic data	Sinigalliano et al., 2019
BioNLP	- Leverages natural language processing - Integrates annotations and metadata	- Relies on annotated literature and structured ontologies - Gaps in poorly described environments	Textual data, annotated literature	Fang et al., 2017
FORENSIC	- Effective in forensic microbiology - Machine learning enhances precision	- Requires extensive training datasets - Computational expertise needed	Genomic, metagenomic, and forensic data	Clancy et al., 2021

GenBank (Ramanan, 2022)	- Comprehensive host-microbiome dataset - Enables comparative genomics and metadata integration	- Dependent on quality and completeness of submitted sequences - Analysis accuracy varies with metadata quality	Genomic, metagenomic, and metadata	Ramanan et al., 2022

Chapter II
Thesis Objectives

Thesis Objectives

The challenges inherent in source tracking, particularly within the realm of microbial source tracking (MST), are multifaceted and demand a nuanced approach. The complexity of environmental matrices, the diversity of microbial communities, and the dynamic nature of microbial interactions all contribute to the intricacies of accurately pinpointing the origins of bacterial contamination.

In light of these challenges, the overarching aim of this thesis is to evaluate the contribution of molecular microbiology techniques to the field of source tracking. Specifically, it seeks to ascertain the degree to which metagenetic method can elucidate the sources of bacterial presence.

To this end, the proposed strategy makes use of the widespread 16S rDNA amplicon sequencing, targeting a phylogenetic marker ubiquitously present across the bacterial domain. By analyzing the sequences retrieved, one can discern the identity and relative abundance of bacterial taxa within a given sample.

The initial phase of this strategy entailed the construction of a database “BiotopeBac” that correlates 16S rDNA sequences, sourced from GenBank, with corresponding information on their origins and biotopes, where available. The utility and validity of this database are critical for its integration with metagenetic analyses.

Proceeding from a novel hypothesis, the second phase questions whether a 16S rDNA amplicon profile can reveal its source matrix. This segment of the research delves into the broader concept of bacterial dispersal in the environment. A newly developed computational tool facilitates the swift determination of a 16S rDNA amplicon profile’s source.

The final segment of the research applies a combined 16S rDNA metagenetic-database approach to pinpoint various potential sources of bacterial contamination. This is exemplified in two case studies: the first investigates the origins of fecal contamination within the sanitary facilities of a veterinary faculty, and the second assesses the water quality of multiple riverine bathing sites in the Walloon Region, with a focus on fecal pollutants.

This thesis, therefore, stands at the intersection of molecular biology and environmental science, endeavoring to advance our understanding of microbial ecology and the practical applications of MST in assessing environmental health.

Chapter III

Database and Software: development and validation

III.1 Introduction

Building on the challenges presented in Chapter I, this project addresses the need for a comprehensive and dynamic classification system tailored to microbial habitats. Such a system must not only capture the vast diversity of bacterial communities but also effectively distinguish their environmental properties. Additionally, it must be adaptable, supporting regular updates and the automated incorporation of new data as scientific knowledge and technologies evolve.

A key element in this endeavor is the accurate annotation of microbial samples with detailed metadata. Critical contextual information such as geographical location, sampling techniques, and environmental characteristics allows for deeper comparative analysis of microbial communities. By enhancing metadata standards, we aim to streamline data analysis, improve cross-study comparisons, and support meta-analyses and predictive model development. ontologies like Environment Ontology (ENVO), OntoBiotope, and Bacteria Biotope (BB) play a pivotal role in organizing and retrieving biotope-related information. However, these tools are predominantly focused on environmental habitats, leaving a significant gap in the representation of host-associated bacterial ecosystems.

To bridge this gap, we developed “BiotopeBac”, a specialized database designed to catalog bacterial-habitat associations by linking annotated sequence data with corresponding biotope metadata. The primary goal of BiotopeBac is to create an open-access resource that connects reliable biotope information with bacterial taxonomic data derived from 16S rDNA and amplicon sequences. By intertwining these essential components, the database enhances our ability to explore bacterial diversity across varied environments.

In this chapter, we outline the methodologies employed to construct BiotopeBac, beginning with the **Preparation Phase**. This phase involves downloading raw sequence data and metadata, manually annotating terms, and constructing the relational database structure. The **Dereplication Phase** follows, which includes filtering and clustering the data to ensure accuracy and consistency. In this section, we also introduce the Metagenomic Source Tool (MGST), which was developed to analyze the data effectively.

The chapter then transitions into the database exploitation and validation section, where we discuss how the database was applied to selected BioSamples, the pre-processing of sequence reads, and the classification of samples. We also describe the construction and use of the

Sequencing Tracking tool (ST), developed to facilitate tracking bacterial sequences through different environments. This tool, alongside MGST, forms the backbone of our biotope identification methods, further detailed in later sections.

Our methodology heavily relies on 16S rDNA sequencing, a well-established approach for bacterial taxonomic identification. The choice of the 16S gene as a marker is deliberate, as it allows for the standardized comparison of bacterial populations across various studies. However, relying solely on species labeling can introduce errors due to misidentifications or database inconsistencies. To address these issues, we propose bypassing species-level identification when sequence similarity is low and instead linking sequences directly to their biotope.

The remainder of this chapter delves into the technical aspects of constructing the database. We begin with the collection of 16S rDNA sequences from public databases such as SILVA and GenBank, followed by the extraction of environmental metadata. This information was organized using a controlled vocabulary to ensure consistency across the dataset. We then describe the process of building the relational database using SQL and the subsequent development of tools for database exploitation. Finally, we validate the database through application to specific datasets, including restroom and natural bathing water environments.

In summary, this chapter offers a detailed overview of the steps taken to construct and validate BiotopeBac, alongside the tools developed to facilitate microbial ecology research. The insights gained from these analyses lay the groundwork for advancing our understanding of bacterial diversity and biotope associations, which are explored further in the Results and Discussion chapters that follow.

III.2. Material and methods

III.2.1 Preparation phase and BiotopeBac database construction

III.2.1.1 Download raw sequence data

Our initiative commenced with the SSU Parc files sourced from the arb-SILVA sequence collection. This dataset, already annotated and quality-checked, encompasses redundant sequences identical sequences with distinct metadata. Our primary objective was to preserve authentic bacterial 16S rDNA sequences within the SILVA database, excluding chimeric sequences (formed from two different template DNAs due to a mismatch during PCR amplification) based on sequence quality scores, and with relevant source information or links to scientific publications.

We successfully retrieved SSU Parc files in fasta format version 132 from the ARB website (https://www.arb-silva.de/documentation/release-132/), witnessing an augmentation in the sequence count to over 6,073,181. Through careful curation, we extracted and retained all sequences attributed to the bacterial kingdom. At the conclusion of this meticulous process, our dataset included crucial information such as Accession numbers, taxonomy paths, and fasta files or sequences (refer to Figure 8). command line related to this phase detail in Appendix 1.

Figure 8. Silva data fields extraction for database construction workflow. *This figure illustrates the extraction and retention of fields from Silva data essential for the subsequent stages of the database construction workflow.*

III.2.1.2 Download metadata

In the subsequent step, we gathered the Accession numbers associated with each retained sequence from the prior stage. These Accession numbers were instrumental in querying and downloading the corresponding XML files from EMBL-Genbank database. Figure 9 provides a visual representation of a sample XML file, offering insight in its structure and content.

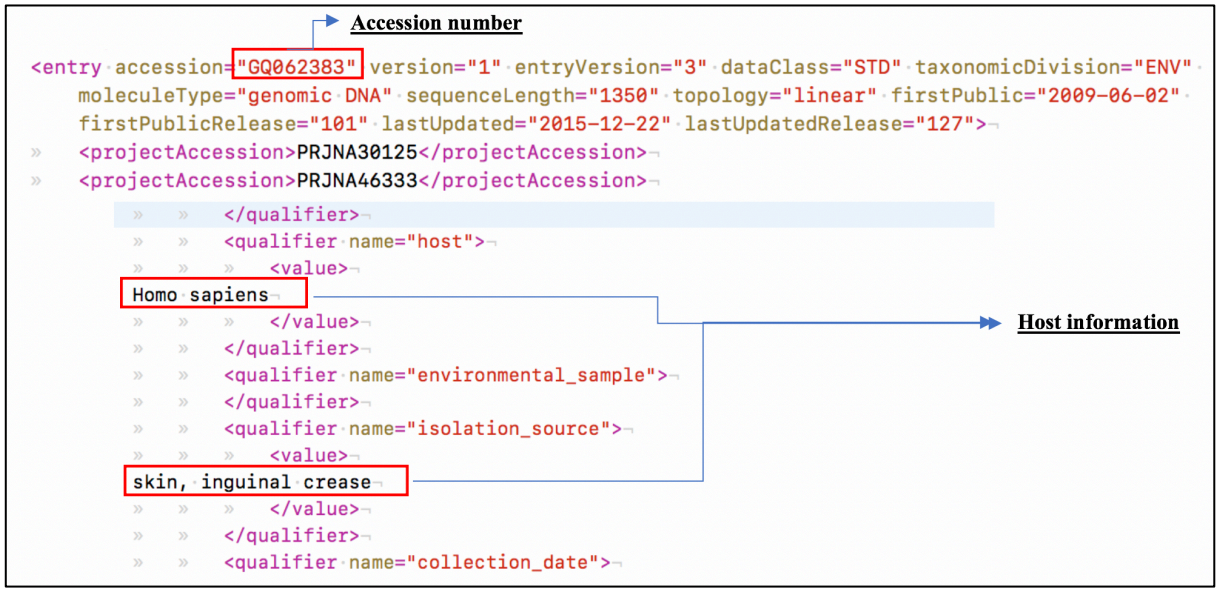


Figure 9. Structure of the XML file and showcases the extraction of host information values from the dataset. *The XML file format, highlighting the key fields with a specific focus on the extraction process of host information, is a critical aspect of our database construction.*

In our data retrieval process, we utilized a specialized bash script, outlined in detail in Appendix 2, to systematically download XML format files from the extensive EMBL-Genbank database. To outline our methodology, we commenced with a dataset encompassing 6,073,181 entries from SILVA's 16S rDNA gene (SSU) Park collection. Our approach focused on selectively retaining sequences with host information, delineated in Figure 2. Unfortunately, 719,945 sequences lacking essential information were excluded during this curation process. Consequently, our dataset underwent refinement, resulting in 5,353,236 sequences that contain relevant information or publication. The aim was to capture information for all 5,353,236 accessions within this vast repository. Following the execution of our script, the outcome was a comprehensive dataset comprising 26,754 XML format output files. Each of these files encapsulated a grouping of 200 complete Accession records, creating a structured and organized resource for our subsequent analyses and investigations.

We curated the data obtained from the XML files, employing a bash script programmed for this specific purpose (refer to Appendix 3). Our focus during this process was on refining the dataset by filtering out any inappropriate or irrelevant information. The extracted fields from the XML files encompassed a range of essential details, including accession numbers, taxonomic divisions (such as BAC, VEC, FUN, EUK), molecule types (genome, DNA sequence, vector), sequence lengths, first publication and last updated dates, study or publication titles, DOI numbers, Pubmed accession numbers (PMID) for associated publications, and host notes. Additionally, where applicable, information from the reference source fields was also included in our refined dataset. This careful curation ensures the dataset's precision and relevance for subsequent analyses.

Within this curated dataset, sequences with host/source information were systematically organized based on associated publications. In cases where host/source details couldn't be deduced from the publication, we preserved the associated project title as a valuable reference for understanding the origins of the sequences deposited on those databases. Within the dataset of 5,353,236 entries, a substantial 95% of the data included relevant source information (Accession number, taxon field, Data of first publication, Title associated to publication, DOI link, Pubmed ID, host information) as shown in figure 10.

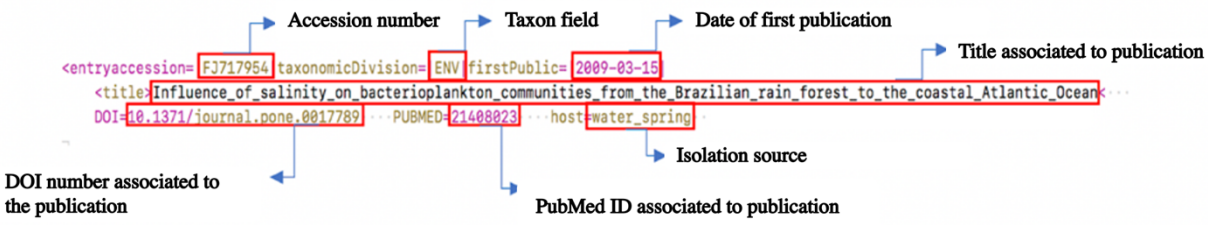


Figure 10. An example of metadata collection from an EMBL file. *The top section of the figure shows the header of the file, which includes general information, such as the accession number and the title of the publication. The bottom section shows the PubMed ID and other important information, such as the isolation source. The metadata collected from these sections can be used to characterize the microbial community and identify the potential sources of environmental samples.*

Only 236,449 sequences were excluded as they lacked pertinent data after the first treatment of those downloaded XML files. Figure 11 visually outlines the workflow, culminating in a refined set of 5,116,787 bacterial sequences with host information sourced from Genbank.

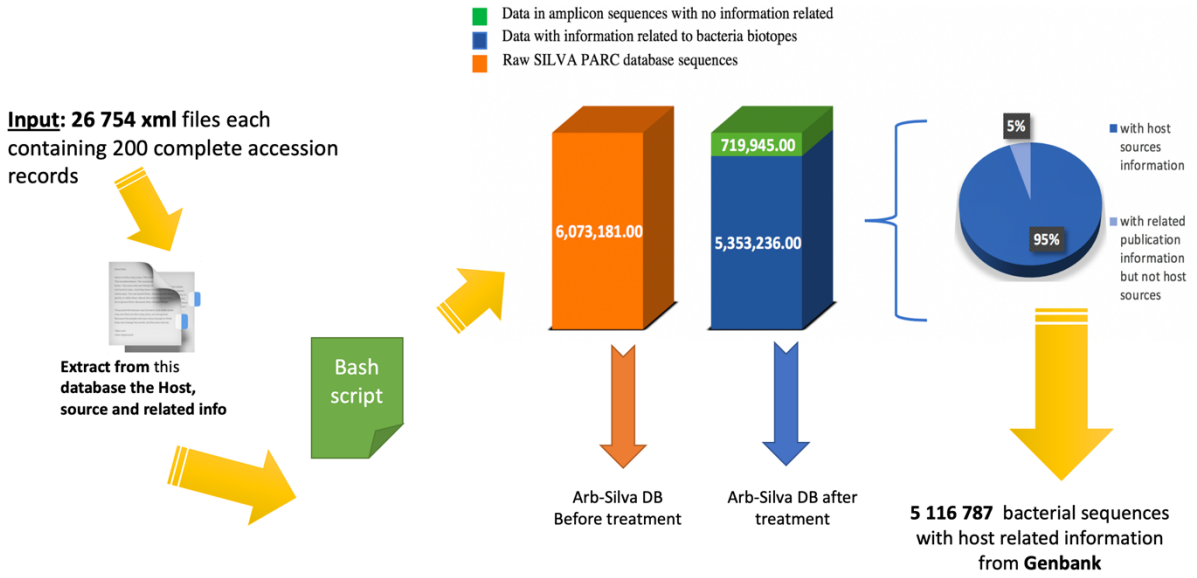


Figure 11. Workflow illustrating the data curation process. *Beginning with a dataset of 5,353,236 entries, the curation process focused on retaining sequences with relevant source information. Sequences lacking such information (236,449 entries) and identified solely by title publication were excluded. The final output comprised 5,116,787 bacterial sequences with host information sourced from Genbank.*

III.2.1.3 Manual terminology annotation

At this juncture, the metadata associated with 16S sequences is in its raw form, having been directly extracted from Genbank. The proliferation of data across diverse web-based platforms poses a challenge, as sourcing information from different datasets often corresponds to several disparate entities. Although a controlled metadata vocabulary, such as the Minimum Information about a MARKer gene Sequence (MIMARKS) system, has been developed for Next-Generation Sequencing (NGS) datasets, it is not obligatory for Sanger DNA databases like Genbank. The unique value proposition of our database lies in the integration of metadata that is formatted to be readily recognized by an associated ontological dictionary. To achieve this, we undertook meticulous control and processing of the metadata to ensure correct structuring.

In our literature review, we identified several studies employing an equivalence search of semantic terms, based on the Environmental Ontology (ENVO) dictionary, which is widely employed in the field of biology (Buttigieg et al., 2016; Walls et al., 2012). However, a limitation of this dictionary is its simplistic semantic identity search, where a term is either recognized or not, without the capability for taxonomic classification (Smith et al., 2007). This inherent limitation underscores the need for a more sophisticated approach to semantic searching in our context, as noted in various studies (Jupp et al., 2016).

We devised a systematic approach grounded in ontological reasoning and linguistic analysis of terms embedded in associated metadata. Our initial efforts focused on synthesizing synonyms and ontological paths to construct a comprehensive reference file. However, complications arose with sequences linked to hosts or clinical metadata, as they were inadequately annotated by existing ontologies like ENVO. Additionally, automatic terminology evaluation overlooked issues like spelling errors and the presence of ambiguous and associative terms (e.g., "sea horse," which, unlike "sea" and "horse," refers to the host and carries a distinct meaning).

We considered two types of terms: **irrelevant terms** (undesirable terms) and **reference terms** (words to be converted to ontological terms). The **irrelevant terms** were composed of: **stop words** (e.g., "a", "about", "above", "after"), **non-relevant words** (e.g., "alpha", "api", "closely"), and **microbial terms** (e.g., "*Acetobacterium*", "anaerobic", "antibiotic", *Candidatus_Carsonella*") present in the title of the publication or the information about the isolation source. Controlled terms were composed of **associative terms or paired terms** (e.g.,

"guinea pig", "jack mackerel", "horse weed"), **entity synonyms and spelling errors** (e.g., "feces", "fece", "faecal", "faeces", "fecal", "stool", "excrements", "excrement", "faces", "fecess") (see Figure 12).

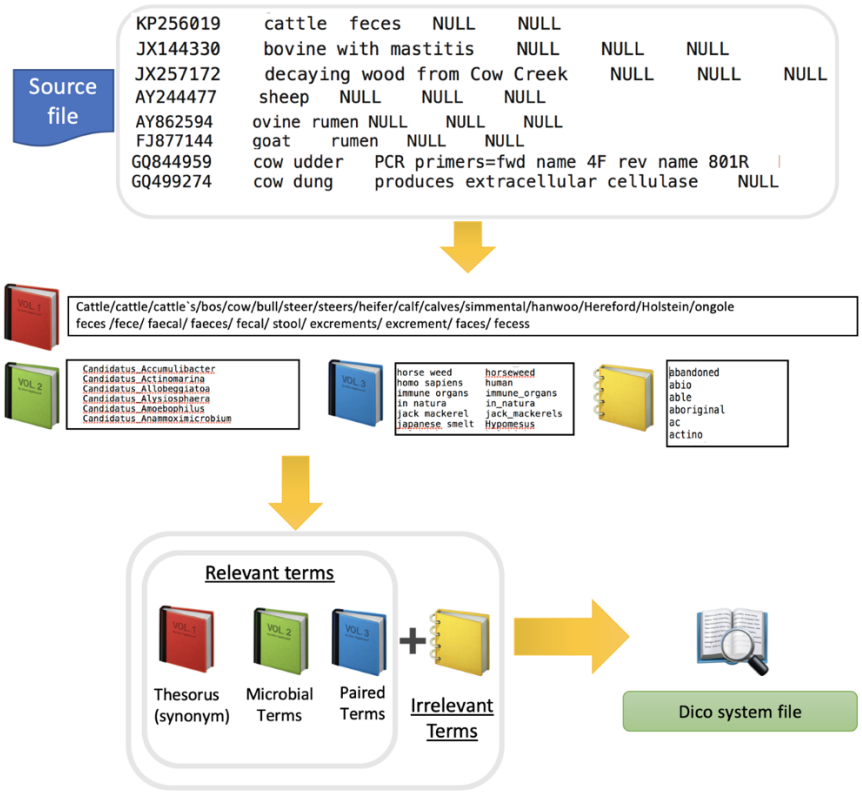


Figure 12. A workflow of our terminology construction process, showcasing the representation of various file types and providing examples of each. *Additionally, it highlights the seamless integration of these file types into our comprehensive dictionary file system.*

Verification of metadata and the establishment of a controlled vocabulary strictly followed the prescribed OBO file format or **Dico system file**, as illustrated in Figure 13. This format encompasses the following key components: **id** (term number), **name** (original term), **def** (terminology or taxonomic classification), **synonym** (alternative terminology, spelling errors, or vernacular names), **is_a** (term format), and **relationship** path (encompassing the complete path of the ontological term).

The culmination of this process involved predominantly utilizing PERL programming, as detailed in Appendix 4, to create our final ontological reference file.

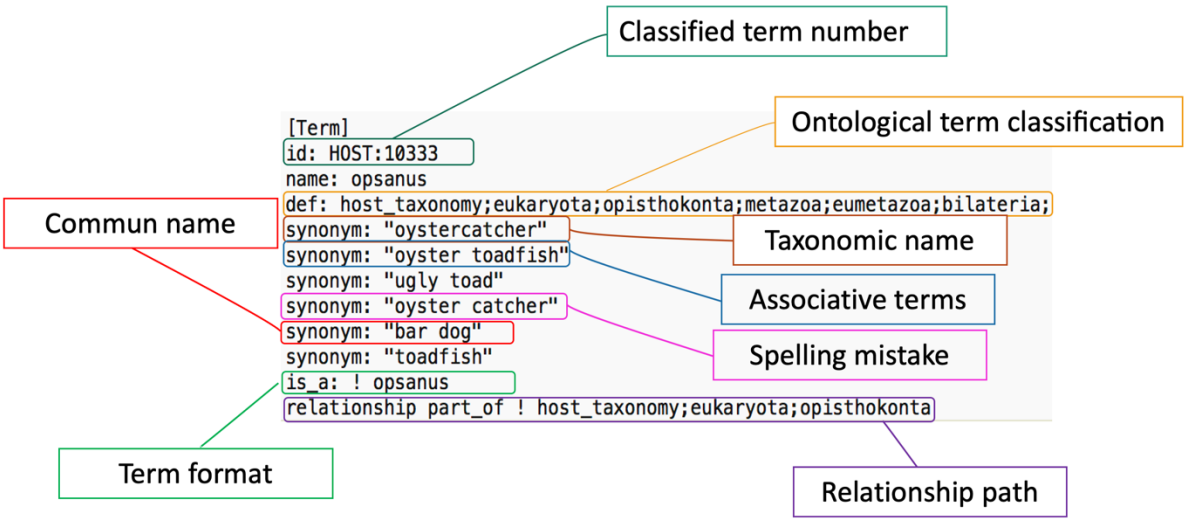


Figure 13. An illustrative example of the reference file used for metadata formatting. *It is accompanied by the curated dictionary file, which standardizes and organizes the data.*

Our ontological annotation strategy underwent a comprehensive three-step evolution. Initially, we examined the metadata to construct a specialized database featuring a controlled vocabulary. Subsequently, we developed a controlled vocabulary comprising word-set terms (FLAGs) that intricately linked previously formatted ontological terms under a more general semantic definition. These word-set terms effectively represented taxonomic categories and grouped host families from the animal, plant, and human kingdoms, streamlining source searches and seamlessly aligning with the structure of metagenetic profiles. The transformation of metadata was executed through a PERL script, created in our laboratory (refer to Appendix 5).

The primary objective of this file is to convert raw metadata into well-referenced keywords. Comprising a total of 99,333 keywords, each with six distinct fields, the file's information content falls into two primary categories. Firstly, we handle nomenclatural information, encompassing scientific names, etymology, related names, and other pertinent details, all classified under the "Organism Name" category. These terms are further organized into word-sets, each denoted by the prefix "**Org::**". Secondly, environmental information is included, categorized under "Environmental Information," with terms also grouped into word-sets marked by the prefix "**Env::**".

The final and concluding step involved processing the input file, incorporating all host sources terms associated with sequence information. This was carried out using a documented file

named “Biotopebac.obo”, which contains curated term alignments with a reference terminology, as described in the previous steps (see Appendix 6).

Initially comprising 5,116,787 sequences with raw host source data predominantly labeled with "sequence" and "dna" terms, our dataset underwent effective filtration. Following the utilization of the OBO file “BiotopeBac.obo”, this process resulted in the successful isolation of 4,264,470 sequences associated with curated keywords, specifically those labeled "human" and "soil" as the prevailing terms. A subsequent, more refined selection process further distilled the dataset, retaining 3,195,698 sequences annotated with curated word-sets. The outcomes of this selection process are visually depicted in Figure 14.

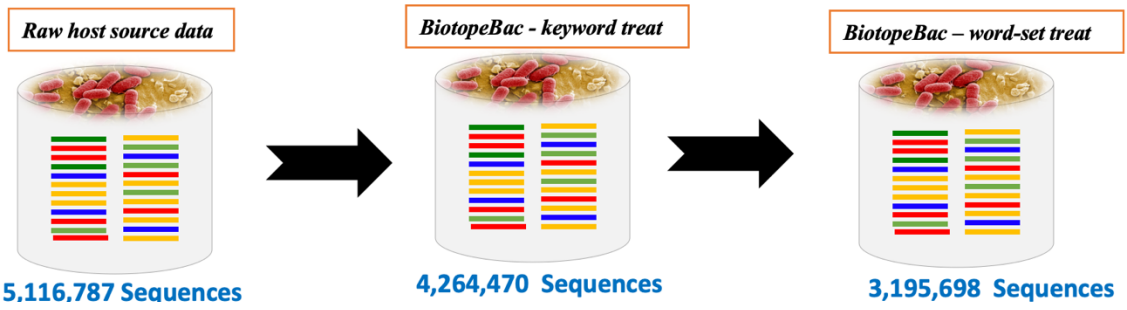


Figure 14. Dataset refinement. *Illustrates the stepwise refinement of the initial dataset (5,116,787 sequences) with raw host source data and 4,264,470 sequences were obtained, emphasizing curated keywords and Further refinement yielded 3,195,698 sequences with annotated word-sets.*

In the following paragraph, we will elucidate our approach to structuring the database, ensuring comprehensive coverage of information pertaining to downloaded sequences and their associated host information.

III.2.1.4 Creation of the relational database structure

Our database constitutes an extensive repository of information, necessitating meticulous structuring and storage on a physical medium. The intricacies of data retrieval, involving multiple interacting criteria, make it a complex operation. The use of an apt language is paramount, and SQL (Structured Query Language) emerges as the widely adopted choice.

Consequently, we employed MySQL as our data structuring tool to effectively manage and navigate this wealth of information.

Information was organized in relation to a singular identity, involving a transformation that eliminated or minimized redundant occurrences of the same information across multiple lines. This restructuring resulted in the uniqueness of each line, often accompanied by the inclusion of a numerical label or identifier. This approach enhanced the efficiency of data referencing in a tabular format.

The organization of data into multiple independent tables not only facilitated the preservation of referencing links between them but also streamlined the process of searching for specific elements. Figure 15 illustrates the nature of these links, distinguishing between direct and indirect connections with the entities gathered for database construction.

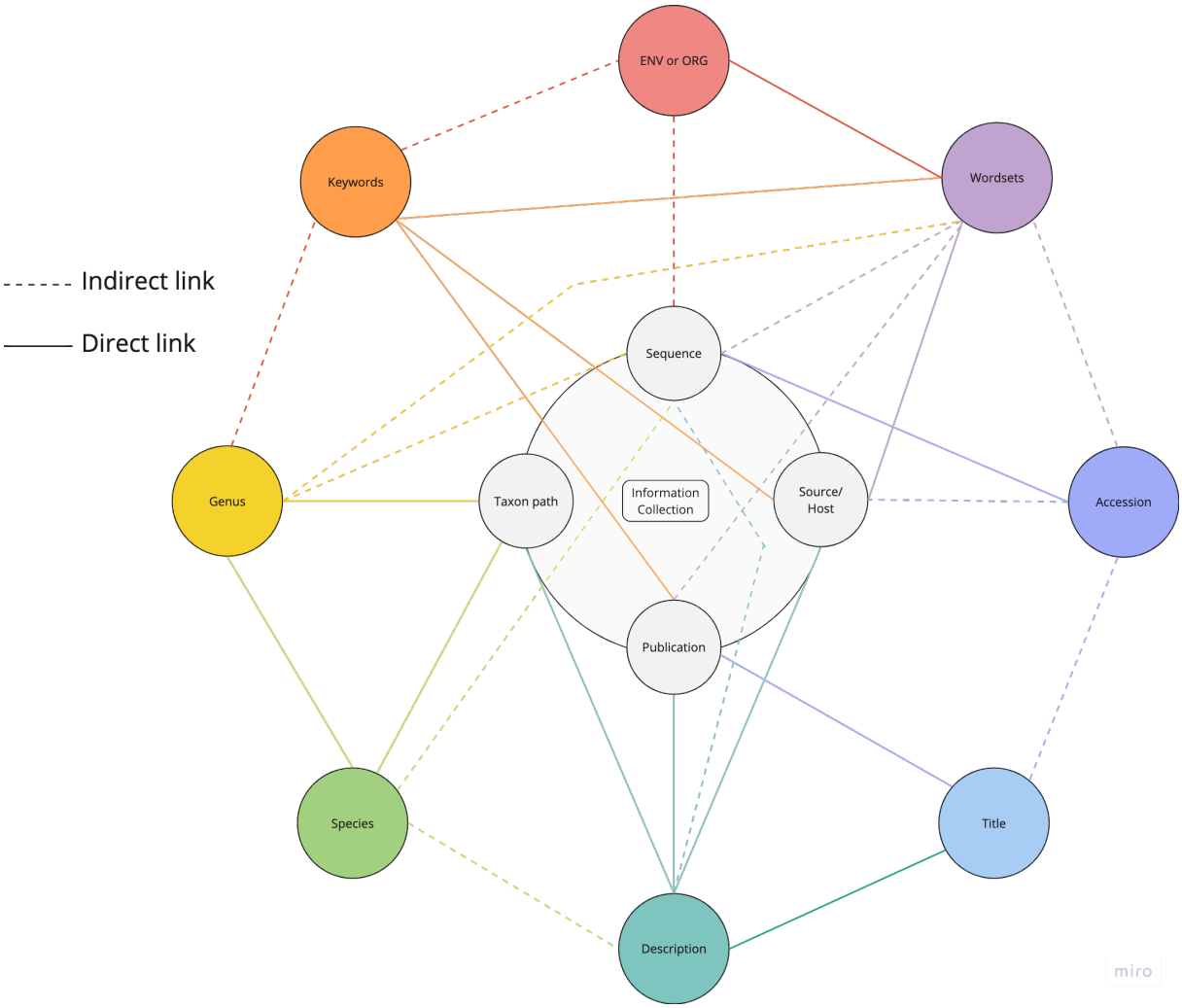


Figure 15. Database relationships. *Illustrates the structured relationships between entities in the database. Direct and indirect links are highlighted, showcasing the interconnected nature*

of the entities collected for database construction. This visualization aids in understanding the intricate web of associations within the data modelusing miro tool .

To elevate clarity and flexibility in portraying information and relationships across tables, we formulated a model leveraging the concept of entities and associations within a conceptual database schema. These entities are equipped with attributes and keys, essential components for establishing meaningful relationships. This method facilitates a more intuitive depiction of information and relationships, eliminating the initial need to worry about the precise structure in terms of tables and columns.

- **Entities**

Entities refer to identifiable groups of elements within the same application or context domain, forming the foundation for our database establishment. In our approach, data is compartmentalized into entities or tables, each containing specific elements known as attributes (see Figure 16).

The database structure hinges on the separation of information based on context and the inherent relationships between data elements. For instance, the "Sequence" entity encompasses data directly associated with sequences along with their respective accession numbers. This deliberate structuring enhances the clarity and relevance of the information within each distinct entity.

In our conceptual database diagram, we have delineated five distinct entities:

- **"Sequences" entity:** This entity encompasses all nucleotide sequences, linked with their accession numbers, identifier numbers for taxonomic affiliation, and details regarding the origin of the sequence sample.
- **"Keywords" entity:** Within this entity, terms are stored in accordance with the reference file, alongside their corresponding tag terms.
- **"Wordsets" entity:** This entity houses the tag terms utilized in conjunction with the "Keyword" entities.
- **"Description" entity:** This entity serves as a repository for information related to publications.
- **"Taxons" entity:** Here, taxonomic affiliations of bacterial species retrieved from our database are stored.

- **Attributes**

Entities are distinguished by properties or attributes, each defined based on their type or cardinality. For instance, all sequences are associated with attributes such as accession number, publication title, publication description, DOI, genus, and species.

- **Identifiers**

Identifiers, or keys, are attributes or sets of attributes that uniquely identify elements within an entity type. In cases where none of the attributes can distinctly differentiate each element, it becomes necessary to create an attribute whose primary function is to serve as a key. Examples include idTaxon, idSequence, idKeyword, idWordset, and idDescription.

The "Sources" entity incorporates linking attributes known as foreign keys, specifically Sequences_idsequence and Keywords_idkeyword in our context. These foreign keys establish connections between different entities, contributing to the relational structure of the database.

- **Associations**

Associations denote relationships between different entities and are expressed intuitively as actions per entity. Various types of associations exist, categorized by their functional class and whether they are mandatory. In our case, we incorporate both one-to-many and many-to-many associations (refer to Figure 16).

- **One-to-many associations:** These link the "Sequences" entity to "Description" (where several sequences share a publication), the "Keywords" entity to "Wordsets" (where several formatted terms belong to a single wordset group), and the "Sequences" entity to "Taxons" (where several sequences share a single taxonomic identity).

- **Many-to-many associations**: These exclusively link the "Sources" entity to "Sequences" (indicating that several sequences can have multiple formatted terms).

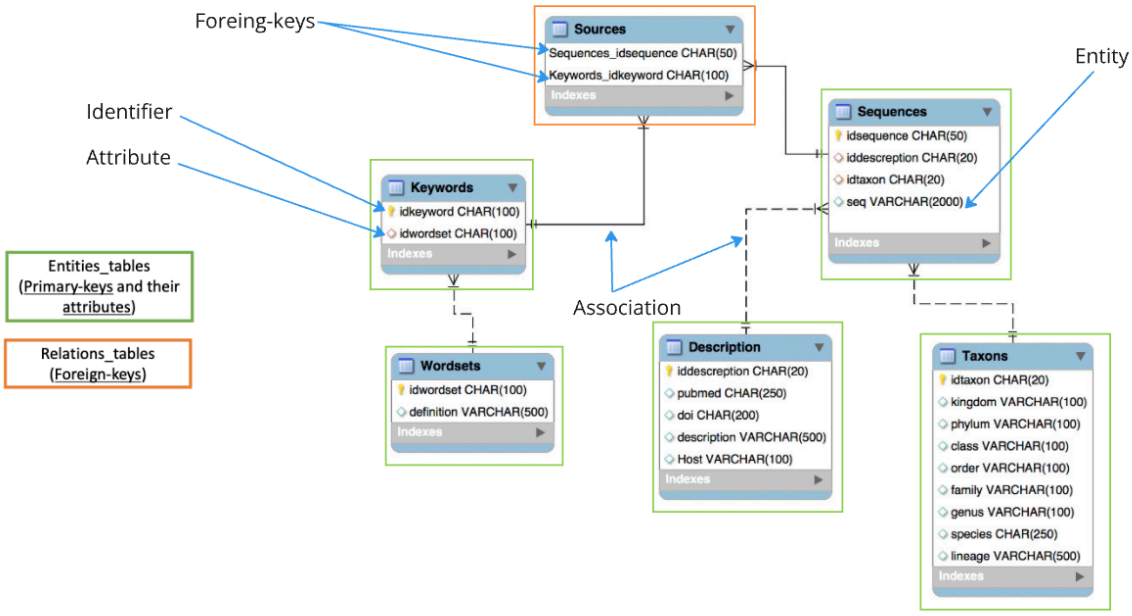


Figure 16. Database associations. *This illustration captures the associations among diverse entities within the database schema. The diagram prominently showcases both one-to-many associations (such as that between "Sequences" and "Description") and many-to-many associations (notably connecting "Sources" to "Sequences"). This visual depiction serves to elucidate the relational structure and connections inherent in the various entities of the database. The Conceptual Relational Schema, incorporating these associations, was designed using the MySQL Workbench tool.*

The schema underwent a conversion into SQL language utilizing the MySQL tool, resulting in an SQL format script (refer to Appendix 7). To query BiotopeBac-DB, we utilized the SQLite command-line interface (http://www.sqlite.org/index.html). To facilitate this process, we added a module to MySQL Workbench that enabled the SQL logical model to be written in SQLite dialect. The source database file, “BiotopeBac.sdb”, was saved in the binary format of SQLite can be downloaded from Figshare under the following DOI (10.6084/m9.figshare.24499804).

III.2.1.5 Keywords and word-sets visualization

A word cloud was used to analyze metadata associated to collected and stored in database in the preparation phase by using the tm package in R (refer to Appendix 8).

III.2.2 Dereplication phase and validation software

III.2.2.1 Filtering, Clustering and Exclusion

- **Filtering of ambiguous sequences**

Upon concluding the preparation phase, out of the 3,195,698 sequences annotated with corresponding wordsets (refer to Figure 17), a nuanced challenge emerged. Identical and partially identical sequences, originating from different records and potentially linked to distinct wordsets, coexist within this dataset. Complicating matters further, certain sequences exhibit ambiguous character states (ambiguous nucleotide annotation) despite being otherwise identical to their counterparts.

To address this complexity, we identified 184,223 sequences containing at least one such ambiguous character state. In response, we generated a "pure-full" version of the database alongside the original "semi-full" database, offering a refined perspective on sequence clustering and dereplication (see Figure 17). This strategic approach not only enhances the precision of the database but also allows for a more nuanced exploration of identical and partially identical sequences with attention to ambiguous nucleotides.

- **Clustering of identical sequences and wordset consolidation**

Following the generation of both the "pure-full" and "semi-full" database versions, a crucial step involved the clusterization of identical sequences in each dataset at a 100% identity threshold utilizing CD-HIT (v4.8.1) through a Perl script (refer to Appendix 9). The outcome was the identification of 2,309,036 and 2,486,039 unique reference sequences for the "pure-full" and "semi-full" databases, respectively.

Clusters comprising identical sequences were further processed as follows. First, the longest sequence within each cluster is marked as the reference sequence for the cluster. Subsequently, all wordset annotations from the sequences in the cluster were assigned to the reference sequence. Indeed, this dereplication process leads to reference sequence with redundancy in wordset annotations. In order to improve the probability distribution of word sets associated with each reference sequence, two strategies to handle the wordset annotations were explored using the "binarize" parameter described in the software construction part.

This approach not only streamlined the database but also facilitated a comprehensive exploration of identical sequence clusters. By selecting the longest sequence as the reference and merging annotations, we aimed to distill the essential information and enhance the robustness of the resulting database.

- **Exclusion of eukaryotic organelle 16S rRNA sequences**

Given that some sequences in the SILVA SSU Parc database originate from eukaryotic mitochondria or chloroplasts, our tool incorporates a mechanism to disregard specific lists of reference sequences associated with these entities. This exclusion mechanism operates during the sample processing step, eliminating the need for a complete rebuild of the databases or remapping of the amplicons.

Conceptually, envisioning these variations in the classification pipeline involves the utilization of distinct "no-eukaryote" databases. Consequently, the BiotopeBac-DB is available in four distinct types, each tailored to specific requirements regarding the inclusion or exclusion of ambiguous and/or organellar sequences (refer to Figure 17). This customization not only ensures precision in classification but also provides flexibility in accommodating diverse research needs by offering variants of the database to suit specific analysis scenarios.

The four databases derived from the dereplication phase were converted into binary format and subsequently employed as essential inputs for the development of our software construction.

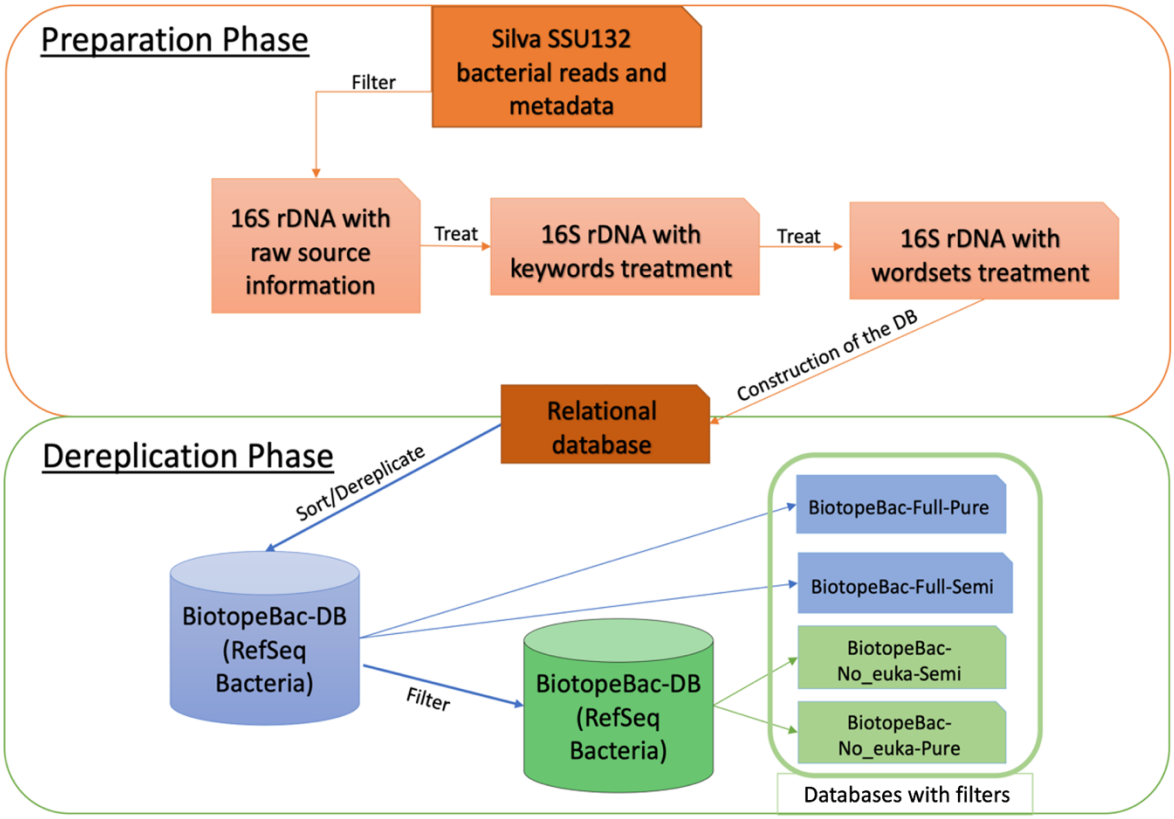


Figure 17. Overview of the database construction process. *The figure provides an overview of the two phases of the database construction process: preparation and dereplication. Preparation phase: During this phase, public 16S rDNA sequences were downloaded from the SILVA database, and sequence metadata required for the second phase were collected. dereplication phase: In this second phase we retrieved the collected metadata for all sequences in the database and used them to cluster the sequences into the BiotopBac-DB database.*

III.2.2.2 MetaGenomic Source Tool (MGST classifier): Software construction and parameters

- **Software construction**

The **MetaGenomic Source Tool** (**MGST** Classifier) was developed in our laboratory to apply a naïve Bayes probabilistic model for analyzing microbial sources. This tool employs probabilistic methods to explore and validate our database, with a particular focus on evaluating datasets refined during the dereplication phase using external public BioSamples. By doing so, MGST facilitates a thorough investigation of potential microbial sources. The complete development process of the MGST classifier is detailed in Appendix 10, which includes the Perl script, while Appendix 11 provides the mathematical framework underpinning the classifier.

The core functionality of the **MGST** classifier lies in its ability to utilize naïve Bayes probability to weigh various parameters, leading to informed predictions and enhancing the accuracy of microbial source identification.

MGST operates by starting with raw amplicon library reads (from the BiotopeBac-DB created during the preparation phase) and processing them through a series of filters applied during the dereplication phase. Each read is then mapped to the database using **BBMap** (Bushnell, 2014), an efficient and fast strict mapping tool, to ensure a strict match with available sequences and to enable the retrieval of associated metadata. This systematic approach increases the precision of microbial classification, reinforcing the robustness of our microbial source identification process.

By incorporating these sophisticated probabilistic methods, the **MGST** classifier aims to improve the accuracy and reliability of microbial source tracking. It serves thus as a suitable tool to contribute to the validition of the database and its applications.

- **Software parameters**

In addition to incorporating two variants of the database (***pure*** vs ***semi-pure***, as described during the dereplication phase), our solution underwent extensive testing with the variation of five other parameters.

The first two parameters are directly associated with the construction of the BiotopeBac-DB, influencing its structure and content. On the other hand, the last three parameters have the potential to impact the behavior of the MGST classifier, either directly or indirectly.

The combination of these 5 parameters, associated with the variants of the database gives a total of 1,200 parameter variations which have be assessed during the development and testing phases. This rigorous testing allowed us to evaluate the robustness and adaptability of the solution and our database under diverse scenarios, ensuring that the MGST classifier could effectively handle the complexities introduced by these varying parameters.

- **Parameter 1: *binarize***

During the process of consolidating multiple identical sequences into a single reference sequence, each individual sequence carried its own set of word-set annotations. A crucial step involved the consolidation of these individual word-set annotations before associating them with the reference sequence. The exact process involved in the incorporation is under the dependance of the binarize boolean option, characterized by **two values: *"multiple"* or *"binary"***.

This option provided the flexibility to determine the weighting assigned to a word-set term associated with more than one individual sequence in a cluster. With the ***"multiple"*** value, the weight was proportional to the occurrence count of the term, emphasizing its frequency in the consolidated probability distribution for the reference sequence. Conversely, the ***"binary"*** value ensured that all word-set terms annotating at least one individual sequence received equal weight, irrespective of their frequency. This strategic choice in weighting allowed for a nuanced and tailored approach in the consolidation of word-set annotations, contributing to the precision and adaptability of the MGST classifier.

- **Parameter 2*: theta***

To allow for some degree of error in the annotations of the reference database, conditional probabilities in BiotopeBac-DB were computed with a varying level of confidence, designed as *theta*, where ***theta* = 1** meant absolute confidence and ***theta* < 1** meant some degree of mis-annotation. In practice, MGST expects an *error-rate* parameter (defined as 1 - theta) when building the database. Here, we explored an error rate ranging from 1e-10 to 1e-01 on a logarithmic scale (**10** steps), with the lower values thus corresponding to a high confidence in the word-set annotations and the higher values corresponding to a low confidence.

- **Parameter 3: *ambiguous***

The nature of amplicons, being shorter than the complete 16S rDNA sequences stored in BiotopeBac-DB, introduces a unique challenge as they can align to several reference sequences. This scenario not only impacts mapping statistics (specifically the ***vote_n***; detailed below) but also introduces the potential for alternative distributions of word-set terms, each associated with a different reference sequence.

To address this issue, **BBMap** was employed with **three** distinct values for the ambiguous parameter. The first approach, labeled "**BEST**," adopts a stringent criterion by retaining only the first strict match, prioritizing the match with the highest alignment score (algorithm-dependent). The second approach, termed "**RANDOM**," randomly selects one match from the list of strict matches. The third and final approach, labeled "**ALL**," retains all strict matches resulting from the mapping process. By systematically testing these three approaches, we sought to rigorously control for any potential effects arising from multiple matches during the mapping process, ensuring a thorough exploration of alternative distributions of word-set terms associated with different reference sequences.

- **Parameter 4: *max-vote-n***

In theory, to ensure consistent statistical power across samples, regardless of their size in terms of number of sequences , the classifier could be configured to process a predetermined number of amplicons independently of their mapping status to the database. This could be achieved by setting an upper bound on the (***read_n)*** parameter (see page 66 for a description of the classifier process). However, after conducting preliminary testing, it became evident that the MGST classifier exhibited the most consistent behavior when processing as many amplicons as required to reach a predefined number of database matches (***vote_n,*** described page 66). As a result, in practical application, the max-vote-n parameter was developped to assume one of **five** discrete values: **10.000**, **5.000**, **1.000**, **500**, and **100**. This strategic choice allowed for a flexible and adaptive approach, accommodating varying dataset sizes and complexities while optimizing the statistical power and reliability of the MGST classifier.

- **Parameter 5: *prior-type***

The execution of Bayesian computations necessitates the definition of prior values (see page 35). Within MGST, the establishment of prior probabilities involves two distinct hypotheses. In the first scenario, referred to as the "**UNIFORM**" hypothesis, an assumption is made that the probability of the sample being derived from any host organism or source environment are equal. Essentially, this hypothesis posits an even distribution of likelihood across all potential sources.

Conversely, the second hypothesis, labeled "**DATABASE**," takes a more nuanced approach. Under this hypothesis, the prior probability distribution is modeled based on the observed frequencies of word-set terms within BiotopeBac-DB. In other words, the prior values are

tailored to reflect the actual distribution of annotated word-set terms present in the reference database. This consideration allows MGST to incorporate empirical observations from BiotopeBac-DB, providing a more context-specific and data-driven foundation for the Bayesian computations.

III.2.3 Database Validation

III.2.3.1 BioSample selection

As part of our evaluation, we proactively searched both public and private samples within the NCBI BioSamples database (https://www.ncbi.nlm.nih.gov/), showcasing a dedication to comprehensive testing for the validation of our classifier's accuracy and effectiveness. Being users of our own methodology, we took the utmost care in testing our system, ensuring its capability to consistently deliver dependable and trustworthy results to its intended audience.

To offer a thorough overview of the samples used in our evaluation, we intentionally selected samples with sources already present in our database. This deliberate selection aimed to focus our evaluation on data of high quality and relevance, allowing for a systematic analysis of the classifier's performance across various sample types. Through such extensive testing, users can confidently make well-informed decisions based on the data obtained, reinforcing the reliability and robustness of our classifier. Building upon this robust validation process, we will now delve into the details of our sample processing methodology, shedding light on the thoroughness and precision applied to ensure the integrity of our results.

To assess the performance of the MGST classifier, we employed a total of 60 publicly available Illumina MiSeq amplicon sequencing libraries. These libraries, acquired as paired-end reads, covered diverse regions of interest, including some with unknown regions. The selection of these regions was informed by the metadata provided by the Bioproject owner for 16S rDNA. Data acquisition was facilitated through the NCBI SRA portal (https://www.ncbi.nlm.nih.gov/sra), and subsequently, the obtained data was converted to FASTQ format for further analysis.

The 60 public BioSamples chosen for this validation represented a wide array of sources, specifically targeting the top 20 word-set terms most prevalent in BiotopeBac-DB. This diverse set included samples originating from various environmental sources such as aquatic, soil, air, and food, as well as organism sources including insects, fish, equids, suids, felids, birds, canids, humans, bovids, reptiles, rodents, amphibians, nematodes, crustaceans, angiosperms, and mollusks. Each sample adhered to a paired-read layout and featured at least 5653 read pairs, ranging between 180 and 300 nucleotides in length.

Additionally, we incorporated 24 additional samples from our private datasets, all in the form of paired-read Illumina MiSeq libraries targeting the V1-V3 region of bacterial 16S rDNA. These samples were sourced from ongoing internal lab projects, serving as examples of private BioSamples. The corresponding libraries are available under BioProject ID PRJNA659639 https://www.ncbi.nlm.nih.gov/bioproject/?term=PRJNA659639. Detailed information for each library, including specific characteristics and details, is available on FigShare under the corresponding DOI: 10.6084/m9.figshare.24499909. This comprehensive selection ensured a robust evaluation of the MGST classifier across a spectrum of diverse samples, both publicly available and internally generated.

III.2.3.2 Biosample mapping

The BioSamples processing pipeline involves a series of crucial steps to ensure the acquisition of high-quality data from the collected BioSamples. After merging reads and removing adapters, we employ a diverse array of analytical tools and techniques to extract meaningful insights from the data.

The first step is the merging step, we utilize BBMerge v38.79 (Bushnell et al., 2017) to merge overlapping paired reads into single reads, employing specific options such as adapter=default, maxloose=t, qtrim=t, and threads=8. To identify potentially problematic BioSamples, we calculate the total number of sequences (x) per sample and the percentage of merged amplicons (y) from the total (x).

These merged amplicons are then mapped onto our databases using BBMap v38.79 with options including reads=20000, fast=t, local=t, mdtag=t, -Xmx20g, and threads=10. The reads undergo local mapping, eliminating the need for a separate trimming process. However, due to the shorter length of merged amplicons compared to reference sequences, they can align perfectly to multiple unique reference sequences. To address this, we explore the effect of the ambiguous option through the ambiguous parameter (using of the following settings: **all**, **best**, and **random**).

The success of the mapping process is evaluated using two metrics: "**read_n**" representing the number of aligned amplicons (up to a maximum of 20,000) and "**vote_n**" indicating the count of matching reference sequences. Alignment begins with a specified number of reads, "reads_n," adjusted for sample size or merging process issues, ultimately yielding a mapping

percentage. Aligned reads can generate multiple votes ("vote_n") due to clustering, allowing a single read to activate multiple word-sets by voting on BiotopeBac-DB sequences. This meticulous approach ensures a thorough assessment of the mapping success and the nuanced interplay between aligned reads and reference sequences during the analysis (refer to Appendix 10).

III.2.3.3 BioSample classification

Ultimately, we deployed a naive Bayes classifier, aptly named "MGST," employing a post-processing approach on SAM files generated from the mapping of merged amplicons to the BiotopeBac-DB database by BBMap (refer to Figure 18 for a visual representation of the process). The primary objective of this classifier is to discern the predominant word-set term, indicative of the isolation source, for each individual sample, where each sample corresponds to a unique amplicon library. Detailed insights into the mathematical procedures underpinning this classification can be found in Appendix 11.

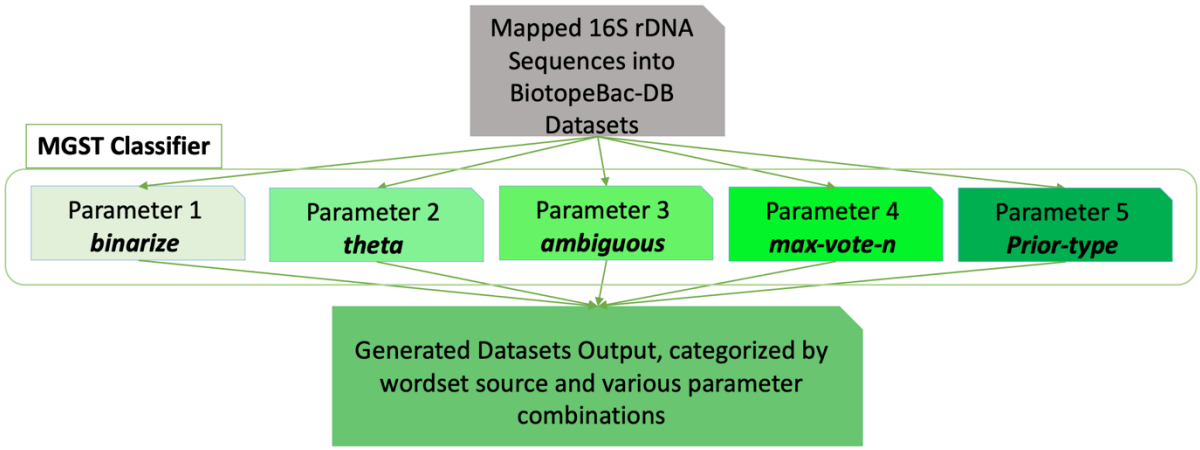


Figure 18. MGST classifier workflow. *This illustration delineates the intricate workflow of the MGST (MetaGenomic Source Tool) classifier. The procedure entails the post-processing of SAM files, a result of mapping merged amplicons to the BiotopeBac-DB database. The outcome yields datasets categorizing each biosample, associating them with the corresponding sources of wordsets. This categorization is achieved through the variation of parameters, providing nuanced insights into the diverse origins of the analyzed samples.*

Our tool is designed to provide either one or two word-set terms per sample, contingent upon the matched sequences in the BiotopeBac-DB database. To achieve this, we explored two variations of the approach. The first, named as the "1 onto" parameter, involves considering all

word-set terms as part of a single terminology and subsequently returning only the dominant word-set term. The second approach, termed the "2 onto" method, involves two distinct wordset ontologies: one for the environment ("Env::" group) and another for organisms ("Org::" group). In MGST, the corresponding option for this method is a boolean argument named "**split-tags**." When this option is selected, the different word-set groups are treated as separate ontologies rather than a unified one. The result is presented in excel file. This flexibility in our approach allows users to tailor the output based on their specific requirements and preferences.

III.2.3.4 Visualization

For the visualization process, we utilized the powerful 'circlize' package in R, renowned for creating insightful chord diagrams, to profile the database. This enabled us to present a comprehensive and visually compelling depiction of the complex relationships between different entities within the database. Additionally, we used the 'tm' package to perform word cloud analysis on the metadata collected during the preparation phase, specifically under manual terminology annotation, revealing patterns and trends. For all other figures, we employed Prism 8 to ensure clear and visual representation.

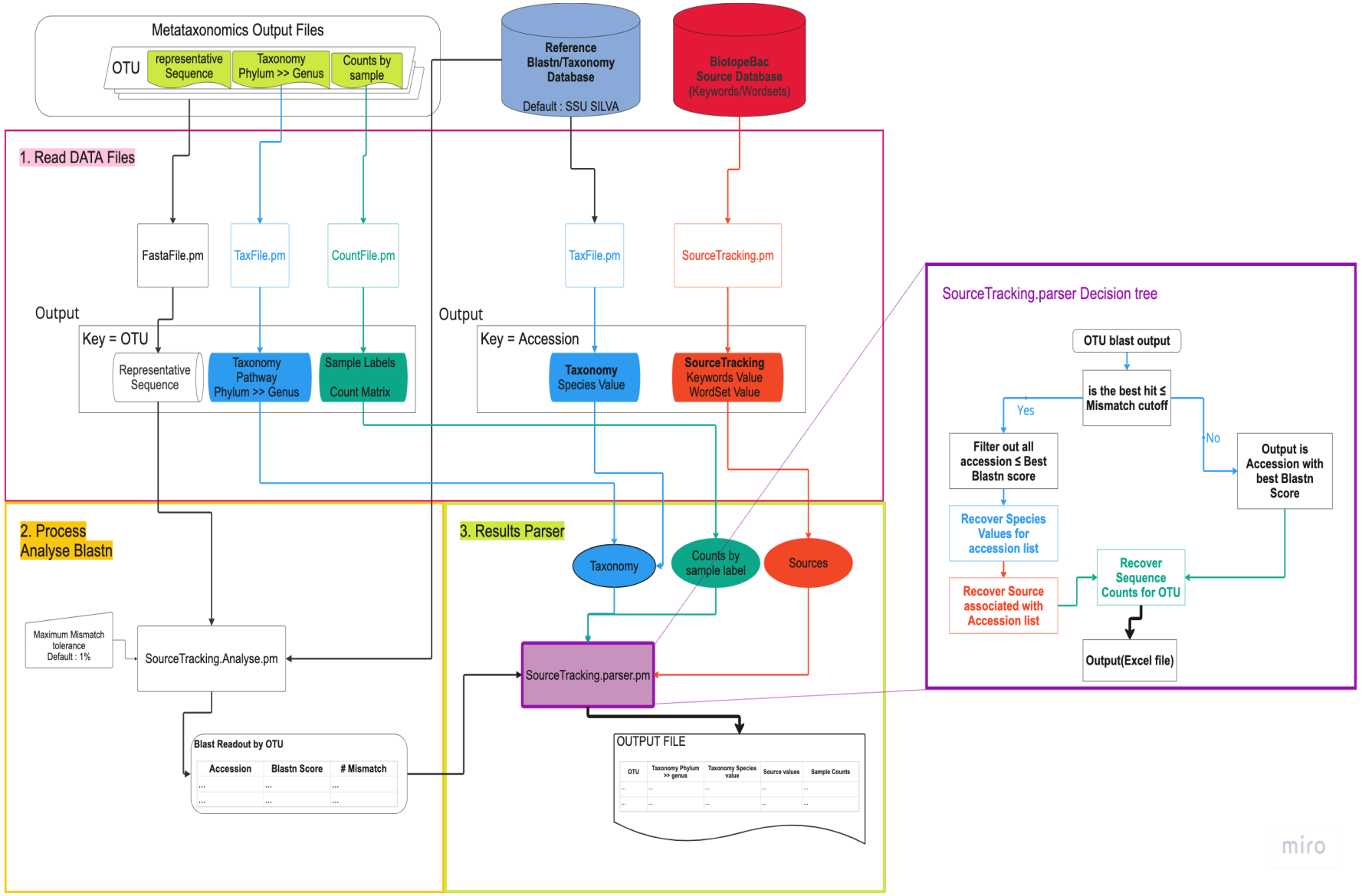
To further elevate the visual appeal and clarity of our representations, we made use of the versatility of the "ggplot2" and "reshape2" packages within the R environment. These packages enhanced the aesthetics and interpretability of our result visualizations, ensuring that the information conveyed was both insightful and accessible. Last, we used Miro (https://miro.com/) for some figures to facilitate interactive and collaborative design, allowing us to refine complex visualizations and enhance the overall clarity of our data presentations.

III.2.4 Sequence Tracking tool (ST): software construction

We developed the Sequence Tracking Tool (ST), an innovative tool designed to enhance the identification of source proportions within microbial communities and improve microbiota analysis. The ST tool achieves this by analyzing the total sequence occurrences and performing parallel analyses to simultaneously determine bacterial taxonomic profiles and their corresponding sources. The tool was developed using custom Perl scripts, which are detailed in Appendices (appendix12, 13, and 14).

Sequence Tracking Tool (ST) is a PERL based script dedicated to enrich metagenetics output with supplementary informations. The default use of ST is to attribute a species value to the taxonomic assignment of the identified OTU. The process uses blastn and OTU representative sequence as query and taxonomy reference database as sequence database. Another use of ST is to add another analysis layer with a sourcetracking tool to extract from BiotopeBac-DB either source “Wordset” or “Keywords” associated with the accessions linked to each OTU.

The figure 19 illustrates the Sequence Tracking Tool (ST) diagram in detail how the workflow is in each phase. The script is divided into 3 operations steps.


Metataxonomics Output Files
OTU
representative Sequence
Taxonomy Phylum >> Genus
Counts by sample
Reference Blastn/Taxonomy Database
Default : SSU SILVA
BiotopeBac Source Database (Keywords/Wordsets)
1. Read DATA Files
FastaFile.pm
TaxFile.pm
CountFile.pm
TaxFile.pm
SourceTracking.pm
Output
Key = OTU
Representative Sequence
Taxonomy Pathway Phylum >> Genus
Sample Labels
Count Matrix
Output
Key = Accession
Taxonomy
Species Value
SourceTracking
Keywords Value
WordSet Value
2. Process Analyse Blastn
Maximum Mismatch tolerance Default : 1%
SourceTracking.Analyse.pm
Blast Readout by OTU
Accession
Blastn Score
Mismatch
3. Results Parser
Taxonomy
Counts by sample label
Sources
SourceTracking.parser.pm
OUTPUT FILE
OTU
Taxonomy Phylum >> genus
Taxonomy Species value
Source values
Sample Counts
SourceTracking.parser Decision tree
OTU blast output
is the best hit ≤ Mismatch cutoff
Yes
No
Filter out all accession ≤ Best Blastn score
Output is Accession with best Blastn Score
Recover Species Values for accession list
Recover Source associated with Accession list
Recover Sequence Counts for OTU
Output(Excel file)

miro

Figure 19. Sequence Tracking Tool (ST) Workflow: **Data Processing and Analysis**. *This figure illustrates the step-by-step workflow of the Sequence Tracking Tool (ST). The process is divided into three phases: (1) Importing and organizing amplicon sequencing output data and reference databases, (2) Identifying homologous sequences using BLASTn, and (3) Parsing results based on a mismatch cutoff to determine taxonomic relevance. The final output combines taxonomy, sample counts, and source information for each OTU based on the analysis in one excel file.*

First, the data needed for the analysis are read and imported into the script. These data are the output files from the metagenetics organized with the OTU as the key values. In addition, the reference taxonomy Database (by default the SILVA SSU database) and BiotopBac database are imported, organized with the Accession as the key values.

During the second phase, the nearest homologous sequences from the reference taxonomy Database to each representative OTU sequence are identified with Blastn program. The readout of the process for each OTU are the first 100 best hits, with the Accession, mismatch and Blastn Score.

Once every OTU representative sequence are analyzed, the output is submitted to the Results parser. This part of the script will combine the taxonomy values (from kingdom to species), the counts value (per sample) and the Source values for each OTU into an output file. The Blast Parser includes a decision step based upon the mismatch cutoff value. The point here is to decide if the best hits from the Blastn are considered as "identical" to the query sequence (OTU representative sequence). By default, the mismatch cutoff is 1% mismatch (that is, 1 nucleotide mismatch for every 100 nucleotide of aligned sequence). For a given OTU representative sequence, if the best hit has a mismatch value superior to the cutoff, the best hit is considered as not identical. The taxonomic species value and BiotopeBac source value for the corresponding accession are thus considered as non-relevant. If the best hit mismatch value is not above the cutoff value, the parser will collect the accession with the best Blastn score. The taxonomic species value and BiotopeBac-DB source value for the corresponding list of accessions will be recovered and associated with the OTU in the parser output.

In summary, ST integrate various bioinformatics tools to process and analyze FASTA files, particularly for sequence alignment and taxonomic classification using the BLAST tool. The execution of the ST tool follows a structured six-step process, ensuring precise and efficient analysis of microbial communities.

- **Input Handling**:

In our taxonomic microbial analysis using MOTHUR (Schloss et al., 2009) and the Flash tool (Magoč et al., 2011), we used a FASTA file containing sequences, along with a taxonomy file and a count file, as input. Additionally, it was necessary to specify the version of the reference database (e.g., SILVA v132) for BLAST analysis.

- **BLASTn Execution**:

If not already provided, the script runs BLASTn to align sequences from the FASTA file against our specified reference database (BiotopeBac-DB). The results are saved in a specific format.

- **BLASTn Results Parsing**:

The script parses the BLAST results to identify the best hits for each sequence, considering mismatches and bit scores. It distinguishes between identical and non-identical hits, and it can flag sequences as chimeric based on the bit score.

- **Taxonomic Assignment**:

For each sequence, the script assigns a taxonomic classification (e.g., Kingdom, Phylum, Class) based on the BLAST results and the taxonomy file provided. It also identifies the nearest hit and determines if the sequence is chimeric.

- **Source Tracking**:

If source tracking is enabled, the script identifies the source of sequences after looking to the correspond accession number identified in the taxonomic assignment step by matching accessions to known sources on our database "BiotopeBac-DB".

After executing our script, an **Excel file is generated** containing detailed taxonomic and analytical data for the BioSamples analyzed. This file includes key taxonomic classifications, such as **Kingdom, Phylum, Class, Order, Family, Genus, and Species**, providing a hierarchical overview of the microbial communities. Additionally, it includes a list of Accession numbers and corresponding OTU (Operational Taxonomic Unit) labels, which are essential for tracking specific microbial sequences.

The output also features a bitscore, which reflects the confidence of sequence alignment, further aiding in the accuracy of source identification. Moreover, the file contains two specialized columns: **Source_tracking_wordset** and **Source_tracking_keywords**, which store metadata related to the origin of the biosample, as identified through our source-tracking methods.

A general workflow integrates in our ST tool using various tools and datasets, with distinct categories highlighted: data files in green, helper tools in orange, and generated data in blue. This organization reflects the coordinated use of computational resources to achieve a comprehensive analysis of the sequence data (refer to Figure 20 for a visual representation of the process).

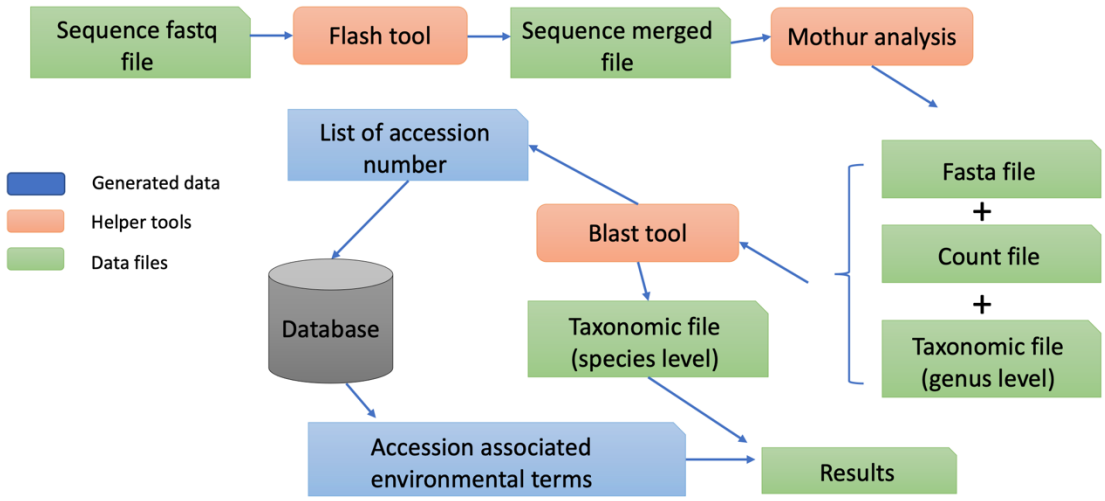


Figure 20. Flowchart illustrates the bioinformatics workflow used for processing sequence data, primarily focusing on taxonomic classification and environmental term analysis. *The process begins with a Sequence Fastq File, which is processed using the Flash Tool to generate a Sequence Merged File. This merged file is then subjected to MOTHUR Analysis to create essential output files, such as the Fasta File, Count File, and Taxonomic File at both the species and genus levels. A parallel process involves generating a List of Accession Numbers from the database, which is further analyzed using the Blast Tool. This tool matches the sequences to known taxa, contributing to the creation of taxonomic files. These files, along with environmental terms data, are consolidated to produce the final results in excel file format. The workflow is categorized by color codes: data files are shown in green, helper tools in orange, and data generated in blue, demonstrating the integration of various tools and datasets to achieve comprehensive sequence analysis*

III.3. Results

III.3.1 Preparation phase

III.3.1.1 Term annotation

The intricate journey of refining our initial dataset of 5,116,787 sequences, enriched with raw host source data and analyzed through the RStudio tool, unfurled a tapestry of seemingly random and non-relevant keywords, vividly portrayed in the intricate landscape of Figure 21.a. Unfazed by this initial challenge, we strategically harnessed the power of the BiotopeBac-DB terminology file, heralding a transformative metamorphosis. The refined dataset emerged, now comprising 4,264,470 sequences, intricately interwoven with carefully curated keywords such as "human," "soil," "skin," and "water" (refer to Figure 21.b for a visual representation).

Buoyed by this initial success, we embarked on a more granular refinement, laser-focused on specific word-sets. The subsequent iteration of refinement yielded a annotated dataset, now reduced to 3,195,698 sequences, each intricately associated with terms like "env::soil," "env::aquatic," and "org::human" (refer to Figure 21.c for a detailed visualization). Encouraged by these promising outcomes and the heightened precision in the composition of our dataset, we found ourselves at a pivotal juncture, prompting a strategic decision to delve deeper into our exploration.

Motivated by the demonstrated potential and validity witnessed in the third refinement of our database, we adopted it as a robust starting point for thorough validation in the subsequent steps of our analytical journey. This strategic approach serves as a meticulous checkpoint, ensuring the reliability and relevance of our refined dataset, thereby laying a formidable foundation for the upcoming phases of our multifaceted investigation. This methodical and strategic refinement not only enhances the quality of our dataset but also positions us with confidence for the intricate analytical strides that lie ahead in our research expedition.

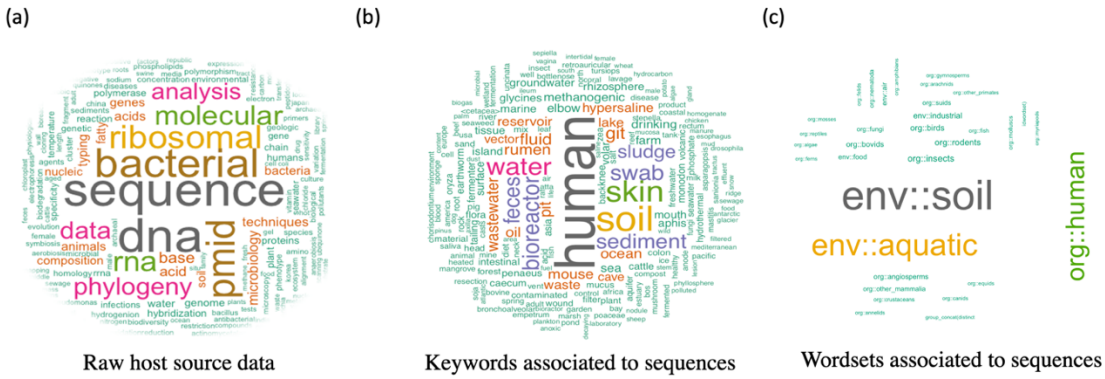


Figure 21. Iterative refinement process. *(a) Initial Dataset Analysis; dataset of 5,116,787 sequences with raw host source data, revealing an array of seemingly random and non-relevant keywords. (b) BiotopeBac-DB Terminology Refinement; Employing the BiotopeBac-DB terminology file led to a transformed dataset of 4,264,470 sequences. Notable associations emerged with carefully curated keywords such as "human," "soil," "skin," and "water."* **(***c) Granular Refinement of Word-sets; A more focused refinement targeted specific word-sets, producing a annotated dataset of 3,195,698 sequences intricately associated with terms like "env::soil," "env::aquatic," and "org::Human."*

Table 4 provides a detailed breakdown of term counts within curated word-sets. The analysis involves two primary ontologies, each marked by specific prefixes ("Org" and "Env") as described in the materials and methods section. The table presents 33 distinct word-sets, containing a total of 99,333 unique keywords. This table offers not just a numerical summary but also insight into the semantic structure of the word-sets based on the selected ontologies.

Table 4. Total number of terms grouped into each word-set. *The two ontologies are distinguished by their prefix, either "Org" or "Env". Thirty-two word-sets associated to a total of 99 333 different keywords.*

Word-set	Number of terms related to word-sets
Org ::INSECTS	34891
Org ::ANGIOSPERMS	14280
Org ::FUNGI	7141
Org ::FISH	5383
Org ::MOLLUSCS	4690
Org ::ARACHNIDS	4406
Org ::CRUSTACEANS	4244
Org ::BIRDS	2461
Org ::REPTILES	1250
Org ::ANNELIDS	1224
Org ::MOSSES	959
Org ::AMPHIBIANS	718

Org ::RODENTS	556
Org ::MYRIAPODS	528
Org ::FERNS	415
Env ::FOOD	206
Env ::AQUATIC	174
Org ::GYMNOSPERMS	115
Org ::Other_PRIMATES	112
Org ::PLANKTON	82
Env ::INDUSTRIAL	64
Org ::BOVIDS	63
Env ::SOIL	61
Org ::ALGAE	57
Org ::NEMATODA	39
Org ::CORALS	19
Org ::FELIDS	18
Org ::Other_MAMMALIA	16
Org ::CANIDS	14
Env ::AIR	11
Org ::EQUIDS	11
Org ::SUIDS	6
Org ::HUMAN	2

III.3.1.2 Database profiling

Our primary goal in analyzing the relationships between taxonomy and word-sets in the database was to highlight the most abundant taxa within the three main source groups across various taxonomic levels, from phylum to genus. To illustrate this, we focused on the taxonomic composition linked to the three-leading major word-set terms in BiotopeBac-DB: soil, human, and aquatic.

Figure 22 provides a clear visualization of the major phyla present in major sources, including *Proteobacteria, Firmicutes, Bacteroidetes, Actinobacteria, Acidobacteria*, and *Cyanobacteria.* This analysis offers valuable insights into the diverse microbial communities associated with these environments.

Within the aquatic and soil sources, certain taxa emerged as particularly noteworthy. *Proteobacteria* stood out as the most abundant phylum, while *Rikenellaceae, Bacillus,* and *Pseudomonas* took the spotlight as the predominant family and genera, respectively. This detailed insight into the taxonomic composition underscores the microbial diversity and ecological significance within aquatic and soil environments.

In contrast, the human sources exhibited a distinct taxonomic profile, with a prevalence of *Firmicutes, Actinobacteria, Nitrosomonadaceae, and Staphylococcus* as the most abundant taxa. This divergence in taxa abundance emphasizes the unique microbial communities associated with human-related environments. Overall, our analysis sheds light on the intricate interplay between taxonomy and word-sets, offering valuable insights into the microbial ecology of diverse environmental sources.

In-depth insights into each major source within our database are presented in the following paragraph.

- **Phyla Observed for the "Human" Source**

The outcomes revealed that 720,456 sequences are associated with this profile, where *Firmicutes* constituted a major group (46.20%) linked to the human host source in our database. Following closely were *Actinobacteria (22.67%), Proteobacteria (16.64%), Bacteroidetes (11.31%), Fusobacteria (1.31%), Epsilonbacteraeota (0.36%), Cyanobacteria (0.30%), Tenericutes (0.26%), and Spirochaetes (0.22%).* Additionally, 32 other phyla were categorized as "Other Phyla" (0.74%) associated with the human host.

- **Phyla Observed for the "Aquatic" Source**

The findings revealed that 1,191,152 sequences are associated with water, predominantly comprising *Proteobacteria (45.07%*) linked to the aquatic environment. Following were *Bacteroidetes (14.35%), Firmicutes (10.35%), Actinobacteria (7.01%), Cyanobacteria (6.10%), Planctomycetes (2.94%), Acidobacteria (2.92%), Chloroflexi (1.68%), and Epsilonbacteraeota (1.56%).* An additional 70 phyla were associated with aquatic sources, constituting 8.02% of the total phyla.

- **Phyla Observed for the "Soil" Source**

The findings indicated that 1,384,966 sequences are linked to this source, with a taxonomic profile predominantly composed of *Proteobacteria (over 30%)*, followed by *Firmicutes (11.12%), Bacteroidetes (10.71%), Actinobacteria (9.79%), Acidobacteria (7.86%), Chloroflexi (4.49%), Planctomycetes (3.78%), Cyanobacteria (1.84%), Verrucomicrobia*

(1.63%), and Patescibacteria (1.58%). The remaining phyla are collectively represented as "Others," comprising 67 different categories.

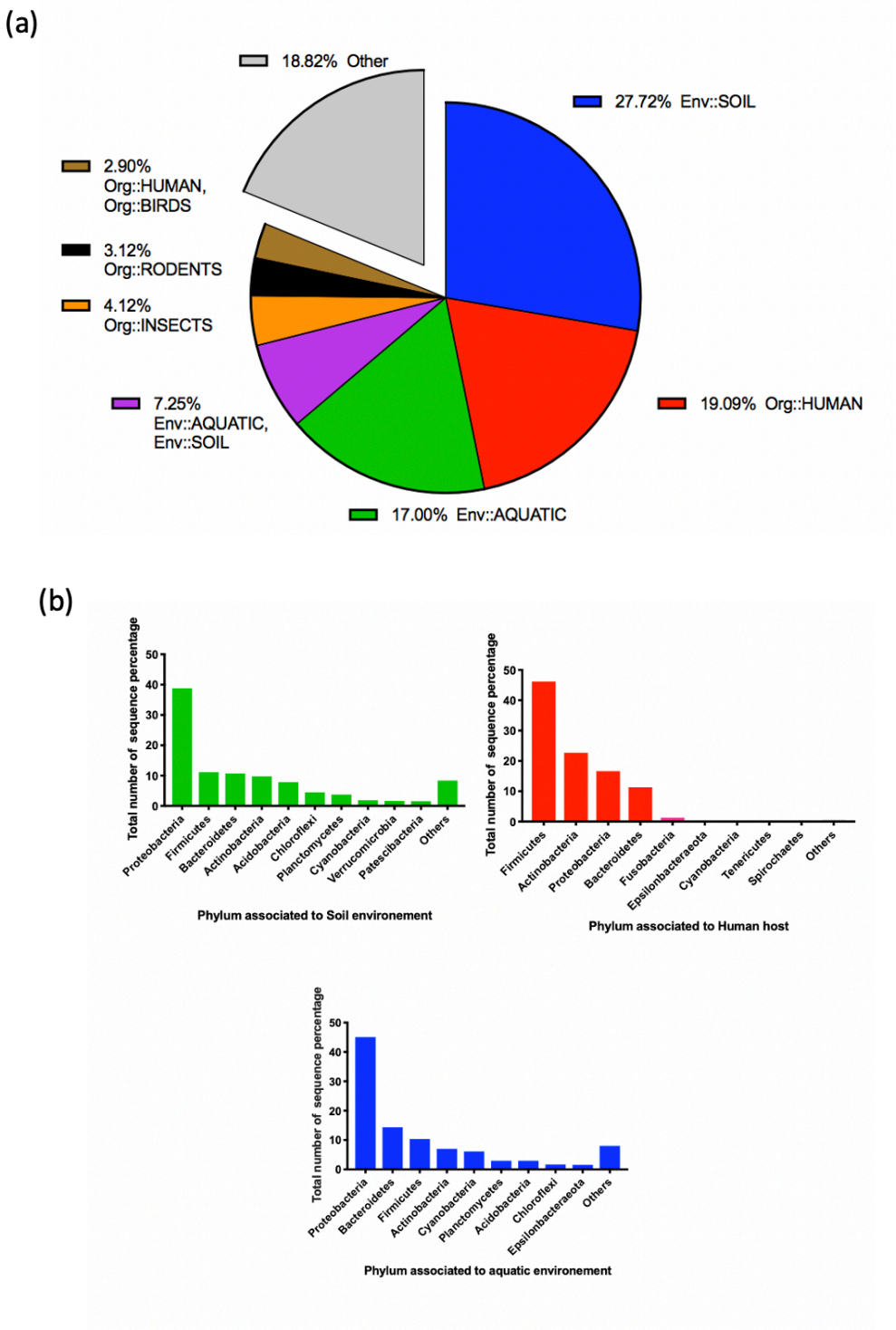


Figure 22. Microbial Diversity Profiling Across Varied Sources. *This figure provides a comprehensive snapshot of microbial diversity and abundance derived from an in-depth analysis of the BiotopeBac-DB database. (a) The pie chart visually encapsulates the dominant sources represented within the extensive database, comprising a total of 3,195,699 reference sequences. (b) Accompanying histograms elucidate the percentage distribution of bacterial phyla intricately associated with these diverse sources. Together, these visualizations offer*

valuable insights into the rich taxonomic landscape and relative abundance of microbial communities across different environmental contexts.

In our quest for a profound comprehension of the intricate tapestry encapsulated within our database, we embarked on a comprehensive analysis, a veritable intellectual journey designed to unravel the nuanced associations between bacterial taxonomic groups and their linked sources. This analytical odyssey, marked by its depth and rigor, involved the establishment of intricate correlations between the carefully curated word-set terms and the expansive bacterial taxonomy. The graphical representation chosen for this endeavor, specifically the Chord plot, emerged as a pivotal tool, providing a visually captivating depiction of the complex associations and relationships governing our dataset.

The Chord plot, as showcased in the visually compelling Figure 23, transcends mere visualization; it becomes a narrative canvas where the interplay between the 80 bacterial phyla and the 19 word-set terms with the highest abundance unfolds. This visualization technique, reliant on cumulative sequence counts across all word-set terms, transcends traditional representation methods, offering a nuanced exploration of interconnections and patterns interwoven among bacterial phyla and the prevailing word-set terms.

The pursuit of meaningful relationships between phylum, family, and genus bacterial levels and their respective sources demanded a meticulous filtering process. In an unwavering commitment to precision, sequences devoid of word-sets were methodically excluded, an indispensable step crafted to eliminate potential noise and ensure the pristine reflection of associations between bacterial taxonomic levels and their corresponding sources. This meticulous analytical scrutiny bore fruit, revealing a staggering identification of over 3 million word-sequence connections.

A tapestry of diversity emerged as we delved into the distribution of phyla across various sources. Soil sources, with their intricate connections, unfurled associations with a remarkable 78 phyla, while aquatic sources displayed a plethora of diversity, encompassing connections to 80 distinct phyla. In contrast, the human host, within our extensive database, exhibited connections to 42 phyla. The nuanced exploration depicted in Figure 23 not only visually encapsulates these variations but provides a robust foundation for deciphering the intricate relationships and patterns characterizing the bacterial taxonomic landscape within our dataset. This depth of exploration transcends numerical figures, becoming a narrative continuum that

beckons researchers to navigate the complex terrain of bacterial taxonomy and ecological associations embedded in our dataset.

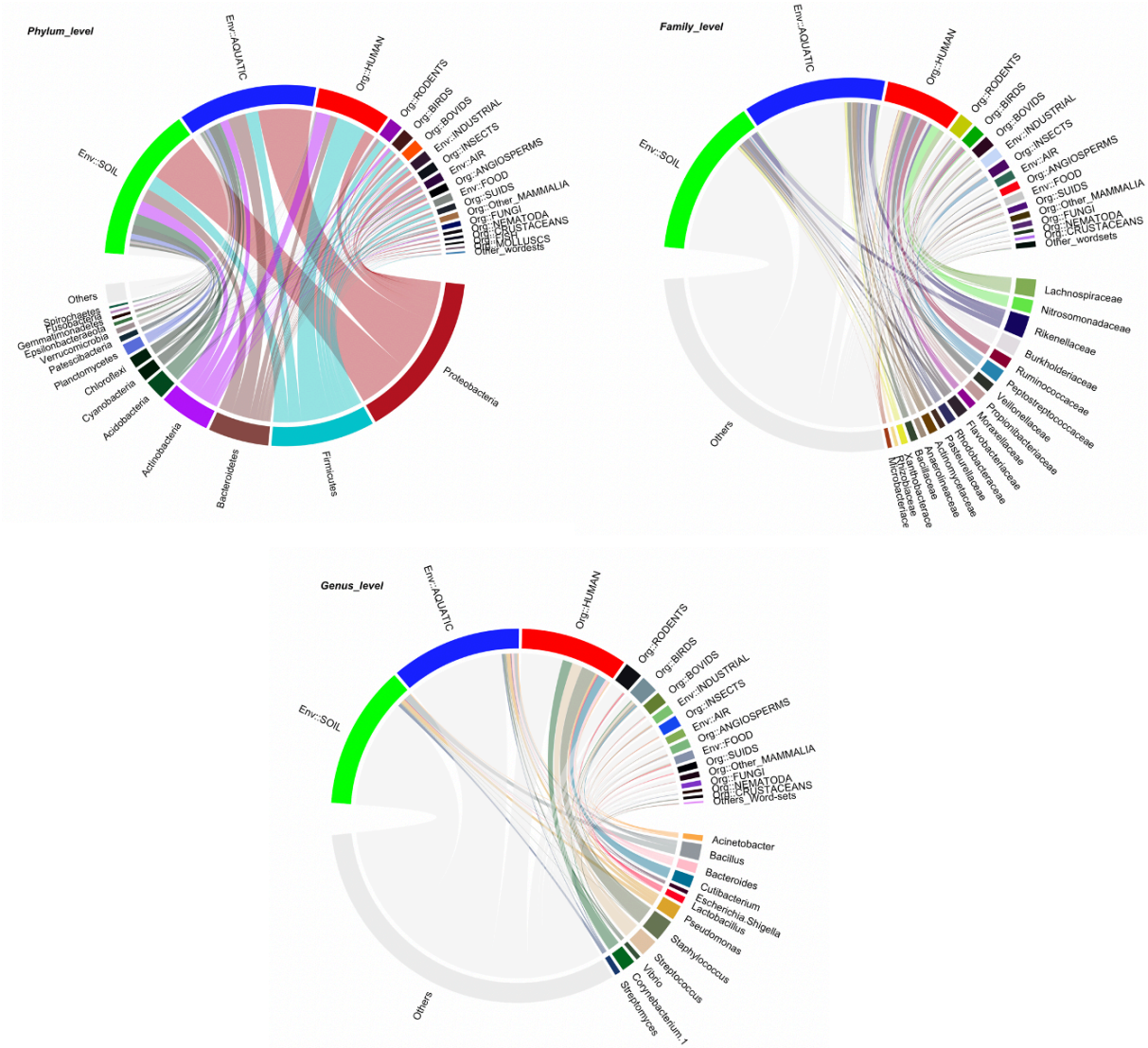


Figure 23. Chord plot illustrating taxa and wordset term connections. *Chord Diagram Overview: The chord diagram offers a comprehensive overview of the connections between different taxa and isolation sources, emphasizing their corresponding spectral abundances. This visualization highlights the intricate relationships among microbial taxa and their sources within the dataset. The figure depicts the linkage between taxa and word-set terms through a chord diagram. This graphical representation showcases the connections between specific taxa (at the phylum, family, or genus level) and isolation sources, focusing on the 19 most abundant taxa and 19 most abundant sources. Abundance is determined by the summed sequence counts across all word-set terms. Circle Segments: The size of each circle segment corresponds to the spectral abundance of a taxon or source. Larger segments indicate higher abundance. Arcs Between Circles: The arcs connecting the circles represent the number of shared spectra between two entities, providing a quantitative measure of their relationship.*

Note: *The Chord plot aids in visualizing the complex interplay between microbial taxa and isolation sources, offering valuable insights into the patterns and associations within the dataset's microbial diversity.*

III.3.2 Dereplicate phase

III.3.2.1 BioSamples mapping

In undertaking a comprehensive exploration to gauge the manifold implications of parameter variations on the myriad BioSamples housed within our extensive database, our investigative journey embarked with a scrupulous scrutiny of the ramifications emanating from the merging read process. The crux of our inquiry rested on unraveling the intricate tapestry of successfully processed samples and discerning their percentage in the vast expanse of the total dataset.

Before the application of our tool, an in-depth analysis to extract critical metrics was performed. Specifically, we delved into the total number of sequences denoted as (x) per BioSample, coupled with a discerning examination of the percentage of merged datasets (y) from the overarching (x). The observation of these metrics, as shown in table 5, revealed trends within the public BioSamples. A predominant theme emerged, characterized by consistently high average values of (y) among a substantial cross-section of datasets within the public BioSamples. The aquatic, fish, and soil datasets, in particular, etched their prominence, boasting average values of 53%, 44%, and 66%, respectively, in the context of the total sequences per dataset. In contrast, the private BioSamples, while mirroring comparably elevated average values of (y), showed lower average values for the industrial and suids datasets (with 61% and 68%, respectively).

Our selection of BioSamples is curated to mirror the diverse source profiles present in our extensive database. This representative sampling is important for ensuring the validity and reliability of our research outcomes. We acknowledge the challenges inherent in locating alternative samples that embody the same breadth of variability. However, our rigorous selection process is designed to encompass a comprehensive range of biological variables, thereby justifying our choices and reinforcing the integrity of our data collection methodology.

Table 5. Total number of sequences in each tested datasets BioSamples (publics and privates) and the percentage of success in merging process. *For each dataset, BBMerge was used to cover the overlapping reads and three replicates of each source type. (a) Result of each 60 datasets of BioSamples downloaded from public data available in NCBI. (b) Result of 24 datasets of BioSamples tested in our own laboratories.*

(a)

Samples	Region of 16 rDNA	Total number of sequences in a sample before merging (x)	Percentage of sequences merged (y)	Total read kept after merging
air1-merged.fasta	V1-V2	40518	95.88%	38849
air2-merged.fasta	V1-V2	45746	95.77%	43811
air3-merged.fasta	V1-V2	53895	95.90%	51685
amphibians1-merged.fasta	unknown*	187540	91.84%	172237
amphibians2-merged.fasta	unknown*	103059	83.99%	86559
amphibians3-merged.fasta	unknown*	125952	84.06%	105875
angiosperms1-merged.fasta	unknown*	154151	97.81%	150775
angiosperms2-merged.fasta	unknown*	90626	98.34%	89122
angiosperms3-merged.fasta	V3-V4	62028	98.26%	60949
aquatic1-merged.fasta	unknown*	114405	53.74%	61481
aquatic2-merged.fasta	unknown*	105662	52.65%	55631
aquatic3-merged.fasta	unknown*	120352	53.55%	64448
birds1-merged.fasta	V4	93573	76.13%	71237
birds2-merged.fasta	V4	84289	77.53%	65349
birds3-merged.fasta	V4	86907	76.35%	66353
bovids1-merged.fasta	V4	95077	81.55%	77535
bovids2-merged.fasta	V4	92864	81.83%	75991
bovids3-merged.fasta	V4	69741	82.86%	57787
canids1-merged.fasta	V4	22712	85.96%	19523
canids2-merged.fasta	V4	22940	86.16%	19765
canids3-merged.fasta	V4	26836	89.28%	23959
crustaceans1-merged.fasta	V1-V2	5626	91.67%	5157
crustaceans2-merged.fasta	V1-V2	6295	92.69%	5835
crustaceans3-merged.fasta	V1-V2	5326	90.48%	4819
equids1-merged.fasta	unknown*	12606	93.51%	11788
equids2-merged.fasta	unknown*	60487	92.95%	56223

equids3-merged.fasta	unknown*	64298	93.44%	60080
felids1-merged.fasta	unknown*	5335	88.60%	4727
felids2-merged.fasta	unknown*	42080	87.65%	36883
felids3-merged.fasta	unknown*	50683	87.16%	44175
fish1-merged.fasta	V4	15007	24.47%	3672
fish2-merged.fasta	V4	5687	15.70%	893
fish3-merged.fasta	V4	36899	94.06%	34707
food1-merged.fasta	V3-V4	177608	96.09%	170664
food2-merged.fasta	V3-V4	199846	96.61%	193071
food3-merged.fasta	V3-V4	46357	73.61%	34123
human1-merged.fasta	V3-V4	28822	77.91%	22455
human2-merged.fasta	unknown*	40707	81.11%	33017
human3-merged.fasta	unknown*	14185	74.02%	10500
insects1-merged.fasta	unknown*	56521	93.06%	52598
insects2-merged.fasta	unknown*	30828	93.22%	28738
insects3-merged.fasta	unknown*	29974	88.36%	26485
molluscs1-merged.fasta	unknown*	35005	96.37%	33734
molluscs2-merged.fasta	unknown*	42676	88.64%	37828
molluscs3-merged.fasta	unknown*	37573	78.01%	29311
nematoda1-merged.fasta	V4	55434	81.34%	45090
nematoda2-merged.fasta	V4	66153	82.06%	54285
nematoda3-merged.fasta	V4	58273	79.83%	46519
reptiles1-merged.fasta	V1-V2	45857	85.51%	39212
reptiles2-merged.fasta	V1-V2	23959	74.14%	17763
reptiles3-merged.fasta	V1-V2	20795	78.16%	16253
rodents1-merged.fasta	unknown*	27348	81.46%	22278
rodents2-merged.fasta	unknown*	23120	81.76%	18903
rodents3-merged.fasta	unknown*	15977	79.42%	12689
soil1-merged.fasta	unknown*	40254	71.23%	28673
soil2-merged.fasta	unknown*	65783	87.44%	57521
soil3-merged.fasta	unknown*	25907	40.87%	10588
suids1-merged.fasta	V1-V2	74747	89.60%	66973
suids2-merged.fasta	V1-V2	208993	88.99%	185983
suids3-merged.fasta	V1-V2	130435	89.17%	116309

*Indicates unknown data due to the unavailability of sequence information in the 16S rDNA region.

(b)

Samples	Region of 16 rDNA	Total number of sequences in a sample before merging	Percentage of sequences merged	Total read kept after merging
Birds1_private_data	V1-V3	81325	91.61%	74502
Birds2_private_data	V1-V3	78654	91.68%	72110
Birds3_private_data	V1-V3	75106	91.23%	68519
Canids1_private_data	V1-V3	248116	81.00%	200974
Canids2_private_data	V1-V3	182506	81.36%	148487
Canids3_private_data	V1-V3	111398	81.67%	90979
Fish1_private_data	V1-V3	132399	90.24%	119477
Fish2_private_data	V1-V3	117555	89.28%	104953
Fish3_private_data	V1-V3	156944	90.40%	141877
Food1_private_data	V1-V3	136889	92.07%	126034
Food2_private_data	V1-V3	163802	91.55%	149961
Food3_private_data	V1-V3	180872	93.34%	168826
Industrial1_private_data	V1-V3	72124	49.74%	35874
Industrial2_private_data	V1-V3	209493	68.29%	143063
Industrial3_private_data	V1-V3	122012	67.22%	82016
Rodents1_private_data	V1-V3	177507	86.12%	152869
Rodents2_private_data	V1-V3	131796	81.00%	106755
Rodents3_private_data	V1-V3	126505	83.55%	105695
Soil1_private_data	V1-V3	89334	80.64%	72039
Soil2_private_data	V1-V3	177442	90.62%	160798
Soil2_private_data	V1-V3	186387	90.30%	168307
Suids1_private_data	V1-V3	63631	66.61%	42385
Suids2_ private_data	V1-V3	103624	72.43%	75055
Suids3_ private_data	V1-V3	45352	68.78%	31193

III.3.2.2 Alignment result

Subsequent to the the merging process, the mapping phase applied to the resultant datasets to an alignment procedure utilizing four distinct databases, each distinguished by its distinctive array of features. Two databases were filtered out of eukaryotic sequences (BiotopeBac-No-Euka-Semi and BiotopeBac-No-Euka-Pure), while their counterparts, BiotopeBac-Full-Semi and BiotopeBac-Full-Pure, operated without such filtration.

We also analyzed the utilization of reads (read_n) by the BBmap tool as well as the distribution of values obtained for both the number of reads aligned to contribute to a vote (align_n) and the cumulative number of votes garnered (vote_n) for both private and public libraries.

With the exception of some samples, BBmap was able to align a maximum of 20,000 reads (read_n). this maximum being set up before the process. In instances where this number of reads is not reached, BBmap was nonetheless able to align all successfully merged sequences from the merging step.

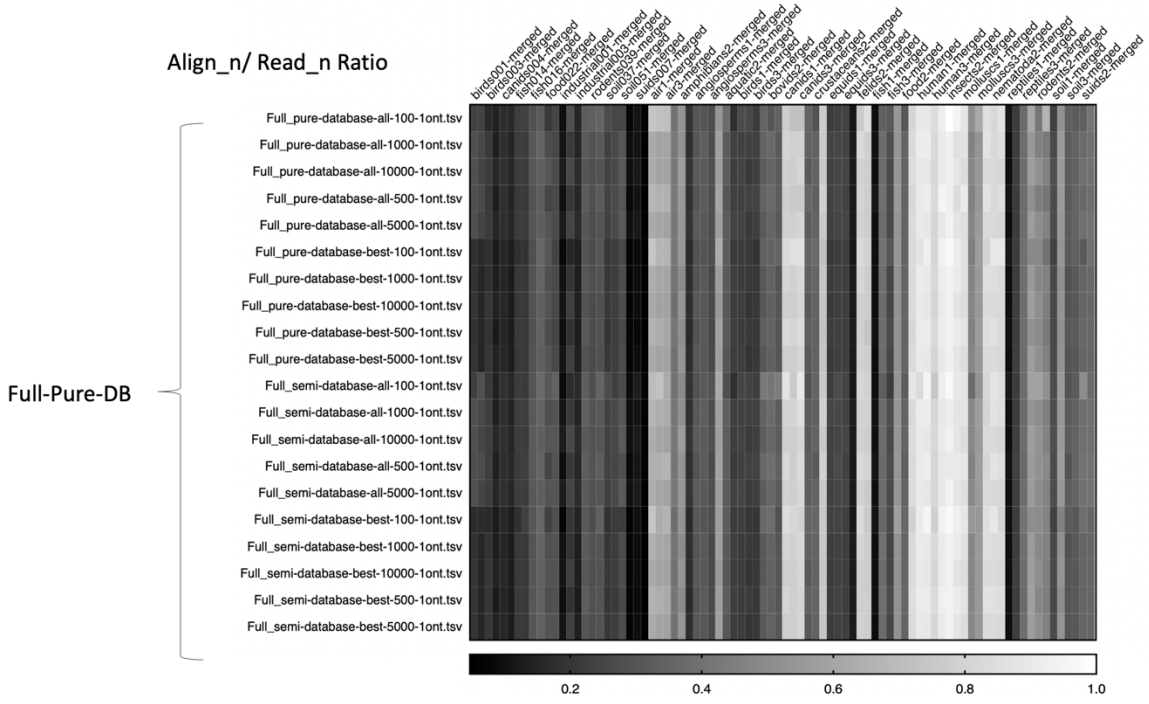
As an example, the air3 BioSample, employing the "semi" database and "all" parameter, where alignment statistics manifested as follows: The attaining of 5000 votes necessitated the alignment of 1764 amplicons, demanding a examination of 3005 pairs of reads (assembled amplicons). It's important to note that while the actual number of reads surpassed these specified figures, the values were selectively chosen to align with the analytical requisites of our study.

Our analytical purview expanded further to encompass an exhaustive analysis of read proportions mapped across diverse databases, homing in on the "full" and "no-euka" databases. In order to ensure both consistency and comparability, we deliberately retained the "pure" parameter, closely aligning with the "semi" parameter. The results are shown in Figure 24, where a ratio of 1 signifies 100% of reads aligning to the respective databases (Figure 24a and 24b). We also analyzed the relationship between "vote_n" and "align_n" ratios under the "best" parameter setting, revealing a consistent ratio of 1. This correlation implies that one best-read aligns seamlessly with one vote, as highlighted in Figure 24c and 24d.

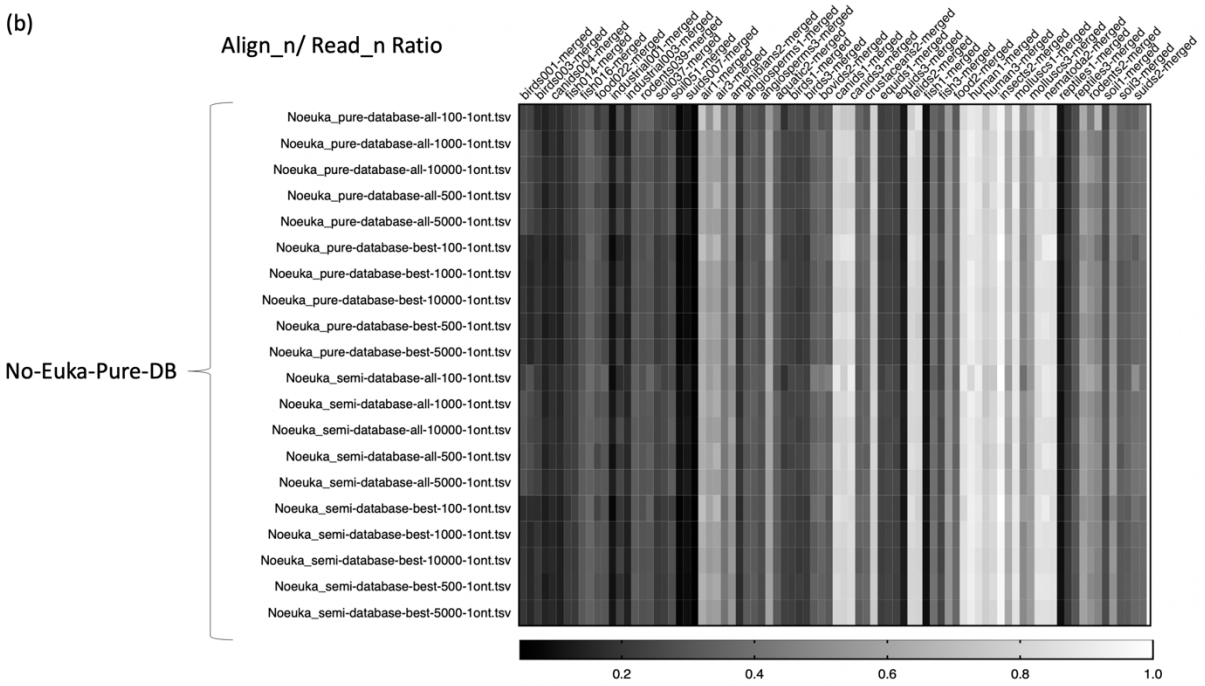
Furthermore, we examined the combinations where "vote_n" could reach to the maximum vote threshold, designated as "vote_max" (100, 500, 1000, 5000, 10000). Notably, a ratio of 1 in this context signifies the culmination of all votes for a given BioSample reaching the prescribed maximum (Figure 24e and 24f).

These analyses, shed light on the dynamic interplay of alignment and voting within the process and on the sensitivity of MGST regarding the variations of the different parameters.

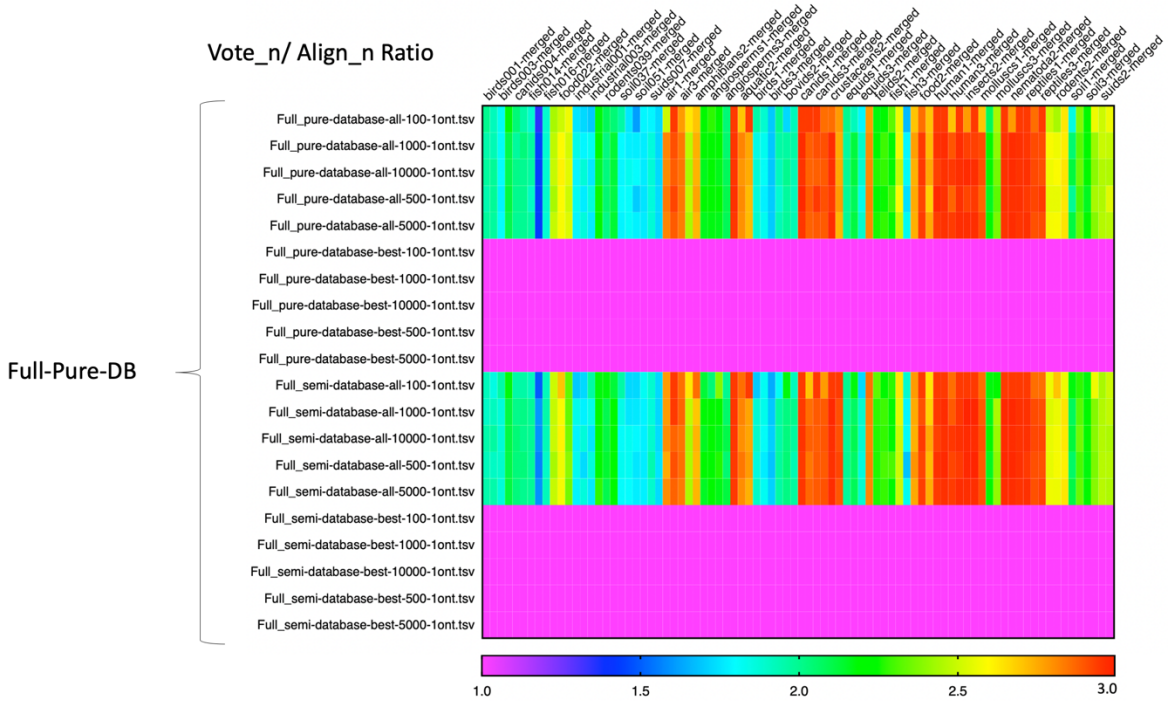
(a)


Align_n/ Read_n Ratio
birds001-merged
birds003-merged
canids004-merged
fish014-merged
fish016-merged
food022-merged
industrial001-merged
industrial003-merged
rodents039-merged
soil037-merged
soil051-merged
suids007-merged
air1-merged
air3-merged
amphibians2-merged
angiosperms1-merged
angiosperms3-merged
aquatic2-merged
birds1-merged
birds3-merged
bovids2-merged
canids1-merged
canids3-merged
crustaceans2-merged
equids1-merged
equids2-merged
felids2-merged
fish1-merged
fish3-merged
food2-merged
human1-merged
human3-merged
insects2-merged
molluscs1-merged
molluscs3-merged
nematode2-merged
reptiles1-merged
rodents2-merged
soil1-merged
soil3-merged
suids2-merged
Full_pure-database-all-100-1ont.tsv
Full_pure-database-all-1000-1ont.tsv
Full_pure-database-all-10000-1ont.tsv
Full_pure-database-all-500-1ont.tsv
Full_pure-database-all-5000-1ont.tsv
Full_pure-database-best-100-1ont.tsv
Full_pure-database-best-1000-1ont.tsv
Full_pure-database-best-10000-1ont.tsv
Full_pure-database-best-500-1ont.tsv
Full_pure-database-best-5000-1ont.tsv
Full_semi-database-all-100-1ont.tsv
Full_semi-database-all-1000-1ont.tsv
Full_semi-database-all-10000-1ont.tsv
Full_semi-database-all-500-1ont.tsv
Full_semi-database-all-5000-1ont.tsv
Full_semi-database-best-100-1ont.tsv
Full_semi-database-best-1000-1ont.tsv
Full_semi-database-best-10000-1ont.tsv
Full_semi-database-best-500-1ont.tsv
Full_semi-database-best-5000-1ont.tsv
Full-Pure-DB
0.2
0.4
0.6
0.8
1.0

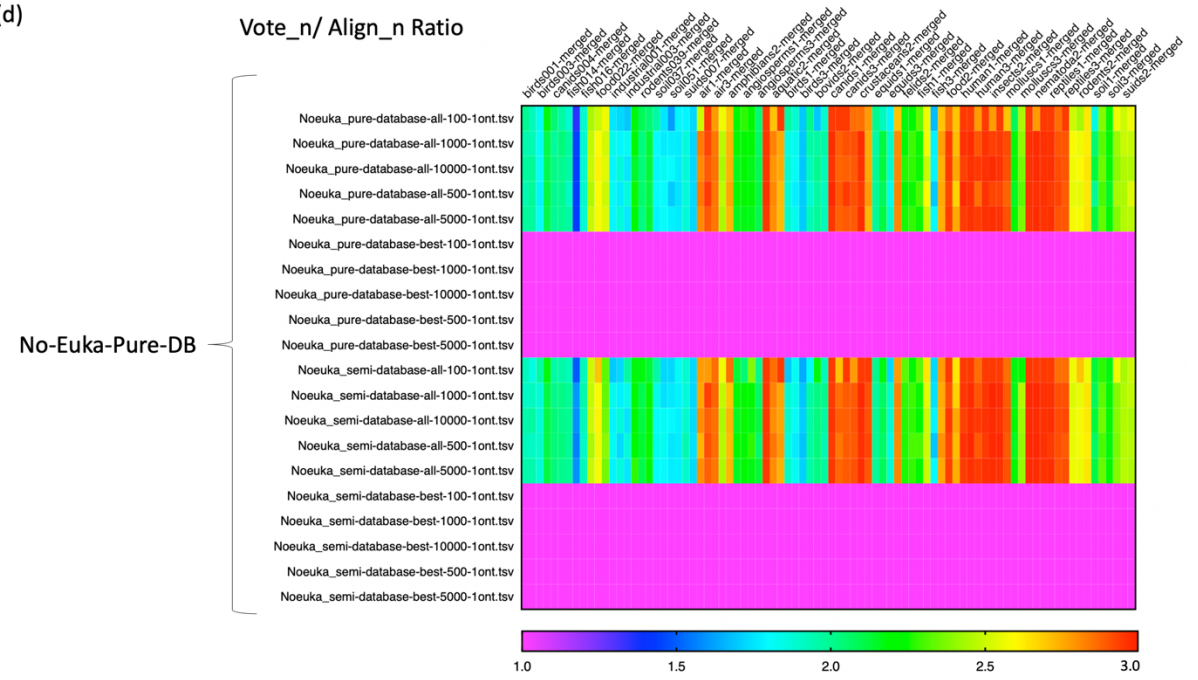
(b)


Align_n/ Read_n Ratio
birds001-merged
birds003-merged
canids004-merged
fish014-merged
fish016-merged
food022-merged
industrial001-merged
industrial003-merged
rodents039-merged
soil037-merged
soil051-merged
suids007-merged
air1-merged
air3-merged
amphibians2-merged
angiosperms1-merged
angiosperms3-merged
aquatic2-merged
birds1-merged
birds3-merged
bovids2-merged
canids1-merged
canids3-merged
crustaceans2-merged
equids1-merged
equids2-merged
felids2-merged
fish1-merged
fish3-merged
food2-merged
human1-merged
human3-merged
insects2-merged
molluscs1-merged
molluscs3-merged
nematode2-merged
reptiles1-merged
rodents2-merged
soil1-merged
soil3-merged
suids2-merged
Noeuka_pure-database-all-100-1ont.tsv
Noeuka_pure-database-all-1000-1ont.tsv
Noeuka_pure-database-all-10000-1ont.tsv
Noeuka_pure-database-all-500-1ont.tsv
Noeuka_pure-database-all-5000-1ont.tsv
Noeuka_pure-database-best-100-1ont.tsv
Noeuka_pure-database-best-1000-1ont.tsv
Noeuka_pure-database-best-10000-1ont.tsv
Noeuka_pure-database-best-500-1ont.tsv
Noeuka_pure-database-best-5000-1ont.tsv
Noeuka_semi-database-all-100-1ont.tsv
Noeuka_semi-database-all-1000-1ont.tsv
Noeuka_semi-database-all-10000-1ont.tsv
Noeuka_semi-database-all-500-1ont.tsv
Noeuka_semi-database-all-5000-1ont.tsv
Noeuka_semi-database-best-100-1ont.tsv
Noeuka_semi-database-best-1000-1ont.tsv
Noeuka_semi-database-best-10000-1ont.tsv
Noeuka_semi-database-best-500-1ont.tsv
Noeuka_semi-database-best-5000-1ont.tsv
No-Euka-Pure-DB
0.2
0.4
0.6
0.8
1.0

(c)


Vote_n/ Align_n Ratio
Full-Pure-DB
Full_pure-database-all-100-1ont.tsv
Full_pure-database-all-1000-1ont.tsv
Full_pure-database-all-10000-1ont.tsv
Full_pure-database-all-500-1ont.tsv
Full_pure-database-all-5000-1ont.tsv
Full_pure-database-best-100-1ont.tsv
Full_pure-database-best-1000-1ont.tsv
Full_pure-database-best-10000-1ont.tsv
Full_pure-database-best-500-1ont.tsv
Full_pure-database-best-5000-1ont.tsv
Full_semi-database-all-100-1ont.tsv
Full_semi-database-all-1000-1ont.tsv
Full_semi-database-all-10000-1ont.tsv
Full_semi-database-all-500-1ont.tsv
Full_semi-database-all-5000-1ont.tsv
Full_semi-database-best-100-1ont.tsv
Full_semi-database-best-1000-1ont.tsv
Full_semi-database-best-10000-1ont.tsv
Full_semi-database-best-500-1ont.tsv
Full_semi-database-best-5000-1ont.tsv
1.0
1.5
2.0
2.5
3.0

(d)


Vote_n/ Align_n Ratio
No-Euka-Pure-DB
Noeuka_pure-database-all-100-1ont.tsv
Noeuka_pure-database-all-1000-1ont.tsv
Noeuka_pure-database-all-10000-1ont.tsv
Noeuka_pure-database-all-500-1ont.tsv
Noeuka_pure-database-all-5000-1ont.tsv
Noeuka_pure-database-best-100-1ont.tsv
Noeuka_pure-database-best-1000-1ont.tsv
Noeuka_pure-database-best-10000-1ont.tsv
Noeuka_pure-database-best-500-1ont.tsv
Noeuka_pure-database-best-5000-1ont.tsv
Noeuka_semi-database-all-100-1ont.tsv
Noeuka_semi-database-all-1000-1ont.tsv
Noeuka_semi-database-all-10000-1ont.tsv
Noeuka_semi-database-all-500-1ont.tsv
Noeuka_semi-database-all-5000-1ont.tsv
Noeuka_semi-database-best-100-1ont.tsv
Noeuka_semi-database-best-1000-1ont.tsv
Noeuka_semi-database-best-10000-1ont.tsv
Noeuka_semi-database-best-500-1ont.tsv
Noeuka_semi-database-best-5000-1ont.tsv
1.0
1.5
2.0
2.5
3.0

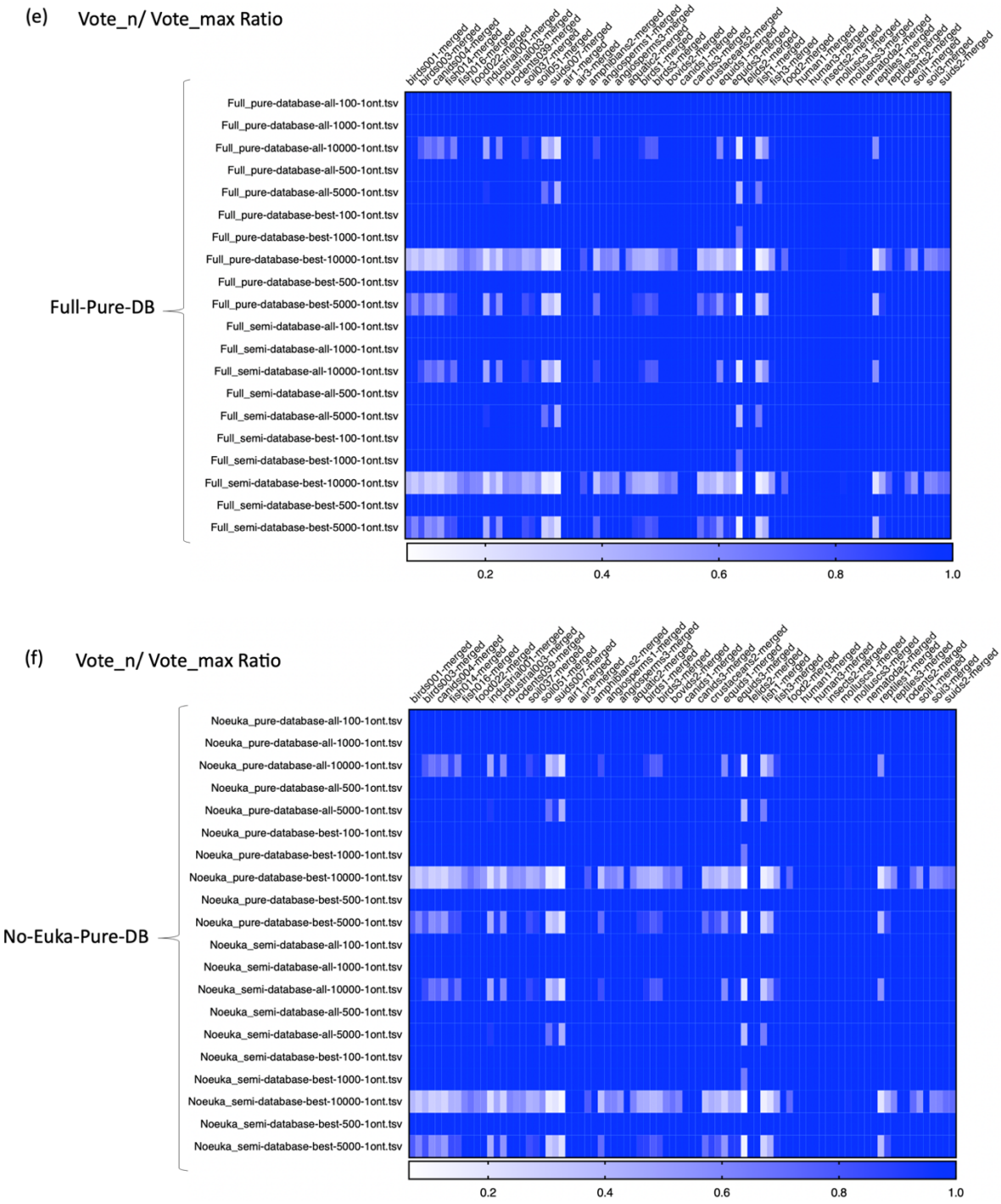


Figure 24. Visualization of alignment results for BioSamples using a heatmap. *The figure illustrates the alignment outcomes for each BioSample, detailing the metrics "read_n," "align_n," and "vote_n" computed across various filtered databases. "read_n" refers to the initial count of reads initiating the alignment process, constrained to a maximum of 20,000 reads. In contrast, "align_n" signifies the total reads successfully aligned to diverse filtered databases for each database category. Conversely, "vote_n" indicates the tally of votes garnered post-aligning reads to different databases, limited to 10,000 reads per BioSample. "vote_max" denotes the maximum possible number of votes that could be obtained, which is set to 10,000 reads per BioSample. Subfigures (a and b), (c and d), and (e and f) portray the ratios between "align_n" and "read_n," "vote_n" and "align_n," and "vote_n" and "vote_max,"*

respectively. The x-axis enumerates the distinct BioSamples, while the y-axis represents the aligned read count. These metrics furnish valuable insights into alignment efficiency and the extent of successful read mapping for each BioSample across the diverse filtered databases.

The analysis of BioSamples sourced from public libraries yielded profound insights, underscoring a significant observation: less than half of the downloaded BioSamples met the stringent criteria stipulating a maximum of 20,000 tested reads (read_n) and 10,000 mapped reads (align_n). This outcome, while not fully expected, finds its origin in the inherent uncertainties associated with the origins of these public BioSamples, predicated on our sole dependence on the accuracy of their metadata. This realization underscores the challenges posed by the intrinsic variability and diversity encapsulated within public BioSamples. Despite efforts to select representative BioSamples, the precision of their annotations can vary, leading to potential discrepancies in sample selection. Consequently, while the chosen BioSamples were deemed suitable for analysis based on available metadata, discrepancies or inaccuracies in annotation could introduce uncertainties in the interpretation of the results. Therefore, future endeavors in metagenomic analysis must address the need for improved metadata standards and robust validation processes to enhance the reliability and reproducibility of research outcomes.

Furthermore, our analysis underlined a noteworthy trend: BBmap exhibited a marked superiority in efficiently mapping a larger number of reads in private BioSamples compared to their counterparts in public library BioSamples. This observation could be linked to differences in terms of average reads quality reads within the private BioSamples when compared with their public counterparts. One plausible explanation for this discrepancy could lie in the differences in sequencing methodologies and quality control protocols between private and public datasets. Private BioSamples are originating from our sequencing protocol and show thus less variability in quality control measures, resulting in higher-quality reads with fewer sequencing errors and biases.

Another factor contributing to the enhanced mapping efficiency in private BioSamples could be the selection of targeted regions for sequencing. In our analysis, private BioSamples predominantly targeted specific hypervariable regions, such as V1-V3, known for their high taxonomic resolution and discriminatory power in metagenomic studies. In contrast, public BioSamples often encompass a broader range of regions, such as V4 or combined regions like V1-V2 or V3-V4. The focus on shorter or less informative regions in public datasets may lead

to decreased mapping efficiency and lower resolution in taxonomic assignment compared to the more targeted approaches employed in private datasets.

In addition, private BioSamples come with known and detailed metadata, including precise habitat definitions and environmental parameters, facilitating more accurate sample selection and interpretation. In contrast, public BioSamples may lack detailed metadata or have inconsistencies in habitat annotations, leading to uncertainties in sample classification and analysis.

Overall, the differential efficiency in mapping between private and public BioSamples underscores the importance of considering dataset characteristics, including sequencing methodologies, targeted regions, and metadata quality, when conducting metagenomic analyses. While public datasets offer extensive diversity and accessibility, private datasets may provide higher-quality data with more targeted sequencing approaches and better-defined sample metadata. Future studies should aim to expand the analysis of public datasets to mitigate the quality issues by relying on a greater number of samples.

Delving into the specifics, the mapping percentages of all BioSamples in Pure-db, spanning a range of 53.55% to 99.88%, and those in Semi-db, ranging from 53.79% to 99.88%, showcased a robust performance across the majority of the BioSamples. An exception surfaced with a lone sample displaying a mapping percentage of 14.20% under both Pure and semi-databases. While anomalous, this outlier underscores the importance of methodological transparency and cautious interpretation of results within the broader context of the dataset. These findings collectively instill confidence in the robustness of our conditions and offer a green light for the subsequent confident analysis of these BioSamples (refer to Table 6 for a detailed breakdown).

Moreover, our comprehensive analysis extended beyond the individual databases, revealing a consistent outcome in the mapping process across all databases employed for aligning BioSamples. This implies that once successful alignment was attained in one database, the utilization of other filtered databases did not exert a statistically significant impact on the resultant mapping outcomes. This uniformity across databases underscores the reliability and consistency of our alignment methodology, fostering a nuanced understanding of the interplay between the databases and the intricacies of BioSample alignment. In essence, this aspect reinforces the robustness of our approach and contributes valuable insights into the intricacies

of utilizing diverse databases for aligning BioSamples, paving the way for a more informed and nuanced interpretation of the ensuing analytical outcomes.

Table 6. Result of the MGST method showing the corresponding source results in public and private BioSamples. *(a) 60 public BioSamples and (b) 24 private BioSamples.*

BioSamples (public samples)	Samples source	Percentage of reads mapped in Pure-db	Percentage of reads mapped in Semi-db
Biosample_SAMEA104119060	Bird1_public_data	87.65%	87.23%
Biosample_SAMEA104119061	Bird2_public_data	80.88%	86.22%
Biosample_SAMEA104119062	Bird3_public_data	82.52%	79.35%
Biosample_SAMEA104324752	Amphibian1 public_data	91.97%	92.11%
Biosample_SAMEA104324753	Amphibian2 public_data	96.55%	96.62%
Bosample_SAMEA104324754	Amphibian3 public_data	86.64%	86.79%
Biosample_SAMEA104569562	Food1 public_data	99.34%	99.32%
Biosample_SAMEA104569563	Food2 public_data	53.55%	53.79 %
Biosample_SAMN14111925	Food3 public_data	99.05%	99.05%
Biosample_SAMEA2628033	Reptile1 public_data	99.77%	99.77%
Biosample_SAMEA2628037	Reptile2 public_data	99.68%	99.67%
Biosample_SAMEA2628038	Reptile3 public_data	99.05%	99.05%
Biosample_SAMEA3718006	Angiosperm3 public_data	71.40%	71.12%
Biosample_SAMEA3718008	Angiosperm1 public_data	73.66%	73.26%
Biosample_SAMEA3718009	Angiosperm2 public_data	70.75%	71.04%
Biosample_SAMEA3730380	Suids3 public_data	99.15%	99.17%
Biosample_SAMEA3730381	Suids2 public_data	99.48%	99.54%
Biosample_SAMEA3730385	Suids1 public_data	99.23%	99.23%
Biosample_SAMEA3898152	Nematode1 public_data	98.55%	98.55%
Biosample_SAMEA3898156	Nematode2 public_data	99.09%	99.09%
Biosample_SAMEA3898157	Nematode3 public_data	99.40%	99.40%
Biosample_SAMEA4731273	Human3 public_data	99.88%	99.88%
Biosample_SAMEA4731274	Human2 public_data	99.78%	99.78%

Biosample_SAMEA4731275	Human1 public_data	99.88%	99.88%
Biosample_SAMEA5190691	Aquatic3 public_data	91.24%	91.31%
Biosample_SAMEA5190693	Aquatic1 public_data	97.15%	97.23%
Biosample_SAMEA5190694	Aquatic2 public_data	92.69%	92.69%
Biosample_SAMN11353304	Fish1 public_data	84.18%	84.31%
Biosample_SAMN11353175	Fish2 public_data	70.33%	71.33%
Biosample_SAMEA5312021	Fish3 public_data	84.63%	84.68%
Biosample_SAMEA5527982	Insect1 public_data	99.69%	99.68%
Biosample_SAMEA5527988	Insect3 public_data	99.57%	99.52%
Biosample_SAMEA5527989	Insect2 public_data	99.79%	99.80%
Biosample_SAMEA5933334	Equids2 public_data	92.23%	92.26%
Biosample_SAMEA5933338	Equids3 public_data	91.99%	92.11%
Biosample_SAMEA5933346	Equids1 public_data	91.96%	91.98%
Biosample_SAMEA6506101	Soil1 public_data	82.80%	83.26%
Biosample_SAMN12634780	Soil2 public_data	97.27%	97.25%
Biosample_SAMN11534979	Soil3 public_data	87.52%	87.79%
Biosample_SAMN11478942	Felids2 public_data	95.94%	96.02%
Biosample_SAMN11478943	Felids3 public_data	95.38%	95.44%
Biosample_SAMN11478944	Felids1 public_data	14.20%	14.20%
Biosample_SAMN12996799	Rodent1 public_data	98.00%	98.07%
Biosample_SAMN12996800	Rodent2 public_data	98.13%	98.16%
Biosample_SAMN12996801	Rodent3 public_data	86.93%	86.94%
Biosample_SAMN13528040	Canids1 public_data	99.79%	99.80%
Biosample_SAMN13528043	Canids2 public_data	99.76%	99.77%
Biosample_SAMN13528046	Canids3 public_data	99.86%	99.87%
Biosample_SAMEA3255619	Molluscs1 public_data	99.61%	99.61%
Biosample_SAMEA3255822	Molluscs2 public_data	97.15%	97.15%
Biosample_SAMEA3255823	Molluscs3 public_data	89.50%	89.49%
Biosample_SAMN11156900	Crustacean1 public_data	97.51%	97.50%
Biosample_SAMN	Crustacean2	97.67%	97.67%

11156901	public_data		
Biosample_SAMN 11156902	Crustacean3 public_data	94.59%	94.72%
Biosample_SAME A4647967	Air1 public_data	99.46%	99.72%
Biosample_SAME A4647971	Air2 public_data	98.86%	99.34%
Biosample_SAME A4647976	Air3 public_data	98.20%	99.57%
Biosample_SAMN 13152663	Bovids1 public_data	92.31%	92.54%
Biosample_SAMN 13152664	Bovids2 public_data	92.43%	92.55%
Biosample_SAMN 13152665	Bovids3 public_data	92.57%	92.82%

BioSamples (private sample)	**Samples source**	**Percentage of reads mapped in Pure-db**	**Percentage of reads mapped in Semi-db**
Biosample_2017/01299/006_001	Suids1_private_data	83.65%	83.74%
Biosample_2017/01299/007_001	Suids2_private_data	87.53%	87.64%
Biosample_2017/01299/008_001	Suids3_private_data	70.66%	70.87%
Biosample_2019/06312/037_001	Soil1_private_data	80.09%	80.19%
Biosample_2019/06312/042_001	Soil2_private_data	83.25%	83.28%
Biosample_2019/06312/051_001	Soil3_private_data	86.08%	86.01%
Biosample_2017/05536/038_001	Rodents1_private_data	73.48%	73.44%
Biosample_2017/05536/039_001	Rodents2_private_data	72.46%	72.50%
Biosample_2017/05536/040_001	Rodents3_private_data	75.72%	75.69%
Biosample_2015/02943/001_003	Industriel1_private_data	68.56%	68.50%
Biosample_2015/02943/001_003	Industriel2_private_data	79.88%	79.91%
Biosample_2015/02943/001_003	Industriel3_private_data	70.91%	70.91%
Biosample_2019/03342/001_001	Birds1_private_data	87.19%	87.23%
Biosample_2019/03342/002_001	Birds2_private_data	86.01%	86.22%
Biosample_2019/03342/003_001	Birds3_private_data	79.21%	79.35%
Biosample_2018/02975/014_002	Fish1_private_data	78.09%	77.99%
Biosample_2018/02975/015_002	Fish2_private_data	60.90%	60.89%
Biosample_2018/02975/016_002	Fish3__private_data	79.43%	79.92%
Biosample_2016/06867/006_001	Canids1__private_data	87.24%	87.61%

Biosample_2016/06867/004_001	Canids2_private_data	84.73%	84.89%
Biosample_2016/06867/002_001	Canids3_private_data	84.19%	84.46%
Biosample_2017/08050/021_001	Food1_private_data	94.28%	94.46%
Biosample_2017/08050/022_001	Food2_private_data	92.71%	92.97%
Biosample_2017/08050/023_001	Food3_private_data	91.74%	92.11%

The alignment process generated files with the results for both public and private BioSamples, which are available on Figshare for review at the following DOI: (10.6084/m9.figshare.24499963).

III.3.3 Biotope identification

III.3.3.1 MGST result

In this investigation phase, we focused on validating our database using validation software by identifying the BioSamples with the expected outcomes in our analysis. A key metric for this validation was the mapping percentage, which indicates how well the reads from each BioSample aligned with our database. A mapping percentage below 50% suggests potential issues, such as low sequencing depth or poor match between the sample reads and the reference sequences in our database.

To evaluate the tool's performance, we analyzed the consistency of responses across different parameter combinations and assessed how often the correct source was identified. Figure 24 illustrates the "1 onto" source result, showing the specific parameter combinations used for each BioSample and their identified sources. A BioSample was successful if it met **two criteria:** correctly **identifying the target source** and aligning with all **1200 parameter** combinations. This stringent requirement ensures the reliability of the results.

The analysis, illustrated in Figure 25, revealed that BioSamples from aquatic environments, birds, bovids, equids, rodents, soil, and suids successfully met both key criteria, accounting for 30% of the total samples. In contrast, a substantial portion of BioSamples did not fully meet both conditions: 48% met only one criterion either the identification of the target source or achieving the same result across the combination of 1,200 parameters, such as those from

human. Additionally, 22% of BioSamples, such as those from reptiles, failed to meet either condition, indicating greater variability or challenges in accurately predicting their biotopes.

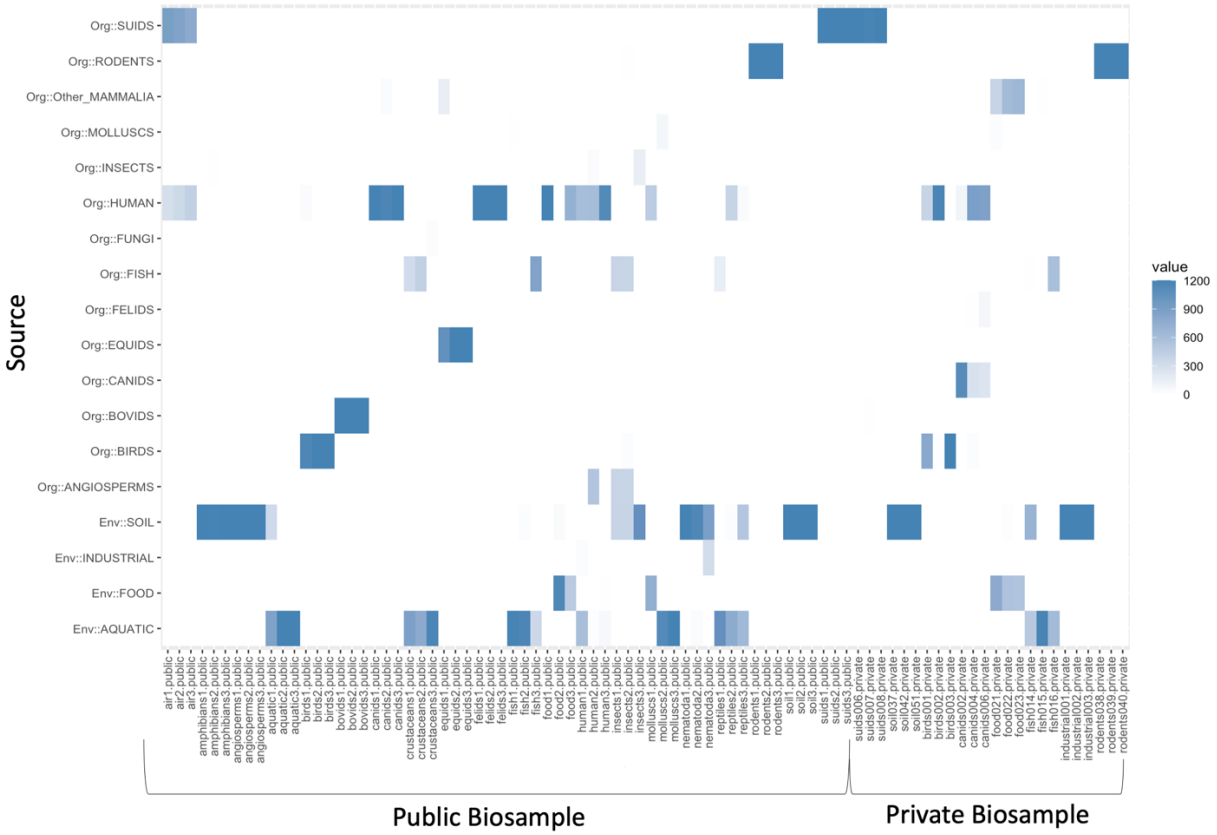


Figure 25. Comprehensive Mapping Results Heatmap. *This heatmap presents the outcomes of our sophisticated mapping tool, which seamlessly integrated 1200 parameters from both public and private BioSamples to predict the data source. Each row represents a specific BioSample, while each column signifies a predicted source.*

- *Color Intensity: The intensity of color within each cell reflects the strength of the prediction. Darker colors indicate a higher number of parameters contributing to the prediction, whereas lighter colors suggest fewer parameters.*
- *Color Scale:A color scale on the right-hand side of the figure provides a visual guide, allowing easy interpretation of the color-coding scheme.*
- *X-Axis: The x-axis represents individual BioSamples, offering a comprehensive overview of the datasets analyzed.*
- *Y-Axis: The y-axis displays the predicted sources, providing insight into the diverse origins anticipated by our mapping tool.*
- *Grouping: BioSamples are thoughtfully grouped based on their origin, with distinct sections for public and private BioSamples. This organization aids in the comparative analysis of predicted sources between the two categories.*

This legend aims to enhance the clarity and interpretability of Figure 25, facilitating a better understanding of the mapping results. BioSamples from the food and human groups identified

the correct source but did not align with all parameter combinations, indicating partial success. Conversely, samples from felids and industrial origins matched all 1200 combinations but failed to identify the correct source. Outliers, such as nematode, reptile, crustacea, and air samples, did not meet either condition. These findings highlight the varying degrees of alignment and compliance among the BioSamples, offering insights into their source identification dynamics.

To provide a clearer understanding of our findings, let's look at specific examples of BioSamples in figure 25. For instance, the amphibian BioSamples has been assigned to “Env::SOIL” source using full compliance with all 1200 parameters. Similarly, the human BioSample linked to “Org::HUMAN” met 1132 parameter combinations. The remaining parameters in human samples showed some alignment with “Env::AQUATIC” and "Org::ANGIOSPERMS”. This variation highlights differences in compliance levels across the BioSamples analyzed.

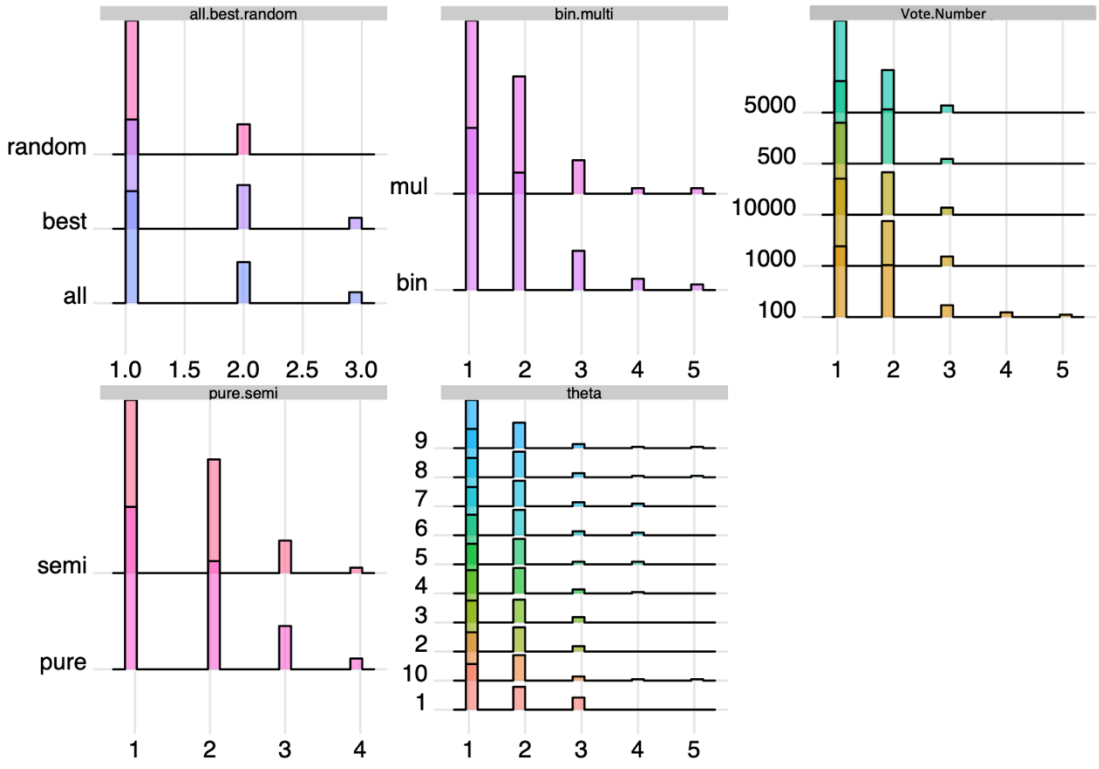


Figure 26. Number of sources among different parameter options. *This figure displays the number of sources found in different parameter options used in the study. Each horizontal bar represents a parameter option, and the x-axis of bars represents the total number of sources detected. The x-axis displays the different parameter options, and the y-axis represents the total number of sources detected for each parameter option. For instance, the “random”, “best”, and “all” parameters each identified one and two sources respectively. Notably, the “random”*

parameter ceased identification after two sources. This visualization reveals a consistent pattern across all parameters, indicating a uniform discovery rate.

Figure 26 illustrates the number of sources detected across various parameter settings used in the study. In this figure and in the top left panel, for the “random”, “best”, and “all” settings, a consistent trend is observed with one or two sources identified, with the “random” setting ceasing detection after two sources. In contrast, the “bin” and “mult” parameters (top middle panel) show broader variability, identifying up to four sources. The “vote_n” parameter (top right panel) reveals a wider range of source detections, capturing between one and five sources, with a notable concentration of samples around three and four sources.

The bottom panels show further parameter comparisons between “pure” and “semi” configurations, where both options display similar trends, with most samples detecting between one and two sources, though the “semi” configuration shows a slightly wider spread, occasionally detecting up to four sources.

Overall, the visualization demonstrates consistency in source detection across most parameters, though some settings (like “vote_n”) reveal greater variability. This consistent pattern suggests a stable identification process across different parameter configurations, with specific settings offering more granular insights into the number of biotopes detected.

These findings illustrate the varying patterns of compliance among the BioSamples, reflecting the complexity and diversity in their classification. To deepen our analysis, we introduced two key metrics: **MajorSource_frequency** and **RightSource_frequency** to compare how frequently the major and correct sources were identified. Figure 27 shows that individual parameters such as "binarize," "theta," "max-vote-x," "prior-type," and "ambiguous" had minimal impact on both MajorSource_frequency and RightSource_frequency. This consistency across all samples suggests that the parameters had little influence on the classification outcomes, reinforcing the stability and reliability of our method despite varying parameter settings.

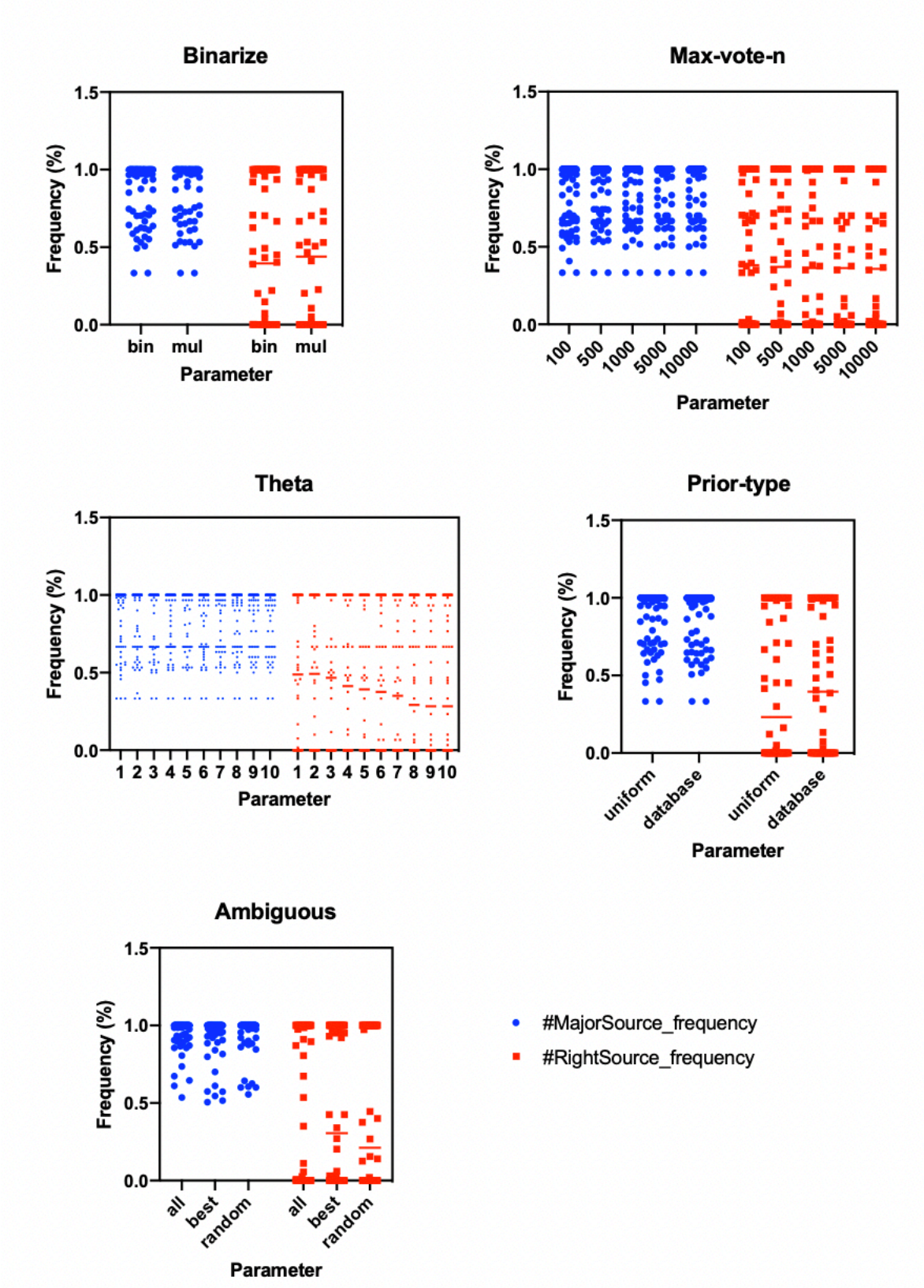


Figure 27. Performance of parameters across different sample groups. *This figure illustrates the performance of parameters across different sample groups, where the influence of parameters on classification was evaluated under two metrics: the frequency of the major source ("MajorSource_frequency") in red and the frequency of the right source ("RightSource_frequency") in blue. Each data point in the figure represents an individual*

sample. The results are shown across five different parameters (binarize, max_vote, theta, Prior-type and ambiguous). The x-axis represents the three replicas in group of source samples used in the study, and the y-axis displays the relative performance of each parameter option, with higher values indicating better performance. The bars are color-coded according to the frequency of the major source of the classifier and the right source of the sample.

The analysis of the validation process highlighted a quite robust independence of source prediction accuracy from the variations in parameter selection within our tool. A meticulous investigation was undertaken to scrutinize the potential impact of alterations to key parameters, including "binarize," "theta," "max-vote-x," and "ambiguous," across both "pure" and "semi" filter databases. Intriguingly, our findings unveiled a consistent trend wherein changes to these parameters exhibited minimal influence on the number of predicted sources, capped at a maximum of five sources.

However, amidst the relative stability observed, it was discerned that specific parameter adjustments could indeed enhance the accuracy of source prediction for select samples. Notably, the inclusion of the "random" option for the "ambiguous" parameter, coupled with setting the maximum vote to 5000 for the "max-vote-x" parameter, yielded notable improvements. This optimized configuration led to a discernible refinement, resulting in a reduced maximum of three different predicted sources, as visually highlighted in Figure 26.

The efficacy of parameters combinations to reach the "vote_max" ratio was assessed for both "1 onto" and "2 onto" parameters across all BioSamples. Figure 28 shows that certain combinations of parameters outperformed individual parameter settings, underscoring synergy between parameters. Further scrutiny of parameter dynamics reveals that the ‘2 onto’ variation exhibit a broader success spectrum in source identification, capturing between 30 to 34 BioSamples. This contrasts with the ‘1 onto’ variation, indicating a more expansive and effective range for the former. The combination of parameters demonstrates a methodology akin to both ‘1 onto’ and ‘2 onto’ approaches. Significantly, while the outcomes of the ‘2 onto’ variation manipulations echo those of the ‘1 onto’, the key distinction lies in the ‘2 onto’ variation ability to yield a higher quantity of dependable results.

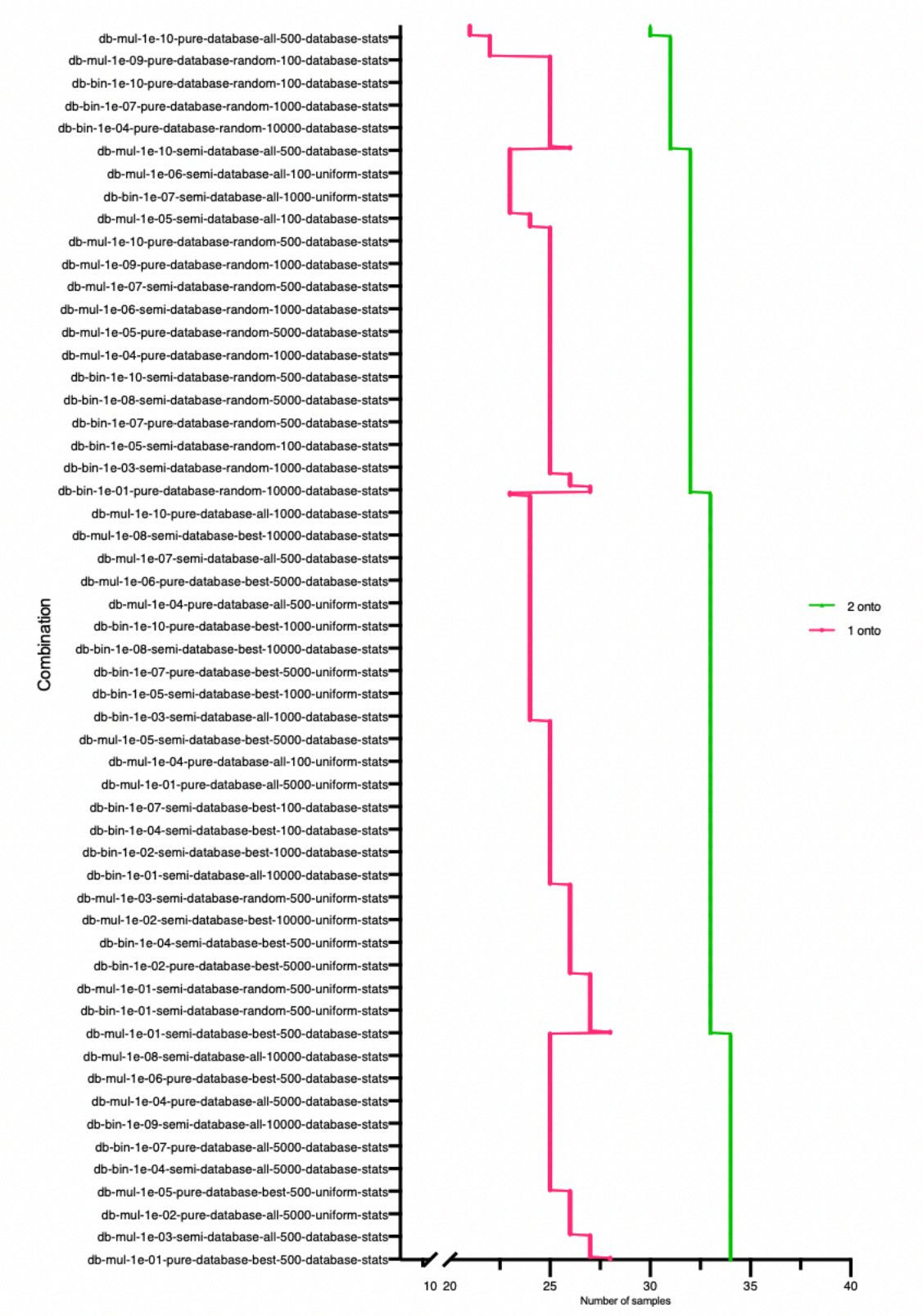


Figure 28. Optimal Parameter Combinations for BioSamples using "1 onto" and "2 onto" Parameters. *In the visualization, the pink shading corresponds to the count of BioSamples where the "vote-max" was achieved for the "1 onto" result. The green shading signifies the count of BioSamples wherein the "vote-max" value was attained for the "2 onto" result. This*

depiction offers a clear and insightful representation of the parameter combinations leading to successful source prediction outcomes.

Intriguingly, certain combinations of parameters exhibited marginal yet discernible enhancements in performance, with incremental improvements ranging from 1% to 10% when transitioning from the ‘1 onto’ to the ‘2 onto’ variation. Yet, our tool shows a consistent efficacy across the parameter combinations. Its resilience to most parameter variations emerges as a noteworthy and valuable characteristic, simplifying the user experience by mitigating the necessity for fine parameter optimization or frequent inquiries.

III.3.3.2 ST result

The results of the Sequence Tracking tool (ST) demonstrated some promising processing outcomes in the BioSamples collected for database validation. However, to further assess the tool’s performance and reliability, we decided to conduct two additional studies, as outlined in the experimental section of this thesis. These studies will provide a more comprehensive evaluation of ST's effectiveness across different datasets and conditions

III.4. Discussion and Conclusion

This work challenges the prevailing notion that bacterial 16S rDNA records in genetic databases are of limited utility without formal taxonomic classifications. To address this, we developed and validated the "BiotopeBac-DB" database, a novel resource that links genetic sequences to environmental or host-related keywords. Complementing this, we introduced the MetaGenomic Source Tool (MGST), which accurately identifies the original biotope of unknown 16S rDNA amplicon profiles, and the Sequence Tracking tool (ST) to quantify the contribution of various sources within a sample, thereby enriching microbiota analysis. While both tools demonstrated significant value when applied to public and private datasets, this discussion focuses on the results generated by the MGST tool, as the outcomes of ST are addressed in separate studies.

Amplicon sequencing is widely employed for bacterial taxonomic identification due to its targeting of specific genomic regions, such as the 16S rRNA gene, which are easier to amplify and have well-established methodologies. This approach is particularly advantageous because high-throughput sequencing platforms achieve higher accuracy with shorter reads. Additionally, partial sequences require less computational power for processing and storage, making them a cost-effective option. The standardization of these partial sequences further facilitates consistent comparisons across diverse studies. In essence, the ease of amplification, established methods, cost-efficiency, and the ability to standardize and compare data are key reasons why amplicon sequences are widely accessible and extensively used in bacterial taxonomy. The selection of 16S rDNA as a genetic marker for taxonomic identification is not arbitrary; bacterial taxonomy is now fundamentally built on this phylogenetic marker. Even when a 16S rDNA sequence is not strictly identical to a known, cultivable bacterial species, its level of nucleotide sequence identity can still be used to assign it to a specific taxonomic group.

Despite these advantages, microbiology faces significant challenges in accurately identifying bacterial species, particularly when relying on conventional methods such as biochemical properties or colony morphology, which are often unreliable. Even with properly processed 16S rDNA sequencing, there is a tendency to assign the identity of the closest homolog, even when the similarity falls below 99%. This approach, though tempting, is fraught with errors, especially when sequences are submitted to databases where species tagging and metadata rely solely on the submitter’s input. Such practices introduce substantial inaccuracies into the system. Furthermore, taxonomic identification methods like BLAST frequently return annotations such as ‘unknown bacteria’ or ‘uncultured bacteria’, particularly when the bacteria

originate from environmental sources that cannot be cultured in the lab. To address these limitations, we propose bypassing species labeling altogether and directly linking 16S sequences to their biotopes, providing a more accurate representation of bacterial population sources.

The construction of the BiotopeBac-DB database involved a meticulous examination of each sequence prior to its inclusion, scrutinizing key criteria such as the exclusion of chimeras, sequence quality scores, and the availability of source information metadata or links to scientific publications from which pertinent details could be extracted. This rigorous methodology ensures that the preserved sequences not only align with the bacterial kingdom but also possess the necessary quality and contextual information, laying a solid foundation for meaningful analysis.

In essence, this work not only introduces a comprehensive 16S database and classifier tool but also challenges the prevailing constraints associated with taxonomic information gaps. By leveraging the relationships between genetic sequences and environmental or host-associated keywords, our methodology offers a practical approach for researchers to extract valuable insights from 16S rDNA records, transcending traditional taxonomic limitations. This chapter lays the groundwork for subsequent explorations, emphasizing the transformative potential of our approach in unlocking the wealth of information embedded in bacterial genetic databases.

By bypassing species labeling and directly linking 16S rDNA sequences to their respective biotopes, we establish a robust foundation for the accurate assignment of bacterial population sources. This strategic decision not only mitigates errors commonly associated with species tagging but also aligns with the broader goals of our research, advancing the exploration of microbial diversity in varied environmental contexts.

In this chapter, we undertook a manual annotation of the terminology, which serves as a foundational step in constructing a well-curated and precise database tailored to the objectives of our research. This method was crucial in ensuring that the database was aligned with our goals, particularly in capturing the complexity of bacterial diversity and their habitat associations. By manually curating the data, we ensured a high level of specificity and reliability, setting the groundwork for subsequent analytical processes aimed at deriving meaningful insights.

Although established ontologies like the Environment Ontology (ENVO) and OntoBiotope offer comprehensive frameworks for environmental and biological contexts, they proved insufficient for our specific requirements. These ontologies lack the granularity needed to capture detailed organism biotope information, particularly concerning host species or the names of living organisms associated with the biotope (Buttigieg et al., 2013; Bossy et al., 2013). As our study necessitated precise biotope information from the metadata, particularly regarding animal hosts and species-level specificity, relying solely on existing ontologies would have introduced gaps in our data.

Therefore, we opted for a manually curated annotation approach to fill these gaps and cover biotope information more comprehensively. This method allowed us to integrate detailed metadata on organism-host interactions and environmental factors, ensuring that our database accurately reflects the complexity of the biological and ecological contexts under investigation. Manual curation, though time-intensive, provided the flexibility needed to address the limitations of pre-existing ontological frameworks, thereby enhancing the depth and utility of the dataset for future analyses (Smith et al., 2007).

The construction of the BiotopeBac-DB database was grounded in the utilization of 16S rDNA sequences sourced from the SILVA 1.32 SSU database. This repository, comprising non-redundant DNA sequences from the GenBank database with comprehensive bacterial taxonomy links, formed the backbone of our innovative approach. Notably, our method distinguishes itself by eliminating the need for read trimming, a distinctive feature facilitated by local mapping using BBMap. This uniqueness allows the MGST to seamlessly accommodate amplicon sequencing strategies targeting various segments of the 16S rDNA (refer to Table 5 for specifics).

Starting with an exploration of our BiotopeBac-DB database, our taxonomic profiling of microbial communities from human, aquatic, and soil sources revealed overlaps and distinctions that align with findings from previous studies in microbial ecology. By analyzing the relationships between word-sets and bacterial taxonomy, we identified key bacterial phyla and genera associated with each environmental source. These results not only demonstrate the distribution of microbial taxa across different biotopes using 16S rDNA but also align with established knowledge on microbial diversity.

Our analysis of human-associated samples revealed that Firmicutes (46.20%) and Actinobacteria (22.67%) were the dominant phyla. These findings are consistent with several studies on the human microbiome. For instance, Lozupone et al. (2012) reported that Firmicutes and Actinobacteria are prominent members of the human gut and skin microbiomes, which play essential roles in digestion, metabolism, and immune regulation. Additionally, the presence of Proteobacteria (16.64%) and Bacteroidetes (11.31%) in our study reflects microbial diversity patterns commonly observed in the human microbiome. Arumugam et al. (2011) similarly identified Bacteroidetes as a significant component of the human gut, where they assist in carbohydrate metabolism. The congruence between our findings and these earlier studies underscores the robustness of our taxonomic profiling, affirming the dominance of these bacterial phyla in human-associated environments.

Our analysis of aquatic environments highlighted Proteobacteria as the most abundant phylum (45.07%), followed by Bacteroidetes (14.35%), Firmicutes (10.35%), and Cyanobacteria (6.10%). These results are in line with the findings of Zinger et al. (2011) and Thompson et al. (2017), who both reported the dominance of Proteobacteria and Bacteroidetes in aquatic ecosystems. Proteobacteria, particularly the *Alphaproteobacteria* and *Gammaproteobacteria* classes, play a key role in nitrogen and carbon cycling within aquatic systems, driving essential biogeochemical processes. The presence of Cyanobacteria, known for their role in primary production through photosynthesis, aligns with studies like those by Paerl and Paul (2012), which emphasized the significance of this group in oxygen generation and nutrient cycling within freshwater and marine environments. Therefore, our profiling of aquatic microbes is consistent with well-established patterns of microbial diversity in water systems.

In soil samples, Proteobacteria again emerged as the dominant phylum (over 30%), followed by Firmicutes (11.12%), Bacteroidetes (10.71%), and Actinobacteria (9.79%). These findings correspond with numerous studies that have characterized soil microbial communities. Fierer et al. (2012) also found that Proteobacteria and Actinobacteria are prevalent in soil ecosystems, where they contribute to organic matter decomposition and nutrient cycling. *Acidobacteria* (7.86%) were also prominent in our soil samples, a result echoed by Janssen (2006), who noted the ecological significance of *Acidobacteria* in soil, particularly in acidic environments where they facilitate the breakdown of complex organic compounds. Our findings further confirm the

taxonomic richness of soil microbiomes, with over 78 distinct phyla identified, consistent with soil's reputation as one of the most diverse microbial ecosystems on Earth.

The microbial diversity observed in our study reflects broader ecological trends seen in microbial biogeography. The distinct taxonomic profiles between human, aquatic, and soil sources demonstrate the unique environmental pressures and biological interactions shaping microbial communities in different biotopes. Our identification of Proteobacteria as a common dominant phylum across all environments underscores its versatility and adaptability, as seen in the work of Lauro et al. (2009), who highlighted the evolutionary success of Proteobacteria in a range of ecological niches. In contrast, the high abundance of Firmicutes and Actinobacteria in human samples and their lower prevalence in soil and aquatic environments reinforces the specialization of these phyla in human-related ecosystems, as previously discussed in studies like Schloss et al. (2012).

While our results largely align with existing research, they also offer novel insights into the microbial taxonomic landscape across these environmental sources. For instance, the detailed breakdown of microbial diversity in the aquatic and soil environments reveals previously underrepresented phyla, such as Planctomycetes and Verrucomicrobia, which together constituted a significant proportion of the microbial communities in our soil and aquatic samples. These findings echo emerging studies, such as those by Sangwan et al. (2015), that suggest these phyla play previously underappreciated roles in biogeochemical cycling and environmental resilience.

Additionally, our use of the Chord plot visualization offers a unique perspective on the complex interplay between microbial taxa and their environmental sources. This innovative approach allowed us to capture not just the abundance of microbial taxa but also their ecological interconnections, providing a more holistic understanding of microbial diversity across biotopes.

In conclusion, our taxonomic profiling of microbial communities across human, aquatic, and soil sources not only corroborates findings from other studies but also extends the current understanding of microbial ecology. By identifying key microbial taxa and highlighting the distinct microbial signatures of different environments, our study contributes valuable insights into microbial biogeography and the factors shaping microbial diversity. Moreover, the robust and diverse microbial profiles observed in our study, combined with the novel visualization

approaches, lay the foundation for future research aimed at exploring the functional roles of these microbial communities in their respective ecosystems.

Our MGST, based on a statistical Bayesian approach, was validated through systematic variation of its parameters, which resulted in a high level of confidence in the classification outcomes. Notably, our extensive investigation of both individual parameters and their combinations revealed a striking observation: parameter variations had no significant effect on the classification of source BioSamples. This robustness highlights the versatility and reliability of our classifier, demonstrating its capacity to maintain consistent performance despite parameter fluctuations. As such, this method offers researchers a powerful and adaptable tool for accurately identifying biotopes associated with 16S rDNA records. The MGST tool proves particularly valuable in scenarios where the original source of a sample is unknown, such as in cases of contamination, by enabling researchers to analyze the predominant source in the sample and appropriately categorize it.

Our analysis begins with a detailed examination of the results produced by the MGST tool, specifically focusing on its efficacy in generating accurate sequence mappings with high mapping percentages. The application of MGST to a diverse set of amplicons from various BioSamples during the validation process yielded a range of outcomes.

In the context of sequence mapping using the MGST tool and our filtered database (Pure-db/Semi-db), we identified three distinct cases, each characterized by different mapping behaviors. Successful classification in MGST is evaluated based on two key criteria: the percentage of sequences mapped to our database (Table 6) and the accuracy of the predicted source (Figure 25).

The first case, representing the optimal scenario, involved BioSamples in which a significant majority—over 70% of sequences—were successfully mapped to our filtered databases (Pure-db/Semi-db) in Table 6, with the predicted source accurately matching the expected biotope (Figure 25). This high mapping percentage, combined with accurate source predictions, underscores the reliability of both the sequencing process and the MGST tool in predicting biotopes. Examples of this scenario include BioSamples from aquatic environments, equids, humans, bovids, birds, and rodents. The strong alignment accuracy observed here suggests that both the quality of the sequencing data and the comprehensiveness of the reference databases were sufficiently robust to produce reliable and accurate results. This consistency in

performance instills a high level of confidence in the biotope classifications, further reinforcing the effectiveness of the MGST tool in correctly identifying diverse habitats and host associations.

In contrast, the second case presented an interesting divergence. Despite similarly high mapping percentages (again exceeding 70%), the predicted source did not align with the anticipated biotope. Although the sequences mapped well, the source classification deviated from the initial hypothesis. This discrepancy warrants further investigation into the nuances of source prediction. Potential explanations include subtle sequence variations, unrecognized similarities between different biotopes, or limitations in the reference database that may obscure the distinctions between certain sources. This finding highlights the need for continuous refinement of both the tool and database, as well as further biological interpretation of unexpected alignments.

The third case involved suboptimal mapping results, where less than 70% of sequences were mapped, and the source predictions were similarly unexpected. For example, BioSample_SAMN11478944 exhibited only a 14% mapping success rate, with a source prediction that did not align with any anticipated biotope. Such cases underscore the need for further investigation into both the quality of the sequencing data and potential gaps in the reference database. Low mapping percentages could be indicative of incomplete or poor-quality sequence data, database limitations, or novel or rare biotopes not adequately represented in the current dataset. These instances call for deeper exploration into the discrepancies, as well as refinement of the annotation process to better capture rare or complex biotopes.

Overall, the analysis of these three cases highlights the strengths and limitations of the MGST tool. While it performs robustly in cases where sequence data and database alignment are well-matched, the second and third cases emphasize the importance of continuous refinement and validation of both the tool and the underlying data. These findings offer a pathway for future work, encouraging improvements in database coverage and accuracy, and the development of more sophisticated algorithms to handle complex or ambiguous source predictions.

Our analysis shows that the MGST tool can overcome some technical challenges presented by BioSamples, but the quality of the samples and the performance of key processes like merging (Table 5) and mapping (Table 6) are critical. While our goal was to achieve high mapping coverage and source accuracy, mapping failures can make source annotation difficult. For

instance, the "fish2" BioSample had a low merging success rate of 15.70%, with 893 reads retained, but 70% of those reads mapped to our database. In contrast, the "felids1" BioSample had an 88.60% merging rate, yet only 14% of reads mapped. This highlights the variability in results depending on both merging and mapping performance.

Biological factors also play a role in biotope identification accuracy with MGST. For example, BioSamples from canids and felids showed associations with human biotopes, consistent with studies showing shared microbiomes between humans and companion animals (Coelho et al., 2018; Wang et al., 2013). Meanwhile, BioSamples from fish, mollusks, and crustaceans mostly mapped to aquatic environments. However, when we split word sets into environmental or host-specific categories ("2 onto" parameters), most results matched expectations except for certain crustacean samples. For instance, BioSamples from a buccal swab of Penaeus monodon hepatopancreas (Devesse et al., 2017) showed inconsistencies, likely due to limited metadata. This lack of detail makes it harder to accurately identify alternative sources.

An unexpected finding in our analysis emerged from BioSamples labeled with an "air" origin, where a surprising association with suid (pig) organism sources was identified. Upon closer examination of the metadata, it became evident that these samples were derived from the nasal regions of pig farmers. This intriguing discovery suggests that individuals in close contact with livestock, such as farmers, may harbor microbial communities characteristic of their animal hosts, potentially due to zoonotic microbial transfer (Kastenbauer et al., 2019; Keestra et al., 2019). Similarly, BioSamples from angiosperms, amphibians, nematodes, and industrial sources exhibited associations with soil-origin bacteria. This observation is consistent with the well-documented phenomenon of bacterial transfer from soil to organisms, particularly those inhabiting soil-rich environments (Fierer et al., 2012; Liu et al., 2016).

Moreover, a parallel pattern was observed in samples from companion animals, including felids and canids, which unexpectedly showed an association with human-origin bacteria. This may reflect the close physical and environmental interactions between humans and their pets, facilitating the exchange of microbiota (McKenzie et al., 2016; Salonen et al., 2017). These nuanced findings highlight the complexity of host-symbiont relationships and underscore the critical impact of accurate metadata annotation. Mislabeling or incomplete metadata, such as indicating the nasal origin of pig farmers without specifying their occupational exposure, can lead to misinterpretation of microbial sources and biotope associations (Knight et al., 2018; Quinn et al., 2016).

The resultant high mapping percentages, despite deviating from expected outcomes, emphasize the importance of accurately determining the true origin of samples for the reliable interpretation of their ecological relevance and probabilistic associations in our analysis. The multifaceted nature of these interactions encapsulates the complexity inherent in biotope mapping, intertwining metadata precision, symbiotic intricacies, and the challenges associated with high-throughput sequencing processes. This understanding enhances our ability to accurately annotate biotopes, thereby advancing the field of microbial ecology and its applications in environmental and veterinary sciences.

A key feature of our database tools is their reliance on detailed metadata for accurate annotations. The success of the classification process is directly linked to the completeness and relevance of the available metadata. Incomplete or missing metadata can hinder the accuracy of classifications, as seen in the case of crustacean BioSamples, where limited data (e.g., only buccal swabs) restricted source detection. Access to additional BioSamples, such as fecal or skin samples from the same organisms, with pertinent metadata, would allow for more accurate annotations.

Our investigation into parameter variations in BioSamples revealed valuable insights into source prediction. We observed consistently high average values of merged reads in public BioSamples, particularly in aquatic, fish, and soil samples, while private BioSamples, such as industrial and suids, presented more variability. The subsequent alignment process, using databases with and without filtered eukaryotic sequences, highlighted BBmap's effective performance, especially in private BioSamples, indicating higher-quality reads compared to public datasets.

In terms of source prediction, the heatmap (Figure 25) shows the predicted sources for each BioSample, reflecting strong prediction accuracy. Despite parameter variations, the analysis showed minimal impact on prediction accuracy, confirming the robustness of our approach. For instance, amphibian BioSamples with 1200 combined parameters showed high compliance, while the accuracy of MajorSource_frequency and RightSource_frequency remained largely unaffected by parameter changes (Figure 27).

However, some BioSamples, particularly in aquatic environments, showed limitations in predicting host biotopes, emphasizing the complex microbial interactions between hosts and

their environments. As noted by Deines et al. (2020), public metadata and sequencing data may not fully capture these interactions, introducing potential uncertainties.

Compared to other tools like SeqEnv, which predicts environmental sources based on 16S rRNA data (Sinclair et al., 2016), our MGST demonstrated greater resilience to parameter variations and offered more consistent performance across diverse BioSample types. While SeqEnv is effective for environmental classifications, our tool's integration of more detailed metadata allowed for improved source predictions in host-associated biotopes, particularly when metadata is complete.

Moreover, our tool's ability to maintain high prediction accuracy despite parameter variations makes it a reliable and adaptable option for researchers. The insights gained from this analysis not only enhance our understanding of BioSample characteristics but also reinforce the strength of our source prediction tool, which remains effective across a range of BioSamples and environmental contexts.

The MGST tool can be particularly valuable for identifying the source of unknown or contaminated samples in various veterinary science applications. One key application is tracking the source of infectious diseases, which is essential for managing zoonotic pathogens capable of crossing between animals and humans. By pinpointing the specific environmental source or intermediate host, MGST offers critical insights that help control disease outbreaks and prevent future occurrences, especially in large-scale farming operations or wildlife populations.

MGST could also play a significant role in classifying microbial communities in animal microbiome studies, which are essential for understanding the overall health and disease resistance of animals. By analyzing environmental samples, such as water, feed, and soil, MGST can help identify microbial communities that influence the gut health of livestock or domestic animals. This information enables more targeted interventions, promoting animal welfare through better management of microbial environments.

Furthermore, the tool is invaluable for ensuring food safety and quality by tracing contamination sources in meat and dairy production. Contamination from bacteria can occur at various stages, including production, storage, or processing. MGST can be used to screen environmental samples to identify the source of contamination, thus helping to ensure the safety

of food products and preventing the transmission of harmful pathogens from animals to humans. For instance, in veterinary science and in cases of mastitis in dairy cows, a common and economically significant disease, identifying the bacterial source is crucial. Mastitis is often caused by bacteria such as Staphylococcus aureus or Escherichia coli, which can originate from various sources, including contaminated bedding, milking equipment, or the cows' own environment. When an outbreak occurs and the bacterial source is unknown, applying the MGST classifier to environmental samples such as feed, water, bedding, or milking equipment can help determine the primary contamination source.

In addition, MGST aids in environmental monitoring, which is a key concern in veterinary science. Monitoring the health of animals across diverse environments such as farms, wildlife reserves, and aquatic ecosystems requires understanding how microbial sources from water, soil, or air impact animal health. MGST helps researchers classify these microbial sources, thereby aiding in the assessment of environmental risks and habitat degradation. For example, when studying emerging or re-emerging diseases, MGST can distinguish between normal microbial flora and external pathogens introduced through contaminated feed, animal migration, or human interaction. This capability is especially critical for biosecurity measures in livestock management and wildlife conservation, providing actionable insights to protect animal populations and mitigate health risks.

By analyzing the majority microbial content in the samples, the MGST tool can reveal whether the bacteria came from external environmental contamination or if it was introduced through improper hygiene practices or contaminated feed. This insight allows for targeted intervention measures, such as adjusting farm management practices, improving sanitation protocols, or altering feed sources to mitigate future contamination risks. Identifying the original contamination source helps limit the spread of the disease, protecting animal health, reducing economic losses, and safeguarding food quality.

Our MGST tool, like any analytical method, has its limitations. A key distinction is that our approach doesn't focus on identifying the exact origins of individual sequences. Instead, we aggregate the most prevalent source information across all sequences in a sample, offering a broad view of the dominant sources shaping the microbial composition. While it's possible to analyze individual sequences, our primary focus was on understanding the broader biotope patterns within each sample.

Our database, while effective, is a preliminary version and contains limitations. It is primarily built from GenBank 16S rDNA records, covering well-studied environments such as aquatic, soil, and human biotopes. This reliance on available data can introduce biases, as the database reflects the existing distribution of sequences rather than a truly comprehensive scope. Moreover, the bacterial taxonomy used is based on the SILVA database (version 1.38.1) as of early 2022, and while it remains relevant, subsequent updates like the inclusion of the Genome Taxonomy Database (GTDB) offer more refined taxonomic insights that are not yet fully integrated into our system.

Compared to existing tools like BioNLP and SeqEnv, MGST offers a streamlined, source-focused approach. BioNLP excels at extracting and linking textual information from the literature to biological entities (Bossy et al., 2013), but its strength lies in textual data mining rather than direct sequence-based biotope identification. SeqEnv, on the other hand, is designed to assign environmental contexts to metagenomic sequences (Sinclair et al., 2016), much like MGST, but focuses more on specific taxonomic groups rather than the broader source-level identification approach that we use. Our approach fills the gap by offering a higher-level source aggregation, particularly useful in cases where the original biotope of a sample is unknown or ambiguous.

The current classifier in MGST, optimized for source prediction, splits biotopes into environmental or host-related categories. While SILVA provides high-quality sequences and metadata, the MGST tool categorizes keywords into wordsets, helping to structure and refine the biotope predictions. However, there is room for improvement, especially in increasing the database's granularity by adding more sequences and refining the wordsets for better accuracy.

Moving forward, refining MGST and expanding the BiotopeBac-DB database are crucial. Future improvements could focus on increasing the precision of source predictions, especially when dealing with hypervariable regions of the 16S rDNA amplicon. Continuous updates to the database will ensure it remains relevant for both current and future microbiome research.

In conclusion, this chapter demonstrates the effectiveness of our approach for biotope identification using 16S rDNA amplicon data. By focusing on majority sequences within a sample, we provide a practical tool for researchers to quickly identify biotopes without the need for exhaustive manual taxonomy or literature interpretation. However, ongoing refinement of the database and further exploration into individual sequence associations with specific

organisms will be critical in advancing the tool's capabilities, contributing to the continued evolution of microbiological research.

Chapite IV : Application

Study 1: Application of the Sequencing Tracking tool (ST) to Restroom Datasets. *Article: Microbiota profiling on veterinary faculty restroom surfaces and source tracking*

Hiba Jabri; Simone Krings; Papa Abdoulaye Fall; Denis BAURAIN; Georges Daube; Bernard Taminiau

Microorganisms. 2023;11(8):2053.

microorganisms

MDPI

Article

Microbiota Profiling on Veterinary Faculty Restroom Surfaces and Source Tracking

Hiba Jabri [1], Simone Krings [2], Papa Abdoulaye Fall [3], Denis Baurain [4], Georges Daube [1] and Bernard Taminiau [1,*]

[1] Laboratory of Food Microbiology, Fundamental and Applied Research for Animals and Health Center (FARAH), Department of Food Sciences, Faculty of Veterinary Medicine, University of Liege, Quartier Vallée 2, B42, Avenue de Cureghem 10, 4000 Liège, Belgium; hibajabri88@hotmail.fr (H.J.); georges.daube@uliege.be (G.D.)
[2] Department of Microbial Sciences, School of Biosciences, Faculty of Health and Medical Sciences, Univesity of Surrey, Guildford GU2 7XH, UK; sk01450@surrey.ac.uk
[3] FoodChainID GENOMICS, Laboratory Manager NGS, Rue Hayeneux, 62, 4040 Herstal, Belgium; abdoulaye.fall@foodchainid.com
[4] Eukaryotic Phylogenomics, InBioS-PhytoSYSTEMS, University of Liège, 4000 Liège, Belgium; denis.baurain@uliege.be
* Correspondence: bernard.taminiau@uliege.be

Abstract: In this study, we aimed to develop a comprehensive microbial source amplicon database tailored for source tracking in veterinary settings. We rigorously tested our locally curated source tracking database by selecting a frequently accessed environment by veterinary students and veterinarians. By exploring the composition of resident microbiota and identifying potential sources of contamination, including animals, the environment, and human beings, we aimed to provide valuable insights into the dynamics of microbial transmission within veterinary facilities. The 16S rDNA amplicon sequencing was used to determine the bacterial taxonomic profiles of restroom surfaces. Bacterial sources were identified by linking our metadata-enriched local database to the microbiota profiling analysis using high-quality sequences. Microbiota profiling shows the dominance of four phyla: Actinobacteria, Bacteroidetes, Proteobacteria, and Firmicutes. If the restroom cleaning process did not appear to impact microbiota composition, significant differences regarding bacterial distribution were observed between male and female users in different sampling campaigns. Combining 16S rDNA profiling to our specific sources labeling pipeline, we found aquatic and human sources were the primary environment keywords in our campaigns. The probable presence of known animal sources (bovids, insects, equids, suids. . .) associated with bacterial genera such as *Chryseobacterium*, *Bergeyella*, *Fibrobacter*, and *Syntrophococcus* was also involved in restroom surfaces, emphasizing the proximity between these restrooms and the exchange of bacteria between people involved in animals handling. To summarize, we have demonstrated that DNA sequence-based source tracking may be integrated with high-throughput bacterial community analysis to enrich microbial investigation of potential bacterial contamination sources, especially for little known or poorly identified taxa. However, more research is needed to determine the tool's utility in other applications.

Keywords: microorganisms; biotopes; 16S rDNA amplicon sequencing; database; restrooms; microbial source tracking

check for updates

Citation: Jabri, H.; Krings, S.; Fall, P.A.; Baurain, D.; Daube, G.; Taminiau, B. Microbiota Profiling on Veterinary Faculty Restroom Surfaces and Source Tracking. *Microorganisms* **2023**, *11*, 2053. https://doi.org/10.3390/microorganisms11082053

Academic Editor: Ana C. Sampaio

Received: 10 July 2023
Revised: 2 August 2023
Accepted: 7 August 2023
Published: 10 August 2023

Copyright: © 2023 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

In the veterinary environment, most animals can be considered potential carriers and probably risk transmitters of bacterial pathogens by professionals. For example, zoonotic bacteria can be transferred from animals to human beings in several ways, including direct and indirect contact with a veterinarian. Buildings with high human activity in the veterinary environments (students or veterinarians) such as restrooms and nests for

bacterial exchange and dissemination (human skin, water, soil, or also animal source). Students or veterinarians can be vectors for various pathogenic bacteria, not only for themselves but also for the animals they come in contact with. Restrooms have always been regarded as potential sources of infectious diseases. Bacterial pathogens such as some strains of *E. coli*, which are often found in restrooms, can be transferred by hands, gowns, and boots to cattle, pigs, horses, dogs, or cats in a veterinary environment [1]. *Staphylococcus aureus (S. aureus)* strains were isolated from cows with mastitis, from horses and dogs with lesions, from human beings, from dogs and cats who were healthy carriers [2], and could also be found in restrooms. Transmission of *S. aureus* between human beings and animals has also been reported [3,4].

It has been demonstrated that human feces can carry a wide range of transmissible pathogens: *Campylobacter*, enterohaemorrhagic *Escherichia coli*, *Salmonella*, *Shigella*, *Staphylococcus*, and *Yersinia* as well as viruses such as norovirus, rotavirus, and hepatitis A and E, just to name a few [5]. Fecal indicator bacteria (FIB) could be used as a marker of fecal pollution and an indication of the pathogen population [6]. However, in terms of source tracking, FIB are members of bacterial groups or taxa that are ubiquitous in human and other animal feces. Therefore, they provide little or no information about specific contaminating hosts. Little is known about the other possible contaminations, either direct or indirect, from environmental or animal sources. Bacterial biogeography is mainly performed with a bacterial identification campaign in a given biotope. With this strategy, the link between microbes and biotopes can only be achieved with proper taxonomical identification, leaving out unknown bacterial populations. Moreover, uncultivable bacteria can also be ruled out if this identification process relies only on microbial culture. Additionally, linking bacteria with biotopes can be further characterized by directly linking bacterial genetic fingerprints with biogeography, even for uncharacterized or unknown populations.

Comparison of collected sequence data (CCSD) approach using a given phylogenetic target (e.g., the 16S rDNA) with existing datasets in genetic databases (using the same sequencing technologies) has already been tested and could be used as a start to identify bacterial environmental origin for anthropogenic microbial communities (e.g., human skin, soil, etc.) [7]. So far, published CCSD campaigns have been restricted to include only well-described bacterial populations and focused mainly on human-associated biotopes. The CCSD approach can be used to detect the probable bacterial contaminants in restroom environments, especially those of animal origin by adding a data connection between bacteria and eukaryotic sources (animals or plants). There is surprisingly no comprehensive database for bacterial biogeography nor a database linking eukaryote organisms as hosts for bacteria. If such databases exist, could they be a support to help us to improve source tracking studies for little known or poorly identified microbial contamination taxa?

To address this question, we designed a sampling campaign targeting restrooms considering criteria like gender, surfaces, and hygiene influence. We created a database where 16S rDNA sequences referenced in the public rDNA database SILVA v.132 [8] were linked to source metadata encoded using a controlled vocabulary (CV). This CV was constructed with an environmental annotation ontology model by adding for the first time the eukaryotic taxonomic classification for their probable host sources organisms using the vernacular term. This allowed us to enrich the microbiota profiling campaign of restrooms with probable sources of contamination in restrooms.

2. Materials and Methods

2.1. Sampled Surfaces

All samples were collected in three sampling campaigns on different dates (March 2017, March 2018, and April 2018) from two restrooms used by potentially 100 to 150 veterinarian students per day, one used by women (n = 48 samples) and one by men (n = 48 samples). Eight surfaces (door handles inside and outside of the restroom, handles inside and outside of the toilet cabin, tap faucet handle, toilet seat, toilet flush handle, and urinal flush used by men's restroom) were sampled in the restroom evenly distributed in

the same buildings at the Faculty of Veterinary Medicine in University of Liège, Belgium. Samples for each surface were taken considering two criteria of hygiene: one after the cleaning hygiene process (Clean) and the other before (Dirty) (Table 1). We replicated the samples three times.

In the cleaning process, a Sani Cud Pur Eco product from the brand Diversey was utilized, containing citric acid and surfactant agents. The cleaner applied the product and added water to effectively clean the restroom. It is essential to clarify that our emphasis on the cleaning hygiene process was to enable a diverse range of sample testing, rather than specifically evaluating its effect on microbial composition. Our primary objective was centered on exploring potential shifts in microbial sources, with a focus on enriching our source analysis, rather than assessing changes in microbial composition.

Table 1. Nature and characteristics of the samples.

Samples	Gender [1]	Surface Type [2]	Cleaning Status [3]
F_Handle_in_Clean	F (Female)	Handle_in	Clean
F_Handle_out_Clean	F (Female)	Handle_out	Clean
F_Handle_Cabin_in_Clean	F (Female)	Handle_Cabin_in	Clean
F_Handle_Cabin_out_Clean	F (Female)	Handle_Cabin_out	Clean
F_Tap_Clean	F (Female)	Tap	Clean
M_Handle_in_Clean	M (Male)	Handle_in	Clean
M_Handle_out_Clean	M (Male)	Hanlde_out	Clean
M_Handle_Cabin_in_Clean	M (Male)	Handle_Cabin_in	Clean
M_Handle_Cabin_out_Clean	M (Male)	Handle_Cabin_out	Clean
M_Tap_Clean	M (Male)	Tap	Clean
F_Flush_Clean	F (Female)	Flush	Clean
M_Flush_Clean	M (Male)	Flush	Clean
M_Flush_M_Clean	M (Male)	Urinal_Flush	Clean
F_Seat_Clean	F (Female)	Seat	Clean
M_Seat_Clean	M (Male)	Seat	Clean
F_Handle_in_Dirty	F (Female)	Handle_in	Dirty
F_Handle_out_Dirty	F (Female)	Handle_out	Dirty
F_Handle_Cabin_in_Dirty	F (Female)	Handle_Cabin_in	Dirty
F_Handle_Cabin_out_Dirty	F (Female)	Handle_Cabin_out	Dirty
F_Tap_Dirty	F (Female)	Tap	Dirty
M_Handle_in_Dirty	M (Male)	Handle_in	Dirty
M_Handle_out_Dirty	M (Male)	Hanlde_out	Dirty
M_Handle_Cabin_in_Dirty	M (Male)	Handle_Cabin_in	Dirty
M_Handle_Cabin_out_Dirty	M (Male)	Handle_Cabin_out	Dirty
M_Tap_Dirty	M (Male)	Tap	Dirty
F_Flush_Dirty	F (Female)	Flush	Dirty
M_Flush_Dirty	M (Male)	Flush	Dirty
M_Flush_M_Dirty	M (Male)	Urinal_Flush	Dirty
F_Seat_Dirty	F (Female)	Seat	Dirty
M_Seat_Dirty	M (Male)	Seat	Dirty

[1] Gender type; F: toilet female user, M: toilet male user. [2] Type surface; all site type used by human hand and skin. [3] Cleaning status; Dirty: before cleaning process, Clean: after cleaning process.

All samples were taken and processed in the same way. The choice of the middle cabin toilet for sampling was based on a study by Christenfeld N in 1995 [9]. Samples were taken with BBL™ CultureSwab™ EZ II (220145, BD, 4000, Belgium), which were moistened with AccuGENE molecular water (BE51200, Lonza, 4800, Belgium) beforehand. Samples were kept at 4 °C and brought to the lab within 1 h.

2.2. Total DNA Extraction and Sequencing Library Preparation

Less than 24 h after the sampling, total DNA extraction was performed with the DNeasy Blood & Tissue kit (69506, QIAGEN, 85764, Germany), following the manufacturer's recommendations. DNA concentration and purity assessment were carried

out with a NanoDrop™ 2000 (Thermo Fisher Scientific, Isogen Life Science B.V.,B-4000, Sart-Tilman, Belgium). PCR-amplification of the 16S rDNA V1–V3 hypervariable region and library preparation was performed with the following primers (with Illumina overhand adapters), forward (5′-GAGAGTTTGATYMTGGCTCAG-3′) and reverse (5′-ACCGCGGCTGCTGGCAC-3′). Each PCR product was purified with the Agencourt AMPure XP beads kit (Beckman Coulter; Pasadena, CA, USA) and submitted to a second PCR round for indexing, using the Nextera XT index primers 1 and 2 (Illumina 2018). After purification, PCR products were quantified using the Quant-IT PicoGreen (ThermoFisher Scientific; Waltham, MA, USA) and diluted to 10 ng/µL. A final quantification of each library was performed using the KAPA SYBR® FAST qPCR Kit (KapaBiosystems; Wilmington, MA, USA) before normalization, pooling and sequencing on a MiSeq sequencer using V3 reagents (Illumina; San Diego, CA, USA) [10]. A positive control using DNA from 20 defined bacterial species and negative control were included in the sequencing run. Samples with too low bacterial DNA content or containing PCR inhibitors were not analyzed. Out of the total 96 samples, 81 samples (34 from women and 43 from men) were categorized into 41 Dirty and 40 Clean samples. During the initial processing, we encountered challenges with insufficient DNA yield in 15 of the samples. Consequently, we had to exclude these 15 samples from further analysis, ultimately resulting in a final dataset of 81 samples. This careful curation was essential to ensure the reliability and accuracy of our data analysis, as samples with insufficient DNA could introduce potential biases or limitations in the interpretation of the results.

2.3. *Bioinformatic Analysis*

2.3.1. Microbial Profiling

During the preprocessing stage, the Illumina adapters and primers were removed from the raw data. Subsequently, to ensure data quality and reliability, we applied essential filtering criteria using the command screen.seqs as a trimming process. These stringent filtering steps were crucial for maintaining the integrity of the sequences, as they enforced a maximum allowance of 1 ambiguous base and ensured that all sequences had a minimum length of 450 nucleotides. This curation process provided a solid foundation for subsequent analyses and interpretations in our scientific investigation.

Additionally, following the data curation, we used the MOTHUR software package v1.39.5 (Schloss et al., 2009) to check for chimeric amplification using the VSearch algorithm [11]. The resulting cleaned reads were then aligned to the SILVA database v1.32 [12]. To reduce computational complexity while preserving data representation, we sub-sampled the aligned reads, retaining 10,000 reads clustered into operational taxonomic units (OTUs) using the average neighbor algorithm from MOTHUR v1.39 with a 0.03 distance cut-off [10,13]. This subsampling approach allowed us to efficiently manage the dataset without compromising the accuracy of our taxonomic assignments.

The combined preprocessing and curation steps ensured that our dataset was well-prepared for subsequent analyses, and we are confident in the reliability of our findings. A taxonomic identity was attributed to each (OTUs) by comparison with the SILVA database using an 80% homogeneity cutoff and a threshold of 0.50. The most abundant sequence for each OTU was compared with the SILVA dataset 1.32 version using the BLASTN algorithm to infer species assignment (1% mismatch threshold for specific labeling). Briefly, the species name if known, or the corresponding NCBI accession number was used. Otherwise, for non-identical OTUs, the population was labeled with its corresponding OTU number. In addition to the taxonomic profiling, the OTU representative sequences were used to recover all source information and host-related information from our source-tracking database (based on sequence data).

2.3.2. Source Tracker Analysis

A local 16S rDNA sequence database was built in our laboratory associating sequences to their curated metadata and biotope information.

- Data collection stored in our Database

Starting with the 16S rDNA v1.32 set from SILVA database, we removed eukaryotic and vector entries. The corresponding GenBank records of the remaining sequences, containing metadata and study titles, were recovered and curated to keep host and environmental habitat information. In this database, we provide a set of 5 million published 16S rDNA sequences for which taxonomic identity was validated and subsequently labeled with publication and biotope information (host and habitat). Briefly, raw metadata recovered from NCBI entries were reviewed to encode animal and plant hosts with eukaryotic taxonomic affiliation and to keep biotope information. The second layer of global key terms (animal, plant, human, soil, water . . .) was added. A catalog of annotated metadata terms using a controlled vocabulary was created. This catalog is available at https://github.com/HibaJabri-project/Host_meta_db/blob/master/Host_Dico.obo.zip (accessed on 9 January 2020).

- Database design

All datasets were organized using an entity-relationship model [14] using the software package MySQL WorkBench Version 1 (available at https://github.com/HibaJabri-project/Host_meta_db/blob/master/HOST_META_model_database.mwb and accessed on 9 January 2020). All tables are appropriately indexed.

- Analysis of restroom data

Sequences are grouped by data partitioning (clustering) according to their similarity and OTUs are defined from a similarity threshold chosen (usually 97%). Using the corresponding accession number, we can deduce the origin source. Bacterial source identification was performed on one side by sequence tracking (ST) (using accession number to find sources) and other side using species name tracking (SNT) as shown in Figure 1. From the final OTU table, populations having known accessions associated with species assignment (keeping the 99% homology threshold) were selected. The corresponding accession IDs were used in the ST approach to recover source tracking keywords associated with these IDs in our source tracking database, returning a reference table containing accession and keywords list pairs. This process was conducted using MySQL query lite version (sqlite3) (query commands available at: https://github.com/HibaJabri-project/Host_meta_db/blob/master/sqlite3_command_restroom.txt and accessed on 9 January 2020).

If the corresponding accessions IDs were not present in our source tracking database, the assigned source labeling was "unknown_source". Source labeling for OTUs without strict homologous sequences in SILVA 1.32 database (pairwise nucleotide identity below the 99% threshold) was "Not_identical_OTU".

OTUs or clusters with sequence similarity at the molecular level are used to deduce the source origin. Here, we faced two types of results; one with an accession number corresponding to several sources, and a second accession number corresponding to only one source. In the first case, we valued by giving each source the total account number, fractionating the sum on the total, and then calculating the percentage. However, if the accession number is corresponding to only one source, then we deduce the sum and the percentage of that source. Data are available at this link: https://github.com/HibaJabri-project/Host_meta_db (accessed on 9 January 2020).

The SNT process is based upon a literature search targeting the known biotopes for the list of defined species names obtained from the species assignment protocol during the amplicon profiling analysis. In order to compare both methods (ST vs. SNT), super keywords representing global types of keywords were created for each approach: Animal, Environment, Human, Ubiquitous, Others, and Unknown_source. Ubiquitous was used for the bacterial population whose associated keyword list belongs to several global super keywords. "Animal" was the name used to cover all bacteria associated only with animal biotopes. "Environment" was the name used to cover all bacteria associated with sources like soil, aquatic, and air origin. "Human" was the name used to cover all bacteria associated with human beings. For bacterial populations without any keywords, the super keyword

"Unknown _source" was used. Principal Coordinate Analysis (PCoA) was conducted to compare source tracking results for both types of search strategies.

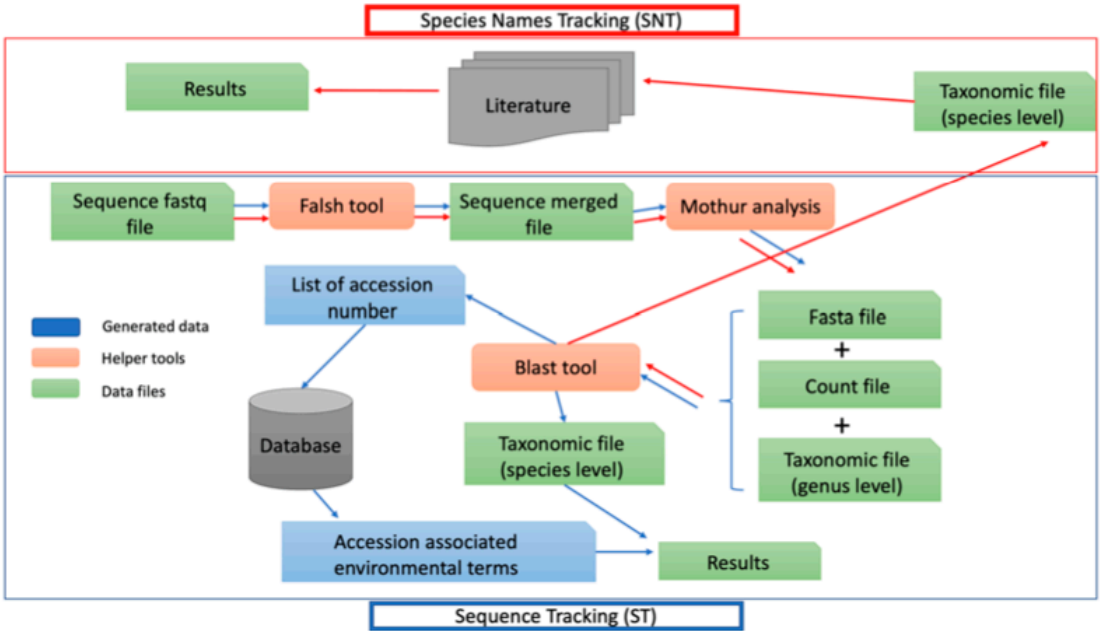


Figure 1. Workflow to analyze restroom data with ST and SNT approach. Microbial profiling analysis was investigated using metagenetic analysis and two kinds of source tracking analysis, one based on the species names (SNT) and the other one based on the frequentist sequence of each OTU to enrich microbial profiling with probable sources of bacterial contamination. Data files in green, helper tools are in orange and data used in the ST pipeline is in blue.

2.4. Statistical and Ecological Analysis

For optimal comparison across all samples, the OTU table was rarefied to 10,000 reads per sample used to evaluate ecological indicators (the richness, microbial diversity and Chao1 richness estimator of the samples). Population structure indices like richness estimation (Chao1 richness estimator) [15], microbial biodiversity (Simpson inverse biodiversity index) [16], and population evenness (Simpson evenness index) [17] were calculated using MOTHUR.

The β-diversity was visualized with the Bray–Curtis dissimilarity-based non-metric multidimensional scaling (NMDS) [18] using the *vegan*, *vegan3d*, and *rgl* packages in R [19]. Significant differences between time points were calculated with MOTHUR v1.39 using AMOVA and HOMOVA tests. The AMOVA test is a non-parametric analysis for testing the hypothesis that genetic diversity within each time point is not significantly different from the genetic diversity in all samples together [20]. The HOMOVA nonparametric test analysis was used to test the hypothesis that the genetic diversity within two or more populations is homogeneous [21].

Differences in bacterial relative abundances between gender users were assessed in STAMP tools using a mixed linear model with Benjamini Hotchberg FDR correction for multiple comparisons [22].

3. Results

3.1. Microbial Profiling

3.1.1. General Characteristics of Microbial Communities

Overall, our study began with 81 DNA samples, which contained a total of 14,915,675 reads. Following the trimming and filtering process, low-quality reads were removed, and all remaining reads had a median length of 495 nucleotides. The reads were then clustered into 17,287 operational taxonomic units (OTUs), with a mean sample coverage of above 97%. The results of the study also revealed a high diversity of microbial biodiversity across the samples, as indicated by a mean index of 15 (Simpson inverse biodiversity index).

According to taxonomic profiling, the bacteria identified in the samples were mainly from four phyla: Actinobacteria (30%), Bacteroidetes (24%), Proteobacteria (22%), and Firmicutes (18%) (Figure 2). The women's restroom, especially the cabin seat and handle outdoor, was dominated by Firmicutes and Actinobacteria (Figure 2). On the other hand, Bacteroidetes were more abundant in men's restrooms than in women's, although the difference was not statistically significant (p-value = 0.09579).

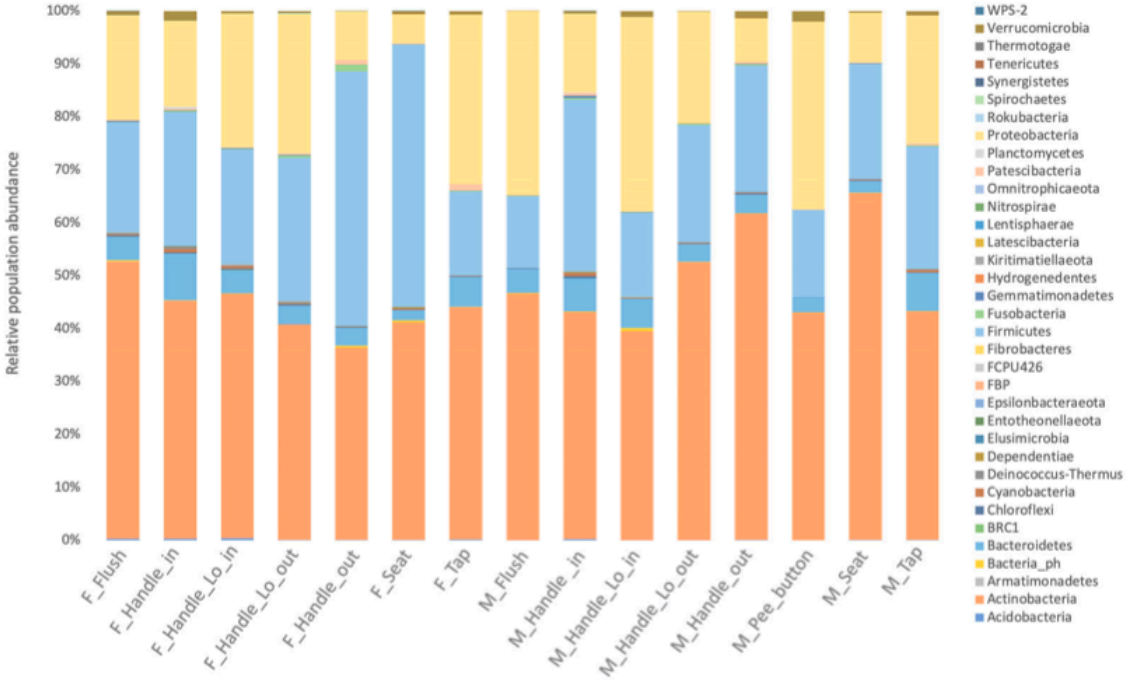


Figure 2. Relative abundances showing the 35 most abundant taxa are shown in phyla for different types of surfaces in men's user (M) and women's (F) user restrooms and between surfaces of toilets.

The six major genera identified by relative abundance were *Corynebacterium* (23%), *Staphylococcus* (10%), *Cutibacterium* (8%), *Acinetobacter* (8%), *Streptococcus* (4%), and *Lactobacillus* (3%) (Figure 3).

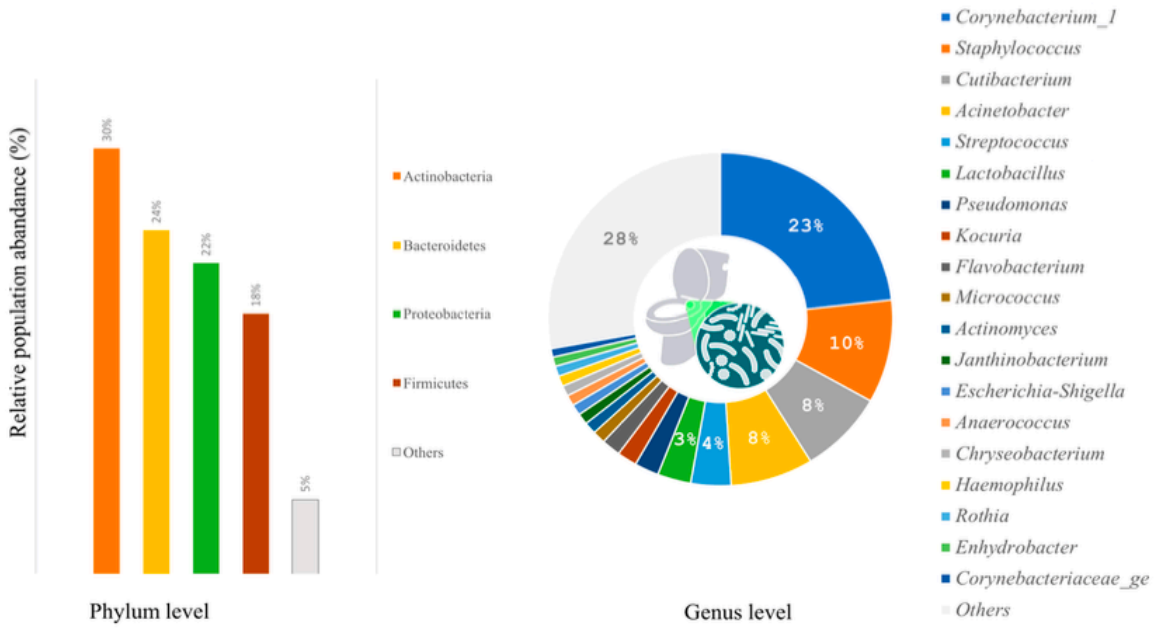


Figure 3. Relative abundances show the 4 most abundant phyla and the 20 most common bacterial genera found on surfaces of toilets.

3.1.2. Characterization of Microbial Communities on Different Surfaces of Restrooms

The relative abundance of bacterial communities was compared between two general categories: those found on seat toilet surfaces and those found on surfaces regularly touched with hands (e.g., door handles, cabin in/out, taps, and flush buttons). This difference was driven by several genera whose abundance showed statistical significance (Figure 4b). Regarding the taxa typically associated with surfaces in direct contact with human hand touch (door handles, taps), significant differences were observed in the presence of *Streptococcus* (p-value = 6.67×10^{-3}) and *Cutibacterium* (p-value = 7.61×10^{-4}) compared to urine surfaces (toilet seat) (Figure 4b). On the other hand, *Anaerococcus* bacteria were significantly associated with surfaces in direct contact with water and urine (p-value = 2.43×10^{-4}), particularly toilet seat and urinal flush surfaces (Figure 4b). The cleaning process did not show a direct effect on the presence or absence of bacterial communities but resulted in different microbial population abundances. The microbiota structure of samples before and after cleaning was not globally different (p-value > 0.05) according to the AMOVA test, suggesting no segregation of samples according to the cleaning criterion. The Dirty group showed a high presence of *Corynebacterium* on all surfaces, while *Acinetobacter* was more widespread in the Clean group.

Finally, classical indicators of fecal contamination were observed, such as *Streptococcus* and *Enterococcus* spp. on restroom surfaces. Pathogenic microorganisms were present at a low level, including *Staphylococcus* (10%), *Streptococcus* (3.2%), *Enterococcus* (0.6%), and *Campylobacter* (0.2%) (Figure 4a).

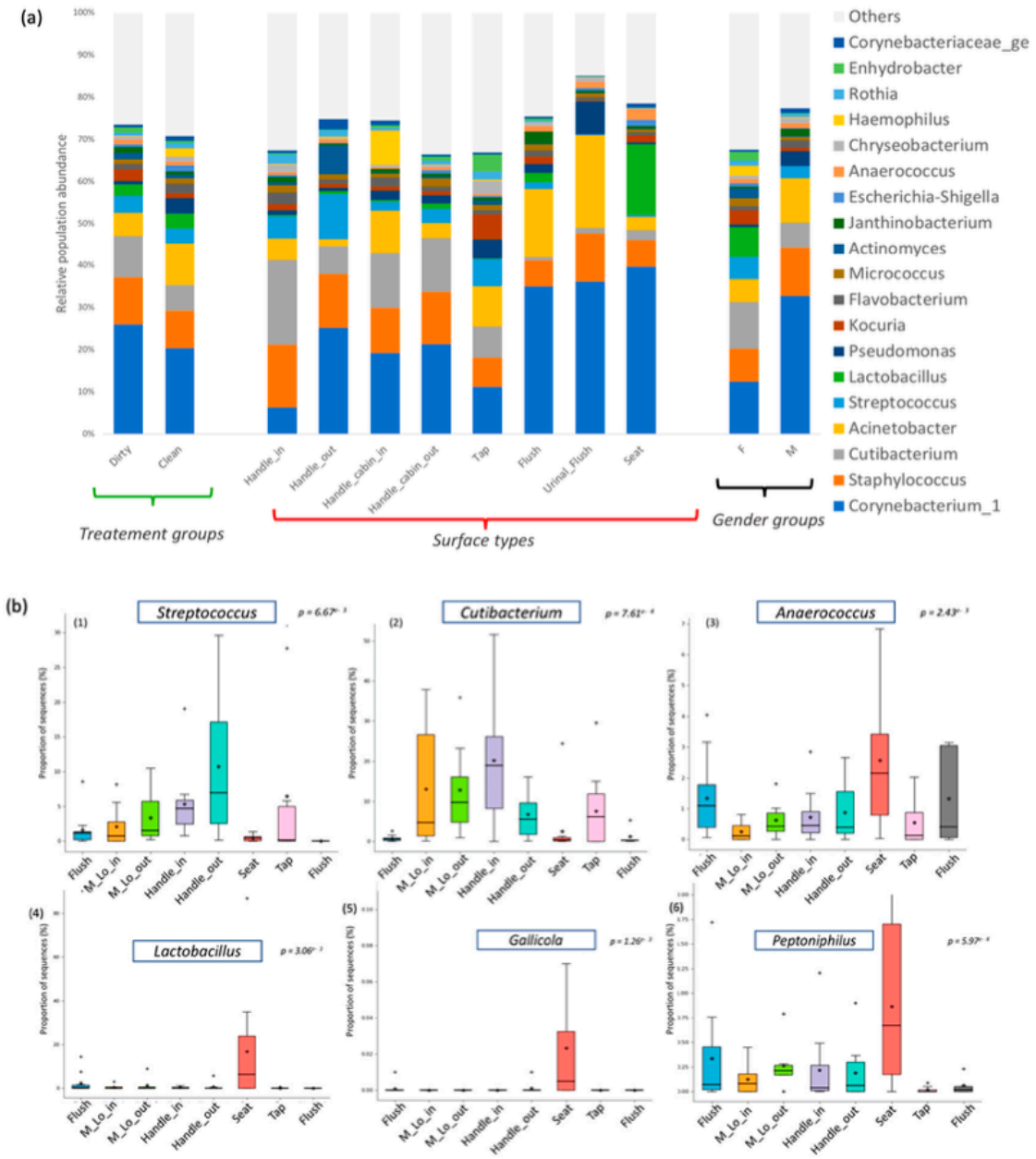


Figure 4. Bacterial diversity with the genus distribution levels expressed as mean cumulative relative abundance in different groups (**a**) Taxonomic composition of bacterial communities (before cleaning process (dirty) or after cleaning process (clean), different sampled locations and different gender users of veterinary faculty restrooms (**b**) Predominance bacterial genera in different restroom surfaces using ANOVA tests. These are default box plot representation of the interquartile range of the relative abundance of the target population in the different groups. Median is indicated as a line inside the box and mean is labelled with a star (*). Outliers values are also indicated (+).

3.1.3. Microbial Diversity between Men's and Women's Users

In the restrooms, there is a noticeable difference in microbial populations between men and women. Women's restrooms have a higher abundance of *Lactobacillus*, *Gallicola*, and *Peptoniphilus*, particularly in the cabin seat area. Non-metric dimensional scaling (NMDS) analysis of microbial populations using a Bray–Curtis dissimilarity matrix shows that the samples cluster distinctly by gender. The analysis of molecular variance (AMOVA test) between men and women users confirms the sample clustering with a p-value < 0.001. Further, the investigation of Lactobacillus at the species level revealed that *Lactobacillus crispatus* is dominant in the cabin seat area of women's restrooms. Moreover, OTUs assigned to *Kocuria rhizophila* were found on surfaces related to women's restroom samples with a p-value < 0.005 (Figure 5).

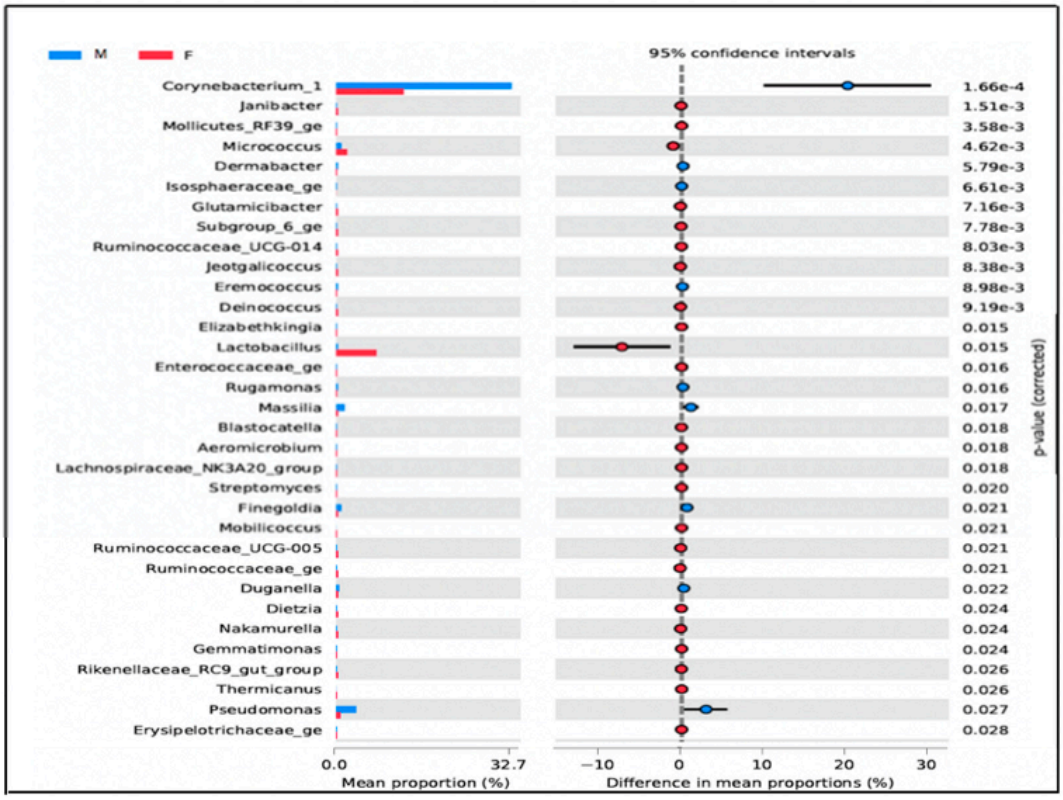


Figure 5. Bacterial Abundance of genera in restroom surfaces for men's (M) and women's (F). This figure shows only OTUs with significant differences in abundance between men's and women's in veterinary restrooms.

Staphylococcus was present on almost all surfaces except for the tap surface of women's toilets, but this difference was not statistically significant (p-value > 0.05) (Figure 6a). Bacterial richness was significantly higher in men's restrooms than in women's restrooms (Figure 6b), but there were no significant differences in α-diversity and evenness between the two groups (Figure 6c; p-value = 0.87). However, β-diversity differed significantly between the groups (p-value = 0.002; Figure 6d).

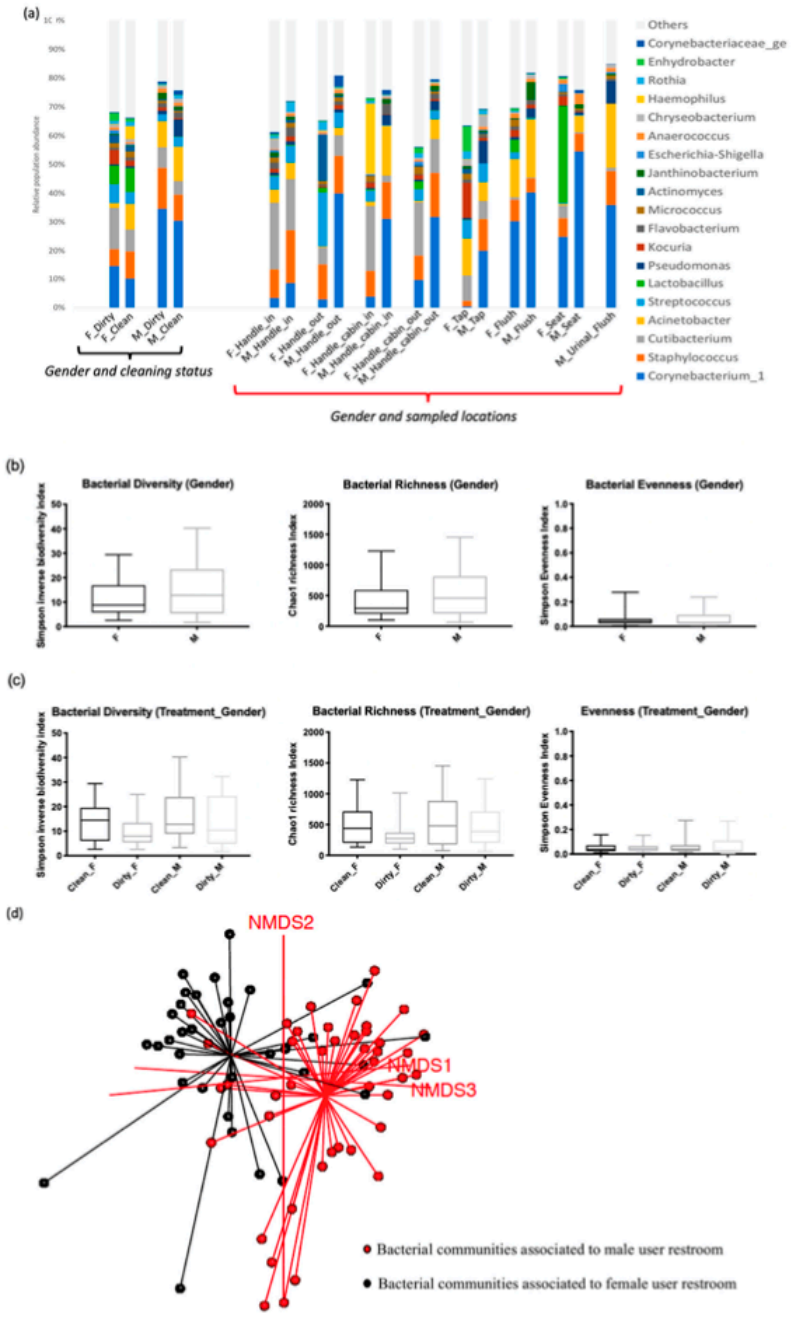


Figure 6. Bacterial diversity between men's and women's users. (**a**) Taxonomic composition of bacterial communities using the interaction of two factors (gender user influence before or after cleaning (Gender_treatment)), gender user influence and sampled locations. (**b**) Bacteria diversity (inverse Simpson Biodiversity Index), bacteria richness (Chaol Richness Index) and bacteria evenness (Simpson Evenness Index) both gender users of restrooms (**c**) For both gender users before and after cleaning of restroom surfaces. (**d**) Spatial ordination, of β-diversity between samples deduced by 16S rDNA profiling. Non-metric dimensional scaling (NMDS, k = 3, stress = 0.1) showing standard deviation, men's user restrooms (M) in red, women's (F) user ones in black.

3.1.4. Bacteria Associated with Animals

To determine the relationship between bacteria and biotopes, a Genbank search was conducted for every known species identified in the microbial profiling using keywords like "Pathology" or "Disease", "Bacteria", and "Animal". Zoonotic bacteria were detected in the bacterial profiling and found on tap-associated surfaces. For example, *Yersinia enterocolitica* species was present on tap surfaces with a very low relative abundance (0.02%) in only one positive sample out of 81 DNA samples. *Haemophilus influenzae* was mostly detected on women's restroom surfaces, while surfaces such as door handles had a low abundance of zoonotic bacteria like *Erysipelothrix* sp. (0.97%) and *Streptococcus canis* (0.1%). In "Table 2", several fecal indicator bacteria (FIB) were found to be associated with animal host sources on toilet surfaces based on bibliographic searches. Out of all *Streptococcus* species detected in our samples, only *S. equinus*, a fecal contamination indicator of animal origin, was found on female restroom surfaces.

Table 2. Fecal indicator bacteria (FIB) were found in veterinary faculty restrooms and main sources using SNT.

Family	Species	Host-Sources in Bibliography Research (SNT)
Bacteroidaceae	*Bacteroides dore*	ANIMAL and HUMAN
	Bacteroides pyogenes	ANIMAL
Bifidobacteriaceae	*Bifidobacterium merycicum*	ANIMAL
	Bifidobacterium pseudolongum	ANIMAL
Clostridiaceae	*Clostridium algidicarnis*	ANIMAL
	Clostridium frigidicarnis	ANIMAL
	Clostridium novyi	ANIMAL
	Clostridium ruminantium	ANIMAL
	Clostridium septicum	ANIMAL
Enterococcaceae	*Enterococcus cecorum*	ANIMAL
Enterobacteriaceae	*Escherichia coli*	ANIMAL and HUMAN
Erysipelotrichaceae	*Faecalicoccus pleomorphus*	ANIMAL
Streptococcaceae	*Streptococcus equinus*	ANIMAL

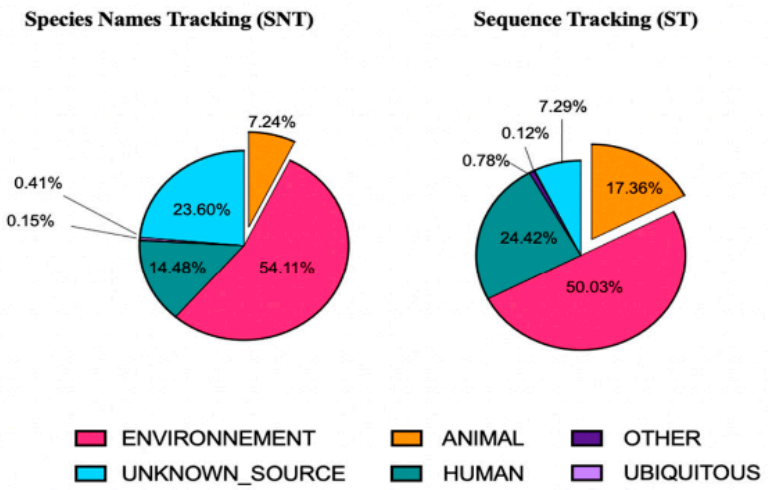
3.2. Source Tracker Analysis

3.2.1. General Characteristics Sources of the Microbial Community in Restrooms

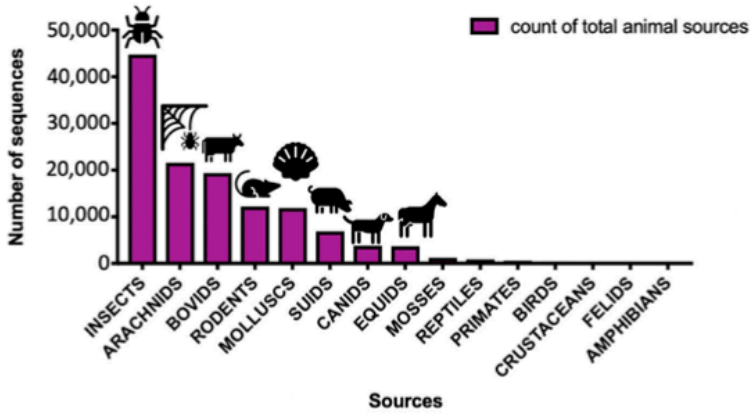
The aim of our investigation was to determine the sources of bacteria in restroom samples. We found that a diverse range of bacterial communities could originate from both human and aquatic sources in restrooms. Using the STN method, we identified the host sources for 76% of the bacterial species found in our samples. The remaining 24% were unknown sources. Our analysis revealed that the bacterial taxa were primarily from environmental sources (54.11%), followed by human (14.48%) and animal (7.24%) sources. Based on the ST method, the six most abundant sources of bacteria in our samples were environment (50.03%), human beings (24.42%), animals (17.36%), other sources (0.78%), ubiquitous sources (0.12%), and 7.29% of unknown sources (Figure 7a).

3.2.2. Bacterial Sources Associated with Animal Hosts

The analysis of animal sources found in restroom surfaces showed that insects were the primary animal sources, followed by bovids, arachnids, mollusks, rodents, suids, and canids, as shown in Figure 7b. Furthermore, the investigation of zoonotic bacteria associated with different animal sources revealed a direct association between Suidae and Equidae keywords and certain bacterial genera, such as *Chryseobacterium, Bergeyella, Fibrobacter*, and *Syntrophococcus*, which were linked to a relatively diverse range of animal hosts, including bovids, insects, equids, and suids.



(**a**)



(**b**)

Figure 7. The relative abundance of different taxa in veterinary faculty restrooms with ST and SNT approaches, respectively (**a**) Results of source tracking analysis showing the average distribution of bacterial communities in different surface veterinary restrooms (ST and SNT) (**b**) Details of animal sources with ST approach.

Interestingly, a comparison between the distribution of sources based on the species level name (nomenclature) and sequence id (accession numbers) showed a considerable difference in the number of source categories using the PCoA analysis (Figure 8). The Principal Coordinate Analysis (PCoA) was utilized to visualize the difference between the number of OTUs obtained using the two types of analyses we conducted, with one providing 556 taxonomic names (Nomenclature sources) and the other providing 3484 accession numbers (Identity sequences sources) in the probable sources in the restrooms. The first component primarily separates human and environmental sequences, while the second component helps identify clusters of animal sources.

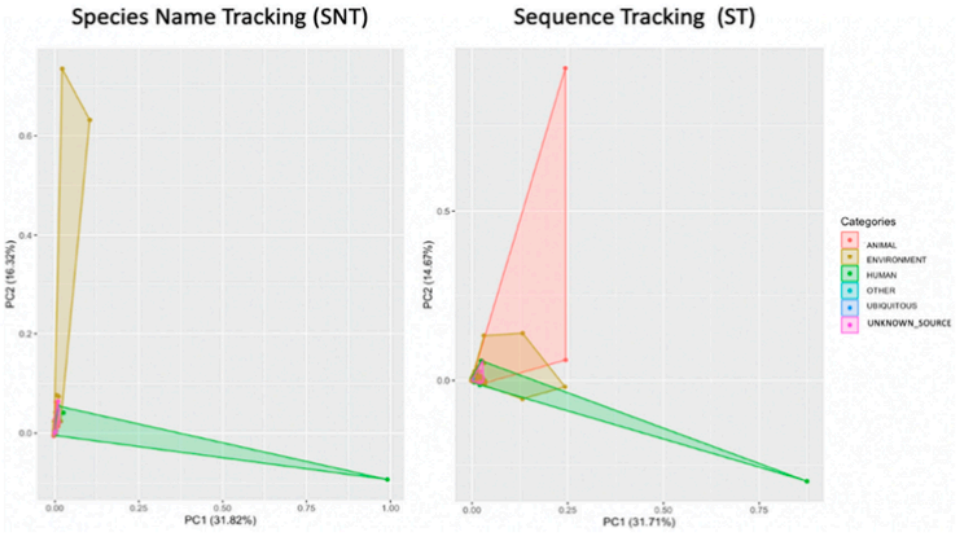


Figure 8. Principle coordinate analysis (PC1 vs. PC2) using PCoA plot (principal components analysis for each species colored by categories ecosystems (animal, environment, human, others, ubiquitous, and unknown sources) with ST and SNT approaches, respectively.

4. Discussion

In restroom environments, the microbiota is closely linked to the human microbiome, with microbial profiles shaped by various factors such as feeding patterns, hand hygiene, and skin microflora [23–25]. Previous studies have explored the concept of defining a bacterial biotope [26], particularly in restrooms [24,27,28]. However, in this study, we employed a new approach and different tools as proof of concept to investigate the direct and indirect contribution of external sources, including animal sources, to the microbial profile of restroom surfaces.

To begin, we conducted a metagenetic analysis to describe the microbial community and identify differences in microbiota between men's and women's restrooms, searched for fecal biomarker bacteria, and studied the impact of hygiene cleaning. Our results showed that Actinobacteria were the most abundant phyla in all samples, followed by Bacteroidetes, Proteobacteria, and Firmicutes, which is consistent with other public restroom studies [27,28]. Notably, it is important to mention that samples were collected in March 2017, March 2018, and April 2018 (primarily during the spring, as the first sample was collected in that season). Despite the temporal variations, the relative abundance of these phyla remained consistent across the different sampling periods.

While we found small amounts of Cyanobacteria in our study, previous research reported higher levels, likely due to "Chloroplast" plant material tracked in from outside [28]. However, our study did detect Melainabacteria, a class of Cyanobacteria associated with mammalian gut environments, indicating fecal contamination in some samples [29]. Additionally, we found *Corynebacterium*, a genus ubiquitous in the environment and closely associated with human skin [30], in many of our samples. Using the source-tracking (ST) approach, we identified a variety of sources for *Corynebacterium*, including fish, insects, canids, suids, bovids, farm animals, food staff, and aquatic environments.

Our investigation has revealed the presence of three classes of Cyanobacteria. Melainabacteria, were detected in association with fecal contamination, while Oxyphotobacteria and Sericytochromatia are linked to water environments and were previously identified by Concha et al. [31]. Of the total samples, twenty exhibited a high proportion of *Corynebacterium*, a bacterium widely distributed in various environments [31] and commonly found

on human skin [23] based on the "SNT approach." Flores et al. [27] used Bayesian classifier SourceTracker model tools and Qiime metagenomic tools [7] to demonstrate that the Corynebacterium genus is mostly associated with human skin. Our developed ST method identified multiple sources of the *Corynebacterium* genus group, indicating its ubiquitous nature and potential origins from various sources such as fish gut, insect swabs, canids, suids, bovids, farm animals, food staff, and aquatic environments. The ubiquitous presence of this genus has been well-documented in a prior publication [30], as mentioned in this paragraph.

For the assessment of potential environmental sources of microbiota in the restroom, we decided to focus primarily on animal-origin sources. As a result, we did not include these environmental sources in the final results. However, during the course of our analysis, we did observe the presence of other environmental sources, such as air, soil, and aquatic samples, showing varying proportions across different samples. Although these environmental sources were not the main focus of our investigation, their presence highlights the complexity of the restroom microbiota and suggests their potential contributions to the overall microbial composition. Further research and analysis specifically targeting environmental sources could provide valuable insights into the broader microbial dynamics within the restroom environment.

Overall, our results showed a high diversity of bacterial communities in restrooms, with aquatic and human sources being the primary contributors. We also distinguished between direct and indirect animal sources and found that insect and arachnid sources were present in restrooms as direct contributors, while larger animals were indirect contributors. Our findings are illustrated in Figure 8b. Through our "SNT approach," we discovered a wide range of bacterial communities in public restrooms, mostly originating from aquatic and human sources such as skin, intestine, and urine, which are commonly found in the restroom environment. However, some keywords overlap, such as bacterial populations associated with fish and the aquatic environment. To address this, we employed the "ST approach" and associated fish source OTUs with aquatic habitats.

The importance of hygiene cleaning in protecting human health from microbial transmission and diseases cannot be overstated. Despite efforts to clean restrooms, studies have shown that they still harbor thousands of types of bacteria and viruses, including common contaminants like fecal bacteria, Influenza, *Streptococcus*, *E. coli*, hepatitis viruses, MRSA, *Salmonella*, *Shigella*, and norovirus [32]. Due to the high number of germs and variables present in restrooms, it may be difficult to remove all contaminants with routine cleaning [33]. Our study found no significant change in microbial profiles before and after cleaning veterinary faculty restrooms (p-value > 0.05).

In this study, no significant difference was observed in the microbiota composition before and after cleaning. However, it is crucial to acknowledge that the analysis did not consider the level of contamination, which may have likely decreased over time. It is also important to note that the analysis focused on genetic traces rather than distinguishing between living and dead bacteria. Prior to cleaning, a higher prevalence of Actinobacteria was observed, while Acinetobacter, a microbe typically found in water and aquatic environments [34], increased after the cleaning process. The increase in relative average abundance of Cyanobacteria after cleaning could be due to the removal of other bacteria [32]. The high abundance of Staphylococcus and *Cutibacterium* found in the restrooms, which are typically found on human skin and fecal matter [35,36], may be attributed to the use of tap water and human hands during surface cleaning. It is important to note, however, that not all environmental surfaces in restrooms are cleaned appropriately in some cases [37], and this could also be a contributing factor to the observed bacterial abundance.

The distinct bacterial signatures observed in our findings (Figure 4a) are consistent with those typically associated with healthy urogenital tracts [29,38,39]. The observed differences in bacterial composition between men's and women's restrooms may be attributed to sex-related differences in gut microbiota, as previously reported [40–47]. Our findings on bacterial composition between men's and women's restrooms are consistent with previous

studies [40–42] showing that the Bacteroides genus is more present in the male gut. In contrast, we found that *Bifidobacterium*, *Lactobacillus*, *Veillonella*, and *Streptococcus* were more abundant in women. Moreover, our findings align with prior research [45–47] in demonstrating that *Lactobacillaceae* bacteria, which are typically found in the healthy vaginal ecosystem, were mostly present in women's restrooms. Specifically, we observed *Lactobacillaceae* bacteria on seat surfaces in women's restrooms, which suggests the presence of healthy vaginal microbiota [48,49].

Women's restrooms also showed the presence of other specific bacterial markers such as *Kocuria rhizophila* (Figure 6a) and *Gallicola* and *Peptoniphilus* (Figure 4b), although in low relative abundance. These bacteria have been associated with urinary tract infections and physiological imbalances in women's bodies, as reported in previous studies [50].

In our investigation of the Staphylococcus genus, we found that *Staphylococcus* aureus is the leading cause of skin and soft tissue infections [51], despite only accounting for 0.13% of the sequences associated with the genus. Other studies have suggested that *Staphylococcus* spp. may be more prevalent on restroom seats, but our findings indicate that the inside handle door harbors more *Staphylococcus* spp. [52] than the outside handle door, likely due to the hand-washing process and the assumption that hands are dirty upon entering the restroom and clean upon leaving (Figure 4a). Our ST approach also revealed that *Staphylococcus* spp. can be found in not only human users but also in animal or aquatic environments, as well as in gut and skin-human bacteria flora, as seen in other studies [27].

Amplicon sequencing strategies provide a comprehensive view of the microbiota in our samples, without the need for culturing, which can be less efficient in terms of bacterial discovery. In this study, we aimed to identify possible sources of bacteria in the restroom, making culture microbiology unnecessary.

Using the ST approach, we found that the most abundant bacteria sources in veterinary faculty restrooms samples were associated with aquatic (12.32%), human (9.04%), soil (6.87%), gymnosperms (4.84%), fish (4.40%), and food sources (2.60%). Our results differed from those of Flores et al. [27], whose SourceTracker module in Qiime tools [7] relied on a statistical model to estimate source proportions by downloading 10 samples for each suspected source of each group (human, aquatic and soil, and others) and using 16S rDNA bacterial communities. Our local database, based on open public sequence information, provided us with more keyword information about bacterial sources, especially related to organisms, increasing our knowledge from 7.24% to 17.36% of general information associated with animals and reducing the unknown source information compared to the SNT approach based only on literature (Figure 7a). Furthermore, providing information about the source origin was useful for comparing our results with the literature [27,28], in addition to the bacterial profiling analysis.

The interactive coloring of the PCoA plot (Figure 8) reveals a divergence in the resource distribution between the two methods of analysis: one based on taxonomic names (SNT) and the other based on sequence identity (ST). The ST approach, in conjunction with our local database, enabled the identification of a vast array of sources, particularly those linked to animal surfaces in restrooms, which were previously unidentified by the SNT method (Figure 8). Our visualization accurately illustrates the distribution of bacterial sources in restrooms based on sequence identity, highlighting the genuine distribution of bacteria sources in these environments. Notwithstanding, systematic errors associated with metadata and sequencing technologies remain a potential concern. As such, our final results table displays the number of supporting sequences for each cluster enrichment, and our interactive visualization provides the ability to inspect PCoA clusters categorized by source groups (animal, environment, human, ubiquitous, other, and unknown sources). Omitting taxonomy can be advantageous because unknown species of ubiquitous genera are uninformative for source tracking.

Comparing bacterial sources in this study using our local database (ST) with those described in the literature (SNT) facilitated the identification of additional sources associated with ubiquitous species. Furthermore, some zoonotic bacteria species, such as *Yersinia*

enterocolitica, *Erysipelothrix*, and *Streptococcus equinus*, were directly linked to swine and horse host animals [53], respectively, based on their presence on restroom surfaces, as previously described in Sullivan 2011 and Sherman 1936 [53,54]. Additionally, known fecal contaminant bacteria found on cabin handles and flush surfaces, such as *Enterococcus cecorum*, were directly linked to the fecal environment, as reported in other studies [55]. Although non-pathogenic bacteria, such as *Erysipelotrichaceae*, are associated with bovid animal hosts through the ST approach and are biomarker bacteria of the animal rumen [56].

The SNT method is efficient for well-known and extensively studied bacterial taxa. However, the ST approach can attribute sources to poorly described or unidentified taxa and can be used in tandem with the SNT method to provide further information on the origin of bacteria. This approach demonstrates a "proof in principle" and validates the significance of sequence-based data in microbial source tracking, with a high probability of yielding correct sources. Nevertheless, some well-known bacteria, such as *Streptococcus canis* [57], were associated with unknown sources due to the absence of source information for sequences deposited in public databases (missing metadata), resulting in the loss of valuable information. To address this limitation, combining the SNT and ST approaches as complementary methods or creating a database with more sequences associated with lesser-known environments and specific biotopes would be advantageous.

Most of the sources in this study were expectedly linked to aquatic, human, and soil environments, as typical of restrooms. However, using a local database and incorporating more information on animal sources provided additional knowledge. We believe that this database could serve as a valuable resource in microbiota profiling, helping the scientific community identify unknown bacteria with regard to their ubiquity or potential biomarker value in key ecosystems. There is a growing interest in the potential use of molecular fingerprinting methods (DNA) not only for detecting but also for identifying contamination sources in various industrial and scientific fields. The insights garnered from this study on bacterial biotopes within veterinary restrooms hold significant potential for detecting the sources of contamination in various research domains, particularly in health sciences and veterinary settings.

5. Conclusions

In conclusion, our ST approach proved to be a useful complement to the STN approach, especially when dealing with poorly characterized microbial taxa such as those found in restrooms. By utilizing a local database, we successfully identified discernible differences in the microbiota associated with direct (human microbiome) and indirect (animal) contributions in veterinary restrooms. However, it is important to acknowledge the limitations of our study, including the relatively low sample size and the absence of viability assessment and rigorous sample treatment process (cleaning process). As such, these results should be regarded as an initial exploration, providing a foundation for future research.

To advance our understanding and overcome the limitations, we propose expanding the scope of our study by incorporating comparisons from a wider range of restrooms. In the specific paragraph mentioned, we intended to emphasize that my focus extended beyond the database itself, encompassing factors like sample preparation control, such as assessing the influence of environmental variations before and after cleaning (e.g., temperature or relative humidity). These aspects served as critical starting points for this study.

Furthermore, conducting larger sample sizes with more stringent control measures will yield more robust and comprehensive data, essential for deeper insights into microbial dynamics. These improvements will enable us to establish a solid groundwork for further research in this area. Therefore, the results presented in our current study should be recognized as an initial step, urging future endeavors to address these crucial aspects and enhance the understanding of the subject matter. Notably, improvements to our local database, such as cross-linking with other databases, could help address the issue of missing sequence information.

The insights gained in this study on bacterial biotopes in veterinary restrooms could be beneficial in various research areas, including health sciences and veterinary environments. Comparing the results from veterinary restrooms to those from human hospital environments, public transport stations, or school public toilets could provide a better understanding of the origins of microbial contamination. Source indication labeling could also be used to investigate sources of contamination associated with animal disease, hygiene management in common places, and public health research. By providing a quick and easy way to enrich the metagenetic analysis, this tool has the potential to be widely adopted in many different fields.

Author Contributions: Conceptualization, H.J. and B.T.; methodology, H.J.; software, H.J. and B.T.; validation, D.B., B.T., and G.D.; formal analysis, H.J.; investigation, H.J.; resources, B.T. and G.D.; data curation, P.A.F., S.K., and H.J.; writing—original draft preparation, H.J.; writing—review and editing, D.B., B.T., and G.D.; visualization, H.J.; supervision, B.T.; project administration, G.D.; funding acquisition, G.D. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: All the biosample raw reads have been deposited at the National Center for Biotechnology Information (NCBI) and are available under de Bioproject ID PRJNA810326 under this URL: https://www.ncbi.nlm.nih.gov/bioproject/?term=PRJNA810326 (Registration date: 25 February 2022). Data results are available at this link: https://github.com/HibaJabri-project/Host_meta_db (accessed on 9 January 2020).

Acknowledgments: The funders had no role in study design, data collection and interpretation or the decision to submit the work for publication.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Burow, E.; Rostalski, A.; Harlizius, J.; Gangl, A.; Simoneit, C.; Grobbel, M.; Kollas, C.; Tenhagen, B.-A.; Käsbohrer, A. Antibiotic resistance in Escherichia coli from pigs from birth to slaughter and its association with antibiotic treatment. *Prev. Vet. Med.* **2019**, *165*, 52–62. [CrossRef]
2. van Duijkeren, E.; Box, A.; Heck, M.; Wannet, W.; Fluit, A. Methicillin-resistant staphylococci isolated from animals. *Vet. Microbiol.* **2004**, *103*, 91–97. [CrossRef]
3. Seguin, J.C.; Walker, R.D.; Caron, J.P.; Kloos, W.E.; George, C.G.; Hollis, R.J.; Jones, R.N.; Pfaller, M.A. Methicillin-Resistant *Staphylococcus aureus* Outbreak in a Veterinary Teaching Hospital: Potential Human-to-Animal Transmission. *J. Clin. Microbiol.* **1999**, *37*, 1459–1463. [CrossRef] [PubMed]
4. Juhász-Kaszanyitzky, É.; Jánosi, S.; Somogyi, P.; Dán, A.; Bloois, L.V.D.G.-V.; Van Duijkeren, E.; Wagenaar, J.A. MRSA Transmission between Cows and Humans. *Emerg. Infect. Dis.* **2007**, *13*, 630–632. [CrossRef]
5. Bojanova, D.P.; Bordenstein, S.R. Fecal Transplants: What Is Being Transferred? *PLoS Biol.* **2016**, *14*, e1002503. [CrossRef] [PubMed]
6. Nguyen, K.; Senay, C.; Young, S.; Nayak, B.; Lobos, A.; Conrad, J.; Harwood, V. Determination of wild animal sources of fecal indicator bacteria by microbial source tracking (MST) influences regulatory decisions. *Water Res.* **2018**, *144*, 424–434. [CrossRef] [PubMed]
7. Knights, D.; Kuczynski, J.; Charlson, E.S.; Zaneveld, J.; Mozer, M.C.; Collman, R.G.; Bushman, F.D.; Knight, R.T.; Kelley, S.T. Bayesian community-wide culture-independent microbial source tracking. *Nat. Methods* **2011**, *8*, 761–763. [CrossRef]
8. Quast, C.; Pruesse, E.; Yilmaz, P.; Gerken, J.; Schweer, T.; Yarza, P.; Peplies, J.; Glöckner, F.O. The SILVA Ribosomal RNA Gene Database Project: Improved Data Processing and Web-Based Tools. *Nucleic Acids Res.* **2013**, *41*, D590–D596. [CrossRef]
9. Christenfeld, N. Choices from Identical Options. *Psychol. Sci.* **1995**, *6*, 50–55. [CrossRef]
10. Ceugniez, A.; Taminiau, B.; Coucheney, F.; Jacques, P.; Delcenserie, V.; Daube, G.; Drider, D. Use of a metagenetic approach to monitor the bacterial microbiota of "Tomme d'Orchies" cheese during the ripening process. *Int. J. Food Microbiol.* **2017**, *247*, 65–69. [CrossRef]
11. Rognes, T.; Flouri, T.; Nichols, B.; Quince, C.; Mahé, F. VSEARCH: A versatile open source tool for metagenomics. *PeerJ* **2016**, *2016*, e2584. [CrossRef]
12. Pruesse, E.; Quast, C.; Knittel, K.; Fuchs, B.M.; Ludwig, W.; Peplies, J.; Glöckner, F.O. SILVA: A comprehensive online resource for quality checked and aligned ribosomal RNA sequence data compatible with ARB. *Nucleic Acids Res.* **2007**, *35*, 7188–7196. [CrossRef] [PubMed]

13. Rodriguez, C.; Taminiau, B.; Brévers, B.; Avesani, V.; Van Broeck, J.; Leroux, A.; Gallot, M.; Bruwier, A.; Amory, H.; Delmée, M.; et al. Faecal microbiota characterisation of horses using 16 rdna barcoded pyrosequencing, and carriage rate of clostridium difficile at hospital admission. *BMC Microbiol.* **2015**, *15*, 181. [CrossRef]
14. Elmasri, R.; Navathe, S.B. *Database Systems*, 6th ed.; Hirsch, M., Ed.; Addison-Wesley: Melbourne, FL, USA, 2017.
15. Chao, A.; Bunge, J. Estimating the Number of Species in a Stochastic Abundance Model. *Biometrics* **2002**, *58*, 531–539. [CrossRef]
16. Chao, A.; Shen, T.-J. Nonparametric estimation of Shannon's index of diversity when there are unseen species in sample. *Environ. Ecol. Stat.* **2003**, *10*, 429–443. [CrossRef]
17. Simpson, E.H. Measurement of diversity. *Nature* **1949**, *163*, 668. [CrossRef]
18. Clarke, K.; Ainsworth, M. A method of linking multivariate community structure to environmental variables. *Mar. Ecol. Prog. Ser.* **1993**, *92*, 205–219. [CrossRef]
19. Zanne, A.E.; Tank, D.C.; Cornwell, W.K.; Eastman, J.M.; Smith, S.A.; FitzJohn, R.G.; McGlinn, D.J.; O'Meara, B.C.; Moles, A.T.; Reich, P.B.; et al. Data from: Three Keys to the Radiation of Angiosperms into Freezing Environments. *Nature* **2014**, *506*, 89–92. [CrossRef] [PubMed]
20. Schloss, P.D.; Westcott, S.L.; Ryabin, T.; Hall, J.R.; Hartmann, M.; Hollister, E.B.; Lesniewski, R.A.; Oakley, B.B.; Parks, D.H.; Robinson, C.J.; et al. Introducing mothur: Open-Source, Platform-Independent, Community-Supported Software for Describing and Comparing Microbial Communities. *Appl. Environ. Microbiol.* **2009**, *75*, 7537–7541. [CrossRef]
21. Schloss, P.D. Evaluating different approaches that test whether microbial communities have the same structure. *ISME J.* **2008**, *2*, 265–275. [CrossRef]
22. Parks, D.H.; Tyson, G.W.; Hugenholtz, P.; Beiko, R.G. STAMP User's Guide v2.0.0. *Bioinformatics* **2014**, *30*, 3123–3124. [CrossRef] [PubMed]
23. Grice, E.A.; Segre, J.A. The skin microbiome. *Nat. Rev. Microbiol.* **2011**, *9*, 244–253. [CrossRef]
24. Nematian, J.; Matini, E.; Shayeghi, F.; Vaghar, M.; Hosseini, S.S.; Mojri, N.; Taherabadi, N.T.; Hakimi, R.; Ahmadi, N.; Badkoubeh, N.; et al. A survey of public restrooms microbial contamination in Tehran city, capital of Iran, during 2019. *J. Fam. Med. Prim. Care* **2020**, *9*, 3131–3135. [CrossRef] [PubMed]
25. Gizaw, Z.; Yalew, A.W.; Bitew, B.D.; Lee, J.; Bisesi, M. Effects of local handwashing agents on microbial contamination of the hands in a rural setting in Northwest Ethiopia: A cluster randomised controlled trial. *BMJ Open* **2022**, *12*, e056411. [CrossRef] [PubMed]
26. Henschel, A.; Anwar, M.Z.; Manohar, V. Comprehensive Meta-analysis of Ontology Annotated 16S rRNA Profiles Identifies Beta Diversity Clusters of Environmental Bacterial Communities. *PLoS Comput. Biol.* **2015**, *11*, e1004468. [CrossRef]
27. Flores, G.E.; Bates, S.T.; Knights, D.; Lauber, C.L.; Stombaugh, J.; Knight, R.; Fierer, N. Microbial Biogeography of Public Restroom Surfaces. *PLoS ONE* **2011**, *6*, e28132. [CrossRef]
28. Pausan, M.-R.; Blohs, M.; Mahnert, A.; Moissl-Eichinger, C. The sanitary indoor environment—A potential source for intact human-associated anaerobes. *NPJ Biofilms Microbiomes* **2022**, *8*, 44. [CrossRef]
29. Dobbler, P.C.T.; Laureano, M.; Sarzi, D.S.; Cañón, E.R.P.; Metz, G.F.; de Freitas, A.S.; Takagaki, B.M.; D'oliveira, C.B.; Pylro, V.S.; Copetti, A.C.; et al. Differences in bacterial composition between men's and women's restrooms and other common areas within a public building. *Antonie Leeuwenhoek Int. J. Gen. Mol. Microbiol.* **2018**, *111*, 551–561. [CrossRef]
30. Roy, M.; Ahmad, S. Rare case of *Corynebacterium striatum* septic arthritis. *BMJ Case Rep.* **2016**, *2016*, bcr2016216914. [CrossRef]
31. Concha, C.-D.; Maestre, F.T.; Eldridge, D.J.; Singh, B.K.; Bardgett, R.D.; Fierer, N.; Delgado-Baquerizo, M. Ecological Niche Differentiation in Soil Cyanobacterial Communities across the Globe. *BioRxiv* **2019**, *15*, 13–23. [CrossRef]
32. Gibbons, S.M.; Schwartz, T.; Fouquier, J.; Mitchell, M.; Sangwan, N.; Gilbert, J.A.; Kelley, S.T. Ecological Succession and Viability of Human-Associated Microbiota on Restroom Surfaces. *Appl. Environ. Microbiol.* **2015**, *81*, 765–773. [CrossRef]
33. Boone, S.A.; Gerba, C.P. The Prevalence of Human Parainfluenza Virus 1 on Indoor Office Fomites. *Food Environ. Virol.* **2010**, *2*, 41–46. [CrossRef]
34. Bifulco, J.M.; Shirey, J.J.; Bissonnette, G.K. Detection of *Acinetobacter* spp. in rural drinking water supplies. *Appl. Environ. Microbiol.* **1989**, *55*, 2214–2219. [CrossRef] [PubMed]
35. Hitzfeld, B.C.; Höger, S.J.; Dietrich, D.R. Cyanobacterial toxins: Removal during drinking water treatment, and human risk assessment. *Environ. Healrh Perspect. J.* **2000**, *108*, 113–122. [CrossRef]
36. Roberts, M. Characterization and Isolation of Fecal Indicator Bacteria, Staphylococcus aureus, and Methicillin-resistant Staphylococcus aureus from Pacific Northwest Marine Beach Samples. *Environ. Health* **2012**, *78*, 50–56. [CrossRef]
37. Drees, M.; Snydman, D.R.; Schmid, C.H.; Barefoot, L.; Hansjosten, K.; Vue, P.M.; Cronin, M.; Nasraway, S.A.; Golan, Y. Prior Environmental Contamination Increases the Risk of Acquisition of Vancomycin-Resistant Enterococci. *Clin. Infect. Dis.* **2008**, *46*, 678–685. [CrossRef] [PubMed]
38. O'neill, A.M.; Gallo, R.L. Host-microbiome interactions and recent progress into understanding the biology of acne vulgaris. *Microbiome* **2018**, *6*, 177. [CrossRef] [PubMed]
39. Fouts, D.E.; Pieper, R.; Szpakowski, S.; Pohl, H.; Knoblach, S.; Suh, M.-J.; Huang, S.-T.; Ljungberg, I.; Sprague, B.M.; Lucas, S.K.; et al. Integrated next-generation sequencing of 16S rDNA and metaproteomics differentiate the healthy urine microbiome from asymptomatic bacteriuria in neuropathic bladder associated with spinal cord injury. *J. Transl. Med.* **2012**, *10*, 174. [CrossRef]
40. Mueller, S.; Saunier, K.; Hanisch, C.; Norin, E.; Alm, L.; Midtvedt, T.; Cresci, A.; Silvi, S.; Orpianesi, C.; Verdenelli, M.C.; et al. Differences in Fecal Microbiota in Different European Study Populations in Relation to Age, Gender, and Country: A Cross-Sectional Study. *Appl. Environ. Microbiol.* **2006**, *72*, 1027–1033. [CrossRef]

41. Li, M.; Wang, B.; Zhang, M.; Rantalainen, M.; Wang, S.; Zhou, H.; Zhang, Y.; Shen, J.; Pang, X.; Zhang, M.; et al. Symbiotic gut microbes modulate human metabolic phenotypes. *Proc. Natl. Acad. Sci. USA* **2008**, *105*, 2117–2122. [CrossRef]
42. Dominianni, C.; Sinha, R.; Goedert, J.J.; Pei, Z.; Yang, L.; Hayes, R.B.; Ahn, J. Sex, Body Mass Index, and Dietary Fiber Intake Influence the Human Gut Microbiome. *PLoS ONE* **2015**, *10*, e0124599. [CrossRef] [PubMed]
43. Singh, P.; Manning, S.D. Impact of age and sex on the composition and abundance of the intestinal microbiota in individuals with and without enteric infections. *Ann. Epidemiol.* **2016**, *26*, 380–385. [CrossRef] [PubMed]
44. Haro, C.; Rangel-Zúñiga, O.A.; Alcalá-Díaz, J.F.; Gómez-Delgado, F.; Pérez-Martínez, P.; Delgado-Lista, J.; Quintana-Navarro, G.M.; Landa, B.B.; Navas-Cortés, J.A.; Tena-Sempere, M.; et al. Intestinal Microbiota Is Influenced by Gender and Body Mass Index. *PLoS ONE* **2016**, *11*, e0154090. [CrossRef] [PubMed]
45. Borgo, F.; Garbossa, S.; Riva, A.; Severgnini, M.; Luigiano, C.; Benetti, A.; Pontiroli, A.E.; Morace, G.; Borghi, E. Body Mass Index and Sex Affect Diverse Microbial Niches within the Gut. *Front. Microbiol.* **2018**, *9*, 213. [CrossRef]
46. Gao, X.; Zhang, M.; Xue, J.; Huang, J.; Zhuang, R.; Zhou, X.; Zhang, H.; Fu, Q.; Hao, Y. Body Mass Index Differences in the Gut Microbiota Are Gender Specific. *Front. Microbiol.* **2018**, *9*, 1250. [CrossRef] [PubMed]
47. Takagi, T.; Naito, Y.; Inoue, R.; Kashiwagi, S.; Uchiyama, K.; Mizushima, K.; Tsuchiya, S.; Dohi, O.; Yoshida, N.; Kamada, K.; et al. Differences in gut microbiota associated with age, sex, and stool consistency in healthy Japanese subjects. *J. Gastroenterol.* **2019**, *54*, 53–63. [CrossRef]
48. Lepargneur, J.-P. Lactobacillus crispatus as biomarker of the healthy vaginal tract. *Ann. Biol. Clin.* **2016**, *74*, 421–427. [CrossRef]
49. You, Y.; Kwon, E.J.; Choi, S.; Hwang, H.; Choi, S.; Lee, S.M.; Kim, Y.J. Vaginal microbiome profiles of pregnant women in Korea using a 16S metagenomics approach. *Am. J. Reprod. Immunol.* **2019**, *82*, e13124. [CrossRef]
50. Josephs-Spaulding, J.; Krogh, T.J.; Rettig, H.C.; Lyng, M.; Chkonia, M.; Waschina, S.; Graspeuntner, S.; Rupp, J.; Møller-Jensen, J.; Kaleta, C. Recurrent Urinary Tract Infections: Unraveling the Complicated Environment of Uncomplicated rUTIs. *Front. Cell. Infect. Microbiol.* **2021**, *11*, 562525. [CrossRef]
51. Claassen-Weitz, S.; Shittu, A.O.; Ngwarai, M.R.; Thabane, L.; Nicol, M.P.; Kaba, M. Fecal Carriage of Staphylococcus aureus in the Hospital and Community Setting: A Systematic Review. *Front. Microbiol.* **2016**, *7*, 449. [CrossRef]
52. Ogba, O.M.; Obio, O.M. Microbial Spectrum on Public Toilet Seats. *Ann. Microbiol. Infect. Dis.* **2018**, *1*, 58–62.
53. Sherman, H.M.H.J.M.; Eqzuinus, S. Streptococcus equinus. *J. Bacteriol.* **1936**, *1910*, 283–289. [CrossRef]
54. O'Sullivan, T.; Friendship, R.; Blackwell, T.; Pearl, D.; McEwen, B.; Carman, S.; Slavić, D.; Dewey, C. Microbiological identification and analysis of swine tonsils collected from carcasses at slaughter. *Can. J. Vet. Res.* **2011**, *75*, 106–111.
55. Lebreton, F.; Willems, R.J.L.; Gilmore, M.S. Enterococcus Diversity, Origins in Nature, and Gut Colonization. In *Enterococci: From Commensals to Leading Causes of Drug Resistant Infection*; Gilmore, M.S., Clewell, D.B., Ike, Y., Eds.; Massachusetts Eye and Ear Infirmary: Boston, MA, USA, 2014. Available online: https://www.ncbi.nlm.nih.gov/books/NBK190427 (accessed on 2 February 2014).
56. Deusch, S.; Camarinha-Silva, A.; Conrad, J.; Beifuss, U.; Rodehutscord, M.; Seifert, J. A Structural and Functional Elucidation of the Rumen Microbiome Influenced by Various Diets and Microenvironments. *Front. Microbiol.* **2017**, *8*, 1605. [CrossRef]
57. Moriconi, M.; Acke, E.; Petrelli, D.; Preziuso, S. Multiplex PCR-based identification of Streptococcus canis, Streptococcus zooepidemicus and Streptococcus dysgalactiae subspecies from dogs. *Comp. Immunol. Microbiol. Infect. Dis.* **2017**, *50*, 48–53. [CrossRef] [PubMed]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.

1

Chapter IV : Application

Study 2: Application of the Sequencing Tracking tool (ST) to Bathing water Datasets. *Mapping Bacterial Profiles in Natural Water Bathing Sites for Identification of Contamination Sources*

Hiba Jabri, Tricot Benoit, Schrooten Didier, Georges Daube, Bernard Taminiau

Ready for submission

Mapping Bacterial Profiles in Natural Water Bathing Sites for Identification of Contamination Sources

Hiba Jabri[1], Tricot Benoit[2], Schrooten Didier[3], Georges Daube[1], Bernard Taminiau [*1]

[1]Laboratory of Food Microbiology, Fundamental and Applied Research for Animals and Health Center (FARAH), Faculty of Veterinary Medicine, University of Liège, Quartier Vallée 2, B42, Avenue de Cureghem 180, 4000, Liège, Belgium.

[2]Departement of water and environment-DEE, Avenue Prince de Liège 15, B, 5100 Namur, Belgium.

[3] Institut Scientifique de Service Public-ISSeP, Rue du Chéra 200, 4000 Liège, Belgium.

*** Correspondence:**
Bernard Taminiau – Laboratory of Food Microbiology, Fundamental and applied Research for Animals and Health center (FARAH), University of Liege, Liege, Belgium.
(Bernard.taminiau@uliege.be)

Keywords: Metagenetics, 16S rDNA amplicon sequencing, biotope information, database, Source tracking.

Abstract

In the Wallonia region of Belgium, this study focuses on three heavily frequented bathing rivers zone, prompting an investigation into microbiological monitoring and microbiome source tracking. Four distinct bathing zones were studied, with one located in River A (Chiny) in 2014, and three zones designated in River B. Subsequently, three critical sampling spots were chosen for each bathing area, with one weekly sample collected per sampling point over a span of three weeks. The objective is to determine the microbiota through amplicon sequencing and to use a host traceability tool developed in the laboratory to identify the origin of microbial populations. In 2017, a comprehensive sampling approach was adopted, involving River A (Chiny) or "A_C" with samples taken every 2 hours over 24 hours, and River A (Lacuisine) or "A_L" downstream of the previous sampling point, sampled every 4 hours for the same duration. In this subsequent campaign, our focus shifted to exploring the stability of the microbiota in a dynamic aquatic environment. Assessing whether sampling water from a river adequately captures the microbiota over a 24-hour period became a central objective. A 16S rDNA metagenomic analysis was conducted to elucidate microbial profiles within these bathing areas, assess the diversity of microbial species through beta diversity comparisons. In 2014, our analysis unveiled a prevailing presence of *Pseudarcicella* and *Limnohabitans* in rivers A_C and B, constituting the entirety of identified sequences. Using our tool, we identified that environmental origin sources accounted for 29.2% of the sequences. Host organism-associated sequences made up 1.68%, with human-related sequences specifically constituting 0.32% in River A_C. This was higher compared to River B, likely due to increased human activities such as camping. Additionally, equids, bovids, suids, birds, and other mammals contributed to the bacterial composition in varying proportions. The remaining 69.63% of sequences were categorized as unknown sources, which is typical for samples collected in natural environments.

Fast forward to 2017, our findings demonstrated a notable shift, with *Limnohabitans* and *Cyanobiaceae* emerging as more prevalent than other bacteria within the total identified sequences. These sequences were associated with environmental sources at 38.51%, unknown sources at 57.73%, and host organisms at 3.76% of the total sequences from all samples. Similar to the observations in 2014, bacteria associated with humans exhibited a higher presence in river A_C (0.36%) compared to river A_L, attributed to the frequent human activity. Furthermore, suids, bovids, equids, birds, and other mammals contributed to the overall bacterial composition, each contributing distinct proportions. These findings underscore the dynamic nature of bacterial presence in rivers over time and highlight the significant impact of environmental factors on microbial communities.

Our analytical tool not only enables us to discern that fecal contamination arises from diverse sources, including both human and animal origins, but it also underscores the predominant contribution of the human source. Additionally, our methodology extends its utility to the evaluation and validation of the long-term stability of microbial communities within aquatic environments, placing particular emphasis on rivers.

1. Introduction

The quality of bathing water is a significant global concern, primarily due to the risks posed by bacterial contamination. With natural water bodies serving as popular recreational hubs, guaranteeing the microbiological safety of these environments is crucial for upholding public health. Bacterial contamination, notably by fecal indicators such as *Escherichia coli* (E. coli) and fecal *Enterococcus*, poses a significant threat, potentially resulting in various waterborne diseases upon exposure. In response to this concern, rigorous monitoring methods have been established to assess and manage the microbial quality of bathing waters.

At the forefront of these efforts, the European Union has pioneered a stringent classification system for bathing areas, ensuring a harmonized approach to water quality assessment. The Bathing Water Directive, a cornerstone of this system, provides a standardized framework for evaluating and categorizing the microbial cleanliness of coastal and inland bathing waters across member states. This directive takes into account levels of fecal indicators and furnishes guidelines for designating bathing areas as compliant, non-compliant, or excellent based on established thresholds.

In the European context, the microbiological quality of river bathing areas is ensured through legislation, with an active monitoring plan in place, particularly in the picturesque Walloon region of Belgium. Nestled in the heart of Wallonia, where the allure of natural waters captivates visitors, the imperative of conscientious environmental stewardship becomes undeniably evident. This communication delves into the examination of two prominent rivers within this landscape, seeking to unravel the microorganisms that populate these aquatic ecosystems. Our paramount objective revolves around the preservation of public health and the perpetuation of environmental excellence.

Monitoring bathing water quality is an ongoing process involving regular assessments to detect potential contamination and prevent health hazards. Traditional methods, encompassing the enumeration of fecal indicator bacteria through culture-based techniques, are, however, time-consuming and may not provide real-time information. To address this, molecular techniques,

with 16S rDNA metagenomic analysis at the forefront, have emerged as powerful tools for assessing microbial communities without the need for taxonomic identification.

In response to the increasing demand for more efficient and accurate methods of tracing bacterial contamination in bathing water, this study presents the novel BiotopeBac-DB database and the Sequence Tracking (ST) tool. These advancements offer precise tracking of contamination sources, enhancing our ability to manage and mitigate bacterial risks in aquatic environments. The primary objective of this research is to improve the interpretation of metagenomic profiles and distinguish between organism-based and environmental sources of bacterial contamination using 16S rDNA metagenomic analysis. By bypassing traditional taxonomic identification, our database links reference sequences directly to their metadata, specifically their source origin. Our tool maps sampled sequences into our database, providing a more precise understanding of contamination sources. Consequently, it facilitates targeted interventions and significantly enhances the overall management of bathing water quality. Through the development and application of the BiotopeBac-DB database, this research aims to contribute to the advancement of methodologies for monitoring and safeguarding bathing water quality, aligning with the broader goals of public health and environmental stewardship.

Beyond unraveling microbial communities within riverine ecosystems, our study ventures into the realm of microbiome source tracking. With the aim of identifying potential sources of fecal contamination, our investigation spans a spectrum of potential hosts, from the animal kingdom to human activities, diligently pursuing the safeguarding of the purity and health of these bathing rivers. In this concise communication, we present our findings, illuminating the intricate microbial ecosystems and potential organism sources origin of contamination within these natural aquatic realms. This study contributes to the ever-growing body of knowledge dedicated to the preservation of the environment, ensuring the safety of bathers, and the enduring legacy of Wallonia's Belgium rivers.

2. Materials and methods

In 2014, two rivers (A and B) were selected for study, along with four distinct bathing zones: H07 (A_C), I14, I15, and I16 (B). The zone H07 (A_C), also known as 'Chiny,' is situated within the waters of River A. The other three zones, I14, I15, and I16, are located within the waters of River B. Samples were collected over three consecutive weeks, with each sampling point being sampled three times to ensure statistical robustness. Over the course of three weeks, we carried out weekly sampling, employing a systematic approach that has unveiled invaluable insights into the microbial communities inhabiting these natural aquatic environments. In 2017, our conducted a thorough sampling of River A to investigate microbial dynamics over time. We focused on two locations: River A (A_C) and River A (Lacuisine) (A_L). At A_C, sampling occurred every two hours over a 24-hour period, while at A_L, downstream from A_C, samples were taken every four hours. This approach aimed to comprehensively understand how microbial communities in these locations change over time. By intensifying the sampling intervals, we were able to identify subtle temporal patterns and fluctuations in microbial composition, providing deeper insights into the dynamic nature of River A's microbial ecosystems

2.2 Sample collection

Samples were collected from two distinct rivers during two separate time periods, 2014 and 2017. In 2014, we sampled four locations (spot 1, spot 2, spot 3, and spot 4) along River A_C in Chiny. Each location was sampled three times in the first week, yielding a total of 12 samples over a three-week period. For River B, situated in the Lesse River within the Ardennes region, we collected 36 samples from three locations over the same three-week period. The Lesse River, known for its dense forests, limestone caves, and periodic flooding issues, flows 89 km from its source near Libramont-Chevigny to its confluence with the Meuse River at Anseremme. It passes through the famous Han-sur-Lesse caves and several rural villages and farmlands. Altogether, we collected a total of 144 samples in 2014.

In 2017, we resumed our investigation with an intensified sampling regimen. For River A_C, samples were collected every 2 hours, resulting in 14 samples. For River A_L in La Cuisine, samples were collected every 4 hours, resulting in 12 samples. The sampling zones varied in distance and were influenced by different human activities such as rural farming, forestry, and small villages or towns not always well-served by wastewater treatment plants. The areas between each station included forests, livestock areas, and residential housing.

Sampling dates for River AC and River B in 2014 were between August and September, while in 2017, samples were collected from July to September during the summer season. Figure 1 illustrates the varying river locations. Samples details in All the corresponding libraries are available under BioProject ID PRJNA1134904 https://www.ncbi.nlm.nih.gov/bioproject/?term=PRJNA1134904 .

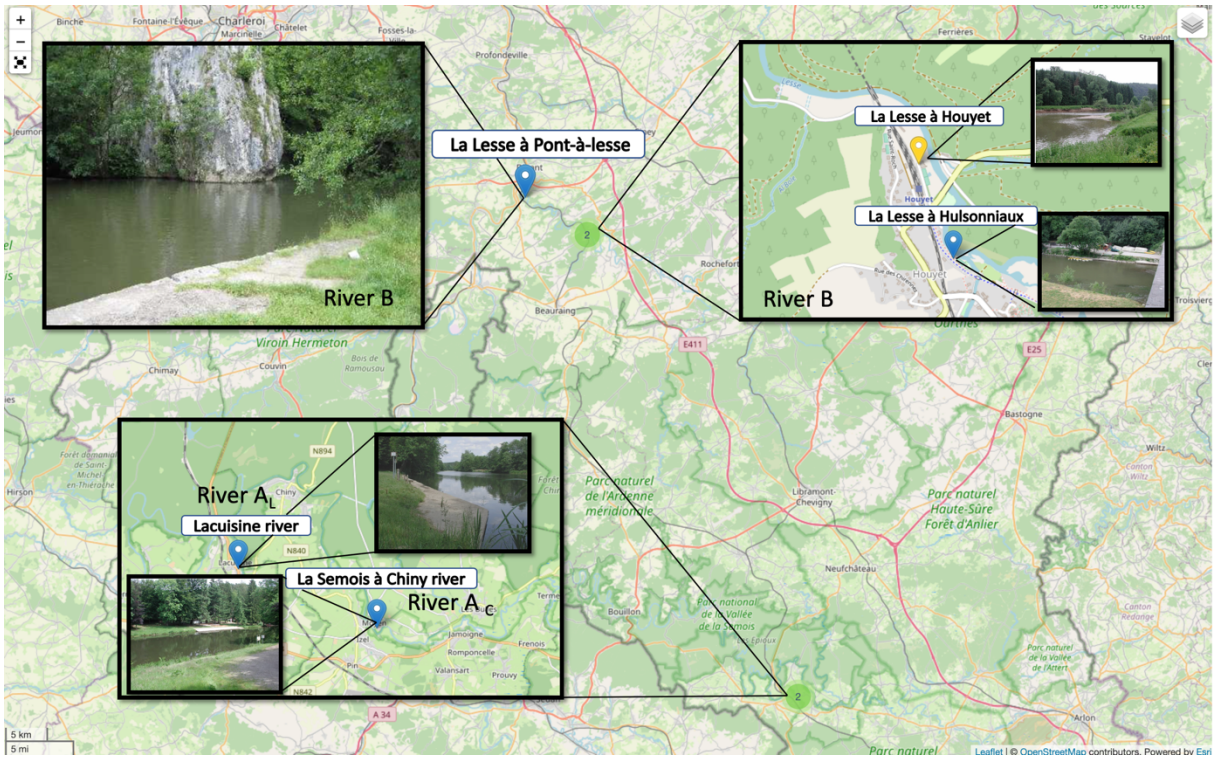


Figure 1. A visual map representation of the sampling locations for the different rivers during the two study periods. *The figure showcases the data collection sites, denoted by distinct markers for each river, offering a clear depiction of the meticulous sampling strategy employed in 2014 and 2017. The chart map as a valuable reference for understanding the temporal and spatial aspects of our research, enhancing the interpretation of the subsequent data and results presented in this study.*

2.3 Total DNA Extraction and Sequencing Library Preparation

One liter of collected water was filtered using 0.45 μm filters. Within 24 hours of sampling, total DNA extraction was performed using the DNeasy Blood & Tissue kit (69506, QIAGEN, 85764, Germany), following the manufacturer's instructions. The extracted DNA was then stored at -20°C. DNA concentration and purity assessment were carried out with a NanoDropTM 2000 (Thermo Fisher Scientific, Isogen Life Science B.V.,B-4000, Sart- Tilman, Belgium). PCR-amplification of the 16S rDNA V1–V3 hypervariable region and library preparation was performed with the following primers (with Illumina overhand adapters), forward (5′-GAGAGTTTGATYMTGGCTCAG-3′) and reverse (5′-AC-CGCGGCTGCTGGCAC-3′). Each PCR product was purified with the Agencourt AMPure XP beads kit (Beckman Coulter; Pasadena, CA, USA) and submi6ed to a second PCR round for indexing, using the Nextera XT index primers 1 and 2 (Illumina). After purification, PCR products were quantified using the Quant-IT PicoGreen (ThermoFisher Scientific; Waltham, MA, USA) and diluted to 10 ng/μL. A final quantification of each library was performed using the KAPA SYBR® FAST qPCR Kit (KapaBiosystems; Wilmington, MA, USA) before normalization, pooling and sequencing on a MiSeq sequencer using V3 reagents (Illumina; San Diego, CA, USA) [1].We conducted a stringent selection process, omitting samples with insufficient bacterial DNA content or evident PCR inhibitors. Initially, we collected 170 samples: 144 from 2014 and 26 from 2017, from Rivers A_C, B, and A_L. During the initial processing phase, we encountered low DNA yields in 84 samples: 59 from 2014 and 25 from 2017. Consequently, we excluded these 84 samples from the analysis, resulting in a final dataset of 86 samples. This careful curation process was crucial for ensuring the integrity, reliability, and accuracy of our data analysis, minimizing potential biases or limitations in interpreting our findings.

2.3. Bioinformatic Analysis

2.3.1. Microbial analysis

During the preprocessing stage, the Illumina adapters and primers were removed from the raw data. Subsequently, to ensure data quality and reliability, we applied essential filtering criteria using the command screen.seqs as a trimming process. These stringent filtering steps were crucial for maintaining the integrity of the sequences, as they enforced a maximum allowance of 1 ambiguous base and ensured that all sequences had a mini- mum length of 450 nucleotides. This curation process provided a solid foundation for sub- sequent analyses and interpretations in our scientific investigation.

Additionally, following the data curation, we used the MOTHUR software package v1.39.5 (Schloss et al., 2009) to check for chimeric amplification using the VSearch algorithm [2]. The resulting cleaned reads were then aligned to the SILVA database v1.32 [12]. To reduce computational complexity while preserving data representation, we subsampled the aligned reads, retaining 10,000 reads clustered into operational taxonomic units (OTUs) using the average neighbor algorithm from MOTHUR v1.39 with a 0.03 distance cut-off [1,3]. This subsampling approach allowed us to efficiently manage the dataset without compromising the accuracy of our taxonomic assignments.

The combined preprocessing and curation steps ensured that our dataset was well- prepared for subsequent analyses, and we are confident in the reliability of our findings. A taxonomic

identity was attributed to each (OTUs) by comparison with the SILVA data- base using an 80% homogeneity cutoff and a threshold of 0.50. The most abundant sequence for each OTU was compared with the SILVA dataset 1.32 version using the BLASTN algorithm to infer species assignment (1% mismatch threshold for specific labeling). Briefly, the species name, if known, or the corresponding NCBI accession number was used. Otherwise, for non-identical OTUs, the population was labeled with its corresponding OTU number. In addition to the taxonomic profiling, the OTU representative sequences were used to recover all source information and host-related information from our source-tracking database (based on sequence data).

The Shannon Diversity Index been measure of biodiversity that takes into account both the number of different species present (species richness) and their relative abundance (evenness) using R. Higher Shannon Diversity values indicate higher diversity within a community. Beta diversity was evaluated analyzing molecular variance (AMOVA) and homogeneity of molecular variance (HOMOVA) using Bray Curtis index.

2.3.2. Source Tracker Analysis

We developed a local 16S rDNA sequence database in our laboratory, linking sequences to their curated metadata and biotope information. Alongside taxonomic profiling, we utilized this database, named "BiotopeBac-DB," and the Sequence Tracking (ST) tool, also developed in our lab, to identify the source of contamination based on the taxonomic profiles of the samples.

Starting with the 16S rDNA v1.32 dataset from the SILVA database, we removed eukaryotic and vector entries. We retrieved the corresponding GenBank records of the remaining sequences, which included metadata and study titles, and curated them to retain host and environmental habitat information. This database contains 5 million published 16S rDNA sequences with validated taxonomic identities, labeled with publication and biotope information (host and habitat).

We reviewed the raw metadata from NCBI entries to encode animal and plant hosts with their eukaryotic taxonomic affiliations and retain biotope information. We added a second layer of global key terms (e.g., animal, plant, human, soil, water). We also created a catalog of annotated metadata terms using a controlled vocabulary, which is available at [this link](https://github.com/HibaJabri-project/Host_meta_db/blob/master/Host_Dico.obo.zip).
The source database file, "BiotopeBac.sdb," is stored in SQLite binary format and can be downloaded from Figshare under the DOI: [10.6084/m9.figshare.24499804](https://doi.org/10.6084/m9.figshare.24499804).

We grouped the sequences collected from the samples by clustering based on similarity, defining operational taxonomic units (OTUs) with a chosen similarity threshold (usually 97%). Using the corresponding accession numbers, we inferred the source origins. Bacterial source identification was conducted using sequence tracking with Sequence Tracking (ST), which can be downloaded from Figshare under the DOI: https://doi.org/10.6084/m9.figshare.25998043. We selected populations with known accessions and species assignments (maintaining a 99% homology threshold). The Sequence Tracking (ST) tool used these accession IDs to retrieve source tracking keywords from our database, generating a reference table with accession and keyword pairs.

The detailed methodology for tracking the source was previously published in an article in 2023 [4].

3. Results and discussion

3.1. Microbial and organism source profiling

In 2014, our investigation revealed heightened levels of *Pseudarcicella* and *Limnohabitans* in both River A_C and B (refer to Figure 2.a), aligning with their recognized prevalence in aquatic environments [5]. The study selected specific populations for analysis, relegating the remaining ones to the 'others' category, as visualized in Figure 2b. These inclusive 'others' category encapsulates populations linked to both human and animal organisms. Examining specific organism sources origin, particularly in spot 3 of River B and in River A_C, a pronounced influence of human activity emerges, emphasizing an intensified impact on River A_C.

Significantly, bacteria such as *Acidaminococcus* or *Akkermansia*, closely associated with human feces and intestinal infections [6], were observed. This observation sheds light on potential implications for water quality and public health. Furthermore, the study revealed a bacterial population associated with food at 0.18%, while bacteria originating from industrial activities represented 0.43%, as determined by our Sequence Tracking (ST) tool.

These findings underscore the diverse sources contributing to the microbial composition in the studied environments. River AC exhibited a more substantial influence of human activity compared to River B, with an observed impact of 0.32%, likely due to intensified human activities such as camping and favorable summer weather conditions. We identified various human-associated bacteria, including *Solobacterium, Bacteroides, Clostridium, Oxalobacteraceae, Streptococcus, Intestinibacter, Blautia, Faecalibacterium, Prevotellaceae, Ruminococcaceae, Romboutsia*, and *Oscillospirales*. Animal sources were also detected, with equids at 0.13%, bovids at 0.11%, other mammals at 0.8%, suids at 0.07%, and birds at 0.07%, implying contributions from farming practices or wildlife activity (see Figure 2.b). Specific bacteria associated with these animals included *Sericytochromatia, Arenimonas, Aquabacterium, Chitinivorax, and Sphingomonadaceae* for equids; *Clostridium, Jeotgalibaca, Proteiniphilum, Fermentimonas, and Bacteroides* for bovids; *Prevotella, Lactobacillus, Lachnospiraceae, Planococcaceae,* and *Sphaerochaeta* for suids; and *Chryseobacterium, Faecalibacterium, Saccharimonadales, Staphylococcus*, and *Microbacteriaceae* for birds. Notably, 1.68% of the organism sources were identified, corresponding to 400 sequences out of a total of 10,000 subsamples, with the remainder distributed among unknown sources (69.63%) and environmental sources (aquatic, soil, air, angiosperms, and gymnosperms) at 29.2%. Analyzing precipitation and temperature data from 2014, as documented in the Meteo Belgique report [7], revealed that the year ranked among the top five hottest in Belgium. This climatic condition likely contributed to increased human activity during that period.
Evaluating alpha diversity between River A_C and River B using the Shannon index, both rivers display variations in diversity across different locations. In specific spots, River A_C exhibits higher diversity than River B but the p-value (0.2617) but was not significant. Therefore, we suggest that there is no statistically significant difference in Shannon diversities between River A_C and River B. The negative t-value ($t = -1.182$) indicates that the mean Shannon diversity in River A_C is slightly lower than in River B, but this difference is not statistically significant based on the p-value given previously by factors such as water quality, habitat characteristics, and anthropogenic impacts.

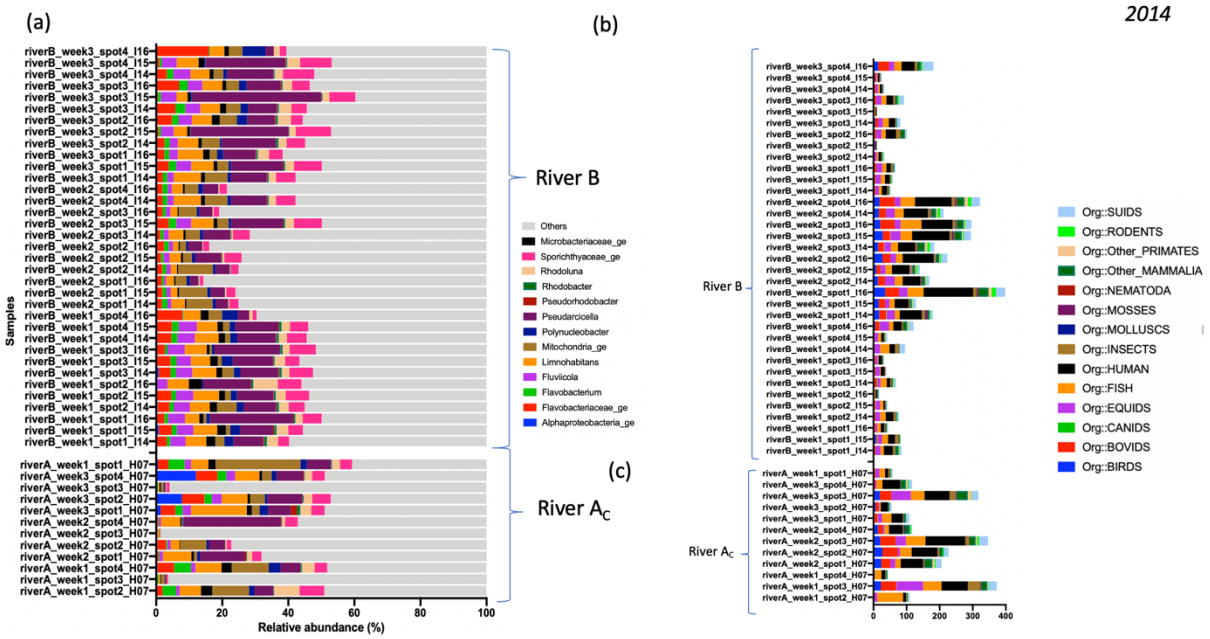


Figure 2: Microbial Composition and Organism Sources origin in Rivers A and B (2014). *Figure (a) illustrates the microbial composition in Rivers A and B during the year 2014, focusing on the prevalence of Pseudarcicella and Limnohabitans. The 13 most abundance bacteria population was presentend in the figure and the rest of the population was put in "others". Figure (b) present a bar chart showcases the distribution of Organism sources, highlighting the notable influence of human activity. Animal sources, including equids, bovids, suids, other mammals, and birds, are depicted, suggesting potential contributions from farming practices or wildlife activity. Additionally, the analysis reveals the identifiability of 1.68% of organism sources origin, with the remaining portion attributed to unknown sources (69.63%) and environmental factors (29.2%), such as aquatic, soil, air, angiosperms, and gymnosperms.*

Beta diversity was evaluated analyzing molecular variance (AMOVA) and homogeneity of molecular variance (HOMOVA) using Bray Curtis index presented in Figure 3. The Beta Diversity between River A_C and River B across different sampling spots in 2014 show a clear difference between the diversity of the two rivers. AMOVA and HOMOVA analysis, testing Bray Curtis dissimilarity index, revealed a significative difference of the beta diversity between River A_C and River B (Fig. 3.a). AMOVA analysis ($P < 0,5$), shown a significantly different clustering of River A_C from River B. While HOMOVA analysis ($P < 0,5$), shown a significative different dispersion between the samples of both matrices. Each spot is a unique ecological niche that influences the composition of microorganisms (Figure 3.a).

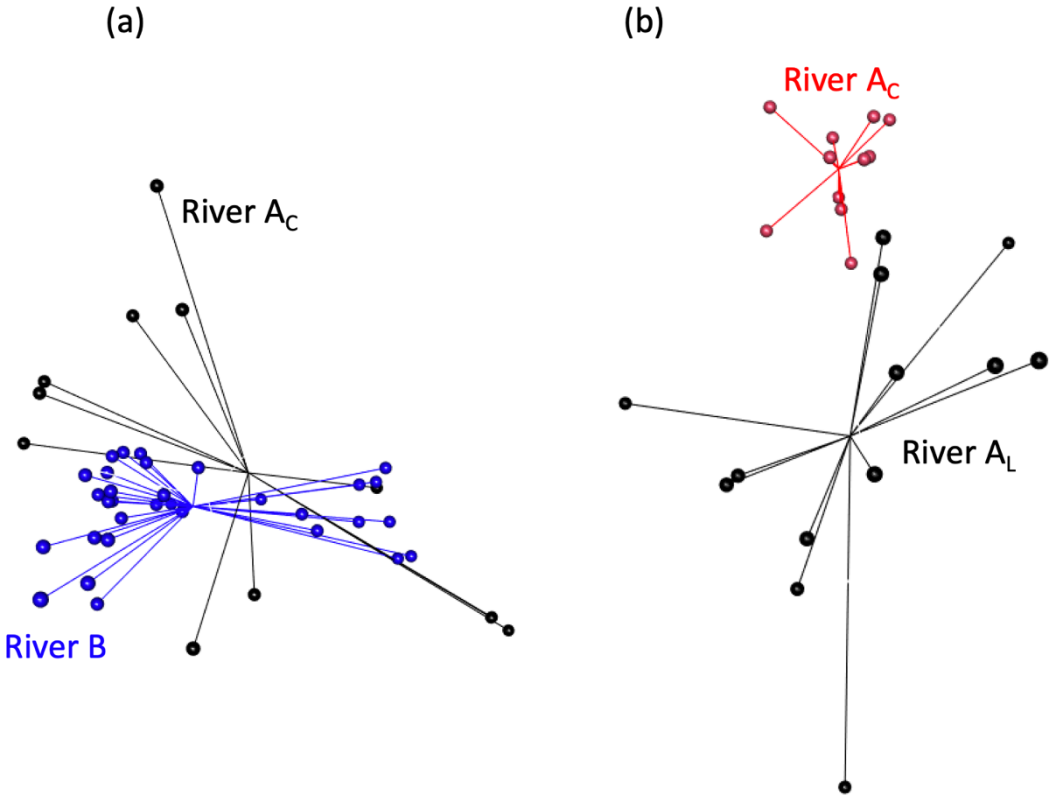


Figure 3. Non-parametric multidimensional scaling (NMDS) representation of (a) River A_C (black balls) and River B (blue balls) samples. (b) River A_L (black balls) and River Ac (red balls) based on a Bray-Curtis dissimilarity matrix. *The accepted model presents 4 dimensions (k) for an acceptable stress value (a) of 0,055 for the study of 2014 and (b) of 0,030 for the study of 2017. The centromere of the two groups (equidistant point between samples of the same group) is illustrated by a "spider web" connecting the samples.*

In 2017, our observations revealed that the microbial profiling in River A_C and A_L did not exhibit significant variations when sampled multiple times per day (Figure 4a), indicating the stability of the microbiota in dynamic aquatic environments. Notably, River A_C displayed a dominance of *Limnohabitans* (30%) and *Cyanobiaceae* (32%), underscoring the resilience and consistent composition of the microbial community in this particular waterway (see Figure 4a). Through observations, we have arrived at the conclusion that there is no apparent influence of temporal variability on the microbial composition, particularly in River A_L compared to River A_C comparing Shannon diversity index. The p-value of 0.01968 is less than the common significance level of 0.05. Therefore, there is a statistically significant difference in the Shannon diversity between River A_C and River A_L. The negative t-value (t = -2.655) and the negative lower bound of the confidence interval indicate that, on average, River A_C has a lower Shannon diversity compared to River A_L.

Our comprehensive examination and analysis furnish compelling evidence that underscores the stability and constancy of the observed microbial profiles over time. In 2017, the notable feature was the unusually high temperatures, complemented by sunshine and precipitation patterns (both in quantity and frequency) that aligned with established norms, as reported on [8]. This climatic context may elucidate the elevated presence of human activity observed on River A_C. However, River A_L exhibited only a 3.76% presence of organism source origin in the water

(Fig 4.b), which we found 57.73% unknow source and 38.51% environmental correlated with observation given the bacterial profile characterized by the predominant abundance of *Limnohabitans* in river A_C and *Cyanobaiaceae* genus in river A_L as known environmental bacteria origin [9,10]. Beta diversity between the two sites in 2017, across different sampling times within a 24-hour period, shows a distinct difference (Figure 3.b). These differences can be attributed to human activities at the two sites, such as the presence of a camping area and a wastewater treatment plant near the bathing water sampling locations. AMOVA and HOMOVA analysis, testing Bray Curtis dissimilarity index, revealed a significative difference of the beta diversity between sites A_C and A_L (Fig. 3.b). AMOVA analysis ($P < 0,5$), shown a significantly different clustering of River A_C from River A_L. While HOMOVA analysis ($P < 0,5$), shown a significative different dispersion between the samples of both matrices. This indicates that locations AC and AL exhibit two distinct bacterial populations in terms of their proportions. However, no significant clustering of bacterial populations was observed within samples from the same river. This redundancy indicates that the microbiota in each part of the river is quite similar in nature. However, the AMOVA and HOMOVA tests reveal distinct differences between the sites (Fig. 2), likely due to varying levels of human activity and environmental influences at different locations along the river (Fig. 4). Therefore, from an ecological perspective, this may correspond two different ecological niches forming part of a complex ecosystem as it has been described for rivers [11,12] specially in Belgium [13,14] and in human health risks [15].

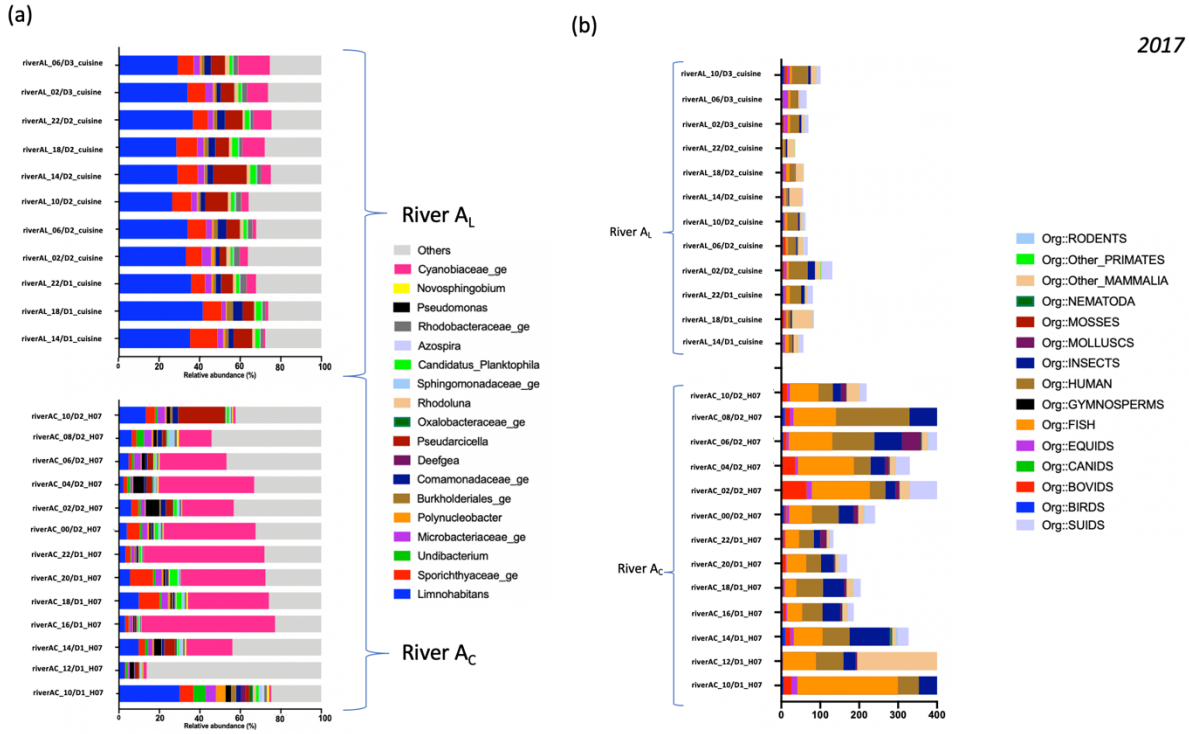


Figure 4. Microbial Composition and Organism Sources origin in Rivers A_C and A_L (2017). (a) *This figure illustrates the microbial composition in sampling sites A_C and A_L during the year 2017 during 24 hours, focusing on the prevalence of Limnohabitans and Cyanobiaceae as major genera present on the different spost of river.* (b) *The bar chart showcases the distribution of organism sources origin, highlighting the notable influence of human activity. Animal sources, including equids, bovids, suids, other mammals, and birds, are depicted, suggesting potential contributions from farming practices or wildlife activity. Additionally, the analysis reveals the identifiability of 40% of organism sources origin, with the remaining*

portion attributed to unknown sources (22%) and environmental factors (45%), such as aquatic, soil, air, angiosperms, and gymnosperms.

Our tool and BiotopeBac-DB database demonstrate that fecal contamination may originate from various organism sources origin, extending beyond just human, as evidenced by the tracking of amplicon sequences without the need for traditional microbiota culture analysis. The subsequent set of samples collected in 2017 was designed to assess and validate the long-term stability of the microbiota in aquatic environments, with a particular focus on rivers. The results unequivocally demonstrated the efficacy of this approach, confirming the enduring stability of the microbiota over time. A clear correlation emerges between the sources of the microbiota and their distribution. The distribution of organism sources origin reveals a significant animal influence, particularly in River A_C, involving bovids, suids, and other mammals (see Fig. 4), indicating animal-related activities within these river ecosystems.

This broader perspective positions our study as a versatile methodology for tracking bacteria sources across different rivers, presenting valuable applications in mitigating potential human health risks and preventing contamination. By testing our tool in diverse environments, we proactively identify and address bacterial sources that may threaten public health. In rural river systems, like the focus of our study, characterized by prevalent human activities, our methodology becomes crucial for pinpointing specific fecal contamination sources, enabling targeted interventions to ensure water safety. Similarly, in agricultural landscapes, our tool's ability to trace bacterial origins aids in implementing measures to minimize runoff and protect water resources from agricultural-related contaminants. This adaptability underscores the potential of our tool, not only for scientific inquiry but also for practical applications in safeguarding public health and environmental quality.

Acknowledgements

The Department of Water and Environment (DEE) and the Institut Scientifique de Service Public (ISSeP) are acknowledged for their invaluable collaboration. Appreciation is also extended to the students who assisted with sampling in 2014 and 2017. This research did not receive any specific funding from public, commercial, or not-for-profit organizations

II. References

1. Ceugniez, A.; Taminiau, B.; Coucheney, F.; Jacques, P.; Delcenserie, V.; Daube, G.; Drider, D. Use of a metagenetic approach to monitor the bacterial microbiota of "Tomme d'Orchies" cheese during the ripening process. *Int. J. Food Microbiol.* **2017**, *247*, 65– 69. h:ps://doi.org/10.1016/j.ijfoodmicro.2016.10.034.
2. Rognes, T.; Flouri, T.; Nichols, B.; Quince, C.; Mahé, F. VSEARCH: A versatile open source tool for metagenomics. *PeerJ* **2016**, *2016*, e2584, h:ps://doi.org/10.7717/peerj.2584.
3. Rodriguez, C.; Taminiau, B.; Brévers, B.; Avesani, V.; Van Broeck, J.; Leroux, A.; Gallot, M.; Bruwier, A.; Amory, H.; Delmée, M.; et al. Faecal microbiota characterisation of horses using 16 rdna barcoded pyrosequencing, and carriage rate of clostridium difficile at hospital admission. *BMC Microbiol.* **2015**, *15*, 181. h:ps://doi.org/10.1186/s12866-015-0514-5.
4. Jabri H, Krings S, Fall PA, Baurain D, Daube G, Taminiau B. Microbiota Profiling on Veterinary Faculty Restroom Surfaces and Source Tracking. Microorganisms. 2023;11(8):2053. Published 2023 Aug 10. doi:10.3390/microorganisms11082053
5. Cruaud, Perrine et al. "Rapid Changes in Microbial Community Structures along a Meandering River." Microorganisms vol. 8,11 1631. 22 Oct. 2020, doi:10.3390/microorganisms8111631
6. Braun, T., Halevi, S., Hadar, R., Efroni, G., Glick Saar, E., Keller, N., ... Haberman, Y. (2021). SARS-CoV-2 does not have a strong effect on the nasopharyngeal microbial composition. Scientific Reports, 11.
7. https://www.meteobelgique.be/article/100-annee-2014/2043-annee-2014-les-chiffres
8. https://www.meteobelgique.be/article/104-annee-2017/2237-bilan-de-l-annee-2017
9. SINGH, Jay Shankar. Cyanobacteria: a vital bio-agent in eco-restoration of degraded lands and sustainable agriculture. Climate Change and Environmental Sustainability, 2014, vol. 2, no 2, p. 133-137.
10. CASTENHOLZ, Richard W., WILMOTTE, Annick, HERDMAN, Michael, et al. Phylum BX. cyanobacteria. In : Bergey's manual® of systematic bacteriology. Springer, New York, NY, 2001. p. 473-599.
11. Yu Z, Zhang J, Wang H, et al. Quantitative analysis of ecological suitability and stability of meandering rivers. Front Biosci (Landmark Ed). 2022;27(2):42. doi:10.31083/j.fbl2702042
12. González-Paz L, Delgado C, Pardo I. How good is good ecological status? A test across river typologies, diatom indices and biological elements. Sci Total Environ. 2022;815:152901. doi:10.1016/j.scitotenv.2021.152901

13. Thijs S, Op De Beeck M, Beckers B, et al. Comparative Evaluation of Four Bacteria-Specific Primer Pairs for 16S rRNA Gene Surveys. Front Microbiol. 2017;8:494. Published 2017 Mar 28. doi:10.3389/fmicb.2017.00494
14. García-Armisen T, İnceoğlu Ö, Ouattara NK, et al. Seasonal variations and resilience of bacterial communities in a sewage polluted urban river. PLoS One. 2014;9(3):e92579. Published 2014 Mar 25. doi:10.1371/journal.pone.0092579
15. Braeye T, DE Schrijver K, Wollants E, van Ranst M, Verhaegen J. A large community outbreak of gastroenteritis associated with consumption of drinking water contaminated by river water, Belgium, 2010. Epidemiol Infect. 2015;143(4):711-719. doi:10.1017/S0950268814001629

1

Chapter V

General Discussion, Conclusion and Perspective

In this thesis, we aimed to establish a direct correlation between bacterial sequences and their habitat sources, moving beyond reliance on taxonomic names alone. By leveraging sequence-related data, we conducted risk assessments to identify specific sources of bacterial organisms. To achieve this, we developed "BiotopeBac-DB", an open relational database constructed using SQL and rigorously validated. This curated database was designed to store bacterial taxonomic annotations alongside detailed metadata about their sampled habitats, integrating annotated sequences from public databases with additional environmental information. The database was organized and generated within our laboratory, with a focus on creating an accessible resource that seamlessly combines reliable biotope information with bacterial taxonomic data derived from complete 16S rDNA sequences. The selection of the 16S rDNA genetic marker for taxonomic identification was a critical aspect of our methodology, balancing cost-effectiveness with widely accepted scientific practices in bacterial identification.

The operationalization of the database followed a systematic approach, involving several key steps: collecting and organizing 16S rDNA sequence datasets, annotating metadata for sampled environments, developing search tools, and utilizing the database for hypothesis testing. This process was divided into multiple phases, each carefully designed to support the construction, validation, and application of the BiotopeBac-DB database. These phases included data extraction and organization, manual annotation of terminology, creation of search tools (Metagenomic Source Tool "MGST" and Sequence Tracking tool "ST"), database validation, and testing methodologies using datasets from specific environments.

Similar tools, such as MetaPhlAn, SourceTracker, and SeqEnv, have also utilized metagenomic data to identify microbial sources. For instance, SourceTracker employs Bayesian methods to predict the sources of microbial communities in environmental samples (Knights et al., 2011), while MetaPhlAn uses marker genes to profile microbial communities (Segata et al., 2012). SeqEnv focuses on linking environmental sequences to habitats by assigning environmental metadata to taxonomic groups, providing insights into their ecological context (Sinclair et al., 2016). While effective, these tools often rely heavily on taxonomic classifications and do not integrate habitat information as extensively as BiotopeBac-DB. Other tools, such as FORENSIC and BioNLP, utilize different methodologies for microbial source tracking. FORENSIC focuses on identifying sources of fecal contamination through machine learning models (Wilkins et al., 2014), while BioNLP provides natural language processing techniques to annotate microbial sequences (Arighi et al., 2011). However, our approach differs by directly

linking 16S rDNA sequences to their respective biotopes, enabling more precise assignment of bacterial population sources and offering substantial clues regarding their origin and characteristics.

During the process of sequence annotation and taxonomic identification of the 16S rDNA gene, certain annotations, such as "uncultured bacteria," were inaccurately assigned due to various factors. These included sequence variability, where the 16S rDNA gene displayed variable regions among different taxa, posing challenges in creating universally applicable reference databases and alignment algorithms. Additionally, limitations in existing databases could result in incomplete identifications or mis-annotations. Misidentified sequences could arise from errors in sample labeling, sequencing artifacts, or inaccuracies in database entries. The presence of uncultured or poorly characterized bacterial species further contributed to annotations like "unclassified/uncultured" or "environmental bacteria." Moreover, the taxonomic resolution provided by 16S rDNA sequences was not always sufficient to differentiate closely related species or strains accurately, leading to ambiguous or inaccurate identifications. Issues with sequence quality, such as sequencing errors or contamination, could further complicate taxonomic annotations. Methodological biases, including the choice of sequencing platform, primer selection, and PCR conditions, could also introduce biases that impacted taxonomic assignments. Given these challenges, the second objective of this thesis aimed to complement metagenomic analysis with a parallel investigation into source identification for annotated and misannotated sequences. This work acknowledged the pitfalls of bacterial identification faced by microbiology, highlighting the inadequacy of relying solely on biochemical properties, poorly annotated sequences, or colony appearance for species assignment. In response, we advocated a new approach that bypassed species labeling and directly linked 16S rDNA sequences to their respective biotopes, enabling more precise assignment of bacterial population sources and offering substantial clues regarding their origin and characteristics.

The third objective of the thesis focused on distinguishing between bacteria of host origin and those originating from the environment. This objective was investigated through two comprehensive studies conducted as part of the thesis. The first study aimed to assess the capability of discerning the origin of bacterial 16S rDNA sequences in veterinary restrooms. By analyzing the genetic sequences present in these indoor environments, the study sought to differentiate between bacteria derived from human occupants and those originating from animals. The second study delved into bacterial source tracking in natural river bathing waters,

aiming to identify and distinguish sources of contamination in these aquatic environments, particularly differentiating between bacteria originating from human activities and those naturally present in the environment. Through these studies, the thesis advanced understanding and methodologies for identifying the origin of bacteria, contributing to the fields of microbial ecology and source tracking.

Our approach incorporated curated annotation and statistical modeling to identify hidden patterns among uncultured or unknown bacteria. By utilizing large datasets and sophisticated algorithms, we predicted the probable habitats or sources of these "unknown" species. This methodology not only refined our understanding of microbial ecology but also enhanced the accuracy of contamination source tracking and environmental monitoring. However, it is critical to emphasize that metagenomic approaches like ours cannot replace traditional culture-based methods, as these techniques serve fundamentally different objectives. Culture-based methods remain indispensable for isolating live organisms, studying their functional traits (e.g., metabolism, pathogenicity), and validating hypotheses about microbial behavior in controlled settings. In contrast, metagenomics provides a broad, community-level perspective of genetic potential and environmental distribution, often uncovering taxa that resist cultivation. Together, these complementary approaches cultivation for functional validation and metagenomics for ecological inference advance a holistic understanding of microbial systems (Steen et al., 2019).

In summary, this thesis contributes to the field of source tracking within 16S rDNA metagenetic analysis, specifically assisting microbiologists in obtaining curated and better annotations of contamination sources. Beginning with the construction and validation of the “BiotopeBac-DB” database, along with the development of the associated methodology detailed in the previous chapters, our work started with foundational efforts in chapter one. This initial phase focused on creating the “BiotopeBac-DB” database, the MGST validation tool, and the ST software, which collectively provide essential resources and tools to enhance source tracking studies in microbiology.

The experimental phase of this study involved two distinct sampling campaigns (2014 and 2017) conducted under consistent methodological conditions, including identical sequencing protocols, filtration techniques, and bioinformatic pipelines to ensure comparability. While the core methodology remained unchanged, the sampling conditions and temporal contexts differed to address two critical questions. First, we aimed to determine whether bacterial biodiversity

and contamination sources (e.g., animal vs. human) varied significantly between two geographically distinct sampling areas (River A and B in bathing water sites). Second, we investigated temporal dynamics within the same sampling area over a 24-hour period to assess short-term fluctuations in bacterial diversity and contamination sources. To achieve this, we applied our newly developed Sequence Tracking Tool (ST) to analyze datasets from both campaigns. This tool enabled precise identification of bacterial sources (e.g., differentiating animal-derived Bacteroides from human-associated variants) while accounting for temporal variations in microbial communities. By maintaining methodological consistency across campaigns, we isolated the effects of spatial heterogeneity (differences between sampling areas) and temporal shifts (changes within a single site over 24 hours) on bacterial diversity. This approach not only validated the utility of the ST tool but also highlighted its capacity to disentangle complex ecological patterns, such as the persistence of contamination signatures or the influence of transient animal presence on microbial communities.

The results from this phase, presented in this thesis, underscore the importance of standardized protocols in longitudinal and spatial microbial studies. They also demonstrate how integrating temporal and spatial dimensions into microbial tracking can refine our understanding of contamination pathways and ecosystem resilience. Our methodology proved robust, showing precise source tracking across diverse environments. Through extensive sampling, amplicon sequencing, and the use of the “BiotopeBac-DB” database, we investigated the microbiological dynamics of frequently visited restrooms and aquatic environments. The results highlighted the database’s effectiveness in differentiating bacterial sources between animal and human origins in varied settings, including aquatic ecosystems. This multi-environment validation underscores the adaptability and reliability of our database, tools, and methodology, illustrating their practical value in enhancing source tracking within metagenomic analysis.

In essence, the trajectory woven across these chapters illustrates a comprehensive and iterative approach to database construction and validation. From its inception and validations in chapter three to the refined practical applications in the experimental section, our research underscores the robustness of the “BiotopeBac-DB” database in precisely discerning microbial sources. As we deepen our comprehension of microbial dynamics in varied environments, the applicability and versatility of our methodology become increasingly apparent. These findings make a substantial contribution to the field of microbial source tracking, providing a valuable tool for environmental management and public health interventions. The chapters, harmonizing with

their predecessors, construct a cohesive narrative, further solidifying the reliability and relevance of the BiotopeBac-DB database in advancing our understanding of bacterial sources across diverse ecosystems.

Our approach to biotope classification relied on constructing a novel ontological framework based on biological and taxonomic data rather than adopting pre-existing dictionaries or ontologies. While this allowed us to tailor the system to our specific research objectives, it also introduces limitations in exhaustiveness. Given the vast diversity of microbial habitats and the dynamic nature of microbial communities, it is challenging to ensure complete coverage of all possible biotopes. Certain environments, particularly understudied or poorly characterized niches (e.g., host-associated microenvironments or extreme habitats), may not yet be adequately represented in our database. Additionally, the absence of standardized terminology for some microbial habitats complicates their unambiguous classification. To address these gaps, future work should integrate collaborative efforts with existing ontological resources (e.g., ENVO, MIMARKS) and expand sampling efforts to include underrepresented environments. This iterative refinement will enhance both the resolution and inclusivity of biotope annotations, aligning with broader goals of ecological and biotechnological research.

This work represents a significant step forward by leveraging the direct correlation between bacterial sequences and their habitat sources, moving beyond traditional taxonomic approaches. We developed “BiotopeBac-DB”, an advanced relational database that integrates bacterial taxonomic annotations with detailed metadata about their sampled habitats. This curated resource provides a comprehensive and precise tool for microbial source tracking. BiotopeBac-DB enhances environmental management and public health interventions by offering valuable insights into bacterial origins. Its integration of sequence data and habitat information makes it an indispensable resource across multiple disciplines, facilitating more accurate and effective analyses.

While the 16S rDNA has been a cornerstone of microbial identification, future advancements could benefit from integrating Metagenome-Assembled Genomes (MAGs) into our methodology. Unlike the 16S rDNA, which targets a single gene, MAGs provide a more comprehensive view of microbial genomes, enabling researchers to study not only taxonomic identity but also functional potential. For example, the presence of genes related to antibiotic resistance (MAR) or other metabolic functions can be identified through MAGs. However, it is

important to note that the presence of a gene does not necessarily indicate its expression or functionality. To fully understand whether a gene is active, complementary RNA sequencing (RNA-seq) is required to analyze gene expression levels. This dual approach combining MAGs for functional potential and RNA-seq for expression would provide a more holistic understanding of microbial communities and their roles in various environments.

While this thesis focuses on bacterial source tracking using 16S rDNA, the ST framework holds potential for adaptation to other domains, such as viral or bacteriophage-based microbial source tracking (MST). For example, bacteriophages like crAssphage a human gut-associated phage are increasingly recognized as stable, indirect markers of fecal contamination due to their environmental persistence and host specificity. Unlike bacterial markers, which may degrade rapidly or exhibit temporal variability, phages like crAssphage offer a more robust signal for tracking contamination sources over extended periods (Ahmed et al., 2015). The ST tool could theoretically be applied to viral DNA extracted from environmental samples, leveraging sequencing data from phage genomes to identify contamination sources. This approach would mirror our bacterial framework but capitalize on the stability of viral markers. For instance, Ahmed et al. (2015) demonstrated that crAssphage markers outperform traditional Bacteroidales-based methods in detecting human fecal contamination in surface water, highlighting the utility of phage-based MST. Integrating such viral markers into the ST framework could enhance its resolution for environments where bacterial signals are transient or ambiguous. However, transitioning to viral targets would require adjustments, such as optimizing viral DNA extraction protocols and expanding reference databases to include phage genomes. Future studies could explore hybrid approaches, combining bacterial and phage markers within the ST framework to improve accuracy in complex environments. This dual-marker strategy would align with emerging trends in MST, where multi-target systems are increasingly advocated to address the limitations of single-marker systems (Harwood et al., 2019).

By incorporating MAGs, our methodology could move beyond taxonomic classification to explore the functional capabilities of microbial communities. This would allow us to identify not only who is present but also what they can do, such as their potential for antibiotic resistance, nutrient cycling, or pathogenicity. However, this approach also introduces new challenges, such as the need for more sophisticated computational tools and larger datasets to handle the complexity of whole-genome analyses. Despite these challenges, the integration of

MAGs and RNA-seq represents a promising direction for future research, offering deeper insights into microbial ecology and enhancing our ability to track and manage microbial sources in diverse environments.

Future enhancements could involve the integration of artificial intelligence (AI) and machine learning (ML) to analyze and interpret the vast amount of data generated. AI algorithms can identify hidden patterns and correlations between bacterial sequences and their habitats, which are not immediately apparent through traditional methods. By applying ML techniques, we could predict the probable habitat or source of "unknown" bacterial species, refining our understanding of microbial ecology and enhancing the accuracy of contamination source tracking. Specifically, we propose training models that can predict bacterial sources without the necessity of building a static database. Instead, these models would dynamically gather public information from various sources, such as genomic repositories, environmental records, and scientific literature. This data would then be structured into a knowledge graph, which the AI system could continuously update and expand. The knowledge graph would enable the AI to link disparate pieces of information, providing a comprehensive contextual understanding of bacterial sequences. For example, the graph could connect specific bacterial genes to known habitats, environmental conditions, and related microbial communities. Using this enriched data structure, the AI could make more informed predictions about the origins of newly sequenced bacterial strains. This approach would significantly enhance the efficacy and flexibility of our system, allowing it to adapt to new information and evolving scientific knowledge.

Additionally, leveraging natural language processing (NLP) techniques would allow the AI to extract relevant data from unstructured sources, such as research papers and reports. This capability would further enrich the knowledge graph, ensuring that the AI has access to the most current and comprehensive data available. By combining these advanced AI methodologies, our system could provide real-time, accurate predictions about bacterial sources, greatly improving the speed and precision of contamination source tracking and microbial ecology studies.

Transitioning our database to a cloud-based platform would enable more efficient data management and accessibility. The cloud can provide scalable storage solutions, ensuring that as our dataset grows, we can maintain performance and accessibility. Additionally, cloud-based

solutions facilitate easier data sharing and collaboration among researchers globally, enhancing the collective effort to improve microbial source tracking.

Employing ETL (Extract, Transform, Load) tools can streamline the process of managing and curating the data. These tools can automate the extraction of raw data from various sources, transform it into a suitable format for analysis, and load it into our database. This automation reduces the risk of errors and improves the efficiency and reliability of our data processing pipeline. Examples of ETL tools that could be particularly beneficial for our study include Apache NiFi, which offers a robust and flexible data flow management system (Apache Software Foundation, 2023), and Talend, known for its user-friendly interface and comprehensive data integration capabilities (Talend, 2023). Another valuable tool is Apache Airflow, which excels in orchestrating complex workflows and can easily integrate with various data sources and transformation processes (Apache Software Foundation, 2023). Using these ETL tools, we can ensure our data pipeline is both efficient and scalable, facilitating more accurate and timely analysis.

Additionally, we can enhance our automated data processing capabilities by incorporating workflow management systems such as Nextflow. Nextflow allows for the efficient orchestration of bioinformatics pipelines, handling complex data workflows and enabling seamless integration with various computational environments (Di Tommaso et al., 2017). By leveraging Nextflow, we can automate and optimize our data processing pipeline further, ensuring that our analysis remains robust, reproducible, and scalable.

Combining multiple ontologies to enrich the annotation of our metadata, particularly through Natural Language Processing (NLP) techniques, can significantly enhance the value of our curated data. By leveraging NLP for annotation, we can accurately extract and standardize terminologies from not only metadata and titles used in our study, but also from other sources such as descriptions and full-text articles in microorganisms reviews and conferences. This comprehensive approach allows us to capture nuanced information present in unstructured text data, complementing structured ontology-based annotations. This integration facilitates better data interoperability and more robust analyses, as it enables us to bridge the gap between structured and unstructured data, ultimately leading to more insightful conclusions and discoveries.

Finally, the creation of a user-friendly website for our database is poised to revolutionize accessibility and usability within the field of source tracking in metagenomic analysis. Beyond simply accessing data, this platform will serve as a hub for comprehensive analysis, offering a range of powerful tools and features. For instance, researchers can utilize advanced search functionalities to pinpoint specific microbial communities or genetic signatures within vast datasets. The integration of interactive visualization tools will enable users to explore complex data structures with ease, facilitating deeper insights into microbial dynamics and interactions. Moreover, the website will support real-time updates, ensuring that researchers have access to the most current information for their analyses. This dynamic nature enhances the relevance and reliability of the data, particularly in rapidly evolving research areas such as microbiome studies. Additionally, the availability of downloadable datasets in standardized formats will promote interoperability and reproducibility across research endeavors, fostering collaboration and knowledge exchange within the scientific community.

In conclusion, this thesis marks a significant step forward in enhancing source tracking methodologies within metagenomic analysis. From the creation of the BiotopeBac-DB database to the development and validation of innovative tools such as the MGST in the third chapter, and the application of the ST tool in studies on restrooms and bathing water, this work establishes a solid foundation for the field. It provides a basis for future research and applications, with great potential for further development. Looking ahead, there are substantial opportunities to build on this work, fostering ongoing progress and innovation to address emerging challenges in understanding and managing microbial ecosystems. By continually improving user-friendly platforms and tools, we can leverage metagenomic data to better understand the complexities of microbial communities and their ecological roles, ultimately advancing scientific knowledge and contributing to human health and environmental sustainability.

Bibliography

Ahmed, W., et al. (2015). Validation of crAssphage as a microbial source tracking marker and comparison with Bacteroidales markers in water. Science of the Total Environment, 542, 976–981.
Almeida, A., Mitchell, A. L., Tarkowska, A., & Finn, R. D. (2018). Benchmarking taxonomic assignments based on 16S rRNA gene profiling of the microbiota from commonly sampled environments. GigaScience, 7(5). https://doi.org/10.1093/gigascience/giy054
Amann, R. I., et al. (1995). Phylogenetic identification and in situ detection of individual microbial cells without cultivation. *Microbiological Reviews*, 59(1), 143-169.
An U, Shenhav L, Olson CA, Hsiao EY, Halperin E, Sankararaman S. STENSL: Microbial Source Tracking with ENvironment SeLection. mSystems. 2022;7(5):e0099521. doi:10.1128/msystems.00995-21
Apache Software Foundation. (2023). Apache NiFi. Retrieved from https://nifi.apache.org
Arumugam, M., Raes, J., Pelletier, E., Le Paslier, D., Yamada, T., Mende, D. R., ... & Bork, P. (2011). Enterotypes of the human gut microbiome. *Nature*, 473(7346), 174-180. https://doi.org/10.1038/nature09944
Ashburner, M., et al. (2000). Gene Ontology: Tool for the unification of biology. Nature Genetics, 25(1), 25–29
Baker, G. C., et al. (2003). Review and re-analysis of domain-specific 16S primers. Journal of Microbiological Methods, 55(3), 541–555.
Benítez-Páez A, Portune KJ, Sanz Y. Species-level resolution of 16S rRNA gene amplicons sequenced through the MinIONTM portable nanopore sequencer. GigaScience 2016;5:4. https://doi.org/10.1186/s13742-016-0111-z.
Bernhard, A. E., & Field, K. G. (2000). A PCR assay to discriminate human and ruminant feces on the basis of host differences in *Bacteroides*-Prevotella genes. *Applied and Environmental Microbiology*, 66(10), 4571-4574.
Bernhard , A. E. , and Field , K. G. (2000) Identification of nonpoint sources of fecal pollution in coastal waters by using host-specifi c 16S ribosomal DNA markers from fecal anaerobes . Appl. Environ. Microbiol. 66 , 1587 – 1594 .
Bernhard , A. E. , Goyard , T. , Simonich , M. , and Field , K. G. (2003) A rapid method for identifying fecal pollution sources in coastal waters . Water Res. 37 , 909 – 913 .
Bergey, D. H., et al. (1923). Bergey's Manual of Determinative Bacteriology. Baltimore: Williams & Wilkins.
Bharti, R., et al. (2019). Current challenges and best-practice protocols for microbiome analysis. Briefings in Bioinformatics, 20(1), 1–15.
Botterel F, Angebault C, Cabaret O, Stressmann FA, Costa JM, Wallet F, et al. Fungal and bacterial diversity of airway microbiota in adults with cystic fibrosis: concordance between conventional methods and ultra-deep sequencing, and their practical use in the clinical laboratory. Mycopathologia. 2018;183:171–83.
Bowen M, Farag IF, Main CR, Biddle JF. Reference library for microbial source tracking in the mid-Atlantic United States. Microbiol Resour Announc. Published online November 30, 2023. doi:10.1128/MRA.00674-23
Bibby, K., & Peccia, J. (2013). Identification of viral pathogen diversity in sewage sludge by metagenome analysis. *Environmental Science & Technology*, 47(4), 1945–1951.
Buchanan, R. E. (1918). Studies in the Nomenclature and Classification of Bacteria. Journal of Bacteriology, 3(1), 27–61.
Bushnell B, Rood J, Singer E. BBMerge - Accurate paired shotgun read merging via overlap. PLoS One. 2017;12(10):e0185056. Published 2017 Oct 26. doi:10.1371/journal.pone.0185056
Bustin, S. A. (2000). Absolute quantification of mRNA using real-time reverse transcription polymerase chain reaction assays. Journal of Molecular Endocrinology, 25(2), 169-193.

Buttigieg, P. L., et al. (2016). The environment ontology in 2016: Bridging domains with increased scope, semantic density, and interoperation. Journal of Biomedical Semantics, 7(1), 57. DOI: 10.1186/s13326-016-0097-y

Buttigieg, P. L., Morrison, N., Smith, B., Mungall, C. J., & Lewis, S. E. (2013a). The environment ontology: contextualising biological and biomedical entities. Journal of Biomedical Semantics. www.environmentontology.org

Chaix, E., Deléger, L., Bossy, R., & Nédellec, C. (2019b). Text mining tools for extracting information about microbial biodiversity in food. Food Microbiology, 81, 63–75. https://doi.org/10.1016/j.fm.2018.04.011

Chakravorty, S., et al. (2007). A detailed analysis of 16S ribosomal RNA gene segments for the diagnosis of pathogenic bacteria. Journal of Microbiological Methods, 69(2), 330–339.

Clarridge, J. E. (2004). Impact of 16S rRNA gene sequence analysis for identification of bacteria on clinical microbiology and infectious diseases. Clinical Microbiology Reviews, 17(4), 840–862.https://doi.org/10.1128/CMR.17.4.840-862.2004.

Clancy, K. K., Voytek, M. A., Knight, R., & Dominguez-Bello, M. G. (2021). Microbial forensics: Source attribution using microbiomes. *Trends in Microbiology*, 29(1), 53–62.

Coelho, L. P., Kultima, J. R., Costea, P. I., Fournier, C., Pan, Y., Czarnecki-Maulden, G., Hayward, M. R., Forslund, S. K., Schmidt, T. S. B., Descombes, P., Jackson, J. R., Li, Q., & Bork, P. (2018). Similarity of the dog and human gut microbiomes in gene content and response to diet. Microbiome, 6(1), 72. https://doi.org/10.1186/s40168-018-0450-3

Collado MC, Rautava S, Aakko J, Isolauri E, Salminen S. Human gut colonisation may be initiated in utero by distinct microbial communities in the placenta and amniotic fluid. Sci Rep. 2016;6:23129.

Dahlberg J, Sun L, Waller KP, Ostensson K, Mcguire M, Agenas S, et al. Microbiota data from low biomass milk samples is markedly affected by laboratory and reagent contamination. PLoS One. 2019;14:1–17.

Dauga, C., Doré, J., & Sghir, A. (2005). Expanding the known diversity and environmental distribution of cultured and uncultured bacteria. In Medecine/Sciences (Vol. 21, Issue 3, pp. 290–296). Elsevier Masson SAS. https://doi.org/10.1051/medsci/2005213290

Deines, P., Hammerschmidt, K., & Bosch, T. C. G. (2020). Microbial Species Coexistence Depends on the Host Environment. https://doi.org/10.1128/mBio

Devesse, L., Ballard, D., Davenport, L., Riethorst, I., Mason-Buck, G., & Court, D. S. (2018). Concordance of theForenSeq™ system and characterisation of sequence-specific autosomal STR alleles across two majorpopulation groups. Forensic Science International-Genetics, 34, 57-61.https://doi.org/10.1016/j.fsigen.2017.10.

Di Tommaso, P., Chatzou, M., Floden, E. W., Barja, P. P., Palumbo, E., & Notredame, C. (2017). Nextflow enables reproducible computational workflows. Nature Biotechnology, 35(4), 316-319. Retrieved from https://www.nature.com/articles/nbt.3820

Dickson R, Erb-Downward J, Huffnagle G. Towards an ecology of the lung: new conceptual models of pulmonary microbiology and pneumonia pathogenesis. Lancet Respir Med. 2014;2:238–46.

Dickson RP, Erb-Downward JR, Freeman CM, Mccloskey L, Falkowski NR, Huffnagle GB, et al. Bacterial topography of the healthy human lower respiratory tract. MBio. 2017;8:e02287–16.

Drengenes C, Wiker HG, Kalananthan T, Nordeide E, Eagan TML, Nielsen R. Laboratory contamination in airway microbiome studies. BMC Microbiol. 2019;19:1–13.

Fang, H., Yang, Z., Steele, J., & Salama, P. (2017). BioNLP approaches for metagenomic annotations. *Journal of Computational Biology*, 24(10), 1016–1027.

Fayer, R., et al. (2000). Cryptosporidium and cryptosporidiosis. *Parasitology Today*, 16(1), 14-20.

Feinerer, I., Hornik, K., & Meyer, D. (2008a). Journal of Statistical Software Text Mining Infrastructure in R. Journal of Statistical Software, 25(5). http://www.jstatsoft.org/

Field D, Garrity G, Gray T, Morrison N, Selengut J, Sterk P, Tatusova T, Thomson N, Allen MJ, Angiuoli SV: The minimum information about a genome sequence (MIGS) specification. Nat Biotechnol. 2008, 26 (5): 541-547. 10.1038/nbt1360.

Field KG, Samadpour M. 2007. Fecal source tracking, the indicator paradigm, and managing water quality. Water Res. 41:3517–3538. 10.1016/j.watres.2007.06.056

Fierer, N., Jackson, J.A., & Lewis, N.E. (2012). Understanding variation in microbial communities across environments. Nature, 488(7410), 531-533. https://doi.org/10.1038/nature11235

Fierer, N., Lauber, C. L., Ramirez, K. S., Zaneveld, J., Bradford, M. A., & Knight, R. (2012). Comparative metagenomic, phylogenetic and physiological analyses of soil microbial communities across nitrogen gradients. *ISME Journal*, 6(5), 1007-1017. https://doi.org/10.1038/ismej.2011.159

Fierer, N. (2017). Embracing the unknown: disentangling the complexities of the soil microbiome. Nature Reviews Microbiology, 15(10), 579-590.

Fong , T. T. , and Lipp , E. K. (2005) Enteric viruses of humans and animals in aquatic environments: health risks, detection, and potential water quality assessment tools . Microbiol. Mol. Biol. Rev. 69 , 357 – 371 .

Frank EM, Ahlinder J, Jephson T, Persson KM, Lindberg E, Paul CJ. Marine sediments are identified as an environmental reservoir for Escherichia coli: comparing signature-based and novel amplicon sequencing approaches for microbial source tracking. Sci Total Environ. 2024;907:167865. doi:10.1016/j.scitotenv.2023.167865

Harwood VJ, Staley C, Badgley BD, Borges K, Korajkic A. Microbial source tracking markers for detection of fecal contamination in environmental waters: relationships between pathogens and human health outcomes. FEMS Microbiol Rev. 2014;38(1):1-40. doi:10.1111/1574-6976.12031

Harwood, V. J., et al. (2013). Microbial source tracking markers for detection of fecal contamination in environmental waters: Relationships between pathogens and human health outcomes. FEMS Microbiology Reviews, 37(2), 257-299.

Hinchliff CE, Smith SA, Allman JF, Burleigh JG, Chaudhary R, Coghill LM, Crandall KA, Deng J, Drew BT, Gazis R, Gude K, Hibbett DS, Katz LA, Laughinghouse HD, McTavish EJ, Midford PE, Owen CL, Ree RH, Rees JA, Soltis DE, Williams T, Cranston KA. Synthesis of phylogeny and taxonomy into a comprehensive tree of life. Proc Natl Acad Sci USA. 2015; 112(41):12764–9.

Hong KH, Hong SK, Cho SI, Ra E, Han KH, Kang SB, et al. Analysis of the vaginal microbiome by next-generation sequencing and evaluation of its performance as a clinical diagnostic tool in vaginitis. Ann Lab Med. 2016;36:441–9.

Human Microbiome Project Consortium. (2012). Structure, function and diversity of the healthy human microbiome. Nature, 486(7402), 207-214.

Huson DH, Beier S, Flade I, Górska A, El-Hadidi M, Mitra S, Ruscheweyh HJ, Tappu R. MEGAN Community Edition - interactive exploration and analysis of large-scale microbiome sequencing data. PLoS Comput Biol. 2016; 12(6):1004957. doi:10.1371/journal.pcbi.1004957.

Huson DH, Beier S, Flade I, Górska A, El-Hadidi M, Mitra S, Ruscheweyh HJ, Tappu R. MEGAN Community Edition - interactive exploration and analysis of large-scale microbiome sequencing data. PLoS Comput Biol. 2016; 12(6):1004957. doi:10.1371/journal.pcbi.1004957.

Janssen, P. H. (2006). Identifying the dominant soil bacterial taxa in libraries of 16S rRNA and 16S rRNA genes. *Applied and Environmental Microbiology*, 72(3), 1719-1728. https://doi.org/10.1128/AEM.72.3.1719-1728.2006

Jensen, L. J., et al. (2012). Challenges in the integration of biological data sources. Nature Reviews Genetics, 13(10), 769–780.

Johnson JS, Spakowicz DJ, Hong B-Y, Petersen LM, Demkowicz P, Chen L, et al. Evaluation of 16S rRNA gene sequencing for species and strain-level microbiome analysis. Nat Commun 2019;10. https://doi.org/10.1038/ s41467-019-13036-1.

Jupp, S., et al. (2016). The Cellular Microenvironment Ontology. Bioinformatics, 32(11), 1699–1706. DOI: 10.1093/bioinformatics/btw063

Karadeniz, I., & Özgür, A. (2015b). Detection and categorization of bacteria habitats using shallow linguistic analysis. BMC Bioinformatics, 16(10). https://doi.org/10.1186/1471-2105-16-S10-S5

Kastenbauer, T., McEwan, A.G., Lathrop, S.K., Stice, S.L., Davis, S.L., & McLean, J.A. (2019). Zoonotic microbial communities in agricultural environments: A review of microbial exposure and health outcomes. Applied and Environmental Microbiology, 85(4), e00407-19. https://doi.org/10.1128/AEM.00407-19

Keestra, A.M., Beebee, T.J., Alexander, P., Rego, M., van Elsas, J.D., & van Veen, J.A. (2019). Soil microbiomes drive microbial community structure and function in the rhizosphere of a bioenergy crop. Environmental Microbiology, 21(9), 3553-3564. https://doi.org/10.1111/1462-2920.14428

Khatib , L. A. , Tsai , Y. L. , and Olson , B. H. (2002) A biomarker for the identifi cation of cattle fecal pollution in water using the LYIIa toxin gene from the enterotoxigenic Escherichia coli . Appl. Environ. Microbiol. 59 , 97 – 104 .

Kim M, Oh HS, Park SC, Chun J. Towards a taxonomic coherence between average nucleotide identity and 16S rRNA gene sequence similarity for species demarcation of prokaryotes. Int J Syst Evol Microbiol 2014;64:346–51. https://doi.org/10.1099/ijs.0.059774-0.

Kirchman, D. L. (2016). Growth rates of microbes in the oceans. Annual Review of Marine Science, 8, 285-309.

Kitajima, M., et al. (2020). Advances in virus source tracking in water. *Water Research*, 183, 116021.

Koch, R. (1882). Die Ätiologie der Tuberkulose. Berliner Klinische Wochenschrift, 19, 221–230.

Koch, R. (1884). Die Ätiologie der Tuberkulose. Mittheilungen aus dem Kaiserlichen Gesundheitsamte, 2, 1–88.

Kornberg, A. (1957). Enzymatic synthesis of DNA. Journal of Cellular and Comparative Physiology, 50(1), 1–22.

Kozich, J. J., Westcott, S. L., Baxter, N. T., Highlander, S. K., & Schloss, P. D. (2013). Development of a dual-index sequencing strategy and curation pipeline for analyzing amplicon sequence data on the miseq illumina sequencing platform. Applied and Environmental Microbiology, 79(17), 5112–5120. https://doi.org/10.1128/AEM.01043-13

Knight, R., Hugenholtz, P., & Jansson, J.K. (2018). Advancing sequencing-based technology and bioinformatics for comprehensive profiling of microbiomes. Nature Reviews Genetics, 19(4), 209-222. https://doi.org/10.1038/nrg.2017.110

Knights, D., Kuczynski, J., Charlson, E. et al. Bayesian community-wide culture-independent microbial source tracking. Nat Methods 8, 761–763 (2011). https://doi.org/10.1038/nmeth.1650

Kleppe, K., et al. (1971). Studies on polynucleotides. Journal of Molecular Biology, 56(2), 341–361.

Kuperman AA, Zimmerman A, Hamadia S, Ziv O, Gurevich V, Fichtman B, et al. Deep microbial analysis of multiple placentas shows no evidence for a placental microbiome. BJOG. 2020;127:159–69.

Khorana, H. G., et al. (1971). Studies on polynucleotides. Journal of Molecular Biology, 56(2), 341–361.

Lauro, F. M., McDougald, D., Thomas, T., Williams, T. J., Egan, S., Rice, S., ... & Cavicchioli, R. (2009). The genomic basis of trophic strategy in marine bacteria. *Proceedings of the National Academy of Sciences*, 106(37), 15527-15533. https://doi.org/10.1073/pnas.0903507106

Le Guyader, H. (2008). la biodiversité : un concept flou ou une réalité scientifique ? http://authors.library.caltech.edu/5456/01/hrst.mit.edu/hrs/evolution/public/papers/hubbylewontin1966/hubbylewontin1966.

Lewis, K. (2007). Persister cells, dormancy, and infectious disease. Nature Reviews Microbiology, 5(1), 48–56.

Lennon, J. T., & Jones, S. E. (2011). Microbial seed banks: The ecological and evolutionary implications of dormancy. Nature Reviews Microbiology, 9(2), 119–130.

Lozupone, C. A., Stombaugh, J. I., Gordon, J. I., Jansson, J. K., & Knight, R. (2012). Diversity, stability and resilience of the human gut microbiota. *Nature*, 489(7415), 220-230. https://doi.org/10.1038/nature11550

Liu, W. T., Marsh, T. L., Cheng, H., & Forney, L. J. (1997). Characterization of microbial diversity by determining terminal restriction fragment length polymorphisms of genes encoding 16S rRNA. Applied and Environmental Microbiology, 63(11), 4516-4522.

Liu, C., Ding, Y., Wang, Q., Luan, Y., & Xu, X. (2016). Microbial diversity and function in soil ecosystems: Insights from metagenomics. Frontiers in Microbiology, 7, 1895. https://doi.org/10.3389/fmicb.2016.01895

Lu, J., Breitwieser, F. P., Thielen, P., & Salzberg, S. L. (2017). Bracken: Estimating species abundance in metagenomics data. *PeerJ Computer Science*, 3, e104

Lloyd-Price J, Abu-Ali G, Huttenhower C. The healthy human microbiome. Genome Med. 2016;8(1):51. doi:10.1186/s13073-016-0307-y.

Magoč, T., & Salzberg, S. L. (2011). FLASH: Fast Length Adjustment of Short Reads to Improve Genome Assemblies. *Bioinformatics, 27*(21), 2957-2963. https://doi.org/10.1093/bioinformatics/btr507.

Margulies, M., Egholm, M., Altman, W. E., Attiya, S., Bader, J. S., Bemben, L. A., ... & Rothberg, J. M. (2005). Genome sequencing in microfabricated high-density picolitre reactors. Nature, 437(7057), 376-380

Martiny, J.B.H., Bohannan, B.J.M., Brown, J.H., Colwell, R.K., Fuhrman, J.A., Green, J.L., Horner-Devine, M.C., Kane, M., Krumins, J.A., Kuske, C.R., Morin, P.J., Naeem, S., Øvreås, L., Reysenbach, A.-L., Smith, V.H., & Staley, J.T. (2006). Microbial biogeography: putting microorganisms on the map. *Nature Reviews Microbiology*, 4(2), 102–112.

Meselson, M., & Stahl, F. W. (1958). The replication of DNA in Escherichia coli. Proceedings of the National Academy of Sciences, 44(7), 671–682.

Metzker, M. L. (2010). Sequencing technologies—The next generation. Nature Reviews Genetics, 11(1), 31-46.

McKenzie, J.A., Miller, M.A., Martinez, K.A., Broadhurst, M.J., & Reid, G. (2016). Characteristics of dog and cat fecal microbiota in relation to lifestyle factors. PLoS ONE, 11(5), e0155377. https://doi.org/10.1371/journal.pone.0155377

Muyzer, G., de Waal, E. C., & Uitterlinden, A. G. (1993). Profiling of complex microbial populations by denaturing gradient gel electrophoresis analysis of polymerase chain

reaction-amplified genes coding for 16S rRNA. Applied and Environmental Microbiology, 59(3), 695-700.
Nam SJ, Kim DW, Lee SH, Koo OK. Assessment of Microbial Source Tracking Marker and Fecal Indicator Bacteria on Food-Contact Surfaces in School Cafeterias. J Food Prot. 2023;86(2):100035. doi:10.1016/j.jfp.2022.100035
Newton, R. J., Jones, S. E., Eiler, A., McMahon, K. D., & Bertilsson, S. (2011). A guide to the natural history of freshwater lake bacteria. Microbiology and Molecular Biology Reviews, 75(1), 14-49.
Needham CJ, Bradford JR, Bulpitt AJ, Westhead DR. A primer on learning in Bayesian networks for computational biology. PLoS Comput Biol. 2007 Aug;3(8):e129. doi: 10.1371/journal.pcbi.0030129.
Notomi, T., et al. (2000). Loop-mediated isothermal amplification of DNA. *Nucleic Acids Research*, 28(12), e63.
Nygaard AB, Tunsjø HS, Meisal R, Charnock C. A preliminary study on the potential of Nanopore MinION and Illumina MiSeq 16S rRNA gene sequencing to characterize building-dust microbiomes. Sci Rep 2020;10:3209. https:// doi.org/10.1038/s41598-020-59771-0.
Parks, D.H., Chuvochina, M., Waite, D.W., and Hugenholtz, P. (2021). A complete domain-to-species taxonomy for Bacteria and Archaea. Nature Biotechnology, 39(12), 1403–1411.
Parks, D.H., Chuvochina, M., Waite, D.W., Rinke, C., Skarshewski, A., Chaumeil, P.A., and Hugenholtz, P. (2018). A standardized bacterial taxonomy based on genome phylogeny substantially revises the tree of life. Nature Biotechnology, 36(10), 996–1004.
Parks, D.H., et al. (2021). "GTDB: an ongoing census of bacterial and archaeal diversity through a phylogenetically consistent, rank normalized and complete genome-based taxonomy." Nucleic Acids Research, https://doi.org/10.1093/nar/gkab776.
Paerl, H. W., & Paul, V. J. (2012). Climate change: Links to global expansion of harmful cyanobacteria. *Water Research*, 46(5), 1349-1363.https://doi.org/10.1016/j.watres.2011.08.002
Pelzer E, Gomez-Arango LF, Barrett HL, Nitert MD. Review: maternal health and the placental microbiome. Placenta. 2017;54:30–7.
Perez-Muñoz ME, Arrieta MC, Ramer-Tait AE, Walter J. A critical assessment of the “sterile womb” and “in utero colonization” hypotheses: implications for research on the pioneer infant microbiome. Microbiome. 2017;5:1–19.
Philippot, L., Raaijmakers, J. M., Lemanceau, P., & van der Putten, W. H. (2013). Going back to the roots: the microbial ecology of the rhizosphere. Nature Reviews Microbiology, 11(11), 789-799.
Pruesse, E., Quast, C., Knittel, K., Fuchs, B. M., Ludwig, W., Peplies, J., & Glöckner, F. O. (2007). SILVA: A comprehensive online resource for quality checked and aligned ribosomal RNA sequence data compatible with ARB. Nucleic Acids Research, 35(21), 7188–7196. https://doi.org/10.1093/nar/gkm864
Quast, C., Pruesse, E., Yilmaz, P., Gerken, J., Schweer, T., Yarza, P., Peplies, J., & Glöckner, F. O. (2013b). The SILVA ribosomal RNA gene database project: Improved data processing and web-based tools. Nucleic Acids Research, 41(D1). https://doi.org/10.1093/nar/gks1219
Quinn, R.J., Johnson, K.P., Tiedje, J.M., & Chain, P.S. (2016). Assessing the impact of metadata quality on microbiome research: Challenges and solutions. Microbiome, 4, 108. https://doi.org/10.1186/s40168-016-0224-3
Ramanan, V., Coskuner-Weber, O., Schoch, C. L., & Shankar, M. (2022). Host-microbiome monitoring using GenBank data. *Frontiers in Microbiology*, 13, 735204

Ramazzotti, D., Angaroni, F., Climente-González, H., & Smucler, E. (2019). Machine learning microbial source tracking (mlST): Improving the resolution of microbial signatures. *Microbiome*, 7, 73.

Rappé, M. S., & Giovannoni, S. J. (2003). The uncultured microbial majority. Annual Review of Microbiology, 57(1), 369–394.

Raza S, Kim J, Sadowsky MJ, Unno T. Microbial source tracking using metagenomics and other new technologies. J Microbiol. 2021;59(3):259-269. doi:10.1007/s12275-021-0668-9

Paez-Espino, D., et al. (2016). Uncovering Earth's virome. Nature, 536(7617), 425–430.

Pasteur, L. (1876). Études sur la bière. Paris: Gauthier-Villars.

Prosser, J. I. (2015). Dispersing misconceptions and identifying opportunities for the use of 'omics' in soil microbial ecology. Nature Reviews Microbiology, 13(7), 439–446.

Robert Bossy, L. D.́eger, E. C. M. B. C. N. (2019). Bacteria Biotope at BioNLP. Association for Computational Linguistics, 121–131.

Rodríguez-Pérez H, Ciuffreda L, Flores C. NanoCLUST: a species-level analysis of 16S rRNA nanopore sequencing data btaa900. Bioinformatics 2020. https:// doi.org/10.1093/bioinformatics/btaa900.

Roux, S., et al. (2015). Viral dark matter and virus-host interactions resolved from publicly available microbial genomes. eLife, 4, e08490.

Roguet A, Esen ÖC, Eren AM, Newton RJ, McLellan SL. FORENSIC: an Online Platform for Fecal Source Identification. mSystems. 2020;5(2):e00869-19. Published 2020 Mar 17. doi:10.1128/mSystems.00869-19

Salonen, A., Oo, M.M., & Kallio, M.J. (2017). Transmission of bacteria between pets and their owners: A longitudinal study. Journal of Clinical Microbiology, 55(7), 2054-2063. https://doi.org/10.1128/JCM.01683-16

Sangwan, N., Verma, H., Kumar, R., Negi, V., Lax, S., Khurana, J. P., ... & Khurana, P. (2015). Reconstructing the bacterial diversity of soil through single-molecule sequencing of DNA enriched for the 16S rRNA gene. *Applied and Environmental Microbiology*, 81(12), 4439-4448.https://doi.org/10.1128/AEM.04191-14

Scheithauer TPM, Dallinga-Thie GM, de Vos WM, Nieuwdorp M, van Raalte DH. Causality of small and large intestinal microbiota in weight regulation and insulin resistance. Mol Metab Elsevier GmbH. 2016;5:759–70.

Schoch, C. L., et al. (2012). Nuclear ribosomal internal transcribed spacer (ITS) region as a universal DNA barcode marker for Fungi. Proceedings of the National Academy of Sciences, 109(16), 6241–6246.

Schloss, P. D., Westcott, S. L., Ryabin, T., Hall, J. R., Hartmann, M., Hollister, E. B., Lesniewski, R. A., Oakley, B. B., Parks, D. H., Robinson, C. J., Sahl, J. W., Stres, B., Thallinger, G. G., Van Horn, D. J., & Weber, C. F. (2009). Introducing mothur: Open-source, platform-independent, community-supported software for describing and comparing microbial communities. *Applied and Environmental Microbiology, 75*(23), 7537-7541. https://doi.org/10.1128/AEM.01541-09

Schloss, P. D., Schubert, A. M., Zachow, J., & Iverson, K. D. (2012). Stability of the murine gut microbiota in different health states. *PLoS ONE*, 7(8), e45021. https://doi.org/10.1371/journal.pone.0045021

Scott, T. M., Rose, J. B., Jenkins, T. M., Farrah, S. R., and Lukasik, J. (2002) Microbial Source Tracking: Current Methodology and Future Directions. *Appl. Environ. Microbiol.* 68, 5796–5803. DOI: 10.1128/AEM.68.12.5796-5803.2002

Scott , T. M. , Jenkins , T. M. , Lukasik , J. , and Rose , J. B. (2005) Potential use of a host associated molecular maker in Enterococcus faecium as an index of human fecal pollution . Environ. Sci. Technol. 39 , 283 – 287 .

Segata, N., Waldron, L., Ballarini, A., Narasimhan, V., Jousson, O., & Huttenhower, C. (2012). MetaPhlAn: Metagenomic phylogenetic analysis. *Nature Methods*, 9(8), 811–814.
Sender, R., Fuchs, S., & Milo, R. (2016). Revised estimates for the number of human and bacteria cells in the body. PLoS Biology, 14(8), e1002533.
Shanks, O. C., et al. (2010). Performance assessment of real-time PCR methods for quantification of fecal indicator bacteria in environmental water samples. Applied and Environmental Microbiology, 76(4), 1342-1353.
Shenhav, L., Thompson, M., Joseph, T.A. et al. FEAST: fast expectation-maximization for microbial source tracking. Nat Methods 16, 627–632 (2019). https://doi.org/10.1038/s41592-019-0431-x
Shin J, Lee S, Go M-J, Lee SY, Kim SC, Lee C-H, et al. Analysis of the mouse gut microbiome using full-length 16S rRNA amplicon sequencing. Sci Rep 2016;6. https://doi.org/10.1038/srep29681.
Sinclair, L., Ijaz, U. Z., Jensen, L. J., Coolen, M. J. L., Gubry-Rangin, C., Chroňáková, A., Oulas, A., Pavloudi, C., Schnetzer, J., Weimann, A., Ijaz, A., Eiler, A., Quince, C., & Pafilis, E. (2016a). Seqenv: Linking sequences to environments through text mining. PeerJ, 2016(12). https://doi.org/10.7717/peerj.2690
Sinigalliano, C. D., Fleisher, J. M., Gidley, M., Solo-Gabriele, H. M., Shibata, T., Plano, L. R. W., Elmir, S. M., Wanless, D., Bartkowiak, J., Boiteau, R., & others. (2019). Linkage of sequence data to environmental conditions. *Frontiers in Microbiology*, 10, 899.
Smith, B., et al. (2007). The OBO Foundry: Coordinated evolution of ontologies to support biomedical data integration. Nature Biotechnology, 25(11), 1251–1255. DOI: 10.1038/nbt1346
Smith, S.A., and Peay, K.G. (2014). Sequence depth, not PCR replication, improves ecological inference from next generation DNA sequencing. PLoS ONE, 9(2), e90234.
Staley, J. T., & Konopka, A. (1985). Measurement of in situ activities of nonphotosynthetic microorganisms in aquatic and terrestrial habitats. Annual Review of Microbiology, 39(1), 321–346.
Steen, A. D., et al. (2019). High proportions of bacteria and archaea across most biomes remain uncultured. Nature Microbiology, 4(10), 603–607.
Stewart-Pularo, J., Daugomah, J. W., Chestnut, D. E., Graves, D. A., Sobsey, M. D., Scott, G. I. (2006) F RNA coliphage typing for microbial source tracking in surface waters. J. Appl. Microbiol. 101, 1015–1026.
Talend. (2023). Talend Data Integration. Retrieved from https://www.talend.com/products/data-integration
Tamames, J., Abellán, J.J., Pignatelli, M. et al. Environmental distribution of prokaryotic taxa. BMC Microbiol 10, 85 (2010). https://doi.org/10.1186/1471-2180-10-85
Thompson, L. R., Sanders, J. G., McDonald, D., Amir, A., Ladau, J., Locey, K. J., ... & Knight, R. (2017). A communal catalogue reveals Earth's multiscale microbial diversity. *Nature*, 551(7681), 457-463.https://doi.org/10.1038/nature24621
Tropini C, Earle KA, Huang KC, Sonnenburg JL. The Gut Microbiome: Connecting Spatial Organization to Function. Cell Host Microbe. 2017;21:433–42.
Vadde KK, Phan DC, Moghadam SV, et al. Fecal pollution source characterization in the surface waters of recharge and contributing zones of a karst aquifer using general and host-associated fecal genetic markers. Environ Sci Process Impacts. 2022;24(12):2450-2464. Published 2022 Dec 14. doi:10.1039/d2em00418f
Vanderzalm J, Currie S, Smith W, Metcalfe S, Taylor N, Ahmed W. Microbial source tracking of fecal pollution to coral reef lagoons of Norfolk Island, Australia. Sci Total Environ. Published online November 26, 2023. doi:10.1016/j.scitotenv.2023.168906

Walls, R. L., et al. (2012). Ontologies as integrative tools for plant science. American Journal of Botany, 99(8), 1263–1275. DOI: 10.3732/ajb.1200222

Wang H, Altemus J, Niazi F, Green H, Calhoun BC, Sturgis C, et al. Breast tissue, oral and urinary microbiomes in breast cancer. Oncotarget. 2017;8:88122–38.

Wang, G. D., Zhai, W., Yang, H. C., Fan, R. X., Cao, X., Zhong, L., Wang, L., Liu, F., Wu, H., Cheng, L. G., Poyarkov, A. D., Poyarkov, N. A., Tang, S. S., Zhao, W. M., Gao, Y., Lv, X. M., Irwin, D. M., Savolainen, P., Wu, C. I., & Zhang, Y. P. (2013). The genomics of selection in dogs and the parallel evolution between dogs and humans. Nature Communications, 4. https://doi.org/10.1038/ncomms2814

Watson, J. D., & Crick, F. H. C. (1953). Molecular structure of nucleic acids: A structure for deoxyribose nucleic acid. Nature, 171(4356), 737–738.

Wesolowska-Andersen, A., Bahl, M. I., Carvalho, V., Kristiansen, K., Sicheritz-Pontén, T., Gupta, R., & Licht, T. R. (2014). Choice of bacterial DNA extraction method from fecal material influences community structure as evaluated by metagenomic analysis. Microbiome, 2(1). https://doi.org/10.1186/2049-2618-2-19

Whitman WB, Coleman DC, Wiebe WJ: Prokaryotes: the unseen majority. Proc Natl Acad Sci USA. 1998, 95 (12): 6578-6583. 10.1073/pnas.95.12.6578.

Woese, C. R., & Fox, G. E. (1977). Phylogenetic structure of the prokaryotic domain: The primary kingdoms. *Proceedings of the National Academy of Sciences*, 74(11), 5088–5090.

Woo PCY, Lau SKP, Teng JLL, Tse H, Yuen K-Y. Then and now: use of 16S rDNA gene sequencing for bacterial identification and discovery of novel bacteria in clinical microbiology laboratories. Clin Microbiol Infect 2008;14:908–34. https://doi.org/10.1111/j.1469-0691.2008.02070.x.

Wood, D. E., Lu, J., & Langmead, B. (2019). Improved metagenomic analysis with Kraken 2. *Genome Biology*, 20(1), 257.

Yatera K, Noguchi S, Mukae H. The microbiome in the lower respiratory tract. Respir Investig. 2018;56:432–9.

Yarza, P., et al. (2014). Uniting the classification of cultured and uncultured bacteria and archaea using 16S rRNA gene sequences. Nature Reviews Microbiology, 12(9), 635–645.

Zeaiter, Z., Fournier, P.-E., Ogata, H., & Raoult, D. (2002). Phylogenetic classification of Bartonella species by comparing groEL sequences. In International Journal of Systematic and Evolutionary Microbiology (Vol. 52). http://ijs.sgmjournals.org

Zhong X, Zhao J, Chen Y, et al. High-Throughput Sequencing Reveals a Dynamic Bacterial Linkage between the Captive White Rhinoceros and Its Environment. Microbiol Spectr. 2023;11(4):e0092123. doi:10.1128/spectrum.00921-23.

Zinger, L., Amaral-Zettler, L. A., Fuhrman, J. A., Horner-Devine, M. C., Huse, S. M., Welch, D. B., ... & Ramette, A. (2011). Global patterns of bacterial beta-diversity in seafloor and seawater ecosystems. *PLOS ONE*, 6(9), e24570. https://doi.org/10.1371/journal.pone.0024570.

Acknowledgements

First and foremost, I thank God, “ALLAH” the Almighty, for granting me the strength and perseverance needed to complete this thesis. My faith helped me overcome obstacles, find courage in moments of doubt, and persist through challenges, especially when balancing my professional responsibilities with my studies. Despite the lack of sleep and countless weekends devoted to this project, it is thanks to this divine strength that I was able to see this journey through to the end.

I want to thank myself for believing in this project and persevering through seven years of challenges and doubt, never giving up despite the obstacles.

To my father *Mohamed Tahar JABRI*, to whom I dedicate this thesis. He has always given me unconditional love and daily support, especially during my time abroad to undertake and complete this research. Since my childhood, he has made tremendous sacrifices to raise me and my brother *Mohamed Zakaria JABRI*. I also wish to express my gratitude to my brother, who took care of our father in my absence, while also contributing to the completion of this research work. The dedication and support of my family have been the pillars of my academic journey, and for that, I am infinitely grateful.

I wish to express my deepest gratitude to Professor *Georges DAUBE* and Doctor *Bernard TAMINIAU* for allowing me to collaborate within their research laboratory at the Faculty of Veterinary Medicine of the University of Liège, specifically in the Laboratory of Food Microbiology. Their unwavering commitment played a vital role in the completion of this thesis. Professor *DAUBE* and Doctor *TAMINIAU* generously shared their knowledge, offered valuable advice, and guided me throughout the entire process. Their constant encouragement helped me improve each day. *Dr. TAMINIAU*’s repeated reminder that a thesis is a personal work that needs to be adapted and developed over time deeply resonated with me. Ultimately, I am now convinced that personal investment in this work yields results, and I am grateful for this opportunity and their continuous support.

I would like to express my deep gratitude to Prof. *Denis BAURAIN* and his team from the Phylo-Genomics Department at the Faculty of Science at the University of Liège, for their collaboration, time, energy, and invaluable assistance in the completion of this work. A heartfelt thanks as well to the members of my thesis committee, Prof. *Véronique DELCENSERIE* and

Prof. *Denis BAURAIN*, who supported me throughout this project and helped me bring it to successful completion.

To the members of the jury, I would like to express my deep gratitude for accepting to evaluate this thesis work, as well as for the time and energy you have devoted to it.

I also wish to express my heartfelt gratitude to the members of the department, my colleagues, and my friends, especially *Barbara Pirard* and her family (her mother *Danielle*, her father *Pier*, and their children). *Barbara* has become an exceptional friend, and her presence has enriched my life in remarkable ways.

Remerciements

Tout d'abord, je remercie Dieu, « ALLAH » le Tout-Puissant, de m'avoir donné la force et la persévérance nécessaires pour mener à bien ce travail de thèse. Ma foi m'a permis de surmonter les obstacles, de puiser du courage dans les moments de doute, et de persévérer face aux défis, notamment lorsque j'ai dû concilier mes responsabilités professionnelles et mes études. Malgré le manque de sommeil et les nombreux week-ends consacrés à ce projet, c'est grâce à cette force divine que j'ai pu aller jusqu'au bout de cette aventure.

Je tiens à me remercier d'avoir cru en ce projet et persévéré à travers sept années de défis et de doutes, sans jamais abandonner malgré les obstacles.

À mon père *Mohamed Tahar JABRI*, à qui je dédie ce travail de thèse. Il m'a toujours offert un amour inconditionnel et un soutien quotidien, en particulier lorsque j'étais à l'étranger pour entreprendre et finaliser cette recherche. Depuis mon enfance, il a consenti d'énormes sacrifices pour m'élever, ainsi que mon frère *Mohamed Zakaria JABRI*. Je voudrais également exprimer ma gratitude envers mon frère qui a su prendre soin de mon père en mon absence, tout en contribuant à la réalisation de ce travail de recherche. Le dévouement et le soutien de ma famille ont été les piliers de mon parcours académique, et je leur suis infiniment reconnaissant.

Je tiens à exprimer ma profonde gratitude envers le Professeur *Georges DAUBE* et le Docteur *Bernard TAMINIAU* pour avoir accepté ma collaboration au sein de leur laboratoire de recherche à la Faculté de Médecine vétérinaire de l'Université de Liège, plus spécifiquement au laboratoire de Microbiologies des Denrées alimentaires. Leur engagement inébranlable a joué un rôle essentiel dans la réalisation de ce travail de thèse. Le Professeur *DAUBE* et le Docteur *TAMINIAU* ont généreusement partagé leurs connaissances, prodigué des conseils éclairés, et m'ont guidée tout au long du processus. Leurs encouragements constants m'ont permis de m'améliorer chaque jour. Les paroles répétées du *Dr. TAMINIAU*, soulignant que la thèse représente un travail personnel à adapter et à développer avec le temps, ont profondément résonné en moi. En fin de compte, je suis désormais convaincue que l'investissement personnel dans ce travail porte des fruits, et je suis reconnaissante pour cette opportunité et leur soutien continu.

Je tiens à exprimer ma profonde gratitude au Pr. *Denis BAURAIN* et à son équipe du département de phylo-génomique de la faculté des sciences de l'Université de Liège, pour leur collaboration, leur temps, leur énergie, et leur précieuse aide dans la réalisation de ce travail. Un grand merci également aux membres de mon comité de thèse, le Pr. *Véronique DELCENSERIE* et le Pr. *Denis BAURAIN*, qui m'ont accompagnée tout au long de ce projet et m'ont aidée à le mener à bien jusqu'à son aboutissement.

Aux membres du jury, je tiens à exprimer ma profonde gratitude pour avoir accepté d'évaluer ce travail de thèse, ainsi que pour le temps et l'énergie que vous y avez consacrés.

Je souhaite exprimer ma profonde gratitude envers les membres du département, mes collègues, et mes amis, tout particulièrement *Barbara Pirard* et sa famille (sa maman *Danielle*, et son papa *Pier* et leur enfants). *Barbara* est devenue une amie exceptionnelle, et sa présence a enrichi ma vie de manière remarquable.

Annexes

Appendix 1

```
# Download the fasta file from silva database version 132
$ Silva 132_SSU_TAX_Parc.fasta
#Keep just the bacterial accession
$ grep “Bacteria;” –a1 <$ Silva 132_SSU_TAX. fasta > silva_Bacteria.fasta
#Delete the sequence and keep the taxonomic fileds with acc Number
$grep “>” silva_Bacteria.fasta | sed ‘s/>//1’ |sed –e ‘s/$/;/’ >Silva_taxon.txt
#keep the good acc number
$ cut –d .  -f1 <Silva_good.txt >acc.txt
```

Appendix 2

```
#!/bin/bash

split -l 200 -a 4 acc.txt database

for file in database ;
do
    Acclist=`tr '\n' , < ${file} | sed ' s/,$// '`

# Download xmlfiles from EBI (EMBL)
   curl "https://www.ebi.ac.uk/ena/data/view/"${Acclist}"&display=xml" > ${file}.xml

   done
```

Appendix 3

```
#!/bin/bash
#remove the spaces

read -p "fichier xml source:" xmlfile
for xmlfile in *.xml
do sed -e 's/ //g' < ${xmlfile}
done

#remove all protein and nucleotide sequences
for xmlfile in *.source.xml
do sed '/<sequence>/,/<\/sequence>/d' <${xmlfile} | sed '/<qualifier name="protein_id">/,/<\/feature>/d' >$
    {xmlfile%.xml}.source_info.xml
rm ${xmlfile}
done
```

```
#!/bin/bash
#commencer par avoir un fichier xml avec des informations liée à accession number | isolation_source | taxonomicDivision | Titre de l'article | l'auteur | First Publication
# modified : le 15 février 2018 v7

for file in *source_info.xml ;
do

    echo ${file}
    sed '/<ROOT/d ; /author>/d ; /journal>/d ; /year>/d ; /volume>/d ; /issue>/d ; /firstPage>/d ; /lastPage>/d ; /referenceLocation>/d ; /<submissionDate>/d ; /reference>/d ; /comment>/d ; /keyword>/d ; /projectAccession>/d ; /description>/d ; /lineage>/d' <${file} | sed -E '/<[/]*taxon/d ; /<[/]*reference/d ; /"MD5"|"RFAM"|"SILVA-SSU"/d' | sed 's/"/|/g' | awk '/accession=/,/feature>/' | awk -F "|" '/accession=/ END {print NR} ; /<title>/ {print "="$0} ; /\|DOI\|/ {print "=" "DOI="$4} ; /\|PUBMED\|/ {print "=" "PUBMED="$4} ; /\|host|\|isolation_source|\|note/ {getline;getline;print "=" "host="$0 }' | awk '{printf "%s%s",(sub(/^=/,"\t") ? "" : "\n"),$0} END {print ""}' | sed 's/ /_/g ; s/<entry_accession=//g ; ' | awk ' NF>=2 ' >${file%xml}out.source.xml
    #fi
done

for file in *out.source.xml ;
do
    echo ${file}
    sed -E 's/^/Accession=/; s/_taxonomicDivision/taxonomicDivision/ ; s/_moleculeType/moleculeType/ ; s/_firstPublic/firstPublic/ ;s/_sequenceLength/sequenceLength/ ' < ${file} | sed -E $'s/\t+/\t/g ; s/entryaccession=//g' | awk -F "|" '{ printf$1"\t"$2"\t"$9"\t"$10"\t"$17"\t"$18 ; for (x=25;x<=NF;x++) { printf "%s%s",(x>27?":""), $x }; printf "\n" }' | sed -E $'s/<//g ; s/>/\t/g ; s/\/title//g ; s/title/title=/g ; s/&.*;//g ; s/[\(\)]//g ' | sort | uniq >> final.source.xml
done

egrep -o "host=.*" < final.source.xml | sort | uniq -c > source.txt
```

Appendix 4

```
#!/usr/bin/env perl
# Script_name: combiner_paired_thesorus_dico_HJ.pl
# version 2 in 25th February 2019 by Hiba & Bernard
#use : perl combiner_paired_thesorus_dico_HJ.pl --paired=paired_terms_v2.txt --thesorus=thesorus_v2.txt --dico=dico.v9.txt --out=obo_data_perl.obo 2>tata


use Modern::Perl '2011';
use autodie;
use Getopt::Euclid;
use Smart::Comments;
use Tie::IxHash;

tie my %obo_data, 'Tie::IxHash';

open my $dico, '<', $ARGV{'--dico'};
my $offset;

while(my $line = <$dico>) {
        chomp $line;
        $line = lc $line;
        my $offset = ++$offset;
        #hash dernier champ du dico a (name :) =
        my @split_dico = split /;/xms, $line;
        # $last_field == $name_dico and is the KEY to the hash
        my $last_field = $split_dico[-1];
        $last_field =~ tr/_/ /;
        my @associated_synonyms = get_synonyms($last_field);
```

```
        my @paired_words = get_paired($last_field);

        ## update @associated_synonyms avec les $paired si il(s) existe(nt)
        push (@associated_synonyms, @paired_words) if @paired_words;


        $obo_data{$last_field} = {id=> $offset, name => $last_field, is_a
=> $line };
        push (@{ $obo_data{$last_field}{synonyms} }, @associated_synonyms);

        }
open my $out, '>', $ARGV{'--out'};

        while (my ($term, $details) = each %obo_data) {
                my ($id, $is_a, $name, @synonyms ) = @{$details}{ qw(id is_a
name synonyms) };

                my @last_fields = split /;/xms, $is_a;
                my     $last_field     =     join(";",     $last_fields[0],
$last_fields[1]);
                ### $last_field
                say {$out}'[Term]';
                say {$out}"id: HOST:$id";
                say {$out}"name: $name" ;
                say {$out}"def: $is_a" ;
                for my $synonym (@synonyms) {
                        push (my @list_synonym , @$synonym);
                        while (my $synonym = shift @list_synonym) {
                                say {$out}"synonym: \"$synonym\"";
                        }
                }
                say {$out}"is_a: ! $name";
                say   {$out}"relationship   part_of   !   $last_field"   if
$last_field ne q{};
                say {$out}q{};
        }

 close $out;


sub get_paired {
        my $last_field = shift;
        my @paired_words;
        open my $in, '<', $ARGV{'--paired'};

        LINE:
        while(my $line = <$in>) {
                chomp $line;
                $line = lc $line;
                my ($paired,$new_paired) = split /\t+/xms , $line;
                if ($paired eq $new_paired){
                        next LINE;
                }
                if ($last_field eq $new_paired ) {
                        push @paired_words, $paired;
                }
        }
```

```
        return @paired_words;
}

sub get_synonyms {
        my $last_field = shift;
        my @associated_synonyms;

        open my $theso, '<', $ARGV{'--thesorus'};

        LINE:
        while(my $line = <$theso>) {
                chomp $line;
                $line = lc $line;
                my ($name_synonym, @synonyms)= split /[,\t]/xms, $line;


                if ($name_synonym eq $last_field) {
                        push (@associated_synonyms, @synonyms);

                        last LINE;

                }
        }
        return @associated_synonyms;
}


=head1 USAGE

$0 --paired=<paired> --thesorus=<thesorus> --dico=<dico> --out=<out>

=head1 REQUIRED ARGUMENTS

=over

=item --thesorus=<thesorus>

 Path to a file with a thesorus file

=for Euclid:
    thesorus.type: readable

=item --paired=<paired>

 Path to a paired file

=for Euclid:
    paired.type: readable

=item --dico=<dico>

 Path to a file with a dico file

=for Euclid:
    dico.type: readable

=item --out=<out>
```

```
Path for output file

=for Euclid:
     out.type : writeable

=back

=head1 OPTIONAL ARGUMENTS

=over

=item --version

=item --usage

=item --help

=item --man

 Print the usual program information

=back

=head1 AUTHOR

Hiba JABRI, C<< <hiba.jabri at uliege.be> >>

=head1 BUGS

There are undoubtedly serious bugs lurking somewhere in this code.
Bug reports and other feedback are most welcome.

=head1 COPYRIGHT

Copyright (c) 2018, Hiba JABRI. All Rights Reserved.
This program is free software. It may be used, redistributed
and/or modified under the terms of the Perl Artistic License
(see http://www.perl.com/perl/misc/Artistic.html)
```

Appendix 5

```
#!/usr/bin/env perl
# Script_name: creation_flag file
# version 1 in 20th March 2019 by Hiba


use Modern::Perl '2011';
use autodie;
use Smart::Comments;
use List::AllUtils 'uniq';


unless (@ARGV == 2) {
die <<"EOT";
```

```
Usage: $0 <infile.txt> <outfile.txt>
This tool creat flag file from dico.V9.txt to flag_list output file
Example: $0 dico.V9.txt flag_list.txt
EOT
}

my $infile = shift;
my $outfile = shift;


my @FISH;
my @HUMAN;
my @PRIMATES;
my @CANIDS;
my @FELIDS;
my @SUIDS;
my @BOVIDS;
my @EQUIDS;
my @RODENTS;
my @BIRDS;
my @AMPHIBIANS;
my @REPTILES;
my @INSECTS;
my @CRUSTACEANS;
my @ARACHNIDS;
my @MYRIAPODS;
my @MOLLUSCS;
my @ANNELIDS;
my @MOSSES;
my @FERNS;
my @GYMNOSPERMS;
my @ANGIOSPERMS;
my @FUNGI;
my @SOIL;
my @AQUATIC;
my @AIR;
my @FOOD;

#all Regex
my $FISH_rx = qr{;(?:fish|cyclostomata|chondrichthyes|actinopterygii);};
my $HUMAN_rx = qr{;(?:human|homo;)};
my $PRIMATES_rx = qr{;primates};
my $CANIDS_rx = qr{;canidae};
my $FELIDS_rx = qr{;felidae};
my $SUIDS_rx = qr{;suidae};
my $BOVIDS_rx = qr{;bovidae};
my $EQUIDS_rx = qr{;equidae};
my $RODENTS_rx = qr{;rodentia};
my $BIRDS_rx = qr{;aves};
my $AMPHIBIANS_rx = qr{;amphibia};
my $REPTILES_rx = qr{;(?:testudines|lepidosauria)};
my $INSECTS_rx = qr{;hexapoda};
my $CRUSTACEANS_rx = qr{;pancrustacea};
my $ARACHNIDS_rx = qr{;chelicerata};
my $MYRIAPODS_rx = qr{;myriapoda};
my $MOLLUSCS_rx = qr{;mollusca};
my $ANNELIDS_rx = qr{;annelida};
```

```
my $MOSSES_rx = qr{;bryophyta};
my $FERNS_rx = qr{;polypodiopsida};
my $GYMNOSPERMS_rx = qr{;acrogymnospermae};
my $ANGIOSPERMS_rx = qr{;magnoliophyta};
my $FUNGI_rx = qr{;fungi};
my $SOIL_rx = qr{;soil};
my $AQUATIC_rx = qr{;aquatic};
my $AIR_rx = qr{;air};
my $FOOD_rx = qr{;food};


open my $dico, '<', $infile;
        LINE:
        while(my $line = <$dico>) {
                chomp $line;
                $line = lc $line;
                $line =~ tr /_/ /;

                if ($line =~ $FISH_rx) {
                        (my $tax_element) = $line =~ m/($FISH_rx.*)/xmsg;
                        my @tax_elements = split /;/xms , $tax_element;
                        push (@FISH, $tax_elements[-1]);
                }
                if ($line =~ $HUMAN_rx) {
                        (my $tax_element) = $line =~ m/($HUMAN_rx.*)/xmsi;
                        my @tax_elements = split /;/xms , $tax_element;
                        push (@HUMAN, $tax_elements[-1]);
                }
                if ($line =~ $PRIMATES_rx) {
                        (my $tax_element) = $line =~
m/($PRIMATES_rx.*)/xmsg;
                        my @tax_elements = split /;/xms , $tax_element;
                        push (@PRIMATES, $tax_elements[-1]);
                }
                if ($line =~ $CANIDS_rx) {
                        (my $tax_element) = $line =~ m/($CANIDS_rx.*)/xmsg;
                        my @tax_elements = split /;/xms , $tax_element;
                        push (@CANIDS, $tax_elements[-1]);
                }
                if ($line =~ $FELIDS_rx) {
                        (my $tax_element) = $line =~ m/($FELIDS_rx.*)/xmsg;
                        my @tax_elements = split /;/xms , $tax_element;
                        push (@FELIDS, $tax_elements[-1]);
                }
                if ($line =~ $SUIDS_rx) {
                        (my $tax_element) = $line =~ m/($SUIDS_rx.*)/xmsg;
                        my @tax_elements = split /;/xms , $tax_element;
                        push (@SUIDS, $tax_elements[-1]);
                }
                if ($line =~ $BOVIDS_rx) {
                        (my $tax_element) = $line =~ m/($BOVIDS_rx.*)/xmsg;
                        my @tax_elements = split /;/xms , $tax_element;
                        push (@BOVIDS, $tax_elements[-1]);
                }
                if ($line =~ $EQUIDS_rx) {
                        (my $tax_element) = $line =~ m/($EQUIDS_rx.*)/xmsg;
                        my @tax_elements = split /;/xms , $tax_element;
```

```
                        push (@EQUIDS, $tax_elements[-1]);
                }
                if ($line =~ $RODENTS_rx) {
                        (my      $tax_element)      =      $line      =~
m/($RODENTS_rx.*)/xmsg;
                        my @tax_elements = split /;/xms , $tax_element;
                        push (@RODENTS, $tax_elements[-1]);
                }
                if ($line =~ $BIRDS_rx) {
                        (my $tax_element) = $line =~ m/($BIRDS_rx.*)/xmsg;
                        my @tax_elements = split /;/xms , $tax_element;
                        push (@BIRDS, $tax_elements[-1]);
                }
                if ($line =~ $AMPHIBIANS_rx) {
                        (my      $tax_element)      =      $line      =~
m/($AMPHIBIANS_rx.*)/xmsg;
                        my @tax_elements = split /;/xms , $tax_element;
                        push (@AMPHIBIANS, $tax_elements[-1]);
                }
                if ($line =~ $REPTILES_rx) {
                        (my      $tax_element)      =      $line      =~
m/($REPTILES_rx.*)/xmsg;
                        my @tax_elements = split /;/xms , $tax_element;
                        push (@REPTILES, $tax_elements[-1]);
                }
                if ($line =~ $INSECTS_rx) {
                        (my      $tax_element)      =      $line      =~
m/($INSECTS_rx.*)/xmsg;
                        my @tax_elements = split /;/xms , $tax_element;
                        push (@INSECTS, $tax_elements[-1]);
                }
                if ($line =~ $CRUSTACEANS_rx) {
                        (my      $tax_element)      =      $line      =~
m/($CRUSTACEANS_rx.*)/xmsg;
                        my @tax_elements = split /;/xms , $tax_element;
                        push (@CRUSTACEANS, $tax_elements[-1]);
                }
                if ($line =~ $ARACHNIDS_rx) {
                        (my      $tax_element)      =      $line      =~
m/($ARACHNIDS_rx.*)/xmsg;
                        my @tax_elements = split /;/xms , $tax_element;
                        push (@ARACHNIDS, $tax_elements[-1]);
                }
                if ($line =~ $MYRIAPODS_rx) {
                        (my      $tax_element)      =      $line      =~
m/($MYRIAPODS_rx.*)/xmsg;
                        my @tax_elements = split /;/xms , $tax_element;
                        push (@MYRIAPODS, $tax_elements[-1]);
                }
                if ($line =~ $MOLLUSCS_rx) {
                        (my      $tax_element)      =      $line      =~
m/($MOLLUSCS_rx.*)/xmsg;
                        my @tax_elements = split /;/xms , $tax_element;
                        push (@MOLLUSCS, $tax_elements[-1]);
                }
                if ($line =~ $ANNELIDS_rx) {
```

```
                        (my       $tax_element)       =       $line       =~
m/($ANNELIDS_rx.*)/xmsg;
                        my @tax_elements = split /;/xms , $tax_element;
                        push (@ANNELIDS, $tax_elements[-1]);
                }
                if ($line =~ $MOSSES_rx) {
                        (my $tax_element) = $line =~ m/($MOSSES_rx.*)/xmsg;
                        my @tax_elements = split /;/xms , $tax_element;
                        push (@MOSSES, $tax_elements[-1]);
                }
                 if ($line =~ $FERNS_rx) {
                        (my $tax_element) = $line =~ m/($FERNS_rx.*)/xmsg;
                        my @tax_elements = split /;/xms , $tax_element;
                        push (@FERNS, $tax_elements[-1]);
                }
                if ($line =~ $GYMNOSPERMS_rx) {
                        (my       $tax_element)       =       $line       =~
m/($GYMNOSPERMS_rx.*)/xmsg;
                        my @tax_elements = split /;/xms , $tax_element;
                        push (@GYMNOSPERMS, $tax_elements[-1]);
                }
                if ($line =~ $ANGIOSPERMS_rx) {
                        (my       $tax_element)       =       $line       =~
m/($ANGIOSPERMS_rx.*)/xmsg;
                        my @tax_elements = split /;/xms , $tax_element;
                        push (@ANGIOSPERMS, $tax_elements[-1]);
                }
                if ($line =~ $FUNGI_rx) {
                        (my $tax_element) = $line =~ m/($FUNGI_rx.*)/xmsg;
                        my @tax_elements = split /;/xms , $tax_element;
                        push (@FUNGI, $tax_elements[-1]);
                }
                if ($line =~ $SOIL_rx) {
                        (my $tax_element) = $line =~ m/($SOIL_rx.*)/xmsg;
                        my @tax_elements = split /;/xms , $tax_element;
                        push (@SOIL, $tax_elements[-1]);
                }
                if ($line =~ $AQUATIC_rx) {
                        (my       $tax_element)       =       $line       =~
m/($AQUATIC_rx.*)/xmsg;
                        my @tax_elements = split /;/xms , $tax_element;
                        push (@AQUATIC, $tax_elements[-1]);
                }
                if ($line =~ $AIR_rx) {
                        (my $tax_element) = $line =~ m/($AIR_rx.*)/xmsg;
                        my @tax_elements = split /;/xms , $tax_element;
                        push (@AIR, $tax_elements[-1]);
                }
                if ($line =~ $FOOD_rx) {
                        (my $tax_element) = $line =~ m/($FOOD_rx.*)/xmsg;
                        my @tax_elements = split /;/xms , $tax_element;
                        push (@FOOD, $tax_elements[-1]);
                }


        }
close $dico;
```

```
open my $out, '>' , $outfile;

                my $list =join ",", @FISH;
                say {$out} 'FISH' . "\t" . "$list" ;
                my $list1 =join ",", @HUMAN;
                say {$out} 'HUMAN' . "\t" . "$list1" ;
                my $list26 =join ",", @PRIMATES;
                say {$out} 'PRIMATES' . "\t" . "$list26" ;
                my $list2 =join ",", @CANIDS;
                say {$out} 'CANIDS' . "\t" . "$list2" ;
                my $list3 =join ",", @FELIDS;
                say {$out} 'FELIDS' . "\t" . "$list3" ;
                my $list4 =join ",", @SUIDS;
                say {$out} 'SUIDS' . "\t" . "$list4" ;
                my $list5 =join ",", @BOVIDS;
                say {$out} 'BOVIDS' . "\t" . "$list5" ;
                my $list6 =join ",", @EQUIDS;
                say {$out} 'EQUIDS' . "\t" . "$list6" ;
                my $list7 =join ",", @RODENTS;
                say {$out} 'RODENTS' . "\t" . "$list7" ;
                my $list8 =join ",", @BIRDS;
                say {$out} 'BIRDS' . "\t" . "$list8" ;
                my $list9 =join ",", @AMPHIBIANS;
                say {$out} 'AMPHIBIANS' . "\t" . "$list9" ;
                my $list10 =join ",", @REPTILES;
                say {$out} 'REPTILES' . "\t" . "$list10" ;
                my $list11 =join ",", @INSECTS;
                say {$out} 'INSECTS' . "\t" . "$list11" ;
                my $list12 =join ",", @CRUSTACEANS;
                say {$out} 'CRUSTACEANS' . "\t" . "$list12" ;
                my $list13 =join ",", @ARACHNIDS;
                say {$out} 'ARACHNIDS' . "\t" . "$list13" ;
                my $list14 =join ",", @MYRIAPODS;
                say {$out} 'MYRIAPODS' . "\t" . "$list14" ;
                my $list15 =join ",", @MOLLUSCS;
                say {$out} 'MOLLUSCS' . "\t" . "$list15" ;
                my $list16 =join ",", @ANNELIDS;
                say {$out} 'ANNELIDS' . "\t" . "$list16" ;
                my $list17 =join ",", @MOSSES;
                say {$out} 'MOSSES' . "\t" . "$list17" ;
                my $list18 =join ",", @FERNS;
                say {$out} 'FERNS' . "\t" . "$list18" ;
                my $list19 =join ",", @GYMNOSPERMS;
                say {$out} 'GYMNOSPERMS' . "\t" . "$list19" ;
                my $list20 =join ",", @ANGIOSPERMS;
                say {$out} 'ANGIOSPERMS' . "\t" . "$list20" ;
                my $list21 =join ",", @FUNGI;
                say {$out} 'FUNGI' . "\t" . "$list21" ;
                my $list22 =join ",", @SOIL;
                say {$out} 'SOIL' . "\t" . "$list22" ;
                my $list23 =join ",", @AQUATIC;
                say {$out} 'AQUATIC' . "\t" . "$list23" ;
                my $list24 =join ",", @AIR;
                say {$out} 'AIR' . "\t" . "$list24" ;
                my $list25 =join ",", @FOOD;
```

```
                say {$out} 'FOOD' . "\t" . "$list25" ;

close $out;
```

Appendix 6

```
#!/usr/bin/env perl
# Script_name: BD_treat_dico.pl version 2
# version 1 in 8th january 2019 by hiba&bernard
# version 2 in 21th february 2020 by Hiba
# ABSTRACT: Script for renaming a list using a Dico_treat_term.txt file.
#use    :    perl    BD_Treat_dico.pl    --dico=BiotopeBac_Dico.obo    --
list=acc_host.csv 2>log

use Modern::Perl '2011';
use autodie;
use Getopt::Euclid;
use Smart::Comments;
use List::AllUtils 'uniq';
use File::Basename;
use Path::Class 'file';
use Storable;

my %def_for;

open my $dico, '<', $ARGV{'--dico'};
        my @names;
while(my $line = <$dico>) {
    chomp $line;
    if ($line =~ /^name:/) {
        my (undef, $name) = split /: /, $line;
        push @names, lc($name);
        ## name: $name
        ## names: @names

    }
    if ($line =~ /^synonym:/) {
        my (undef, $name) = split /\"/, $line;
        push @names, lc($name);
                ## synonym: $name

    }
    if ($line =~ /^flag:/) {
        my (undef, $flag) = split /! /, $line;
        ## traduction: $flag
        foreach my $new_name (@names) {
            $def_for{$new_name} = $flag;
            }
    ## @names
    ## %def_for
        @names = ();
    }

}
## %def_for
```

```
open my $host, '<', $ARGV{'--list'};

my %ontology_for;

while(my $line = <$host>) {
        chomp $line;
        my ($Acc,@line_content) = split /\s+/xms, $line;
        @line_content = map lc, @line_content ;

        my @definitions;

DEFINITION:
        while ( my $term = shift @line_content ) {

                ## $term
                ## $line_content[0]:$line_content[0]
                ## $line_content[1]:$line_content[1]


                unless (@line_content) {
                ## ne fait que si line content est vide
                                push   @definitions,   $def_for{$term}   if
$def_for{$term} ;
                        last DEFINITION ;
                }

                if (scalar @line_content > 1) {

                        my  $definition  =  $def_for{"$term  $line_content[0]
$line_content[1]"} ?

        $def_for{"$term $line_content[0] $line_content[1]"} :

        $def_for{"$term $line_content[0]"} ?

        $def_for{"$term $line_content[0]"} :

        $def_for{$term} ?

        $def_for{$term} :
                                                        undef ;

                        my      $splice_offset      =      $def_for{"$term
$line_content[0] $line_content[1]"} ?
                                                        2 :

        $def_for{"$term $line_content[0]"} ?
                                                        1 :
                                                        0 ;


                        push @definitions, $definition if $definition;

                        splice (@line_content, 0 , $splice_offset);

                        ## @line_content
```

```
                        next DEFINITION ;
                }


                my $definition =        $def_for{"$term   $line_content[0]"}
?
                                                $def_for{"$term
$line_content[0]"} :
                                                $def_for{$term} ?
                                                $def_for{$term} :
                                                undef ;

                push @definitions, $definition if $definition;

                shift @line_content if $def_for{"$term $line_content[0]"};
        }

push @{$ontology_for{$Acc}} , uniq(@definitions);


}

my ($basename, $dir, $suffix) = fileparse($ARGV{'--list'}, qr{\.[^.]*}xms);
my $outfile = file($dir, ($basename . '.out'));

#treat
open my $out, '>', $outfile;

print {$out} map { "$_\t" . (join ',' ,"'@{$ontology_for{$_}}'") . "\n" }
keys %ontology_for ;

close $out;


__END__

=head1 USAGE

$0 --dico=<dico> --list=<list>

=head1 REQUIRED ARGUMENTS

=over

=item --list=<list>

 Path to a file with a list to rename.

=for Euclid:
    list.type: readable

=item --dico=<dico>

 Path to a dico file

=for Euclid:
```

```
    dico.type: readable

=back

=head1 OPTIONAL ARGUMENTS

=over

=item --version

=item --usage

=item --help

=item --man

 Print the usual program information

=back

=head1 AUTHOR

Hiba JABRI, C<< <hiba.jabri at uliege.be> >>

=head1 BUGS

There are undoubtedly serious bugs lurking somewhere in this code.
Bug reports and other feedback are most welcome.

=head1 COPYRIGHT

Copyright (c) 2018, Hiba JABRI. All Rights Reserved.
This program is free software. It may be used, redistributed
and/or modified under the terms of the Perl Artistic License
(see http://www.perl.com/perl/misc/Artistic.html)
```

Appendix 7

```
-- Creator:       MySQL Workbench 6.3.10/ExportSQLite Plugin 0.1.0
-- Author:        Hiba Jabri
-- Project:       Name of the project
-- Changed:       2020-01-24 11:34
-- Created:       2019-08-06 11:11
PRAGMA foreign_keys = OFF;

-- Schema: BiotopeBac
ATTACH "BiotopeBac.sdb" AS "BiotopeBac";
BEGIN;
CREATE TABLE "BiotopeBac"."Description"(
  "iddescreption" CHAR(20) PRIMARY KEY NOT NULL,
  "pubmed" CHAR(250) DEFAULT NULL,
  "doi" CHAR(200) DEFAULT NULL,
  "description" VARCHAR(500) DEFAULT NULL,
  "Host" VARCHAR(100)
);
```

```
CREATE TABLE "BiotopeBac"."Taxons"(
  "idtaxon" CHAR(20) PRIMARY KEY NOT NULL,
  "kingdom" VARCHAR(100),
  "phylum" VARCHAR(100),
  "class" VARCHAR(100),
  "order" VARCHAR(100),
  "family" VARCHAR(100),
  "genus" VARCHAR(100),
  "species" CHAR(250) DEFAULT NULL,
  "lineage" VARCHAR(500) DEFAULT NULL
);
CREATE TABLE "BiotopeBac"."Sequences"(
  "idsequence" CHAR(50) PRIMARY KEY NOT NULL,
  "iddescreption" CHAR(20) DEFAULT NULL,
  "idtaxon" CHAR(20) DEFAULT NULL,
  "seq" VARCHAR(2000) DEFAULT NULL,
  CONSTRAINT "fk_SEQ_Source1"
    FOREIGN KEY("iddescreption")
    REFERENCES "Description"("iddescreption"),
  CONSTRAINT "fk_SEQ_Taxon1"
    FOREIGN KEY("idtaxon")
    REFERENCES "Taxons"("idtaxon")
);
CREATE INDEX "BiotopeBac"."Sequences.fk_SEQ_Source1_idx" ON "Sequences"
("iddescreption");
CREATE INDEX "BiotopeBac"."Sequences.fk_SEQ_Taxon1_idx" ON "Sequences"
("idtaxon");
CREATE TABLE "BiotopeBac"."Wordsets"(
  "idwordset" CHAR(100) PRIMARY KEY NOT NULL,
  "definition" VARCHAR(500) DEFAULT NULL
);
CREATE TABLE "BiotopeBac"."Keywords"(
  "idkeyword" CHAR(100) PRIMARY KEY NOT NULL,
  "idwordset" CHAR(100) DEFAULT NULL,
  CONSTRAINT "fk_Key_Word_SET1"
    FOREIGN KEY("idwordset")
    REFERENCES "Wordsets"("idwordset")
);
CREATE INDEX "BiotopeBac"."Keywords.fk_Key_Word_SET1_idx" ON "Keywords"
("idwordset");
CREATE TABLE "BiotopeBac"."Sources"(
  "Sequences_idsequence" CHAR(50) NOT NULL,
  "Keywords_idkeyword" CHAR(100) NOT NULL,
  PRIMARY KEY("Sequences_idsequence","Keywords_idkeyword"),
  CONSTRAINT "fk_Sequences_has_Keywords_Sequences1"
    FOREIGN KEY("Sequences_idsequence")
    REFERENCES "Sequences"("idsequence"),
  CONSTRAINT "fk_Sequences_has_Keywords_Keywords1"
    FOREIGN KEY("Keywords_idkeyword")
    REFERENCES "Keywords"("idkeyword")
);
CREATE                                                    INDEX
"BiotopeBac"."Sources.fk_Sequences_has_Keywords_Keywords1_idx"      ON
"Sources" ("Keywords_idkeyword");
CREATE                                                    INDEX
"BiotopeBac"."Sources.fk_Sequences_has_Keywords_Sequences1_idx"     ON
"Sources" ("Sequences_idsequence");
```

```
COMMIT;
```

Appendix 8:

```
#!/usr/local/bin/R

library("tm")
library("SnowballC")
library("wordcloud")
library("RColorBrewer")
filePath <- "title_host_perl.txt"
text <- readLines(filePath)
docs <- Corpus(VectorSource(text))
inspect(docs)
dtm <- TermDocumentMatrix(docs)
m <- as.matrix(dtm)
v <- sort(rowSums(m),decreasing=TRUE)
d <- data.frame(word = names(v),freq=v)
head(d, 20)
set.seed(1234)
wordcloud(words = d$word, freq = d$freq, min.freq = 1,max.words=200, random.order=FALSE,rot.per=0.35,colors=brewer.pal(8,
 "Dark2"))
findFreqTerms(dtm, lowfreq = 4)
findAssocs(dtm, terms = "human", corlimit = 0.3)
head(d, 10)
barplot(d[1:10,]$freq, las = 2, names.arg = d[1:10,]$word,col ="lightblue", main ="Most frequent words",ylab = "Word
 frequencies")
```

Appendix 9:

```
#!/usr/bin/env perl

use Modern::Perl '2011';
use autodie;

use Getopt::Euclid qw(:vars);
use Smart::Comments '###';

use Const::Fast;
use List::AllUtils qw(all count_by);
use Storable qw(lock_store lock_retrieve);

use Bio::MUST::Core::Utils qw(:filenames);
use aliased 'Bio::FastParsers::CdHit';

const my $STORABLE => '.storable';
const my $EPSILON  => 1e-10;             # needed for large $M


### Processing TSV seq-tags file: $ARGV_seq_tags
my $tags_for;
my $tags_for_store = append_suffix($ARGV_seq_tags, $STORABLE);

if (-e $tags_for_store) {
    ### Loading from cache file: $tags_for_store
    $tags_for = lock_retrieve $tags_for_store;
}
else {
    open my $tag_in, '<', $ARGV_seq_tags;
    while (my $line = <$tag_in>) {      ### Processing seqs |===[%]
        chomp $line;
        my ($acc, $tag_str) = split "\t", $line;
        my @tags = split ',', $tag_str;
        $tags_for->{$acc} = \@tags;
    }
```

```
    ### Storing to cache file: $tags_for_store
    lock_store $tags_for, $tags_for_store;
}
##### $tags_for


### Processing CD-HIT cluster file: $ARGV_clusters
my $report;
my $report_store = append_suffix($ARGV_clusters, $STORABLE);

if (-e $report_store) {
    ### Loading from cache file: $report_store
    $report = lock_retrieve $report_store;
}
else {
    $report = CdHit->new( file => $ARGV_clusters );
    ### Storing to cache file: $report_store
    lock_store $report, $report_store;
}
my @representatives = $report->all_representatives;

### Collecting observations (bij) for representative seqs
my @details;
my @B;
my %S;

### $ARGV_binarize

my $clus_id = 0;
my $seq_n = 0;
my $space = 0;

for my $repr_id (@representatives) {    ### Collecting |===[%]
    #### $clus_id
    #### $repr_id
    my $members = $report->members_for($repr_id);
    my $clus_sz = @$members + 1;
    #### $members
    #### $clus_sz
    $seq_n += $clus_sz;
    my @tags = map { @{ $tags_for->{$_} } } $repr_id, @$members;
    #### @tags
    my $tag_n = @tags;
    $space += $tag_n;
    push @details, [ $clus_id++, $clus_sz, $repr_id, $tag_n ];
    my %count_for = count_by { $_ } @tags;
    if ($ARGV_binarize) {
        %count_for = map { $_ => 1 } keys %count_for;
    }
    #### %count_for
    push @B, \%count_for;
    $S{$_} += $count_for{$_} for keys %count_for;
    #### %S
}
#### @B

my $M = @B;
```

```
my @utags = sort keys %S;
my $N = @utags;

### $seq_n
### $space

### $M
### $N

### %S


### Computing and storing cond probs (log Pij) to TSV outfile: $ARGV_outfile
my @P;

my $theta = 1 - $ARGV_error_rate;
### $ARGV_error_rate
### $theta

open my $out, '>', $ARGV_outfile;
say {$out} join "\t", qw(clus_id clus_sz repr_id tag_n),             @utags;
say {$out} join "\t",     $M,      $seq_n, 'S',     $space, map { $S{$_} }
@utags;

my @sums;

for (my $i = 0; $i < @B; $i++) {         ### Computing |===[%]
    my @Pi = map {
        ( (1 - $theta) / $M ) + ( $theta * ($B[$i]{$_} // 0) / $S{$_} )
    } @utags;
    $sums[$_] += $Pi[$_] for 0..$#Pi;
    @Pi = map { log $_ } @Pi;
    say {$out} join "\t", @{ $details[$i] }, @Pi;
    push @P, \@Pi;
}
#### @sums
### assert: @P == $M
### assert: @sums == $N
### assert: all { abs(1 - $_ < $EPSILON) } @sums

my $database = {
    theta    => $theta,
    binarize => $ARGV_binarize,
    seq_n    => $seq_n,
    space    => $space,
    utags    => \@utags,
    details  => \@details,
    M  => $M,
    N  => $N,
    S  => \%S,
    P  => \@P,
};

my $database_store
    = insert_suffix( change_suffix($ARGV_outfile, $STORABLE), '-database'
);
### Storing to cache file: $database_store
```

```
lock_store $database, $database_store;


__END__

=head1 USAGE

    build_seq_tag_probs.pl --seq-tags=<file> --clusters=<file> \
        --outfile=<file> [optional arguments]

=head1 REQUIRED ARGUMENTS

=over

=item --seq-tags=<file>

Path to input file with sequence => tag(s) pairs in TSV format.

=for Euclid:
    file.type: readable

=item --clusters=<file>

Path to input file with cluster => sequences(s) pairs in CD-HIT format.

=for Euclid:
    file.type: readable

=item --outfile=<file>

Path to TSV outfile (and database) with conditional probs for sequences
given
tags.

=for Euclid:
    file.type: writable

=back

=head1 OPTIONAL ARGUMENTS

=over

=item --error-rate=<n>

Error rate (0 to 1) to be applied to observations when computing conditional
probs [default: n.default]. Confidence (theta) will be derived as 1 -
error-rate.

=for Euclid:
    n.type:    +number
    n.default: 1e-9

=item --binarize

Build a purely binary database of conditional probs (still taking into
account
```

```
--theta) [default: no]. Otherwise, tags attached to multiple sequences of
a
cluster are given more weight.

=item --version

=item --usage

=item --help

=item --man

Print the usual program information

=back
```

Appendix 10:

```
#!/usr/bin/env perl

use Modern::Perl '2011';
use autodie;

use Getopt::Euclid qw(:vars);
use Smart::Comments '###';

use Const::Fast;
use File::Basename;
use List::AllUtils qw(sum max mesh bundle_by partition_by);
use POSIX qw(log1p);
use Storable qw(lock_store lock_retrieve);

# should be fine considering small enough $N
const my $EPSILON  => 1e-12;


### Loading database from cache file: $ARGV_database
my $database = lock_retrieve $ARGV_database;
my ($theta, $binarize, $seq_n, $space, $utags, $M, $N, $S, $P, $details)
    = @{$database}{ qw(theta binarize seq_n space utags M N S P details)
};

### database properties
### $seq_n
### $space
### $binarize
### $theta
### $M
### $N
### $S

### configuration
### $ARGV_split_tags
### $ARGV_prior_type
### $ARGV_max_read_n
```

```
### $ARGV_max_vote_n

# (optionally) handle multiple ontologies
my %tags_for                    # by ontology (or for all)
    = $ARGV_split_tags
    ? partition_by { (split '::')[0] } @$utags
    :                (     'all'     => $utags )
;
#### %tags_for
my @onts = keys %tags_for;
### @onts
my $ont_n = @onts;
### $ont_n


### Computing tag priors (log Pe)

my %num_for = map {                 # by tag
       $_ => $ARGV_prior_type eq 'uniform'
           ? 1                          # 1 tag  (equal weight)
           : $S->{$_}                   # N obs for one tag
} @$utags;
#### %num_for

my %sums_for = bundle_by {          # by ontology (or for all)
    $_[0] => $ARGV_prior_type eq 'uniform'
           ? scalar @{$_[1]}            # N tags (equal weight)
           : sum @{$S}{ @{$_[1]} }      # N obs for tags
} 2, %tags_for;
#### %sums_for

my %den_for  = bundle_by {          # by tag
    map { $_ => $sums_for{$_[0]} } @{$_[1]}
} 2, %tags_for;
#### %den_for

my %Pe = map { $_ => $num_for{$_} / $den_for{$_} } @$utags;      # by tag
### assert: abs($ont_n - sum values %Pe) < $EPSILON
   %Pe = map { $_ =>  log $Pe{$_}               } @$utags;     # log scale
### %Pe


# build indices
# because Pij are ArrayRef[ArrayRef]
my $i = 0;
my %ridx_for = map { $_->[2] => $i++ } @$details;
### assert: $i == $M
my $j = 0;
my %tidx_for = map { $_      => $j++ } @$utags;
### assert: $j == $N

#### %ridx_for
#### %tidx_for


my %posts_for;
```

```
for my $infile (@ARGV_infiles) {
    ### Processing read mapping file: $infile

    my %matches_for;
    my $read_n = 0;
    my $vote_n = 0;
    my $curr_read = q{};

    open my $in, '<', $infile;

    LINE:
    while (my $line = <$in>) {
        # skip header
        next LINE if $line =~ m/^@/xms;

        chomp $line;
        my ($read, undef, $repr, $sam_fields) = split "\t", $line, 4;

        # track unique read count...
        # ... and stop when enough reads
        if ($read ne $curr_read) {
            last LINE if defined $ARGV_max_read_n
                    && $read_n >= $ARGV_max_read_n;
            $read_n++;
            $curr_read = $read;
         }

        # skip unexact matches (for now)
        next LINE unless $sam_fields =~ m/\b MD:Z: \d+ \b/xms;

        # fetch representative seq for aligned read
        $vote_n++;
        push @{ $matches_for{$read} }, $repr;

        # stop when enough votes
        last LINE if defined $ARGV_max_vote_n
                && $vote_n >= $ARGV_max_vote_n;
    }

    #### %matches_for

    my %product_for = %Pe;          # by tag
    my $alig_n = 0;

    while (my ($read, $matches) = each %matches_for) {
        for my $repr_id (@$matches) {    # in case of multiple exact matches
            my $ridx = $ridx_for{$repr_id};
            $product_for{$_} += $P->[ $ridx ][ $tidx_for{$_} ] for @$utags;
        }   # sum of log P (joint P for sample seqs)
        $alig_n++;
    }
    #### %product_for

    ### $read_n
    ### $alig_n
    ### $vote_n
```

```
    # compute marginal likelihood (in log scale)
    # (optionally) handle multiple ontologies
    my %sums_for = bundle_by {      # by ontology
        $_[0] => lse_all( @product_for{ @{$_[1]} } )
    } 2, %tags_for;
    #### %sums_for
    my %sum_for  = bundle_by {      # by tag
        map { $_ => $sums_for{$_[0]} } @{$_[1]}
    } 2, %tags_for;
    #### %sum_for

    # compute log posteriors using Bayes' theorem
    my %post_for = bundle_by {      # by tag
        $_[0] => $_[1] - $sum_for{$_[0]}
    } 2, %product_for;
    #### %post_for

    # store log posteriors for sample
    my ($sample) = fileparse( $infile, qr/\.sam$/xms );
    $posts_for{$sample} = {
        read_n => $read_n,
        alig_n => $alig_n,
        vote_n => $vote_n,
        map { $_ => $post_for{$_} // 'NA' } @$utags     # no NA if priors
    };

}


### Storing posterior probs to TSV outfile: $ARGV_outfile

my @samples = sort keys %posts_for;
my @columns = ( qw(read_n alig_n vote_n), @$utags );

open my $out, '>', $ARGV_outfile;
say {$out} join "\t", ('sample', @columns);

for my $sample (@samples) {
    say {$out} join "\t", $sample,
        map { $posts_for{$sample}{$_} } @columns;
}


#    from:    https://stats.stackexchange.com/questions/379335/adding-very-
small-probabilities-how-to-compute
sub lse {
    my $l1 = shift;
    my $l2 = shift;
    return max($l1, $l2) + log1p(exp(-abs($l1-$l2)));
}

sub lse_all {
    my $l1 = shift;
    while (my $l2 = shift) {
        $l1 = lse($l1, $l2);
    }
    return $l1;
```

```
}

__END__

=head1 USAGE

    comp_tag_probs_in_samples.pl    <infiles>    --database=<file>    --
outfile=<file> \
         [optional arguments]

=head1 REQUIRED ARGUMENTS

=over

=item <infiles>

Path to input SAM files [repeatable argument].

=for Euclid:
    infiles.type: readable
    repeatable

=item --database=<file>

Path to input database (cache) file.

=for Euclid:
    file.type: readable

=item --outfile=<file>

Path to TSV outfile with posterior probs for tags given sample seqs.

=for Euclid:
    file.type: writable

=back

=head1 OPTIONAL ARGUMENTS

=over

=item --prior-type=<str>

Type of tag priors to use [default: str.default]. The following types are
available:

    - database (priors computed from tag distribution in database)
    - uniform  (flat priors computed as 1/N)

=for Euclid:
    str.type:    /database|uniform/
    str.type.error: <str> must be one of database or uniform (not str)
    str.default: 'uniform'

=item --split-tags
```

```
Consider tags as belonging to distinct ontologies [default: no].
Ontology-aware tags are encoded as C<Ont::Tag>. Enabling this option also
affects the way tag priors are computed.

=item --max-read-n=<n>

Maximum number of input reads (aligned or not) to process [default:
n.default].

=for Euclid:
    n.type: +number

=item --max-vote-n=<n>

Maximum number of votes (= read mappings) to process [default: n.default].
This number can grow faster than the number of aligned reads when multiple
mappings per read are allowed in the SAM file(s).

=for Euclid:
    n.type: +number

=item --version

=item --usage

=item --help

=item --man

Print the usual program information

=back
```

Appendix 11:

Determining the environment of a DNA sample from a database (March 4, 2020)

Let suppose a list of *M* different DNA sequences $s_1 \ldots s_M$ and a list of *N* different environments $e_1 \ldots e_N$ be considered. An array **B** can link a sequence to an environment if the sequence was found in it, written as

$$\mathbf{B} = \begin{bmatrix} b_{11} & \cdots & b_{1N} \\ \vdots & \ddots & \vdots \\ b_{M1} & \cdots & b_{MN} \end{bmatrix} \quad (1)$$

with $b_{ij} = 0$ (resp. 1) if the sequence *i* was not found (resp. was found) in the environment *j*. One of the rows of B correspond to one DNA sequence and one of its columns to one environment. Let S_j be the number of sequences found in the environment e_j , i.e.

$$S_j = \sum_{k=1}^{M} b_{kj}. \quad (2)$$

Let an array **P** be written as

$$\mathbf{P} = \begin{bmatrix} p_{11} & \cdots & p_{1N} \\ \vdots & \ddots & \vdots \\ p_{M1} & \cdots & p_{MN} \end{bmatrix} \quad (3)$$

With $p_{ij} = P(s_i|e_j)$, the probability that a sequence s_i is found in a sample taken from the environment e_j. The column $[p_{1j} \ldots p_{Mj}]^T$ therefore defines a probability distribution function. In other words,

$$\sum_{k=1}^{M} p_{kj} = \sum_{k=1}^{M} P(s_k|e_j) = 1. \quad (4)$$

If $\theta = 1$, there is an absolute confidence in **B** and the hypothesis is made that finding any sequence that has been found in environment e_j is equiprobable and finding any sequence that has not been found in e_j has null probability:

$$p_{ij} = \frac{b_{ij}}{S_i}, \quad (5)$$

If $\theta = 0$, there is absolutely no confidence in **B** and finding any sequence in the environment is considered equiprobable:

$$p_{ij} = \frac{1}{M}. \quad (6)$$

The conditional probability P (e | s) was calculated from the Bayes theorem using the following statement

$$p_{ij} = \frac{1-\theta}{M} + \theta\frac{b_{ij}}{S_i}. \quad (7)$$

Let a new sample be considered. In it, the sequence s_1^* is detected. One can compute P (e_j | s_1^*), the probability that the sample was taken in the environment e_j given that the sequence s_1^* has been detected. Bayes theorem states that

$$P(e_j|s_1^*) = \frac{P(s_1^*|e_j)P(e_j)}{P(s_1^*)}. \quad (8)$$

P (s_1*|e_j) can be found in the array P as the element p_1j. One can express the total probability of finding sequence s_1^* in any environment

$$P(s_1^*) = \sum_{k=1}^{N} P(s_1^*|e_k)P(e_k). \quad (9)$$

Combining Eqs 8 and 9 gives

$$P(e_j|s_1^*) = \frac{P(s_1^*|e_j)P(e_j)}{\sum_{k=1}^{N} P(s_1^*|e_k)P(e_k)}. \quad (10)$$

If nothing is known about the environment in which the new sample has been taken, the hypothesis is made that the probability of having taken the sample in any environment is constant.

$$P(e_j) = \frac{1}{N}. \tag{11}$$

Finally,

$$P(e_j|s_1^*) = \frac{P(s_1^*|e_j)}{\sum_{k=1}^{N} P(s_1^*|e_k)}. \tag{12}$$

Finally, the probability of finding every sample independently from the environment is

$$P(e_j|s_1^* \cap \cdots \cap s_m^*) = \frac{\prod_{l=1}^{m} P(s_l^*|e_j)}{\sum_{k=1}^{N} \prod_{l=1}^{m} P(s_l^*|e_k)}. \tag{13}$$

Appendix 12:

```
#!/usr/bin/env perl
# V0.0.8
# added --no_blast ARGUMENT, in case the blast report already done
#       unlink $infile from @filenamelist after blast job
#               new version of blast results parsing - head Accession based
#               with corresponding new functions... details pending
#                               --sourcetracking functions to create source
field in output
#                       sourcetracking module created
use Modern::Perl '2011';
use autodie;

use Getopt::Euclid;
use Smart::Comments '###';
use Template;
use File::Basename;
use Path::Class 'file';

use Tie::IxHash;
use File::Temp;
use List::AllUtils qw( pairwise uniq mesh any first);

use BAF::FastaFile qw(:io);
use BAF::CountFile qw(:io);
use BAF::TaxFile qw(:io);
use BAF::BlastAnalyse qw(:io);
use BAF::sourcetracking_flag qw(:io); ## code of this function is provided

#### Arguments: %ARGV
### reading mismatch: $ARGV{'--mismatch'}
### Reading infile: $ARGV{'<infile>'}
### no blast argument: $ARGV{'--no_blast'}
### sourcetracking : $ARGV{'--sourcetracking'}
### list_acc : $ARGV{'--list'}
```

```
# need only $seq_for, $infile
my $seq_for = read_fasta($ARGV{'<infile>'});

### Reading infile: $ARGV{'--tax-file'}
my $tax_for = read_taxonomy($ARGV{'--tax-file'});

### Reading infile: $ARGV{'--count-file'}
my ($count_for, @label_samples) = read_count($ARGV{'--count-file'});

my                                $source_file                                =
'/Users/dda/Documents/DATABASES/sourcetracking/Flag_sources.txt';
#my                                $source_file                                =
'/Users/dda/Documents/DATABASES/sourcetracking/Acc_id_source.txt';
my $source_for = read_source($source_file) if $ARGV{'--sourcetracking'};
#### %$source_for: %$source_for

my $max_target_seqs = $ARGV{'--max_target'};
my $pgm = 'blastn';     # in my program, no other BLAST program can be used
# my $refile is the path to either silva128 or further ref datafile.
#'128'                                                                      =>
'/Users/dda/Documents/DATABASES/SilvaDB/SILVA_128_SSU.FINAL.fasta'
my %version_for = (
        '128'                                                              =>
'/Users/dda/Documents/DATABASES/SilvaDB/SILVA_128_SSU.FINAL.fasta',
        'test'                                                             =>
'/Users/bernard_taminiau_1/Documents/BD_DATA/BIO_INFO/Cours_perl/mod_perl
/bd_temp.txt',
        '132'                                                              =>
'/Users/dda/Documents/DATABASES/SilvaDB/SILVA_128_SSU.FINAL.fasta',
        '138'                                                              =>
'/Users/dda/Documents/DATABASES/SilvaDB/SILVA_138/SILVA_138_SSURef_tax_si
lva_trunc.DB.FINAL.fasta',
        'ITS'                                                              =>
'/Users/dda/Documents/DATABASES/UNITE/V8/BLAST/UNITE_public_mothur_full_a
ll_02.02.2019.treated.fasta',
        '26S'=>
'/Users/dda/Documents/DATABASES/silva_28S/LSU_132.tax.nr.filter.ng.fasta'
,
);

my $refile = $version_for{$ARGV{'--ref-file-version'}};
### $refile
### read Ref_taxonomy for species assignement
my $taxheader_for = read_reftaxonomy($refile);

my $report = "report.blast.sfmt6";

#my $blast_status = $ARGV{'--no_blast'};
my $noblast_status = $ARGV{'--no_blast'} ? '1' : '0';
### $noblast_status: $noblast_status
if ($noblast_status == '0') {
        # define command template
        # pay attention that multiple report will be passed to parser as
FASTA is likely to be splitted every 100 sequences.
```

```
        # Therefore, no need to label $report with different names, each
time $report will smashed by the previous one.

        my $template = <<'EOT';
        [% pgm %] -query [% infile %] -db [% refile %] -outfmt '6 qacc sacc
mismatch bitscore ' -num_threads 28 -max_target_seqs [% max_target_seqs %]
-max_hsps 1 >> [% report %]
EOT

        my @filenamelist = write_tmp_fasta($seq_for);
        ### Performing BLAST...

        while (my $infile = shift @filenamelist) {

                # build command
                my %vars = (
                        pgm => $pgm,
                        infile => $infile,
                        refile => $refile,
                        report => $report,
                        max_target_seqs => $max_target_seqs,

                );
                my $command;
                my $tt = Template->new;
                $tt->process(\$template, \%vars, \$command);
                #### $command

                system($command);
                unlink $infile;
        }
}

### Parsing BLAST report version 1...

my $parser = get_BAF_parser($report);

my %species_for;
my %hid_for;
my $curr_otu = q{};
my $skip_otu = q{};
my $curr_bitscore = q{};
my @hidlist = ();
my $chimeric_status = 'NA';

HSP:
while ( my $hsp_ref = $parser->() ) {

        my ($qid, $hid, $mismatches, $bitscore) = @{ $hsp_ref }{ qw(query_id
hit_id mismatches bitscore) };
        $hid = $hid =~ s/\|.*//r;

        next HSP if $skip_otu eq $qid; # skip next hit for same OTU
        if ($curr_otu ne $qid){        # next OTU

                if (@hidlist) {
                        my $nearest_hit = join "/", @hidlist ;
```

```
                        my     $species     =     get_species($taxheader_for,
\@hidlist);
                        $species_for{ $curr_otu } = {
                                species => $species,
                                chimeric_status => $chimeric_status,
                                nearest_hit      =>      $nearest_hit      =~
s/\.\d+\.\d+//gr,
                        };
                }
                @hidlist = ();
                if ($mismatches > $ARGV{'--mismatch'}) { # Hit non identique
                        $curr_bitscore = q{};
                        $chimeric_status = 'NA';
                        $skip_otu = $qid;       # to skip next hit for same
OTU

                        # compute $species value
                        my  $species  =  %$tax_for{$qid}->{genus}  ."_"  .
"$qid";
                        # compute chimeric status
                        $chimeric_status  =  'chimeric'  if  $bitscore  <=
$ARGV{'--chimeric'};
                        # capture species and chimeric values for OTU

                        $species_for{ $qid } = {
                                species => $species,
                                chimeric_status => $chimeric_status,
                                nearest_hit => $hid ,
                        };
                        next HSP;
                }
                $chimeric_status = 'NA';
                $curr_otu = $qid;
                $curr_bitscore = $bitscore;

                # compute chimeric status
                $chimeric_status  =  'chimeric'  if  $bitscore  <=  $ARGV{'--
chimeric'};
                # stock hid until species computation
                push @hidlist, $hid;
                next HSP;
        }

        # $curr_otu eq $qid ie 2nd and following hits for same OTU
        if  ($curr_bitscore eq $bitscore) {    # second hid has the same
bitscore, showing redundancy

                push @hidlist, $hid;

                next HSP;

        }

        my $nearest_hit = join "/", @hidlist ;

        # $bitscore lower than $curr_bitscore
        # compute $species value
```

```
        my $species = get_species($taxheader_for, \@hidlist);
        # capture species and chimeric values for OTU
        $species_for{ $qid } = {
                species => $species,
                chimeric_status => $chimeric_status,
                nearest_hit => $nearest_hit,
        };
        $curr_bitscore = q{};
        @hidlist = ();
        $chimeric_status = 'NA';
        $skip_otu = $qid;
}
if (@hidlist) {
        my $nearest_hit = join "/", @hidlist ;

        my $species = get_species($taxheader_for, \@hidlist);
        $species_for{ $curr_otu } = {
                species => $species,
                chimeric_status => $chimeric_status,
                nearest_hit => $nearest_hit,
        };
}

my ($basename, $dir, $suffix) = fileparse($ARGV{'<infile>'},
qr{\.[^.]*}xms);
my $outfile = file($dir, $basename . '.parsed.txt');

### Writing report 1

open my $out, '>', $outfile;

say {$out}
"Kingdom\tPhylum\tClass\tOrder\tFamily\tGenus\tSpecies\tOTU_label\tchimer
a\tnearestID\t" . join "\t", @label_samples;

my @keylist = keys %$seq_for;
while (my $Otu = shift @keylist) {
        my $Kingdom = %$tax_for{$Otu}->{kingdom};
        my $Phylum = %$tax_for{$Otu}->{phylum};
        my $Class = %$tax_for{$Otu}->{class};
        my $Order = %$tax_for{$Otu}->{order};
        my $Family = %$tax_for{$Otu}->{family};
        my $Genus = %$tax_for{$Otu}->{genus};
        my $Species = $species_for{$Otu}->{species} ? $species_for{$Otu}-
>{species} : "no_hit";
        my $chimera = $species_for{$Otu}->{chimeric_status}?
$species_for{$Otu}->{chimeric_status} : "chimeric";
        my $nearest_hit = $species_for{$Otu}->{nearest_hit}?
$species_for{$Otu}->{nearest_hit} : "NA";
        my $counts = %$count_for{$Otu};

        say {$out} join "\t", $Kingdom, $Phylum, $Class, $Order, $Family,
$Genus, $Species, $Otu, $chimera, $nearest_hit, @{$counts};

}

### Parsing BLAST report version 2...
```

```
my $headOTU_for = get_hid_matrix($report, $ARGV{'--mismatch'});

#### OUTPUT_2
my $outfile2 = file($dir, $basename . '.digested.txt');

### Writing report 2
open $out, '>', $outfile2;
#open my $out_acc, '>', $ARGV{'--list'};

say {$out} "Kingdom\tPhylum\tClass\tOrder\tFamily\tGenus\tSpecies\tlist_Accession\tOTU_label\tbitscore\tSource_tracking\t" . join "\t", @label_samples;

my @list_key = sort keys %$headOTU_for;
#### %$headOTU_for: %$headOTU_for

while (my $head_accession = shift @list_key) {
        #### $head_accession: $head_accession
        my $head_OTU = %$headOTU_for{$head_accession}->{head_OTU};
        #### $head_OTU: $head_OTU
        my $head_mismatch = %$headOTU_for{$head_accession}->{head_mismatch};
        my @hidlist = @{ %$headOTU_for{$head_accession}->{Accession_list}};
        #test: $source = get.source($source_for,\@hidlist)
        ## @hidlist: @hidlist
        my @Otulist = @{ %$headOTU_for{$head_accession}->{OTU_list}};
        my $firstOTU = shift @Otulist;

        #generate sum of counts
        my @counts= @{%$count_for{$firstOTU}};

        # could cook a function instead
        COUNTGREP:
        for my $cc (@Otulist) {
                my @tmp_count = @{%$count_for{$cc}};
                @counts = pairwise { $a + $b } @counts, @tmp_count;
        }

        my $list_acc = join "/", @hidlist;

        #generate Taxonomy for $head_accession
        my $Kingdom = %$tax_for{$head_OTU}->{kingdom};
        my $Phylum = %$tax_for{$head_OTU}->{phylum};
        my $Class = %$tax_for{$head_OTU}->{class};
        my $Order = %$tax_for{$head_OTU}->{order};
        my $Family = %$tax_for{$head_OTU}->{family};
        my $Genus = %$tax_for{$head_OTU}->{genus};
        my $Species = get_species_V2($taxheader_for, \@hidlist, $head_OTU, $head_mismatch, $ARGV{'--mismatch'});
        ## @hidlist: @hidlist

        my $Source_tracking = 'unknown_source';
        $Source_tracking = get_source($source_for,\@hidlist, $head_mismatch, $ARGV{'--mismatch'}) if $ARGV{'--sourcetracking'};
        ## @hidlist: @hidlist
```

```
        say {$out} join "\t", $Kingdom, $Phylum, $Class, $Order, $Family,
$Genus, $Species, $list_acc, $head_OTU, %$headOTU_for{$head_accession}-
>{bitscore}, $Source_tracking, @counts;

}

#### OUTPUT_3
my $outfile3 = file($dir, $basename . '.source.txt');

### Writing report 3
open $out, '>', $outfile3;

say {$out} "Sources\t" . "Genus\t" . join "\t", @label_samples ;
#say {$out} "Sources\t" . join "\t", @label_samples . "Genus"
## %$tax_for : %$tax_for


my @list_head_accessions = sort keys %$headOTU_for;

my %occurence_for;     # occurence for source

while (my $head_accession = shift @list_head_accessions) {
        #### $head_accession: $head_accession
        my      $head_mismatch      =      %$headOTU_for{$head_accession}-
>{head_mismatch};
        my @hidlist = @{ %$headOTU_for{$head_accession}->{Accession_list}};
        my $head_OTU = %$headOTU_for{$head_accession}->{head_OTU};
        ## $head_OTU
        my $Genus = %$tax_for{$head_OTU}->{genus};
        ## $Genus

        #build list of sources for head_accession
        my     $Source_tracking     =     get_source($source_for,\@hidlist,
$head_mismatch, $ARGV{'--mismatch'});
        #### $head_accession
        #### $Source_tracking


        #generate sum of counts for head_accession
        my @Otulist = @{ %$headOTU_for{$head_accession}->{OTU_list}};
        ## @Otulist
        my $firstOTU = shift @Otulist;
        my @counts= @{%$count_for{$firstOTU}};
        ## $firstOTU
        ## @counts



        # could cook a function instead

        for my $cc (@Otulist) {
                my @tmp_count = @{%$count_for{$cc}};
                @counts = pairwise { $a + $b } @counts, @tmp_count;

        }

        if ($occurence_for{$Source_tracking}) {
```

```
                my    @tmp_count    =    @{$occurence_for{$Source_tracking}-
>{count}};
                @counts= pairwise { $a + $b } @counts, @tmp_count;
        }
        @{$occurence_for{$Source_tracking}->{count}} = @counts;

        push @{$occurence_for{$Source_tracking}->{genus}}, $Genus;


}
## %occurence_for


my @list_sources = sort keys %occurence_for;
while (my $source = shift @list_sources) {
        my $genus = join "|", uniq(@{$occurence_for{$source}->{genus}});
        say {$out} join "\t", $source, $genus,  @{$occurence_for{$source}-
>{count}} ;
}
#### OUTPUT_4
my $outfile4 = file($dir, $basename . '.source_uniq.txt');
my $outfile5 = file($dir, $basename . '.source_uniq_binaire.txt');

### Writing report 4
open $out, '>', $outfile4;
open my $out1, '>', $outfile5;
say {$out} "Sources\t" . join "\t", @label_samples;
say {$out1} "Sources\t" . join "\t", @label_samples;

@list_head_accessions = sort keys %$headOTU_for;

my %totalcount_for;    # occurence for source

while (my $head_accession = shift @list_head_accessions) {
        #### $head_accession: $head_accession
        my      $head_mismatch      =      %$headOTU_for{$head_accession}-
>{head_mismatch};
        my @hidlist = @{ %$headOTU_for{$head_accession}->{Accession_list}};


        #build list of sources for head_accession
        #generate sum of counts for head_accession
        my @Otulist = @{ %$headOTU_for{$head_accession}->{OTU_list}};
        my $firstOTU = shift @Otulist;
        my @counts= @{%$count_for{$firstOTU}};
        # could cook a function instead
        for my $cc (@Otulist) {
                my @tmp_count = @{%$count_for{$cc}};
                @counts = pairwise { $a + $b } @counts, @tmp_count;
        }
        #### @counts

        my     $Source_tracking     =     get_source($source_for,\@hidlist,
$head_mismatch, $ARGV{'--mismatch'});
        #### $Source_tracking
        my @Keywordset = split qr/\|/xms, $Source_tracking;
```

```
        #### @Keywordset

        KEY:
        while (my $keyword = shift @Keywordset) {
        #### $keyword
                if ($totalcount_for{$keyword}) {
                        #### $keyword
                        my @tmp_count = @{$totalcount_for{$keyword}};
                        my @specific_count = @counts;
                        @specific_count   =   pairwise   {   $a   +   $b   }
@specific_count, @tmp_count;
                        @{$totalcount_for{$keyword}} = @specific_count;
                        next KEY;
                }

                @{$totalcount_for{$keyword}} = @counts;
                #### %totalcount_for
        }

}

@list_sources = sort keys %totalcount_for;

while (my $source = shift @list_sources) {

        if  (@{$totalcount_for{$source}}){

        my @binaire_count = @{$totalcount_for{$source}};
## @binaire_count


                $_ eq 0 and $_ = '0' for  @binaire_count ;
                $_ ne 0 and $_ = '1' for  @binaire_count ;

         say {$out1} join "\t", $source,  @binaire_count ;

        }

        say {$out} join "\t", $source,  @{$totalcount_for{$source}};

}
 say "End of this analyse ... this is it  :)" ;

### done


sub get_species_V2 {
                my $taxheader_for = shift;
                my $hidlist = shift;

                my $head_OTU = shift;
                my $head_mismatch = shift;
                my $mismatch_cutoff = shift;
                my $species_value = undef;
                $species_value    =    $head_OTU    if    $head_mismatch    >
$mismatch_cutoff; #non identical OTU
```

```
                unless ($species_value) {

                        my @species_list;

                        speciesHIT:
                        for my $hid (@$hidlist){
                                my $species = %$taxheader_for{$hid};
                                next    speciesHIT    if    $species    =~
qr{uncultured_bacterium};
                                push @species_list, $species;
                        }

                        $species_value = join "/", uniq(@species_list);
                        #### $head_OTU
                        #### $species_value: $species_value
                        # if $species_value is still undef
                        $species_value   =   join   "/",   @$hidlist   unless
$species_value;
                        #### $species_value: $species_value

                }

                return $species_value;
}

sub get_Hotu_count {
                my $headOTU_for = shift;
                my $count_for = shift;
                my $head_accession = shift;

                my    @Otulist    =    @{    %$headOTU_for{$head_accession}-
>{OTU_list}};
                my $firstOTU = shift @Otulist;

                my @counts= @{%$count_for{$firstOTU}};

                COUNTGREP:
                for my $cc (@Otulist) {
                        my @tmp_count = @{%$count_for{$cc}};
                        @counts = pairwise { $a + $b } @counts, @tmp_count;
                }
        return @counts;
}


=head1 NAME

BAF_anablast.pl

=head1 VERSION

        This documentation refers to annotate version 0.0.2

=head1 USAGE

        BAF_anablast.pl  <infile>  --tax-file  <infile>  --count-file  [=]
<infile> --ref-file-version <string> [options] --list
```

```
=head1 REQUIRED ARGUMENTS

=over

=item <infile>

        Path to input FASTA file.

=for Euclid:
        infile.type: readable

=item --tax-file [=] <infile>

        Path to cons TAXONOMY file.

=item --count-file [=] <infile>

        Path to cons count file.

=for Euclid:
        infile.type: readable

=item --count-file [=] <infile>

                Path to MOTHUR OTU count_table file.

=for Euclid:
                infile.type: readable

=item --ref-file-version [=] <string>

        version of the reference database to use [default: string.default].

The supported fields specifiers are:

        128     means SILVA SSU 1.28 for 16S

        132     means SILVA SSU 1.32 for 16S

        ITS     means UNITE for ITS

        26S means SILVA LSU 1.32 for 26S

        test    for debugging

=for Euclid:
                string.type: string
                string.default: 128

=back

=head1 OPTIONS

=over

=item --mismatch [=] <integer>
```

```
        Mismatch threshold for declaring hit as identical [default:
integer.default].

=for Euclid:
        integer.type: +integer
        integer.default: 5

=item --max_target [=] <integer>

                maximun target sequences hits [default: integer.default].

=for Euclid:
        integer.type: +integer
        integer.default: 100

=item --chimeric [=] <score>

        Bitscore threshold for declaring hit as likely chimeric [default:
score.default].

=for Euclid:
        score.type: +integer
        score.default: 500

=item --no_blast

        use yes if blast report is already done and you want to skip
directely to parsing results. [Default: no].

=item --sourcetracking

        use argument if source tracking for identical accession is needed.
[Default: no].
=item --list [=] <outfile>

=for Euclid:
                outfile.type: writeable

=item --usage

=item --help

=item --man

        Print the usual program information

=back

=head1 AUTHOR

        Bernard Taminiau & Hiba JABRI (bernard.taminiau@uliege.be)

=head1 BUGS

        There are undoubtedly serious bugs lurking somewhere in this code.
        Bug reports and other feedback are most welcome.
```

```
=head1 COPYRIGHT

        Copyright (c) 2013, Bernard Taminiau. All Rights Reserved.
        This program is free software. It may be used, redistributed
        and/or modified under the terms of the Perl Artistic License
        (see http://www.perl.com/perl/misc/Artistic.html)
```

Appendix 13:

```
package BAF::sourcetracking_flag;

use Modern::Perl '2011';
use autodie;
use List::AllUtils qw( uniq );
use Template;
use DBI;


use Exporter::Easy (
        OK => [ qw(:io) ],
        TAGS => [
                io => [ qw(read_source get_source) ],
        ],
);

=head1 NAME

BAF::sourcetracking_flag - The great new BAF::sourcetracking!

=head1 VERSION

Version 0.02

=cut

our $VERSION = '0.02';


=head1 SYNOPSIS

Quick summary of what the module does.

Perhaps a little code snippet.

    use BAF::sourcetracking_flag;

    my $foo = BAF::sourcetracking->new();
    ...

=head1 EXPORT

A list of functions that can be exported.  You can delete this section
if you don't export anything, such as for a purely object-oriented module.
```

```
=head1 SUBROUTINES/METHODS

=head2 read_source

  Read a source tracking file containing the source keywords associated to
NCBI accessions ids.
  Return a reference to a hash of arrays containing accession/keywords
pairs.

=cut

sub read_source{
        my %source_for;
        #my                                $dbfile                                =
"/Users/dda/Documents/DATABASES/Source_DATABASES/MyHOST_METAV4.sdb";
        my                                $dbfile                                =
"/Volumes/Talvza_Pegasus/DOSSIER_HIBA_J/Source_DATABASES/BiotopeBac/Bioto
peBac.sdb";
        my   $dbh   =   DBI->connect("DBI:SQLite:dbname=$dbfile","","",   {
RaiseError => 1 })
                or die $DBI::errstr;
                print "Opened database successfully\n";

        my $stmt = qq(select Sequences_idsequence, idwordset from Sources,
Keywords where idkeyword = Keywords_idkeyword;);
        my $sth = $dbh->prepare( $stmt );
        my $rv = $sth->execute() or die $DBI::errstr;

        if($rv < 0) {
                print $DBI::errstr;
        }

        while(my @row = $sth->fetchrow_array()) {
                my $Accession = $row[0];
                my $Keywordset = $row[1];
                push @{$source_for{$Accession}}, $Keywordset;

         }
        return \%source_for;
}

=head2 get_source

read a reference to source Hash and a list of NCBI Accession and return a
scalar containg the dereplicated list of source keywords.

=cut

sub get_source {
        my $source_for = shift;
        my $hidlist = shift;
        my $head_mismatch = shift;     # mismatch value for head OTU blast
best hit
        my    $mismatch_cutoff    =    shift;    #$ARGV{'--mismatch'}    from
BAF_anablast.pm
```

```
        my $source_valeur = undef;
        $source_valeur    =    'Not_identical_OTU'    if    $head_mismatch    >
$mismatch_cutoff; #non identical OTU

        unless ($source_valeur) {


                my @source;

                source:
                for my $hid (@$hidlist) {
#                       $hid = $hid =~ s/\.\d+\.\d+//gr;
                        next source unless (%$source_for{$hid});

                        #old_command
                        for my $keyword (@{%$source_for{$hid}}){

                        #for my $keyword (%$source_for{$hid}){
                  push @source, $keyword;
                        }
                }
                #trier et uniq le @source
                my @source_lc = sort (@source);
                $source_valeur = join "|", uniq(@source_lc);
                ### $source_valeur: $source_valeur
                $source_valeur = "uknown_source" unless $source_valeur;

        }


        return $source_valeur;
}

=head1 AUTHOR

Hiba Jabri, C<< <hiba.jabri at uliege.be> >>

=head1 BUGS

Please report any bugs or feature requests to C<bug-baf-sourcetracking at
rt.cpan.org>, or through
the                  web                  interface                  at
L<https://rt.cpan.org/NoAuth/ReportBug.html?Queue=BAF-sourcetracking>.  I
will be notified, and then you'll
automatically be notified of progress on your bug as I make changes.




=head1 SUPPORT

You can find documentation for this module with the perldoc command.

    perldoc BAF::sourcetracking


You can also look for information at:
```

```
=over 4

=item * RT: CPAN's request tracker (report bugs here)

L<https://rt.cpan.org/NoAuth/Bugs.html?Dist=BAF-sourcetracking>

=item * AnnoCPAN: Annotated CPAN documentation

L<http://annocpan.org/dist/BAF-sourcetracking>

=item * CPAN Ratings

L<https://cpanratings.perl.org/d/BAF-sourcetracking>

=item * Search CPAN

L<https://metacpan.org/release/BAF-sourcetracking>

=back


=head1 ACKNOWLEDGEMENTS


=head1 LICENSE AND COPYRIGHT

This software is Copyright (c) 2019 by Hiba Jabri.

This is free software, licensed under:

  The Artistic License 2.0 (GPL Compatible)


=cut

1; # End of BAF::sourcetracking
```

Appendix 14:

```
package BAF::sourcetracking;

use Modern::Perl '2011';
use autodie;
use List::AllUtils qw( uniq );



use Exporter::Easy (
        OK => [ qw(:io) ],
        TAGS => [
                io => [ qw(read_source get_source) ],
        ],
);
```

```
=head1 NAME

BAF::sourcetracking - The great new BAF::sourcetracking!

=head1 VERSION

Version 0.02

=cut

our $VERSION = '0.02';


=head1 SYNOPSIS

Quick summary of what the module does.

Perhaps a little code snippet.

    use BAF::sourcetracking;

    my $foo = BAF::sourcetracking->new();
    ...

=head1 EXPORT

A list of functions that can be exported.  You can delete this section
if you don't export anything, such as for a purely object-oriented module.

=head1 SUBROUTINES/METHODS

=head2 read_source

  Read a source tracking file containing the source keywords associated to
NCBI accessions ids.
  Return a reference to a hash of arrays containing accession/keywords
pairs.

=cut

sub read_source{
        my $infile = shift;
        open my $in, '<', $infile;
        my %source_for;
        while (my $line = <$in>){
                chomp $line,
                my ($Acc, @sources) = split /[\|,]/xms, $line;
                for my $source (@sources) {
                        push @{$source_for{$Acc}}, $source;
                }
        }
        return \%source_for;
}

=head2 get_source
```

```
read a reference to source Hash and a list of NCBI Accession and return a
scalar containg the dereplicated list of source keywords.

=cut

sub get_source {
        my $source_for = shift;
        my $hidlist = shift;
        my $head_mismatch = shift;     # mismatch value for head OTU blast
best hit
        my    $mismatch_cutoff    =    shift;    #$ARGV{'--mismatch'}    from
BAF_anablast.pm

        my $source_valeur = undef;
        $source_valeur = 'NA' if $head_mismatch > $mismatch_cutoff; #non
identical OTU

        unless ($source_valeur) {


                my @source;

                source:
                for my $hid (@$hidlist) {
#                       $hid = $hid =~ s/\.\d+\.\d+//gr;
                        next source unless (%$source_for{$hid});
#                       my @keywords = @{%$source_for{$hid}};
                        for my $keyword (@{%$source_for{$hid}}){
                  push @source, $keyword;
                        }
#                       #
                }
                #trier et uniq le @source
                $source_valeur = join "|", uniq(@source);
                ### $source_valeur: $source_valeur
                $source_valeur = "NA" unless $source_valeur;

        }


        return $source_valeur;
}

=head1 AUTHOR

Hiba Jabri, C<< <hiba.jabri at uliege.be> >>

=head1 BUGS

Please report any bugs or feature requests to C<bug-baf-sourcetracking at
rt.cpan.org>, or through
the                  web                  interface                  at
L<https://rt.cpan.org/NoAuth/ReportBug.html?Queue=BAF-sourcetracking>.   I
will be notified, and then you'll
automatically be notified of progress on your bug as I make changes.
```

```
=head1 SUPPORT

You can find documentation for this module with the perldoc command.

    perldoc BAF::sourcetracking

You can also look for information at:

=over 4

=item * RT: CPAN's request tracker (report bugs here)

L<https://rt.cpan.org/NoAuth/Bugs.html?Dist=BAF-sourcetracking>

=item * AnnoCPAN: Annotated CPAN documentation

L<http://annocpan.org/dist/BAF-sourcetracking>

=item * CPAN Ratings

L<https://cpanratings.perl.org/d/BAF-sourcetracking>

=item * Search CPAN

L<https://metacpan.org/release/BAF-sourcetracking>

=back


=head1 ACKNOWLEDGEMENTS


=head1 LICENSE AND COPYRIGHT

This software is Copyright (c) 2019 by Hiba Jabri.

This is free software, licensed under:

  The Artistic License 2.0 (GPL Compatible)


=cut

1; # End of BAF::sourcetracking
```