Albert-Ludwigs University Freiburg
Department of Computer Science
Bioinformatics Group

Master Thesis

What is in the microbiome of beers? Aggregation of public beer microbiome data, development and evaluation of FAIR beer microbiome workflows, and implementation of BeerMicroDB, a comprehensive database for beer microbiome

Author:
Yedil Serzhan

Examiner:
Prof. Dr. Rolf Backofen

Second Examiner:
Prof. Dr. Wolfgang R. Hess

Advisors:
Dr. Bérénice Batut, M.Sc. Teresa Müller

Submission date:
09.07.2023

Declaration

I hereby declare that I am the sole author and composer of my thesis and that no other sources or learning aids, other than those listed, have been used. Furthermore, I declare that I have acknowledged the work of others by providing detailed references of said work.
I hereby also declare that my Thesis has not been prepared for another examination or assignment, either wholly or excerpts thereof.

Place, Date

Signature

Acknowledgments

First of all, I would like to express my gratitude to Prof. Rolf Backofen. Having the opportunity to complete my thesis in the Bioinformatics Group at the University of Freiburg is a huge step for me. Moreover, it's an honor to have him as my examiner.

I'm also grateful to Prof. Dr. Wolfgang R. Hess for considering the role of my second examiner. His expertise and willingness to participate in this process are greatly appreciated.

My supervisors Bérénice Batut and Teresa Müller have been anchors in my research journey. Looking back to my beginnings and to the present, their influence on me at every step of the way is evident; Bérénice's insights often illuminated the direction of my research, while Teresa's pragmatism served as a compass when I felt lost. The patience they showed, especially when I was skeptical, was commendable. And the countless hours of discussion? They were the building blocks of my confidence. Their trust in me, even when I was hesitant, pushed me to aim higher and achieve more. I could not have asked for better guides on this journey than them.

I appreciate my dear friends who have been there for me throughout this process. A heartfelt shoutout to Polina – introducing me to the opportunities at the Bioinformatics Group was a turning point. I'm ever grateful for that.

The support of my family has always surrounded me, even from a distance. My parents and sister, your strong belief in me pushes me forward. Not only do you applaud my academic progress, but you are always concerned about my mental health and remind me of the importance of balance. For all this and more, I am grateful beyond words.

Thank you all for being a part of this incredible journey.

Abstract

Beer is a globally consumed fermented beverage, with its unique attributes largely attributed to the complex microbial communities involved in fermentation. Advancements in DNA sequencing have brought more opportunities to analyze these microbial communities, providing insights into beer fermentation, flavor, and microbiome. This research introduces innovative workflows on the Galaxy platform for standardized beer microbiome analysis, utilizing QIIME 2 for metabarcoding and Kraken 2 for shotgun approaches. Within the context of prior studies, reproducibility analysis was conducted using workflows implemented on collected data. Preliminary findings underscore the rich microbiome diversity in beers subjected to spontaneous fermentation and aging, particularly in traditional African beer, sesotho, and craft lagers. Distinctive bacterial and fungal species prevalent in the beer samples are identified. Concurrently, the study presents the BeerMicroDB — a public database encompassing 56 beer types and 301 samples. This endeavor aims to bolster research in beer microbiomes, shedding light on microbial influence in beer characteristics and promoting collaboration in the beer microbiome community.

Contents

1 Introduction

Beer is one of the most consumed fermented beverages in the world and an integral part of human food culture. According to statistical data, in 2021, the global brewing industry experienced a growth of around four percent in beer production, reaching approximately 1.86 billion hectoliters internationally [1]. Beer owes many of its unique qualities to the complex communities of microorganisms involved in the fermentation processes. The identification and analysis of these microbial communities can be helpful for understanding the production and flavor of these beverages and their impact on human health. Over the years advances in DNA sequencing technology have made it possible to easily and affordably monitor microbial communities during fermentation. There are some studies on the beer microbiome, such as on American beer [2], sour beers [3] and Swiss beers [4].

This thesis aims to enhance and establish new workflows on the Galaxy platform for beer microbiome analysis, addressing the current lack of standardization in the field and improving the reproducibility of results. The motivation behind this endeavor stems from our hypothesis that the beer microbiome varies across different beer types, with specific microbiomes influencing certain beer qualities, including flavor and aroma. To elucidate these relationships, the study will gather both shotgun and metabarcoding data from the European Nucleotide Archive (ENA) [5] and MG-RAST [6]. Based on the data and results produced by the workflows above, the development of a comprehensive beer microbiome database is set to provide a crucial resource, enabling researchers and brewers to delve deeper into the role of microorganisms in beer production and their impact on quality. This database not only serves as a repository of knowledge but also promotes information sharing, fostering collaboration and innovation within the beer microbiome community. Ultimately, by standardizing beer microbiome analysis workflows and creating this database, this thesis sets the groundwork for a more rigorous and collaborative future in beer microbiome research.

The database will be publicly hosted on a website. Users will be granted the capability to explore a collection of beer samples. Each entry will provide detailed metadata about the specific beer sample, as well as the microbiome composition determined by our established workflows. With the insights generated from this database, we seek to elucidate the following research questions: (i) Which microbial populations are predominantly found in beers? (ii) Can the workflow implemented in the thesis be able to reproduce the results in previous studies? (iii) Does a recognizable trend exist between beer types and their microbiome?

2 State of the art

2.1 Microbiome

Microorganisms also known as microbes are microscopic life forms. They pervade virtually every dimension of our planet. These tiny creatures exhibit a remarkable range of diversity. They include bacteria, archaea, fungi, protozoa, microalgae, and non-living viruses. And they are present in almost every environment on Earth. This includes soil, water, and air, as well as within and on the human body. Microbes play a vital role in maintaining ecological stability as they participate in processes like nutrient cycling, organic matter breakdown, and the formation of symbiotic connections with an array of other life forms. Furthermore, their influence on human health is substantial, with some microorganisms functioning as pathogens that induce disease, while others foster general well-being[7].

Since the advent of the microscope by Levenhoek, the investigation of microorganisms has been a continuous pursuit, leading to numerous scientific and technological advancements. Researchers have harnessed microbes for diverse objectives, including the generation of antibiotics, the enhancement of fermentation processes, and the development of groundbreaking biotechnologies. As our comprehension of microbes continues to evolve, their capacity to impact and improve numerous facets of human life and the environment is further magnified.

The microbiome is the collective genome of all microorganisms that live in a given environment. The study of the microbiome is very important. To take the human microbiome as an example, the microbes that live in and on our bodies play a key role in maintaining homeostasis, influencing metabolism, and regulating the immune system. In fact, it is estimated that there are 10 times more microbial cells in the human body and that the combined (meta) genome of the microbes is more than 100 times the human genome [8].

In recent years, the study of the microbiome has had more opportunities with advances in high-throughput sequencing technologies, which have made it possible to describe the details of microbial populations[9]. Researchers can now use technologies such as 16S ribosomal RNA (rRNA) gene sequencing, Internal Transcribed Spacer (ITS) rRNA gene sequencing, and metagenomics to identify and quantify the microbes present in a given environment and their functional capabilities.

Understanding the composition and function of the microbiome is important for a wide range of applications, from human health and disease to environmental protection and biotechnology to fermented foods, such as beer and wine. For example, researchers are investigating the role of the gut microbiome in diseases such as obesity, diabetes, and inflammatory bowel disease, and are exploring the potential of using microbiome-based therapies to treat these diseases[10]. In addition, the study of microbial communities in the environment is important for understanding ecosystem function and biodiversity, and for developing sustainable agricultural and industrial processes [11]. Many microbiomes are also involved in the production and fermentation of products such as beer, and understanding these microbiomes will be of great help in the development of fermentation for beer manufacturing.

2.2 Microbiome analysis

Microbiome analysis is a complex process that involves multiple steps from sample collection to data analysis. Each step and the different tools or protocols that can be used are outlined in this section.

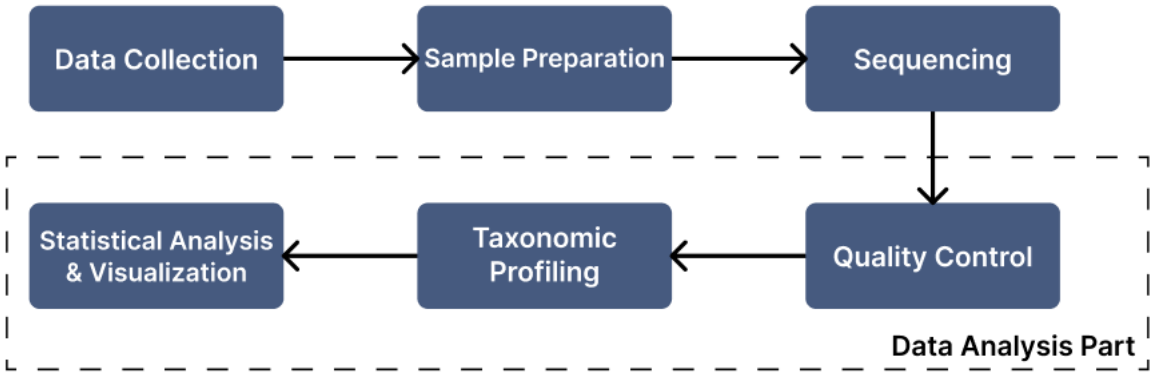


Figure 1: Common microbiome analysis steps
Microbiome analysis typically follows a structured methodology. The process begins with data collection, followed by sample preparation, sequencing, and ultimately, data analysis. The latter, as depicted by the dashed section in the figure, encompasses steps such as quality control, taxonomic profiling, as well as diversity analysis, and visualization. For methods like shotgun sequencing and metabarcoding, the tools used in the steps above can be different.

Data collection

Sample collection is the first step in microbiome analysis and involves collecting samples from the environment of interest, such as soil, water, or host organisms. The use of appropriate sampling techniques is essential to maintain the integrity of the microbial community. Methods such as wiping, filtration, and direct sampling are commonly used and protocols vary depending on the sample type and study objectives.

Sample preparation

Sample preparation involves a series of procedures to prepare the collected samples for downstream high-throughput sequencing and analysis. Sample preparation encompasses processes that isolate and concentrate microbial DNA or RNA from complex sample matrices, convert it into a format suitable for sequencing, and generate libraries for high-throughput sequencing.

For metabarcoding, which is also referred to as amplicon sequencing, the sample preparation process involves extracting genomic DNA from the collected samples, amplifying targeted DNA regions using PCR with specific primers (usually 16S rRNA gene for bacteria and ITS region for fungi), and preparing amplicon libraries for sequencing. This targeted approach enables the identification of specific taxa present in the sample, based on the sequenced marker genes. The advantages of this method are its simplicity, speed, low cost, and mature analytical techniques,

while the disadvantage is its limitation to the target marker genes so no global genomic functional context.

In contrast, for shotgun metagenomics, the sample preparation process involves extracting total DNA or RNA from the collected samples, without any targeted amplification. The extracted nucleic acids are then fragmented and converted into sequencing libraries, capturing the entire genetic information within the sample, including both coding and non-coding regions. This approach allows for the characterization of the complete microbial community composition and functional potential within the sample. The limitation of shotgun sequencing is that it is of high cost, often includes contamination including the host, and requires high-performance computing, and high memory.

Sequencing

As for sequencing technologies, there are 2 popular techniques: (1) Illumina and (2) Nanopore. Illumina sequencing is a second-generation sequencing technique that uses reversible dye terminator technology to detect the sequence of DNA molecules. Solexa company, now a part of Illumina company, was founded in 1998. This company invented a sequencing method based on reversible dye terminator technology and engineered polymerases.

In the Illumina sequencing method, the sample is first cleaved into short Fragments. Therefore, in Illumina sequencing, about 100-150bp long reads or fragments are created at the beginning. These fragments are then ligated to generic adaptors and annealed to a slide. Bridge amplification is done to amplify each fragment. This creates a spot with many copies of the same fragment. Later, they are separated into single-stranded fragments and subjected to sequencing. The sequencing slide contains fluorescently labeled nucleotides, DNA polymerase, and a terminator. Because of the terminator, only one base is added at a time. Each cycle terminator has washed away, and it allows the addition of the next base to the site. Furthermore, based on fluorescent signals, the computer detects the base added in each cycle. Illumina sequencing technology constructs the sequence within 4 to 56 hours[12].

Nanopore sequencing, a third-generation sequencing method, employs the protein nanopore to determine the nucleic acid sequence of a nucleotide sequence. The technology has seen considerable expansion in both basic and applied research since Oxford Nanopore Technologies (ONT) introduced the first Nanopore sequencer, MinION, in 2014. This technique relies on a nanoscale protein pore, or "nanopore," which serves as a biosensor and is embedded in an electrically resistant polymer membrane [13]. A motor protein controls the translocation speed by guiding the nucleic acid molecule through the nanopore in a step-by-step fashion. Changes in the ionic current during translocation correspond to the nucleotide sequence in the sensing region, which is then decoded using computational algorithms, allowing for real-time sequencing of individual molecules. Besides regulating translocation speed, the motor protein possesses helicase activity, facilitating the unwinding of double-stranded DNA or RNA-DNA duplexes into single-stranded molecules that can pass through the nanopore.

2.2.1 Data analysis

As shown in figure 1, the data analysis part in the general microbiome analysis contains 3 steps: quality control, taxonomy classification, and diversity analysis and visualization.

Quality control and pre-processing

After sequencing, raw sequence data is subjected to quality control and pre-processing steps by detecting and removing adapter contamination and low-quality reads which could cause reduced accuracy in downstream analysis. Tools such as FastQC [14] assess read quality, while Trimmomatic [15] and Cutadapt [16] trim low-quality bases and remove spliced sequences. Pre-processing ensures that only high-quality reads are used for downstream analysis.

Taxonomic profiling

Taxonomic profiling is a crucial step in the microbiome analysis pipeline that aims to identify and classify the microorganisms present in a sample. For metabarcoding (also referred to as amplicon) data, tools such as QIIME2 [17], Mothur [18] and USEARCH [19] assign operational taxonomic units (OTUs) or amplicon sequence variants (ASVs) and assign taxonomy. OTUs are generated through clustering methods. These methods are founded on the principle that organisms with similar target gene sequences are likely to be closely related. In contrast, the ASV approach takes a different route. It begins by identifying the exact sequences that were observed in the sequencing data and quantifying their frequencies. This information is then combined with an error model. For shotgun data, classification can be performed using tools such as MetaPhlAn [20] and Kraken [21].

Below are some state-of-the-art tools and methods used for analyzing metabarcoding data:

QIIME 2

`QIIME 2` represents a cutting-edge, extensible, open-source, and community-developed microbiome bioinformatics platform [17].

QIIME 2 has integrated and automatic data provenance tracking that ensures decentralized tracking of analyses. It eliminates guesswork about the executed commands, as QIIME 2 artifacts, comprising data and metadata, facilitate automatic tracking of data type, format, and provenance. This enhances the replicability and reproducibility of analyses.

Each QIIME 2 artifact is associated with a semantic type that helps identify appropriate inputs for analysis and prevents incompatible artifacts from being used. This semantic type system aids users in avoiding semantically incorrect analyses.

QIIME 2 features a plugin-based system, allowing users to access a broad range of microbiome analysis methods. Users can install plugins developed by both the QIIME 2 team and third-party developers, expanding the availability of tools and techniques.

QIIME 2 supports various user interfaces, including a command line interface (`q2cli`), a web-based graphical user interface (`q2galaxy`), and a Python 3 API (Artifact API). The `q2cli` requires command line interaction but has comprehensive documentation. The `q2galaxy` is

user-friendly and requires no command inputs, while the Artifact API is suited for advanced users, supporting interactive computing using Python 3.

Mothur

`Mothur`, an open-source and extensible microbial community analysis tool, was developed by Schloss et al. at the University of Michigan's Department of Microbiology and Immunology [18]. The tool caters to the bioinformatics analysis needs of microbial community ecology.

`Mothur` integrates algorithms from tools such as `DOTUR`, `SONS`, `TreeClimber`, `LIBSHUFF`, `∫-LIBSHUFF`, and `UniFrac`, in addition to its own features.

`Mothur` offers additional features, including diversity assessment, visualization tools like Venn diagrams, heat maps, and dendrograms, sequence quality screening functions, a NAST-based sequence aligner, and a pairwise sequence distance calculator.

Users can run `Mothur` in three modes - interactive, batch, and command line. The interactive mode provides prompt feedback and is ideal for beginners or working with individual files.

`Mothur` has extensive community-supported documentation through a MediaWiki-based wiki and a phpBB-based discussion forum. The wiki format allows users to read, edit, and expand content, promoting diverse experiment design and data analysis.

DADA2

`DADA2` is a cutting-edge bioinformatics tool designed for fast and accurate sample inference from metabarcoding data with single-nucleotide resolution [22]. It offers several advantages over traditional methods, making it an essential tool for researchers in the field of microbial ecology and other disciplines that rely on metabarcoding sequencing data.

`DADA2` is known for its ability to infer exact amplicon sequence variants (ASVs) from amplicon data, resolving biological differences down to 1 or 2 nucleotides. This level of resolution allows for the detection of subtle differences in microbial communities.

`DADA2` is highly accurate in identifying sequence variants, reporting significantly fewer false-positive sequence variants compared to other methods that produce false OTUs. This ensures that researchers can rely on the data generated by `DADA2` for robust downstream analyses.

`DADA2`'s output of ASVs can be directly compared between studies without the need for reprocessing pooled data, facilitating meta-analyses and cross-study comparisons.

`DADA2`'s computational efficiency is another advantage, as the software's compute time scales linearly with the sample number, and memory requirements remain flat. This allows for the analysis of large datasets without compromising on processing time or computing resources.

Comparions of QIIME 2, Mothur and DADA2

Feature	QIIME 2	Mothur	DADA2
Resolution	High, with the ability to distinguish closely related microbial taxa	High, but inferior to DADA2's single-nucleotide resolution	Single-nucleotide resolution
Comparability	Extensive interoperability and compatibility with other tools	Limited compatibility with external tools	ASV output allows direct comparisons between studies
Flexibility	Plugin-based system for expanding microbiome analysis functionality	Allows users to run individual commands, batch files, or directly from the command line	Less flexibility in creating customized analysis pipelines
User Interface	Offers command line interface, graphical user interface, and Artifact API	Interactive mode, batch mode, and command line mode	Primarily designed for use with the R programming language
Documentation	Comprehensive documentation and user support	Extensive community-supported documentation	Thorough documentation and community support

Table 1: Comparison of QIIME 2, Mothur, and DADA2

Despite the strengths of each of the three tools (QIIME 2, Mothur, and DADA2), QIIME 2 stands out for its versatility, ease of use, flexibility, and community support as shown in the table 1. The plugin-based system and wide variety of user interfaces make it accessible to both novice and experienced users. Moreover, its efficient computational scaling, high accuracy, and interoperability make it an attractive choice for microbial community analysis. The integrated and automatic data provenance tracking feature also greatly enhances the reproducibility and transparency of analyses conducted with QIIME 2, making it a more attractive choice for academic and research settings. While Mothur and DADA2 have their own strengths, they lack the versatility and comprehensive user support that QIIME 2 provides. Therefore, QIIME 2 is considered a superior tool for metabarcoding data analysis in microbial community studies.

some state-of-the-art tools and methods used to analyze shotgun data are described below:

Kraken 2

`Kraken 2` is a sequence classification system that uses a k-mer based approach to assign taxonomic labels to DNA sequences [21]. It is designed to be fast and efficient, making it well-suited for large datasets. `Kraken 2` uses a pre-built database of microbial genomes to classify sequences. `Kraken 2` uses k-mers from the DNA sequences to classify them to the lowest taxonomic rank possible, based on the reference database. The output of `Kraken 2` is a list of taxonomic labels

and their relative abundance, which can be used for downstream analysis. `Kraken 2` can also generate reports to help interpret the results.

Metaphlan

`Metaphlan` is a metagenomic profiling tool that uses unique clade-specific marker genes to profile the taxonomic composition of microbial communities [20]. It is designed to be highly accurate and sensitive and can identify both bacteria and archaea. `Metaphlan` includes a pre-built database of marker genes and can generate both abundance and presence/absence profiles. The marker genes used by `Metaphlan` are specific to different taxonomic groups and are used to identify the presence and abundance of those groups in metagenomic datasets. `Metaphlan` 's accuracy and sensitivity make it well-suited for detecting low-abundance microbial taxa that may be missed by other methods. The output of `Metaphlan` includes taxonomic abundance estimates and a taxonomic tree, which can be visualized and used for downstream analysis.

mOTUs

`mOTUs` is a metagenomic clustering tool that groups similar sequences into OTUs based on their genomic content[23]. It is designed to be highly accurate and specific and can identify both bacteria and archaea. mOTUs includes a pre-built database of microbial genomes and can generate both abundance and presence/absence profiles. `mOTUs` algorithm uses a clustering approach that is based on the pairwise similarity of genomic content between sequences. This approach is useful for providing information on the functional potential of the microbial community. The output of `mOTUs` includes a list of OTUs and their relative abundance, which can be used for downstream analysis.

Comparison of Kraken 2, Metaphlan and mOTUs

In comparing these tools, `Kraken 2` emerges as the superior option for several reasons. Primarily, `Kraken 2`'s speed and efficiency in handling large datasets make it highly suitable for high-throughput sequencing studies. Its k-mers based approach and comprehensive microbial genome database allow for accurate classification to the lowest taxonomic rank possible. Additionally, the detailed reports generated by `Kraken 2` facilitate the interpretation of results and downstream analysis. While `Metaphlan` and `mOTUs` have their own unique advantages and are both highly accurate and specific, the overall speed, efficiency, and comprehensive reporting offered by `Kraken 2` make it the more desirable choice for shotgun metagenomic analysis.

Diversity analysis and visualization

Finally, various diversity analysis methods and visualizations of the analyzed data were performed to determine differences in microbial community structure and function. Alpha diversity and beta diversity are commonly used to measure the diversity within samples and between samples. In addition, specialized software such as Pavian [24] and Krona [25] provide user-friendly interfaces to generate information visualizations.

By following these general steps above and employing appropriate tools and protocols, researchers can perform comprehensive microbiome analyses to gain valuable insights into the complex microbial communities that underpin various ecosystems and host organisms.

2.2.2 Workflow management systems

Even with the right analysis tools and methods, the complexity of these studies can often be overwhelming, rendering efficient data management and process automation critical to their successful implementation. Scientists often employ different tools and software at various stages of the analysis, which might not be entirely compatible or require different computational environments to function optimally. Thus, in managing and manipulating the generated datasets, particularly in the context of high-throughput sequencing, we are brought to the doorstep of another fundamental aspect of bioinformatics: Workflow Management Systems.

Workflow management pertains to the process of designing, documenting, monitoring, and refining a sequence of operations required to fulfill a particular task.

In the context of bioinformatics, workflow management systems are pivotal as they streamline and automate multistep computational investigations. They offer a standard lexicon for outlining analysis workflows, which in turn aids in replicability and in the establishment of libraries with reusable components.

The primary advantage of these systems is that they enable bioinformaticians to delegate the reproducibility of workflows and the execution of workflows to the management system. Consequently, researchers can focus their attention on the actual functionality of these workflows, the interpretation of their data, and the advancement of their scientific inquiries.

Snakemake

Snakemake is a workflow management system designed for creating reproducible and scalable data analyses. The system uses a human-readable, Python-based language to describe workflows, enabling seamless scaling to various environments, including servers, clusters, grids, and clouds, without needing to alter the workflow definition. Additionally, Snakemake workflows can include descriptions of necessary software, automatically deploying them to any execution environment[26].

Snakemake's core concept involves specifying workflows by decomposing them into rules or steps. Each rule dictates the process of deriving a set of output files from a set of input files, which can be accomplished through a shell command, Python code block, external script (Python, R, or Julia), or a Jupyter notebook. The example Snakemake workflow consists of a workflow definition, a directed acyclic graph (DAG) of jobs representing the automatically derived execution plan, and a script referred to from a rule.

The workflow definition language of Snakemake emphasizes maximum readability, which is vital for transparency and adaptability. Factors influencing readability include the use of known words, intuitive punctuation, and operator usage.

Snakemake is easily deployable through the Conda package manager, Python package, or Docker container.

During workflow processing, Snakemake tracks input files, output files, parameters, software, and code for each executed job. Upon completion, this information can be accessed through self-contained, interactive, HTML-based reports, facilitating the exploration of results alongside their

provenance information. Snakemake reports are portable and archivable, as their presentation does not rely on server backends.

Like many other advanced workflow management systems, Snakemake enables workflow execution to scale across various computational platforms, from single workstations to large compute servers, common cluster middleware, grid computing, and cloud computing. Snakemake's design ensures that scaling a workflow to a specific platform only requires modifying command-line parameters, leaving the workflow itself unchanged. Configuration profiles allow for the persistence and sharing of Snakemake command-line setup for any computing platform.

Nextflow

Nextflow is a platform that facilitates scalable and reproducible scientific workflows using software containers, allowing for the adaptation of pipelines written in widely-used scripting languages [27]. Its fluent domain-specific language (DSL) streamlines the implementation and deployment of intricate parallel and reactive workflows on clouds and clusters, with Linux serving as the foundation for data science.

Nextflow simplifies the process of creating computational pipelines by seamlessly integrating various tasks. It allows for the reuse of existing scripts and tools, eliminating the need to learn a new language or API to begin using the platform.

Nextflow supports Docker and Singularity container technologies, enabling seamless integration with the GitHub code-sharing platform. This feature allows users to create self-contained pipelines, manage versions, and quickly reproduce previous configurations.

Nextflow offers an abstraction layer between the pipeline logic and the execution layer, enabling execution on multiple platforms without modification. Out-of-the-box executors are provided for GridEngine, SLURM, LSF, PBS, Moab, and HTCondor batch schedulers, as well as Kubernetes, Amazon AWS, Google Cloud, and Microsoft Azure platforms.

Built on the dataflow programming model, Nextflow greatly simplifies the development of complex distributed pipelines. Parallelization is implicitly defined by the processes' input and output declarations, resulting in inherently parallel applications that can scale up or out without adapting to a specific platform architecture.

Nextflow automatically tracks all intermediate results generated during pipeline execution. This feature enables users to resume execution from the last successful step, regardless of the reason for stopping.

Nextflow enhances the Unix pipes model with a fluent DSL, simplifying the handling of complex stream interactions. This approach promotes functional composition-based programming, leading to resilient and easily reproducible pipelines.

Common Workflow Language (CWL)

The Common Workflow Language (CWL) constitutes an open standard for outlining the execution of command line tools and their integration to form workflows. CWL-based tools and workflows are compatible with various platforms that adhere to CWL standards, facilitating the

scaling of intricate data analysis and machine learning workflows from an individual developer's laptop to massive parallel clusters, cloud, and high-performance computing environments[28].

CWL's execution model is explicit, with each tool's runtime environment clearly defined, and any necessary elements specified by the CWL tool-description author. Every tool invocation employs a separate working directory, populated in accordance with the CWL tool description. Additional constructs in the CWL Command Line Tool standard cater to applications requiring specific filenames, directory layouts, and environment variables.

The CWL Workflow Description Standard builds on the CWL Command Line Tool Standard, utilizing the same YAML- or JSON-style syntax and featuring explicit workflow-level inputs, outputs, and documentation. Workflow descriptions consist of a list of steps, including CWL Command Line Tools or CWL sub-workflows, each re-exposing their tool's essential inputs. Inputs for each step are connected by referencing the name of either the common workflow inputs or outputs from other steps. Workflow outputs reveal selected outputs from workflow steps, explicitly indicating which intermediate-step outputs will be returned from the workflow. All connections contain identifiers that CWL document authors are encouraged to name meaningfully.

CWL execution requires the implementation of CWL standards rather than a specific software product. Workflow/tool runners that comply with CWL standards can flexibly execute a user's CWL documents as long as they fulfill the requirements outlined in those documents. However, aspects not defined by CWL standards include Web APIs for workflow execution and real-time monitoring.

The CWL standards impose no technical constraints on file sizes processed or parallel tasks run, with major scalability bottlenecks being hardware-related. As technology advances, CWL standards should keep pace without limiting capabilities.

The CWL ecosystem, developed over the past six years, comprises tools, software libraries, connected specifications, and shared CWL descriptions for popular tools. For example, software development kits (SDKs) for both Python and Java are generated automatically from the CWL schema, enabling programmers to load, modify, and save CWL documents using an object-oriented model corresponding to the standards themselves. Other languages' CWL SDKs can be developed by extending the code generation routines.

Galaxy

Galaxy is a versatile open-source bioinformatics platform designed to facilitate the analysis and interpretation of complex biological data. Developed by a collaborative community of researchers, Galaxy is designed to address the challenges faced by life scientists in the era of high-throughput genomics, proteomics, and other histological data[29]. By providing a comprehensive set of tools and resources in a user-friendly interface, Galaxy facilitates a collaborative and reproducible approach to biological data analysis, ultimately contributing to the advancement of scientific knowledge.

One of Galaxy's key features is its modular architecture. The architecture allows users to create and share custom workflows that can be customized to address specific biological questions or experimental designs. By integrating numerous analytical tools and resources, Galaxy allows for

the seamless processing of data across multiple steps and formats. This supports both novice and experienced researchers in gaining meaningful insights from complex datasets.

The Galaxy web user interface offers a comfortable setting for conducting data analysis. However, its utility can be challenged when it comes to handling repeated or looped tasks. Here, Bioblend, a tool designed to interface with Galaxy's API, becomes invaluable[30].

An Application Programming Interface (API) is a set of protocols established by a software program, outlining how it can be operated by an external program. Galaxy offers a comprehensive API that facilitates developers in accessing its functionalities through high-level scripts. Notably, the Galaxy user interface is transitioning to be fully based on the Galaxy API.

Currently, there are three methods to interact with the Galaxy API. The first method involves plain HTTP requests, which can be cumbersome. The second method uses an object-oriented package, which, however, is still under development. The third method is using Bioblend. Bioblend is the superior choice for bioinformaticians seeking to automate extensive data analyses from the ground up. This tool enhances collaboration by empowering users to both establish the necessary infrastructure and automate complex analyses over vast datasets within the familiar Galaxy environment.

Utilizing Bioblend offers numerous advantages. It allows for programmatic interaction with the server and offers complex control functionalities like branching and looping, which are currently not achievable in workflows. Moreover, Bioblend helps automate repetitive tasks and promotes integration with external resources.

Comparison of workflow management systems

Among Snakemake, Nextflow, and CWL, Galaxy emerges as an excellent choice for managing workflows that involve the tools mentioned above for several reasons. This selection is primarily influenced by a few distinguishing attributes of Galaxy.

Firstly, Galaxy is distinguished by its user-friendly interface. In contrast to many other workflow management systems that necessitate advanced technical skills, Galaxy's interface caters to users with varying levels of expertise. The construction of workflows is simplified through intuitive drag-and-drop operations, making it accessible even to novices in the field. Moreover, the visualization capabilities of Galaxy further enhance its usability, enabling clear representation and understanding of the constructed workflows.

Secondly, Galaxy stands out for its robust support system, especially when compared to platforms like Snakemake. While both platforms facilitate various bioinformatics tools, what differentiates Galaxy is its extensive community support. This community backing allows users, including those who may not have advanced expertise, to readily adopt and adapt standardized workflows for their data. For instance, Galaxy integrates widely used tools such as QIIME 2, predominantly for metabarcoding analyses, and Kraken 2 for shotgun data analyses. The seamless integration of these tools within Galaxy's framework simplifies the execution of intricate bioinformatics analyses, thereby promoting efficiency and accuracy in research workflows. Conversely, many alternative systems may necessitate a deeper level of expert knowledge for effective interaction.

Lastly, an integral feature of Galaxy that lends it an edge is the provision of Bioblend, a Python library for Galaxy and BioMart. Bioblend acts as a facilitator for automating extensive data analyses, and simplifying the management of large and complex bioinformatics projects. The integration of a Python-based library within Galaxy's ecosystem offers researchers a powerful tool for scripting and automating data-intensive tasks, enhancing productivity and the reproducibility of analyses.

In summary, Galaxy, with its user-friendly interface, broad tool support, and the ability to automate extensive data analyses via Bioblend, emerges as an excellent choice for managing workflows in bioinformatics. It offers a balanced combination of power, flexibility, and ease of use, making it a reliable and robust choice to be used for workflow implementation.

2.2.3 Microbiome databases

After executing workflows on the data using workflow management systems, microbiome results, and collected data samples can be processed for storage and presentation in a microbiome database. Microbiome databases store microbial community data, metadata, and results of microbial community analyses. These databases are essential to deepening our understanding of microbial ecology, functional dynamics, and their impact on human health and environmental sustainability.

A well-structured microbiome database should contain the following:

Sequence data The database should aggregate and organize various types of sequence data, including 16S rRNA gene sequences, macro-genomic and macro-transcriptomic data. This facilitates comparative analysis and allows researchers to identify taxonomic and functional patterns within microbial communities.

Metadata Comprehensive metadata should accompany sequence data detailing relevant information about the sample, such as collection method, location, time, and environmental conditions. This context allows for a deeper understanding of the microbial community and its interactions with the environment.

Accessibility Microbiome databases should ensure accessibility, organization, and standardization of data. This facilitates comparisons between studies and increases the reproducibility and transparency of scientific research.

Visualizations Visualizations of the data should be provided. This allows researchers to gain meaningful biological insights more easily instead of just plain text-based results.

Collaboration and data sharing Microbiome databases should foster a collaborative research environment by providing mechanisms for data sharing and submission that facilitate the exchange of information and advance scientific knowledge.

In summary, microbiome databases are a valuable resource for bioinformatics research, providing centralized access to a large number of selected microbial data and essential analytical tools. By continuously developing and expanding these databases, researchers can deepen our understanding of the intricate relationships between microbial communities and their environments.

2 popular microbiome databases are outlined below: (1) Mgnify[31] and (2) BacDive[32].

Mgnify

MGnify is a robust, free-access platform from the European Molecular Biology Laboratory-European Bioinformatics Institute (EMBL-EBI) that facilitates the comprehensive exploration, analysis, and preservation of diverse microbiome datasets. Users can access the data under (`https://www.ebi.ac.uk/`). This includes metagenomic, metatranscriptomic, metabarcoding, and assembly data. Essentially, it serves as a hub where researchers submit their data, which MGnify then processes using standardized pipelines for taxonomic and functional analysis[31].

The platform provides users with the capacity to submit their microbiome studies for analysis, to navigate through a wide variety of analyzed microbiome studies, and to visualize and download the results of these analyses. Furthermore, users can access raw data from the European Nucleotide Archive (ENA). MGnify structures its data according to projects, samples, runs, and analyses, thereby ensuring an orderly and easy-to-navigate resource for microbial genomics research.

MGnify boosts advanced search and browsing functions, along with a comprehensive API for complex inquiries and data downloads. This enables users to search for particular organisms or proteins with certain functions and to switch seamlessly between projects, samples, and runs. Moreover, MGnify provides a search function within its vast non-redundant database of predicted proteins.

One key feature of MGnify is its utilization of standardized analysis pipelines and its close cooperation with the European Nucleotide Archive (ENA). This affords a context for interpreting results in relation to other datasets and provides a platform for researchers to analyze their pre-publication data. Additionally, MGnify supports the assembly and reanalysis of any pertinent public dataset from the International Nucleotide Sequence Database Collaboration (INSDC) initiative.

The team behind MGnify is consistently working to expand and enhance the platform to accommodate the ever-evolving field of microbiome research. Their most recent updates are geared towards simplifying access to MGnify's analyses and derived data products. These include improvements to the website and its associated API, the provision of advanced analysis options directly from the web pages, and a significant revamp of the MGnify protein database.

However, it has its limitations when compared to a dedicated beer microbiome database since MGnify is a general platform for microbiome data. A dedicated beer microbiome database would provide more extensive metadata relating to beer production parameters, such as geographical origin, beer types, and breweries that are interesting for understanding the influence of these factors on the microbiome composition and dynamics.

BacDive

BacDive(`https://bacdive.dsmz.de`) - the Bacterial Diversity Metadatabase - is a comprehensive resource providing information linked to bacterial and archaeal biodiversity. The database covers a wide spectrum of data, including taxonomy, morphology, physiology, environment sampling conditions, and molecular biology. As of its 2021 update, BacDive hosts information on 82,892 strains. A majority of the data is manually curated and annotated, ensuring reliable information for researchers[32].

A unique feature of BacDive is its user-friendly dashboard(`https://bacdive.dsmz.de/dashboard`). This feature, designed to guide newcomers, provides statistical overviews on various aspects of the data, including total strains, species count, average cell size, and optimal environmental conditions. In addition, the dashboard includes an interactive display of 49 graphs across eight categories. Clicking on specific data points initiates a search in the advanced search, the isolation source search, or the TAXplorer, making the dashboard an effective launchpad for data exploration.

Additionally, BacDive has introduced RESTful web services, offering immediate access to machine-readable data in XML and JSON formats. These services enhance access to BacDive's data and improve linkage to external life science web resources. Moreover, the recent update introduces new data fields and features such as gene name search, plasmid search, and 16S rRNA search.

Despite its strengths, BacDive does have certain limitations when compared to a dedicated beer microbiome database. Notably, BacDive concentrates solely on bacterial diversity and does not include the fungal microbiome, which is an integral part of the beer microbiome. Additionally, similar to MGnify, BacDive is designed for general bacterial biodiversity and lacks a specific focus on the beer microbiome. Consequently, it may not provide in-depth metadata tailored to beer.

2.2.4 Open science

Apart from the above necessary steps of microbiome analysis, and different workflow management systems, the analysis in the thesis follows the spirit of Open Science. Since the metabarcoding and metagenomics workflows to analyze the beer microbiome data will be on an open-source platform which will be introduced later. And all the data, code, and deliverables for this thesis will be publicly available on GitHub as well, which welcomes the participation of other enthusiasts.

It is an innovative movement that seeks to promote the free and accessible dissemination of scientific knowledge and has gained substantial traction in the research community[33]. The principles of Open Science, including collaboration, reproducibility, transparency, and accessibility, hold immense potential to transform the landscape of scientific research across various disciplines. Bioinformatics, an interdisciplinary field that combines biology, computer science, mathematics, and statistics to analyze and interpret biological data, is particularly well-suited to benefit from the adoption of Open Science practices.

Reproducibility

Reproducibility, the ability to recreate and validate the findings of a study, is fundamental to the scientific method. In bioinformatics, the reproduction of results often requires access to the original data, computational tools, and detailed documentation of the analyses performed. Open Science promotes reproducibility by advocating for open access to data, software, and protocols, enabling researchers to independently verify and build upon the work of others. Implementing

reproducible research practices in bioinformatics not only strengthens the credibility and reliability of scientific findings but also minimizes the duplication effort, and thereby enhances the overall efficiency of the research process.

Collaboration

Collaboration is the cornerstone of scientific progress, allowing researchers from diverse backgrounds to pool their expertise, knowledge, and resources to solve complex problems. In bioinformatics, interdisciplinary collaborations are crucial, as they enable the development of novel algorithms, computational tools, and statistical methodologies that facilitate the analysis of large-scale biological data. Open Science fosters a collaborative environment by encouraging data sharing, cross-disciplinary interactions, and the formation of global research networks. By embracing collaboration, bioinformatics researchers can address the challenges posed by the increasing complexity of biological data and accelerate the discovery of novel insights.

Transparency

Transparency in research encompasses the clear and comprehensive reporting of methodologies, data, and results, ensuring that the scientific process is open to scrutiny and evaluation. Bioinformatics is inherently dependent on complex data and computational analyses, which may inadvertently introduce biases or errors. Open Science emphasizes the importance of transparent reporting, including the provision of raw data, detailed descriptions of analytical pipelines, and the use of open-source software, facilitating critical assessment and enabling further research. Greater transparency in bioinformatics strengthens the validity of research findings and contributes to the development of robust methodologies and best practices.

Accessibility

Accessibility is a key principle of Open Science, promoting the democratization of knowledge by ensuring that research outputs are freely available to all, regardless of geographical location or institutional affiliation. In bioinformatics, access to data, tools, and computational resources is essential for the generation of new knowledge and the advancement of the field. By advocating for open-access publication, data sharing, and the development of open-source software, Open Science endeavors to remove barriers to knowledge dissemination and foster a more inclusive research community. Increased accessibility in bioinformatics empowers researchers from diverse backgrounds and resource settings to contribute to the global body of knowledge, fueling scientific innovation and discovery.

2.3 Beer Microbiome

Beer, likely the oldest and most widely consumed alcoholic beverage globally, traces its origins back to the Middle East and Egypt [34]. Historical records reveal detailed accounts of mass production in ancient Babylonian Mesopotamia around 1800 BC [35].

As a testament to its cultural significance, beer continues to play a pivotal role in daily life across various societies. The brewing process has evolved over time, with innovations and regional variations contributing to the development of diverse beer styles.

Comprising more than 90% water, beer also contains carbohydrates and alcohol, which, when metabolized in the human body, release a certain amount of energy. The alcoholic content in various beer types varies, typically falling within an estimated range of approximately 3.5 to 10% weight/volume (w/v) [36]. In Western societies, malted barley serves as the primary ingredient in beer production, although other grains such as wheat, corn, and rice can also be used. Beer can be broadly classified into two main categories: ales and lagers. While the differences between these categories might appear subtle, they have a crucial impact on the beer's characteristics, including flavor, aroma, and appearance.

Regardless of the specific type, all beers undergo the same essential steps in the brewing process. Each stage plays a vital role in shaping the beer's final profile, allowing brewers to create a wide range of flavors and styles to cater to diverse consumer preferences.

The beer microbiome is a complex ecological system. It consists of different microbial communities. These communities play a vital role in the production, flavor, and stability of beer. Microorganisms such as bacteria and fungi contribute to the beer fermentation process. In recent years, advanced bioinformatics techniques have helped analyze these microbial communities. This allows scientists to better understand their dynamics and optimize the brewing process.

Beer bacterial microbiome

The bacterial component of the beer microbiome includes various *lactic acid bacteria* (LAB) and *acetic acid bacteria* (AAB). These microorganisms produce different organic acids. These acids contribute to the beer's sourness and overall flavor profile [3]. LAB mainly comes from the genera *Lactobacillus* and *Pediococcus*. They produce lactic acid during fermentation. AAB primarily comes from the *Acetobacter* and *Gluconobacter* genera. They are involved in the synthesis of acetic acid. The balance of these bacterial populations is essential. It helps develop desired flavor profiles and maintain product stability.

Beer fungal microbiome

The fungal component of the beer microbiome consists mostly of yeasts. *Saccharomyces cerevisiae* and *Saccharomyces uvarum* are the most prevalent species used in brewing. Yeasts have a crucial role in the fermentation process. They metabolize sugars to produce ethanol and carbon dioxide. This shapes the beer's alcoholic content and carbonation levels.

Aside from *Saccharomyces*, other types of yeast also play crucial roles. To illustrate, *Kluyveromyces* and *Brettanomyces* species are often detected in lambic and gueuze beers. Notably, *Brettanomyces bruxellensis* has a significant impact by producing the unique aroma that develops during the beer's maturation phase[37].

2.3.1 Beer fermentation

Conventionally, beer production techniques are classified into two groups: bottom-fermenting (lager-fermenting) and top-fermenting (ale-fermenting) processes. By broadening this classification to encompass mixed fermentations, two additional categories can be integrated: non-spontaneous fermentation and spontaneous fermentation [38].

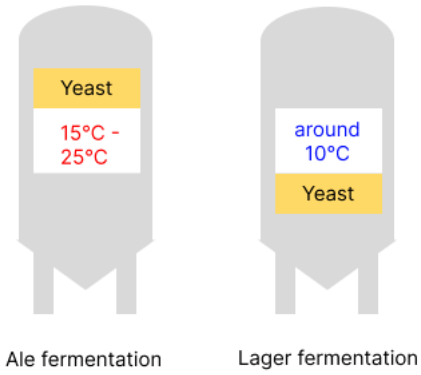


Figure 2: Two types of beer fermentation
Temperature in Ale (Top) fermentation ideally is between 15°C to 25°C and temperature in Lager (bottom) fermentation ideally is between 6°C and 8°C which is colder compared to Ale fermentation[39].

Top fermentation

Top fermentation, also referred to as Ale Fermenting, entails introducing yeast directly onto the wort, where it starts fermenting from the top down. The yeast employed in ale fermentation is *Saccharomyces cerevisiae*, which is also utilized in wine and bread making.

The ideal fermentation temperature for ale ranges from 15°C to 25°C. Most yeast types perish if the temperature exceeds 40°C [39]. The yeast used in lager fermentation can tolerate higher alcohol concentrations, which is why top-fermented beers generally have higher alcohol content than bottom-fermented beers. This yeast accelerates the fermentation process, allowing ale to be ready for consumption within a week. In general, top-fermented beers exhibit more robust flavors.

Bottom fermentation

In bottom fermentation or Lager Fermenting, the yeast in the wort starts working from the bottom up. The yeast typically employed for lager fermentation is *Saccharomyces uvarum*. The optimal temperature range for lager fermentation lies between 6°C and 8°C [39].

Lager yeast becomes inactive when alcohol content surpasses a certain threshold, resulting in lager generally having a lower alcohol content than ale. The fermentation process is considerably slower than that of top fermentation beers, yielding a clearer and more refreshing beer.

Spontaneous fermentation

Spontaneous fermentation is an ancient and natural method of beer production, harnessing the power of wild, ambient microorganisms, predominantly yeast, and lactic acid bacteria, to convert the fermentable sugars in the wort into alcohol, carbon dioxide, and various flavor compounds. This uncontrolled and serendipitous process relies on the unique microbial ecosystem of a given environment, imparting distinctive, complex, and often unpredictable flavors to the resulting beer. The most notable example of spontaneously fermented beer is the Belgian Lambic, a sour and wild-fermented ale that reflects the rich microbiota of the Senne River Valley[40]. Spontaneous fermentation is revered for its ability to capture the essence of a specific terroir and its innate capacity to produce unparalleled flavor profiles. However, due to the uncontrollable nature of wild microorganisms, the process can yield inconsistent results and may require lengthy maturation periods.

While wild yeast, lactic acid bacteria (LAB), and certain other bacteria are typically deemed contaminants in many beer fermentation procedures, these same microorganisms are often desired in the crafting of specific sour beers and wild beer styles. Currently, the international beer market is witnessing a renewed interest in sour beers, as breweries of various sizes globally are exploring new product varieties and complex flavors[3]. For instance, American Coolship Ales, a beer undergoing a complex series of spontaneous fermentations, is gaining popularity for its distinctive flavor profile[41].

This trend underscores the role of these traditionally "undesirable" microorganisms in creating novel and intriguing brews. More and more interest has been focused on identifying these microorganisms.

Non-spontaneous fermentation

In contrast, non-spontaneous fermentation is a controlled and deliberate process that employs selected strains of yeast and, occasionally, bacteria to ferment the wort. This method allows brewers to manipulate various parameters such as fermentation temperature, yeast strain, and wort composition, thereby enabling the production of a vast array of beer styles with consistent and reproducible characteristics. Non-spontaneous fermentation is the predominant method used in modern brewing, given its capacity for precision and the ability to tailor specific flavor profiles to cater to consumer preferences. Examples of beers produced through non-spontaneous fermentation include lagers, and ales, each showcasing distinct flavors and attributes attributed to the particular yeast strains employed and the conditions under which fermentation occurs[42].

2.3.2 Beer microbiome studies overview

In this thesis, we have picked five studies related to beer microbiome analysis. Investigating these previous studies is beneficial for our beer microbiome analysis because it allows us to identify and understand the microbial compositions documented in earlier research and we can identify common trends and patterns. By reproducing the results using the data from the studies and comparing them with the original results, we can have a better understanding of our workflows' capability and performance. Among these, we have chosen to focus in detail on three studies that exhibited a higher variety of beers analyzed and were also among the most recent. They are

BeerDeCoded: the open beer metagenome project[4], Bacterial and Fungal Dynamics During the Fermentation Process of Sesotho, a Traditional Beer of Southern Africa[43], Characteristics of bacterial and yeast microbiomes in spontaneous and mixed-fermentation beer and cider [2], A Culture-Independent Comparison of Microbial Communities of Two Maturating Craft Beers Styles[44] and Description of the temporal dynamics in microbial community composition and beer chemistry in sour beer production via barrel aging of finished beers [3]. 3 of the 5 studies mentioned had a higher variety of beers, so they are discussed in detail here.

BeerDeCoded: the open beer metagenome project

The BeerDeCoded project analyzed the targeted metagenomic profile of 39 bottled beers using ITS sequencing of fungal species. These 39 commercial beers originated from 5 different European countries: 30 were from Switzerland, five from Belgium, two from Italy, one from France, and one from Austria.

After extraction and sequencing the ITSs, a refined set of ITS sequences from the RefSeq database[45] (Targeted Loci) was utilized to construct an ITS index for the `Burrows-Wheeler Aligner (BWA)` [46]. Using standard parameters, the `BWA` was applied to align the beer sample reads, which are stored in `FASTQ` format to this ITS index. Subsequently, the files were sorted and indexed using `samtools` [47]. These BAM files underwent a quality control assessment with the aid of `SAMstat`[48]. To ensure accuracy and to remove low-quality, non-uniquely mapped reads, a minimum mapping quality (MAPQ) score of 3 was set as the threshold. Following this, the count of ITS per beer and per species was calculated, with only species having more than 10 reads considered for further analysis.

The analysis revealed 42 distinct fungal species, intriguingly, 24 of these species were exclusive to just one type of beer. The extensive diversity of wild yeasts present in commercial beers was unanticipated, with some beers exhibiting evidence of containing over 10 different fungal species.

Waldbier 2014 Schwarzkiefer, an Austrian beer incorporating pine cones sourced from local forests in its brewing process, stood out with the highest internal transcribed spacer (ITS) diversity, comprising 19 fungal species. In addition, two other beers both showed over 12 fungal species: La Nébuleuse Cumbres Rijkrallpa, a sour/wild ale that incorporates cranberries and fermented corn known as "Chicha", and Chimay Red Cap, a traditional Belgian Trappist beer.

Bacterial and Fungal Dynamics During the Fermentation Process of Sesotho, a Traditional Beer of Southern Africa

Sesotho, a widely consumed spontaneously fermented beer in Lesotho, Southern Africa, is brewed from milled maize, sorghum, or wheat flour, sometimes a blend of these. This beer is recognized for its cloudy appearance, light body, and characteristic sour taste. The goal of this study was to scrutinize the microbial diversity, covering both bacterial and fungal species, across five distinct stages of Sesotho fermentation. This was done at five unique locations in Lesotho, using Next-Generation Sequencing (NGS) methodologies.

Ensuring data integrity and precision, all data sets underwent initial processing and trimming, with the aim of achieving an average quality score of at least 20. Sequences shorter than 200

base pairs were excluded. Subsequently, paired-end reads were amalgamated. A demultiplexing and quality filtering script in `QIIME` was executed with default parameters, producing a FASTA output file. Chimeric sequences, or abnormal sequences composed of two distinct sequences, were recognized and removed. Then representative Operational Taxonomic Units (OTUs) were taxonomically classified, adhering to a 97% sequence identity standard against the SILVA 132 database for the 16S rRNA bacterial data, and the UNITE database for the fungal ITS data.

For the bacterial results, 9,885 bacterial OTUs were identified, with individual samples containing between 600 and 2,543 OTUs. *Proteobacteria* and *Firmicutes* emerged as the dominant bacterial phyla across all samples and locations. Regarding fungi, 46 OTUs were detected at the same sequencing depth across all samples, with individual samples yielding between 5 and 13 OTUs. The prevailing fungal phyla across all samples and locations were *Ascomycota* and *Mucoromycota*, with *Saccharomyces spp.* being an example of the Ascomycota. This research offers valuable insights into the microbial ecology involved in Sesotho beer fermentation.

Characteristics of bacterial and yeast microbiomes in spontaneous and mixed-fermentation beer and cider

This study concentrates on the microbial communities within 14 commercially available Russian beers, known for their mixed or spontaneous fermentation. These beer samples were either acquired from retail stores or directly supplied by the manufacturers. The analysis hinged on High Throughput Sequencing of 16S rRNA and ITS regions from these beer samples.

The data was processed utilizing the `Knomics-Biota system`, which was enhanced by authors to scrutinize the prokaryotic and yeast microbiomes associated with food products. Firstly, the reference databases were expanded to accommodate 16S rRNA sequencing data from any niche. Secondly, a new module was incorporated for analyzing the fungal ITS region. This module is based on the UNITE database [49], utilizes the Deblur algorithm [50] for amplicon-sequence variant (ASV) identification, and employs the `QIIME2 naive-Bayes classifier`.

Given the diversity in the samples, which encompassed mixed-fermentation beer, it was anticipated that the yeast composition would not be restricted to *Saccharomyces cerevisiae*. As hypothesized, the microbiome was primarily influenced by the balance between *Dekkera* (also known as *Brettanomyces*) and *Saccharomyces*, while other less prevalent contributors included *Issatchenkia*, *Pichia*, *Hanseniaspora*, and *Candida*. On a species level, *Dekkera* was predominantly represented by *D. bruxellensis*, with minor presence of *D. custersiana* and *D. anomala*. *Saccharomyces* was chiefly represented by *S. cerevisiae* and *S. bayanus/pastorianus* among others. On the other hand, the bacterial community was predominantly composed of members from the *Lactobacillales* order. Other bacterial families were generally found in low abundances in the beer samples, except for three instances where the *Enterobacteriaceae*, *Leuconostocaceae*, and *Acetobacteraceae* families were present in considerable (>5%) quantities.

Summary of prior beer microbiome studies

These 3 studies focus on different types of beer, but all studies employ metagenomic profiling, targeting fungal ITS and bacterial 16S rRNA regions, to investigate the microbial communities in beers. And the studies utilize similar data analysis steps just like discussed previously, such as

quality control, taxonomic profiling, and diversity analysis & visualization. Across the studies, the predominant fungal phylum observed is *Ascomycota*, within which *Saccharomyces spp.* are identified as key contributors to the beer microbiome. This genus is known for its pivotal role in the fermentation process. The studies report *Proteobacteria* and *Firmicutes* as dominant bacterial phyla, with members of the *Lactobacillales* order frequently detected. These bacteria are responsible for the production of lactic acid and other metabolites that impact beer flavor profiles.

3 Methods

This section provides a comprehensive overview of the methodologies employed in various stages of our research. These stages are i) collection of beer microbiome data samples; ii) implementation, evaluation, and automation of bioinformatics analysis workflows for the data and reproducibility analysis; and iii) the construction of the beer microbiome database, BeerMicroDB, which was designed to serve as a comprehensive overview of beer microbiome data and the reproduced results using the workflows created.

3.1 Data collection

In the beginning, due to uncertainty about the adequacy of beer-related data samples, the decision was made to collect samples not only of beer but also of cider, Kombucha, kefir, and wine. The procedure for data collection was bifurcated. The first avenue involved seeking out research articles on the microbiome of fermented drinks via Google Scholar. The availability of public sample data was then verified for each paper. If such data was available, it was added to a spreadsheet along with relevant metadata. The second method involved a direct search for microbiome samples related to fermented drinks in databases such as ENA or MG-RAST [6]. Any suitable datasets found this way were also added to the spreadsheet. In total, 768 samples were collected.

Once data collection was completed, the Python libraries `pandas` and `matplotlib` were used in a `Jupyter Notebook` to graphically represent the data, facilitating a preliminary understanding of its composition.

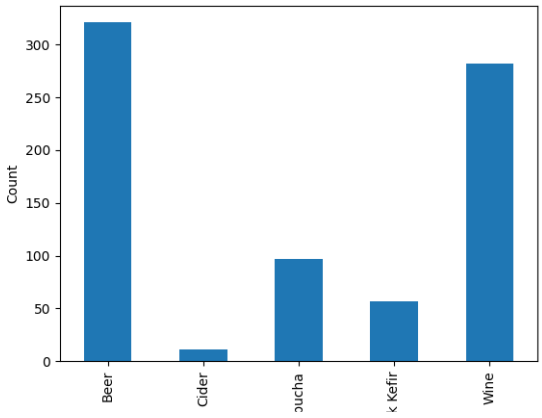
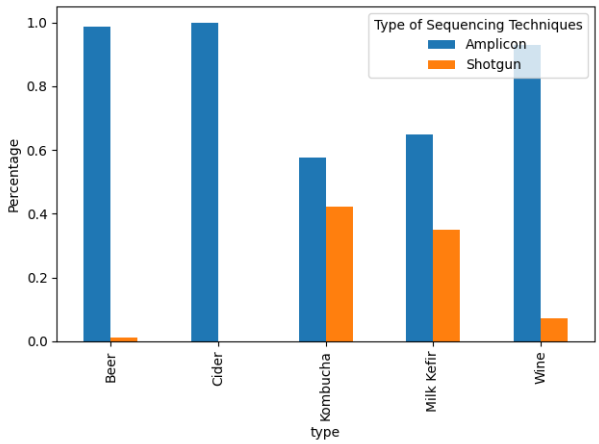


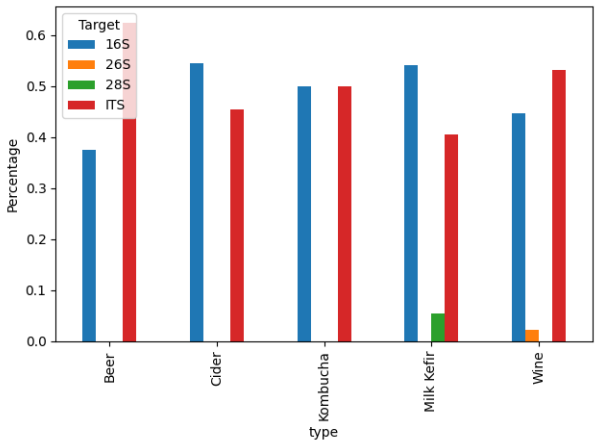
Figure 3: Overview of fermented drinks
Number of data samples by fermented drink types, which consist of Beer, Cider, Kombucha, Milk Kafir, and Wine.

To deepen our understanding of the collected data, we first examine the number of samples associated with each type of fermented drink shown in figure 3. Subsequently, to enhance our insight, we explore how these samples are distributed across various extraction techniques, such as shotgun and metabarcoding techniques. For the metabarcoding approach, we pay special

attention to the distribution of sequencing targets, including but not limited to ITS and 16S shown in figure 4.



(a) Data composition by sequencing techniques



(b) Metabarcoding data composition by targets

Figure 4: Data composition by sequencing techniques and metabarcoding targets

Through meticulous data collection, we have assembled a substantial array of beer samples. Our observation indicates that the majority (higher than 95%) of our metabarcoding data can be categorized as either ITS or 16S. The beer sample did not even have 26S and 28S data. With this in mind, and given the adequate amount of data for the beer samples, which is over 300, we have chosen to focus only on these beer samples at the moment.

While we're concentrating on beer samples now, we're open to the prospect of incorporating data on other fermented drink types in the future. Our current database comprises 317 beer samples. We found that a small subset of these samples (16 to be precise) are housed on the MG-RAST platform, in the form of FASTA files, while all other data samples are in the form of FASTQ files, for keeping consistency, we decided to omit these from our immediate analysis.

In conclusion, our study will concentrate on the remaining 301 beer samples. The volume of data contained within these samples establishes a robust foundation for our continuing investigations. We have made the scripts for plotting the figures, as well as information about the datasets, accessible to the public. These resources can be accessed at the following URL: `https://github.com/asdsd/sdfdfs/blob/main/data_plot_n_processing/data_plot.ipynb`.

3.2 Microbiome analysis workflows

This section shows the implementation and evaluation of the workflows used in the thesis and also the methodologies and tools used in the reproducibility analysis.

3.2.1 Implementation of workflows

This section illustrates the implementation of metabarcoding and shotgun workflows accordingly.

Metabarcoding workflow

Based on the characteristics and comparison of state-of-the-art tools for analyzing metabarcoding data discussed in the section "State of the art: Data Analysis" we use QIIME 2 as the workflow for analyzing metabarcoding data.

QIIME 2 Set up on Galaxy Europe

QIIME 2 was available as 109 Galaxy wrappers and initially not available on Galaxy Europe and therefore needed to be installed. This involved extracting the QIIME 2 Galaxy tools from the qiime2/galaxy-tools GitHub repository, which contains Official QIIME 2 tools for Galaxy, by using scripts designed to extract all tool IDs into a YAML file and subsequently merging them into the usegalaxy-eu/usegalaxy-eu-tools repository.

After completing this process, the QIIME 2 tools can be accessed through the Galaxy Europe instance. However, to ensure that these tools are executed using docker when invoked, the associated tool IDs must also be added to the job_conf.yml file in the usegalaxy-eu/infrastructure-playbook repository.

Thus, with the above procedure, the QIIME 2 tools have been successfully configured on the Galaxy Europe instance. The scripts for setting up QIIME 2 on Galaxy Europe can be accessed at the following URL: `https://github.com/YedilSerzhan/beer_microbiome/tree/main/qiime2_tools_crawler`.

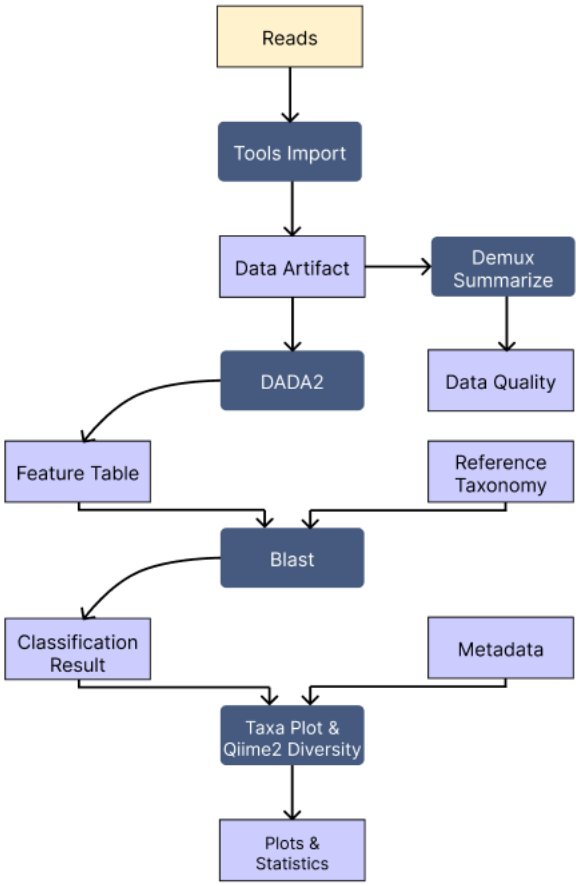


Figure 5: The metabarcoding workflow
The metabarcoding workflow is implemented using QIIME 2. The steps contain data importing, DADA 2 denoising, taxonomy profiling using BLAST, and visualization & diversity analysis. (Elements with dark blue color are for tools and purple color are for results of the intermediate steps)

Data importing

Firstly data files from ENA should be uploaded to Galaxy using the upload function. However, all QIIME 2 tools interact with artifacts instead of traditional data files like FASTQ or FASTA files. A QIIME 2 artifact is a file with .qza extension. In order to convert the input data to an artifact, the tool `qiime2 tools import` needs to be used. There is one parameter "Type of data to import" that needs to be set. For a FASTQ data file, we should choose the "SampleData[PairedEndSequencesWithQuality]" option. Then we need to specify the data file and the file format to import from. Normally in the QIIME 2 CLI version, we can choose the files and format by creating a manifest file, which declares the name and location of each data file to be imported. However, in Galaxy, the location of uploaded data files can't be accessed for security reasons. So in QIIME 2 Galaxy version, we can do it by choosing the option "Casava One Eight Single Lane Per Sample Directory Format" and renaming each data file to the format like "L2S357_15_L001_R1_001.fastq.gz".The underscore-separated fields in this file name are (1) the sample identifier (2) the barcode sequence or a barcode identifier (3) the lane number

(4) the direction of the read (i.e. only R1, because these are single-end reads) and (5) the set number. This certainly has set some limitations for the variety of types of data to process, for example, FASTA files can not be properly processed using QIIME 2 Galaxy.

Quality control and feature table construction

After importing the data, the quality of the data can be accessed using the tool `qiime2 demux summarize`, which can provide an interactive plot of the quality of the data. Next depending on the data quality information, parts of the data can be truncated or trimmed.

The quality control and amplicon sequence variant (ASV) feature table construction can be done using the tools `DADA2`. Deblue is also available for feature table construction but it needs additional quality filtering steps, so dada2 is chosen here. Rather than adopting the conventional approach of "OTU-picking" that is commonly used in metabarcoding workflows, DADA2 [22] introduces a unique algorithm that simulates the inaccuracies generated in the course of metabarcoding. By employing this error model, the algorithm can deduce the actual composition of a sample. It generates tables of amplicon sequence variants (ASVs), which provide a higher level of resolution in comparison to traditional methods. Based on whether the data is paired or not, `qiime2 dada2 denoise-single` or `qiime2 dada2 denoise-paired` will be used.

Options: For DADA2 denoising, depending on the quality of the data, a few parameters can be set: 1. "trunc_len": Position at which sequences should be truncated due to a decrease in quality. 2."trim_left": Position at which sequences should be trimmed due to low quality 3. "max_ee": Reads with a number of expected errors higher than this value will be discarded. This procedure takes normally around 10 minutes depending on the size of the data. These options should be set depending on the quality information gotten from the result of `qiime2 demux summarize`.

Outputs: The outputs consist of 3 files. 1. "feature table": The resulting feature table is a tabular representation of the data that maps samples to ASVs. 2. "representative sequences": The resulting feature sequences. Each feature in the feature table will be represented by exactly one sequence. 3. "denoising stats": stats about the denoising procedure like how many reads are removed, the percentage of reads that are preserved, and so on. This information can help you have an understanding of how well the denoising is conducted and whether you want to change the parameters and do it again.

Visualization for outputs: In order to view the output results, we can use the following tools. For a feature table, we can use `qiime2 feature-table summarize` to generate visual and tabular summaries of a feature table. For the denoising stats, we can use `qiime2 metadata tabulate` to generate a visualization that shows a tabular view of data. And it supports interactive filtering, sorting, and exporting to common file formats.

Taxonomy profiling

In this step, we'll perform annotation of the features that were observed in the last denoising step by performing taxonomic classification of the sequences. The tool used in the workflow is `qiime2 feature-classifier classify-consensus-blast`, the function of which is assigning taxonomy to query sequences using BLAST+. Performs BLAST+ local alignment between query

and reference reads, then assigns consensus taxonomy to each query sequence from among max accepts hits, min consensus of which share that taxonomic assignment.

Inputs and reference database: The required inputs for this tool are 1. Query sequences. 2. Reference sequences. and 3. reference taxonomy labels. The query sequences come from the output representative sequences of DADA 2 denoising. The reference sequences and reference taxonomy labels come from a reference database that we can choose. Until now, for QIIME 2, the available options are UNITE for fungal ITS, Silva for 16S/18S rRNA, and Greengenes for 16S rRNA. In our case, we stick to UNITE for fungal ITS and Silva for 16S rRNA.

Optional parameters: There are 4 important optional parameters that can be set for consensus-blast. (1) maxaccepts: the maximum number of hits to keep for each query. The first N hits will be chosen in the reference database that is similar to the query the default value is 10. (2) perc_identity: A float number from 0.0 to 1.0 to reject a match if the percent identity to query is lower. The default value is 0.8. (3) query-cov: A float number from 0.0 to 1.0 with a default value of 0.8. The algorithm rejects matches if query alignment coverage per high-scoring pair is lower than this value. (4) min-consensus: A float number ranging from 0.5 to 1.0 represents the minimum fraction of assignments that must match the top hit to be accepted as consensus assignments. The default value is 0.51. Throughout the processing of all data, most of them are set to use the default values.

Algorithm outputs: There are 2 outputs generated in the end. They are classification and search results. The classification file is the taxonomy classifications of query sequences. And the search results file shows the top hits for each query.

Both files can be visualized using the `qiime2 feature-table summarize` tool.

Diversity analysis and visualization

For a more comprehensive grasp of the results, we can employ diversity analysis to assess the richness and diversity within and among the samples. The use of visualization tools can further enhance our insight into the results, offering a more detailed and intuitive understanding of the data.

There are 2 tools that can be used to do the diversity analysis. The tool `qiime2 diversity alpha` is for computing a user-specified alpha diversity metric for all samples in a feature table. And tool `qiime2 diversity beta` can be used to compute a user-specified beta diversity metric for all pairs of samples in a feature table. The input is the feature table generated by the dada2-denoising tool. And we can choose a metric out of dozens of options.

Apart from visualizing the intermediate results from the previous steps using tools `qiime2 metadata tabulate` and `qiime2 feature-table summarize`. We can also use `qiime2 taxa barplot` to generate an interactive bar chart visualizing the taxonomic classification results.

The inputs are (1) Feature table: the feature table generated by the DADA2 denoising tool. And (2) Taxonomy: Taxonomic annotations for features in the provided feature table, which is the file generated by the tool `qiime2 feature-classifier classify-consensus-blast`.

The output file is a barplot.qzv file which can be visualized using the QIIME 2 View website to generate an interactive bar chart plot can include multi-level sorting from level 1 kingdom to

level 7 species, plot recoloring, sample relabeling, and SVG figure export. If a proper metadata file is also provided as input, you can also group and sort by the metadata columns.

Shotgun workflow

As we transition to the more comprehensive shotgun sequencing approach, which indiscriminately sequences all the DNA within a sample, we are met with a change in both the scope and the complexity of the analysis. Consequently, the workflow for shotgun sequencing requires a shift in the analytical tools and methods employed. This section talks about the workflow implementation for shotgun data.

Existing workflow

There is already a Galaxy tutorial available on workflows on identifying beer micro-organisms before this thesis (`https://training.galaxyproject.org/training-material/topics/metagenomics/tutorials/beer-data-analysis/tutorial.html`).

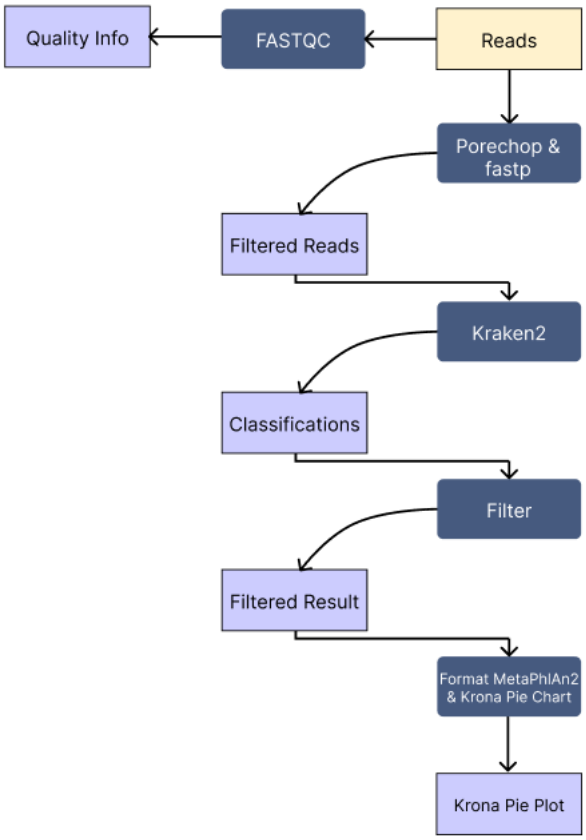


Figure 6: The existing shotgun workflow
The existing workflow initiates with quality control measures employing tools such as FASTQC, Oirechop, and fastp. Subsequently, taxonomic profiling is undertaken using Kraken 2. Once profiled, the filtered results are visualized utilizing the Format MetaPhAn2 and Krona Pie charts.

The established workflow illustrated in figure 6 begins with the utilization of `FASTQC` [14] to evaluate the quality of the reads. The resulting HTML file vividly portrays the quality scores distributed over all bases. `Porechop` [51] is subsequently employed to eliminate sequencing adapters and chimeras, or contaminants. Moreover, the tool `fastp` [52] aids in filtering sequences possessing low-quality scores. One can review the HTML report generated by `fastp` to ascertain how the data quality has been enhanced.

The subsequent step is taxonomic classification, for which we utilize `Kraken2`. The database incorporated in this phase is the "Prebuilt Refseq indexes: PlusPF". If the "Create Report" option in Kraken2 is selected, a report file encompassing detailed information is produced. The report's first column displays clades, extending from taxonomic domains (like Bacteria, Archaea, and so forth) to species. The second column discloses the number of reads allocated to the clade rooted at that particular taxon.

For visualizing the results, we employ `MetaPhlAn2` [20] to transfigure the report into the Krona format. Subsequently, the `Krona pie chart` tool can process the converted report file to construct a Krona pie chart. This chart is a multilayered pie chart that exhibits the community profile. The central part of the chart represents higher taxonomy levels (such as the domain), whereas the outer parts provide more detailed information (such as species).

Updates to the shotgun workflow

The existing workflow for shotgun sequencing analysis has been in place for an extended period. During this time, there have been significant updates to both the tools and databases involved in the process. Notably, the advent of the Kraken software suite [53] has necessitated further updates to the shotgun workflow.

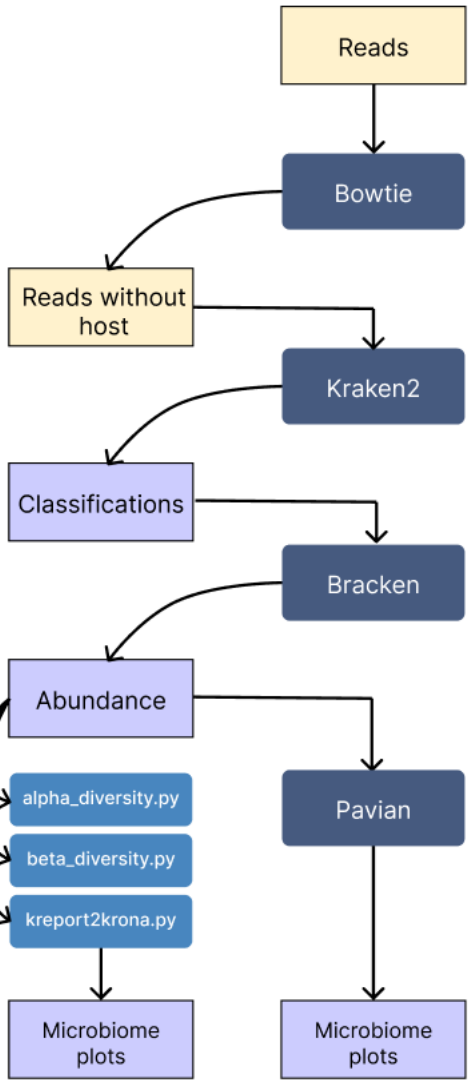


Figure 7: The shotgun workflow
Relative to the previous workflow, the updated procedure incorporates Bowtie2 at the outset to eliminate host DNA from the reads. Following the classification by Kraken 2, the data is processed by Bracken to yield abundance information. This information is subsequently visualized using Pavian and KrakenTools, as depicted by the three scripts presented in the figure.

The process of the shotgun workflow illustrated in figure 7 commences with the removal of human host DNA using Bowtie2. For some datasets collected, the host DNA is already removed, then this step can be skipped.

Then it is followed by the taxonomic classification of sequencing reads using `Kraken2`[21]. Each read is then allocated to a taxonomic group (species, genus, or higher-level taxa). The database used here is Standard plusPFP (Standard plus protozoa, fungi, and plant) version 2022-06-07. Optionally a report can be generated showing the aggregate counts of each classified result.

And then subsequent step involves calculating the relative abundance of different species within the sample. `Bracken` was devised to function alongside `Kraken 2` for species abundance computation, utilizing the classification results from `Kraken 2`. `Bracken` [54] then processes the classified read counts and estimates the abundance of each taxon in the sample.

In the final analysis, `KrakenTools` and `Pavian` [24] offer a comprehensive toolkit for downstream statistical analysis and visualization of the classification and abundance estimation outcomes.

`Pavian` can be employed to examine and visualize the sample to identify differences. In addition, `alpha_diversity.py` can be utilized to measure the diversity in a sample, and `beta_diversity.py` can be applied to compare diversity across samples. `kreport2krona.py` can convert the Kraken report into the Krona format, which can be visualized using `Krona` by generating a Krona plot.

`KrakenTools` are not available in the suite of Galaxy tools, hence there is a need to integrate them into the Galaxy EU instance. The process of doing so involves several steps:

1. Construct the Wrappers

`Planemo` is a command-line utility employed in creating Galaxy and Common Workflow Language artifacts, which encompass tools, workflows, and training materials. The procedure commences with the command planemo tool_init –id 'id' –name 'name', which generates an XML file containing significant tags such as requirements, command, inputs, outputs, and help.

The requirements tag should include the necessary packages to execute the tool that requires wrapping. For `KrakenTools`, this mainly pertains to Python.

The command tag houses the command line necessary to operate the tool, comprising input files, parameters, and outputs. These options will subsequently be defined in the inputs and outputs sections.

The inputs tag encompasses the tool's input, which will be displayed on the tool's page within Galaxy for users to select input files or parameters. The input encompasses several parameter types, including data, integer, float, and boolean values. Advanced tools may also permit the setting of conditional parameters.

The outputs tag involves the tool's output, which should also be encompassed within the commands.

The help tag should offer valuable information to assist new users in learning how to employ the tool.

Proper configuration is essential for creating a user-friendly interface that facilitates easy understanding of the tool. This consideration is not only vital for the current workflow but also for all researchers utilizing the Galaxy platform.

2. Construct Test Cases

A proficient tool wrapper should include tests to ensure the smooth operation of the tool. These tests can also be set within the XML file. Multiple tests can be constructed, each involving example inputs and example outputs.

`Planemo` utilizes the commands in the XML file to generate results using the inputs. If the generated result aligns with the anticipated outputs, it is declared that the tool is functioning

correctly. All potential edge cases should be taken into account in these tests. Importantly, the example input and output files should not exceed 1MB in size.

3. Submit the Pull Request on Tools-IUC

Tools-IUC (https://github.com/galaxyproject/tools-iuc) is a GitHub repository that contains a selection of Galaxy repositories used in the Tool Shed (https://toolshed.g2.bx.psu.edu/). These repositories are maintained and developed by the Intergalactic Utilities Commission. Tools-IUC maintains high standards for Galaxy tools, further ensuring the smooth operation of the tool.

4. Submit the Pull Request on Galaxy EU

Following submission to Tools-IUC, this does not automatically make the tools usable. They must also be added to a Galaxy instance, in this case, Galaxy EU. Hence, a pull request should also be submitted to the Galaxy Europe. This is achieved by adding all the tool ids to the job_conf.yml in the Galaxy infrastructure playbook, thereby instructing the Galaxy job dispatcher to pick up the tools.

Finally, alpha_diversity.py, beta_diversity.py, and kreport2krona.py are integrated into the Galaxy EU instance and are prepared for use.

3.2.2 Evaluation of workflows

Despite the user-friendly features of the Galaxy platform, executing workflows on a large dataset, such as the more than 300 beer samples mentioned earlier, can be a ton of work. Fortunately, through the use of Bioblend, a Python library designed for interfacing with the Galaxy API, it is possible to automate the execution of workflows. The automation process shown in figure 8 is carried out through several discrete steps:

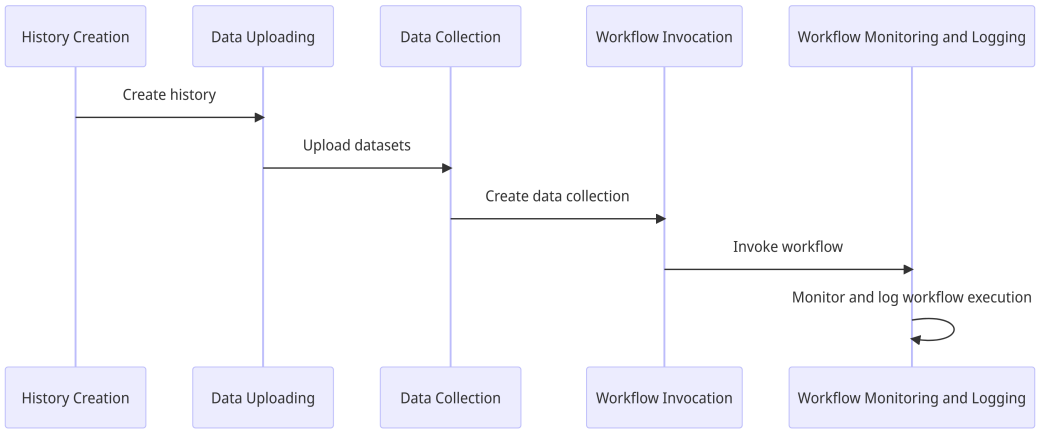


Figure 8: Automation of workflows execution
The figure illustrates the automation of workflow execution via Python scripts, leveraging Bioblend. The encompassed steps entail the creation of history, data upload procedures, formulation of data collection, the invocation of the workflow, and diligent workflow monitoring complemented by logging.

History Creation: In the Galaxy system, the concept of a 'history' denotes a unique computational environment that encapsulates specific datasets and analyses. For every distinct study,

the script generates a new history, establishing an isolated workspace for data storage and manipulation. This new history is then associated with its corresponding project ID in a locally stored JSON file, providing easy access for future retrieval.

Data Uploading: The next step in the process involves uploading the project's datasets to the newly created history. This is achieved by forming a single string of the dataset links, which is then uploaded through the upload_dataset function. This function returns the result of the upload operation, which is further utilized in the subsequent steps. To ensure the robustness of the automation, the script actively monitors the status of the dataset upload operation. If the data is not yet available, the script pauses for a minute before rechecking, thereby preventing premature progression to the next steps.

Data Collection: The datasets are then systematically paired if needed and organized into a Galaxy collection using the create_paired_dataset_collection function. This preparation facilitates efficient downstream analyses by streamlining the data structure.

Workflow Invocation: The datasets having been suitably prepared, the script proceeds to invoke the desired workflow. This is facilitated by mapping the datasets to the inputs of the workflow, identified by its unique ID. The workflow is then run on these datasets within the current history via the gi.workflows.invoke_workflow function.

Workflow Monitoring and Logging: Perhaps the most crucial step in this automated pipeline is monitoring the execution of the workflow. The updated monitor_workflow_execution function keeps track of the workflow status and logs it into a CSV file. This consistent tracking enables the prompt detection of potential issues and the successful completion of the workflow, thereby enhancing the reliability of the process.

By doing these steps, it allows for the efficient handling and processing of multiple projects simultaneously and minimizes manual intervention and potential errors. The scripts for doing these steps are available here `https://github.com/YedilSerzhan/beer_microbiome/tree/main/data_plot_n_processing`.

3.2.3 Reproducibility analysis

This section describes the methodologies and tools used to examine workflow reproducibility, with a particular focus on the three previous studies chosen. I employed `Python` programming language (version 3.9) and `Jupyter Notebook` to conduct the analysis.

Python libraries `Pandas`[55], `Matplotlib`[56], and `Seaborn`[57] are utilized for data manipulation and visualization. `Pandas` is a robust and widely used library that offers versatile data structures for efficient data manipulation and cleaning. `Matplotlib` is a highly customizable plotting library that provides a vast array of tools to create high-quality figures for data visualization. `Seaborn`, a statistical data visualization library built on `Matplotlib`, streamlines the creation of informative and attractive plots.

To assess the difference of diversity between original and reproduced data, I applied the Shapiro-Wilk test for normality and the paired-sample t-test, which were implemented using functions from the `scipy.stats` module. The Shapiro-Wilk test is a widely used statistical method for determining if a given dataset is normally distributed. The paired-sample t-test, on the other

hand, is employed to compare the means of two related groups, which is particularly useful for assessing the effects of different treatments on a given sample set. The `shapiro` and `ttest_rel` functions from the `scipy.stats`[58] module were employed to conduct the Shapiro-Wilk test and paired-sample t-test, respectively. The metrics of the alpha diversity include Chao1, Shannon, and Simpson. Chao1 metric primarily estimates species richness, which is the number of different species present in a sample. Shannon Index encapsulates both species richness and evenness within a community. While also capturing species richness and evenness, the Simpson Index places more weight on the most abundant species.

3.3 Beer microbiome database: BeerMicroDB

The results generated from the analysis of beer data samples collected using the workflows mentioned above serve as the foundation for the beer microbiome database, MicroBeerDB.

3.3.1 Database implementation

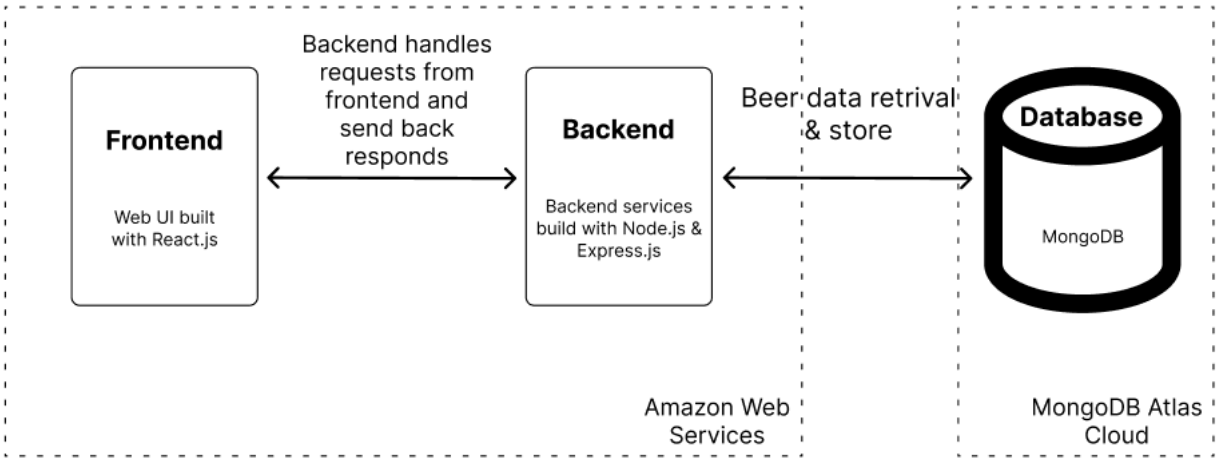


Figure 9: The system architecture diagram of BeerMicroDB
The frontend of BeerMicroDB is constructed using React.js. Resources for this frontend are sourced from requests to the backend, developed with Express.js. This backend interfaces with a database established in MongoDB. Both the frontend and backend are deployed on Amazon Web Services, while the database resides on the MongoDB Atlas Cloud.

The MicroBeerDB consists of 3 parts: 1. a frontend website built with React.js providing a web user interface. 2. Backend services built with Node.js and Express.js providing API services. 3. MongoDB hosted on MongoDB Atlas Cloud to store required data.

Frontend implementation

The frontend implementation of the Beer Microbiome Database website is primarily built using React, a powerful JavaScript library for constructing user interfaces, along with Material-UI, a popular React UI framework.

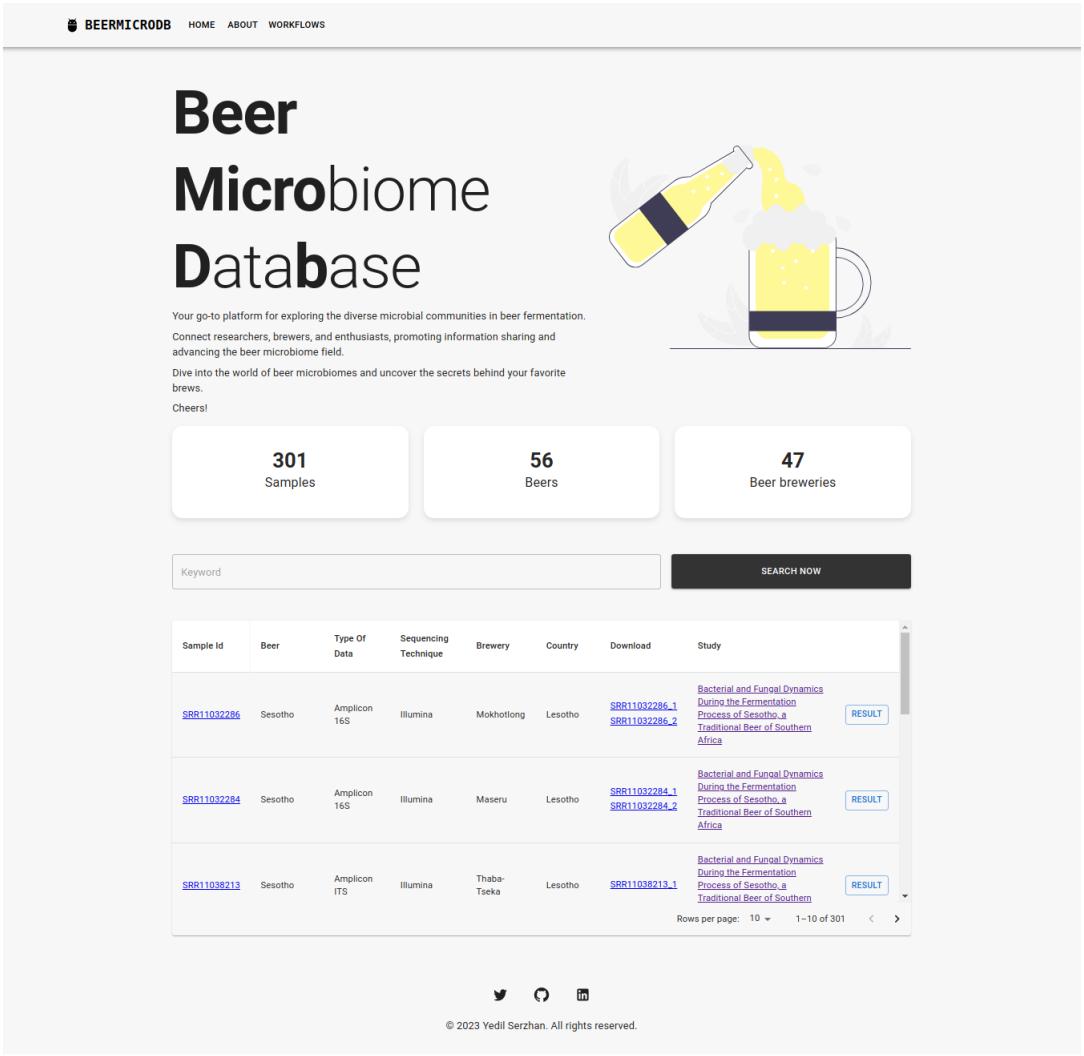


Figure 10: BeerMicroDB Home page
The homepage features a welcoming and explanatory message accompanied by an image of beer at the top. This is followed by a section presenting statistics pertaining to BeerMicroDB. Further down, a search bar allows users to seek specific information, with the results displayed in a table format beneath it.

The Home page shown in figure 10 serves as the primary entry point to the application. Here, a combination of Material-UI components is used to create a visually appealing layout. The Home page contains a search bar, and a table constructed using Material-UI components displaying various microbiome samples. This page makes an asynchronous fetch request to a server upon mount to retrieve the microbiome sample data. This data is then managed using the useState and useEffect hooks from React to implement a search functionality, allowing users to filter through the microbiome samples based on their search criteria. The sample table includes pagination, offering a user-friendly way to navigate through potentially large sets of data. Each row in the table represents a single sample and provides a link to the study page that the sample belongs to.

The Study page, a dynamic component of the application, employs React Router's useParams

hook to capture the study ID from the URL. This ID is subsequently used to dynamically import corresponding study results. This page shows the community profiling result at the species level for each sample belonging to the study.

The About page delivers valuable insights about the project's purpose, context, methodologies, and data sources. The Workflows page exhibits the shotgun and metabarcoding sequencing workflow diagram, elucidating the process involved in analyzing the microbiome samples.

All pages have a consistent layout and are equipped with a navigation bar and a Footer component, providing guidance to different components of the website. It utilizes the BrowserRouter, Routes, and Route components from the react-router-dom library to handle the navigation between different parts of the website.

In summary, the tech stack for the Beer Microbiome Database frontend implementation allows for the creation of dynamic, interactive web pages. And it provides a smooth user experience. The user interface is divided into distinct components, promoting the reusability and maintainability of the code. The frontend code base is available at `https://github.com/YedilSerzhan/beer_microbiome/tree/main/frontend`.

Backend implementation

The backend is built with Express, a minimalistic and flexible Node.js web application framework, coupled with the Mongoose library for MongoDB [59].

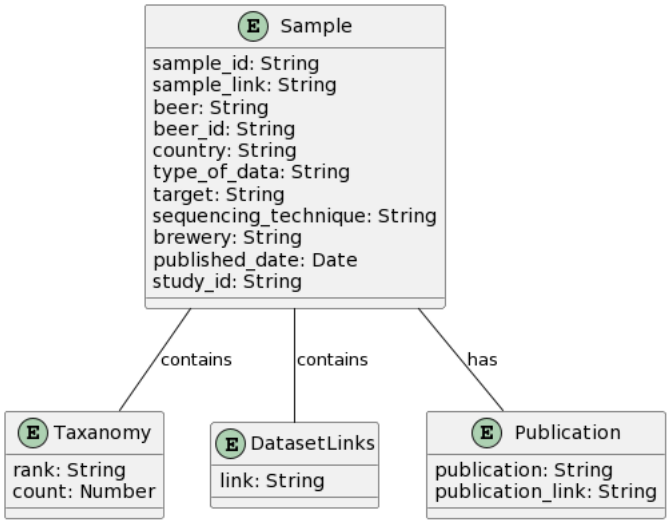


Figure 11: Database models
In the MongoDB, the sample document model encompasses both metadata and the outcomes of microbiome analysis pertinent to each respective sample.

The mongoose-based 'Sample' schema illustrated in figure 11 stands as the cornerstone of our database model. This schema was carefully designed to accommodate a plethora of attributes intrinsic to the beer microbiome samples, such as 'sample_id', 'beer', 'brewery', 'taxonomy', and more. Notably, 'taxonomy' is an array of objects, capturing information on various taxonomic

ranks and their respective counts - a fitting representation of the complexity of microbiome data. In addition to these, the model carries metadata attributes including 'sequencing_technique', 'country', 'published_date', 'dataset_links', and more, thereby encapsulating the diverse spectrum of data associated with each microbiome sample.

The implementation of RESTful API operations is achieved through distinct route handlers for each HTTP method (GET, POST, PUT, DELETE). Each route corresponds to a specific action - obtaining all samples, getting a single sample, creating a new sample, updating an existing sample, and deleting a sample. The utilization of asynchronous functions in each route ensures non-blocking code execution.

The backend code base is available at `https://github.com/YedilSerzhan/beer_microbiome/tree/main/backend`.

3.3.2 Data ingestion

Data ingestion is the process of populating a database. In our case, we need to populate the results of workflows on the beer sample data to the microbiome database we are building.

Firstly we perform data preprocessing on a DataFrame 'df' that contains the beer microbiome result. The taxonomy, dataset_links, and publication related to the sample are objectified and saved to 'db_ready.json', ready for ingestion.

Then a code snippet initiates the later process by setting up a connection with the MongoDB database using the 'mongoose' library. Once the connection is established, the 'fs' (File System) module reads the 'db_ready.json' file, containing the data to be populated. The JSON file is read asynchronously to prevent the blocking of subsequent operations.

Each object in the JSON array is instantiated as a 'Sample' model, which is defined by the 'Sample' schema in Mongoose. This operation ensures that the data being saved adheres to the structure and validation rules defined in the schema. A save operation is then executed asynchronously on the instantiated 'Sample' model. If the save operation is successful, a console message confirming the ID of the saved record is printed. Otherwise, an error message is logged.

In this way, the beer microbiome results are populated into the database. The scripts to do the data ingestion can be accessed at `https://github.com/YedilSerzhan/beer_microbiome/tree/main/backend/data`.

4 Results

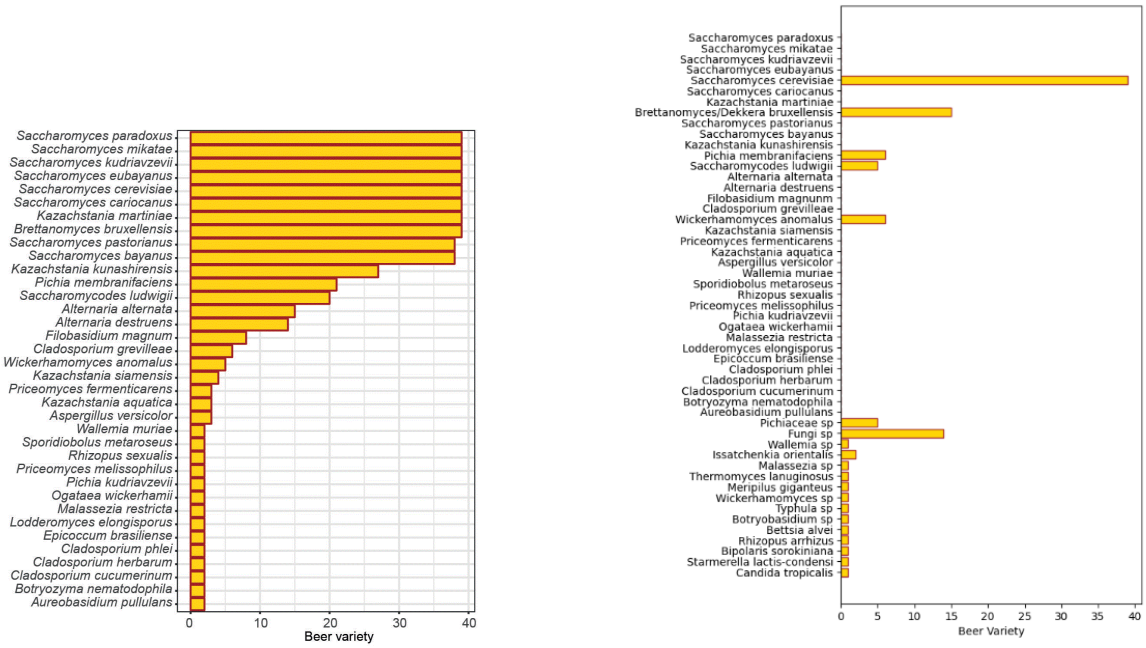
The results section is divided into two subsections. The first subsection presents the reproduced results from three previous beer microbiome studies that were introduced earlier, along with a comparison between the reproduced results and the original findings. The second subsection provides an overview of BeerMicroDB, where the overall fungal and bacterial microbiome results from the beer samples are presented.

4.1 Results of reproducibility analysis

Here are the reproduced results and comparisons between them and the original results from the prior three beer microbiome studies respectively.

4.1.1 BeerDeCoded: the open beer metagenome project

In the BeerDecoded project [4], as shown in the figures 12, analysis of 39 ITS samples revealed a total of 20 distinct fungal species, fewer than the 42 species identified in the larger scope of the BeerDEcoded study. The beer sample exhibiting the greatest ITS diversity, comprising 6 unique fungal species, was the Waldbier 2014 Schwarzkiefer. This observation aligns with the general trend observed in the BeerDEcoded project, albeit with fewer species when compared to the maximum of 19 identified in a single sample there. Four other beers, namely Stirling, La Fourbe and BE from Switzerland, and Chimay Red Cap from Belgium, also exhibited high ITS diversity with 5 species each, showing a notable variety of ITS compared to other samples within the BeerDEcoded project.

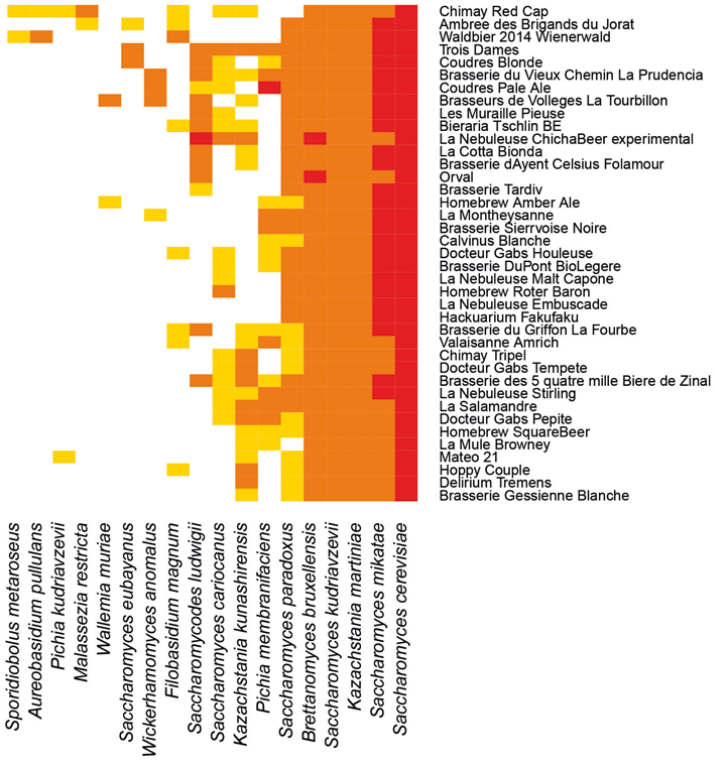


(a) Orginal result

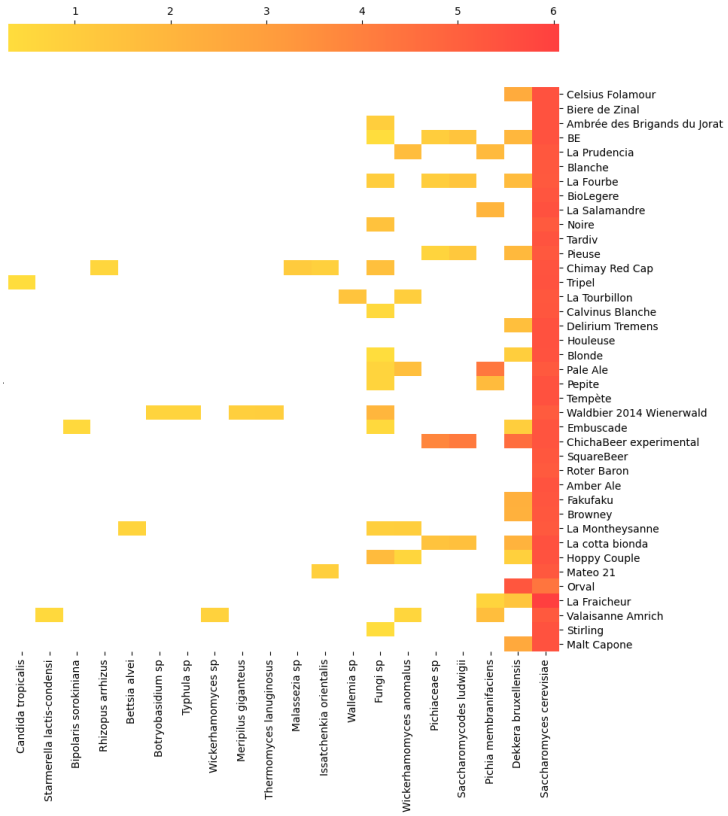
(b) Reproduced results

Figure 12: BeerDecoded beer variety diagram
The figures present data regarding the quantity of beers associated with each species identified in both the original and reproduced results. For clarity, the figure depicting the reproduced results on the right retains the species order as presented in the original figure. This arrangement aids in a more transparent comparison between the two sets of results.

Upon examining the heatmap 13, we observe, as expected, that *Saccharomyces* is the most prevalent species, identified in every beer sample, a finding that is in line with the broader BeerDEcoded project results. The next most common species is *Brettanomyces bruxellensis*, appearing in 15 samples, a species found in all samples within the BeerDEcoded project. Unexpectedly, an organism labeled as *Fungi sp* is also found in high abundance. Further investigation revealed that *Fungi sp* is labeled as 'unidentified' in the database. At present, we are uncertain whether this status is due to the workflow's inability to accurately identify it as a known species, or if it truly remains unclassified in the database.



(a) Original results



(b) Reproduced results

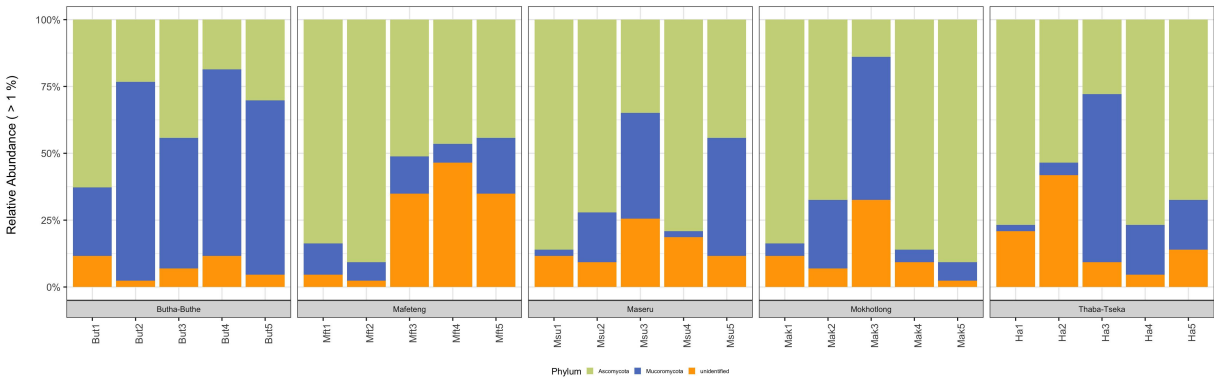
Figure 13: Heatmap of the number of reads per ITS per beer
Beer names are shown on the right and species names are shown at the bottom.

In contrast to the BeerDecoded project, our analytic process was unable to detect 31 species that were identified in the BeerDecoded study. However, we did discover 7 species not previously identified. This discrepancy could potentially stem from the relatively low quality of the samples, which averaged a quality score of around 20. As we utilized the QIIME 2 and BLAST consensus methodologies for taxonomy classification, the less-than-ideal quality could lead to ambiguous results. In instances where one feature sequence may be linked to more than two species, it could result in the feature being labeled as unidentified. This classification approach is notably different from the strategy adopted in the BeerDecoded study. In the latter, they followed an unconventional methodology of constructing a database with assumed taxonomies presumed to be present in the samples. They then proceeded to map the sequences directly to their custom-built database.

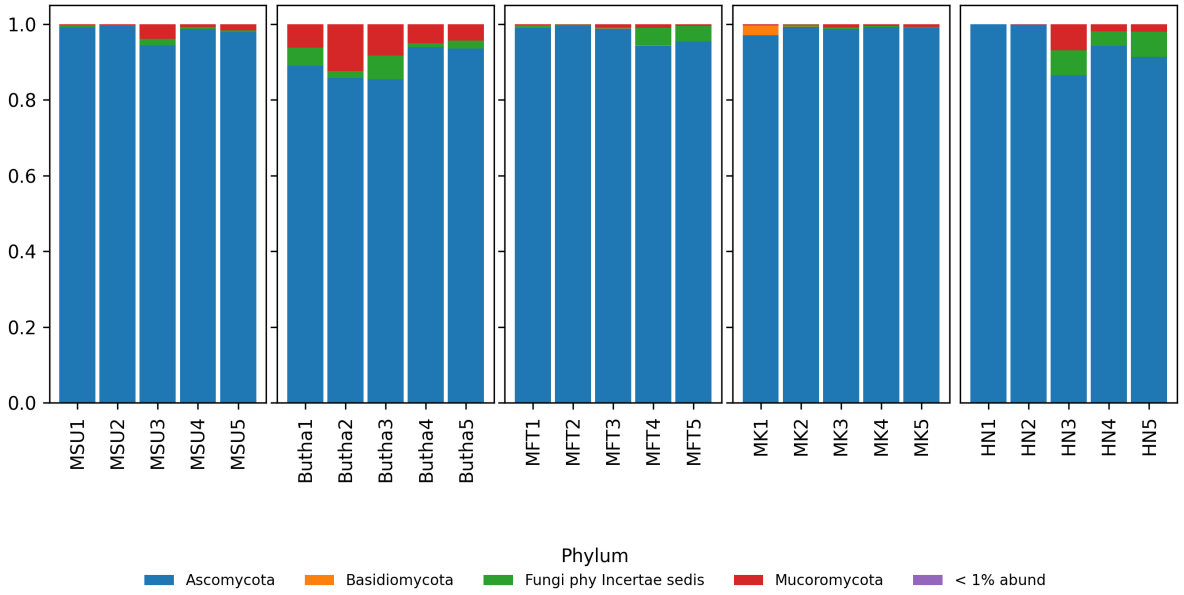
4.1.2 Bacterial and Fungal Dynamics During the Fermentation Process of Sesotho, a Traditional Beer of Southern Africa

The collection of Sesotho samples took place from five different districts (breweries) namely, Maseru (MSU), Mafeteng (MFT), Thaba-Tseka (HN), Butha-Buthe (Butha), and Mokhotlong (MK). Within each location, five samples were obtained, representing various stages of fermentation. The first sample (1) was gathered one hour after the initial starter culture was added to commence the first fermentation phase. Subsequently, the second sample (2) was obtained at approximately eight hours into the fermentation process, after the completion of the first fermentation phase. The third sample (3) was collected one hour after introducing the second starter culture, which initiated the second and final fermentation phase. Following that, the fourth sample (4) was acquired approximately eight hours after the second fermentation, prior to the beer undergoing sieving, which involves the separation of sorghum malt from the beer. Lastly, the fifth sample (5) was taken from the final product, around eight hours into the maturation stage.

Figures 14, 15, and 16, presented in this thesis and the original study, respectively, depict the distribution of fungal taxa in Sesotho beer at the Phylum, Family, and Genus levels.



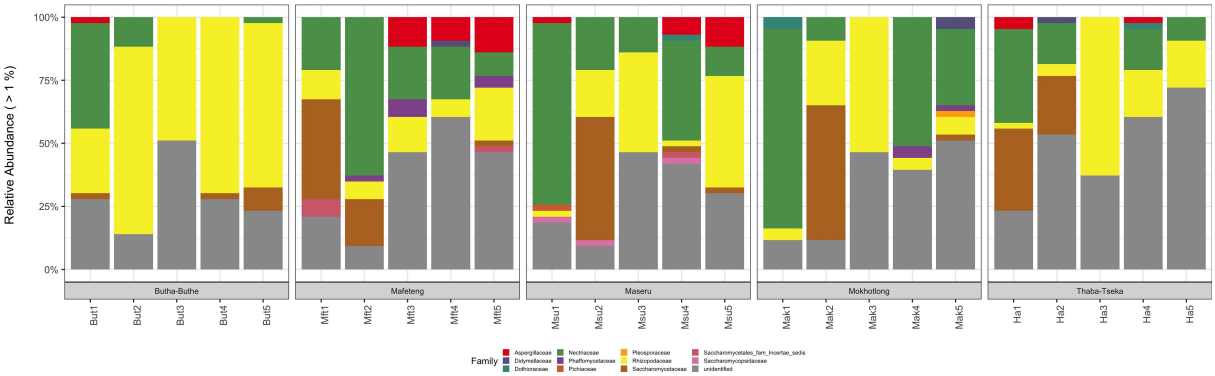
(a) Original results



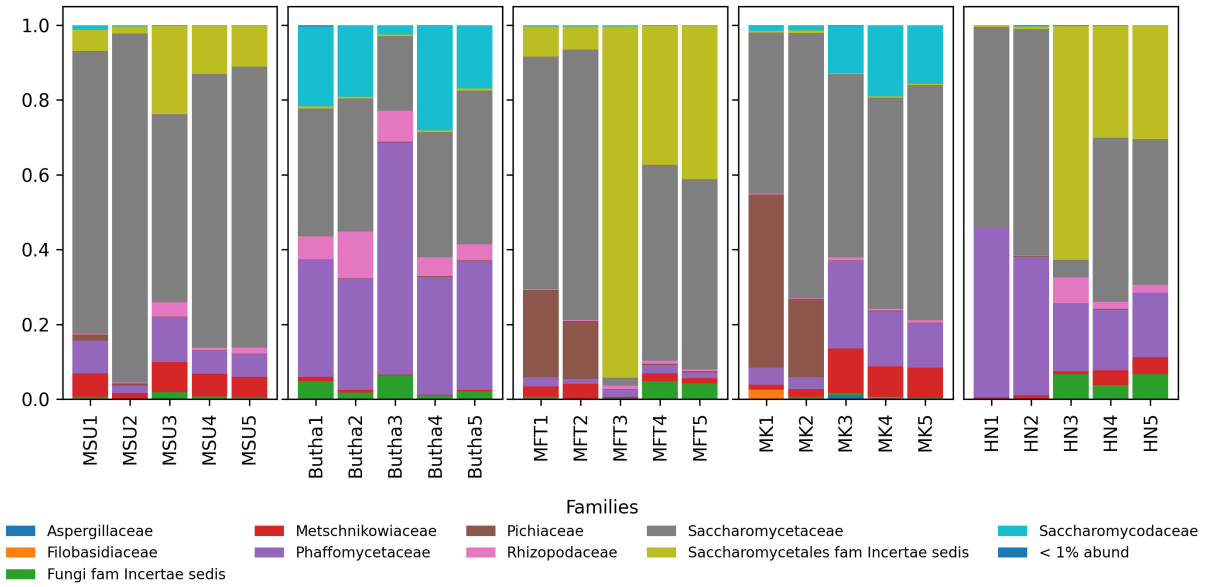
(b) Reproduced results

Figure 14: Distribution of fungal Phylum in Sesotho.
In the graphical representation, the x-axis delineates the various breweries, labeled as Maseru (MSU), Mafeteng (MFT), Thaba-Tseka (HN), Butha-Buthe (Butha), and Mokhotlong (MK). To illustrate, the label "MK1" denotes a sample sourced from Mokhotlong during the first stage of fermentation. The fungal phyla Ascomycota and Mucoromycota emerged as the predominant groups in the study. Notably, Ascomycota displayed a higher dominance in the reproduced results compared to the original findings.

Similar to the original findings, Ascomycota and Mucoromycota were the predominant fungal phyla observed in all samples and locations as shown in figure 14. However, in this thesis, the presence of Basidiomycota was also identified. Basidiomycota fungi are widely distributed in nature and can be found in various habitats such as soil, plants, and air. If the raw ingredients, such as sorghum or maize, are contaminated with Basidiomycota spores during the brewing process, these spores can enter the beer mixture and subsequently proliferate during fermentation.



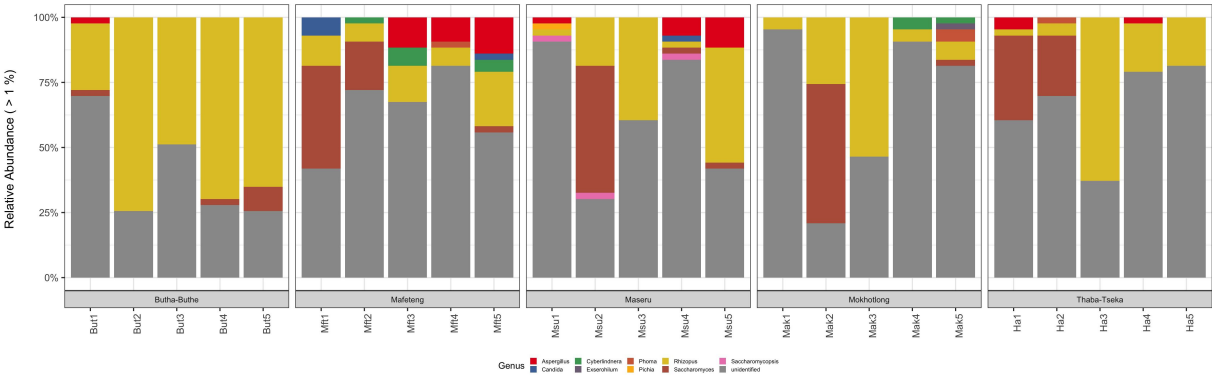
(a) Original results



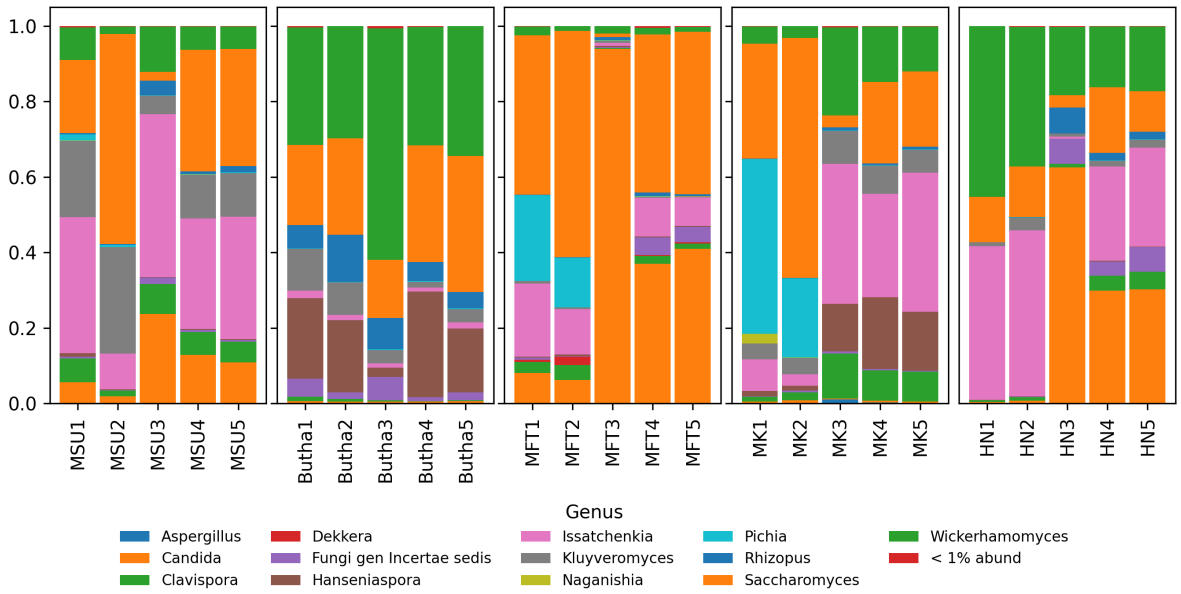
(b) Reproduced results

Figure 15: Distribution of fungal Family in Sesotho.
In alignment with the original findings, the reproduced data also identified the presence of *Phaffomycetaceae* and *Pichiaceae*.

As seen in figure 15, consistent with the Original results, we observed the presence of *Saccharomycetaceae* and *Nectriaceae* from the Phylum Ascomycota, as well as *Rhizopodaceae* from the Phylum Mucoromycota. Additionally, we detected a substantial abundance of *Phaffomycetaceae* and *Pichiaceae*, which are families of yeasts belonging to the order Saccharomycetales.



(a) Original results



(b) Reproduced results

Figure 16: Distribution of fungal Genus in Sesotho
Based on the analysis of the distribution of fungal genera in Sesotho, *Rhizopus* emerges as the dominant genus in the original findings. In contrast, *Saccharomyces* is more prevalent in the reproduced data.

Unlike the original findings, where *Saccharomyces* was most prevalent in the 1st and 2nd stages of the brewing process, our analysis revealed the presence of *Saccharomyces* throughout all stages of brewing (Figure 16). Moreover, *Wickerhamomyces* was also found in significant abundance, which is a genus of fungi within the *Saccharomycetales* order. *Wickerhamomyces anomalus* is known to produce a faintly pleasant odor.

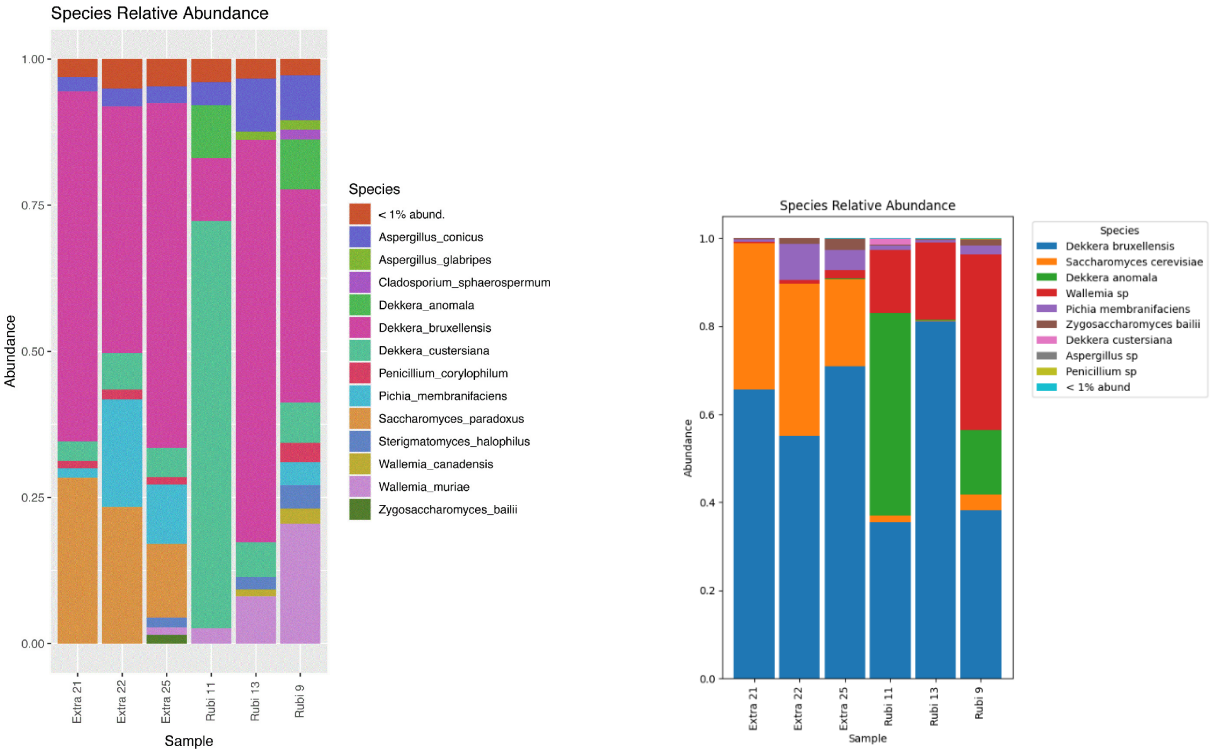
Overall, this thesis demonstrates a greater diversity of fungal taxa at all taxonomic levels compared to the original study. This disparity could be attributed to the utilization of QIIME 2 instead of QIIME 1, as well as the incorporation of an updated version of the UNITE database, which likely contributed to enhanced taxonomic resolution and accuracy in our analysis.

4.1.3 A Culture-Independent Comparison of Microbial Communities of Two Maturating Craft Beers Styles

For this study, two beer styles produced by a commercial craft brewery were selected. The first beer style, named Extra, is classified as a Doppelbock Lager and has an alcohol content of 8%. The second beer style, called Rubi, falls under the category of a Märzen Lager and has an alcohol content of 6.3%.

Fungal diversity

Contrary to other studies shown above, the analysis revealed that *Saccharomyces* was not the dominant genus in spontaneously-fermented craft beer, consistent with previous findings[60]. Instead, the most abundant genus in both beer styles was *Dekkera*, represented by the species *D. bruxellensis*, *D. anomala*, and *D. custersiana* (Figure 17). This suggests that a souring process occurs during maturation, contributing to the flavor profile. Additionally, *Zygosaccharomyces*, known for its association with fruity flavors [61], was also detected in the reproduced result, consistent with the original findings.



(a) Original results

(b) Reproduced results

Figure 17: Fungal species relative abundance
The relative abundance of fungal taxa at the species level across various beer styles is depicted. Notably, *Dekkera* stands out as the most abundant genus in both beer styles examined.

In addition to the presence of *Pichia*, the Original results showed the detection of *Aspergillus* and *Penicillium*. Notably, *Hanseniaspora uvarum*, a yeast commonly found on grapes[62], was

also identified, indicating its presence in the samples.

Table 2: Comparison of Original and reproduced estimates and T-test results

(a) Species richness and diversity of fungal communities in Extra and Rubi beer barrels

Sample	Chao1		Shannon		Simpson	
	Original	Reproduced	Original	Reproduced	Original	Reproduced
Extra #21	48	42	2.07	2.50	0.83	0.76
Extra #22	67.5	61	2.45	2.94	0.87	0.80
Extra #25	98.6	94	2.53	3.31	0.86	0.82
Rubi #11	70.5	65	1.88	3.46	0.72	0.82
Rubi #13	90.2	89	2.16	2.71	0.77	0.65
Rubi #9	117.2	104	3.14	4.16	0.92	0.89

(b) T-test results for Original vs Reproduced estimates

Index	T-statistic	p-value
Chao1	-3.238	0.0230
Shannon	-8.748	0.0003
Simpson	7.000	0.0009

Note: Chao1 represents the Chao1 species richness estimator.
Shannon denotes the Shannon index of biodiversity. Higher values indicate higher diversity.
Simpson represents the Simpson diversity index. Values range from 0 (simplest) to 1 (most diverse).

Upon reviewing the results from the original study and the reproduced results of diversity indices for the beer samples on the diversity of fungal communities, several notable points of comparison and contrast emerge.

First, looking at the Chao1 biodiversity index, the reproduced results indicate generally lower estimates across all samples, with a noticeable decrease in Rubi #9 (from 117.2 to 104) and Extra #21 (from 48 to 42). Chao1 is a non-parametric estimator used to predict the true species richness of a community, and thus these lower estimates suggest that the reproduced results predict fewer unique species in each sample than the original study.

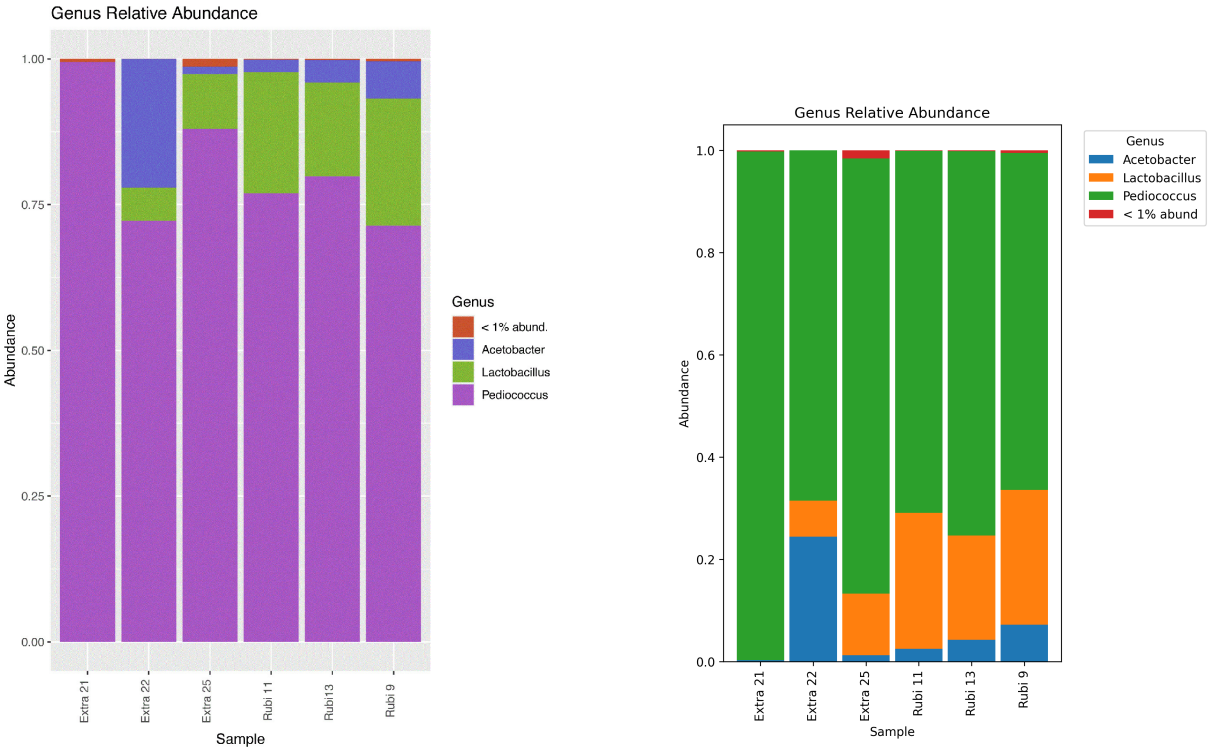
In contrast, the Shannon index, which measures the entropy (or unpredictability) of species diversity within a community, is consistently higher in the reproduced results across all samples. This suggests that the reproduced results found a greater degree of diversity, possibly with more evenly distributed species within each sample. The most pronounced differences can be observed in Rubi #11 (from 1.88 to 3.46) and Extra #22 (from 2.45 to 2.94).

Similarly, the Simpson index, which is a measure of species diversity that gives more weight to the abundant species, also shows an increase in all but one sample (Rubi #13, where it decreased from 0.77 to 0.65) in the reproduced results. This suggests an increased evenness and diversity, as a higher Simpson's index implies that individuals are distributed more evenly among species.

In summary, with these differences in individual measures and t-test results, the data suggests that the reproduced results have more evenly distributed diversity but less richness.

Bacterial diversity

In our reanalysis of the bacterial communities present in two distinct styles of beer, we found the overall bacterial distribution to be consistent with the original study (Figure 18). The primary microbial constituents encompassed just three genera, each with a relative abundance of 1% or more. These pivotal genera were identified as *Pediococcus*, *Lactobacillus*, and *Acetobacter*.



(a) Original results

(b) Reproduced results

Figure 18: Bacterial genus relative abundance
Regarding the relative abundance of bacterial taxa at the genus level within the beer styles, the produced results align well with the findings of the original study.

Discrepancies in the prevalence of these bacterial genera were more pronounced in the samples of Extra beer when compared to the Ruby variant. A stark example was found in Extra barrel #21 where *Pediococcus* appeared to be the predominant microorganism, accounting for a staggering 99% of the bacterial population. Conversely, Extra barrel #22 was characterized by the highest percentage of *Acetobacter* across all beer samples, recorded at 22%. Extra barrel #25 was notable for exhibiting the greatest abundance of *Lactobacillus* among the Extra samples. Interestingly, this marked variability was not reflected in the Ruby beer samples. Instead, we observed a much more uniform distribution of bacterial genera across the Ruby barrels.

The Chao1 index, an estimator of species richness, appears to show higher values in the reproduced results than in the original, for all samples. The increase in Chao1 values implies that the re-analysis might have identified more species within the samples. A T-test comparing the original and reproduced Chao1 values reveals a significant difference with a p-value of 0.012. This

Table 3: Comparison of Original and reproduced estimates and T-test results

(a) Species richness and diversity of bacterial communities in Extra and Rubi beer barrels

Sample	Chao1		Shannon		Simpson	
	Original	Reproduced	Original	Reproduced	Original	Reproduced
Extra #21	10	13	1.45	1.73	0.74	0.66
Extra #22	10	13	1.86	2.45	0.82	0.78
Extra #25	18	31	1.88	2.57	0.81	0.77
Rubi #11	9	19	1.72	2.27	0.80	0.76
Rubi #13	17	24	1.96	2.69	0.83	0.79
Rubi #9	10	11	1.75	2.30	0.81	0.77

(b) T-test results for Original vs reproduced estimates

Index	T-statistic	p-value
Chao1	3.845	0.012
Shannon	-4.537	0.006
Simpson	1.257	0.264

Note: Chao1 represents the Chao1 species richness estimator.
Shannon denotes the Shannon index of biodiversity. Higher values indicate higher diversity.
Simpson represents the Simpson diversity index. Values range from 0 (simplest) to 1 (most diverse).

indicates that the difference in species richness estimates between the original and reproduced data is not likely due to chance.

For the Shannon index, which incorporates both species richness and evenness, the reproduced results also display higher values for all samples. The T-test further corroborates this observation with a statistically significant p-value of 0.006. Higher Shannon values suggest that the bacterial communities identified in the re-analysis are not only richer in species but also display a more even distribution among the identified species.

In contrast, the Simpson index, which places greater weight on dominant species, reveals no statistically significant difference between the original and reproduced estimates (p-value = 0.264). This suggests that, despite variations in species richness and evenness, the overall dominance structure of the bacterial communities remains relatively consistent between the original and reproduced results.

In summary, the comparison between the original and reproduced diversity estimates reveals a significant difference in the Chao1 and Shannon indices, but not in the Simpson index. This suggests that while species richness and evenness may be affected by the improved workflow, the overall dominance structure of the communities remains stable across studies.

4.2 BeerMicroDB overview

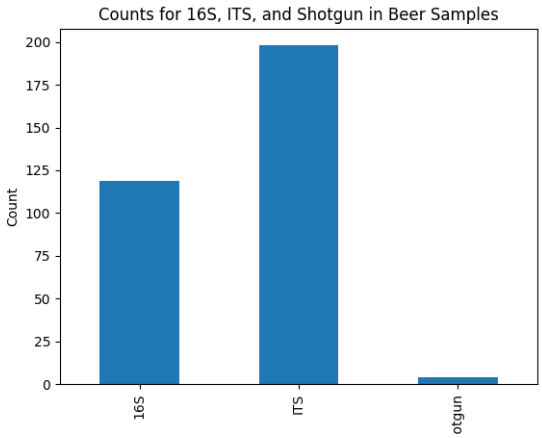


Figure 19: Beer sample count

Overall, the BeerMicroDB database comprises a total of 119 samples analyzed via 16S rRNA gene sequencing, 198 samples examined using ITS sequencing, and a significantly smaller subset of just 4 samples studied using shotgun metagenomic sequencing (Figure 19). It's clear that our current database is predominantly populated with data from metabarcoding methods, namely 16S and ITS sequencing.

Despite the robustness and insights these methods offer, the relative scarcity of shotgun metagenomic data in our collection does present a limitation in terms of the depth and granularity of our microbial community analysis. We acknowledge this disparity in data types and emphasize the need to enhance the proportion of shotgun metagenomic data within BeerMicroDB. As such, we are hopeful that with the recommencement of the BeerDEcoded and Street Science Project, which help public citizens aware of bioinformatics by learning from sequencing to analyzing beer microbiome, we will be able to enrich our database with more shotgun metagenomic data, thus improving the breadth and depth of our analyses and leading to more nuanced insights into the beer microbiome.

4.2.1 Fungal microbiome overview

This section gives an overview of fungal microbiome results.

As shown in figure 20, the beer sample with the highest fungal species count was "Sesotho," which contained an impressive number of 70 distinct species. This was followed by "Rubi Marzen Lager" and "Extra Doppelbock Lager," hosting 51 and 46 species respectively. These beers exhibited significantly higher fungal biodiversity compared to others in the analysis. For instance, "Blond Beer (8.88%)," "Chimay Red Cap," "Waldbier 2014 Wienerwald," "La Fourbe," "Valaisanne Amrich," and "La Montheysanne" exhibited species counts ranging from 7 to 21.

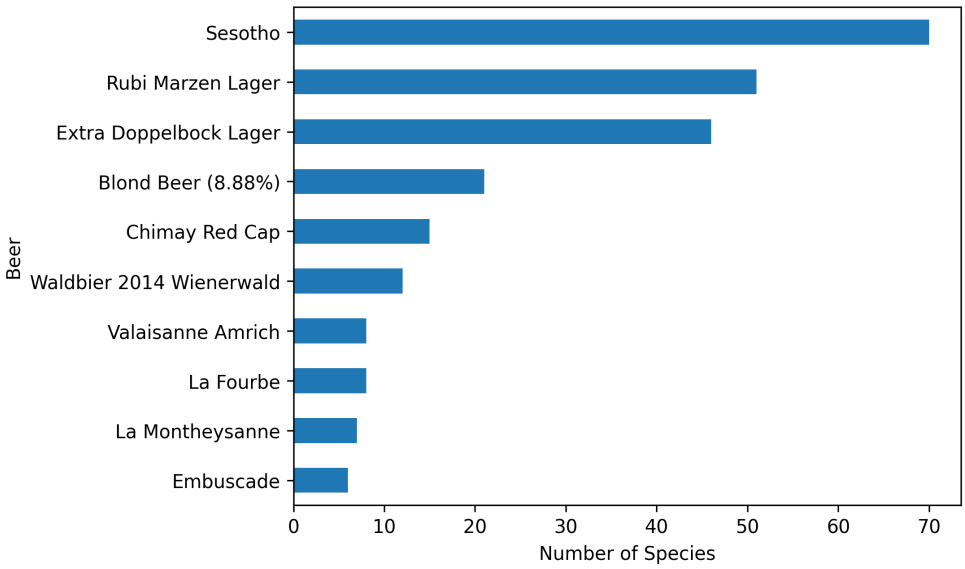


Figure 20: Top 10 Beers with the Highest Species Count
Among all beer types, Sesotho, which undergoes spontaneous fermentation, exhibits the highest count of fungal species.

The data suggest that beer type and presumably the specific brewing process might significantly influence the diversity of fungal species in the final product. The top beers all have less regulated brewing processes and conditions.

The fungal species with the highest frequency of occurrence across the beer samples was *Saccharomyces cerevisiae* identified in 126 instances (Figure 21). This is unsurprising, given that *S. cerevisiae* is a commonly used yeast in the brewing industry due to its robust fermentation capabilities and its contribution to beer's distinctive flavor profiles.

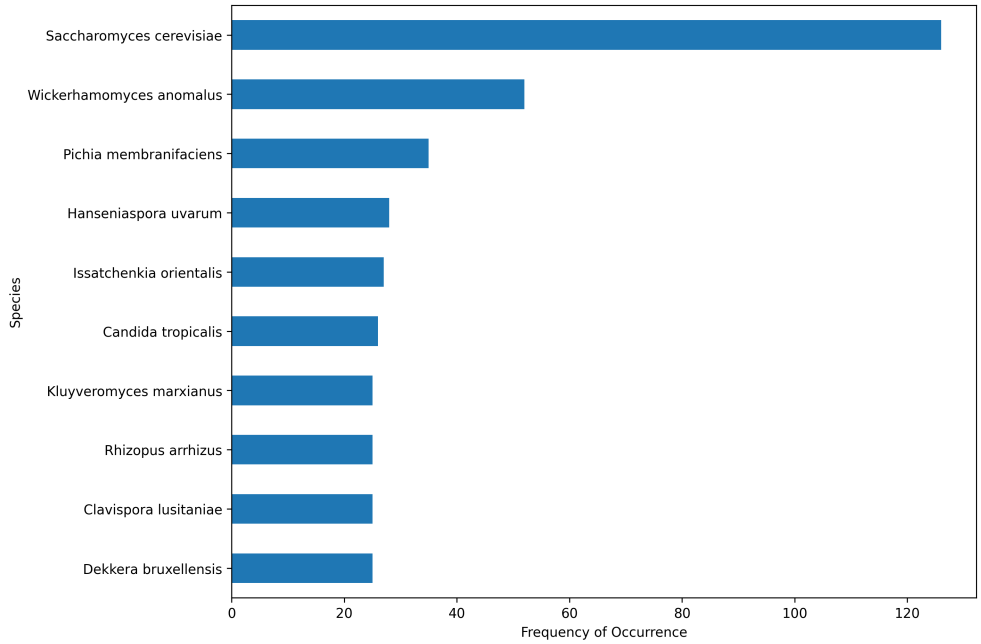


Figure 21: Top 10 Species with the Highest Frequency of Occurrence
Saccharomyces cerevisiae unequivocally emerges as the most prevalent species. It is closely followed by *Wickerhamomyces anomalus*, which holds a slight advantage over other fungal species in the dataset.

Other recurrent species include *Wickerhamomyces anomalus* and *Pichia membranifaciens* noted in 52 and 35 instances respectively, and a group of species including *Hanseniaspora uvarum*, *Issatchenkia orientalis*, *Candida tropicalis*, *Dekkera bruxellensis*, *Clavispora lusitaniae*, *Rhizopus arrhizus* and *Kluyveromyces marxianus* each of which appeared 25-28 times.

These findings highlight the prevalence and importance of these species in the beer brewing ecosystem. It should be noted that while some of these species, such as *D. bruxellensis* can potentially impart off-flavors to beer, others like *K. marxianus* and *W. anomalus* are known to have potentially beneficial effects on flavor and aroma complexity as shown before. Therefore, these species' presence and frequency could play a significant role in shaping the unique characteristics of the beers analyzed.

4.2.2 Bacterial microbiome overview

This section gives an overview of bacterial microbiome results from the 119 16S samples. The data revealed a diverse range of bacterial species associated with the beers, indicating the presence of complex microbial communities that potentially contribute to the unique flavors and characteristics of each brew. Two key aspects were explored: the top 10 beers with the highest species count and the top 10 species with the highest frequency of occurrence.

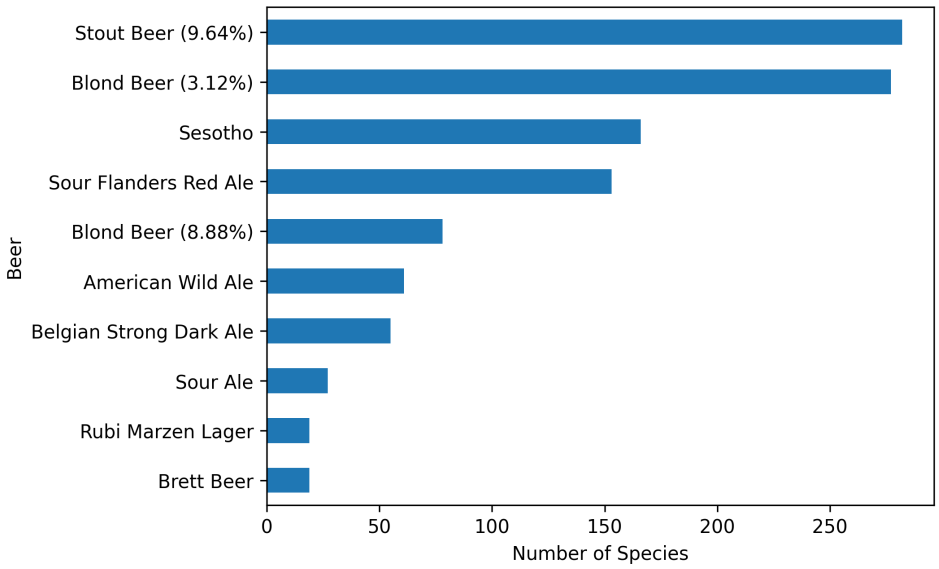


Figure 22: Top 10 Beers with the Highest Species Count
The stout beer (9.64%) and blond beer (3.12%) lead in terms of bacterial species richness. They are closely followed by Sesotho and sour Flanders red ale in the subsequent rankings.

The analysis of the bacterial microbiome in different types of beer unveils a complex and diverse bacterial composition. Notably, Stout Beer with a 9.64% alcohol content exhibits the highest species count (282), followed closely by Blond Beer at two different alcohol percentages, 3.12% (277 species) and 8.88% (78 species) shown in the figure 22.

Interestingly, the presence of a beer called Sesotho, containing 166 species, indicates a potential regional influence on microbial diversity. Other notable inclusions are Sour Flanders Red Ale (153 species), American Wild Ale (61 species), Belgian Strong Dark Ale (55 species), Sour Ale (27 species), Brett Beer (19 species), and Rubi Marzen Lager (19 species).

The stour beer and blond beer went through a barrel aging process, this could be the reason why they have a high bacterial species count. And Sesotho which is spontaneously fermented in the wild environment, could be the reason why it is with high diversities. Overall, as we can see, the sour beer, ale has higher bacterial diversity.

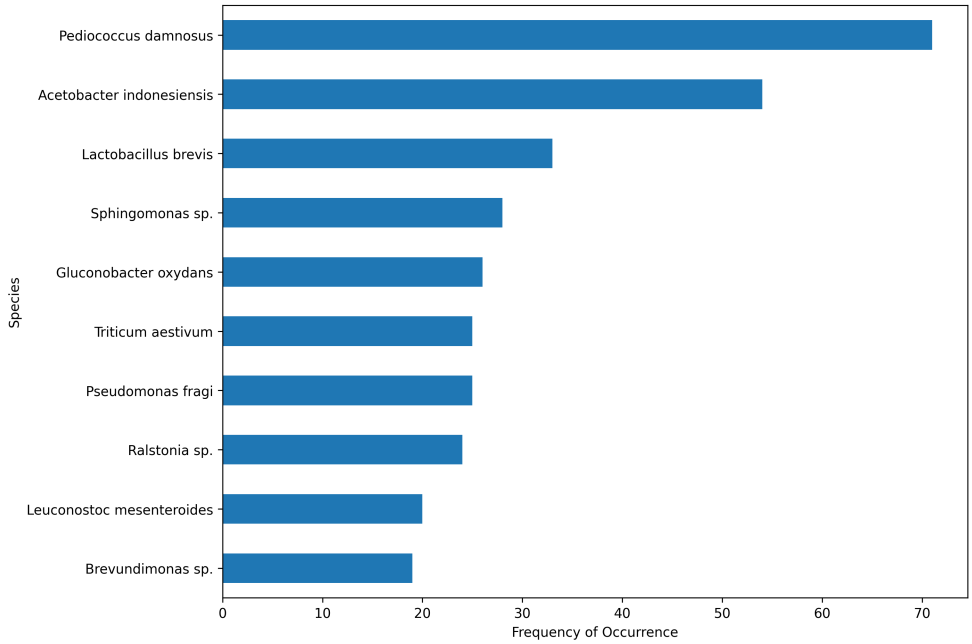


Figure 23: Top 10 Species with the Highest Frequency of Occurrence
The species *Pediococcus damnosus* emerged as the most dominant, closely followed by *Acetobacter indonesiensis*. Among the subsequent species observed, eight exhibited fewer differences in their frequency of occurrence.

A close inspection of the top ten bacterial species, in terms of frequency of occurrence, highlights the significant presence of certain bacteria across different beer samples (Figure 23). The most prevalent species is *Pediococcus damnosus*, found in 71 samples, followed by *Acetobacter indonesiensis* (54), *Lactobacillus brevis* (33), and *Sphingomonas sp.* (28).

The presence of *Gluconobacter oxydans* (26), *Pseudomonas fragi* (25), *Triticum aestivum* (25), *Ralstonia sp.* (24), *Leuconostoc mesenteroides* (20), and *Brevundimonas sp.* (19) also suggests their common role in various beer types. These species are often involved in fermentation processes and may contribute to the unique flavors, aromas, and characteristics of the beer.

Furthermore, the occurrence of *Triticum aestivum* might be indicative of wheat's role in certain beer formulations. The identification and characterization of these bacteria not only provide insights into the microbiological profile of different beers but also raise potential considerations in quality control, flavor development, and potential health aspects.

In conclusion, these findings serve as an essential resource for understanding the intricate bacterial ecology within various beer types, contributing to ongoing research in brewing science, quality assurance, and sensory analysis. Further studies may elucidate the specific roles and interactions of these microbial communities in shaping the attributes of the diverse world of beer.

5 Conclusion

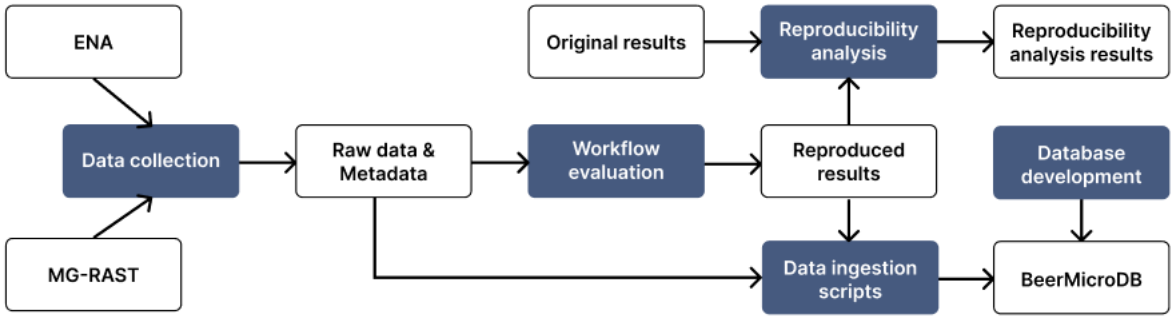


Figure 24: Thesis overview
In the overview of the thesis, the white box stands for the data and results, and the blue box symbolizes the action taken. The research starts with data collection, advancing to the evaluation of workflows, followed by an in-depth reproducibility analysis, and culminating in the construction of the BeerMicroDB.

The thesis commences with the gathering of raw data and metadata from two sources: ENA and MG-RAST. Subsequently, two distinct workflows are employed: a metabarcoding workflow developed using `QIIME 2` and a shotgun workflow facilitated by `Kraken 2`. Both workflows have been implemented on the `Galaxy` platform and automated through the Galaxy API Python library, known as `BioBlend`. For the comparison of reproduced results with original findings, `Jupyter Notebook` combined with `Python` libraries such as `Pandas`, `Matplotlib`, and `Seaborn` was utilized to undertake the reproducibility analysis. Based on the raw data metadata and the reproduced results, BeerMicroDB was constructed, integrating `Express.js` and `MongoDB` for backend operations and `React.js` for frontend development.

Upon the execution of the workflows, while these workflows had some variances, their overarching outcomes were in harmonious alignment with the conclusions of prior studies that we have collected. One of the principal outcomes drawn from these workflows was the evident richness in microbiome diversity observed in certain beer types, particularly those subjected to spontaneous fermentation and aging procedures. The spontaneously fermented beers include sesotho - a traditional African beer, along with Extra Doppelbock Lager and Rubi Marzen Lager. Concurrently, the aging process was predominantly linked with beers such as Blond Beer and stout beer. Contrastingly, when these beers were measured against their industrial counterparts, the distinction was clear. The aforementioned beers showcased a significantly more diverse microbiome profile. This resonates with the idea that traditional and specific brewing methods can indeed have a pronounced impact on the microbial diversity of the final product.

Within the beer samples, bacterial species such as *Pediococcus damnosus*, *Acetobacter indonesiensis*, and *Lactobacillus brevis* were predominant. On the fungal front, the dominance was held by species like *Saccharomyces cerevisiae*, *Wickerhamomyces anomalus*, and *Pichia membranifaciens*.

Our initiative to create a specialized microbiome database, named BeerMicroDB, has accumulated metadata and microbiome compositions encompassing 56 distinct beers and a total of 301

samples that are publicly available for users to browse and have an insight into beers and beer microbiomes. However, the research encountered certain limitations. The number of shotgun datasets is disproportionate relative to the metabarcoding datasets. Additionally, the `QIIME 2` tool on `Galaxy` currently lacks support for manifest files for data importation, given the obscured nature of data pathways in the cloud environment to its users. It's also worth noting that the present iteration of BeerMicroDB is limited to read-only access. The in-depth analysis of content in the database provided by BeerMicroDB appears to be lacking.

Anticipating future advancements, there exists a keen interest in expanding the dataset spectrum, both in terms of quantity and the variety of beer types represented. Introducing more advanced statistical methods, such as principal component analysis (PCA), and potentially harnessing machine learning techniques could offer valuable insights into distinguishing the core beer microbiome associated with varying beer types or regional beer variations. Moreover, an objective lies in refining BeerMicroDB by incorporating user authentication mechanisms. This would potentially facilitate user-driven contributions, enabling them to append additional datasets and share their microbiome analysis outcomes, thereby enriching the repository.

References

[1] BarthHaas. "BarthHaas Report 2021/2022". In: *BarthHaas Report* (2022).

[2] Alexander Tyakht et al. "Characteristics of bacterial and yeast microbiomes in spontaneous and mixed-fermentation beer and cider". In: *Food Microbiology* 94 (2021), p. 103658.

[3] Sofie Bossaert et al. "Description of the temporal dynamics in microbial community composition and beer chemistry in sour beer production via barrel ageing of finished beers". In: *International Journal of Food Microbiology* 339 (2021), p. 109030.

[4] Jonathan Sobel et al. "BeerDeCoded: the open beer metagenome project". In: *F1000Research* 6 (2017), p. 1676.

[5] Rasko Leinonen et al. "The European nucleotide archive". In: *Nucleic acids research* 39.suppl_1 (2010), pp. D28–D31.

[6] Kevin P Keegan, Elizabeth M Glass and Folker Meyer. "MG-RAST, a metagenomics service for analysis of microbial community structure and function". In: *Microbial environmental genomics (MEG)* (2016), pp. 207–233.

[7] Thomas Dale Brock et al. *Brock biology of microorganisms*. Upper Saddle River (NJ): Prentice-Hall, 2003., 2003.

[8] Xing Yang et al. "More than 9,000,000 unique genes in human gut bacterial community: estimating gene numbers inside a human body". In: *PloS one* 4.6 (2009), e6074.

[9] Richa Bharti and Dominik G Grimm. "Current challenges and best-practice protocols for microbiome analysis". In: *Briefings in bioinformatics* 22.1 (2021), pp. 178–193.

[10] Juliana Durack and Susan V Lynch. "The gut microbiome: Relationships with disease and opportunities for therapy". In: *Journal of experimental medicine* 216.1 (2019), pp. 20–40.

[11] Paton Vuong, Sandy Chong and Parwinder Kaur. "The little things that matter: how bioprospecting microbial biodiversity can build towards the realization of United Nations Sustainable Development Goals". In: *npj Biodiversity* 1.1 (2022), p. 4.

[12] Michael A Quail et al. "A large genome center's improvements to the Illumina sequencing system". In: *Nature methods* 5.12 (2008), pp. 1005–1010.

[13] Yunhao Wang et al. "Nanopore sequencing technology, bioinformatics and applications". In: *Nature biotechnology* 39.11 (2021), pp. 1348–1365.

[14] Simon Andrews et al. *FastQC: a quality control tool for high throughput sequence data*. 2010.

[15] Anthony M Bolger, Marc Lohse and Bjoern Usadel. "Trimmomatic: a flexible trimmer for Illumina sequence data". In: *Bioinformatics* 30.15 (2014), pp. 2114–2120.

[16] Marcel Martin. "Cutadapt removes adapter sequences from high-throughput sequencing reads". In: *EMBnet. journal* 17.1 (2011), pp. 10–12.

[17] Evan Bolyen et al. "Reproducible, interactive, scalable and extensible microbiome data science using QIIME 2". In: *Nature biotechnology* 37.8 (2019), pp. 852–857.

[18] Patrick D. Schloss et al. "Introducing mothur: Open-Source, Platform-Independent, Community-Supported Software for Describing and Comparing Microbial Communities". In: *Applied and Environmental Microbiology* 75.23 (2009), pp. 7537–7541. ISSN: 0099-2240. DOI: 10.1128/AEM.01541-09. URL: https://aem.asm.org/content/75/23/7537.

[19] Robert C Edgar. “Search and clustering orders of magnitude faster than BLAST”. In: *Bioinformatics* 26.19 (2010), pp. 2460–2461.

[20] Aitor Blanco-Mıguez et al. “Extending and improving metagenomic taxonomic profiling with uncharacterized species using MetaPhlAn 4”. In: *Nature Biotechnology* (2023), pp. 1–12.

[21] Derrick E Wood, Jennifer Lu and Ben Langmead. “Improved metagenomic analysis with Kraken 2”. In: *Genome biology* 20 (2019), pp. 1–13.

[22] Benjamin J Callahan et al. “DADA2: High-resolution sample inference from Illumina amplicon data”. In: *Nature methods* 13.7 (2016), pp. 581–583.

[23] Alessio Milanese et al. “Microbial abundance, activity and population genomic profiling with mOTUs2”. In: *Nature communications* 10.1 (2019), p. 1014.

[24] Florian P Breitwieser and Steven L Salzberg. “Pavian: interactive analysis of metagenomics data for microbiome studies and pathogen identification”. In: *Bioinformatics* 36.4 (2020), pp. 1303–1304.

[25] Brian D Ondov, Nicholas H Bergman and Adam M Phillippy. “Interactive metagenomic visualization in a Web browser”. In: *BMC bioinformatics* 12.1 (2011), pp. 1–10.

[26] Felix Mölder et al. “Sustainable data analysis with Snakemake”. In: *F1000Research* 10 (2021).

[27] Paolo Di Tommaso et al. “Nextflow enables reproducible computational workflows”. In: *Nature biotechnology* 35.4 (2017), pp. 316–319.

[28] Michael R Crusoe et al. “Methods included: Standardizing computational reuse and portability with the common workflow language”. In: *Communications of the ACM* 65.6 (2022), pp. 54–63.

[29] “The Galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2022 update”. In: *Nucleic Acids Research* 50.W1 (2022), W345–W351.

[30] Clare Sloggett, Nuwan Goonasekera and Enis Afgan. “BioBlend: automating pipeline analyses within Galaxy and CloudMan”. In: *Bioinformatics* 29.13 (2013), pp. 1685–1686.

[31] Lorna Richardson et al. “MGnify: the microbiome sequence data analysis resource in 2023”. In: *Nucleic Acids Research* 51.D1 (2023), pp. D753–D759.

[32] Lorenz Christian Reimer et al. “Bac Dive in 2022: the knowledge base for standardized bacterial and archaeal data”. In: *Nucleic Acids Research* 50.D1 (2022), pp. D741–D746.

[33] Ruben Vicente-Saez and Clara Martinez-Fuentes. “Open Science now: A systematic literature review for an integrated definition”. In: *Journal of business research* 88 (2018), pp. 428–436.

[34] CW Bamforth. “A brief history of beer”. In: *Proceedings of the 26th Convention of the Institute of Brewing, Asia-Pacific Section, Singapore*. 2000, pp. 5–12.

[35] Victor R Preedy. *Beer in health and disease prevention*. Academic Press, 2011.

[36] Giovanni De Gaetano et al. “Effects of moderate beer consumption on health and disease: A consensus document”. In: *Nutrition, Metabolism and Cardiovascular Diseases* 26.6 (2016), pp. 443–467.

[37] Gabriela A Miguel et al. “Non-Saccharomyces yeasts for beer production: Insights into safety aspects and considerations”. In: *International journal of food microbiology* (2022), p. 109951.

[38] Renan Eugênio Araujo Piraine, Fábio Pereira Leivas Leite and Matthew L Bochman. “Mixed-culture metagenomics of the microbes making sour beer”. In: *Fermentation* 7.3 (2021), p. 174.

[39] Claudia Gonzalez Viejo et al. “Chemical characterization of aromas in beer and their effect on consumers liking”. In: *Food chemistry* 293 (2019), pp. 479–485.

[40] Freek Spitaels et al. “The microbial diversity of traditional spontaneously fermented lambic beer”. In: *PloS one* 9.4 (2014), e95384.

[41] Nicholas A Bokulich, Charles W Bamforth and David A Mills. “Brewhouse-resident microbiota are responsible for multi-stage fermentation of American coolship ale”. In: *PloS one* 7.4 (2012), e35507.

[42] Chris White and Jamil Zainasheff. *Yeast: the practical guide to beer fermentation*. Brewers Publications, 2010.

[43] Errol D Cason et al. “Bacterial and fungal dynamics during the fermentation process of sesotho, a traditional beer of Southern Africa”. In: *Frontiers in Microbiology* 11 (2020), p. 1451.

[44] João Costa, Isabel N Sierra-Garcia and Angela Cunha. “A Culture-Independent Comparison of Microbial Communities of Two Maturating Craft Beers Styles”. In: (2022).

[45] Nuala A O’Leary et al. “Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation”. In: *Nucleic acids research* 44.D1 (2016), pp. D733–D745.

[46] Heng Li and Richard Durbin. “Fast and accurate short read alignment with Burrows–Wheeler transform”. In: *bioinformatics* 25.14 (2009), pp. 1754–1760.

[47] Petr Danecek et al. “Twelve years of SAMtools and BCFtools”. In: *Gigascience* 10.2 (2021), giab008.

[48] Timo Lassmann, Yoshihide Hayashizaki and Carsten O Daub. “SAMStat: monitoring biases in next generation sequencing data”. In: *Bioinformatics* 27.1 (2011), pp. 130–131.

[49] Rolf Henrik Nilsson et al. “The UNITE database for molecular identification of fungi: handling dark taxa and parallel taxonomic classifications”. In: *Nucleic acids research* 47.D1 (2019), pp. D259–D264.

[50] Amnon Amir et al. “Deblur rapidly resolves single-nucleotide community sequence patterns”. In: *MSystems* 2.2 (2017), pp. 10–1128.

[51] Ryan R Wick et al. “Completing bacterial genome assemblies with multiplex MinION sequencing”. In: *Microbial genomics* 3.10 (2017).

[52] Shifu Chen et al. “fastp: an ultra-fast all-in-one FASTQ preprocessor”. In: *Bioinformatics* 34.17 (2018), pp. i884–i890.

[53] Jennifer Lu et al. “Metagenome analysis using the Kraken software suite”. In: *Nature protocols* 17.12 (2022), pp. 2815–2839.

[54] Jennifer Lu et al. “Bracken: estimating species abundance in metagenomics data”. In: *PeerJ Computer Science* 3 (2017), e104.

[55] The pandas development team. *pandas-dev/pandas: Pandas*. Version latest. Feb. 2020. DOI: 10.5281/zenodo.3509134. URL: https://doi.org/10.5281/zenodo.3509134.

[56] John D Hunter. "Matplotlib: A 2D graphics environment". In: *Computing in science & engineering* 9.03 (2007), pp. 90–95.

[57] Michael L. Waskom. "seaborn: statistical data visualization". In: *Journal of Open Source Software* 6.60 (2021), p. 3021. DOI: 10.21105/joss.03021. URL: https://doi.org/10.21105/joss.03021.

[58] Pauli Virtanen et al. "SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python". In: *Nature Methods* 17 (2020), pp. 261–272. DOI: 10.1038/s41592-019-0686-2.

[59] Kyle Banker et al. *MongoDB in action: covers MongoDB version 3.0*. Simon and Schuster, 2016.

[60] Avi Shayevitz, Keisha Harrison and Chris D Curtin. "Barrel-induced variation in the microbiome and mycobiome of aged sour ale and imperial porter beer". In: *Journal of the American Society of Brewing Chemists* 79.1 (2020), pp. 33–40.

[61] Yvonne Methner et al. "Screening for the brewing ability of different non-Saccharomyces yeasts". In: *Fermentation* 5.4 (2019), p. 101.

[62] Graham H Fleet. *Wine microbiology and biotechnology*. CRC press, 1993.