SCUOLA DI DOTTORATO
UNIVERSITÀ DEGLI STUDI DI MILANO-BICOCCA

Department of Biotechnology and Biosciences

PhD program “Converging Technologies for Biomolecular Systems” (TeCSBi), Cycle XXXIV

Data-driven approaches for biodiversity exploration via DNA metabarcoding data analysis

Agostinetto Giulia

Registration number: 781520

Tutor: Prof. Maurizio Casiraghi

Co-tutor: Prof. Dario Pescini

Coordinator: Prof. Paola Branduardi

ACADEMIC YEAR 2020/2021

Dedicated to the people I love.

Dedicated to Lorenzo, the nicest person I've ever met.

Declaration

I, Giulia Agostinetto, confirm that this is my own work and the use of all material from other sources has been properly and fully acknowledged.

Table of contents

Abstract

Metagenomic approaches have changed the way to study biology and biodiversity in several fields. In particular, technology advancement enables us to determine taxa composition and to study complex biodiversity patterns in very different environments. Nowadays, DNA metabarcoding is a standard procedure, applied on a wide range of fields, from human health to ecology, to industry applications.

In the last few years, 16S rRNA metabarcoding was widely used to study the bacterial community, leading to routine analysis which created huge amounts of data, bringing researchers to develop data mining strategies in order to answer complex biological questions. On the other hand, DNA metabarcoding can be applied also to study Plants, Animals or Fungi, as very different molecular markers have been identified.

In both cases, considering the huge amount of data produced by researchers and available in repositories, a data-driven perspective in managing and exploring DNA metabarcoding data could be useful to collect hidden information and potentially determine undiscovered aspects.

In this PhD dissertation, I focused the attention on a data-centered perspective of DNA metabarcoding data, touching four main points that can enhance and ameliorate the current strategies: i) consider the molecular information obtained from high-throughput DNA sequencing (HTS) and available in public repositories, ii) enhance taxonomy assignment step, iii) investigate new methods for pattern reconstruction and iv) use data as a valuable resource for research.

These four steps can enhance at different levels the potentials of DNA metabarcoding applications, paving the way for standardization procedures for non-bacterial markers and the integration of new data mining and data reuse strategies of metabarcoding data.

1. Introduction to the dissertation contents

1.1 Aims of the dissertation

In this work, I focused on a data-centered perspective of DNA metabarcoding data. Starting from DNA metabarcoding as a method to track species in food and environmental samples, I subsequently dealt with four main points: i) use molecular information as a main source of information when non-bacterial molecular markers are used, ii) issues related to the taxonomy assignment and development of strategies to enhance it, iii) pattern reconstruction via data mining methods and iv) public data as a valuable resource for meta-analysis and data integration projects.

These four aspects can be summarized into two main sections that I will introduce in the next paragraph (1.2 “Introduction to the dissertation”): i) issues related to the taxonomy assignment and ii) data mining approaches particularly focused on 16SrRNA metabarcoding data. Both can contribute to ameliorate different phases of the metagenomics framework, such as the experimental design, data analysis and data interpretation, with the idea to integrate new methods that can enhance experimental strategies and reveal new aspects extracted from DNA metabarcoding data.

1.2 Introduction to the dissertation

DNA metabarcoding is a genetic-based technique used to study a community of organisms through a gene or a set of genes, also called molecular markers, able to define the taxonomy of the individuals from DNA extraction, amplification and sequencing (Porter and Hajibabaei, 2018).

Beside this, other techniques can be used. In particular i) microarrays, which allow to detect the presence of predefined markers from an individual specimen or a community sample; ii) quantitative or digital PCR, exploiting a single marker; iii) organelle sequencing; iv) genome skimming, a low-coverage

genome sequencing strategy for an individual specimen; v) whole-genome sequencing, vi) shotgun metagenomics, where all the genes are sequenced and then analysed with specific bioinformatics strategies (Quince et al., 2017) and vii) metatranscriptomics, allowing the study of community "function" thanks to RNA sequencing (Shakya et al., 2019).

Currently, gene marker surveys are the most popular method to study the biodiversity of a sample. Thanks to the compromise obtained with costs, scalability and coverage (Porter et al., 2018), DNA metabarcoding changed the way of studying biodiversity in several research fields (Deiner et al., 2017; Makiola et al., 2020; McGee et al., 2019). Supported by advances in high throughput sequencing technologies, DNA metabarcoding introduced surprising progresses in surveying prokaryotic and eukaryotic diversity from any type of environments (Makiola et al., 2020; McGee et al., 2019). Due to the implementation of multiplex protocols, highly sample parallelization is nowadays the rule (Herbold et al., 2015) increasing data yield with costs reduction (Cordier et al., 2020; Porter et al., 2018; Pimm et al., 2015; Thomsen et al., 2015; Shokralla et al., 2012).
Advancements currently allow taxa exploration at unprecedented extent, for a time and cost-effective biodiversity tracking (Westfall et al., 2019; Ruppert et al., 2019; Deiner et al., 2017). Moreover, several molecular markers have been studied and used to explore biodiversity, spanning from Plant surveys to the most studied human microbiome (HMP, 2012; Deiner et al., 2017; Nilsson et al., 2019; Ruppert et al., 2019). As different markers, and consequently different taxa, can be used, also pipelines and issues can be affected by the type of marker considered (Deiner et al., 2017; Ruppert et al., 2019).

In this section, I will introduce the main and current issues related to DNA metabarcoding data analysis, which will be discussed in the dissertation based on the work presented. In particular, the work considers two main points: i) issues related to markers linked to the detection of non-bacterial biodiversity

(as 18S, for example), and ii) 16SrRNA analysis. Around these two points, both case studies and the development of new strategies and tools will be presented, in order to improve the current data analysis techniques of metabarcoding data.

1.2.1 A step forward taxonomy assignment of DNA metabarcoding data

First, non-bacterial markers will be considered (**Chapter 2, 3 and 4**). Considering metabarcoding investigations of non-bacterial biodiversity, which comprehend for example ITS2, 18S, COX1 or 16S (but not only; Deiner et al., 2017; Ruppert et al., 2019), a lack of standardization procedures is still present (Deiner et al., 2017; Ruppert et al., 2019). Bioinformatic pipelines have now started to be implemented for standards and reproducibility (Ruppert et al., 2019; Wood-Charlson et al., 2020), also with the obligation to submit the raw data to ENA or SRA (Wood-Charlson et al., 2020) during the publication process. However, as these surveys range from ecology to food applications, a few steps must be taken. In general, metabarcoding bioinformatic pipelines involved four main phases: i) raw sequencing data cleaning, ii) extraction of reliable sequences (such as OTUs or ESVs; Callahan et al., 2017; Deiner et al., 2017), iii) taxonomy assignment and iv) data analysis, intended as post-processing analysis of reliable sequences or assigned sequences (Deiner et al., 2017).

Currently, the main issues related to non-bacterial markers is the taxonomy assignment step (Ruppert et al., 2019; Porter et al., 2018; Deiner et al., 2017). After the extraction of reliable sequences, taxonomy assignment consists in using a reference database to give a name to the sequences obtained, where the final results will be integrated in statistical analysis and used for data interpretation. Beside the choice of algorithm, which is not the focus of this work, a great importance relies on reference databases: SILVA 18S, UNITE, BOLD or PLANITS are a few examples of specific reference database to assign, respectively, sequences obtained from 18S, ITS Fungi, COX and ITS Plants

surveys (Pruesse et al., 2007; Ratnasingham et al., 2007; Nilsson et al., 2019). Specific reference databases are a resource of great value, as they often include sequences checked by experts of the field (Pruesse et al., 2007; Ratnasingham et al., 2007; Nilsson et al., 2019). However, some issues exist: i) currently, reference databases exist only for few molecular markers, such as SILVA for 16S and 18S genes (Pruesse et al., 2007), BOLD for animals and plants (Ratnasingham et al., 2007) or UNITE for Fungi domain (Nilsson et al., 2019); ii) information is not always updated, as the work behind the collections is immense and often non-automatic (Pruesse et al., 2007). Consequently, these data resources are not representative of all the genomic and taxonomic diversity collected to date.

Recently, it has been demonstrated the potential to use NCBI as a primary source of reference sequences (Keller et al., 2020; Ankenbrand et a., 2015; Benson et al., 2008). Also depicted in **Chapter 2**, applications of ITS2 metabarcoding for food traceability and pollinator network reconstructions were reported. In both cases, the NCBI database was used to assign the taxonomy (Frigerio, Agostinetto et al., 2020; Frigerio et al., 2021; Tommasi et al., 2021). In fact, GenBank still resumes the majority of genetic data and their related metadata currently available and it can be considered a constantly updated database of all the sequences deposited all over the world (Keller et al., 2020; Ankenbrand et a., 2015; Benson et al., 2008). As part of the INSDC consortium (Arita et al., 2021), NCBI is also constantly interconnected with ENA and DDBJ databases (Harrison et al., 2019; Ogasawara et al., 2020; Sayer et al., 2019). Thanks to the tools developed to interact with it, such as Entrez (Kans, 2021), it has become easy to access and query. However, such information is not always easy to access without specific bioinformatics skills, which is a limiting factor to a large audience of scientists.

As the taxonomy assignment step remains a pivotal phase to interpret data obtained from sequencing runs, the absence of sequences in the reference database or the biases linked to sequencing itself, may affect the taxonomy

assignment, leading to incorrect or imprecise results (Deiner et al., 2017; Ruppert et al., 2019). In **Chapter 3**, I present a case study in which a multi-marker strategy on two different matrices (air and water) was applied to explore eukaryotic biodiversity at microscale level in a built environment, the case of EXPO2015. Currently, the work is unpublished and under evaluation for submission, but it is an interesting large-scale analysis in which the main metabarcoding issues can be addressed and used as an example for future research. In this work, the taxonomy assignment step was deeply explored, with several attempts (here, only one was mentioned, but several methods were applied). As the markers involved were 18S, ITS2 for Fungi and the intron trnL, for some of them a reference database exists (Pruesse et al., 2007; Nilsson et al., 2019). However, the coverage of the taxonomy was very low in all attempts. Reaching the species level was very difficult, a lot of Unassigned sequences remain. In **Chapter 3**, I demonstrated that in this case, using the reliable sequences (in particular, ESVs; Callahan et al., 2017) achieve a good result in the prediction of the sample area, leading the taxonomy to another level of investigation. Here, the two phases of sample prediction and taxonomy exploration were kept separately and deeply explored, considering the potential of molecular information in microscale biodiversity investigation.

However, giving a name to sequences still remains crucial, also considered the wide variety of applications in which reaching a species level investigation is mandatory to, for example, studies of food fraud detection, diet characterization or using DNA metabarcoding to reconstruct pollinator networks (Frigerio, Agostinetto et al., 2020 a; Frigerio, Agostinetto et al., 2020 b; Tommasi et al., 2021; Bruno et al., 2019). For this reason, I present here the ExTaxsI project (**Chapter 4**). ExTaxsI, which means “Exploring Taxonomy Information”, is a Python project, developed with the aim to help biologists to improve their experimental designs and to promote data exploration and exploitation. It is linked to NCBI taxonomy database (Federhen et al., 2012) and ETE toolkit (Huerta et a., 2016), in order to produce standard formats readable

by most common software that deal with taxonomic information (Bolyen et al., 2019; Rognes et a., 2016; Bengtsson et al., 2015; Mahe et al., 2015; Camacho et al., 2009; Wang et al., 2007), such as QIIME2 platform (Bolyen et al., 2019).
The tool is available as a command line software package (https://github.com/qLSLab/ExTaxsI) and it is under revision. Currently, my colleagues and I are working on making a full-fledged Python library, in order to allow its integration directly into pipelines.
To reach a wide range of researchers and applications, ExTaxsI can be used with any molecular marker, gene name or taxonomic group, making possible to create non-standard marker genes database usable in metagenomic/metabarcoding taxonomic assignment tools (Bolyen et al., 2019). Thanks to the integration of the NCBI query tool (NCBI, 2014), ExTaxsI can reorganize personal datasets in a standardized format in order to easily describe taxonomic variability and geographic provenance of records. In this context, a great importance has been given to data visualization strategies, including in the outputs several plots that not only aggregate and present the research results, but also guide advanced investigations (Kaur et al. 2018; Hardisty et al., 2013).

1.2.2 Enhancing data-driven strategies on 16S rRNA microbiome data

Setting aside non-bacterial markers and taxonomic issues, 16S rRNA metabarcoding data have been widely established as the main source of microbiome information in all the environments, including the human one (Layeghifard et al., 2017; Kyrpides et al., 2020; Wood-Charlson et al., 2021; Su et al., 2020; Bokulich, 2020; Knight, 2018; Gonzales et al., 2018; Mitchell et al., 2020). Consequently, the number of researchers and developers related to this field guaranteed a deep resource of tools and methods (e.g. QIIME2 Bolyen et al., 2019; Greathouse et al., 2019; Bharti and Grimm, 2021; Liu et al., 2021; Amos et al., 2020; Pollock et al., 2018; Callahan et al., 2016; Bolyen et al., 2019; She et al., 2019). Thanks to this, best practices have been partially

standardized, focusing the attention of microbiome research on other aspects, such as post-processing analysis and data mining applications (Bolyen et al., 2019; Kyrpides et al., 2020).

Bioinformatic pipelines have reached a certain uniformity: QIIME2 improvements and diffusion is an example of how the research related to microbiome data has reached a great milestone (Bolyen et al., 2019). Beside this, other tools and pipelines have been implemented (Greathouse et al., 2019; Bharti and Grimm, 2021; Liu et al., 2021; Amos et al., 2020; Pollock et al., 2018; Callahan et al., 2016; Bolyen et al., 2019; She et al., 2019). Reference databases are few and usually updated and curated, as for example SILVA or Greengenes (Pruesse et al., 2007; DeSantis et al., 2006).

Beside this, large amounts of data have been produced and deposited in public databases and more is going to be produced in the near future, as the number of sequencing experiments is exponentially growing (Kyrpides et al., 2020; Wood-Charlson et al., 2021; Vangay et al., 2019). As a consequence, we are facing an increasing adoption of novel large-scale data science approaches to address challenges in microbiome science (Duvallet, 2020; Longo and Drazen, 2016), shifting the attention to post-processing analysis and data mining strategies to districate the complexity of microbiome data (Galimberti et al., 2021; Wood-Charlson, 2020; Ghannam et al., 2021).

One of the main topics related to microbiome research is defining the associations and interactions between species detected with sequencing technologies (Faust et al., 2021; Faust and Raes, 2012; Pasolli et al., 2016; Qu et al., 2019). In the past, great efforts were undertaken to extrapolate relevant associations that could be integrated into biological contexts (Faust et al., 2012; Faust et al., 2021). In general, it was demonstrated that microbial co-occurrence analysis may be an extraordinarily promising approach for studying microbiome associations (Faust and Raes, 2012). Several works explained how co-occurrences reveal indications about ecological processes shaping community structure (Lima-Mendez, 2010), exploring hub species and

potential microorganisms relationships (Berry, 2014). Further, Ma and colleagues (2020) showed how global microbial co-occurrence analysis and network reconstruction may be an encouraging strategy to reveal patterns and explore new mechanisms. However, besides these promising results, transform microbiome data into purposeful biological insights remain challenging, as also demonstrated by different evaluations (Faust et al., 2012; Berry et al., 2014), and open questions still remain (Faust et al., 2021; Ma et al., 2020; Layeghifard et al., 2017; Faust et al., 2012).

Recently, association rule mining (ARM) emerged as a promising technique to study microbiome patterns (Tandon et al., 2016; Naulaerts et al., 2015). Specifically, Tandon and colleagues (2015) have demonstrated the potentials of this technique on two microbiome datasets, in particular the HMP dataset (Turnbaugh et al., 2007) and two prebiotic studies (Xiao et al., 2014; Kato et al., 2014). Despite the apparent simplicity of use, large datasets can produce high numbers of patterns, making their calculation and exploration difficult (Karpinets et al., 2012; Naulaerts et al. 2015; Agrawal et al., 1993; Han et al., 2004). In addition, considering the size and complexity of High-Throughput Sequencing (HTS) 16SrRNA metabarcoding data, interpretation and summarization are not straightforward (Naulaerts et al., 2015).

Due to the large amount of data constantly produced, pattern mining strategies have become essential for researchers to disentangle the high amount of information (Ghannam et al., 2021; Wood-Charlson et al., 2020; Kyrpides et al., 2016). At the same time, tests to establish specific best practices of ARM applications for 16SrRNA metabarcoding data do not exist.

In **Chapter 5** I report and discuss the use of association rule mining (ARM) strategy to study microbial associations. In particular, I implemented a new tool, microFIM (microbial Frequent Itemset Mining; https://github.com/qLSLab/microFim) to promote the use of ARM to explore microbiome patterns. Currently, data mining approaches seem to be newfangled solutions for disclosuring and understanding microbial ecosystems

(Galimberti et al., 2021; Wood-Charlson, 2020; Ghannam et al., 2021). Investigating patterns and exploring their role in functional and predictive aspects are now pivotal to proxy the knowledge of microbial associations, both disentangling interactions and niche specialization (Faust et al., 2012; Ma et al., 2020). Recently, different works related to pattern mining applied to microbiome studies were published, such as MITRE (Bogart et al., 2019), MANIEA framework (Liu et al., 2019) and the work of Tandon and colleagues (2016). Nevertheless, as also highlighted by the work of Faust (2021), applying such an algorithm still has its limitations and, despite the efforts of recent works, guidelines for microbiome data applications have not been completely defined (Faust et al., 2021; Naulaerts et al., 2015).

With this work, I wanted to shed light on the strengths and weaknesses of pattern mining strategy into the study of microbial patterns, in particular from 16SrRNA microbiome datasets. In detail, I report the key steps that must be considered to apply ARM consciously on 16SrRNA microbiome data and propose a user-friendly and open source Python tool that accepts as input common microbiome file formats, such as taxa tables. In addition, microFIM merges the results of ARM analysis with the typical microbiome outputs, aiming at creating a bridge between microbial ecology research and ARM technique.

However, the development of new strategies and the need to make results statistically sound are strictly dependent from the availability of data to use as tests, both to create standard datasets and perform meta-analysis (Duvallet et al., 2017; Bisanz et al., 2019; Kosti et al., 2020). Recent advancements in data integration and data reuse strategies may enhance the exploration of microbial patterns from large-scale studies (Ghannam et al., 2021; Jordan et al., 2015; Ma et al., 2020; Su et al., 2020).

Machine learning strategies can be applied to perform powerful prediction tasks on metagenomics data (e.g. disease-prediction based on microbiome composition). However, these strategies require a large amount of data to train and test models, making the integration and harmonization of multiple datasets

a necessary step (Jordan and Mitchell, 2015; Ghannam and Techtmann, 2021). In this way, the availability of large-scale sequencing data can enable microbiology researchers to ask new questions and develop new strategies to study the human-associated microbial communities (Wood-Charlson et a., 2021; Su et al., 2020).

In detail, several research groups have been proposing different sources of microbiome data: initiatives like the Human Microbiome and the Integrative Microbiome Projects (Gevers et al., 2012; Proctor et al., 2019), MicrobiomeDB (Oliveira et al., 2018), HumanMetagenomeDB (Kasmanas et al., 2021), curatedMetagenomicData (Pasolli et al., 2017), the ML Repo (Vangay et al., 2019), QIITA portal (Gonzales et al., 2018), or the MG-RAST portal (Wilke et al., 2016) suggested both data management infrastructures and frameworks to guarantee data accessibility and reuse.

However, it is hard to have a comprehensive collection of all the dataset regarding a specific topic, as also the process of data FAIRification is still in its infancy (Wood-Charlson et al., 2020; Vangay et al., 2019).

In **Chapter 6** I report a case study related to these issues specifically focusing on a type of microbiome data: the skin microbiome (Dimitriu et al., 2019). My colleagues and I developed a skin microbiome collection of datasets, called SKIOME Project, including all the sequencing datasets of 16S rRNA publicly available and sequenced from 2012. Alongside the main objective of the work, we provide insights related to the metadata collection and harmonization. Through a framework that integrates different tools to access INSDC databases, we have been working on the retrieval of metadata important to evaluate the datasets from a biological and technical point of view. Consequently, we noticed the inconsistencies and biases related to microbiome data submission to public repositories.

The lack of metadata and the presence of datasets with missing or inconsistent information can reduce the interpretability of the data generated, influencing the understanding of microbial dynamics and ecological patterns

(Wood-Charlson et al., 2020; Su et al., 2020; Greenhouse et al., 2019). Moreover, this fact impact negatively on the idea of FAIR (Findable, Accessible, Interoperable, and Reusable) principles, supported within the National Microbiome Data Collaborative and FAIR Microbiome community (https://www.go-fair.org/implementation-networks/overview/fair-microbiome) (Wood-Charlson et al., 2020; Vangay et al., 2019) to promote data discovery and reuse in the microbiome field.

Beside this, the main output of our work constitutes a valuable resource for researchers interested in performing meta-analyses with human skin microbiome data, who can explore our collection to find a list of datasets that can be integrated to answer old and new biological questions.

1.3 Structure of the dissertation

This dissertation is organized into seven main chapters, of which the first is the current introductory chapter.

Chapter 2 is focused on DNA metabarcoding applications and the current issues related to non-bacterial markers and, separately, the application of 16S rRNA marker, discussing two joint first-authored published manuscripts, three published collaborative studies and a collaborative review to which I contributed for a specific paragraph. In detail, I reported:

- Frigerio, J., **Agostinetto, G.**, Galimberti, A., De Mattia, F., Labra, M., & Bruno, A. (2020). Tasting the differences: microbiota analysis of different insect-based novel food. Food Research International, 137, 109426.
- Frigerio, J., **Agostinetto, G.**, Sandionigi, A., Mezzasalma, V., Berterame, N. M., Casiraghi, M., ... & Galimberti, A. (2020). The hidden 'plant side' of insect novel foods: a DNA-based assessment. *Food Research International*, *128*, 108751.

- Bruno, A., Sandionigi, A., **Agostinetto, G.**, Bernabovi, L., Frigerio, J., Casiraghi, M., & Labra, M. (2019). Food tracking perspective: DNA metabarcoding to identify plant composition in complex and processed food products. Genes, 10(3), 248.
- Tommasi, N., Biella, P., Guzzetti, L., Lasway, J. V., Njovu, H. K., Tapparo, A., …, **Agostinetto, G.**, ... & Galimberti, A. (2021). Impact of land use intensification and local features on plants and pollinators in Sub-Saharan smallholder farms. Agriculture, Ecosystems & Environment, 319, 107560.
- Frigerio, J., **Agostinetto, G.**, Mezzasalma, V., De Mattia, F., Labra, M., & Bruno, A. (2021). DNA-Based Herbal Teas' Authentication: An ITS2 and psbA-trnH Multi-Marker DNA Metabarcoding Approach. *Plants*, *10*(10), 2120.
- Galimberti, A., Bruno, A., **Agostinetto, G.**, Casiraghi, M., Guzzetti, L., & Labra, M. (2021). Fermented food products in the era of globalization: tradition meets biotechnology innovations. Current Opinion in Biotechnology, 70, 36-41.

"The hidden 'plant side' of insect novel foods: a DNA-based assessment" and "Tasting the differences: microbiota analysis of different insect-based novel foods". Here, we performed a DNA metabarcoding and bioinformatic pipeline to trace both microbiome and plant species in the context of insect novel foods, in order to explore both contaminants, frauds and determine microbial and plant signatures related to insect species. Moreover, three collaborative studies are discussed: "DNA-based herbal teas authentication: a ITS2 and *psbA-trnH* multi-marker DNA metabarcoding approach", "Impact of land use intensification and local features on plants and pollinators in Sub-Saharan smallholder farms" and "Food tracking perspective: DNA metabarcoding to identify plant composition in complex and processed food products". With these works, I show the potentials of DNA metabarcoding in biomonitoring projects and food industry applications. Finally, I report my main contribution

for the review “Fermented food products in the era of globalization: tradition meets biotechnology innovations”, discussing the potentials and future perspectives of microbiome research applied to fermented foods.

Chapter 3 reports one first-authored submitted work, consisting of a multi-marker DNA metabarcoding case study applied to a large-scale event, the EXPO2015, in which molecular information seemed to be a better strategy in the study of eukaryotic communities in a built environment. In detail:

- **Agostinetto, G.**, Bruno, A., Sandionigi, A., Brusati, A., Manzari, C., Chiodi, A., … & Casiraghi, M. Dealing with the promise of metabarcoding in mega-event biomonitoring: EXPO2015 unedited data. bioRxiv. Under revision.

Chapter 4 reports one first-authored accepted manuscript updated with the final analysis related to the exploration of molecular data to enhance taxonomy assignment step. In detail:

- **Agostinetto, G.**, Sandionigi, A., Chahed, A., Brusati, A., Parladori, E., Balech, B., ... & Casiraghi, M. (2021). ExTaxsI: an exploration tool of biodiversity molecular data. GigaScience.

In particular, with “ExTaxsI: an exploration tool of biodiversity molecular data”, I addressed caveats of taxonomy assignment for non standard molecular markers. Through three main case studies, we demonstrated the potentials of ExTaxsI to explore molecular data and visualize molecular information via a taxonomy centered perspective, in order to both ameliorate experimental design and data interpretation.

Chapter 5 is focused on pattern mining reconstruction, reporting one first-authored published manuscript. In detail:

- Agostinetto, G., Sandionigi, A., Bruno, A., Pescini, D., & Casiraghi, M. Extending association rule mining to microbiome pattern analysis: tools

and guidelines to support real applications. Frontiers in Bioinformatics, 77.

Through this work, I introduce the potentials and caveats of association rule mining technique (or frequent itemset mining) to extract microbiome patterns. In addition, I present microFIM tool (microbial Frequent Itemset Mining) a bioinformatic tool implemented to easily integrate ARM on common microbiome pipelines.

Chapter 6 reports one first-authored submitted manuscript related to the specific field of skin microbiome data. In detail:

- **Agostinetto, G.**, Bozzi, D., Porro, D., Casiraghi, M., Labra, M., & Bruno, A. (2021). SKIOME Project: a curated collection of skin microbiome datasets enriched with study-related metadata. bioRxiv. Under revision.

In this work, I present the SKIOME Project, a collection of 16S rRNA metabarcoding datasets obtained with a semi-automatic framework developed in this work to reconstruct metadata of sequencing datasets. In parallel, I introduce data reuse strategies to perform meta-analysis and data mining approaches.

Chapter 7 is dedicated to main conclusions and future perspectives.

2. DNA metabarcoding: applications and issues based on real case studies

In this chapter I will introduce two main works related to DNA metabarcoding applications into insect-novel food characterization, both considering plant and bacterial markers (Frigerio, Agostinetto et al., 2020; Frigerio, Agostinetto et al., 2020). In addition, I will discuss my contribution in three collaborative works (Frigerio et al., 2021; Tommasi et al., 2021; Bruno et al., 2019), of which I report the abstracts. Full papers are provided at Supplementary Data link.
For a complete list of all the works cited in this section, see **Chapter 1.3** "Structure of the dissertation".

2.1 The hidden 'plant side' of insect novel foods: a DNA-based assessment

2.1.1 Introduction

Traditionally, edible insects are consumed in large parts of the world like Africa and Asia. In the last few years, they have increased in popularity as trendy foods in many Western countries (Sun-Waterhouse et al., 2016). Being rich in essential nutrients, they represent an important source of energy for human diets (Rumpold & Schlüter, 2013). Mean estimates show that the energy level of insects is around 400–500 kcal per 100 g of dry matter, making it comparable with other protein sources (Payne, Scarborough, Rayner, & Nonaka, 2016). Protein is probably the most significant component of edible insects, with an average value ranging from 30% to 65% of the total dry matter. However, edible insects are also rich in micronutrients such as iron, zinc, and calcium (Dobermann, Swift, & Field, 2017), and preliminary studies have shown that insect farming has a lower environmental footprint compared to other livestock animals (Oonincx & de Boer, 2012). Moreover, a recent study demonstrated that edible insects could represent a potential source of antioxidants with

positive effects on human health (Di Mattia, Battista, Sacchetti, & Serafini, 2019).

In very recent years, the use of insects in food fortification is emerging as a means of producing nutritious and acceptable food products for human consumption (Myers & Pettigrew, 2018). In many countries a lot of insect-based products have recently become commercialized (e.g. pasta, biscuits, and energy bars based on insect flour in combination with fruits, nuts, and other ingredients), and their consumption trend is expected to grow steadily.

In Europe, edible insects are placed in the category of novel foods, and since the beginning of 2018, the Regulation (EU) 2015/2283 entered into force in attempt to regulate the production and safety of novel foods in Europe. This regulation establishes the requirements that enable Food Business Operators to bring new foods into the EU market, while ensuring high levels of food safety to the consumers. However, concerning the possible risks for human health caused by insect-based products, the European Food Safety Authority has published an initial assessment (EFSA, 2015), and it concluded that attention should be placed to the possible occurrence of biological and chemical hazards caused by novel foods. Insects ingredients may cause allergic symptoms (Mazzucchelli et al., 2018; Pali-Scholl et al., 2019; Van der Fels-Klerx, Adamse, Punt, & Van Asselt, 2018), and there is a cross-reactivity/cosensitisation between edible insects and crustaceans (Ribeiro, Cunha, Sousa-Pinto, & Fonseca, 2018). Moreover, at the microbiological level, the microbiota of insects is highly complex and, apart from the body surface and the mouthparts, the maximum microorganism diversity is in the mid-gut with poor or no data about its effect on consumer health (Schluter et al., 2017; Walia, Kapoor, & Farber, 2018).

Another element that can negatively impact the safety of insect-based novel foods is the composition and quality of feed used to raise the insects and their rearing conditions. In many cases, these feeds are composed of vegetables and very few ingredients of animal origin, such as fishmeal and egg and milk

based-products (EU No. 2017/893). However, to date no analytical systems are available to control the diet of marketed insects.

In general, according to a recent review (Schlüter et al., 2017), there is a lack of scientific knowledge about insect processing to ensure safe novel foods. Moreover, most insect-based products are consumed as flour or processed items (e.g. pasta and bars); therefore the insect morphological traits cannot be used to verify the product authenticity and consequently its safety. Considering the high price of insect flours, we cannot exclude the occurrence of deliberate or intentional counterfeits (i.e. mix with low-cost flours, such as maize) as happens with other high-value food products, such as saffron (Petrakis, Cagliani, Polissiou, & Consonni, 2015).

Based on these assumptions, in this study we adopted, for the very first time, a DNA-based approach to analyse commercial novel food products to verify the identity of declared insect species and characterize trace amounts of plant material occurring in the same products in order to derive the source of the substrate, such as from insect diets and litter, with particular attention paid to putative elements of allergenic concern. Particularly, we used a region of the mtDNA COI marker to identify the insect species (DNA barcoding), and the nuclear ITS2 region to characterize insect diets and the vegetal composition of the tested products (High Throughput Sequencing HTS DNA meta-barcoding). Moreover, the same approach can be used to verify product contamination and counterfeiting insect flour with low cost vegetable flours. We prepared six mock mixtures composed of wheat and soybean flours at different concentrations to also verify the limit of detection of our approach in the context of possible contamination of insect-based novel foods.

2.1.2 Materials and methods

2.1.2.1 Insect commercial food products

A total of 13 commercial insect-based products, namely flour (n = 3), pasta (n = 3), crackers (n = 2), protein bars (n = 4), and pet food (n = 1) were purchased

via e-commerce from six different companies (Table 1). These categories offer an almost complete representation of the insect novel foods available in Europe. Based on the label information, these products contained only one insect species each, for a total of five species belonging to the orders Orthoptera (Acheta domesticus and Gryllodes sigillatus), Diptera (Hermetia illucens) and Coleoptera (Alphitobius diaperinus and Tenebrio molitor). Reference insect samples (RI) for each species were also retrieved (Supplementary Table S1) from a certified pet shop (AGRIPETGARDEN S.r.l., Conselve, Italy).

Code	Category	Label Declared insect	Label Declared Ingredients
F_001	Flour	*Tenebrio molitor*	_
F_002	Flour	*Gryllodes sigillatus*	_
F_003	Flour	*Alphitobius diaperinus*	_
FP_004	Pasta	*Alphitobius diaperinus* (14%)	*Triticum durum, Ocimum basilicum* (1.5%); organic powdered egg whites.
FP_005	Pasta	*Alphitobius diaperinus* (14%)	*Triticum durum*; organic powdered egg whites.
FP_006	Pasta	*Tenebrio molitor* (10%)	*Oryza sativa* (43); *Cicer arietinum* (43%); organic powdered egg whites (4%).
FP_007	Cracker	*Acheta domesticus* (14%)	*Triticum aestivum*; *Sesamum indicum* (6%); *Olea europaea.*
FP_008	Cracker	*Tenebrio molitor* (10%)	*Triticum aestivum*; *Cocos nucifera; Avena sativa*; *Sesamum indicum* (12%); *Porphyra* sp. (1.2%).
FP_009	Protein bar	*Acheta domesticus* (5.2%)	*Phoenix dactylifera*; *Prunus dulcis*; *Musa* spp. (11%); *Theobroma cacao* (9%); *Vaccinium macrocarpon* (8%); *Anacardium occidentale*; *Cannabis sativa.*
FP_010	Protein bar	*Acheta domesticus*(5.5%)	*Phoenix dactylifera*; *Prunus dulcis*; *Prunus armeniaca* (22%); *Pisum sativum*; *Helianthus annuus*; *Lycium barbarum*(4.5%); *Salvia hispanica* (3.5%).
FP_011	Protein bar	*Acheta domesticus* (20%)	*Arachis hypogaea* (34%); *Cannabis sativa*; *Theobroma cacao*; *Agave* sp; *Beta vulgaris*; *Cinnamomum* sp. (1%).
FP_012	Protein bar	*Acheta domesticus* (10%)	*Ananas comosus* (30%); *Phoenix dactylifera*; *Anacardium occidentale*; *Cocos nucifera*; *Plantago* sp.; *Citrus limon.*
FP_013	Pet food	*Hermetia illucens* (25%)	*Ipomoea batatas*; *Saccharomyces cerevisiae*; *Lycium barbarum* (1%); *Rosmarinus officinalis*; plant base glycerin.

Table 1. List of analysed insect-based products. For each sample, information found on the label about the category, the species of insect, and the plant ingredients are reported. F (flour); FP (Food Product).

2.1.2.2 Mock mixtures

To test for the efficacy of DNA metabarcoding in characterizing the composition of the insect-based products at the qualitative and semi-quantitative levels, six mock mixtures were prepared (Supplementary Table S2). These were composed of insect flour (T. molitor) mixed with wheat (Triticum aestivum) flour (20 ppm, 200 ppm and 500 ppm of gluten) or soybean (Glycine max) flour (50 ppm, 200 ppm and 500 ppm of soybean proteins). Wheat and soybean flour concentrations were defined based on the alert threshold for allergens according to the EU regulation (No. 828/2014) and to Ballmer-Weber et al. (2007), respectively.

2.1.2.3 DNA extraction

For insect-based products and mock mixtures (see Table 1 and Supplementary Table S2), purified gDNA was obtained starting from 250 mg of samples by using the DNeasy PowerSoil Kit (QIAGEN, Hilden, Germany) following the manufacturer's instructions. All samples were prepared in three replicates. For reference insects (see Supplementary Table S1), purified gDNA was obtained starting from 25 mg of samples by using the DNeasy Blood & Tissue Kit (QIAGEN, Hilden, Germany) following the manufacturer's instructions. Purified DNA was checked for concentration and purity by using a Qubit 2.0 Fluorometer and Qubit dsDNA HS Assay Kit (Invitrogen, Carlsbad, California, United States).

2.1.2.4 Insect identification by DNA barcoding

The 658 bp mtDNA COI region was used to authenticate the animal component in the purchased insect-based products. This region was amplified and sequenced for all 13 samples and for the reference insects by using primer pairs described by Folmer, Black, Hoeh, Lutz, and Vrijenhoek (1994)

(Supplementary Table S3) and the protocol described in Bellati et al. (2014). The obtained sequences were submitted to the international GenBank through the EMBL platform (see Supplementary Table S1 for accession numbers). Each sequence was taxonomically assigned to the reference species (or to the declared one in the case of food items) by looking at the nearest matches with the BLAST algorithm using the following cut-off values/maximum identity > 99% and query coverage of 100%.

2.1.2.5 Library preparation and sequencing

To characterize the plant composition of the investigated insect-based products and mock mixtures, the obtained gDNA extracts were sequenced at the DNA barcode ITS2 region (Chen et al., 2010). Amplicons were obtained using the same approach described by Biella et al. (2019) with Illumina adapter (Supplementary Table S3) using puReTaq Ready-To-Go PCR beads (GE Healthcare Life Sciences, Italy) following the manufacturer's instructions in a 25 µL reaction containing 1 µL 10 µM of each primer and up to 50 ng of gDNA. PCR cycles consisted of an initial denaturation step for 5 min at 94 °C, followed by 40 cycles of denaturation (30 s at 94 °C), annealing (30 s at 56 °C), and elongation (1 min at 72 °C), and, hence, a final elongation at 72 °C for 10 min. Amplicon DNA was checked for concentration by using a Qubit 2.0 Fluorometer and Qubit dsDNA HS Assay Kit (Invitrogen, Carlsbad, California, United States) (Supplementary Table S4) and amplicon length was measured by comparison against QX DNA Size Marker using the Qiaxcel Automatic electrophoresis system (QIAGEN, Hilden, Germany). Samples were sequenced by the Center for Translational Genomics and BioInformatics (Milan, Italy). The sequencing was carried out on the MiSeq sequencing platform (Illumina, San Diego, CA, USA) with a paired-end approach (2 × 300 bp).

2.1.2.6 Bioinformatic analysis

Raw Illumina reads were paired and merged using the PEAR algorithm (Zhang, Kobert, Flouri, & Stamatakis, 2013). Pre-processing was performed using VSEARCH 2.0 algorithm (Rognes, Flouri, Nichols, Quince, & Mahé, 2016): reads were filtered out if ambiguous bases were detected and lengths were outside the bounds of 100 bp; moreover, an expected error = 1 was used as an indicator of read accuracy. Sequences were then dereplicated using –derep_fulllength. In order to decrease the false positive rate in the sequence population, a chimera detection analysis was performed on the obtained reference sequences. Since there is no reference database for ITS2 region for chimera detection, we used –uchime_denovo algorithm that carries out a de novo analysis without a reference. Plant features were obtained using –cluster_fast algorithm with a 100% sequence identity with at least a depth of 500x for each feature. A random sequence was chosen as the representative sequence of the cluster. Subsequently, DNA metabarcoding analysis was performed using the plugins of the QIIME2 suite (https://docs.qiime2.org/). The taxonomic assignment of the representative sequences was carried out using the classify-consensusblast plugin implemented in QIIME2 (Camacho et al., 2009) against the local database, built with downloaded ITS2 sequences available in NCBI at 29th of January 2019, adopting a percent identity > 0.99 and a query coverage > 0.90. To evaluate the occurrence of contaminants (T. aestivum and G. max) in insect flour (F_001) and their relative abundance, we generated a heat map representation of the significant discriminatory features (plant species) obtained with the bioinformatic pipeline. Sample and feature axes were also organized using a clustering approach. The heat map was generated with the feature-table QIIME2 plugin (McDonald et al., 2012). To evaluate the sensitivity of the approach in detecting species based on feature depth, we performed a qualitative analysis considering the results of the taxonomy assignment described previously, assuming a depth of 500×, 100×, and 25× for each OTU, respectively.

Python script (v.3; Pandas and NumPy libraries) was used to calculate sequence abundance weighted OTU and taxa overlap respectively (Wen et al., 2017) among the technical replicates. To evaluate significative differences among samples belonging to mock mixtures, a PERMANOVA test (permutation-based ANOVA, PerMANOVA) with 999 permutation-based Bray Curtis distance metrics (Faith, Minchin, & Belbin, 1987) was performed using the diversity QIIME2 plugin, considering both OTUs and taxa composition. PerMANOVA Pairwise contrast was performed through the beta-group-significance command of diversity plugin (Anderson, 2001).

2.1.2.7 ELISA assays

The three flour samples (F001, F002 and F003) and the mock samples prepared with T. aestivum (MT_20 MT_500) were also analysed by RIDASCREEN® Gliadin kit (R-Biopharm AG, Darmstadt, Germany, prod. no. R 7001), a sandwich enzyme-linked immunosorbent assay (ELISA) kit for gluten detection. The assay is based on the monoclonal antibody R5 (Méndez, Vela, Immer, & Janssen, 2005; Valdés, García, Llorente, & Méndez, 2003), which is specific for gliadin-fractions from wheat. The detection limit for gluten is 3 ppm (mg/kg). The manufacturer's instructions were followed.

2.1.3 Results

2.1.3.1 DNA barcoding authentication of insect ingredients

Good DNA yield (20–40 ng/µl) was obtained from all the replicates of the 13 collected samples and from the reference insect samples as well (see Supplementary Table S4). The mt COI DNA barcoding sequencing results indicate that all the tested insect-based products were correctly labelled concerning insect composition with the only exclusion of FP009 and FP010 (protein bars). In both cases, the DNA barcoding analysis did not find

occurrences of Acheta domestica as expected, but the sequences matched with the COI of the food parasite species Ectomyelois ceratoniae (Insecta: Pyralidae).

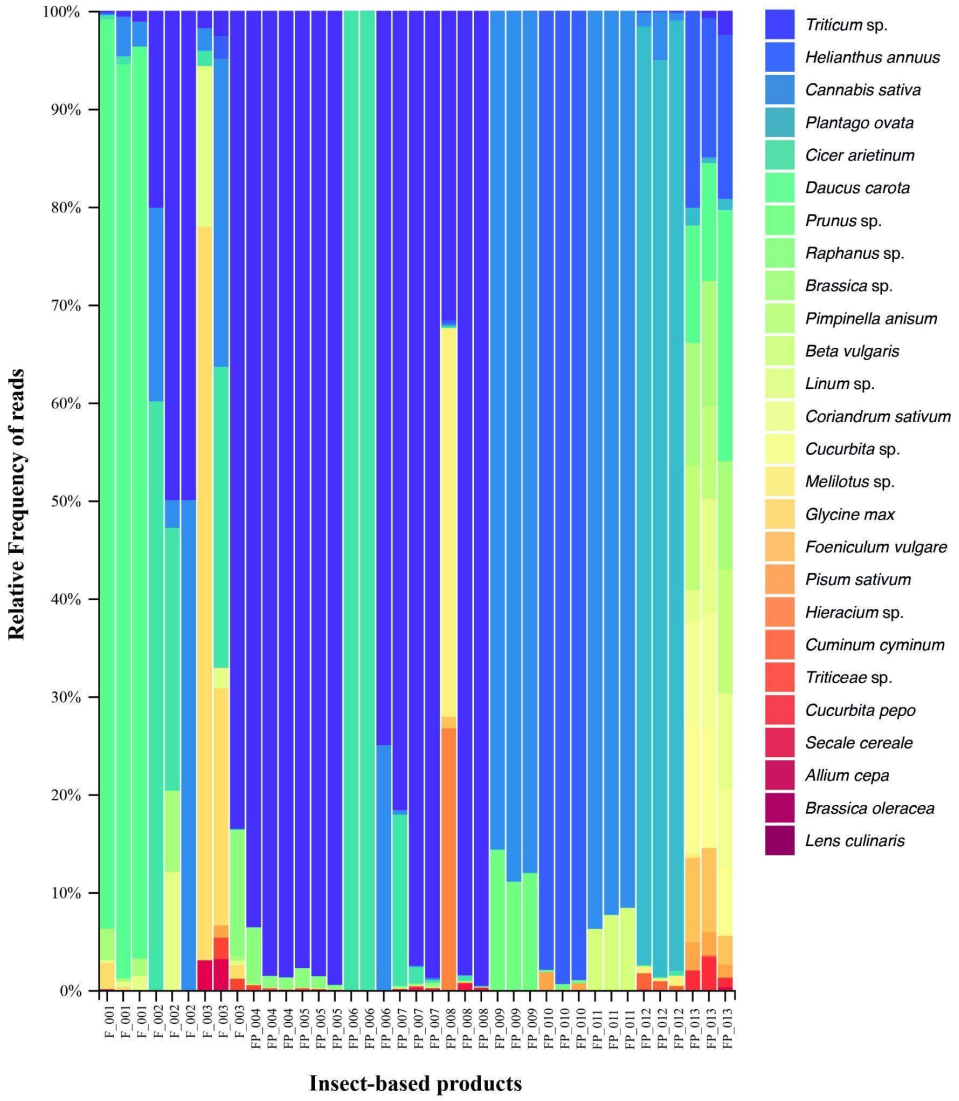

Figure 1. Relative abundance of the plant taxa recovered in the 13 insect-based products through ITS2 metabarcoding sequencing.

2.1.3.2 DNA metabarcoding characterization of plant composition

The High Throughput Sequencing step produced a total of 8,142,444 raw reads, with an average of 126,299 reads (SD = 108,789; range 334–482,500) per sample. After the merging, quality filtering and dereplication steps, we retrieved a total of 868,414 reads, with an average of 13,677 reads (SD = 11,453; range 5–45,066) per sample. Details on the average and standard deviation statistics about raw and filtered reads obtained for each sample, considering replicates, are provided in Supplementary Table S4. After chimera detection and 100% cluster identity, with a depth of at least 500x for each feature, we obtained 120 OTUs (Operational Taxonomic Units). Negative controls for library sequencing were not included in the analysis since the very low amount of DNA copies. OTUs and taxa diversity were analysed separately for technical replicates at each sample. Both OTUs and taxa overlap (calculated with the weight of reads per OTUs and taxa) maintained a mean of 90% for all insect-products (with a standard deviation of 0.26).

The obtained taxonomic assignment and the distribution of the assigned taxa among the sample data are depicted in Figure 1. Overall, 120 OTUs were assigned to at least 26 plant species. Most of the assigned OTUs reached the species taxonomic level, however, in some genera, such as Triticum and Brassica, the low interspecific variability did not allow the species to be identified.

The three flour samples were largely different. Specifically, F_001 is rich in Daucus carota reads (93.10%) followed by Brassica sp. (1.80%) and Glycine max (0.99%), F_002 contains mainly Cicer arietinum (43%), Triticum sp. (35%), and Brassica sp. (4%) reads. Finally, the F_003 flour shows many reads of Glycine max (33%), Triticum sp. (30%), and Cicer arietinum (11%). Moreover, all the flour samples showed a variable relative read abundance of Cannabis sativa (range 2.18–11%) and Linum sp. (range 0.76–6%).

The assigned plant taxa in the 13 samples were grouped in Expected species (E), that include taxa listed on the product label, Rearing Substrate (RS) which

includes the putative plant used both as feed and litter for insect farming, and Not Expected (NE) encompassing all the remaining species. Table 2 indicates the distribution of the assigned species among the above-mentioned categories.

Concerning the processed insect-based products (from FP_004 FP_013), the plant composition was clearly different among the tested categories. In the case of pasta, the first two samples (FP_004 and FP_005) mainly consisted of Triticum sp. (97% and 99% of reads respectively) according to the label information (Table 1). Interestingly, reads of the NE Raphanus sp. (3% and 1% respectively) were found. Regarding the F_005 pasta sample, two out of the three replicates were largely dominated by reads of the expected species Cicer arietinum (67%). Strangely, no OTUs belonging to Oryza sativa were found.

In the two cracker products, the reads of Triticum sp. were mainly detected (range 74–92.68%), followed by several NE species such as Cicer arietinum (range 1–6.5%), Secale cereale (range 0.19–1%), and other taxa occurring in traces (see Table 2).

In the four protein bars (belonging to two different companies), very few OTUs were obtained, and in most cases reads of E species were not found. Finally, the composition of pet food was very complex, and OTUs belonging to > 12 species were detected (see Table 2).

Overall, the plant taxa distribution among the five insect-based product categories is schematized in Figure 2. The contribution of plant diet is appreciable in the insect flours which are the purest and least processed products. In Pasta, Crackers, and Protein bars the HTS analysis reveals the expected ingredient as the most abundant. Finally, in the pet food we detected the highest percentage of rearing substrate plant species.

Code	Detected ingredients		
	Expected (E)	Rearing substrate (RS)	Not expected (NE)
F_001	*T. molitor*	*Daucus carota* (93.1%); *Cannabis sativa* (2.2%); *Brassica* (1.8%); *Glycine max* (1%); *Linum* sp. (0.8%); *Cicer arietinum* (0.4%); *Triticum* sp. (0.6%); *Helianthus annuus* (0.1%).	*Melilotus* sp. (0.01%).
F_002	*G. sigillatus*	*Cicer arietinum* (43%); *Triticum* sp. (35%); *Cannabis sativa* (11%); *Linum* sp. (6%); *Brassica* (4%).	_
F_003	*A. diaperinus*	*Glycine max* (33%); *Triticum* sp. (30%); *Cannabis sativa* (11%); *Cicer arietinum* (11%); *Linum* sp. (6%); *Helianthus annuus* (1%); *Pisum sativum* (1%).	*Raphanus* sp. (4%); *Allium cepa* (1%).
FP_004	*A. diaperinus; Triticum* sp.(97%).	_	*Raphanus* sp. (3%).
FP_005	*A. diaperinus; Triticum* sp. (99%).	_	*Raphanus* sp. (1%).
FP_006	*T. molitor; Cicer arietinum* (67%).	*Triticum* sp. (25%); *Cannabis sativa* (8%).	_
FP_007	*A. domesticus; Triticum* sp. (92.7%).	*Cicer arietinum* (6.5%); *Cannabis sativa* (0.2%); *Brassica* sp. (0.2%); *Helianthus annuus* (0.06%).	*Beta vulgaris* (0.01%); *Raphanus* sp. (0.2%); *Secale cereale* (0.2%).
FP_008	*T. molitor; Triticum* sp. (74%).	*Cicer arietinum* (1%).	*Melilotus* sp. (13%); *Hieracium* sp. (9%); *Foeniculum vulgare* (1%); *Secale cereale* (1%).
FP_009	*Cannabis sativa* (87%); *Prunus* (12%)	_	*Ectomyelois ceratoniae*
FP_010	*Helianthus annuus* (97%); *Pisum sativum* (2%); *Prunus* sp. (1%).	_	*Ectomyelois ceratoniae*
FP_011	*A. domesticus; Cannabis sativa* (93%); *Beta vulgaris* (7%).	_	_
FP_012	*A. domesticus; Plantago ovata* (96%).	*Cannabis sativa* (3%).	*Melilotus* sp. (1%); *Cuminum cyminum* (1%).
FP_013	*H. illucens*	*Helianthus annuus* (17%); *Daucus carota* (17%); *Pimpinella anisum* (12%); *Brassica* sp. (12%); *Linum* sp. (8%); *Pisum sativum* (2%); *Triticum* sp. (2%).	*Cucurbita* (11%); *Coriandrum sativum* (10%); *Foeniculum vulgare* (7%); *Cucurbita pepo* (2%); *Plantago ovata* (1%).

Table 2. List of the detected ingredients based on DNA metabarcoding assignment. For each insect-based product, the Expected species (E), the taxa belonging to the Rearing Substrate (RS), and the Not Expected (NE) species are indicated. The percentage values refer to the relative abundance of HTS ITS2 reads for each recognized ingredient.

To better characterize the composition of the processed products (from FP_004 to FP_013), we tested the identification performance, reducing the reads filtering parameter from 500 to 100 and 25 reads per OTU. Figure 3 shows the distribution of plant taxa among the three categories (i.e. E, RS, and NE) at different filtering thresholds. As expected, the whole number of assigned plant taxa increased with the decreasing threshold. Furthermore, the number of NE species increases dramatically compared to the increase of the Expected species.

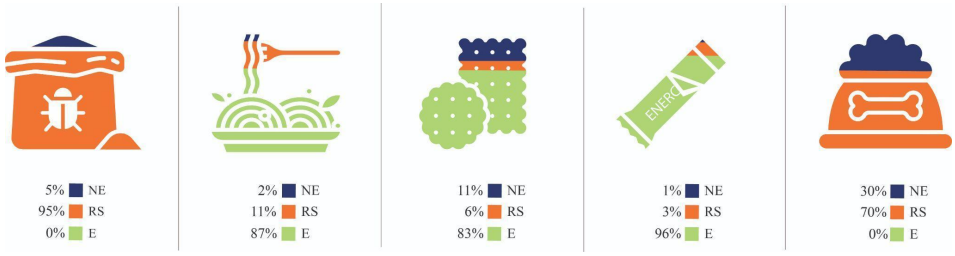

Figure 2. Comprehensive graphic summary of the detected plant taxa in the five insect-based product categories (from left: Flour, Pasta, Crackers, Protein Bars and Pet Food). E: expected species, RS: Rearing substrate species, NE: Not expected species.

2.1.3.3 DNA metabarcoding characterization on flour mock mixtures

The results of the HTS DNA metabarcoding analysis performed on the six flour mock mixtures are shown in the heat map diagram of Figure 4. Pure insect flours (F_001) cluster together and show the occurrence of several RS plants, especially D. carota. In these samples, the two contaminants T. aestivum and G. max are absent. Conversely, in both the types of flour mock mixture, the abundance of wheat and soybean reads increases along with the amount of the admixed contaminant flours. In the case of samples MT_200 and MT_500, the frequency of wheat reads reaches the maximum level.

The PerMANOVA analysis shows that there is no statistical difference between samples belonging to the two categories of mock mixtures, both considering OTUs and taxa composition.

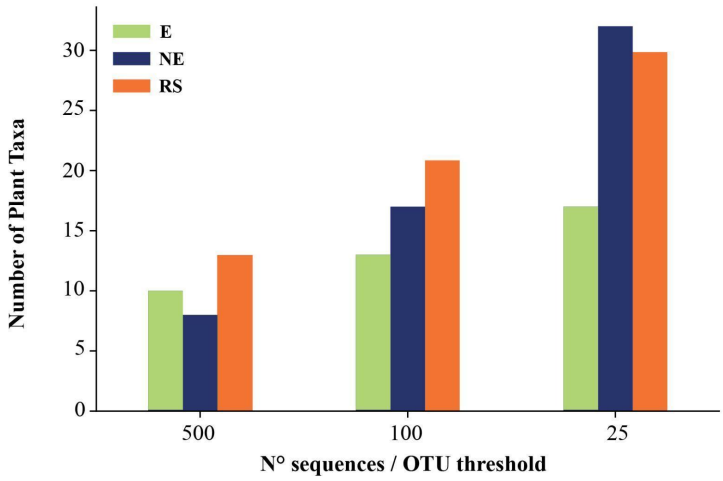

Figure 3. Numbers of plant taxa recovered in the 13 insect-based products and mocks (E, RS, and NE) through ITS2 metabarcoding sequencing using different thresholds of numbers of sequences per OTU (500, 100, and 25, respectively).

2.1.3.4 ELISA assays

Flour samples (i.e., F_001, F_002 and F_003) and the mocks prepared with T. aestivum flour (i.e., MT_20, MT_200 and MT_500) were analysed for gluten detection. Sample F_002, the mock MT_200, and MT_500 were positive to gluten content. (Details are provided in Table 3).

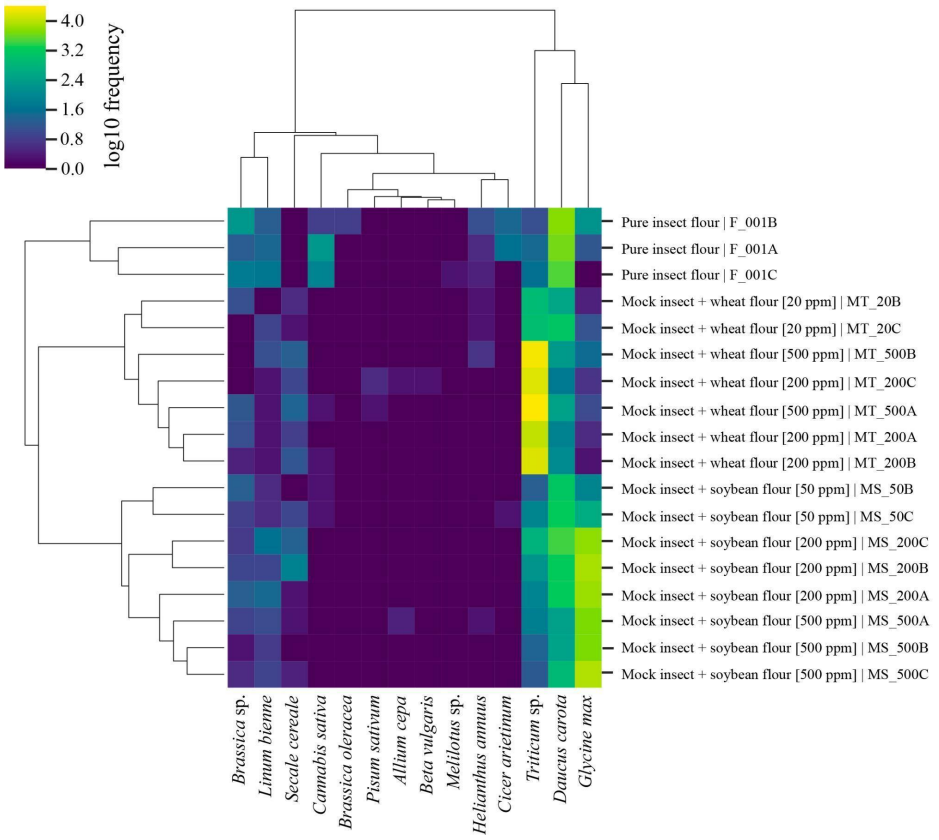

Figure 4. Heat map diagram showing the abundance of plant taxa in the pure insect flours and mock mixtures. All the replicates, with the only exception of MT_20A and MS_50A, are shown. Color shading in the heat map indicates the abundance, expressed as log10 frequency, of each species in the sample.

Sample	ELISA results (mg/kg)	Gluten-free claim
F_001	<3	Yes
F_002	>20	No
F_003	11	Yes
MT_20	15	Yes
MT_200	>20	No
MT_500	>20	No

Table 3. Results of ELISA analyses on insect flours and wheat mock mixtures.

2.1.4 Discussion

The characterization of novel food composition is essential to ensure safety (Patel, Suleria, & Rauf, 2019) and our results suggest that the mtDNA COI region is useful for identifying the declared insect species in flours and in almost all the analysed insect-based products. However, this approach showed some limitations, since in two protein bars, the COI DNA barcoding analysis identified the parasite Ectomyelois ceratoniae instead of the declared Acheta domestica. The occurrence of E. ceratoniae is not unexpected as its larvae typically parasitize raw food material, especially when almond occurs as an ingredient (Mortazavi, Samih, Ghajarieh, & Jafari, 2015). We cannot exclude that the industrial processing steps fragmented or degraded the A. domestica DNA in some way, making it not targetable by the used primer pair. The use of species-specific DNA probes (Tramuta et al., 2018) or another DNA metabarcoding approach targeting at the insect ingredients using short genetic regions (< 400 bp), would have detected the occurrence of both the moth and the insect species declared on product label (Frigerio et al., 2019).

Concerning the plant components, the ITS2 region was efficient in providing information on species composition including those taxa probably belonging to the insect rearing substrate. There are several open questions about the composition of the feeds used for insects farming and its influences on the quality and safety of the final products (Magara et al., 2019; Van Broekhoven, Oonincx, Van Huis, & Van Loon, 2015). The selection of suitable feeding substrates is very important to enhance the nutritional characteristics of the insects (Magara et al., 2019; Oonincx & de Boer, 2012; Oonincx, Laurent, Veenenbos, & van Loon, 2019; Van Broekhoven et al., 2015) and can also affect the total farming yield (Ganda, Zannou-Boukari, Kenis, Chrysostome, & Mensah, 2019). To date, the EU regulation (EC no 1069/2009) clarifies that the insect rearing substrate has to contain only products of non-animal origin. The circular economy strategies support the adoption of biowaste and by-products of different agricultural and industrial origins claiming for potential benefits in terms of sustainability (EFSA, 2015). Therefore, the insect feeds are usually composed of vegetable ingredients derived from different agricultural supply chains which are likely impossible to be identified using morphological parameters. This was confirmed by our data which highlighted that the insect feed was almost completely characterized by horticultural plant sources, such as carrots, cabbages, and chickpea (Magara et al., 2019). Therefore, the proposed molecular approach offers a universal diagnostic system to identify the composition of the rearing substrate and to verify compliance with the current and future regulations.

In agreement with Schlu̇ter and co-workers (Schlu̇ter et al., 2017), insects must be reared under a defined substrate to avoid contamination and possible food borne outbreak for the consumer (e.g., due to pathogenic microorganisms, toxins, and antinutrients). Our data suggest that the ITS2 barcode region was also able to identify the plant-based substrate used for insect rearing. For example, in the analysed flour samples, we detected DNA from hemp (Cannabis sativa) and linen (Linum sp.), which are commonly used as litter material (data confirmed by interviewed companies). We underline that

Cannabis sativa is also used as an ingredient in some of the tested protein bars. Unfortunately, our method is not able to distinguish between the two sources.

In many processed products, we obtained most reads belonging to the expected highly abundant species, such as wheat and chickpeas, which likely hide the less represented OTUs. For example, the species that possibly constitute the rearing substrate are not very evident when compared to the analysed pure flours. Therefore, the main limitation of our analysis certainly resides in the sensitivity to detect the less abundant species due to primer bias. This issue has already been discussed by Bista and colleagues (Bista et al., 2018) and Krehenwinkel and colleagues (Krehenwinkel et al., 2017). Both studies agreed that a PCR-free whole genome sequencing could permit to avoid this effect. Furthermore, we demonstrated that different thresholds of OTUs size (in terms of number of reads) dramatically affect the final list of plant taxa recovered in the analysed samples. It is important to underline that in a context of food authentication and traceability, the adopted conservative criteria (i.e., n° sequences/OTU > 500) is essential to preserve information on the most representative species. Conversely, a deeper data exploration, using fewer conservative parameters (i.e., n° sequences/OTU (500), retrieves information on trace species but increases the risk of including false positives. Probably, a multi-marker approach, coupled with a dedicated reference molecular database encompassing the expected plants, as well as the most common contaminants (or species prohibited by law) could improve the plant characterization of insect-based products and could be useful to exclude the false positives. Other authors (e.g., Zhang, Chain, Abbott, & Cristescu, 2018) demonstrated that the combined use of at least two barcode markers improves species detection. Another possible limitation of DNA metabarcoding resides in the completeness and reliability of the reference dataset that could lead to incorrect reads assignment (Murali, Bhargava, & Wright, 2018). In our study, we chose the ITS2 barcode due to its higher capability of distinguishing congeneric plant species due to a higher mutation rate (Al-Juhani, 2019; Yu et

al., 2017). Moreover, in recent years, the ITS2 database is growing exponentially and this improves the suitability of this locus to taxonomically assigning DNA metabarcoding data.

2.1.4.1 DNA metabarcoding to identify and quantify allergens

In our study, we tested the ability of DNA metabarcoding to find plant contaminants in edible insect flours. The obtained results suggested that we are able to identify DNA of T. aestivum or G. max in the tested mock mixtures, starting from 20 ppm and 50 ppm of allergenic proteins, respectively. According to European regulations (EU No. 828/ 2014), these concentrations are the maximum limits for commercial gluten and soybean-protein free products.

Therefore, the HTS DNA metabarcoding analysis detects low amounts of contaminant products and allergens, with a limit of detection even lower than the ELISA analysis (i.e., 3 ppm). We underline that the three tested insect-based flours are declared as gluten-free products, but only F_001 and F_003 comply with the limit established by the European Commission.

Concerning the ability of DNA metabarcoding to quantify the 'putative plant contaminants', our study seems to indicate a weak relationship between the dry weight and the number of reads. However, there is a fervent debate about the effectiveness of providing quantitative inferences using HTS data. Some recent studies reported their findings in a quantitative manner where the relative read abundance is interpreted as the relative abundance of biomass (Lamb et al., 2019). Others use a frequency of occurrence approach, also referred to as weighted occurrence (Deagle et al., 2018), where the proportion of samples in which a given sequence was detected is used to infer a different sort of quantitative measure (De Barba et al., 2014).

As already noticed in our previous paper on processed food (Bruno et al., 2019), the amplicon DNA metabarcoding efficacy could be biased by the PCR amplification step using “universal” markers. The occurrence of bias during

PCR amplification may cause the inaccurate estimation of quantities, and this was at least partially demonstrated for metazoans and plants (Balech et al., 2018; Thudi, Li, Jackson, May, & Varshney, 2012). This bias generates a variable number of template–primer mismatches across species, resulting in a final amplified DNA mixture that does not always reflect the original proportion of each species, limiting the quantitative potential of DNA metabarcoding (Bista et al., 2018; Piñol, Senar, & Symondson, 2019). Nevertheless, our analysis suggests that DNA metabarcoding has a relative quantitative ability, as already demonstrated by Lamb and colleagues (Lamb et al., 2019), and this methodology can be intended as an early warning method for allergen detection in food products.

2.1.4.2 DNA metabarcoding of insect-based novel food: An overview

DNA metabarcoding is currently used for food authentication (Galvin-King, Haughey, & Elliott, 2018; Prosser & Hebert, 2017; Utzeri et al., 2018). Moreover, Haynes and colleagues (Haynes, Jimenez, Pardo, & Helyar, 2019) recommend this approach to enhance the quality control along the food supply chain. In order to present an overview of DNA metabarcoding applied to insect food authentication and safety, we developed a SWOT (Strengths, Weaknesses, Opportunities, Threats) analysis (Figure 5). The strengths are related to the high efficiency of the technique. DNA metabarcoding can detect traces of ingredients due to its high sensitivity and allows to obtain simultaneously different information about food safety and quality. The main weaknesses are the higher cost compared to the current available analytical approaches like ELISA tests or target PCR assays. However, we cannot exclude that in the very next future, the panel of targets required to assess authenticity and safety of insect-based products will be so wide to make the HTS DNA metabarcoding approach much more convenient than the use of multiple single target tests. DNA tests indeed are related to DNA quality and quantity, so highly processed insect products like protein bars can be challenging to analyse. Finally, the

results depend on the database which may be incomplete with a subsequent incorrect assignment. The opportunities are related to the novelty of insect food products. It is possible to create new economic opportunities in the analytical field with the goal to guarantee a safe product, stimulating the insect-based novel food market. HTS techniques, due to the completeness of the results, could also quickly support the compliance to forthcoming regulations. Finally, the first threat is that information on insect feed composition is not reported on the product label and this impedes comparing the detected composition with the declared one. Furthermore, there is currently no scientific reference to DNA metabarcoding applied to the quality and safety control of the new insect-based foods. In addition to that, DNA analyses are currently not mandatory, so this can interfere with the spread of DNA metabarcoding as a routine analysis.

Figure 5. Overview of the main strengths, weaknesses, opportunities and threats (SWOT analysis) related to the use of DNA metabarcoding as a tool for insect-based novel food products.

2.1.5 Conclusions

Novel foods demand is increasing, and their consumption is expected to grow in the next few years. This condition encourages an increase in the number of ingredient species and enhances the risk of misidentification, contamination, and counterfeiting. This is well documented in the case of fish, where more new

species are available on the market, and in many countries the DNA barcoding approach is considered an essential tool to avoid fish frauds (Fox, Mitchell, Dean, Elliott, & Campbell, 2018).

Considering that many food products contain a mixture of species, we strongly encourage the adoption of DNA metabarcoding to better elucidate not only the food composition but also to assess trace elements belonging to different steps of the food supply chains. Unfortunately, this approach does not accurately estimate the biomass of the ingredient taxa, and although in our case the HTS DNA meta-barcoding approach highlighted the occurrence of allergenic species (even at limit concentration values), we cannot use this method as an alternative to the standard ELISA test. This is also because the presence of DNA of a species is not necessarily correlated with the occurrence of allergens. However, we proposed the DNA metabarcoding analysis as a preliminary screening, especially for novel foods, because this method offers the ability to identify, at the semiquantitative level, several potential allergenic plants with a single analysis. Therefore, DNA-based analysis can be used to select which ELISA tests, and in general which of the more reliable toxicological assays for the detection of plant contaminants to use. In this sense, the DNA metabarcoding approach offers an opportunity to enhance the food safety of novel food products, such as those based on insect ingredients.

2.1.6 Data availability statement

The dataset generated for this study was submitted to the EBI metagenomics portal (https://www.ebi.ac.uk/metagenomics/; Study ID: PRJEB34990).

Supplementary Materials are available through the main paper (https://doi.org/10.1016/j.foodres.2019.108751).

2.2 Tasting the differences: microbiota analysis of different insect-based novel food

2.2.1 Introduction

Entomophagy is an emerging and fashionable diet issue in western countries. Insects are an important source of energy for human diets, because of their richness in essential nutrients (Rumpold & Schlüter, 2013). They have a protein content average value ranging from 30% to 65% of the total dry matter, and they are also rich in micronutrients such as iron, zinc and calcium (Dobermann, Swift, & Field, 2017). Insects like Alphitobius diaperinus and Tenebrio molitor L. can be also used as a source for the production of fortified foods (Roncolini et al., 2019, Roncolini et al., 2020) facing the problem of the food demand of the growing world population (Baiano, 2020). Moreover, preliminary studies of Oonincx and de Boer (2012) stated that, compared to other livestock animals, insect farming has a lower environmental footprint.

Safety, traceability and quality of edible insects are of great interest both for the producers and the consumers, heavily affecting the acceptance of edible insects in the human diet (House, 2016). New tools for safety controls on these food items could also benefit institutions like food agencies, customs and health departments in the evaluation of new product development based on processed insects. In the European Union, the regulation (Regulation EU /2015, 2283) has classified edible insects as novel foods, which follow specific rules and require specific authorizations before allowing them to be distributed (Klunder et al., 2012, Schlüter et al., 2017, Van Huis, 2013). Besides, food safety authorities and the scientific community are discussing whether edible insects can be a reliable solution or a problem to the food security (Belluco et al., 2015, Di Mattia et al., 2019).

The potential safety risks of edible insects are chemical hazards including pesticides, heavy metals, allergens, toxins (mycotoxin and bacterial toxins) (Garofalo et al., 2019). There is a risk that harmful insect microbes are transmitted through the consumption of insect products (Van der Spiegel, Noordam, & van der Fels-Klerx, 2013). Most of the insect microbiota are associated with gut (e.g., the intrinsic insect symbionts in the intestinal tract and the proximity of other anatomical compartments) or related to extrinsic sources, such as environment and rearing conditions (substrates and feed), handling, processing and preservation (ANSES, 2014). Especially, as stressed recently by the European Food Safety Authority (EFSA, 2015), spore-forming bacteria in processed edible insects (including freeze-dried, boiled and dried varieties) can be considered a dangerous source of biological contamination as well.

Garofalo et al. (2017) explored the microbiota of marketed processed edible insects using culture-based methods and pyrosequencing. They described, among others, the microbiota of whole dried small crickets (Acheta domesticus) and whole dried mealworm larvae (Tenebrio molitor), revealing a great bacterial diversity and variability among individual insect species: some of the identified microbes may act as opportunistic pathogens in humans, such as Listeria spp., Staphylococcus spp., Clostridium spp. and Bacillus spp., while others represent food spoilage bacteria, as well as Spiroplasma spp. in mealworm larvae. The insect diet and social behaviour have a great impact on the composition of the gut microbial community (Tinker & Ottesen, 2018), therefore different insect farm conditions result in different microbiological ecosystems. Although some authors such as Stoops and co-workers (Stoops et al., 2017) suggested that the microbial taxonomic composition varies mainly with insect species, the additional factors such as the growing substrates or contact with soil may play an important role in the composition of the insect gut microbiota (Risk profile, 2015, Klunder et al., 2012, Li et al., 2016). Considering the insect production system, industrial practices, such as

post-harvest starvation and rinsing, can affect the microbial quality of the final insect products too (Wynants et al., 2018). Since all food products, including those insect-based, undergo processing, the risk for human safety should be measured throughout the various stages, from raw materials (i.e. insect flour) to final food products (Osimani et al., 2018). High-Throughput DNA Sequencing (HTS) offers a standardized and sensitive method to evaluate the microbial community changes by analysing a wide range of food products (De Filippis, Parente, & Ercolini, 2018). The search for a microbial signature represents an opportunity to verify both food safety and food traceability strategy, indeed the microbial variation gives insight about rearing and processing products. The microbial variability allows obtaining more information besides the identification of the insect species, like the hygienic and sanitary conditions concerning the rearing systems. Moreover, the insect microbiota can be used to identify the geographical origin of a food product and used as a tracing signature, as previously demonstrated by recent studies (Bokulich et al., 2016, Mezzasalma et al., 2017). The microbial signature can then eventually be applied to management and control systems (Galimberti et al., 2019).

In this study, we evaluated the microbiota composition of insect-based commercial food products, applying HTS with complementary bioinformatics analysis. The aim of this preliminary study was to analyse the microbiota variability of different categories of insect-based products made of A. domesticus (house cricket), T. molitor (mealworm beetle), and A. diaperinus (lesser mealworm or litter beetle) (including commercial raw materials like flours and processed food items), purchased via e-commerce from different companies. We sought to define if HTS can be a useful tool for insect-based novel food quality assessment.

2.2.2 Materials and methods

2.2.2.1 Insect food products

A total of 12 commercial insect-based products were purchased via e-commerce from five different companies. Referring to the label information, these products contained only one insect species each: Acheta domesticus (Order: Orthoptera), Alphitobius diaperinus (Order: Coleoptera), and Tenebrio molitor (Order: Coleoptera) (S1 Table).

Four out of 12 samples were pure insect flours, belonging to the species A. diaperinus (n = 1) and T. molitor (n = 3), and they have been categorized as insect raw material (dried insect product without other ingredients). In the case of T. molitor, flour samples derived from three different batches of the same product. Eight out of 12 samples represented processed food products: pasta (n = 3), crackers (n = 2) and protein bars (n = 3). A detailed description of the samples can be found in Table 1.

Sample type	Code	Label declared insect	Label declared ingredients	Company origin	Company name
Flour	R_001	*T. molitor* (100%)*	–	Netherlands	Company 1
	R_002	*T. molitor* (100%)*	–	Netherlands	Company 1
	R_003	*T. molitor* (100%)*	–	Netherlands	Company 1
	R_004	*A. diaperinus* (100%)	–	Netherlands	Company 1
Pasta	FP_005	*A. diaperinus* (14%)	*Triticum durum, Ocimum basilicum* (1.5%); organic powdered egg whites	France	Company 2
	FP_006	*A. diaperinus* (14%)	*Triticum durum*; organic powdered egg whites	France	Company 2
	FP_007	*T.o molitor* (10%)	*Oryza sativa* (43%); *Cicer arietinum* (43%); organic powdered egg whites (4%)	France	Company 3
Cracker	FP_008	*A. domesticus* (14%)	*Triticum aestivum*; *Sesamum indicum* (6%); *Olea europaea*	Great Britain	Company 4
	FP_009	*T. molitor* (10%)	*Triticum aestivum*; *Cocos nucifera*; *Avena sativa*; *Sesamum indicum* (12%); *Porphyra* sp. (1.2%	France	Company 3
Protein bar	FP_010	*A. domesticus* (5.2%)	*Phoenix dactylifera*; *Prunus dulcis*; *Musa* spp. (11%); *Theobroma cacao* (9%); *Vaccinium macrocarpon* (8%); *Anacardium occidentale*; *Cannabis sativa*	France	Company 2
	FP_011	*A. domesticus*(5.5%)	*Phoenix dactylifera*; *Prunus dulcis*; *Prunus armeniaca* (22%); *Pisum sativum*; *Helianthus annuus; Lycium barbarum* (4.5%); *Salvia hispanica* (3.5%)	France	Company 2
	FP_012	*A. domesticus* (20%)	*Arachis hypogaea* (34%); *Cannabis sativa*; *Theobroma cacao*; *Agave* sp; *Beta vulgaris*; *Cinnamomum* sp. (1%)	Great Britain	Company 5

Table 1. List of analysed insect-based products. For each sample, the information found on the label about the category, the species of insects, the percentage of insects present in the food product, the other ingredients declared on the label and the company origin are reported. R (Raw food products); FP (Processed Food product). *Different batches of the same product of T. molitor flour.

2.2.2.2 DNA extraction

High-quality genomic DNA was obtained starting from 250 mg of each sample of Table 1 using DNeasy PowerSoil Kit (QIAGEN, Hilden, Germany), according to manufacturer's instructions. Three replicates of DNA extraction were generated for each sample plus a negative control. Purified DNA was checked for concentration and purity by using a Qubit 2.0 Fluorometer and Qubit dsDNA HS Assay Kit (Invitrogen, Carlsbad, California, United States).

2.2.2.3 DNA barcoding characterization of insect samples

The 658 bp mtDNA COI region was used to validate the animal species declared on the label in the sampled insect-based products. This region was amplified and sequenced for all 12 samples according to the primer pairs presented by Folmer, Black, Hoeh, Lutz, and Vrijenhoek (1994) and the protocol described in Bellati et al. (2014). Each sequence was defined as the nearest match with the BLAST algorithm using the following cut-off values: maximum identity > 99% and query coverage of 100%.

2.2.2.4 HTS library preparation and sequencing

To characterize the bacterial composition of the investigated insect-based products, 16S rRNA genes (V3 and V4 hypervariable regions) of the obtained gDNA extracts were sequenced using a High-Throughput DNA Sequencing approach. Amplicons were generated following the protocol described by Caporaso et al. (2012) with Illumina adapters (S2 Table), with minor modifications as described in Frigerio et al. (2020): we used PuReTaq

Ready-To-Go PCR beads (GE Healthcare Life Sciences, Italy) according to manufacturer's instructions in a 25 µL reaction, containing 1 µL 10 mM of each primer and up to 50 ng of gDNA. The amplification profile consisted of an initial denaturation step for 5 min at 95 °C, followed by 25 cycles of denaturation (30 s at 95 °C), annealing (30 s at 55 °C), and elongation (30 s at 72 °C), and finally elongation at 72 °C for 5 min. Amplicon DNA was checked for concentration by using a Qubit 2.0 Fluorometer and Qubit dsDNA HS Assay Kit (Invitrogen, Carlsbad, California, United States) and amplicon length was measured by comparison against QX DNA Size Marker using the Qiaxcel Automatic electrophoresis system (QIAGEN, Hilden, Germany). Samples were sequenced by the Center for Translational Genomics and BioInformatics (Milan, Italy). The sequencing was performed on the MiSeq sequencing platform (Illumina, San Diego, CA, USA) with a paired-end approach (MiSeq Reagent Kit v3, 2 × 300 bp).

2.2.2.5 Bioinformatic analysis

Illumina reads were analysed with QIIME2, Quantitative Insights Into Microbial Ecology 2 program (ver. 2019.4; https://qiime2.org/) (Bolyen et al., 2019). Sequences were demultiplexed with native plugin and DADA2 (Divisive Amplicon Denoising Algorithm 2) (Callahan et al., 2016) was applied to obtain ASVs sequences (or features) (Callahan, McMurdie, & Holmes, 2017), trimming primers and performing a quality filter with an expected error of 2.0. Chimeric sequences were removed using the consensus method. Features with at least 10 representatives associated and detected in at least two samples were kept. The taxonomic assignment of representative sequences was carried out using the feature-classifier (https://github.com/qiime2/q2-feature-classifier) plugin implemented in QIIME2, using classify-consensus-vsearch method against the SILVA SSU non-redundant database (132 release), adopting a consensus confidence threshold of 0.8. Taxa bar plots were generated with the QIIME2 dedicated plugin taxa (https://github.com/qiime2/q2-taxa). As ASVs assigned to Cyanobacteria phylum (Class: Chloroplast) were considered potential plant

contaminants, they were removed from the downstream analysis. Reads of mitochondrial or eukaryotic origin were also excluded. Overlap among technical replicates was calculated considering taxa at the family level weighted for abundances (Wen et al., 2017). Alpha diversity was carried out considering the presence/absence of ASVs and Shannon index. Statistical differences among samples belonging to the same insect species were calculated using alpha-group-significance plugin by QIIME2, performing also a pairwise contrast (Kruskal & Wallis, 1952). Beta diversity, instead, was carried out considering qualitative (Jaccard and unweighted UniFrac) and quantitative (Bray-Curtis and weighted UniFrac) distance metrics (Lozupone, Lladser, Knights, Stombaugh, & Knight, 2011), using QIIME2 core-metrics plugin (https://github.com/ qiime2/q2-diversity). Statistical differences were calculated by permutation based ANOVA (PerMANOVA) functions of beta-group-significance plugin (Anderson, 2001), with 999 permutations, considering insect species and sample type categories. A PerMANOVA Pairwise contrast was performed with beta-group-significance plugin. Principal coordinates plots (PCoA) method was used to explore the structure of microbial communities. The phylogenetic tree necessary to calculate UniFrac distances was built on the alignment of ASVs representative sequences using align-to-tree-mafft-fasttree method by phylogeny plugin (https://github.com/qiime2/q2-phylogeny). Heatmap visualization was used to explore the abundance of bacteria families among samples and was generated by QIIME2. Core microbiota among insect samples was calculated considering the ceiling of the mean of species frequencies among samples and keeping a core threshold of 0.7 (minimum fraction of samples that a species must be observed in), performed with core-features plugin (https://github.com/qiime2/q2-feature-table). A Venn diagram was created starting from core microbiota results setting the threshold = 1, by calculating the number of shared and unique taxa per insect collapsed at the genus level. ANCOM analysis (Analysis of composition of microbiomes; Mandal et al., 2015) was performed to test differential abundances among genera distribution in the

dataset, comparing samples with different insect composition. To avoid false discovery rates, only features shared in at least 25% of samples were considered.

2.2.3 Results

2.2.3.1 Sequencing output

All the replicates of the 12 collected samples showed good DNA quality (i.e., A260/A230 and A260/A280 absorbance ratios within the range 1.6–2.2) and good yield (20–40 ng/µl). The DNA barcoding (mt COI) sequencing results indicated that all the tested samples were composed of insects. Moreover, the BLAST analysis against reference insect DNA barcoding sequences confirmed that all samples corresponded with the declared insect species (i.e., maximum identity > 99% with the declared species).

HTS analysis produced about 8,571,836 raw reads from the analysed samples, with an average of 119,053.,28 reads per sample (DS = 62,045.83). After quality filtering, merging reads, chimaera and contaminants removal, we obtained a total of 590 ASVs (Amplicon Sequence Variants). Negative controls (deriving from DNA extraction and amplification step) for library sequencing were not included in the analysis since they encompassed a very low number of DNA reads.

2.2.3.2 Microbial diversity analysis

From overlap calculations for technical replicates, family overlap resulted in a mean of 96%, with a standard deviation of 0.06.

Both considering ASVs and Shannon metric, differences among samples derived from different insects were observed ($H = 22.13$, p-value < 0.01 and $H = 29.93$, p-value < 0.01, for ASVs and Shannon respectively; pairwise comparisons are visible in Table S3).

Samples belonging to raw material (flour) and food products (crackers, pasta and protein bars) showed a significant difference, considering both qualitative (Jaccard and Unweighted UniFrac) and quantitative metrics (Bray-Curtis and Weighted UniFrac) (p-value < 0.01). Overall, we observed a significant difference among samples belonging to different insects (Jaccard metric: F-statistic = 10.59, p-value = 0.001; Unweighted UniFrac metric: F-statistic = 10.57, p-value = 0.001; Bray-Curtis metric: F-statistic = 16.79, p = 0.001; Weighted UniFrac metric: F-statistic = 25.38; p-value = 0.001). Results of pairwise comparisons are visible in Table S4.

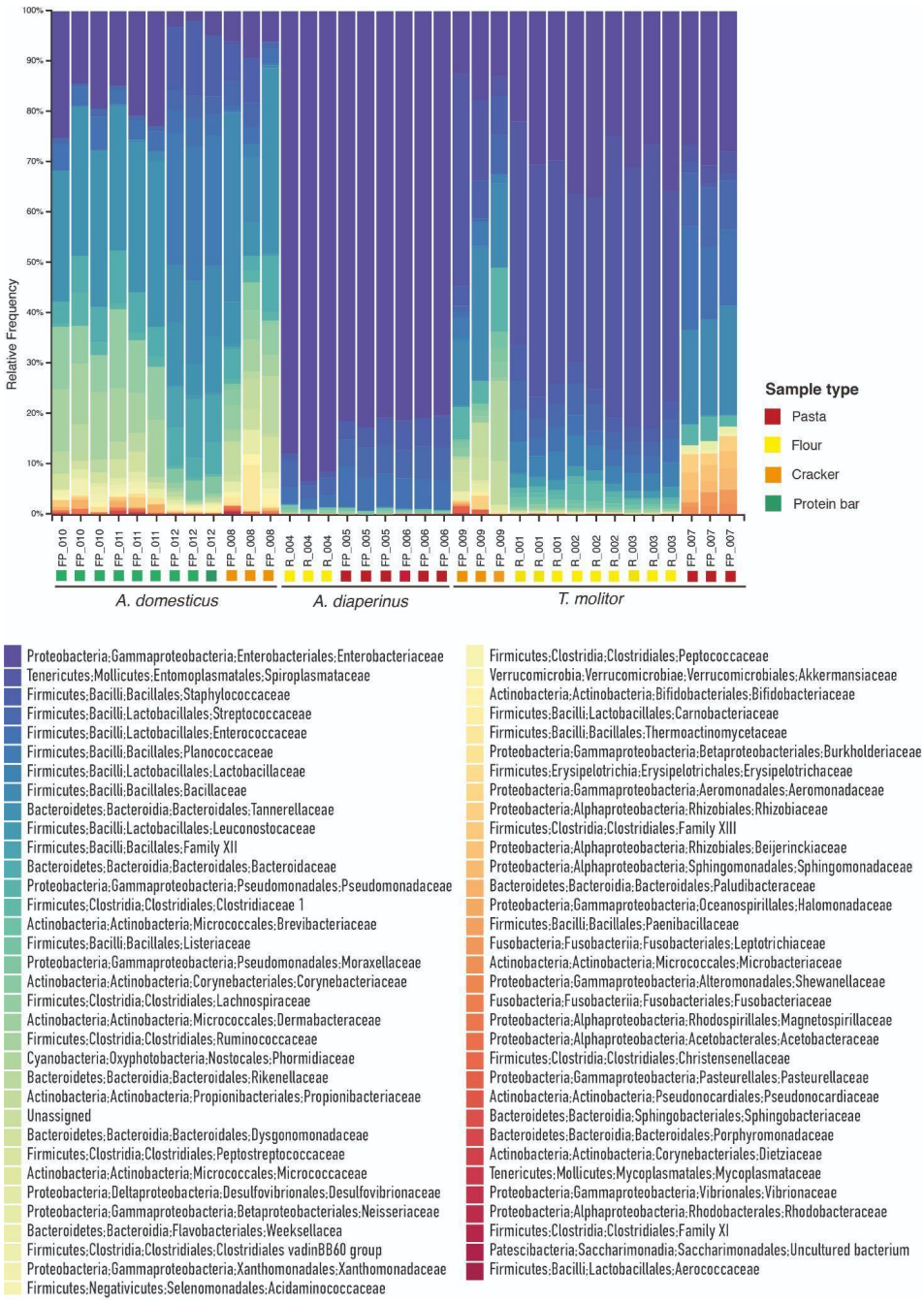

Figure 1. Relative abundance of bacteria families recovered in the insect-based products through 16S metabarcoding sequencing. Bacteria families are reported in gradient colors indicating relative abundances. For each sample, the sample type is reported (pasta: red square; flour: yellow square;

cracker: orange square; protein bar: green square). For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.

2.2.3.3 Taxonomic composition analysis

A total of 9 bacterial phyla, 14 classes, 34 orders, and 66 families were identified (Figure 1, S5 Table).

Taxonomic analysis revealed that most of the sequences in all the samples were associated with the phyla Proteobacteria (47%) and Tenericutes (26%), followed by Firmicutes (23%). 0.13% resulted in Unassigned taxa. Looking inside the taxonomic rank of class, the most abundant were Gammaproteobacteria, with 47% of sequences, followed by Mollicutes (26%) and Bacilli (22%). Enterobacteriales was the most abundant order, encompassing 45% of the sequences, distributed across all the samples, followed by Entomoplasmatales (26%), Lactobacillales (12%), Bacillales (10%), and Bacteroidales (2.6%). On the whole, the remaining 29 orders covered 4.4% of sequences. The Enterobacteriaceae family accounted for 45% of sequences, whereas Spiroplasmataceae represented 26% of sequences.

Considering taxa distribution per insect (Figure 2), we can notice differences in microbial composition, spanning from the phylum level to a deeper taxonomic resolution. Considering taxonomy per insect species, at the taxonomic level of order, we found that A. domesticusbased samples were dominated by Bacillales (54%), followed by Bacteroidales (21.2%), and Lactobacillales (8.9%), representing 84.1% of 28 orders. However, food products made with A. diaperinus had most of the sequences assigned to Enterobacteriales (89.6%), with the remaining 7% and 2.1% assigned to Lactobacillales and Bacillales, respectively, and 1.3% of sequences distributed in 11 orders. T. molitor-based food products showed 45.4% of sequences corresponding to Entomoplasmatales order, 29.5% to Enterobacteriales, 14.5% to Lactobacillales, and the remaining 10.6% to 21 different orders.

Focusing on specific features, we observed that the most abundant feature was assigned to an uncultured Spiroplasma (25%), reported exclusively in T. molitor samples. The sixth most abundant feature (3%), assigned to the genus Kurthia (Planococcaceae; Bacillales; Bacilli; Firmicutes) was detected only in A. domesticus protein bars produced by the British company 5, but not in samples belonging to the British company 4. Moreover, all and only the food products deriving from British company 5 are characterized by the presence of a specific feature assigned to Exiguobacterium (Family XII; Bacillales; Bacilli; Firmicutes). Considering features shared between protein bars belonging to British company 5 and French company 2, some of the most abundants were assigned to Tannerellaceae (12,3%) (Bacteroidales; Bacteroidia; Bacteroidetes), followed by Bacteroidaceae (6%) (Enterobacteriales; Gammaproteobacteria; Proteobacteria), Enterobacteriaceae family (5%) (Enterobacteriales; Gammaproteobacteria; Proteobacteria) and Lachnospiraceae (2%) (Clostridiales; Clostridia; Firmicutes).

A feature assigned to an uncultured Parabacteroides (Tannerellaceae; Bacteroidales; Bacteroidia; Bacteroidetes) is unique for A. domesticus samples, whereas features assigned to Enterobacter (Enterobacteriaceae; Enterobacteriales; Gammaproteobacteria; Proteobacteria), a different microorganisms belonging to Enterobacteriaceae, and Enterococcus (Enterococcaceae; Lactobacillales; Bacilli; Firmicutes) were highly prevalent in A. diaperinus food products.

To better visualize the microbial variation among different food products, and which family mostly contribute distinguishing food products, a heatmap based on relative abundances was generated (Figure 3). Analyzing the sample cluster dendrogram, two main clusters separate samples based on insect order, composed by A. domesticus (Orthoptera) food products and T. molitor plus A. diaperinus (both Coleoptera) food products. Subclusters differentiated raw food products (flour) from processed food products (pasta, crackers and protein bars): flour made by the two insects of the Coleoptera order (i.e., T. molitor and A. diaperinus) formed a distinct cluster that separated pasta and crackers

samples based on the same insects. Moreover, the same food products constituted by different insects can be distinguished by family abundances in the heatmap: A. diaperinus pasta clustered separately from T. molitor pasta. Conversely, protein bars composed by the same insect (A. domesticus), but produced by different companies, are scattered in two different clusters, as also shown by microbial diversity analysis represented in the PCoA plot (Table S4).

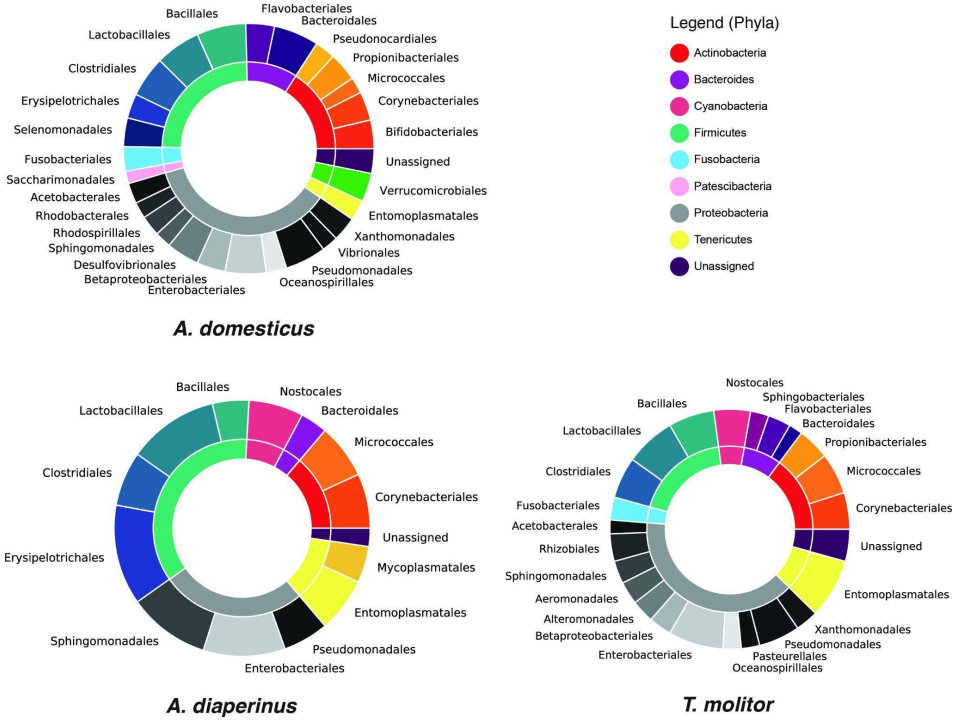

Figure 2. Donut charts of A. domesticus, A. diaperinus, and T. molitor microbial composition. Phyla in the inner circle and Orders in the outer circle are reported. Abundances are expressed as log frequency, in order to better show underrepresented taxa.

2.2.3.4 Preliminary analysis on microbial signature

The preliminary analysis on core microbiota, defined as a group of shared microbial taxa or genes (Hamady & Knight, 2009; Turnbaugh et al., 2007), revealed the taxa shared by at least 70% and the 100% of amples of the category representing the insect used in the food products. Venn diagram, calculated from core microbiota results of the most conserved taxa (100% of samples per insect), highlighted the presence of unique and shared taxa considering insect species used in the food products analysed (Figure 4).

In the case of T. molitor-based food products, we observed a core microbiota constituted by 21 taxa shared among > 70% of the samples and 10 taxa shared by all the samples. The 10 most conserved taxa (100% of samples) belonged to uncultured Spiroplasma sp., a taxon from Enterobacteriaceae family, Enterococcus, Staphylococcus, Enterobacter, uncultured Lactococcus, Pseudomonas, Bacillus, Serratia and Pantotea (S6 Table).

On the other hand, A. diaperinus-based food products showed 14 shared taxa, both subsampling the 70% of samples or considering all the samples, indicating a highly conserved core microbiota. In contrast to T. molitor-based products, we reported the presence not only of Enterococcus, Staphylococcus, Enterobacter, Lactococcus, but also of Enterococcus faecalis, Listeria, Brevibacterium, Corynebacterium, Brachybacterium, Acinetobacter, and Bacillus pumilus. We reported as well the absence of Spiroplasma, Pseudomonas, Serratia and Pantotea.

Considering A. domesticus-based food products, all the samples share 29 taxa, and 44 taxa are shared by 70% of samples. Among these, all the samples reported the presence of bacteria belonging to the family Lachnospiraceae and the genus Parabateroides (Family: Tannerellaceae).

Venn diagram analysis showed that, if four genera are shared among all the samples (a genus belonging to Enterobacteriaceae family, Lactococcus, Enterobacter, Enterococcus), 28 genera were unique considering the insect species.

Figure 3. Heatmap diagram showing the relative abundance of families for each sample. Color shading in the heatmap indicates the abundance, expressed as log10 frequency, of each family in the sample. Sample type categories are flour (yellow), pasta (red), cracker (orange) and protein bar (green). Companies are represented in fuchsia (Company 1), blue (Company 2), aquamarine (Company 3), pink (Company 4) and light blue (Company 5). For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.

In particular, twenty genera were exclusively detected in all the samples of A. domesticus-based food products, and, among them, the three most abundant were Parabacteroides, Bacteroides, and a genus belonging to Lachnospiraceae family (see S6 Table for the complete list), thus confirming the explorative analyses described in the previous section. Brevibacterium, Acinetobacter, Brachybacterium, Listeria, and Corynebacterium were the genera unique for A. diaperinus-based food products, whereas T. molitor-based food products showed as unique genera Spiroplasma, Pantoea, and Serratia.
In order to explore variations in genera abundances among insect samples, ANCOM analysis was performed. Considering features shared in at least 25% of the dataset, the analysis comprehended a total of 31 genus. ANCOM results showed 16 differential abundant genera among samples (S7 Table).

In particular, ten were detected as insect-specific genera, according to core microbiome analysis. Further, for T. molitor samples, Spiroplasma, Lactobacillus, Pediococcus and a genus belonging to the Clostridiaceae family were identified, with a W-statistic of 30, 30, 24 and 23 respectively. Genera Parabacteroides, two uncultured bacteria belonging to the Ruminococcaceae family, Bacteroides, a genus belonging to Lachnospiraceae family and Citrobacter were peculiar of A. domesticus samples, with a W-statistic of 30, 30, 28, 29, 29, and 27, respectively. Enterobacter (W-statistic = 30), Corynebacterium (W-statistic = 25), and Listeria (W-statistic = 25) were differentially distributed among the dataset, characterizing only A. diaperinus and T. molitor samples. Regarding genera that were shared among all insect species, Lactococcus (W-statistic = 25), Staphylococcus (W-statistic = 24), and a genus belonging to the Enterobacteriaceae family (W-statistic = 23) were differentially distributed. Further, median abundances of Lactococcus were 35.5 in A. domesticus samples, 547.0 in A. diaperinus samples and 1,917.0 in T. molitor samples. Staphylococcus, instead, showed a median abundance of 501.0 in A. diaperinus samples and 1,475.0 in T. molitor samples, while Enterobacteriaceae medians were 144.0, 2,108 e 11,659, for A. domesticus, A.

diaperinus and T. molitor samples, respectively (ANCOM results and distribution of genera among insect samples are visible in details in S7 Table).

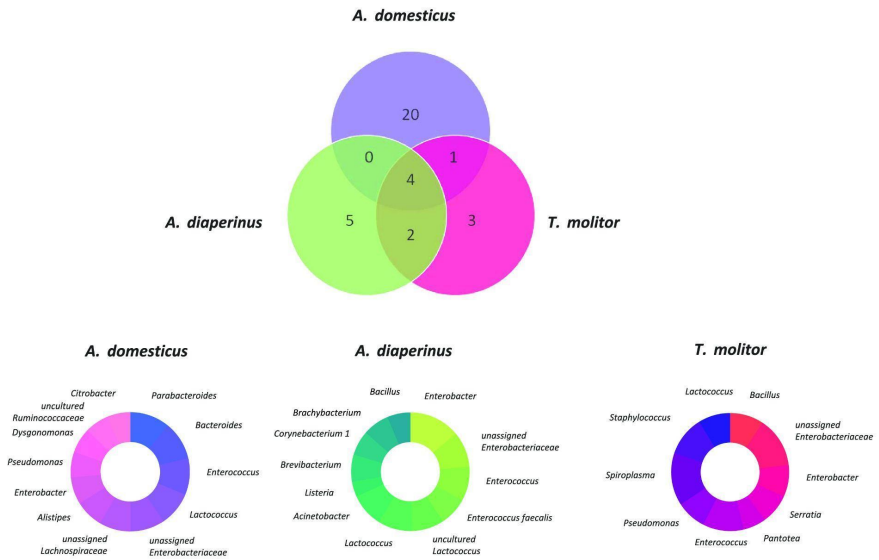

Figure 4. Venn diagram and donut charts of A. domesticus, A. diaperinus, and T. molitor core microbial composition. The Venn diagram in the upper part of the figure shows shared and unique taxa per insect. Taxa identified through core microbiota analysis are reported in the lower part of the figure. We considered the taxa found in 100% of the samples. In the case of A. domesticus and A. diaperinus the first twelve hits are reported, according to the frequency values listed in S4 Table.

2.2.4 Discussion

In this study, we characterized through the application of HTS techniques the microbial composition of insect-based food products made of A. domesticus, T. molitor, and A. diaperinus, purchased via e-commerce. We selected both raw

and processed food products, considering the availability on the market, from different selling companies.

Our preliminary data revealed that a small number of prevalent bacteria formed a “core microbiota” for each insect, which can potentially be used as biomarkers to identify insect ingredient origin in food products.

A recent study (Cambon, Ogier, Lanois, Ferdy, & Gaudriault, 2018) showed that a resident microbiota in T. molitor gut exists, thus supporting our hypothesis tested with core microbiota analysis. In particular, this study identified a resident T. molitor microbiota consisting of Pseudomonas, Serratia and genera belonging to the Enterobacteriaceae family. Noteworthy, this evidence is in accordance with the data we obtained in our study, as a further confirmation of our results.

If there was a significant insect component, the core microbiota would reflect the physiology of the organisms, the diet and rearing conditions. By contrast, if the level of food processing affected the microbiota, the organism could be difficult to identify searching for a microbial signature. Nevertheless, we identified shared features constituting the core microbiota of specific insects. In addition to that, despite the processing level, we found exclusive taxa in all the samples of specific insects. Noteworthy, our results showed that in A. domesticus processed food (i.e. protein bars and crackers) microbiota is composed by a robust core of microorganisms that is conserved and is similar in composition to what was reported in other studies on raw food (i.e. fresh crickets): Vandeweyer and colleagues (Vandeweyer, Crauwels, Lievens, & Van Campenhout, 2017) showed that A. domesticus is abundantly colonised by (Para)bacteroides species (Johnson, Moore, & Moore, 1986), confirming the first two hits we obtained through core microbiota analysis.

Interestingly, in this study A. domesticus core microbiota harbored bacteria belonging to the Lachnospiraceae family too. This evidence may prove beneficial when edible insects will be introduced in the western diet and it is worth further studies: Lachnospiraceae are found, among others, in our digestive tract and are involved in fibre digestion. Menni and colleagues indeed

discovered the association between Lachnospiraceae and lower long term weight (Menni et al., 2017). Furthermore, the exposure to antibiotics (such as β-lactam antibiotics and fluoroquinolones) eliminates Lachnospiraceae from gut microbiota. This lead to the gut becoming a prime target for opportunistic infections such as the one caused by Clostridium difficile, but restoring Lachnospiraceae into the intestines of infected patients has been shown to help cure C. difficile infections (Lagier, Million, Hugon, Armougom, & Raoult, 2012; Seekatz et al., 2018; Segata et al., 2012; Song et al., 2013). It is conceivable that in the processed food we found only DNA and not viable cells and more investigations are needed, also focusing on prebiotic effects. In a recent study, the impact of an insect-based diet (cricket) on the human gut microbiota revealed increased levels of Bifidobacterium animalis. This could be due to cricket chitin which may function as a prebiotic (Stull et al., 2018). T. molitor flour in in vitro fecal models promoted the growth of Bacteroidaceae and Prevotellaceae, but not of Clostridium histolyticum group or Desulfovibrionales and Desulfuromonales (De Carvalho et al., 2019).

On the other hand, exclusively all the samples based on T. molitor source are dominated by Spiroplasmataceae family (Phylum: Tenericutes; Class: Mollicutes), in particular bacteria belonging to Spiroplasma genus. Spiroplasma are found in the gut or hemolymph of insects where they can act as endosymbionts, impacting on host reproduction or host defence system. These findings are consistent with studies on fresh mealworm larvae (Vandeweyer et al., 2017) deriving from different companies.

A. diaperinus samples are dominated by Enterobacter, both flour and pasta, produced by different companies. These findings are in agreement with previous studies on fresh larvae (Wynants et al., 2017) and minced meat-like products (Stoops et al., 2016). A. diaperinus-based pasta clustered separately from flour samples made of the same insect, but in the same main cluster including food products belonging to Coleoptera. A similar behaviour can be seen in the case of T. molitor pasta and flour samples.

Concerning food safety, it is worth mentioning the presence, considering the 20 most abundant bacteria classified at the genus level, of sequences assigned to Bacillus in most of the samples (80%). The capacity to form endospore, resistant to heat and desiccation, deserve attention even if there is no confirmation of viability assay. There are currently no regulations for microbiological criteria of edible insects or their products in Europe, but a 5 Log10 (CFU/g) was defined as a safety threshold. Fasolato and colleagues found the presence of vital Bacillus in edible processed insects. Even if median values were lower than 4 Log10 CFU/g, some products showed higher lever (maximum 6.6 log10 CFU/g) (Fasolato et al., 2018).

Considering differential abundances of shared genera among samples, such as genera belonging to Enterobacteriaceae family or Lactococcus, differences may be caused by matrix peculiarities. In particular, food treatments as freezing or boiling processes can cause cell lysis and DNA degradation, thus affecting High-Throughput DNA Sequencing (HTS) output (De Filippis et al., 2018; Osimani et al., 2018). Further, high abundances of features were found in T. molitor and A. diaperinus samples, where we have a predominance of flour samples, a matrix obtained from only grinding treatment.

With the increasing availability of insect-based processed food products in the market, including a higher number of samples in the analyses will help in disentangling the microbial dynamics behind food processing, and allowing the food products traceability at a finer scale.

Overall, our results showed that insect-based food products cluster based on their microbial signature. Even in the case of processed food in which there is more than one constituent (i.e., plant ingredients, see Table 1) that could interfere with its microbial contribution in the clustering process, we identified a shared pattern highlighted by core microbiota analysis and unique taxa that can be used as biomarkers. We also showed that differences exist in comparing raw vs processed food considering both qualitative and quantitative metrics. Recent studies (Bruno et al., 2019) reported the possibility to track the composition of plant processed food despite critical issues mostly deriving

from the starting composition (i.e., variable complexity in taxa composition) of the sample itself and the different processing level (i.e., high or low DNA degradation). Other studies (Garofalo et al., 2017), investigating the microbial composition of commercial food products based on insects, never explored if any variability can be correlated with highly processed food such as pasta, crackers or protein bars. Our data clearly showed that processed food can be analysed searching for a microbial signature and that raw food products (i.e., flours) had a significant different microbiota compared to the processed ones (i.e., pasta, crackers and protein bars), even if maintaining unchanged a core of bacteria.

Highly processed food products represent one of the challenges of food traceability because of DNA degradation during food processing and, as a consequence, the limits in applying the common DNA barcoding techniques. Thus, DNA metabarcoding, based on HTS techniques combined with powerful tools for data analysis, can provide new perspectives for unveiling the composition of processed food, to retrace food origin and food quality control (Bruno et al., 2019; De Filippis et al., 2018; Parente, De Filippis, Ercolini, Ricciardi, & Zotta, 2019).

The identification of a microbial signature for traceability purposes was suggested also by forensic scientists as a natural consequence of the application of HTS technologies in a wider perspective (Bishop, 2019): with the globalisation of trade, food traceability is a hot topic and identifying a microbial signature in these products can provide a deeper insight into the “food ecosystem” (Bokulich et al., 2016; Galimberti et al., 2015, 2019; Parente et al., 2019).

2.2.5 Conclusions and future perspectives

The application of high-throughput molecular techniques coupled with bioinformatic analyses allowed us to detect and identify the diversity of

microbial communities in some raw and processed novel food products available on e-commerce. This study shows the value of the application of HTS analysis for unveiling the composition of microbiota in processed food containing insect ingredients. We were able to identify with our preliminary analysis a microbial signature, depending on the insect, suggesting that a resident microbiota is conserved despite the different food processing levels, rearing conditions, and selling companies. We are now facing a striking imbalance between available technologies and knowledge gaps on "food ecosystem": especially in the case of insect flour and insect-based products, as a future perspective we should consider the whole food production chain, taking into consideration that the microbial communities inhabiting surfaces, interacting with foods and being part of food themselves are influenced all along the supply chain, from rearing, in the case of insects, to the final processed product. HTS approach is a valuable tool to protect food safety as routine monitoring analysis, from the identification of insect microbiota along the food production processing chain and characterization of the raw ingredients to the final processed food products. This tool can be applied to a wider range of food products to improve food source traceability too. Further studies are needed to improve our knowledge on the influence of rearing conditions and processing on the edible insect associated with the microbiota.

2.2.6 Data availability statement

The dataset generated for this study was submitted to the EBI metagenomics portal (https://www.ebi.ac.uk/metagenomics/; Study ID: PRJEB35480).
Supplementary Materials are available through the main paper (https://doi.org/10.1016/j.foodres.2020.109426).

2.3 Additional contributions: food and ecology applications

In this chapter, abstracts of the work in which I contribute are provided. For complete manuscripts, see Supplementary Data link.

2.3.1 DNA-Based Herbal Teas' Authentication: An ITS2 and psbA-trnH Multi-Marker DNA Metabarcoding Approach

Medicinal plants have been widely used in traditional medicine due to their therapeutic properties. Although they are mostly used as herbal infusion and tincture, employment as ingredients of food supplements is increasing. However, fraud and adulteration are widespread issues. In our study, we aimed at evaluating DNA metabarcoding as a tool to identify product composition. In order to accomplish this, we analyzed fifteen commercial products with DNA metabarcoding, using two barcode regions: psbA-trnH and ITS2. Results showed that on average, 70% (44–100) of the declared ingredients have been identified. The ITS2 marker appears to identify more species (n = 60) than psbA-trnH (n = 35), with an ingredients' identification rate of 52% versus 45%, respectively. Some species are identified only by one marker rather than the other. Additionally, in order to evaluate the quantitative ability of high-throughput sequencing (HTS) to compare the plant component to the corresponding assigned sequences, in the laboratory, we created six mock mixtures of plants starting both from biomass and gDNA. Our analysis also supports the application of DNA metabarcoding for a relative quantitative analysis. These results move towards the application of HTS analysis for studying the composition of herbal teas for medicinal plants' traceability and quality control.

2.3.2 Impact of land use intensification and local features on plants and pollinators in Sub-Saharan smallholder farms

Sub-Saharan African crop production largely relies on smallholder farms, located both in urban and agricultural landscapes. In this context, the investigation of plant and pollinator diversity and their interactions is of primary importance since both these factors are threatened by land use intensification and the consequent loss of natural habitats. In this study, we evaluated for the first time how plant and pollinator insect assemblages and interactions in Sub-Saharan farming conditions are shaped by land use intensification. To do that, we complemented biodiversity field surveys in Northern Tanzania with a modern DNA metabarcoding approach to characterize the foraged plants and thus built networks describing plant-pollinator interactions at the individual insect level. Moreover, we coupled this information with quantitative traits of landscape composition and floral availability surrounding each farm. We found that pollinator richness decreased with increasing impervious and agricultural cover in the landscape, whereas the flower density at each farm correlated with pollinator richness. The intensification of agricultural land use and urbanization correlated with a higher foraging niche overlap among pollinators due to convergence of individuals' flower visiting strategies. Furthermore, within farms, the higher availability of floral resources drove lower niche overlap among individuals, while a greater flower visitors abundance shaped higher generalization at the networks level (H2′), possibly due to increased competition. These mechanistic understandings leading to individuals' foraging niche overlap and generalism at the network level, could imply stability of interactions and of the pollination ecosystem service. Our integrative survey proved that plant-pollinator systems are largely affected by land use intensification and by local factors in smallholder farms of Sub-Saharan Africa. Thus, policies promoting nature-based solutions, among which the introduction of more pollinator-friendly practices by smallholder farmers, could be effective

in mitigating the intensification of both urban and rural landscapes in this region, as well as in similar Sub-Saharan contexts.

2.3.3 Food Tracking Perspective: DNA Metabarcoding to Identify Plant Composition in Complex and Processed Food Products

One of the main goals of the quality control evaluation is to identify contaminants in raw material, or contamination after a food is processed and before it is placed on the market. During the treatment processes, contamination, both accidental and economically motivated, can generate incongruence between declared and real composition. In our study, we evaluated if DNA metabarcoding is a suitable tool for unveiling the composition of processed food, when it contains small trace amounts. We tested this method on different types of commercial plant products by using tnrL marker and we applied amplicon-based high-throughput sequencing techniques to identify plant components in different food products. Our results showed that DNA metabarcoding can be an effective approach for food traceability in different types of processed food. Indeed, in the vast majority of our samples, we identified the species composition as the labels reported. Although some critical issues still exist, mostly deriving from the starting composition (i.e., variable complexity in taxa composition) of the sample itself and the different processing level (i.e., high or low DNA degradation), our data confirmed the potential of the DNA metabarcoding approach also in quantitative analyses for food composition quality control.

2.4 Main conclusions and future perspectives

2.4.1 Insights into bioinformatic frameworks of non-bacterial markers

This first chapter is dedicated to the main works in which I contributed during the experimental design, bioinformatic analysis and paper writing phases. In general, I observed the following issues: non-bacterial markers, such as ITS2 for Plants traceability, do not have a clear standardization into the data analysis process. In my experience, there is a great effort to communicate the results to non-bioinformatic people and to find solutions to correctly define the threshold able to recapitulate the molecular information and species detected. Considering the food industry, the abundance of a species, in terms of reads and reliable sequences identified, assume a great importance: what is the difference between a contamination, a bias or a fraud? Is it possible to identify the correct threshold to discriminate between them? It is clear that answering is not an easy task. In addition, it is a delicate point, both considering marketing and industry implications. In our work, expertise of the field of study and the communication between researchers and partners become fundamental to conduct a scientifically sound, clear and useful study.

In addition, the second important step of the bioinformatic process is the taxonomy assignment. Considering the pipelines reported above and the work in which I contributed, taxonomy assignment was performed using NCBI as a reference database. A smaller collection of sequences available in NCBI were downloaded and used as reference. Further, a manual curation step was performed by colleagues that have prepared samples and sequencing libraries. A clear interconnection between the case study under investigation, laboratory issues and sequencing biases exist. This leads to difficulties to guarantee a fully standardized procedure and framework, also from the bioinformatic point of view. Currently, the main difficulties lie on a precise standardization of processes, both considering experimental and bioinformatic frameworks. From primer selection to the extraction of reliable sequences, a deep work must be

done considering the very different matrices we can analyse. Probably, tests starting from sequencing run will be necessary to clearly disentangle the issues described above. Of course, as DNA metabarcoding has great potentials both in environmental and food-related projects, continuing to work will help to propose new and better solutions. For the taxonomic assignment, for example, I invite you to read **Chapter 4.**

In the following chapter, I present the potentials of omics tools into fermented food products research area. Specifically, fermented foods can be viewed as a subcategory where foods and microorganisms exploration are strictly connected. I report below my contribution to two main chapters of the Review "Fermented food products in the era of globalization: tradition meets biotechnology innovations" (Galimberti et al., 2021), where me and my colleagues depict the recent advancements of DNA metabarcoding in this field and the future perspectives related with the potentiality of omic tools.

2.4.2 Fermented food products in the era of globalization: tradition meets biotechnology innovations

Omics tools offer the opportunity to characterize and trace traditional and industrial fermented foods. Bioinformatics, through machine learning, and other advanced statistical approaches, are able to disentangle fermentation processes and to predict the evolution and metabolic outcomes of a food microbial ecosystem. By assembling microbial artificial consortia, the biotechnological advances will also be able to enhance the nutritional value and organoleptics characteristics of fermented food, preserving, at the same time, the potential of autochthonous microbial consortia and metabolic pathways, which are difficult to reproduce. Preserving the traditional methods contributes to protecting the hidden value of local biodiversity, and exploits its potential in industrial processes with the final aim of guaranteeing food security and safety, even in developing countries.

2.4.2.1 Microbial ecosystem as a valuable signature of fermented food typicalities

The metagenomics revolution started to provide researchers with catalogues of gene (or genome) sequences of bacteria and yeasts from many fermented food categories (Parente et al., 2019; Pasolli et al., 2020). Metagenomics-based approaches identified groups of functional microorganisms able to (i) enhance the bioavailability of nutrients and the sensory quality of fermented foods, (ii) impart bio-preservative effects, (iii) improve the safety of food products, and (iv) provide positive effects to human gut microbiota and health conditions (Marco et al., 2017). However, in each fermentation phase there are also 'non-functional' microorganisms playing a key role in maintaining the stability of the whole microbial ecosystem. Therefore, both functional and non-functional microorganisms are essential for providing more standardisation and precision to the industrial biotechnological processes and for shaping the food flavour and taste to reach the desired gold standards (Tamang et al., 2016; De Filippis et al., 2018).

As a matter of fact, the fermented food market focuses on increasing the sensory reward in the consumer by branding a unique microbial-mediated signature in terms of flavour, taste, and appearance. This microbial fingerprint permits consumers to be able to distinguish a product from others similar that are available on the market (Van Reckem et al., 2019; Kamimura et al., 2020). In this context, wine is one of the fermented items that has the greatest link to the territory:its flavour (and brand) is shaped by the influence of pedoclimatic characteristics of the vineyard and wine cellar and of their associated microbial consortia (i.e. the microbial terroir, sensu (Bokulich et al., 2014; Bokulich et al., 2016). Bokulich et al. (2014) also demonstrated that this microbial signature of the zone of production and/or biotransformation also persists in the bottle, allowing it to be traced back to the production cellar. These bacteria belong to the soil of the vineyard but also come from the local associated

pollinator/frugivore metazoans, pathogens, and fertilization management (Mezzasalma et al., 2018).

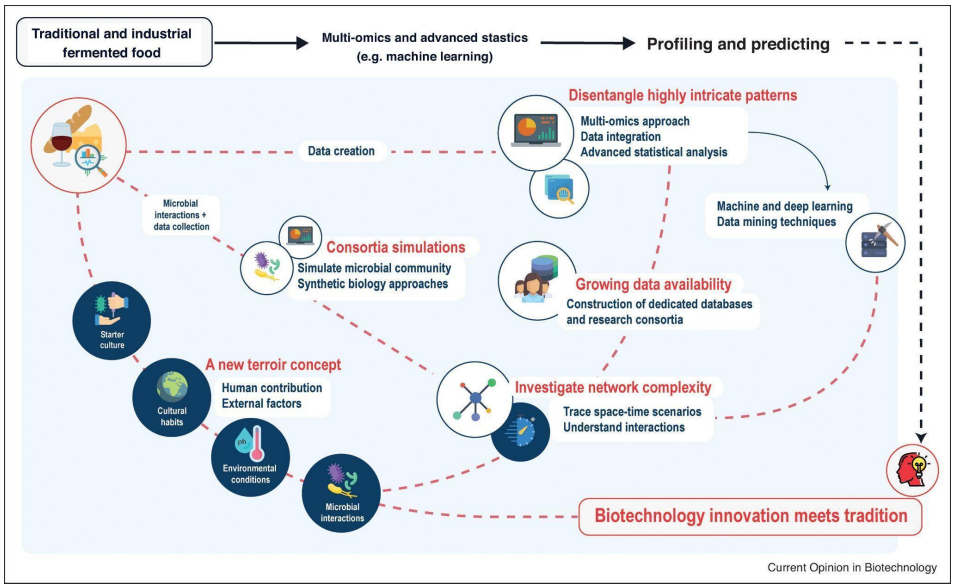

Figure 1. The multiple facets of fermentation processes. Multi-omics and advanced statistical analysis combined with ad-hoc database and data integration are the frontiers into exploring fermented microbial worlds, with the aim of profiling food from microbial interactions to space-time scenarios and predicting community behaviour, both considering environmental conditions and starter culture contribution. Technology leads to harnessing the potential of autochthonous microorganisms and driving biotechnological advances to improve not only the process efficiency and guaranteeing safety and quality, but also preserves the ancestral intrinsic biodiversity at the micro-scale to govern the outcomes at the macro-scale.

A similar scenario also occurs in dairy production where the autochthonous microorganisms are overwhelmingly dominant over commercial strains, suggesting that starter inoculation may have a limited influence on driving cheese microbiota (Bokulich et al., 2013). Again, although relying on the same ingredients, bread flavour and taste diversity around the globe reflect the contribution of peculiar microbial taxa belonging not only to the used flour and

yeast, but also to the microbial community of the baker's hands (Reese et al. 2020). Since some flavours and tastes of fermented foods should also be attributed to variations in spontaneous microbial consortia (Johansen et al., 2019; Akaike et al., 2020), the current frontier in food fermentation biotechnologies is the accurate reproduction of a traditional microbial consortium and its dynamics that direct the entire biotransformation process in order to obtain high-quality food products of increased nutritional value. Moreover, both the scientific community and the consumers are paying increasing attention to the effects of fermented food on the gut microbiome. It has largely been established that the human gut ecosystem covers a special interest since it deals with nutrients adsorption, energy availability, hormone balance, immune system activation, and even behaviour. Subtle, yet persistent signatures associated with the consumption of fermented food impact the gut microbiota structure and enrich it with beneficial compounds (Taylor et al., 2020). The existence of a gut–brain axis is also well documented and growing evidence suggests that fermented foods can positively impact mood and brain activity through a gut cross-talk mediated by microorganisms (Aslam et al., 2020). In this context, the technological challenge is to develop suitable tools to accurately characterize, model and mimic the complex world of food-related microbial ecosystems (Wolfe et al., 2014). Such tools could also be used to predict the changes occurring in the microbial consortia over time, their role in fermentation processes, their effects on the final transformed products, and therefore on human health.

2.4.2.2 Predicting the microbial ecosystem dynamics of fermented food

The characterization of interactions occurring within the food microbial community is a challenging task (Smid et al., 2013). The emerging multi-omics (e.g. genomics, proteomics, meta-bolomics) data have highlighted the molecular mechanisms (i.e. metabolic pathways) occurring in microorganisms and the effects of interactions taking place among these microorganisms and

with external factors, which usually affect the evolution of the whole microbial community (Parente et al., 2018; Afshari et al., 2020). The application of association network analysis in food ecosystems showed that the most frequent interactions mainly regard co-occurrence relationships involving starter and spontaneous microorganisms, that can vary depending on environmental variables that are difficult to disentangle (Layeghifard et al., 2017). In addition, mutual exclusion relationships have been observed between beneficial and spoilage microorganisms with obvious implications in terms of food safety (Parente et al., 2018). Combining molecular data and culture-based methods validates the functional association networks and identifies the key players involved in each metabolic pathway (Shetty et al., 2019). Unfortunately, this approach does not work with the uncultivable microorganisms. These usually constitute a predominant fraction of the microbial community and are often responsible for unique metabolic pathways enriching foods with peculiar flavours and aromas (Solden et al., 2016; Bruno et al., 2017). Alternatively, the multi-omics approaches offer the opportunity to reconstruct the key metabolic pathways of microbial ecosystems regardless of the exact taxonomic knowledge of the involved microorganisms. Once the relevant and intermediate metabolites have been identified through a metabolomics approach, it is then possible to hypothesize which microorganisms are responsible for their production or are the most likely involved. As an example, more than one thousand microbial protein clusters were identified in bean sauce mash, a traditional fermented soybean product. Metabolomic analysis suggested that these were expressed mainly by members of two genera of the microbial consortium, leaving the role of the vast remaining microbial community of the fermented sauce almost unsolved (Xie et al., 2019 - a; Xie et al., 2019 - b). The same approach was also used to elucidate the microbiota changes during fermentation and their effect on the metabolite contents of some traditional products, such as Chinese Pu-Erh tea (Zhao et al., 2019), to assess the role of peculiar taxa other than lactic acid bacteria in the progression of fermentation in dairy products (Afshari et al., 2020; O'Donnell et al., 2020), or to characterize

the molecules shaping the flavour and taste of specific wine products(Sirén et al., 2019). Concerning the definition of the roles of external inputs on the fermentation processes, the increasing amount of biological data, including those derived from unique local contexts, coupled with integrative data analysis (e.g. multiple co-inertia analysis McIA; Sankaran and Holmes, 2019) are considered promising. A virtuous case-study in this context is that by Afshari et al. (2020) who were able to characterize the microbiota and metabolite fingerprints involved in the ripening process of different brands of artisanal and industrial cheddar cheeses to a deeper extent. Moving forward in this direction, it will be possible to disentangle the complex networks, the relevant hubs, and the bottlenecks of the whole fermentation ecosystem.

Data-driven and knowledge-based approaches are now the cutting-edge solutions for exploring and understanding microbial ecosystems and are mainly based on advanced statistical approaches and data mining techniques. Ranging from classification and microbial signature prediction to interaction and features associations (i.e. organisms, metabolites) (Pasolli et al., 2016; Qu et al., 2019), machine learning strategies have the potential to identify hidden structures (Noor et al., 2019) and highly intricate patterns that may help predict biological functions (Thompson et al., 2019). These approaches pave the way for tracking microbial landscapes and temporal dynamics and unveiling the added value of geographical diversity and environmental contribution (Bokulich et al., 2016). Computational simulations have become of fundamental importance to address the difficulties of understanding complex microbial ecosystems. Combined with machine and deep learning methods, these approaches are emerging into discovering the microbial community structure and dynamics, also uncovering new putative interactions among organisms (Marsland et al., 2020). This is the case of a new inference method proposed by Lee et al. (2020) that is able to predict microbial community interactions from spatially distributed data. To disentangle elusive metabolic interactions among microorganisms, Dimucci et al. (2018) combined partially known networks with trait-level information, exploiting the potential of machine learning algorithms

and providing an innovative procedure to find underlying associations. Great efforts were also provided by advances in data mining research. The need to get more information boosted the development of text mining tools to explore microbes and take advantage of non-bacterial strategies to extract new infra-community patterns (Tandon et al., 2016; Chaix et al., 2019).

All of these predictive systems increase their own accuracy and reliability depending on the growing availability in data resources (Parente et al., 2019) fed by projects cataloguing microbial life at an unprecedented vast scale (e.g. The Earth Microbiome Project EMP, http://www.earthmicrobiome.org) and by the contribution of crowdsourced campaigns. Citizen science projects, such as the American Gut Project (http:// americangut.org/), the Microsetta Initiative (https:// microsetta.ucsd.edu/) and the Global Sourdough Project (http://robdunnlab.com/projects/sourdough/), conjugate microbiome analysis services to general consumers, thus providing a huge and precious source of data, with the lateral effect of filing microorganisms typical of traditional/artisanal fermentation processes (Ryan et al., 2018) and investigating the relationship between the consumption of fermented food and the equilibrium of the gut microbiota.

3. Dealing with the promise of metabarcoding in mega-event biomonitoring: EXPO2015 data report

3.1 Introduction

Environmental degradation due to anthropic activities have increased the scale and frequency of biodiversity assessments. Certainly, the environmental degradation is particularly dramatic in the highly anthropogenic areas. Governments and international organisations are issuing and thus including in their agendas new strategies to protect and restore biodiversity, such as the Intergovernmental Science-Policy Platform on Biodiversity and Ecosystem Services (IPBES 2019; Bongaarts, 2019; Lanzen et al., 2016; Baird et al., 2012). These circumstances have required a renovation in monitoring techniques, encouraged by the necessity to develop more rapid and accurate tools supporting timely observations of ecosystems structure and functions (Taylor et al., 2016; Pimm et al., 2015). In this framework, supported by Next Generation Biomonitoring (NGB) initiatives (Makiola et al., 2020), DNA metabarcoding introduced surprising signs of progress in surveying prokaryotic and eukaryotic diversity from any type of environment (Makiola et al., 2020; McGee et al., 2019). After a few years from the adoption of DNA metabarcoding, many worldwide molecular data collection projects include DNA metabarcoding data, as a natural progression for biodiversity assessment e.g. Earth BioGenome Project (EBP 2021), The European Reference Genome Atlas initiative (ERGA 2021), the BIOSCAN project (BIOSCAN2021), the Vertebrate Genomes Project (VGP) (Rhie et al. 2021), the i5k Arthropod Genomes Initiative (i5K Consortium 2013), the 10KP Plant Genomes Project (Cheng et al. 2018), and others (Waterhouse et al 2021). DNA metabarcoding widespread adoption has also been supported by the advances in high-throughput DNA sequencing (HTS) technologies, increasing data yield with costs reduction (Cordier et al., 2020; Porter et al., 2018; Pimm et al., 2015; Thomsen et al., 2015; Shokralla et al., 2012), allowing taxa exploration at unprecedented extent, for a time and

cost-effective biodiversity monitoring (Westfall et al., 2019; Ruppert et al., 2019; Deiner et al., 2017).

Several studies exploited the potential of DNA metabarcoding to improve the understanding of anthropogenic impacts (Cordier et al., 2021; Tommasi et al., 2021; Frontalini et al., 2018; Lanzen et al., 2016), monitoring alien species introduction (e.g., Westfall et al., 2019; Comtet et al., 2015), even in the context of regulatory policymaking (Cordier et al., 2021; van der Heyde et al. 2020; Pawlowski et al., 2018).

However, the DNA metabarcoding data analysis and interpretation still requires great efforts to achieve full data exploitation and not standard procedures could be applied for each taxa domain (Ruppert et al., 2019; Porter et al., 2018; Deiner et al., 2017). In particular, difficulties remain related to the lack of information in reference taxonomic databases (Curry et al., 2018; Weigand et al., 2019), taxonomic resolution and misidentification (Bush et al., 2019), leading also to the implementation of taxonomy free approaches (Vasselon et al., 2017).

Nevertheless, we harnessed the advantages and the huge amount of information that DNA metabarcoding can generate to investigate the influence of massive human-induced activities on biological communities, also considering the issues related to marker choice, reads processing and the information contained in reference databases, a fundamental part of data interpretation.

In this study, we focused on the monitoring of the World Exposition (from now “EXPO2015”) hosted in Milan from May to October 2015. This global event was categorized as a mega-event (Muller, 2015), which can be defined as an acute environmental stressor, possibly generating biodiversity alteration and disturbance.

During the six months of EXPO2015, exhibitors from more than 135 countries and 22 million visitors insisted on a 1.1 million square meters exhibition area.

Faced with such a massive event, a wide-range analysis of biodiversity could be reliable for addressing biomonitoring purposes (Cordier et al., 2021; Cristescu et al., 2018; Alberdi et al., 2018; Trebitz et al., 2017; Comtet et al. 2015). To overcome restrictions of traditional biomonitoring, which is limited to observations on small sets of bioindicators and/or flagship species (Cordier et al., 2021; Dequiedt et al. 2011; Magurran et al. 2010; Reavie et al. 2010; Bonada et al. 2006), we applied a DNA metabarcoding approach targeting three different molecular markers and involving two different sampling strategies (i.e., water and air) to obtain a comprehensive overview of the impact of the exhibition on environmental community assemblages. In this context, both overall and microscale investigations were conducted. Specifically, we monitored the water canalization system, which connects two local rivers across the exhibition area, the two local rivers and the air biodiversity collected at two representative sites. We chose three mini-barcode regions allowing the assessment of a broad taxonomic spectrum of the eukaryotic community: the V9 hypervariable region of 18S SSU rRNA (Harrison et al., 2021; Fernández-Álvarez et al. 2018; Chariton et al. 2015; Cowart et al. 2015; Lallias et al. 2015; Zimmermann et al. 2015; Edgcomb et al., 2011), the plastid trnL intron (Deiner et al., 2017; Fahner et al., 2016; Quéméré et al., 2013; Taberlet et al., 1991), and the internal transcribed spacer ITS2 of rRNA (Nilsson et al., 2019; Blaalid et al., 2013; Toju et al., 2012; White et al., 1990).

Overall, our main intent was to validate DNA metabarcoding as a biomonitoring strategy to understand the environmental impact of global events, such as EXPO2015, on eukaryotic community diversity. In parallel, we tried to deepen the following questions: i) if DNA metabarcoding can track biodiversity communities in a mega-event context, ii) which are the pros and cons of using multi-marker strategies, considering the absence of common procedures and the issues related to the taxonomy assignment and iii) if machine learning strategies can help in predicting sample origin, overcoming the uncertainty of the taxonomy assignment.

Our results showed that DNA metabarcoding coupled with machine learning approach is a powerful genomic-based tool to monitor biodiversity at the microscale, allowing us to capture exact fingerprints of specific event sites and to explore in a comprehensive manner the eukaryotic community alteration. We discussed in the work the crucial issues related to the generalization of the approach and the degree of taxa identification. We provided a case-study application of DNA metabarcoding to an urban context, monitoring biodiversity at micro-scale, but also with a focus on the changes starting from the laying of the first stone. As well as the great potential of genomic-based tools and data related to genetic biodiversity are growing, machine learning approaches could give the decisive breakthrough to the application of DNA-based monitoring 3.0 at a broader extent.

3.2 Materials and methods

3.2.1 Study area and sampling design

The EXPO 2015 exhibition site is located northwest of Milan. The site occupies an area of 110 hectares, with approximately 250,000 m2 of vegetation, 6,000 m of canals, more than 70 exposition pavilions, for the exhibitors coming from more than 135 countries, built in three and a half years, and was completed just hours before the opening ceremony (Expo Milano 2015 Official Report 1).

It had long been an industrial zone before its conversion to logistical and municipal services and agriculture. The area is characterized by two parallel water canals and it is crossed by two rivers, Guisa and Olona.

Figure 1. Map of sampling sites. Blue circles indicate the two sampling points of air (S2 closer to the exposition site). Red circles indicate the four sampling points of water canals (P1 P2 ring water canal; C1 C2 water canal parallel).

Within the EXPO area, four main sampling points were considered (Fig. 1):

- P1: localized in the ring water canal, upstream of P2 (inlet water);
- C1: located along the water canal parallel to the area, receives incoming water from the Guisa river and enters more times in contact with the area;
- P2: localized in the ring water canal, downstream of P1, collects outlet water derived from the whole area and from P1;
- C2: located along the water canal parallel to the area, receives the water from the exhibition area and enters the river Olona.

Considering the water sampling, samples were collected using one-liter sterile single-use bottles (LP Italia) in PET from the two rivers crossing the area of the exhibition, Guisa and Olona, and at four sites localized within the EXPO area (P1, P2, C1 and C2; see Figure 1). Sampling began in October 2014 and ended in March 2016. Since the works for the construction of the exhibition site have continued up to the days immediately prior to the opening of the event, the

sampling of water perimeter channels was not possible in the ante operam phase (i.e., before May 2015) in the EXPO area.
Regarding the post operam phase (i.e., after October 2015), the analyzed samples were collected at the same sampling sites, since the exposition area was no longer accessible. In total, Guisa was sampled six times (6 samples) and Olona three times (3 samples) as the river was dry, once a month (for details about sampling dates, see Supplementary S1).
The sites C1, C2, P1, and P2 were sampled monthly during the EXPO event (in operam phase), obtaining 30 samples of P1, 18 for C1, 33 for P2 and 26 for C2.

Considering the air sampling campaign, samples were collected monthly from October 2014 to March 2016 through two different methods: a Tauber Trap approach (Tauber, 1974) and Lanzoni VPPS 2010 (Lanzoni, Bologna, Italy) instrument (based on Hirst model; Hirst, 1952; Núñez et al., 2017)
Sites sampled were:

- S1, located at the company Tarenzi s.p.a, 600 meters north of EXPO (a total 44 samples);
- S2, located on the roof of c.m.p. Poste Roserio, 100 meters south of EXPO (a total of 47);

The S1 site was investigated using the Tauber Trap method, instead of the S2 site in which both instruments were installed. Sites were carefully selected for their geographical position, near the exhibition area and opposite each other, in order to collect the biological component considering wind direction. The different distance of sampling sites from the EXPO area allowed both short-range (100 meters) and long-range (600 meters) monitoring (c.m.p Roserio and Tarenzi s.p.a., respectively). Overall, a total of 228 samples were collected from water (137 samples) and air (91 samples), covering the period from October 2014 to March 2016 (for time point list see Tables in Supplementary S1), using three molecular markers. The sample distribution was conducted as follows: 34 air and 47 water samples belonging to 18S V9

region, 30 air and 45 water samples to trnL and 27 air and 45 water samples belonging to ITS2.

3.2.2 Samples pre-processing and environmental DNA extraction

Each liter of water belonging to each site was pre-processed with serial filtrations with the use of nitrocellulose and acetate membrane filters with 8 μm and 0.45 μm pore sizes (Jamwal et al., 2021; Valsechi et al., 2021; Capo et al., 2020), respectively. For the air sampling campaign, each Tauber trap sample (composed by a solution of ethanol and glycerol) was pre-processed with serial filtrations with the same strategy used for water samples.

Filters belonging to both media were initially crushed with Tissue-Lyser and liquid nitrogen. Subsequently, the DNA was extracted using the EuroGold Plant DNA Mini Kit (EuroClone). DNA extraction from samples subjected to mechanical lysis was carried out following the protocol for dry material with the following modifications: instead of starting from 250 mg of dry material, all the filters obtained for each sample were processed together, so that the DNA extracted corresponded to the volume of filtered water. DNA elution was carried out with 100 μl of elution buffer.

Three genetic markers (i.e, the nuclear V9 region of 18S rDNA and ITS2 and the plastid intron trnL) have been selected. The V9 region of 18S rDNA was used as a generalist genetic marker to explore the eukaryotic community (Harrison et al., 2021; Fernández-Álvarez et al. 2018; Chariton et al. 2015; Cowart et al. 2015; Lallias et al. 2015; Zimmermann et al. 2015; Edgcomb et al., 2011). The plastid intron trnL and the internal transcribed spacer ITS2 were used specifically to identify Plantae (Deiner et al., 2017; Fahner et al., 2016; Quéméré et al., 2013; Taberlet et al., 1991) and Fungi (Nilsson et al., 2019; Blaalid et al., 2013; Toju et al., 2012; White et al., 1990), respectively.

Raw reads were generated in an eighteen month assessment (from October 2014 to March 2016; Supplementary Files S1), collecting a total of 228 samples

(i.e., 137 water and 91 air), sequenced at the three selected loci markers (Supplementary Table S3).

3.2.3 Illumina library preparation and sequencing

V9 hypervariable region of 18S rRNA gene, intron trnL and ITS2 (primer details are provided in Supplementary Table S3) libraries were generated following the standard protocol (16S Metagenomic Sequencing Library Preparation, Part # 15044223 Rev. B). Amplicon PCRs were performed using the primer pairs used for qPCR quantification plus the adapter sequence. Libraries were quantified with a 2100 Bioanalyzer (Agilent Technologies) and sequenced with the Illumina MiSeq platform (five runs, v2 chemistry, 2x150bp). Library preparation and sequencing were carried out at IBIOM-CNR (Bari, Italy). Quantification protocol and primer list are available in Supplementary Data S2.

3.2.4 Bioinformatic workflow, biodiversity and machine learning analysis

For each marker gene, the raw paired-end FASTQ reads were imported into the Quantitative Insights Into Microbial Ecology 2 program (QIIME2, ver. 2020.8; Bolyen et al., 2019) and demultiplexing native plugin. Illumina runs were processed independently with the Divisive Amplicon Denoising Algorithm 2 (DADA2) plugin (Callahan et al., 2016). DADA2 was used to filter, trim, denoise, merge, remove of chimeras and calculate ESVs (Exact Sequence Variants; Callahan et al. 2017). In particular, an expected error = 2.0 was used as an indicator of read accuracy. Primers were trimmed and low-quality bases were removed. ESVs sequences with at least 10 representatives were taxonomically assigned using OBITools (Boyer et al., 2015) by ecotag tool, comparing sequences with an ecoPCR database extracted from the EMBL database version r139 (Kanz et al., 2004).

For each marker gene, the results of the taxonomy assignment were analysed considering the percentage of rank assigned at different levels (Kingdom, Phylum, Class, Order, Family, Genus, Species).

In order to estimate the biodiversity variation, we calculated alpha and beta diversity index for each marker gene separately. In detail, differences among sample types (water and air), sites (sampling points) and the macro category (air: S; rivers: R; internal canals: P; external canals: C) were tested For alpha diversity, we considered Shannon metric and presence/absence observations. Differences were tested using the pairwise Krustall-Wallis test implemented in the alpha-group-significance QIIME2 plugin (Kruskal and Wallis, 1952). To assess how volatile a dependent variable (alpha diversity measured as Shannon diversity) is over an independent variable (time) in water and air medium, a volatility plot was generated for each marker. For beta diversity, we calculated Jaccard metric to test differences among sample types, sites and macro categories using a PERMANOVA analysis performed with beta-group-significance plugin (Anderson, 2001).

Subsequently, the Random Forest classifier implemented in the sample-classifier QIIME2 plugin (Bokulich et al., 2018) was used to classify samples based on sites and macro categories metadata. The number of trees to grow for estimation was set to 1,000. Overall accuracy (i.e., the fraction of times that the tested samples are assigned the correct class) was calculated for each factor. K-fold cross-validation was performed during automatic feature selection and parameter optimization steps. A fivefold cross-validation was also performed. Further, machine learning analysis was carried out considering the genetic information of all the three marker regions, based on sites and macro categories metadata.

Figures and plots were created through QIIME2 plugins (Bolyen et al., 2019; Anderson, 2001; https://github.com/qiime2/q2-taxa) and ExTaxsI tool (Agostinetto et al., 2021; Agostinetto et al., 2020; https://github.com/qLSLab/ExTaxsI) to give an overview of biodiversity collected during the sampling campaign, with the aim to summarize the great

amount of data generated and help data interpretation. Raw reads were submitted to the ENA database (PRJEB45249).

3.3 Results

3.3.1 Sequencing results

Nine Illumina MiSeq sequencing runs for the three markers selected (18S SSU rRNA, trnL and ITS2) produced a total of 127,971,220 reads (63,985,610 pair-end reads), belonging to 228 samples. After the filtering steps, a total of 44,193,721 sequences were retained for the downstream analysis. As the DADA2 R package implements a full amplicon workflow (Callahan et al., 2016), we obtained a total of 19,304 ESVs (Callahan et al., 2017) for V9 raw reads, 3,630 ESVs for trnL and 8,471 ESVs for ITS2. Complete ESVs tables are available in Supplementary S8.

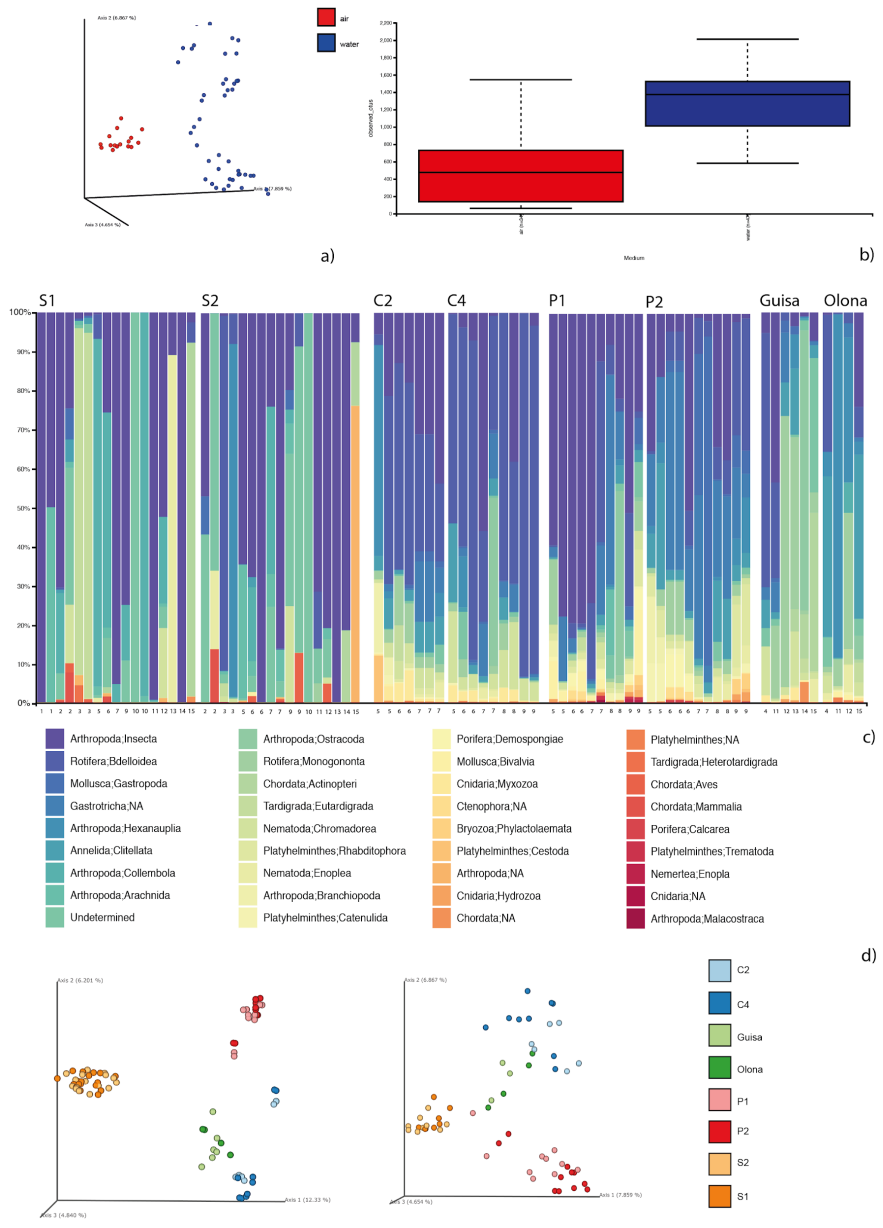

Figure 2. Resume figure showing the V9 18S results with: a) volatility plot and b) sunburst plot representing the taxa distribution of taxonomy assignment; c) taxa-bar-plot considering only Metazoa assignments; d) PCoA analysis based on Jaccard metric considering ESVs (left) and ESVs assigned only to Metazoa (right) on sampling sites.

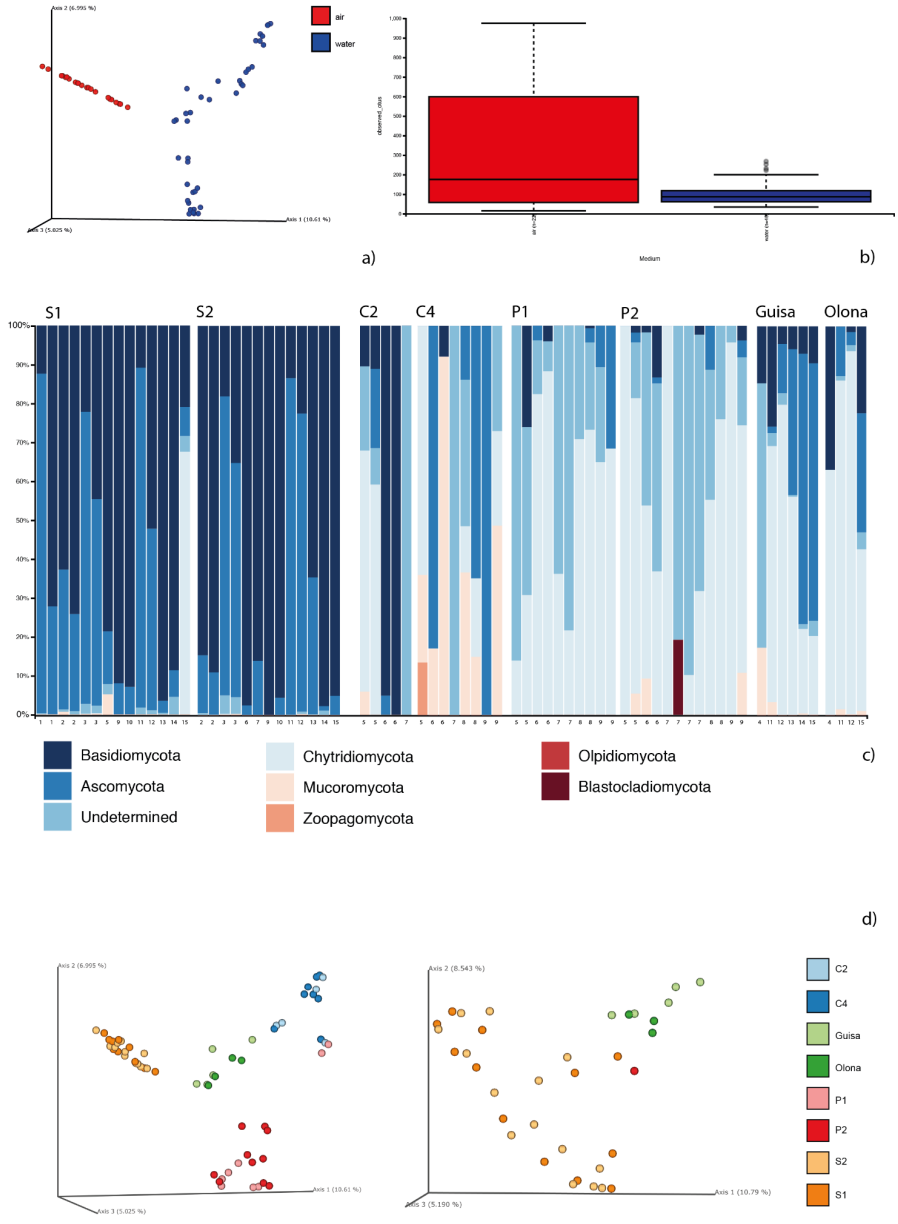

Figure 3. Resume figure showing the ITS2 results with: a) volatility plot and b) sunburst plot representing the taxa distribution of taxonomy assignment; c) taxa-bar-plot considering only Fungi assignments; d) PCoA analysis based on Jaccard metric considering ESVs (left) and ESVs assigned only to Fungi (right) on sampling sites.

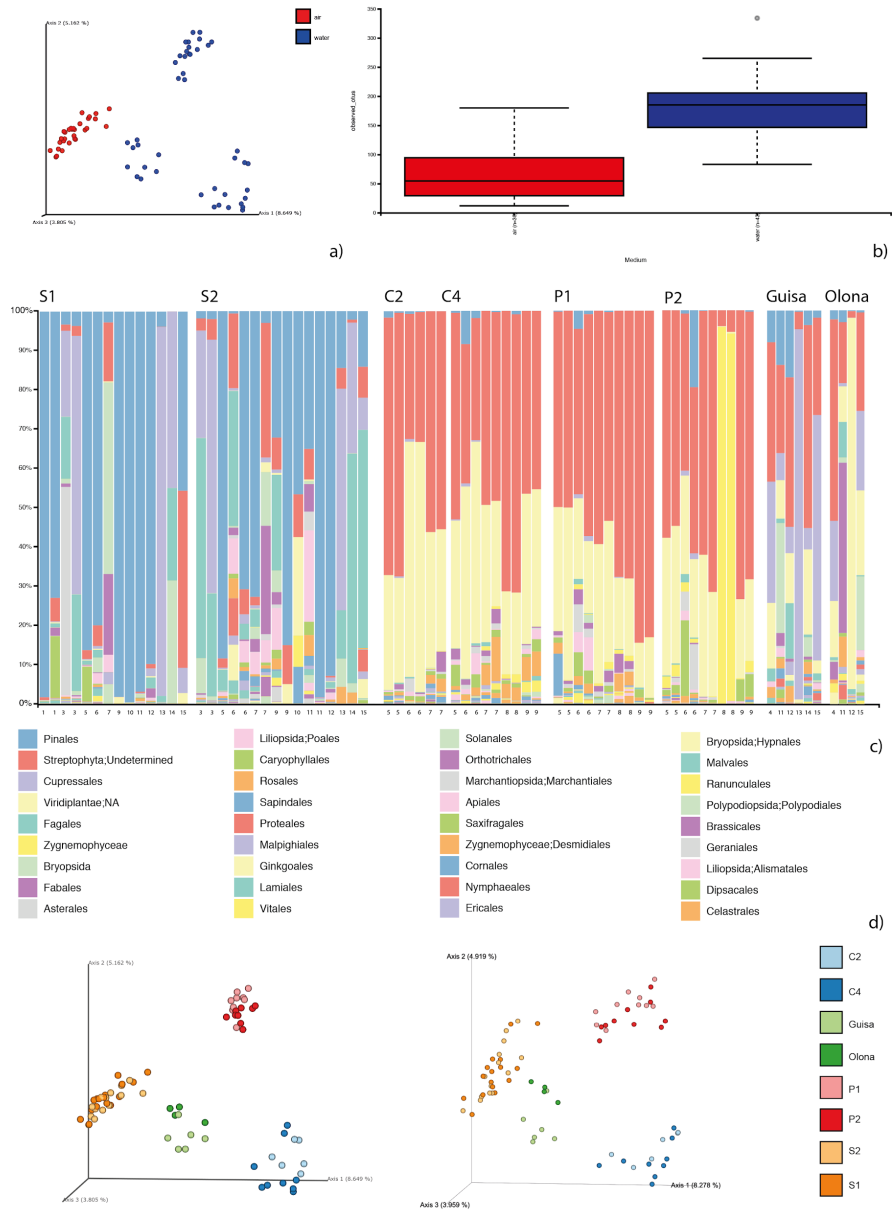

Figure 4. Resume figure showing the trnL results with: a) volatility plot and b) sunburst plot representing the taxa distribution of taxonomy assignment; c) taxa-bar-plot considering only Streptophyta assignments; d) PCoA analysis based on Jaccard metric considering ESVs (left) and ESVs assigned only to Streptophyta (right) on sampling sites.

3.3.2 Taxonomy results

V9 18S sequences resulted in 77.43 % assigned to Unicellular Eukaryotes, 13.20% to Fungi, 6.9% to Viridiplantae group, 3.58% to Metazoa and 2.54% to Bacteria, with 20.01% Unassigned sequences. Overall, 44.28% of assignments reached the genus level (Figure 5). Among Metazoa assignments, 39.65 % was composed by Arthropoda, 13.60% by Nematoda and 9.26 % by Rotifera. These taxa were followed by Platyhelminthes (7.8%), Unknown sequences (6.80%), Gastrotricha (6.80%), Annelida (4.62%), Cnidaria (3.76%), Chordata (2.31%), and Tardigrada (2.02%). A small fraction of assignments collected Mollusca (1.44%), Porifera (1.15%), Bryozoa (0.28%), Ctenophora (0.28%) and Nemertea (0.14%). Among Metazoa sequences, 45.5% of them reached a genus level assignment.

ITS2 sequences resulted in 64.29% of Fungi assignments, followed by 17.99% of Unclassified Eukaryotes, 11% of Unassigned sequences, 8% of Viridiplantae and 0.36% of Metazoa sequences. Overall, 46.86% of sequences reached a genus level assignment. Among Fungi sequences, 29.99% of them were assigned to Ascomycota phylum and 28.54% to Basidiomycota phylum, followed by 2.23% of Chytridiomycota, 0.56% of Mucoromycota, 0.04 % of Zoopagomycota, 0.01% of Olpidiomycota and 0.01% of Blastocladiomycota.

Plastid trnL intron sequences resulted in 51.81% of Streptophyta assignments, followed by 42.17% of Viridiplantae Unassigned sequences and 6% of Chlorophyta sequences. Overall, 14.38% of sequences reached the genus level assignment. Among Streptophyta, 63.42% remained Unassigned.

For each marker gene, the distribution of taxa among sites can be consulted into the respective resume figures, in particular considering Metazoa for V9 18S, Fungi for ITS2 and Streptophyta for trnL (Figure 2-3-4, section “c”). In addition, tables with the complete taxonomy assignments are available in Supplementary S9.

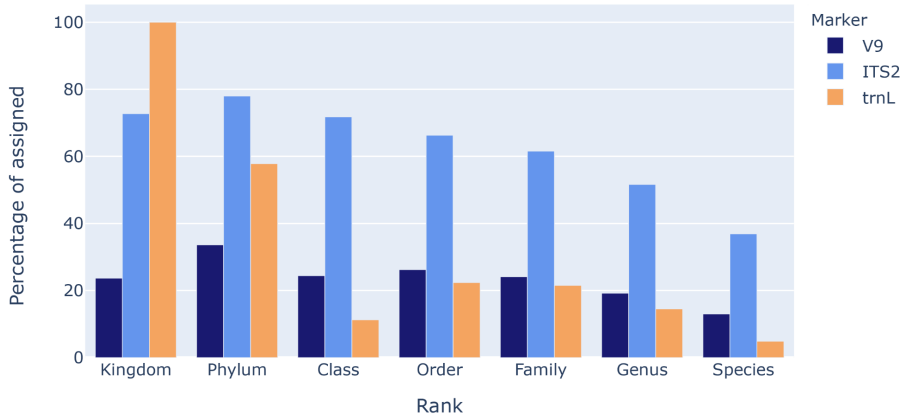

Figure 5. Percentage of sequences assigned for each marker region considering a 7 rank NCBI taxonomy (complete data are available in Supplementary S4).

To explore the results of the taxonomic assignment for each rank, we provide in Figure 5 a report with the percentage of sequences assigned for each marker region (complete data are available in Supplementary S4). In detail, we explored for each marker gene the taxa reached during the taxonomy assignment. The trnL intron was the only marker for which we assigned the 100% of the sequences obtained, in particular to Viridiplantae Kingdom. In addition, at least the 10% of ESVs with a complete Phylum and Order rank lacked Class information. Considering the number of species assigned, trnL was the marker which performed the worst.

From ITS2, we obtained the best results considering the taxonomy assignment, since the number of assigned ranks gradually decreased without gaps between one rank and another. Issues were found considering the Kingdom level: the Kingdom rank was not specified for taxa belonging to the Unicellular Eukaryotes group and a lower percentage of sequences were assigned if we consider the Kingdom rank instead of the Phylum.

V9 18S had the same issues of ITS2 considering Kingdom and Phylum ranks, as sequences belonging to V9 18S were also assigned to the Unicellular Eukaryotes group. For this reason, issues related to Class information were observed. In general, it was the marker with the highest percentage of Unassigned sequences.

A summarization of assignments was represented with sunburst plots in Figure 2-3-4 section “b”, accompanied by the results related to the ranks assigned in Figure 5. In addition, for each marker we added interactive sunburst charts obtained via ExTaxsI tool to explore dynamically the taxonomy obtained (Supplementary S7; Agostinetto et al., 2021; Agostinetto et al., 2020).

3.3.3 Biodiversity analysis

For each genetic marker, a biodiversity analysis was performed considering the type of sample, the sampling site and the macro category (air: S; rivers: R; internal canals: P; external canals: C).

The data analysed consisted not only of the ESVs calculated with the bioinformatics analysis, but also filtering the assigned ESVs considering different taxonomic levels, based on the results of the taxonomy assignments. In particular, for V9 18S we considered sequences assigned only to the Metazoa group and also sequences assigned to eukaryotic taxa, overall. For ITS2, sequences assigned to Fungi were considered. For trnL, Streptophyta sequences and sequences excluding Chlorophyta were considered. Finally, A PERMANOVA test was used to assess statistical significance.

Considering alpha diversity analysis, a significant difference was observed between air and water samples of V9 18S and trnL. In addition, considering V9 18S, a difference was found between macro category samples of each group considered, in particular between C and P. For ITS2 sequences a difference was detected among macro category sites (C-P and C-R). A stronger difference was detected considering Fungi sequences, also regarding water samples sites (C2-GU; GU-P1; GU-P2). For trnL, all the three types of analysis (only ESVs,

Streptophyta sequences and sequences excluding Chlorophyta) showed a difference among water and air samples. For details about alpha diversity results, see Supplementary S5a.

Considering beta diversity, sample medium (air and water), sampling sites and macro category were analysed. A significant difference was observed between the two different sampling media, for all the markers tested, both considering ESVs and taxa detected. In addition, significant differences were found comparing both macro category and sampling sites, both considering ESVs and taxa detected. For details about PERMANOVA results, see Supplementary Material S5b.

Overall, PCoA plots (Figure 2-3-4, section “d”) showed a significant structuration (model results are reported in Supplementary Materials S5) based on sampling site (different sampling point in EXPO2015 area), with the Internal sites (P1-P2) clustering close to each other, as well as the External sites (C1-C2) and Rivers (OL-GU). The same significant structure is also visible considering the taxonomic information, for which we reported in the main figures the results about Metazoa group (18S V9), Fungi (ITS2) and Streptophyta (trnL) (Figure 2-3-4 section “d”).

3.3.4 Machine learning analysis

DNA metabarcoding monitoring studies often aim to differentiate samples based on their biodiversity composition, a task that can be efficiently performed by Supervised Learning methods (Knights et al., 2011, Bokulich et al, 2018). We used a supervised machine learning approach to evaluate the potential of DNA metabarcoding data to classify sampling sites and macro category outcomes, considering the different types of information that we obtained based on the taxonomy assignments results (see the section above).

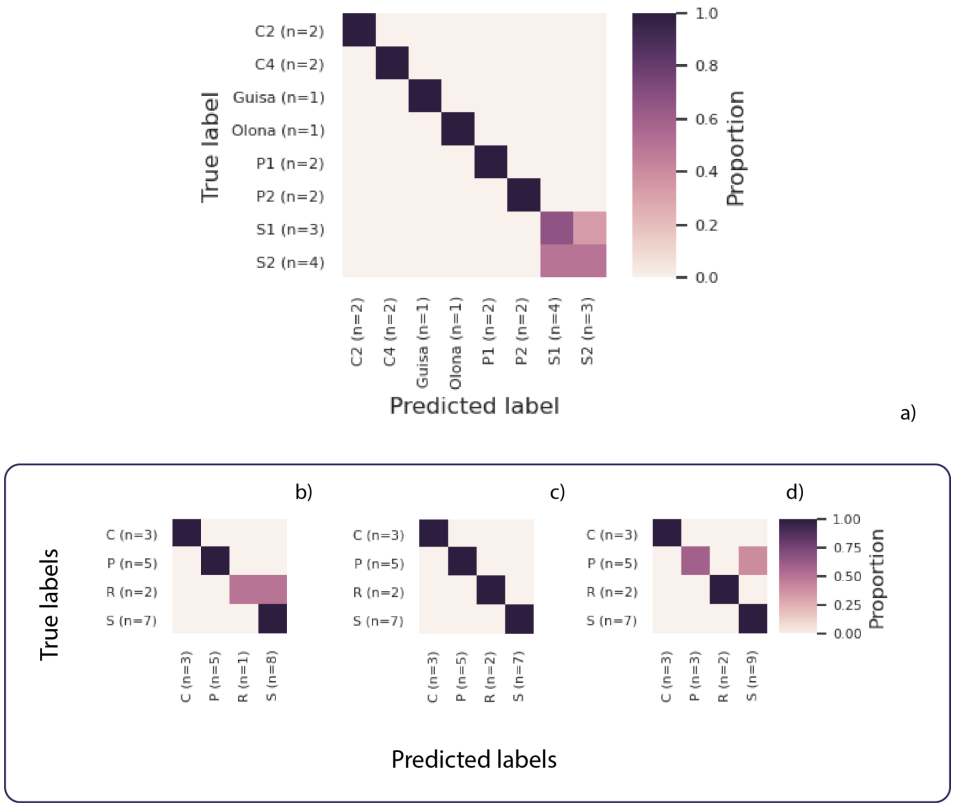

Figure 6. SL results of V9 18S marker: results were reported considering a) sequences per sites and considering the division of sites into macro categories of b) sequences, c) Eukaryota, d) Metazoa. The Figure shows a scatter plot of true vs. predicted values for regression results.

For all the three markers analysed, an improvement of classification was seen passing from sites to macro category metadata prediction. This trend is visible in Figure 6 for V9 18S and in Supplementary S10 for trnL and IT2.

Considering V9 18S marker, the set of ESVs was able to discriminate between water sampling sites with high precision and recalling, difficulties remained in the classification of different air sites, but without bias in defining the macro category air site (S). Considering the use of data filtered, an optimal result was obtained considering Eukaryota sequences (Figure 6c), obtaining a higher recall. Using only ESVs (Figure 6b) or Metazoa sequences (Figure 6d), the

macro category site prediction resulted in a lower performance, in particular regarding Rivers and Air sites prediction for sequences, and P and Air sites for Metazoa.

The trend described above was quite similar also for the other two marker regions. In detail, ITS2 sequences and Fungi worked well considering macro category prediction; predicting sampling sites using ESVs, instead, did not obtain optimal results, in particular for sites belonging to Guisa and Olona and P canals. Air sites, instead, were not correctly classified considering the division into the two sites, but the macro category was maintained (for details, see Supplementary S10).

Overall, machine learning analysis considering trnL markers reached good results, both considering ESVs and data filtered by taxa. Macro category prediction was reached. Streptophyta filtering showed difficulties in distinguishing River and Air sites. In general, the recalling was high (for details, see Supplementary S10).

Considering the information obtained by the three marker region sequencing, we decided to integrate all the data obtained from the ESVs calculated and run the machine learning classification considering all the ESVs obtained from the nine runs. The results are shown in Figure 6, representing the scatter plots obtained considering sampling sites and macro category prediction (Fig. 6b and c). In addition, the heatmap with the ESVs selected was represented, accompanied by the taxonomy at the Order rank (Fig. 6a). Also in this case, it is possible to observe a clear distinction between water and air medium type, a clear discrimination of canal water sites (C and P).

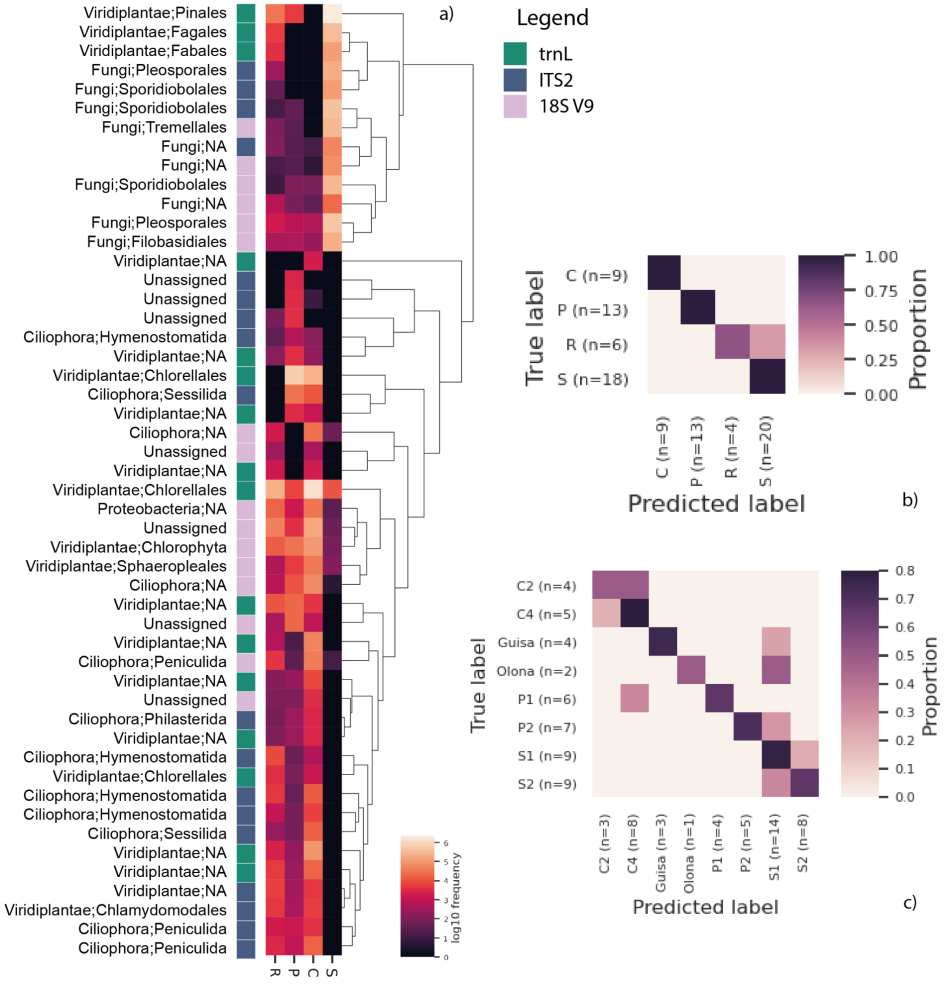

Figure 7. a) scatter plot representing ESVs with the additional information of marker and taxonomy (in particular: pink for V9 18S, green for trnL and blue for ITS2); scatter plot representing ESVs both using macro category sites (b) and sampling sites (c).

3.4 Discussion

In this study we have tested the reliability of DNA metabarcoding in capturing the environmental effects on eukaryotic diversity of a mega-event. Based on

the data obtained from the study, we tried to answer the following questions: is it possible to use DNA metabarcoding to track biodiversity communities in a mega-event context? Which are the pros and cons of using multi-marker strategies, considering the absence of common procedures and the issues related to the taxonomy assignment? Lastly, can machine learning procedures help in predicting sample origin, overcoming the taxonomic gap?

According to Bayraktarov et al. 2019, a common trend in most eDNA studies is the accumulation of data following the widespread opinion that "the more data there is, the better" (van Dorst et al., 2014). Nevertheless, this statement may be true only in specific contexts, since it disproportionately prioritizes data quantity over data quality. The value of data collected depends on how effective they are in achieving the solution to the key problem addressed (e.g., improving environmental management, Field et al., 2005, Miller et al., 2019).
Currently, DNA metabarcoding is the only reliable solution to collect large scale biodiversity data, as stated in the review paper of Cordier et al. 2021. The first question is if it is possible to use it as a routine tool for environmental biomonitoring. Several scientific studies demonstrated the feasibility of applying DNA metabarcoding in monitoring strategies, implementing both wet lab and bioinformatic pipelines in their workflows (Zafeiropoulos et al., 2020).
In this study, the event EXPO2015 provided an ideal system to test the effectiveness of DNA metabarcoding as a biomonitoring tool to check biodiversity variations in a critical ecosystem. The area is located in a highly urbanized environment, close to large suburban parks. These areas are proved to be an important reservoir of biodiversity, especially larger parks that could contribute more to the conservation of biodiversity (Cornelis and Hermy, 2004; McKinney, 2008; Beninde et al. 2015). Preserving these habitats is a fundamental point in the conservation of biodiversity, especially in a fragile context such as the urban one. For these reasons, validating eDNA metabarcoding tools is pivotal to monitor environments exposed to changes that could burden their equilibrium.

Going back to our first question, the capacity of our tool to detect variations is really powerful. Our results clearly showed a difference among the different sites (Figure 2-3-4, section “c” and “d”), but not a strong difference among the most distant sampling dates, considering both types of medium (Supplementary S6).
At the same time, the data collected by air and water were very different, both considering ESVs and taxa detected, as we can state by statistical analysis performed for all the markers investigated (see Supplementary S5 for details). Further, it was possible to individuate a fingerprint that made P sites different from C sites and rivers sites. In particular, PCoA analysis (Figure 2-3-4, section “d”) showed a clusterization among sites belonging to these three categories, identifying patterns of biodiversity that characterized distinct regions of the EXPO area, supported also by statistical tests and obtained considering both ESVs and taxonomic assignment (Figure 2-3-4, section “d”).

The approach addressed above opens our discussion to another question: is DNA metabarcoding a valid taxonomic identification tool? Several research papers demonstrated its application in diet characterization, water assessments, pollen identification and many other relevant fields (Porter et al., 2018; Deiner et al., 2017; Zhang et al., 2020). The taxonomic assignment step is still a delicate phase, as there are still no well-defined standards for each marker used in metabarcoding studies (Porter et al., 2018). In particular, the choice of the marker is still a compromise between two main aspects: i) the length of the DNA region that can be sequenced and, as a consequence, the genetic information that can be obtained; ii) the reference databases completeness and accuracy.
The first mostly regards the fact that any kind of matrix has its own characteristic. The key aspect to keep in mind is, usually, the DNA degradation, that can affect the reliability of the study. Shorter fragments are more likely to be detected, considering for example diets characterization or water

assessments, where DNA is exposed to multiple degrading sources, like chemical compounds (Deiner et al., 2017) and temperature (Krehenwinkel et al., 2018). At the same time, the marker chosen will influence taxa detection (Deiner et al., 2017): a gap of references recorded versus the known biodiversity exists for several relevant taxa, impeding a complete and correct taxonomic assignment (Cordier et al., 2021; Weigand et al., 2019; McGee et al., 2019). If this aspect is not considered, important biases could be included into experiments, leading to misinterpretations and excluding crucial information.
The markers that we evaluated as suitable for our experiments are a compromise among all these issues. Selecting short length markers (of about 150-200 base pairs), such as 18S V9 and the intron region of trnL, allowed us to collect a great number of information about eukaryotic and plant groups, even considering the highly degraded matrices we collected in our sampling campaign. Similarly, the longer region of the internal transcribed spacer ITS2 represented a good trade-off between low length variation and universality of primer sites (Nilsson et al., 2018), thus providing an overview of the Fungi Kingdom. In order to completely explore the potential of DNA metabarcoding, we decided to show not only the taxonomic assignments of ESVs, but also to include the analyses evaluating the ability of each marker to reach the taxa group for which they are recommended (Porter et al., 2018; Deiner et al., 2017), considering also researches already conducted (e.g. for trnL Quemere et al., 2013; 18S V9 Fernández-Álvarez et al. 2018; ITS2 Banchi et al., 2018). Sunburst plots of taxonomic distribution of taxa detected are shown in Figure 2-3-4 section “b”. The category of Metazoa, Plants and Fungi was extracted to show the balance of variations across sites in Figure 2-3-4 section “c”.

Some critical issues that emerged in our study are still at the center of the current scientific debate.
Despite the huge amount of data obtained from the sampling campaign, taxonomy assignment remains a difficult task. In general, the most of V9 18S sequences were assigned to unicellular eukaryotes taxa, followed by Fungi,

Viridiplantae and Metazoa, with a 20% of Unassigned sequences. Considering ITS2 sequences, the majority were assigned to Fungi, followed by Unicellular Eukaryotes taxa, with 11% of Unassigned sequences. Lastly, plastid trnL intron sequences resulted in Streptophyta assignments, with no Unassigned sequences. Overall, exploring taxonomy results helped us to consider sequencing outputs from different points of view. In particular, interactive sunburst in Supplementary S7 were created to enable a correct comprehension of the data obtained. Basically, there are issues with the standardization of taxonomy through different taxonomic groups (in particular Unicellular Eukaryotes and Viridiplantae). Knowing the difficulties of markers to reach genus or species ranks, it happened to investigate diversity considering for example families or orders gaps into the description of taxonomy that will not allow the data to be interpreted correctly.

For this reason, we decided to evaluate both biodiversity and machine learning analysis considering not only the taxa of interest, but also the genetic information that we obtained. In general, the analysis was coherent considering all the markers used, also subgrouping the ESVs based on particular taxa. But for the sake of interpretability and standardization, we believe that a focus on ESVs without the taxonomic assignment must be taken into account for a reliable and correct analysis of DNA metabarcoding data.

Overall, PCoA (Figure 2-3-4 section “d”) clearly showed a significant structuration based on sampling site, with Internal sites (P1-P2) cluster closely as External sites (C1-C2) and Rivers (OL-GU), demonstrating that ESVs composition could be the key to identify site types among the EXPO area (rivers, sites C and sites P), overcoming the gap in reference databases.

Aside from the spatial information, we explored the effect of the sampling month (see Supplementary S6 for details). We think that the effect of Site was predominant, though samples belonging to the same site and period of sampling were much more similar, suggesting the presence of a fingerprint due to both spatial and period. As the difficulties related to collecting samples during the EXPO event, we cannot ascertain the importance of sampling time.

But, considering several previous works related to temporal biomonitoring with DNA metabarcoding, we do not exclude time effects (Deiner et al., 2017; Pawlowski et al., 2018; Ruppert et al., 2019; Porter et al., 2018).
In general, our strategy suggests that the molecular information collected during the sample campaign was universally different in the sampling area and this trend was observed for all the three genetic surveys that we performed. The additional analysis carried out considering only the taxonomy demonstrated the strength of information collected.

Our last evaluation took into account if the DNA metabarcoding can be applied for predictive purpose. For this task, we used a machine learning approach both considering ESVs and taxa assigned.
We confirmed the biodiversity analysis conducted for all the markers. In addition, results indicated the use of sequences can be predictive, passing the taxonomy assignment that can be misleading. At the same time, a filter based on specific kingdoms suggests a peculiar structure for each taxa explored. Regarding this point, we are perfectly aware of the complexity of the communities that we analysed, but the recall is high, considering the vicinity of sampling sites and the medium of investigation.
In general, results obtained from machine learning classification showed three main aspects: the importance of sequences as a baseline pattern information of sites, the strength of the patterns considering also different taxonomic levels of analysis and, lastly, the optimization of the classification considering the macro site category. The use of taxonomy filtering for machine learning demonstrated the role of molecular fingerprinting, suggesting that the method can also be applied without reaching specific taxa information. This fact suggests two things: the first one is that DNA metabarcoding with middle-short region can be used for finding molecular fingerprinting in large-analysis and, in some cases, also taxonomy fingerprinting can be obtained and exploit (unfortunately, as we mentioned above, this really depends on the molecular marker used and the reference databases used for the assignment) (Schloss

and Westcott, 2011). The second, instead, demonstrated the difficulties in using DNA metabarcoding in reaching species-level information.
We think that the complexity should not be underestimated. Considering the data that we showed and the results related, the level of investigation may be very different, allowing researchers to answer several questions. The type of matrices, sampling method and marker used may lead to a real selection of communities under studies.

From the end of EXPO2015, no alien species were detected by state control agencies (ARPA) next to the exposition area. From the biomonitoring point of view, it is clear that advances in collecting data and contributing to public repositories could make a difference in interpreting these results. However, large amounts of biodiversity data may be useful for the generation of hypotheses (Bayraktarov et al., 2019). In the last few years, an increase of publications of 'DNA metabarcoding' in monitoring, biosurveillance and species invasions was observed (Piper et al., 2019). Though the advent of new sequencing technologies bring us the possibility to collect longer reads, therefore more genetic information, DNA metabarcoding with mid-short marker genes is still an important methodology in biodiversity assessment (Piper et al., 2019). From the biomonitoring point view, traces of ESVs could be informative alone, in order to study patterns and without focusing on specific taxonomic groups, for which it is possible to implement taxa-specific markers (Elbrecht et al., 2019) or classic monitoring strategies (Cordier et al., 2021).

3.5 Conclusions

DNA metabarcoding is nowadays widely used for very different purposes. Several research papers demonstrated its applicability into the monitoring of biodiversity. In this research paper, we wanted to highlight not only the power, but also the limitations that have to be considered in order to manage the data

and to give a conscious interpretation of data generated. As a genomic approach, limitations can be due to both markers chosen and the molecular information registered into the reference databases. Despite these issues, we demonstrated that the power of DNA metabarcoding is related not only to the molecular fingerprint obtained with sequencing ESVs, but also to the ability to collect a large amount of data, achieving a sort of freeze frame of the environment under study.

For these reasons, bioinformatics and post-processing analysis is still a pivotal process. Mining information from genomic data is still an important task, not without difficulties, and in this context collecting information and submitting datasets to reference databases will only ameliorate the comprehension of biodiversity all around the world, implementing both our current knowledge and future research. Considering the trends related to open science and our ability to sequence and produce data, data mining approaches (e.g. machine learning) will become more and more important, helping in disentangling high amounts of data, detecting biodiversity patterns and integrating additional information that give an edge to future studies.

3.6 Data availability statement

The dataset generated for this study was submitted to the EBI metagenomics portal (https://www.ebi.ac.uk/metagenomics/; Study ID: PRJEB45249).

Supplementary Materials are available through the main preprint paper (https://doi.org/10.1101/2022.01.02.474438).

4. ExTaxsI: an exploration tool of biodiversity molecular data

4.1 Introduction

In recent years, studies investigating biodiversity at large scale have started to create and incorporate molecular data. In particular, the spread of metagenomic studies (e.g. metabarcoding) have contributed to an exponential increase in genomic data availability. Thanks to this large amount of new information it is possible to expand our knowledge and enhance our scientific investigation capacity in many fields of research (Porter et al., 2018), ranging from macro-ecology and ecosystem monitoring, to food safety control, forensics applications and microbiome identification (Ruppert et al., 2019; Porter et al., 2018; Deiner et al., 2017).
Different groups of researchers emphasized the wealth of information collected in biological and molecular databases, with the aim to improve usefulness and reusability of data (Hampton et al., 2017; Whine et al, 2013; Michener et al, 2012). Therefore, building experimental designs that consider the totality of the data present in such databases could certainly increase the efficiency of these studies, and lead to more robust results (Mitchell et al., 2020; Almeida et al., 2019).

Biodiversity data retrieval and exploration has become a big data issue, forcing researchers to use Information Technologies (IT) tools to manage those data. In particular, the interpretation of results derived from metagenomic experiments, requiring computational pipelines and IT infrastructures that improve over time, is strongly linked to the availability of pre-existing data stored in online databases (e.g. ENA www.ebi.ac.uk/ena; and NCBI https://www.ncbi.nlm.nih.gov/).

In this context, data visualization represents an effective strategy not only to aggregate and present the research results, but also to guide advanced investigations (Kaur et al. 2018; Hardisty et al., 2013). At this moment, reference databases, where molecular and taxonomic data are friendly explorable and punctually updated, exist only for few molecular markers, such as SILVA for 16S and 18S genes (Pruesse et al., 2007), BOLD for animals and plants (Ratnasingham et al., 2007) or UNITE for Fungi domain (Nilsson et al., 2019). However, these data resources are not representative of all the genomic and taxonomic diversity collected to date.
On the other hand, although GenBank still resumes the majority of genetic data and their related metadata currently available (Keller et al., 2020; Ankenbrand et a., 2015; Benson et al., 2008), such information is not always easy to access without specific bioinformatics skills, which is a limiting factor to a large audience of scientists.

With the aim to help biologists to improve their experimental designs and to promote data exploration and exploitation, we have developed a tool, ExTaxsI (Exploring Taxonomy Information), able to facilitate the molecular data integration with its associated metadata, eventually retrieved from heterogeneous sources. Moreover, its ease of use interface will help researchers and practitioners in the visualization phase. ExTaxsI can both query the NCBI Nucleotide database for molecular data and accept data from an external source, exploiting the standard taxonomy notation.
To our knowledge, tools that provide user-friendly instruments to download and explore taxonomic data from NCBI have not been implemented yet. There are a few works to cite that perform parts of this task, focusing on slightly different goals. For example, NCBImeta (Eaton, 2020) allows querying data from NCBI databases via command line scripts, favoring in particular the exploration of metadata associated with records investigated, but it does not integrate scripts or libraries to promote data visualization and exploration, neither integrates NCBI taxonomy database data (Federhen et al., 2012). On the other hand,

TaxonTableTools (Macher et al., 2021) integrates workflows to analyse data produced by the user, focusing on metabarcoding common approaches. ExTaxsI, instead, implements NCBI data retrieval, in order to create formatted databases for taxonomy assignment and explore the results from a taxonomic point of view. In addition, the command line script is user-friendly, as it is built to make the tool interactive, using questions and explanations to help users use.

ExTaxsI, in detail, is linked to NCBI taxonomy database (Federhen et al., 2012) and ETE toolkit (Huerta et a., 2016), in order to produce standard formats readable by most common software that deal with taxonomic information (Bolyen et al., 2019; Rognes et a., 2016; Bengtsson et al., 2015; Mahe et al., 2015; Camacho et al., 2009; Wang et al., 2007), such as QIIME2 platform (Bolyen et al., 2019). The tool is applicable to any molecular marker, gene name or taxonomic group data, where it is also possible to create non-standard marker genes database usable in metagenomic/metabarcoding taxonomic assignment tools (Bolyen et al., 2019). In addition, thanks to the integration of the NCBI query tool (NCBI, 2014), ExTaxsI can reorganize personal datasets in a standardized format in order to easily describe taxonomic variability and geographic provenance of records.

4.2 ExTaxsI at work

ExTaxsI is a bioinformatic open-source tool aimed to elaborate and visualize molecular and taxonomic information via a simple interface. It is developed in Python 3.7 both as a command line and as a python library. The command line scripts are available through a user-friendly console, as they are built to make the tool interactive, helping the users via questions and explanations. Instead, the Python module was built for IT advanced users to facilitate its integration into specific analytical pipelines (e.g., genomics, metagenomics).

As illustrated in Fig. 1, this open-source instrument starting from a list of taxa or gene name/s, allows to i) search for taxonomic, genetic and biogeographical data through NCBI databases, ii) create a local and formatted nucleotide sequences (FASTA format) dataset and iii) their related taxonomy classification paths/datasets, thanks to the integration of NCBI taxonomy data, iv) generate genetic markers lists coming from different studies, and finally v) produce interactive plots starting from NCBI query search results or directly from offline taxonomic files, including representative graphs for the exploration of taxonomy and refinement of biogeographical data by creating geographical maps with the locations of the species analyzed (Figure 1).

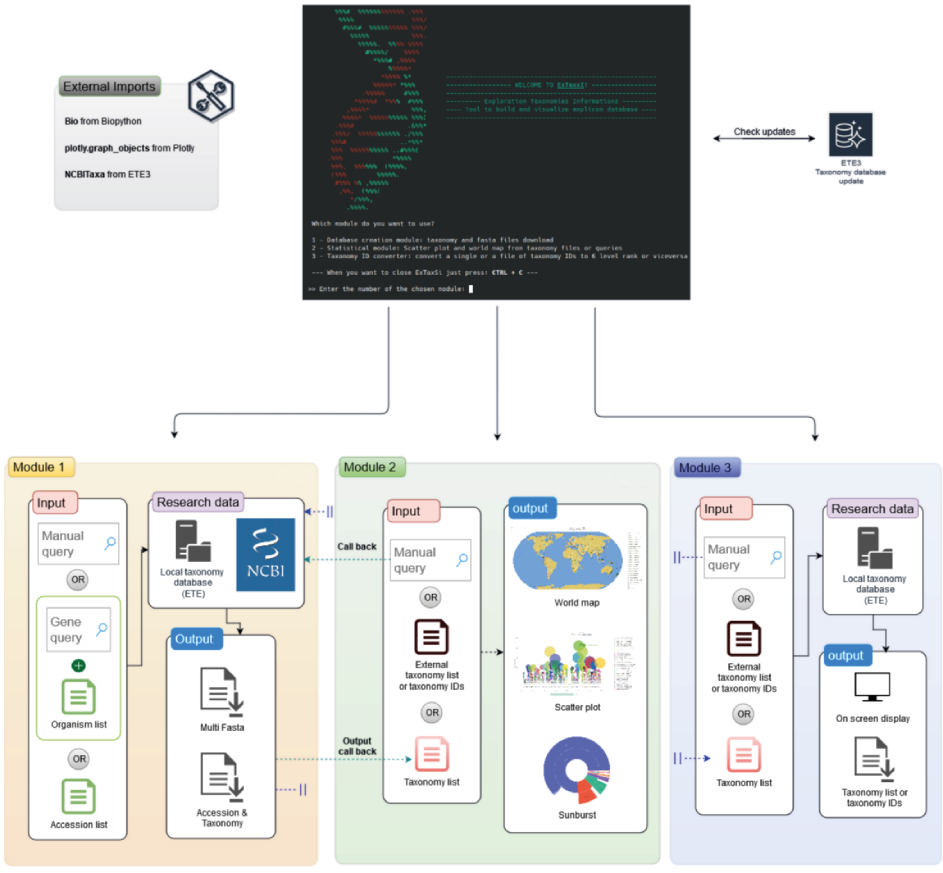

Figure 1. ExTaxsI pipeline: module 1 (orange) searches and creates files and databases; module 2 (green) processes georeferenced or taxonomic data for the creation of graphs and plots; module 3 (blue) converts taxonomic data into NCBI taxonomy ID (TaxID) and vice versa.

It is important to note that ExTaxsI outputs are compatible with other tools for taxonomic assignment purposes (Rognes et al., 2016; Bengtsson et al., 2015; Mahe et al., 2015; Camacho et al., 2009; Wang et al., 2007), such as the QIIME2 platform (Bolyen et al., 2019).

The communication with NCBI server is mediated by the Entrez module (NCBI, 2014), implemented in Biopython library (Cock et al., 2009), which allows to search, download and parse query results. To help NCBI interaction, when the requests are less than 2500, the search key is composed by a single query, otherwise the query will be splitted in groups of 2500 generating temporary files, which are then merged into a single output file at the end of the process.

Regarding taxonomy handling, the ETE toolkit was exploited (Huerta et al., 2016). In particular, ETE allows to create and maintain a local taxonomy database up to date by extrapolating the 6 main ranks (phylum, class, order, family, genus, and species).
If the organism is poorly described or it is an unknown species, the NCBI taxonomy ID (i.e. TaxID) of its ancestor (known as parent TaxID) in ETE taxonomic tree is then used and converted into its scientific correspondent name. It is important to underline that all queries are carried out locally, avoiding unnecessary online response delays.
Finally, the extracted data are visualized through scatter plot and interactive sunburst chart for the taxonomy exploration, and world map plot for the geographic metadata plotting.

4.3 Use cases

Being a taxonomy focused data exploration tool, we designed three possible scenarios of increasing complexity, to challenge it with increasing taxonomic variability and dimension of accession entries.

The first scenario hypothesizes a query to explore data with i) low taxonomic variability and a high number of expected entries (1 species, more than 300,000 entries). The second scenario provides ii) a high taxonomic variability and a large expected number of entries (about 500 species, more than 300,000 entries). The third and more complex scenario explores a iii) complete case study with taxonomic input intersected by molecular data.

Considering the case studies of the first two scenarios, we focused on taxa of interest in marine fisheries: 1) the cod fish species *Gadus morhua*, for which a worldwide economic interest exists, and 2) its taxonomic group at order level the Gadiformes order which supports long-standing commercial fisheries and aquaculture. These two case studies evaluate the capacity to explore data and to fill the geographic distribution of a species, prospecting also the available genes information to perform a genetic survey (e.g. in a potential DNA metabarcoding study).

With the third use case, we aimed at demonstrating the flexibility of ExTaxsI in different contexts: a genetic exploration of the available data in NCBI associated to SARS-CoV-2 virus a very recent topic that involved many research groups, leading to huge amounts of data collected and deposited in public sources (Blomberg et al., 2020). A large-scale exploration of data related to this topic can potentially improve the reliability of results and can provide valuable evidence to inform decisions on public health protection, both now and most importantly in the future.

4.3.1 Insights into two taxonomic groups of commercial interest

The first scenario is the case of *Gadus morhua* species (Gadidae; Gadiformes), also called Atlantic cod. In detail, *Gadus morhua* is a large, cold-adapted teleost fish that supports long-standing commercial fisheries and aquaculture (Jorde et al., 2018; Knudsen et al., 2019; Star et al., 2011;Kurlansky et al., 2006; Johansen et a., 2009).

ExTaxsI retrieved a total of 367,455 accessions (18 of June, 2021) using the Taxonomy ID through the following query: “txid8049[ORGN]” (where 8049 is the *Gadus morhua* TaxID). Only 54,061 entries showed a ‘gene’ tag investigable by ExTaxsI. As a unique species, we decided to represent the results obtained from a gene survey (Figure 2) and the world map plot (Figure 3).

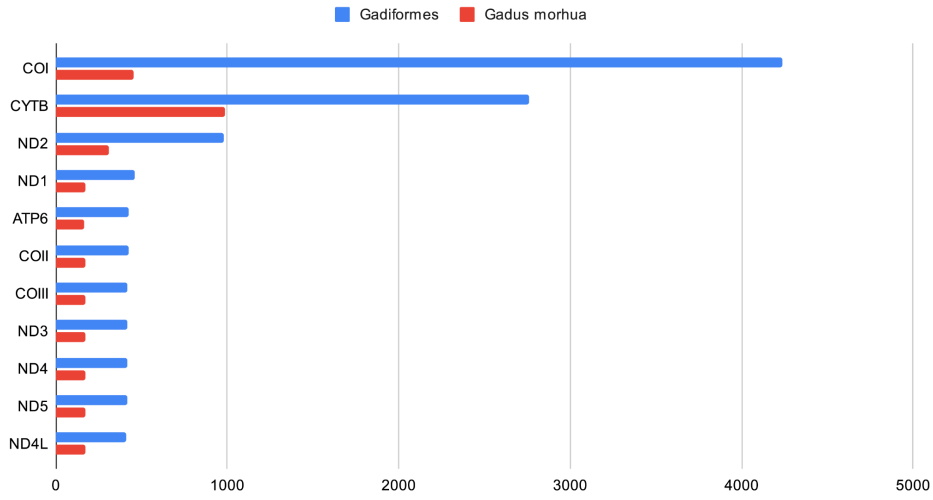

Figure 2. Gene distribution of accessions with complete ‘gene’ tag among Gadus morhua and Gadiformes taxon.

Regarding gene distribution, the most abundant gene is CYTB cytochrome b (with 985 accessions), followed by COI cytochrome c oxidase subunit I (455) and ND2 (311). These results are in line with those obtained by Knudsen and

colleagues (2019), where they personally developed specific primers for CYTB amplification, as it is a widely used marker in fish molecular characterization. The remaining most abundant genes are the other ND portions and Cytochrome Oxidase fragments (COIII and COII), belonging to the mitochondrial genome. These results show the increased effort in sequencing "standard" barcoding markers, while moderately sequencing larger portions of mitochondrial genomes. The remaining genes in the retrieved list and their relative accession frequency distribution (see the complete list in Additional file 1) demonstrate that many entries of the genome were investigated.

Regarding the geographic area, the Gadidae family has a circumpolar distribution, comprising species occurring principally in northern and cool seas (Jorde et al., 2018). Further, as reported by Jorde and colleagues (2018), in Norway we can recognize four distinct stocks of the Atlantic cod: (1) the oceanic Northeast Arctic cod, (2) coastal cod north of 62°N, (3) coastal cod south of 62°N, and (4) a North Sea/Skagerrak stock, the most densely populated region in Norway (Jorde et al., 2018). This geographic distribution is partly visible via the metadata extracted by ExTaxsI, as shown in the world map plot in Figure 3b (Additional file 2).

Via ExTaxsI, this Order was explored using the following query “txid8043[ORGN]”, yielding 389,640 accessions (where 8043 is the specific Gadiformes TaxID; 21 of June, 2021), where 61,249 showed the ‘gene’ tag. As a group spread on different taxonomic levels, both taxonomy and gene lists were created. In detail, in order to explore taxa distribution and accessions abundances across the entire order, the tool created scatter plot and sunburst plot HTML outputs. In Figure 3a genera across families are documented in scatter plot modality, while sunburst plot and entirely interactive plots are available in the Supplementary Material section (Additional files 3 and 4).

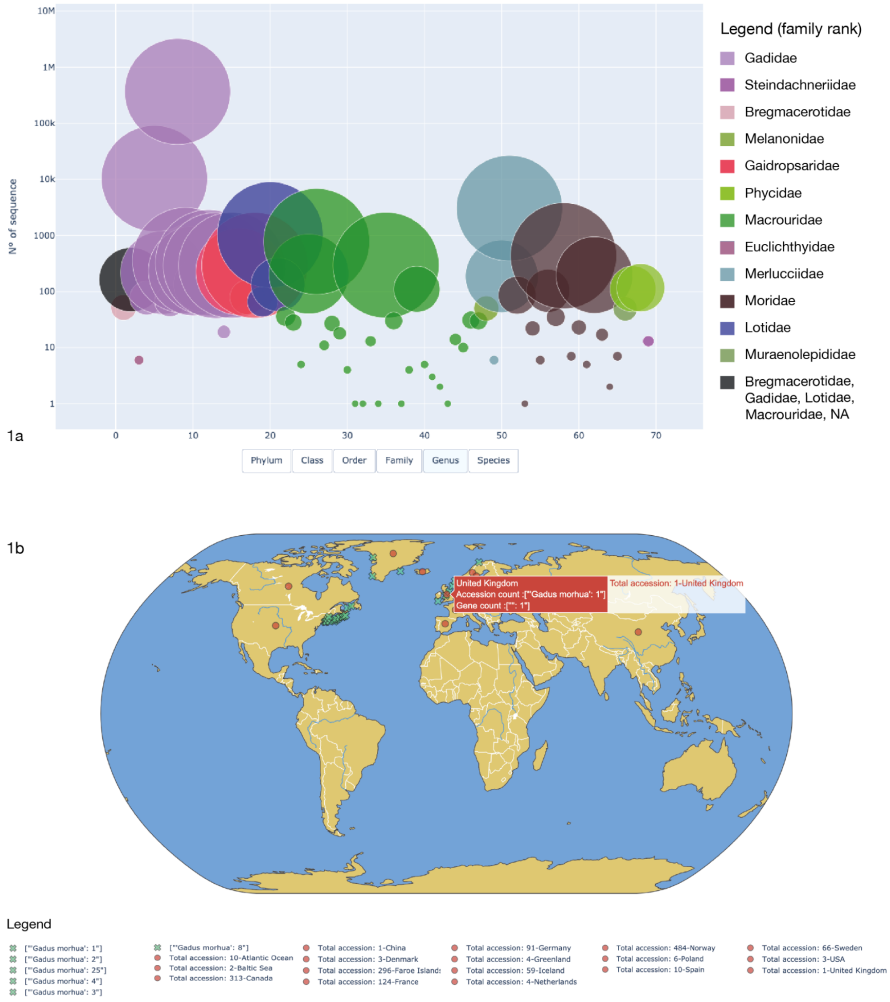

Figure 3. 3a) Scatter plot of Gadiformes accessions to represent sequence abundances among families; 3b) World map plot of Gadus morhua distribution considering geographic metadata extracted from records.

As shown in Figure 3a, Gadidae is the most abundant family, considering the number of accessions available. In detail, a total of 381,460 accessions populate this group, followed by Merlucciidae (3,252) and Macrouridae (1,673) families. These results are in accordance with the literature, as Gadidae family

is a primary marine, bottom-dwelling family of fishes in the Order of Gadiformes with great commercial power (Nelson et al., 2016; Knudsen et al., 2019).

Further, considering the scatter plot in Additional file 3, the interactive visualization allowed to visualize the taxonomy distribution among the accessions available, changing dynamically the rank to explore. This feature permitted to disclose that the genus Gadus is the most abundant of the entire dataset, highlighting that *Gadus morhua* species corresponded to 94.3% of the accessions in the entire dataset. This is an expected result, as *Gadus morhua* is documented to be a key species both in the North Atlantic ecosystem and commercial fisheries, with an increasing aquaculture production in several countries (Jorde et a., 2018).

Considering the genetic information reached by ExTaxsI, a total of 28,850 unique genes were found from the 61,249 completely tagged accessions. A classification of the most ten abundant genes is reported in Figure 2. As shown in the figure, at the first position we can find the COI gene, a widely used marker gene in metabarcoding projects (Knudsen et al., 2019), dealing mainly with animals detection (Porter et al., 2018), followed by CYTB and ND2 (Porter et al., 2018).

Concluding with these two case studies, the tool was able to accurately portrait the state of the art of the genetic information available in NCBI. Comparing the most abundant genes found among the records, it is possible to see a thin discrepancy between the two taxa explored (Figure 2), highlighting the disclosures that the survey can report. In general, the detection of mitochondrial genes, coding for COI and CYTB, is in accordance with the reliability of these DNA barcodes, principally used in the discrimination of animal species (Hebert et a., 2003; Hellber et al., 2014; Mueller et al., 2015). To date, considering the subjects of our case studies, diverse studies have used COI or CYTB barcoding to identify seafood products and explore broad

patterns in fish mislabelling (Fernandes et al., 2017; Cline et al., 2012; Di et al, 2013; Miller et al., 2010; Rasmussen et al., 2008; Yancy et al., 2008).

Regarding the extraction of geographic metadata from NCBI records, the completeness and collection of data can improve drastically the biogeographic and ecological research, allowing not only to explore sampling areas, but also to improve phylogeography investigations, biodiversity monitoring and environmental genomics strategies (Porter et al., 2018; Cordier et al., 2020).

The unbalance between the number of records and the number of genes explorable is in some cases due to the incompleteness of the 'gene' tag. In the very recent years genome sequences started to play a key role in public repositories, making sequences available for sharing and reuse. Submission process can be challenging and errors can affect the availability of the data. For this reason, there is a wide interest to integrate standardized procedures into the annotation process (Geib et al., 2018). The promotion of FAIR principles and best practices can certainly avoid the error propagation in sequence databases (Wilkinson et al., 2016; Pirovano et al., 2017), making the data fully explorable in the future.

4.3.2 Explore biodiversity data in pandemic outbreak: the case of SARS-CoV-2

The severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) is an enveloped, positive-sense, single-stranded RNA virus that causes coronavirus disease 2019 (COVID-19).

RNA and structural proteins are included into virus particles mediating host cell invasion. After cell infection, RNA encodes structural proteins that make up virus particles. Virus assembly, transcription, replication and host control are mediated by nonstructural proteins (Lu et al., 2020).

The pandemic linked to SARS-CoV-2 highlighted hidden virus reservoirs in wild animals and their potential to occasionally spillover into human populations (Lu

et al., 2020). A detailed understanding of this process is crucial to prevent future spillover events. As reported in the seminal paper of Andersen and colleagues (2020) (Andersen et al., 2020), the risk of future re-emergence events increases if SARS-CoV-2 pre-adapted in another animal species. SARS-CoV-2 probably originated from *Rhinolophus affinis* bats, with pangolin (*Manis javanica*) as intermediate host (Andersen et al., 2020). Recently, other animal species were supposed to be possible intermediate hosts in between bats and humans (Liu et al., 2020; Zhou and Shi, 2021).

To date, ACE2 (Angiotensin-converting enzyme 2), the receptor which binds to the receptor-binding domain (RBD) of SARS-CoV-2 S protein (Letko et al., 2020), is reported as crucial in host invasion.

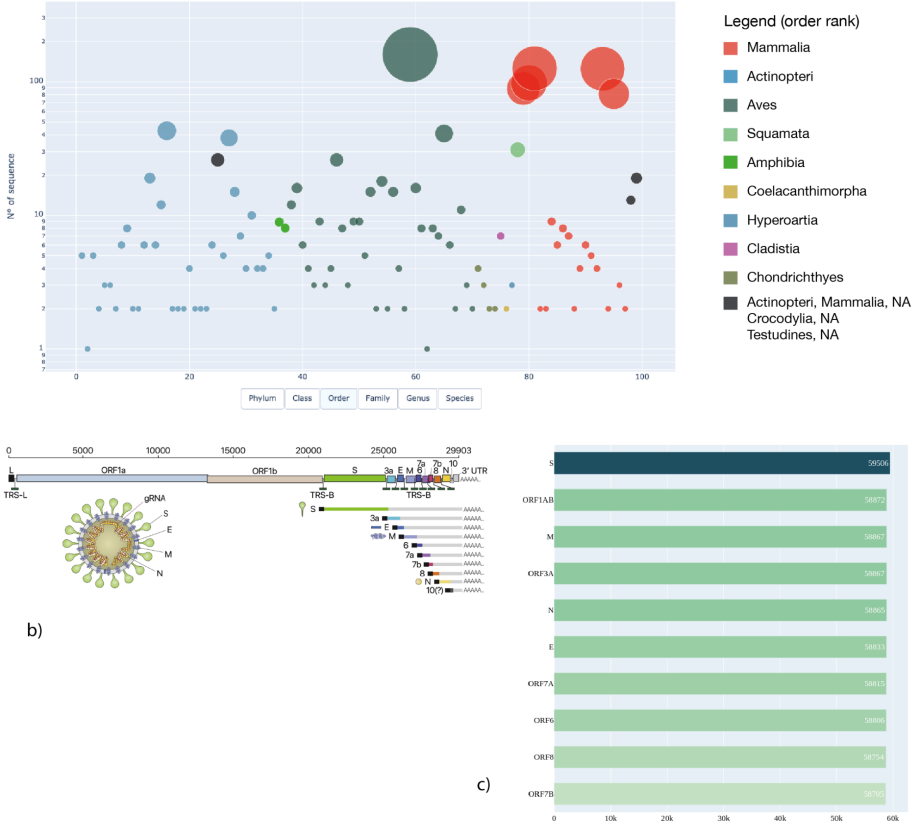

Figure 4. 4a) Scatter plot of ACE2 accessions representing sequence abundances among taxa at order level; 4b) SARS-CoV-2 representation, from Kim et al., 2020; 4c) gene distribution across accessions of SARS-CoV-2 data.

To test our approach and explore the genetic information available in NCBI, we decided to extrapolate information of the ACE2 gene from the Vertebrata taxonomic group, with the following query: "txid7742[ORGN] AND ACE2[gene]" (where 7742 is the specific Vertebrata TaxID). The results show that the ACE2 gene is widely distributed throughout Vertebrata: we obtained a total of 1,391 accessions (20 of June, 2021), distributed mainly among the Mammalian Class, with a high representation in Actinopteri and Aves groups (Figure 4a; Additional files 5 and 6 for an interactive exploration). In detail, Chiroptera, Primates and Rodentia are the most represented, with 126, 125 and 81 accessions respectively. Supporting the exploitation of molecular data survey, Luan and colleague (2020) (Luan et al., 2020) analyzed the affinity to S protein of the 20 key amino acid residues in ACE2 from mammal, bird, turtle, and snake, and suggested that Bovidae and Cricetidae should be included in the screening of intermediate hosts for \textit{SARS-CoV-2}. In addition, thanks to the analysis of spike glycoprotein sequences from different animals, the study of Dabravolski and Kavalionak (2020) suggested that the human SARS-CoV-2 could also come from yak as an intermediate host.

ExTaxsI has the advantage to provide the complete list of taxa, allowing an exhaustive exploratory research. It allows downloading all the sequences available for the query input, generating in turn the input for downstream analyses, such as the calculation of sequence similarities among different taxa. Further, investigating shared features with other species can have important implications for understanding potential natural reservoirs, zoonotic transmission, and human-to-animal transmission. Noteworthy, the survey can give researchers an instrument to download specific data related to Covid-19 disease, with a user-friendly approach, to explore interactively the data,

including biodiversity related information, and to design informed scientific experiments.

Lastly, we explored the data available for SARS-CoV-2 (Figure 4) using the following query “txid2697049” (where 2697049 is the specific Severe Acute Respiratory Syndrome Coronavirus 2 TaxID). We obtained a total of 773,293 accessions (28 of June, 2021). The top ten genes retrieved are shown in Figure 4c.

In particular, the genes most represented genes are: S (59,506), the spike or surface glycoprotein fragment, ORF1AB (58,872), followed by M (58,867), ORF3A (58,867) and N fragments (58,865), the nucleocapsid protein. These results are in line with the recently published scientific data highlighting the functional aspects of viral proteins. Considering the ORF1AB, several studies demonstrated its pivotal role among coronaviruses (Wan et al., 2020), providing a clinical target to break down SARS-CoV-2 infection (Khailany et al., 2020).
Regarding the first and fifth results, the nucleocapsid phosphoprotein is involved in packaging the RNA into virus particles and protects the viral genome. For these reasons, it has been suggested as an antiviral drug target (Wu et al., 2020; Gordon et al., 2020). The spike glycoprotein, instead, is located outside the virus particle, mediating its attachment and promoting the entry into the host cell. It also gives viruses their crown-like appearance. In the very last research, the S protein was found as an important target for diagnostic antigen-based tests, antibody therapies and vaccine development (Salvatori et al., 2020; Pillay et al., 2020).
The entry of SARS-CoV-2 is mediated by further processes, for example the activity of the protease TMPRSS2 (Hoffmann et al, 2020). Also in this case, the use of ExTaxsI can unearth similar proteases in possible intermediate hosts, revealing new insights into the mechanism of infection.

As also documented in Khailany et al., 2020 (Khailany et al., 2020), the emergent and huge amounts of data collected in the last few months necessitates a large scale exploration of the data. The rapid increment of data releases may give some important insights about SARS-CoV-2 behaviour in its host species, helping in improving not only our knowledge, but also models to predict COVID-19 outbreaks and new drug targets.

4.4 Conclusions and future perspectives

ExTaxsI provides an easy-to-use standalone tool able to interact with NCBI databases and personal datasets, offering instruments to standardize taxonomy information and visualize vast amounts of data distributed on different taxonomic levels. It also provides interactive visualization plots, easily shareable through HTML formats.

The user-oriented interrogation of NCBI databases may help researchers involved in environmental genomics fields, from phylogeographic studies to DNA metabarcoding surveys, and also in projects related to human health, as demonstrated with the SARS-CoV-2 case study.

With this work, we hope to meet the needs of a broad group of researchers, providing an instrument easy to install either on common laptops or on high performance servers and directly connected with NCBI databases. In parallel to the command-line tool, a python library containing all ExTaxsI functions has been implemented, favoring a direct incorporation of such functions into data analysis and exploration pipelines.

In addition, as data volume is increasing over time and NCBI databases still have a few constraints regarding the queries results dimension and their retrieval time required, an automatic management of large queries will be

organized in future releases. Finally, we will also consider further data visualization strategies and additional metadata (e.g. GBIF country information) to enhance data interpretation and to provide comprehensive sets of relevant scientific-focused information.
In our opinion, ExTaxsI data management ability with its visual interactive exploration can really improve the experimental design phase and the awareness of the information available, facilitating data examination and sharing.

4.5 Implementation

ExTaxsI is a bioinformatic tool aimed to explore, elaborate and visualize molecular and taxonomic information via a simple user interface without specific bioinformatic or programming skills. The tool can be run, via command line interface, where the user is guided by the appropriate documentation of each script, avoiding the implementation of ad hoc python code.
ExTaxsI is developed in three separate modules, which can be used either interconnected as workflow or independent according to the user needs. The main modules are listed as follows: i) Database creation, ii) Visualization and iii) Taxonomy ID converter.

ExTaxsI is also available as a python library that can be installed through *pip* (package installer for Python), containing the same functions and parameters as those of the command-line tool. A detailed description of each module is provided below.

4.5.1 Database creation module

The module ‘Database’ allows the user to create multi FASTA files composed of nucleotide sequences, taxonomic lists, genes names and their related

accessions, starting from either a single or a batch query mode using csv/tsv files (Figure 1). After indicating the type of input, the tool asks, with the exception of the file accession, whether the user wants to integrate the query with one or more gene name/s (or other details). This step allows to restrict the research in NCBI if needed.

In general, the output formats are i) a multi-FASTA file (widely used format for molecular sequences) and ii) text file in TSV format, with two columns composed by the accessions code followed by the taxonomy path of each accession at the six main levels separated by semicolons: phylum, class, order, family, genus and species.

When requested by the user, the output file of genes names is provided in TSV format consisting of a table with two columns, the first is the list of genes and the other is the frequency values of the respective genes found in the retrieved records.

The tool also provides a summary table containing the most popular genes from a list of taxid, accessions or organisms. In addition, it is possible to create a barplot with the top ten of the summary table, downloadable as a PNG file.

4.5.2 Visualization module

The module 'Visualization' allows the user to create interactive plots, starting from the Database module output or from external sources (e.g., Additional files 3, 4 and 5) containing taxonomic lists. Before producing the plots, a dialogue box will ask the user to choose a filter value on the data based on the frequency. If the chosen filter value is 0, the tool processes all the data. Otherwise, all the taxonomic units that have not reached the minimum value are inserted into an additional text file, specifically created with a name containing the filter used.

The available plots generated by ExTaxsI are i) scatter plot (Additional file 3), ii) sunburst plot (Additional file 4) and iii) world map plot (Additional file 2). All

figures created by the Visualization module can be downloaded as HTML format files.

In detail, scatter plot uses taxonomy as input to produce a graph that indicates the quantity of each individual taxonomic unit; the interactive plot enables the user to i) choose the taxonomic level to be displayed using the buttons located under the graph and ii) hover over points to show details, such as the number of records within taxa, names of selected taxa and name of the higher taxon from which they derives. The plot allows also to compare more data on mouse-over, highlight an area of interest with zoom function and view of a specific group or remove specific taxa from the graph.

Sunburst plot, instead, from a taxonomy input creates an expansion pie that allows exploring taxonomy by clicking on the taxonomic group of interest and showing the underlying taxa within a new sunburst plot. Also in this case, hovering over points shows the number of records within taxa.

Regarding world map plot, the initial input is processed in order to obtain geographic data. The tool exploits the 'Country' metadata stored in the NCBI records to produce a map indicating the position of each entry. In this step, based on the type of geographic data obtained, ExTaxsI divides results into two different arrays: i) a specific array of coordinates (if the coordinates are present in the record) or ii) a specific array of country names (if the coordinates are not present in the record). It is also possible to add external sources data to the map. In each created map, the coordinates are indicated by green X signs, while countries by red circles. Thinking of multiple taxa plotting, each symbol can have a legend that summarizes the data downloaded with the same country name or coordinates description. Further, it is possible to see both genes and counts available among the represented accessions.

4.5.3 Taxonomy ID converter module

This module allows to convert TaxID to the main six ranks taxonomy and vice versa (phylum, class, order, family, genus and species); it can convert single manual inputs or multiple inputs from a tsv/csv file containing a list of Tax IDs.

4.6 Availability of source code and requirements

No specific system requirements are needed for the installation of ExTaxsI, however for the correct functioning of the software we suggest a minimum of 4GB of RAM.

To successfully run ExTaxsI, the following python libraries must be installed: Biopython (Cock et al., 2009), NumPy (Harris et al., 2020array), SciPy (Virtanen et al., 2020), Matplotlib (Hunter et al., 2007), ipython (Perez et al., 2007), Pandas (Mckinney et al., 2011), SymPy (https://www.sympy.org/en/index.html), nose (https://nose.readthedocs.io/en/latest/), genutils (https://pypi.org/project/genutils/), requests (Chandra et al., 2015) and Plotly (https://plotly.com/), in addition to Plotly-Orca and ETE toolkit (Huerta et al., 2016ete). To install all the dependencies compatible versions, we provide a requirement list at the GitHub page https://github.com/qLSLab/ExTaxsI, with a detailed guideline to set directly a conda environment.

The Python library extaxsi is available both in the Github page: (https://github.com/qLSLab/ExTaxsI/tree/master/library) and in PyPI repository (https://pypi.org/project/extaxsi/).

General information:

- Project name: ExTaxsI
- Project home page: https://github.com/qLSLab/extaxsi
- Operating system(s): Platform independent
- Programming language: Python
- License: GNU GPL version 3
- bio.tools ID (https://bio.tools/): extaxsi
- Research Resource Identification Initiative ID (RRID) (https://scicrunch.org/): SCR_021846

4.7 Availability of supporting data and materials

Supplementary data are available at Supporting data for "ExTaxsI: an exploration tool of biodiversity molecular data" GigaScience Database. http://dx.doi.org/10.5524/100959 (Agostinetto et al., 2021). In particular, supplementary files are the following:

- Additional file 1: Gene list in TSV format obtained through ExTaxsI for the species Gadus morhua. Gene counts were extracted by 367,455 accessions (query: “txid8049[ORGN]”; 18 of June, 2021).
- Additional file 2: World map plot in HTML format created via ExTaxsI extracting the values of 'Country' tag contained into 367,4553 accessions of Gadus morhua (query: “txid8049[ORGN]”; 18 of June, 2021). Coordinates are indicated by green X signs, while States by red circles.
- Additional file 3: Scatterplot in HTML format created via ExTaxsI extracting the taxonomy of 389,640 accessions of Gadiformes Order (txid8043[ORGN]”; 21 of June, 2021).
- Additional file 4: Sunburst plot in HTML format created via ExTaxsI extracting the taxonomy of 388,603 accessions of Gadiformes Order (txid8043[ORGN]”; 21 of June, 2021).
- Additional file 5: Scatterplot in HTML format created via ExTaxsI extracting the taxonomy related to 1,391 accessions of ACE2 genes belonging to the Vertebrata taxonomic group (query: “txid7742[ORGN] AND ACE2[gene]”; 20 of June, 2021).
- Additional file 6: Sunburst plot in HTML format created via ExTaxsI extracting the taxonomy related to 1,391 accessions of ACE2 genes belonging to the Vertebrata taxonomic group (query: “txid7742[ORGN] AND ACE2[gene]”; 20 of June, 2021).

4.8 List of abbreviations

SILVA: High quality ribosomal RNA databases;

BOLD: Barcode of Life Data System;

UNITE: Database and sequence management environment centered on the eukaryotic nuclear ribosomal ITS region;

ETE: Environment for Tree Exploration;

QIIME2: Quantitative Insights Into Microbial Ecology;

FASTA: Text-based format for representing either nucleotide sequences or peptide sequences;

TAXID: Taxonomy ID;

HTML: Hyper-Text Markup Language;

COI: Cytochrome Oxidase I;

COI: Cytochrome Oxidase II;

COIII: Cytochrome Oxidase III;

CYTB: Cytochrome B;

ND2: NADH dehydrogenase 2;

ACE2: Angiotensin-Converting enzyme 2;

RBD: Receptor-Binding Domain;

PNG: Portable Network Graphics;

NCBI: National Center for Biotechnology Information;

ENA: European Nucleotide Archive

5. Extending association rule mining to microbiome pattern analysis: tools and guidelines to support real applications

5.1 Introduction

Studying microbiome patterns is now a hot-topic in different fields of application (Kyrpides et al., 2016; Wood-Charlson et al., 2020). From ecology to medicine, microbiomes are undoubtedly a cornerstone of research, acknowledged as being key participants in all ecosystems, including the human one (Layeghifard et al., 2017). In recent years, DNA sequencing strategies have become one of the main sources for studying microbial communities (Wood-Charlson, 2020). Further, 16S rRNA metabarcoding is currently the preferential method to obtain great amounts of information in a time and cost effective manner (Wood-Charlson, 2020), becoming one of the primary sources of data regarding microbiome studies (Bokulich, 2020; Knight, 2018; Gonzales et al., 2018; Mitchell et al., 2020).
In this context, data mining approaches seem to be newfangled solutions for disclosuring and understanding microbial ecosystems (Galimberti et al., 2021; Wood-Charlson, 2020; Ghannam et al., 2021). Spanning from classification and signature extraction to interaction and trait associations (Pasolli et al., 2016; Qu et al., 2019), data mining strategies can identify hidden patterns that may help to predict biological functions (Noor, 2019; Thomposon, 2019). Investigating patterns and exploring their role in functional and predictive aspects are now pivotal to proxy the knowledge of microbial associations, both disentangling interactions and niche specialization (Faust et al., 2012; Ma et al., 2020).

Considering the size and complexity of High-Throughput Sequencing (HTS) 16S rRNA metabarcoding data, interpretation and summarization are not straightforward (Naulaerts et al., 2015) and, for this reason, pattern mining strategies have become essential for researchers to disentangle the high amount of information (Ghannam et al., 2021; Wood-Charlson et al., 2020;

Kyrpides et al., 2016).

Recently, association rule mining (ARM) emerged as a promising technique to study microbiome patterns (Tandon et al., 2016; Naulaerts et al., 2015). Specifically, Tandon and colleagues (2015) have demonstrated the potentials of this technique on two microbiome datasets, in particular the HMP dataset (Turnbaugh et al., 2007) and two prebiotic studies (Xiao et al., 2014; Kato et al., 2014).

From the classic application on market basket problems (Agrawal et al., 1993), association rule mining started to be applied to answer a wide range of biological questions. From annotation tasks (Manda et al., 2020; Manda et al., 2013; Manda et al., 2012) to protein interaction networks (Koyuturk et al., 2006), ARM was applied to a wide range of research fields, including genetics (Ong, 2020; Karpinets et al., 2012; Alves, 2010; Carmona-Saez et al., 2007), molecular biology (Boutorh et al., 2016; Agapito et al., 2015; Naulaerts et al., 2016), and biochemical disciplines (Zhou et al., 2013; Naulaerts et al., 2016). Noticeably, the expression ‘association rule mining' comprehends two main phases: i) frequent itemset mining, the extraction of patterns intended as elements often co-occur together in a dataset (Agrawal et al., 1993), and ii) rule calculation, to identify strong association between patterns previously extracted (Agrawal et al., 1993).

Despite the apparent simplicity of use, large datasets can produce high numbers of patterns, making their extraction difficult (Karpinets et al., 2012; Naulaerts et al. 2015; Agrawal et al., 1993; Han et al., 2004). Beside several algorithms have been developed to better capture reliable patterns, as for example Eclat (Agrawal et al., 1996), FP-Growth (Han et al., 2004) or Apriori (Agrawal et al., 1993), avoiding uninformative or spurious information is still a current issue (Naulaerts, 2015). Interesting measures such as support (frequency of a pattern) or pattern length are pivotal to control the generation and the evaluation of patterns discovered (Karpinets et al., 2012; Naulaerts et al. 2015; Agrawal et al., 1993). Still, a few issues exist in setting these parameters (Naulaerts et al., 2015). Considering the support, setting a low

value leads to a high amount of patterns, difficult to explore and visualize. At the same time, setting a high support value can be detrimental for finding rare but informative patterns. Over and above, researchers try to identify metrics that can be used to pinpoint patterns of interest (and so called “interest measures”). In detail, several metrics have been implemented (Omiecinski et al., 2003; Franceschini, 2012; Tan, 2002; Tang et al., 2012), as for example lift or maximal entropy (Hussein et al., 2015; Tatti et al., 2010). Nevertheless, extracting effective information is not an easy task as the definition of interestingness is strictly associated with the biological question and the research field under study (Karpinets et al., 2012; Naulaerts et al., 2015; Koyutürk et al., 2006). Considering the rule calculation phase, issues regarding the evaluation of reliable rules remain (Karpinets et al., 2012; Naulaerts et al., 2015). In general, taking into account previous works, the most widely used parameters to evaluate both patterns and rules are support and confidence, where confidence is a measure that describes the strength of the association between the two elements of the rule (Nauleaerts et al., 2015).

Recently, different works related to pattern mining applied to microbiome studies were published, such as MITRE (Bogart et al., 2019), MANIEA framework (Liu et al., 2019) and the work of Tandon and colleagues (2016). Nevertheless, as also highlighted by the work of Faust (2021), applying such an algorithm still has its limitations and, despite the efforts of recent works, guidelines for microbiome data applications have not been completely defined (Faust et al., 2021; Naulaerts et al., 2015). Different libraries have been implemented, such as pyfim (Muino and Borgelt et al., 2014), mlxtend (Raschka, 2018) and arules (Hahsler et al., 2011). A few frameworks have been recently developed and applied on real case studies (Liu et al., 2019; Tandon et al., 2016). However, tests to establish specific best practices for 16S rRNA metabarcoding data do not exist.

Apart from the availability of tools, the application of pattern mining to study microbiome patterns must consider the intrinsic biological aspect of microbiome data (Gloor et al., 2017; Balint et al., 2016). Beside the issues

related to species abundances that should be filtered to obtain a solid input dataset, also metadata composition and taxonomy level should be considered. Further, microbiome matrices can be large and complex: composed of thousands of taxa and hundreds of samples (Faust et al., 2021; Ghannam et al., 2021), microbiome data can affect pattern mining approaches, sometimes obliging to set high but improper interest measures. This last point is crucial if we consider that 16S rRNA metabarcoding data can describe putative ecological properties and sparse microbial associations (Faust et al., 2021).

Given these premises, our work wants to shed light on the strengths and weaknesses of pattern mining strategy into the study of microbial patterns, in particular from 16S rRNA microbiome datasets. In detail, we show pitfalls of ARM applied on real case studies, highlighting issues related to the type of input and the use of metadata. Then, we identify the key steps that must be considered to apply ARM consciously on 16S rRNA microbiome data. Moreover, to facilitate the integration of ARM technique into microbiome pipeline, we developed microFIM (microbial Frequent Itemset Mining), a versatile user-friendly and open source Python tool that promotes the use of ARM integrating common microbiome practices, such as taxa tables and distance matrix visualizations. Besides the conventional parameters, microFIM implements interest measures to remove spurious information. Moreover, it merges the results of ARM analysis with the typical microbiome outputs, aiming at creating a bridge between microbial ecology research and ARM technique.

5.2 Materials and methods

This section comprehends two main paragraphs: i) description of microFIM (microbial Frequent Itemset Mining) tool to promote microbiome pattern exploration with two simulated dataset and ii) microFIM analysis on real case microbiome datasets to highlight ARM potentials and caveats. microFIM was developed on the basis of Frequent Itemset Mining (Naulaerts et al.. 2015), in

which patterns of elements that co-occur can be extracted from a transactional dataset, typically (Naulaerts et al.. 2015). A pattern (or itemset) is called frequent if its support value within the dataset is greater than a given minimal support threshold. For an overview of the method and its translation in terms of bacterial composition instead of elements, please see Fig. 1. A complete description of the approach with formalized expression can be found in the works of Naulaerts et al.. 2015, Goethals, 2005 and Tan et al., 2016 (Chapter 6).

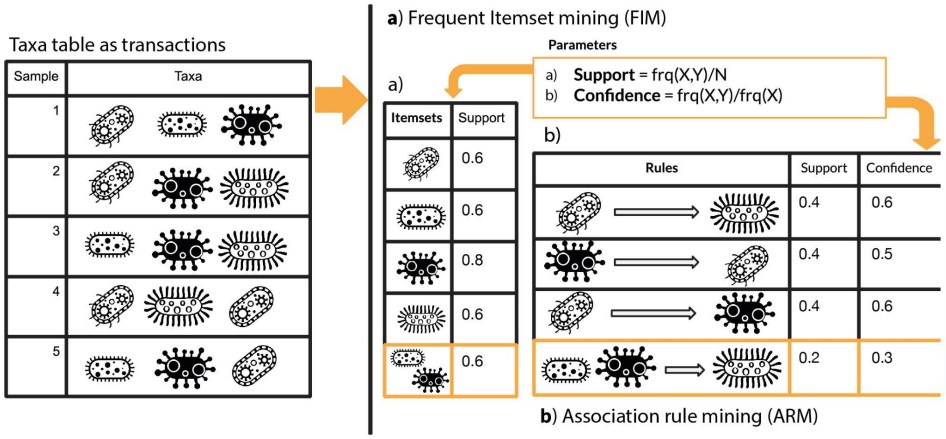

Figure 1. Graphical overview of Frequent Itemset Mining (A) and Association Rule mining (B) approach integrated with elements related to microbiome analysis.

5.2.1 microFIM implementation

To promote and integrate the use of ARM in microbiome studies, we developed microFIM (microbial Frequent Itemset Mining), a versatile open-source user-friendly tool implemented in Python (v. > 3; https://github.com/qLSLab/microFim).

microFIM receives as input the taxa table and the metadata file used during the microbiome bioinformatic analysis. In particular, a taxa table is composed of rows and columns representing the taxa and their abundances for each sample. It derives from the conversion of the BIOM file into a CSV or TSV file

(https://biom-format.org/). In general, considering the well-established QIIME2 microbiome platform (https://qiime2.org/; Bolyen et al., 2019), complete frameworks and scripts to analyse and obtain taxa tables are implemented.

To promote the usage to a wider group of researchers, the tool can be used both via Python functions and running the pre-settled scripts, which allow interactivity through the command-line, avoiding coding implementations. To favour easy integration in Python scripting and future implementation of additional functions and metrics, Python functions were divided into thematic sections.

microFIM is composed by 6 main steps: i) filtering taxa table with metadata, ii) converting taxa table into a transactional database to be read by ARM algorithms, iii) extract microbiome patterns, iv) calculate additional interest measures to evaluate the patterns extracted, v) create the pattern table (a taxa table improved with patterns, presence-absence information among samples and interest measures) and vi) visualization of results.

Template files are provided to run microFIM scripts. Considering interest measures, we integrated support, pattern length and all-confidence metrics, which generates 'hyperclique patterns' (Agrawal et al., 1993; Xiong et al., 2006; Tan et al., 2016; Omiecinski et al., 2003). Considering a pattern 'X' composed of different items, all-confidence is calculated as the ratio between the support of 'X' and the highest support retrieved from the elements of the pattern 'X'. For example, a pattern X is composed of 3 elements that, considering the entire dataset, have the following support threshold: 0.3, 0.6 and 0.8. Overall, the pattern X has a support of 0.3. All-confidence will be calculated as the ratio between the support of X - 0.3 - and the higher support within X - 0.8, resulting in 0.37. All-confidence, in this way, is defined as the smallest confidence of all rules which can be produced from a pattern, i.e., all rules produced from a pattern will have a confidence greater or equal to its all-confidence value (Tan et al., 2016; Omiecinski et al., 2003). In detail, confidence is an indication of how often a rule has been found to be true, so it is considered as a measure of rule reliability (Hashler, 2005; Hashler, 2011; Naulaerts, 2015).

In order to show the usage and the potentials of microFIM, we tested the tool on simulated matrices (available in Supplementary Table 1 and 2) and on real case studies. In particular, the cases selected are: i) the ECAM dataset, (Bokulich et al., 2016), ii) the vaginal microbiome dataset of Ravel et al. (2011) and iii) the Montassier dataset (Montassier et al., 2016). Details about the application of microFIM on real case studies are described in the next sections. Parameters used to run microFIM on simulated matrices are the following: 0.3 as minimum support threshold, a minimum of 2 elements and a maximum of 10 to create patterns.

In the Results section, a complete scheme of the tool is provided. microFIM is mainly based on four Python libraries: fim (Muino and Borgelt et al., 2014), Pandas (McKinney et al., 2010), Numpy (Harris et al., 2020), and plotly (https://plotly.com/). It is available as a conda environment (https://docs.anaconda.com/) and all the details about tutorials and installation are available in our Github repository (https://github.com/qLSLab/microFim). Python notebooks and an example of microFIM usage via scripting are also reported in the repository. In general, beside the focus of this work, microFIM may potentially be used for a wide range of applications. As the primary resource input consists in a matrix describing the presence-absence of an element (rows) in a dataset (columns, representing samples), fields of study in which it can be applied may be various, also merely consider the analysis of OTU (Operational Taxonomic Unit) or ESV (Exact Sequence Variants) instead of taxa (Schloss and Westcott, 2011; Callahan et al., 2017) of 16S rRNA metabarcoding data.

5.2.2 Real case studies analysis

To show the caveats and potentials of association rule mining, we used microFIM on three real case studies the ECAM dataset (Early Childhood Antibiotics and the Microbiome; Bokulich et al., 2016), the vaginal microbiome case study of Ravel et al. (2011) and Montassier case study (Montassier et al., 2016). Different input types were selected based on taxonomy level and

metadata composition. In detail, the ECAM dataset collects a total of 875 samples, describing the gut microbiome of the first 2 years of life of 43 infants. Presence-absence tables were created taking account of the taxonomic rank. In particular, we used: i) the taxa table obtained directly from QIIME2 datasets (Bolyen et al., 2019) in which only taxa assigned to genus level, with a relative abundance > 0.1 % in more than 15% of samples, are considered (Input 1 - data are available in Supplementary Table 3); ii) family table obtained from collapsing the previous Input 1 via QIIME2 plugins (https://github.com/qiime2/q2-taxa; Input 2 - Supplementary Table 4); iii) a taxa table consisting only of taxa with complete taxonomy at the genus level (Input 3 - Supplementary Table 5). Metadata as type of delivery and antibiotic exposition were considered to evaluate patterns extraction.
Considering the vaginal microbiome dataset (Ravel et al., 2011), we obtained from MLRepo repository (Vangay et al., 2019) the taxa table obtained via the MLRepo pipeline (Vangay et al., 2019). The dataset collects 388 samples, investigating the vaginal microbiome of 396 asymptomatic North American women. Additional presence-absence tables were created taking account of the taxonomic rank, in particular from the original dataset obtained from MLRepo, also family and genus levels were considered. Low and high nugent score values (a scoring system for vaginal swabs to diagnose bacterial vaginosis) were considered for the evaluation regarding metadata filtering.

Finally, the dataset of Montassier et al. (2016) was included. The dataset collects 28 samples from patients with non-Hodgkin lymphoma undergoing allogeneic hematopoietic stem cell transplantation (HSCT) in order to identify microbes that predict the risk of BSI (bloodstream infection). OTU table and taxa table obtained with MLRepo pipeline were selected (Vangay et al., 2019).
For the ECAM and Ravel et al. (2011) datasets, minimum support threshold of 0.2, minimum length of 3 and a maximum length of 15 elements were used. Montassier et al. (2016) datasets were analysed considering a minimum support of 0.9, a minimum length of 5 and a maximum length of 10. After

pattern extraction, interest measures as support, pattern length and all-confidence were calculated (Xiong et al., 2006; Tan et al., 2016; Omiecinski et al., 2003). Distributions of number of patterns, length and support were evaluated considering both ARM analysis and interest measures filtering. A minimum of 0.5 and 0.8 of all-confidence were used to evaluate hypercliques patterns (Xiong et al., 2006; Tan et al., 2016; Omiecinski et al., 2003). Considering metadata filtering, pattern extraction was performed with the previous settings. A minimum of 0.8 of all-confidence was used to evaluate hypercliques patterns (Xiong et al., 2006; Tan et al., 2016; Omiecinski et al., 2003).

Visualizations were created with plotly and pandas Python libraries. Both datasets, results and metadata files are available in Supplementary Materials.

5.3 Results

5.3.1 microFIM tool: extending ARM to microbiome pattern analysis

Association rule mining demonstrates its useful properties in different contexts (Tandon et al., 2016; Naulaerts et al., 2015). To promote the use of ARM in the microbial community field, we implemented microFIM, a versatile open-source project developed in Python and freely available at https://github.com/qLSLab/microFim.

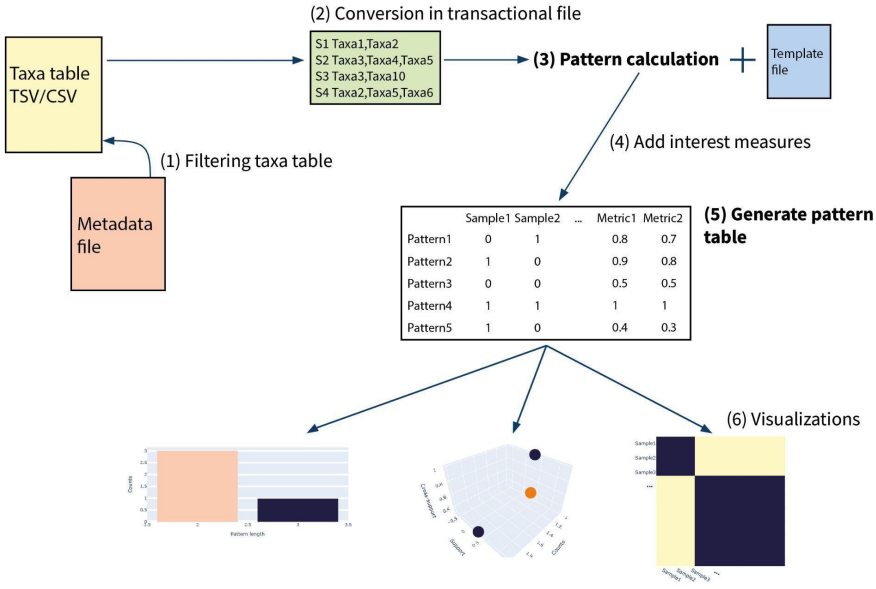

Figure 2. Scheme of microFIM framework. 1) Filtering taxa table; 2) Conversion of taxa table into transactional file; 3) Calculate patterns with template file filled with minimum support threshold, minimum and maximum length; 4) Adding of interest measures as support, pattern length and all-confidence (Xiong et al., 2006; Omiecinski et al., 2003); 5) Generating pattern table, composed by presence-absence of patterns within samples and interest measures; 6) Generating visualizations.

In this section, we explain the framework of usage, the main steps of pattern extraction and filtering and insights of visualizations available. In addition, two main examples are reported, in order to show the workflow of the tool. In Fig. 2 a scheme of microFIM framework is reported. In particular, microbiome data (taxa table) can be filtered (step 1) and then converted into a transactional dataset (step 2), in order to be read as input by association rule mining algorithm. Subsequently, patterns can be generated setting parameters via a template file to be filled (tutorials and templates are available at https://github.com/qLSLab/microFim) (step 3). In detail, minimum support

threshold, minimum and maximum length of patterns must be specified. Pattern extraction was implemented via pyfim library (Muino and Borgelt et al., 2014). At this stage, the default algorithm used is Eclat (Muino and Borgelt et al., 2014), but other algorithms are available within the pyfim library (Apriori or FP-Growth; Muino and Borgelt et al., 2014). The set of interest measures initially calculated are 'support' and 'pattern length' (which describes the number of elements belonging to a pattern). Further, other interest measures are added (step 4) and can be used to filter patterns. In microFIM implementation, all-confidence interest measure was included, in order to help remove spurious information (Xiong et al., 2006; Omiecinski et al., 2003; Tan et al., 2016). As described in Materials and Method section, all-confidence can be used to set the smallest confidence of all rules that can be produced from a pattern, i.e., all rules produced from the pattern will have a confidence greater or equal to its all-confidence value, creating the basis for rule reliability exploration at the pattern level (Xiong et al., 2006; Omiecinski et al., 2003; Tan et al., 2016; Hashler, 2005; Hashler, 2011; Naulaerts, 2015).

The main result of this step is the creation of the pattern table (step 5). Conceptually similar to the microbiome taxa table, the pattern table described the presence of a pattern for each sample, integrating the interest measures previously calculated (step 4). microFIM visualizations comprehend distributions of patterns considering support, length and interest measure values. To describe the relationships between samples considering patterns found, a Jaccard matrix can be also obtained and visualized (step 6).

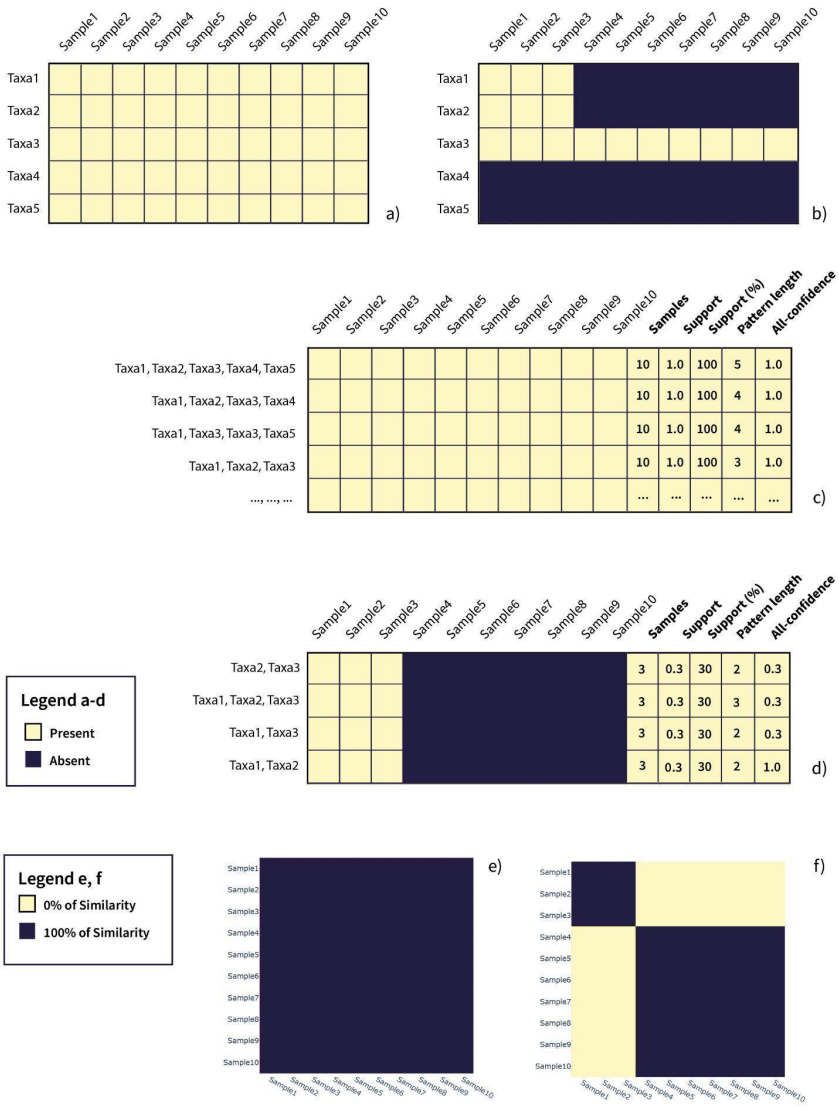

Figure 3. a) Graphical representation of Table 1; b) Graphical representation of Table 2; c) Pattern table generated from Table 1; d) Pattern table generated from Table 2; e) Jaccard heatmap plot of Table 1; f) Jaccard heatmap plot of Table 2.

To better show the potentials of microFIM, we included a demonstrative

analysis of both simulated data and data belonging to real case studies (see the next Results section). In particular, as also described in the Materials and Methods, simulated data are composed of two main matrices with a dimension of 10 samples and 5 taxa. In Fig. 3a and 3b a graphical representation of the simulated matrices is shown. Through microFIM, ARM analysis was performed. The final output of the analysis is the pattern table, represented in Fig. 3c and 3d and available in Supplementary Tables 6 and 7, respectively. The pattern table integrates the interest measures of length, support and all-confidence and, as it is a dataframe, patterns can be filtered and further visualized with Python libraries or other data analysis tools easily. In addition, results of the pattern table can be visualized with microFIM through the following plots: scatter plot, bar chart and heatmap. In Fig. 3e and 3f, heatmaps built on Jaccard distance results are shown.

In detail, Dataset 1 (Fig. 3a and Supplementary Table 1) is a full-presence dataset. This means that ARM can potentially generate all the combinations of patterns from a length of 1 to a length of 5. All patterns will have a 1.0 of support and a 1.0 of all-confidence, as they are all associated with each other. In this case, considering only the pattern composed by Taxa1, Taxa2, Taxa3, Taxa4 and Taxa5, with a length equal to 5 and a support equal to 1.0, can be sufficient to resume the information within the dataset. In addition, these settings can be adjusted directly by running the algorithm, avoiding the creation of uninformative patterns and reducing calculation time. In Fig. 3e, Jaccard heatmap shows also the 100% similarity between Dataset 1 samples. The complete pattern list obtained by Dataset 1 is available in Supplementary Table 6.

Considering Dataset 2 (Fig. 3b and Supplementary Table 2), instead, a different composition can be observed. In particular, Taxa1, Taxa2 and Taxa3 co-occur in samples 1, 2 and 3. In addition, Taxa3 is present in all the samples (Fig. 3b). As we ran an ARM analysis considering a minimum length of 2, the pattern composed by only Taxa3 was not detected. However, the pattern built by Taxa1, Taxa2 and Taxa3 was detected, with a pattern length of 3 and a support

of 0.3. Focus the attention on Taxa1-Taxa2 pattern, the value of all-confidence is equal to 1.0, meaning that there is a strong association between them and the rules generated from this pattern will have a minimum confidence of 1.0. Details about patterns extracted from Dataset 2 are available in Supplementary Table 7.

5.3.2 microFIM applied on real case studies

Association rule mining is a data mining technique widely used in very different research fields and applications. This chapter is dedicated to the use of ARM, in particular the pattern mining step, on real microbiome case studies . In detail, three case studies was chosen to demonstrate the potentials of ARM and microFIM: the ECAM dataset (Bokulich et al., 2016), the vaginal microbiome case study of Ravel et al. (2011) and the Montassier case study (Montassier et al., 2016) (see Material and Methods section for details). Considering the potential of ARM to reconstruct patterns, we focused the analysis on three main aspects: the type of input used, the filter of patterns whose elements are highly related to each other (also called hyperclique patterns; Xiong et al., 2006) and the use of metadata to filter and apply ARM.

To evaluate how ARM can be used on microbiome data, different types of inputs were considered. In particular, for the ECAM case study, we used: i) the ECAM taxa table obtained directly from QIIME2 datasets (Bolyen et al., 2019) in which only taxa assigned to genus level, with a relative abundance > 0.1 % in more than 15% of samples, are considered (Input 1 - data are available in Supplementary File 3); ii) family table obtained from collapsing the original one via QIIME2 plugins (Input 2 - Supplementary File 4); iii)a taxa table consisting only of taxa with complete taxonomy at the genus level (Input 3 - Supplementary File 5).

Minimum support thresholds of 0.2, minimum length of 3 and maximum length of 15 were considered. In Fig. 4 we show the results about the number of patterns retrieved considering three levels of analysis: output after the analysis previously described, patterns filtered with a minimum all-confidence of 0.5 and

patterns filtered with a minimum all-confidence of 0.8. In Fig. 4, for each filter, the distribution of support values and pattern length are provided.

In detail, Input 1 (Supplementary File 3) generated a total of 1,844,696 patterns. The mean support achieved by the patterns generated is 0.3 and a median of 0.2, with a minimum value of 0.2 and maximum value of 0.7. Regarding the pattern length, the mean value is 8.45, while the median is 8, with a minimum value of 3 and maximum value of 16.

Family table (Input 2 - Supplementary File 5) generated a total of 23,997 patterns. The mean support achieved by the patterns generated is 0.28 and a median of 0.24, with a minimum value of 0.2 and maximum value of 0.85. Regarding the pattern length, the mean value is 6.38, while the median is 6, with a minimum value of 3 and maximum value of 12.

Regarding genus table (Input 3 - Supplementary File 6), ARM analysis generated a total of 25,250 patterns. The mean support achieved by the patterns generated is 0.25 and a median of 0.23, with a minimum value of 0.2 and maximum value of 0.85. Regarding the pattern length, the mean value is 6.14, while the median is 6, with a minimum value of 3 and maximum value of 11.

All the results are available in Supplementary Table 6, 7 and 8, respectively, and can be visualized in Fig. 4.

In order to consider the putative informative patterns, a framework involving hypercliques patterns (Xiong et al., 2006) was applied. In particular, the all-confidence metric was considered at 0.5 and 0.8 thresholds for all the datasets analysed (Input 1, 2 and 3).

Regarding the Input 1 (Supplementary File 3), a total of 2,213 patterns were extracted considering an all-confidence of 0.5, while no patterns were obtained with 0.8 threshold. First all-confidence threshold resulted in patterns with a mean and a median support value was 0.43, with a minimum value of 0.21 and a maximum of 0.72. Pattern length consisted in a mean of 3.9, a median length of 4, with minimum and maximum of 3 and 7, respectively.

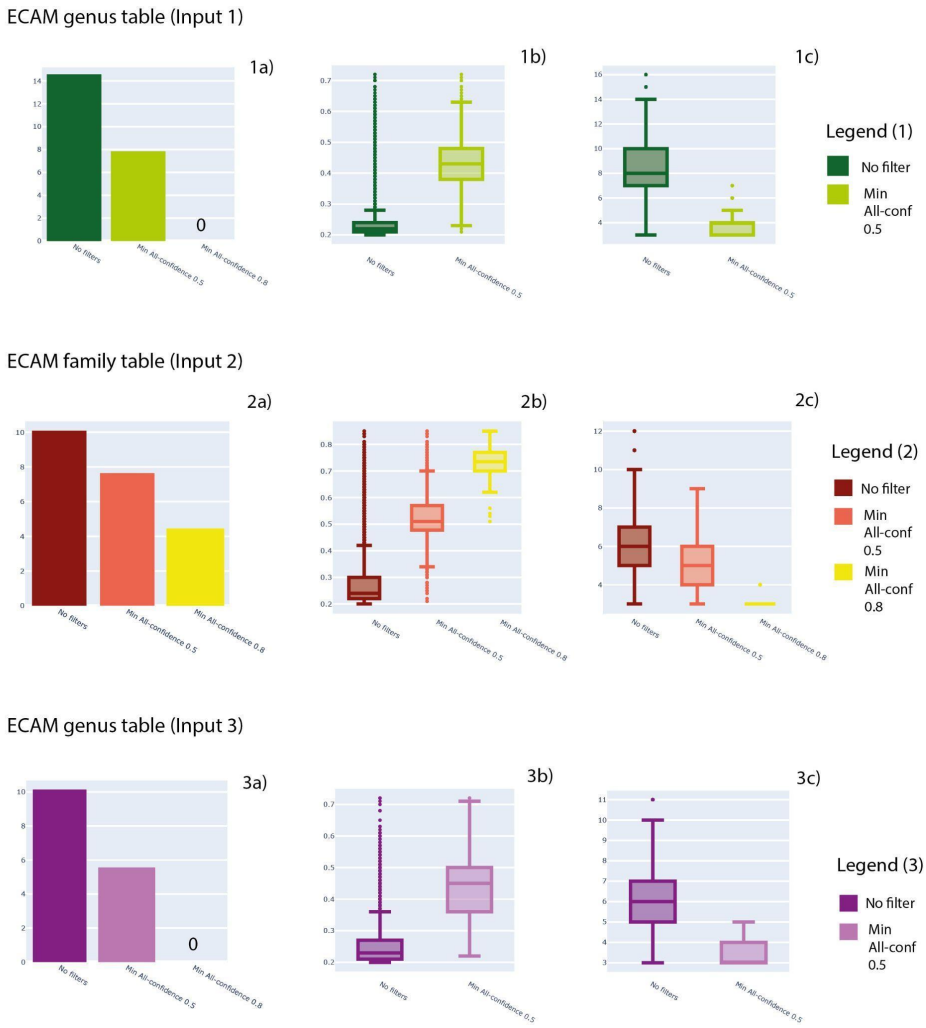

Figure 4. For Input 1, 2 and 3, here number of patterns obtained (1a, 2a, 3a), distribution of support values (1b, 2b, 3b) and distribution of pattern lengths (1c, 2c, 3c) are shown. In particular, three levels of analysis are shown: no filters applied to patterns, a minimum all-confidence of 0.5 and a minimum all-confidence of 0.8.

Regarding the Input 2 (Supplementary File 4), a total of 2,081 patterns were extracted considering an all-confidence of 0.5. A mean support of 0.53 and a median support was 0.51 were observed, with a minimum value of 0.21 and a maximum of 0.85. Pattern length consisted of a mean of 4.98, a median length

of 5, with minimum and maximum of 3 and 9, respectively. A total of 78 patterns were extracted considering an all-confidence of 0.8. A mean support of 0.72 and a median support was 0.73 were observed, with a minimum value of 0.51 and a maximum of 0.85. Pattern length consisted of a mean of 3.23, a median length of 3, with minimum and maximum of 3 and 4, respectively.

Regarding the Input 3 (Supplementary File 5), instead, a total of 25,250 patterns were extracted considering an all-confidence of 0.5, while no patterns were obtained with 0.8 threshold. First all-confidence threshold resulted in patterns with a mean of 0.25 and a median support value of 0.23, with a minimum value of 0.2 and a maximum of 0.72. Pattern length consisted in a mean of 6.14, a median length of 6, with minimum and maximum of 3 and 11, respectively.

For demonstrative purposes, a Jaccard heatmap considering samples belonging to the first sampling date of the ECAM dataset of the Input 3 table (Supplementary Table 5) was generated, in order to show a potential use of Jaccard distance on pattern analysis (available in Supplementary Figure 11). In general, results are summarized in Fig. 4 and tables are available in Supplementary Table 8, 9 and 10, respectively.

Overall, Input 1 obtained the maximum number of patterns, achieving 1,844,696 patterns. The support distribution has a great range of values for all the three datasets, from 0.2 to almost 0.8. Also length achieved a wide range of values, considering patterns from 3 elements length to almost 16. In general, a great reduction in the number of patterns was observed considering the all-confidence filtering (Fig. 4 - sections 1a, 2a and 3a). In parallel, this filter resulted in higher support values (Fig. 4 - sections 1b, 2b and 3b) and lower pattern length (Fig. 4 section 1c, 2c and 3c).

Metadata filtering was applied to the genus ECAM dataset, considering two category types: antibiotic administration and type of delivery. The complete results of the pattern analysis are available in Supplementary Table 12. Overall, a total of 141,480 patterns were obtained from the data belonging antibiotic administration, while the opposite obtained a total of 8,223. Vaginal delivery resulted in a total of 45,412 patterns, while cesarean delivery samples resulted

in 10,288. Also in this case, the usage of all-confidence filtering drastically reduced the number of explorable patterns, achieving the following results: 2 and 1 patterns for antibiotic administration and vaginal delivery, respectively, and 0 patterns for the opposites.

microFIM was also applied to other two real case studies: vaginal microbiome obtained by the work of Ravel et al. (2011) and the dataset of Montassier case study (Montassier et al., 2016). Considering the first one, different input types and metadata filtering were used: in particular, the dataset was obtained from the MLRepo collection (Vangay et al., 2019). Then, family level and genus level dataset were obtained. Dataset can be identified as Input 4 (dataset available in MLRepo; Vangay et al., 2019 - Supplementary File 15a), Input 5 (dataset at the family level - Supplementary File 15b) and Input 6 (dataset at the genus level - Supplementary File 15c). As for the ECAM analysis, results are presented considering the three main input types and the number of distribution of patterns are evaluated as the previous scheme.
In particular, Input 4 (Supplementary File 15a) generated a total of 83 patterns. The mean support achieved by the patterns generated is 0.2 and a median of 0.2, with a minimum value of 0.2 and maximum value of 0.5. Regarding the pattern length, the mean value is 3.1, while the median is 3, with a minimum value of 3 and maximum value of 4. Family table (Input 5 - Supplementary File 15b) generated a total of 226 patterns. The mean support achieved by the patterns generated is 0.25 and a median of 0.23, with a minimum value of 0.2 and maximum value of 0.55. Regarding the pattern length, the mean value is 3.68, while the median is 4, with a minimum value of 3 and maximum value of 6. Regarding genus table (Input 6 - Supplementary File 15c), ARM analysis generated a total of 225 patterns. The mean support achieved by the patterns generated is 0.25 and a median of 0.24, with a minimum value of 0.2 and maximum value of 0.46. Regarding the pattern length, the mean value is 3.77, while the median is 4, with a minimum value of 3 and maximum value of 6. All the results are available in Supplementary Table d, e and f, respectively, and

can be consulted in Supplementary Table 14.

Minimum all-confidence of 0.5 and 0.8 were considered to evaluate hypercliques patterns.

Regarding the Input 4 (Supplementary File 15a), 16 patterns were extracted considering an all-confidence of 0.5, while no patterns were obtained with 0.8 threshold. First all-confidence threshold resulted in patterns with a mean of 0.23 and a median support value was 0.21, with a minimum value of 0.2 and a maximum of 0.48. Pattern length consisted in a mean of 3.06, a median length of 3, with minimum and maximum of 3 and 4, respectively.

Input 5 (Supplementary File 15b) obtained 2 patterns were extracted considering an all-confidence of 0.5, while no patterns were obtained with 0.8 threshold. The 0.5 all-confidence threshold resulted in patterns with 0.46 and 0.55 support values. Both patterns have a length of 3.

Regarding the Input 6 (Supplementary File 15c), 15 patterns were extracted considering an all-confidence of 0.5, while no patterns were obtained with 0.8 threshold. First all-confidence threshold resulted in patterns with a mean and a median support value was 0.3, with a minimum value of 0.25 and a maximum of 0.38. Pattern length consisted in a mean of 3.13, a median length of 3, with minimum and maximum of 3 and 4, respectively.

Overall, the support distribution has a low range of values for all the three input files, from 0.2 to almost 0.5. Length is around 3 elements per pattern. In general, also in this case a great reduction in the number of patterns was observed considering the all-confidence filtering (Supplementary Table 14).

Metadata filtering was applied to the dataset, considering the nugent category, low and high levels. The complete results of the pattern analysis are available in Supplementary Table 14. Overall, a total of 15,836 patterns were obtained from the data belonging to high nugent score value, while the opposite obtained a total of 21. The usage of all-confidence filtering drastically reduced the number of explorable patterns, obtaining 16 patterns for high nugent score value.

Finally, Montassier dataset (Montassier et al., 2016) was tested considering the OTU table and taxa table obtained from MLRepo pipeline (Vangay et al., 2019).

A minimum support threshold of 0.9 was considered, with a minimum length of 5 and a maximum length of 10. A total of 446 patterns were obtained considering the taxa table, while 9 patterns were obtained considering the OTU table.

Distributions of pattern and length are similar between the two input files. In particular, a mean support of 0.93 and a mean length of 5.1 (5-6) were detected.

5.4 Discussion

Pattern mining strategies are now newfangled solutions for disclosure of microbial patterns (Tandon et al., 2015; MANIEA et al., 2021). However, besides the power of these techniques, great efforts must be undertaken to extrapolate relevant patterns that can be integrated into biological contexts (Naulaerts et al., 2015; Faust et al., 2021).

Basically, the strategy consists of two main phases: i) extraction of patterns (also known as 'frequent itemset mining') and ii) rules calculation. In this work, we focused in particular on the first phase, as great potential can be achieved considering the exploration of patterns at any length and subsequently be filtered to create reliable associations.

In detail, our Discussion section will touch two main topics: i) considerations about parameter settings to perform pattern mining strategies in the context of 16S rRNA metabarcoding data and ii) guidelines and future perspectives to support real applications. In order to present an overview of frequent itemset mining as a tool for microbiome pattern analysis, we developed a SWOT (Strengths, Weaknesses, Opportunities, Threats) analysis (Fig. 5).

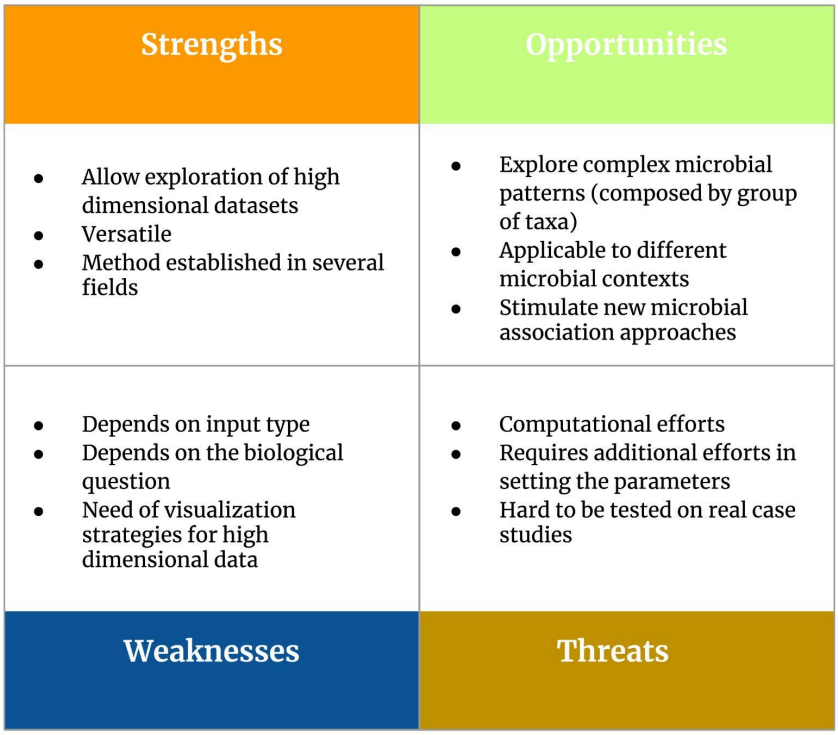

Figure 5. Overview of the main strengths, weaknesses, opportunities and threats (SWOT analysis) related to the use of frequent itemset mining as a tool for microbiome pattern analysis.

5.4.1 Run ARM could not be enough without care in setting parameters

As described above, pattern mining strategies can be powerful to get insights from large and complex datasets (Naulaerts et al., 2015). However, pattern analysis may have limitations (Faust et al., 2021). In this work, we provide ARM analysis on both simulated and real datasets and propose microFIM (https://github.com/qLSLab/microFim), a Python tool specifically suited for microbiome pattern analysis. Our results will consider the pattern composition obtained through our framework (Material and Methods section) without

considering their biological implications, as it is beyond the scope of this work.
Considering the application of ARM on simulated datasets, we showed that initial settings can reduce the amount of information retrievable, both considering interest measures as support or length and all-confidence metric.
Regarding the application on the real case studies, a few considerations can be made. First of all, the type of input can change the reliability of results: different numbers of patterns have been generated considering different input types. In particular, both considering aspects related to data visualization and interpretation, the taxonomy level of investigation must be considered.
A second point that arises is the minimum support threshold to choose. The choice can be both related to biological questions, as for example which is the minimum number of samples to retain a pattern interesting, but also on technicalities. In detail, exploring all the potential patterns cannot be reliable and useful, as the number of patterns can be very high, related also to great computational efforts and visualization issues (Naulaerts et al., 2015). For this reason, we started using a support of 0.2, that means that only the taxa that co-occur in at least the 20 % of samples were considered (up to 175 of 875 for the ECAM dataset and up to 77 of 388 for the Ravel case study). However, this is a case-specific threshold as no guidelines exist to set a correct support threshold in this research field. The wrong value can potentially hide information and, at the same time, create spurious patterns. In addition, it can generate misleading results without taking into account the Simpson's paradox (Tan et al., 2016), a phenomenon in which a pattern appears frequently but disappears or drastically changes when the data are combined differently, as for example considering only a set of samples (Tan et al., 2016).
Nevertheless, once patterns are generated, filtering steps can be added, in order to both reduce the information and better evaluate specific patterns, with peculiar characteristics. Filters can include the length of patterns or additional interest measures (Karpinets et al., 2012; Naulaerts et al. 2015; Agrawal et al., 1993).
Pattern length, in particular, can be also included before running the analysis,

as algorithms take into account a minimum and a maximum value of pattern length, in order to reduce the number of explorable patterns (Agrawal et al., 1993). However, this choice must be done before exploring the results. Of course, it is possible to reduce the number of patterns after extraction, but computational efforts and running time must be considered (Naulaerts et al. 2015; Agrawal et al., 1993). Pattern length can also vary based on the research field of application and the biological questions. In the ECAM case study, for example, we observed different median values of pattern length, from minimum values of 3 to maximum of 16, suggesting also different levels of analysis.

However, other metrics can be included to filter patterns (Omiecinski et al., 2003; Franceschini, 2012; Tan, 2002; Tang et al., 2012). Usually they are called 'interest measures' and are generally used to evaluate a set of peculiar patterns, in order to filter the interesting ones (Naulaerts et al., 2015; Hussein et al., 2015; Tatti et al., 2010). Also in this case, the biological question can guide how to properly set the filtering step. In this work, we used all-confidence metrics, which generate hyperclique patterns (Omiecinski et al., 2003; Xiong et al., 2006). The application of this metric helps to find groups of items (in this case species or taxa) where items belonging to the same pattern are highly affiliated with each other and can generate rules with the minimum threshold chosen. Using this approach reduces drastically the number of patterns and, in addition, allows to filter only strong associated groups. In this case, the amount of information was drastically reduced considering the two thresholds of all-confidence considered (0.5 and 0.8). This reduction can promote a manual exploration of results and pave the way for exploring strong associations and putative rules.

Clearly, other interest measures can be applied. All-confidence may not be the only interest measures useful for microbiome analysis. Other metrics can be selected to filter patterns, but they must be identified based on specific questions related to the research field of application (Naulaerts et al. 2015).

5.4.2 Fitting ARM for microbiome studies: guidelines to support real applications

Frequent itemset mining and, subsequently, association rule mining, is a pattern mining technique able to explore items that co-occur with a certain frequency, as sets of commercial products that customers buy together in the classic supermarket basket problem (Naulaerts et al. 2015; Agrawal et al., 1993). The flexibility of frequent itemset mining techniques is demonstrated by the wide range of bioinformatics applications, from for example SNPs association studies to annotations and motif association exploration (Manda et al., 2020; Manda et al., 2013; Manda et al., 2012; Koyuturk et al., 2006; Ong, 2020; Karpinets et al., 2012; Alves, 2010; Carmona-Saez et al., 2007; Boutorh et al., 2016; Agapito et al., 2015; Naulaerts et al., 2016; Zhou et al., 2013). It is a powerful instrument to explore patterns from large and complex data sets (Karpinets et al., 2012; Naulaerts et al. 2015; Agrawal et al., 1993), providing different algorithms and a wide range of parameters to filter patterns of interest. Besides the most used, as support (frequency of a pattern or a rule in the dataset) or length (the number of species contained in a pattern), other metrics can be included in the pattern analysis (Naulaerts et al. 2015; Agrawal et al., 1993; Hashler, 2005). Beside its potentials, great efforts have to be made to perform pattern mining strategies on microbiome data and obtain reliable and interpretable results, with sound biological implications. As mentioned above, a few points raised from the works done. From threshold choices to input data types, setting pattern analysis is not an easy task. Considering the peculiarities of microbiome data and the flexibility of the technique, here we propose five statements to guide researchers before starting ARM analysis.

Setting the input data. This point highlights the importance of the type of pattern to be considered. In the microbial ecology field, a lot of interest probably regards the investigation of species patterns, in order to evaluate community patterns and putative ecological processes. However, this is not straightforward if we consider 16S rRNA metabarcoding data: taxonomy does

not always reach a species level and this uncertainty can negatively impact pattern reconstruction. In addition, noise derived from contamination or sequencing biases can be present (Faust et al., 2012; Faust et al., 2021; Gloor et al., 2017; Balint et al., 2016). However, precautions can be taken: removing uncertain taxa or cleaning the table based on abundance thresholds or statistical methods is possible (Faust et al., 2012; Gloor et al., 2017; Balint et al., 2016). Different levels of taxonomy can be used as input, as we also demonstrated in the previous sections. Of course, choices must be taken with conscience as they will impact on the final result and therefore the interpretation must be correctly contextualized.

Consider the use of metadata. The inclusion or filtering considering metadata information can improve the reliability of the method, both looking for specific patterns linked to metadata and also to better explore the dataset. In this way, we can reduce the information to be explored, lowering the support value, retaining rare or patterns related to specific metadata, and preventing Simpson's paradox issues (Naulaerts et al. 2015; Agrawal et al., 1993).

Individuate what is interesting for the specific case study. The definition of what is interesting depends on the biological context at issue. No simple guidelines exist, as the application of pattern mining on microbiome data is still in its infancy (Naulaerts et al. 2015). Testing and developing new metrics is an important field of research and can make a difference to track reliable patterns that can be further used for classification tasks or functional analysis. In this work, we applied the all-confidence metric (Omiecinski et al., 2003; Xiong et al., 2006). However, we believe that other interest measures can be applied and a wide variety of them are available in other tools already developed (Hahsler et al., 205; Hahsler et al., 2011). In general, this step allows to drastically reduce the number of explorable patterns (Tan et al., 2016; Omiecinski et al., 2003; Xiong et al., 2006).

Basically, length can be used to clean the information extracted via ARM. As ARM can generate patterns at any length, single items or only pairs of items can be pruned, in order to find interesting associations composed by 3 or more

elements. From a biological point of view, exploring longer microbial patterns can enhance microbial community investigations and pave the way for high-order interactions exploration (Faust et al., 2021).

Consider computational time. As fully described in previous works, data dimensions and density drastically increase time calculation and memory usage (Naulaerts et al. 2015; Agrawal et al., 1993). Reducing input data can make ARM more reliable and faster to be performed (Naulaerts et al. 2015; Agrawal et al., 1993). In addition, beside the common concept of pattern, closed and maximal patterns exist. Both result in a faster extraction, but with a reduction of information (Naulaerts et al. 2015; Agrawal et al., 1993).

Overall, the inclusion of interest measures directly into the ARM framework may favour the development of new faster algorithms, leading the technique directly to the exploration of specific patterns (Omiecinski et al., 2003; Xiong et al., 2006; Naulaerts et al. 2015).

Tools and visualization strategies. To better suit pattern mining for microbiome data applications, tools and visualization techniques are essentials (Naulaerts et al. 2015). In detail, in this work we tried to concept a new pattern mining output combining the common microbiome output with pattern analysis. The pattern table can be an important resource to perform and visualize pattern results in a microbial perspective. In addition, it allows further statistical analysis that is usually performed for microbiome data. Considering the visualization process, we set up different plots to have an overview of pattern distributions and create a Jaccard matrix to show the distance between samples. However, different visualization methods exist, based on tables, matrices and graphs (Naulaerts et al. 2015). Here we cite the R packages arulesviz, FPViz and WiFIsViz (Hashler, 2005; Hashler, 2011; Naulaerts, 2015). Even though these visualizations allow different strategies to explore data, issues related to high dimensional dataset remain and none of them are conceptualized for microbiome analysis. At the same time, collecting human readable information can facilitate data visualization strategies and interpretation (Naulaerts et al. 2015), but of course interesting measures must

be considered. Finally, considering practicality of use, several ARM implementations can be utilized (Nauleaerts, 2015). Moreover, frameworks have been implemented, often accompanied by GUI (Graphical User Interface) or interactivity components (Naulaerts, 2015). However, a deepening in the microbiome field has not been established yet.

Evaluation and benchmarking strategies. From a computational point of view, the complexity and dynamics of microbial communities leads to difficulties in developing and testing methods to evaluate them. In general, it was demonstrated that microbial co-occurrence analysis may be an extraordinarily promising approach for studying microbiomes (Faust and Raes, 2012). Several works explained how co-occurrences reveal indications about ecological processes shaping community structure (Lima-Mendez, 2010), exploring hub species and potential microorganisms relationships (Berry, 2014). Further, Ma and colleagues (2020) showed how global microbial co-occurrence analysis and network reconstruction may be an encouraging strategy to reveal patterns and explore new mechanisms. However, besides these results, transform microbiome data into purposeful biological insights remain challenging, as also demonstrated by different evaluations (Faust et al., 2012; Berry et al., 2014), and open questions still remain (Faust et al., 2021; Ma et al., 2020; Layeghifard et al., 2017; Faust et al., 2012). The use of ARM on microbiome data models or datasets created in-silico will be necessary to disentangle the potentials of ARM in the microbiome research field, also considering the range of microbiome aspects that can be considered (Faust et al., 2021; Hosoda et al., 2020; Weiss et al., 2016). In particular, tests should examine how the technique is affected by noise signals, both related to sequencing and laboratory protocols (Weiss et al., 2016). In addition, as microbiome data may potentially describe a complex and intricate ecological community, several ecological aspects can be evaluated with ARM, both describing the generation of redundant information and the difficulty associated with extracting patterns due to specific ecological behaviors, as for example competition, exclusion or symbiosis (Faust et al., 2021; Weiss et al., 2016; Faust and Raes, 2012).

In general, recent advancements in data integration and data reuse strategies may enhance the exploration of microbial patterns from large-scale studies (Ghannam et al., 2021; Jordan et al., 2015; Ma et al., 2020; Su et al., 2020). Microbiome simulators and in vitro studies can be a great instrument for benchmarking works and improve guidelines to apply ARM (Faust et al., 2021). Beside the potential of ARM on large scale analysis, giving a great overview of data under investigation (Naulaerts et al. 2015), these advancements may contribute to developing tests and benchmarking strategies in order to set ARM for microbial pattern research looking at biological implication, specifically.

5.5 Conclusions

Concluding, all the challenges mentioned above can disentangle ARM analysis for microbiome pattern exploration. As the output of the analysis can be extensive and redundant, results should be interpreted with caution. The associations extracted do not necessarily imply causality. Instead, it suggests a strong co-occurrence relationship between species. Causality, on the other hand, requires knowledge about the causal and effect attributes in the data (Tan et al., 2016). There are several approaches to evaluate the robustness of an output. In this first work, pattern length, support and all-confidence were explored and included in the microFIM tool. From a biological perspective, filtering results with these parameters could help to highlight meaningful patterns, but may not be enough. Further, we tried to depict issues that we think must be considered before using an ARM approach for specifical biological traits. As there is an interest in research to exploit data mining techniques, citing for example the works of Srivastava et al., 2019 or Zakrzewski et al., 2016, we also think that suiting ARM for microbiome analysis will be a great resource in the future. Considering the huge amount of data available and produced with the advent of High-Throughput DNA Sequencing (HTS) technologies, an increasing selection of large-scale data science

strategies seems to have enormous potential in resolving challenges in microbiome pattern exploration (Kypides et al., 2020; Jordan et al., 2015). Association rule mining and microFIM tools may have great potential not only with 16S rRNA metabarcoding data, but also in a wide range of applications. As also supported by Naulaerts et al. (2016), ARM analysis is a versatile technique: the integration of files such as taxa tables guarantees the usage also on a wide variety of datasets belonging from different sources, as for example the QIITA platform (https://qiita.ucsd.edu/; Gonzales et al., 2018) or the MLrepo (https://knights-lab.github.io/MLRepo/; Vangay et al., 2019), but not only. Beside the main focus of this work and microFIM development, very different types of data can be analysed and integrated with ARM framework. From gene associations to merely metabarcoding projects, whose output has the same structure of 16S rRNA taxa table, microFIM may potentially pave the way for multiple usages, creating a bridge with several research fields and applications.

5.6 Supplementary

Supplementary data are available in the main paper (https://doi.org/10.3389/fbinf.2021.794547). In particular, supplementary files are the following:

- Supplementary material 1: Table describing a simulated dataset 1 composed of 5 taxa and 10 samples (CSV format).
- Supplementary material 2: Table describing a simulated dataset 2 composed of 5 taxa and 10 samples (CSV format).
- Supplementary material 3: ECAM taxa table obtained directly from QIIME2 datasets (Bolyen et al., 2019) in which only taxa assigned to genus level, with a relative abundance > 0.1 % in more than 15% of samples, are considered (TSV format).
- Supplementary material 4: Family ECAM taxa table obtained collapsing the ECAM dataset (Supplementary Table 3; Bolyen et al., 2019) to the

family level via QIIME2 plugins (https://github.com/qiime2/q2-taxa) (TSV format).

- Supplementary material 5: Genus ECAM taxa table obtained collapsing the ECAM dataset (Supplementary Table 3; Bolyen et al., 2019) consisting only of taxa with complete taxonomy at the genus level (TSV format).
- Supplementary material 6: Pattern table generated performing microFIM on simulated dataset 1 (Supplementary Table 1) with the minimum support of 0.3, a minimum length of 2 and a maximum length of 10 (CSV format).
- Supplementary material 7: Pattern table generated performing microFIM on simulated dataset 2 (Supplementary Table 2) with the minimum support of 0.3, a minimum length of 2 and a maximum length of 10 (CSV format).
- Supplementary material 8: Table generated performing microFIM on ECAM dataset (Supplementary Table 3) with a minimum support of 0.2, a minimum length of 3 and a maximum length of 15 (CSV format).
- Supplementary material 9: Pattern table generated performing microFIM on ECAM dataset at family level (Supplementary Table 4) with a minimum support of 0.2, a minimum length of 3 and a maximum length of 15 (CSV format).
- Supplementary material 10: Pattern table generated performing microFIM on ECAM dataset at genus level (Supplementary Table 5) with a minimum support of 0.2, a minimum length of 3 and a maximum length of 15 (CSV format).
- Supplementary figure 11: Heatmap representing Jaccard distance matrix was generated via microFIM visualization phase on the ECAM dataset, considering Input 3 and samples belonging to the first sampling date.

6. SKIOME Project: a curated collection of skin microbiome datasets enriched with study-related metadata

6.1 Introduction

Directly in contact with the environment, the skin microbiome is a tangled and dynamic ecosystem that interacts with both the host and its surroundings (Dimitriu et al., 2019). It is characterized by diverse ecological niches, where the microbiota, the host skin cells and the host immune system are involved in the maintenance of skin health. In the last decade, numerous studies have investigated the composition of the human skin microbiome under very different conditions (Swaney et al., 2021; Luna et al., 2020; Callewaert et al., 2020; Sa et al., 2015).

The advent of high-throughput DNA sequencing (HTS) technologies has revolutionized numerous research fields, and the study of the human microbiome was no exception. Following the introduction of HTS technologies, the number of studies investigating the human microbiome has increased, expanding our knowledge about its implications for human health. In particular, it was demonstrated its pivotal linkage with diet and age (Sa et al., 2015; Leyden et al., 1975) and specific microbiome patterns were shown to relate to the body region sampled (Capone et al., 2011; Bouslimani et al., 2015). Geography and ethnicity have also been shown to affect the skin microbiome (Gupta et al., 2017) and numerous diseases have been associated with an altered microbial state (Byrd et al., 2018), as in the cases of atopic dermatitis (Williams et al., 2015) and psoriasis (Langan et al., 2018).

Since their adoption, the new sequencing strategies have been getting cheaper and cheaper, becoming available for researchers and companies on a global scale. In recent years, large amounts of data have been deposited in public

databases and more is going to be produced in the near future, as the number of sequencing experiments is exponentially growing.

There are three major databases used to store nucleotide sequence data: the NCBI's Sequence Read Archive (SRA) (Sayer et al., 2019), the EBI's European Nucleotide Archive (ENA) (Harrison et al., 2019), and the DDBJ Sequence Read Archive (DRA) (Ogasawara et al., 2020). These three databases are brought together by the International Nucleotide Sequence Database Collaboration (INSDC) and are constantly synchronized to share their data (Arita et al., 2021). The publicly available datasets deposited in these databases represent a valuable resource for the microbiome research community. Public available data can be now accessed and downloaded to be re-analysed or integrated to perform meta-analysis studies (Duvallet et al., 2017; Bisanz et al., 2019; Kosti et al., 2020).

As a consequence, in the last few years, we are facing an increasing adoption of novel large-scale data science approaches to address challenges in microbiome science (Kyrpides et al., 2020). For example, machine learning strategies can be applied to perform powerful prediction tasks on metagenomics data (e.g. disease-prediction based on microbiome composition). However, these strategies require a large amount of data to train and test models, making the integration and harmonization of multiple datasets a necessary step (Jordan and Mitchell, 2015; Ghannam and Techtmann, 2021). In this way, the availability of large-scale sequencing data can enable microbiology researchers to ask new questions and develop new strategies to study the human-associated microbial communities (Wood-Charlson et al., 2021; Su et al., 2020).

However, this huge amount of microbiome data still lacks harmonization and is far from being completely exploited to its full potential. Guidelines have been proposed and tools have been developed to promote the standardization of sample processing, sequencing and data analysis across the microbiome field

(Greathouse et al., 2019; Bharti and Grimm, 2021; Liu et al., 2021; Amos et al., 2020; Pollock et al., 2018; Callahan et al., 2016; Bolyen et al., 2019; She et al., 2019) but achieving global standardization is not an easy task. Initiatives such as the Human Microbiome (Turnbaugh et al., 2007) and the Earth Microbiome Projects (Gilbert et al., 2014) have favored the development of standardized procedures. In addition, important field-specific databases were created, such as the Human Oral Microbiome Database (Chen et al., 2010) or the GMrepo, a database of curated and consistently annotated human gut metagenomes (Wu et al., 2020).

Several research groups have been proposing different sources of microbiome data: initiatives like the Human Microbiome and the Integrative Microbiome Projects (Gevers et al., 2012; Proctor et al., 2019), MicrobiomeDB (Oliveira et al., 2018), HumanMetagenomeDB (Kasmanas et al., 2021), curatedMetagenomicData (Pasolli et al., 2017), the ML Repo (Vangay et al., 2019), QIITA portal (Gonzales et al., 2018), or the MG-RAST portal (Wilke et al., 2016) suggested both data management infrastructures and frameworks to guarantee data accessibility and reuse.

Despite the contribution of groups involved in this field, the lack of metadata and the presence of datasets with missing or inconsistent information can reduce the interpretability of the data generated, influencing the understanding of microbial dynamics and ecological patterns (Wood-Charlson et al., 2020; Su et al., 2020; Greenhouse et al., 2019). Inconsistency and uncontrolled metadata filling were demonstrated by Gonçalves and Musen (2019), revealing the necessity of standardized metadata compilation (Bernasconi, 2021).

FAIR (Findable, Accessible, Interoperable, and Reusable) principles are supported within the National Microbiome Data Collaborative and FAIR Microbiome community (https://www.go-fair.org/implementation-networks/overview/fair-microbiome) (Wood-Charlson et al., 2020; Vangay et al., 2019) to promote data discovery

and reuse in the microbiome field, and allow for broader dissemination of knowledge and compliance for both humans and machines.

Thus, making microbiome data and metadata accessible is a key aspect to guarantee a concrete opportunity to perform meta-analyses and data reuse (Vangay et al., 2019; Ching et al., 2018; She and Schloss, 2016). In this context, well-curated and FAIR microbiome datasets are now a necessity to explore microbiome patterns, apply data science techniques and promote data reusability (Duvallet, 2020; Longo and Drazen, 2016).

In order to help researchers interested in performing meta-analyses with human skin microbiome data and exploring the context-specific information related to potentially useful datasets, we focused our work on published human skin microbiome datasets, creating a curated skin microbiome collection accompanied by a state-of-the-art analysis of the last 10 years of the skin microbiome field.

In particular, during the last decade, most of the studies have relied on amplicon sequencing approaches, where different regions of the 16S rRNA gene are amplified and sequenced to identify the microbial taxa present in a sample (Bokulich et al., 2020; Knight et al., 2018). For this reason, we built a comprehensive human skin microbiome collection enriched with detailed metadata information, focusing on existing 16S rRNA amplicon-sequencing microbiome datasets from the human skin biome.
To achieve our goal, we first collected datasets from the INSDC, which store the majority of the publicly available nucleotide sequencing datasets together with their associated metadata (Arita et al., 2021). As the availability of these metadata and the possibility of recovering them is crucial for ensuring the reusability of the available datasets (Gonçalves and Musen, 2019), we dedicated special attention to maximize the amount of metadata information that can be recovered. To do so, we combined different metadata retrieval approaches enriched with a manual curation step. Then, we generated

explorable data frames at different curation levels containing all the retrieved datasets together with the associated metadata. Further, we highlighted some of the shortcomings of the current approaches for data and metadata retrieval and we called attention to some of the issues that currently aict the re-usability of the deposited data. Overall, the output of our work constitutes a valuable resource for researchers interested in performing meta-analyses with human skin microbiome data, who can explore our collection to find a list of datasets that can be integrated to answer old and new biological questions.

6.2 Materials and methods

6.2.1 Metadata retrieval and manual curation procedures

To obtain a comprehensive list of skin microbiome studies derived from amplicon approaches with the associated metadata, we built a three-step framework (Figure 1) based on:

- Step 1: dataset retrieval from INSDC;
- Step 2: metadata retrieval and enrichment;
- Step 3: output curation with the removal of redundant and spurious information.

In the sections below, all the steps are described together with the methods and strategies used.

6.2.1.1 Step 1: dataset retrieval from INSDC

To generate a comprehensive list of datasets of human skin microbiome derived from 16S rRNA amplicon sequencing available on the INSDC public databases, we decided to rely on two different approaches: i) an automatic

search, which allows querying the INSDC databases automatically using keywords and ii) a manual approach on the SRA and ENA portals.

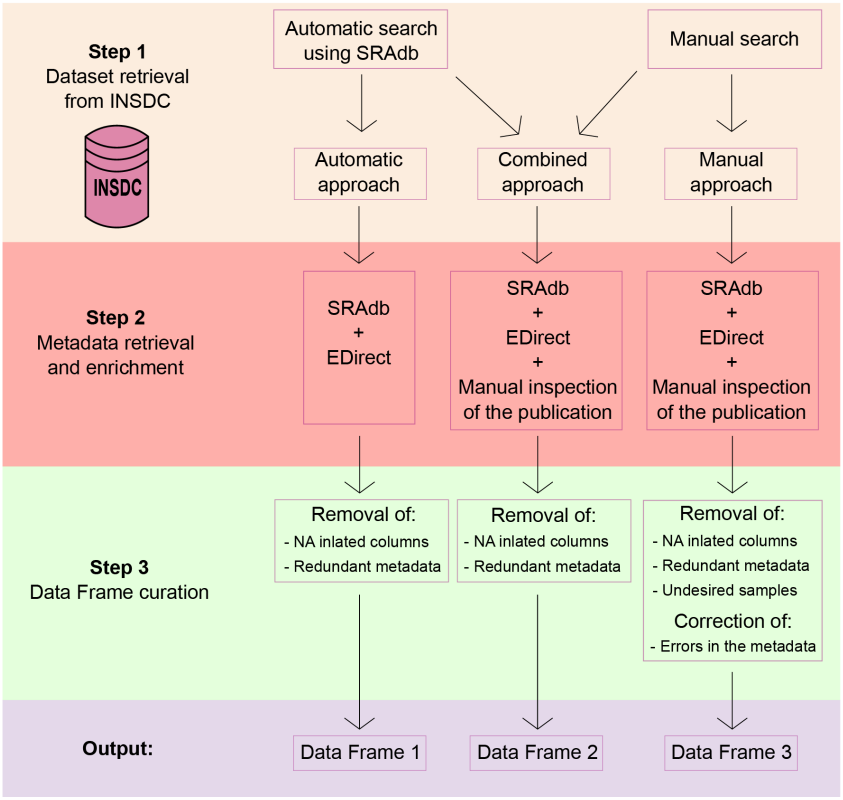

Figure 1. Schematic representation of the three-step framework adopted in the study to collect datasets and metadata and generate three differently curated data frames.

The automatic search of the datasets was performed with the R package "SRAdb" (Zhu et al., 2013). SRAdb relies on a SRAdb SQLite database, a regularly updated database of metadata associated with the raw reads deposited on SRA and its interconnected databases (ENA, DRA). The SRAdb

database (up to 36 Gb) was downloaded and stored locally on the 17 of June, 2021. We performed a full-text search with the following query: “human skin microbiome OR human skin microbiota OR human skin metagenome”.
For the manual approach, instead, we performed a search on the NCBI’s SRA and EBI’s ENA databases with the following criteria: datasets coming from 16S rRNA amplicon sequencing, containing only human skin samples that were deposited from 2012 onwards and that presented an associated publication.

6.2.1.2 Step 2: metadata retrieval and enrichment

An enrichment step was performed on both automatic and manual outputs in order to recover the largest amount of metadata associated with the datasets previously found. For this step, we integrated three different strategies: i) SRAdb was used to collect all the possible information from the retrieved list of studies and samples; ii) for some run-associated metadata that could not be retrieved with SRAdb, we used the Entrez Direct (EDirect) tool (Kans, 2021); iii) for the list of manually recovered studies, we collected study-specific metadata from the associated publication, including information that cannot be found on the INSDC databases. We focused our attention on the sample origin, the laboratory and bioinformatics strategies and the data related to the context in which the studies were performed. In particular, we retrieved study-specific information related to the collection method used, the 16S rRNA gene hypervariable region sequenced, the clustering method used (OTUs, ASVs/RSVs), the number of recovered units/variants reported in the study, the database used for taxonomic assignment and its version, the disease condition investigated (if any), the location of the sampling, the presence of a MGnify analysis (Mitchell et al., 2020), the DOI and the year and journal of publication.
In addition, a bibliometric analysis of published papers related to the datasets retrieved was performed. Research areas and categories from the Web of Science (WoS) collection and Elsevier’s Scopus classifications were added to each publication. Notably, since Scopus reported multiple subject areas for each publication, we included multiple columns in the data frame to keep all

the information. We further generated a column categorizing a scientific journal as a medicine-related journal (Medicine_Journal) or not depending on the presence of 'Medicine' among the Scopus subject areas. Lastly, an additional column containing any useful notes related to the study was added. A comprehensive list of the manually curated metadata with description is available in Supplementary File 1., also available in our Github repository (https://github.com/giuliaago/SKIOMEMetadataRetrieval).

6.2.1.3 Step 3: outputs curation and metadata correction

Once all the information was stored into three data frames that differed in the way the datasets and the metadata were retrieved, we proceeded to reorganize them by removing redundant metadata and NA-inflated columns. For the smallest and most refined data frame, we further inspected the data frame rows to remove undesired samples and to correct wrongly assigned metadata. In detail, we removed samples that were not obtained from amplicon sequencing and corrected metadata by double-checking with the related publications.

6.2.2 Script and data availability

For all the steps of datasets and metadata retrieval, a list of studies and associated metadata were kept (Dataframe 1, Dataframe 2 and Dataframe 3). All the outputs will be available in our Github repository (https://github.com/giuliaago/SKIOMEMetadataRetrieval), accompanied by the scripts used for the retrieval framework. In particular, scripts describe the use of SRAdb, Edirect tool, the entire R pipeline to obtain the final outputs and codes for plot creation and data frame exploration.

6.3 Results

Following the three steps presented in the Methods section (dataset retrieval from INSCD; metadata retrieval and enrichment; data frame curation), we first

tested two approaches to retrieve datasets of the human skin microbiome from the INSCD databases (Step 1): a manual search of the datasets and an automatic search with SRAdb (Zhu et al., 2013). We then collected metadata information for the retrieved datasets (Step 2) using three different approaches: automatic search with SRAdb (Zhu et al., 2013), EDirect (Kans, 2021) and a manual search from the associated publication for the manually retrieved studies. In this way we obtained three data frames:

- Data Frame 1, containing only datasets retrieved with SRAdb and metadata collected automatically with SRAdb and EDirect;
- Data Frame 2, containing all the datasets identified with both the strategies (manual and automatic) together with all the metadata that could be recovered with SRAdb, EDirect and manual inspection of the publication;
- Data Frame 3, a subset of Data Frame 2, containing only the manually retrieved datasets together with all the metadata that could be recovered both manually and automatically with SRAdb and EDirect.

Data Frame 2 and Data Frame 3 both contain 61 metadata columns (from manual and automatic metadata search), while Data Frame 1 only contains 37 metadata columns obtained from the automatic search. All three data frames were curated to remove redundant columns and NA-inflated columns (Step 3). Among the redundant metadata, we observe columns containing the IDs of Run, Experiment, Submission, Sample/BioSample and Study/BioProject. Other metadata recovered by both methods were the spots, the bases, the library strategy, the sequencing platform used and the Taxon ID. Data Frame 3 was further curated to remove undesired samples coming from whole-genome sequencing experiments and to correct wrongly assigned metadata.

The following sections will show the results, starting from a comparison between the data collection approaches used and then moving to describe the state-of-the-art of metadata related to the submission process and the

metadata obtained from our manual curation step, in particular regarding the bioinformatic strategies used and the skin data characteristics retrieved directly from the published studies.

6.3.1 Comparison of datasets collection approaches and metadata retrieval

The automatic search with SRAdb recovered a total number of 97,182 samples from 203 studies (Data Frame 1) with 8,492 samples that were uploaded before 2012. The manual search, instead, recovered a total of 21,958 samples from 68 studies (Data Frame 3) starting from 2012.

We compared the ability of the two approaches in identifying the desired datasets. Notably, the automatic search failed to identify 47 studies that were recovered by the manual search, indicating that SRAdb does not perform an exhaustive search of the available datasets. The automatic search identified 182 studies not found by the manual search. Based on these observations we generated a data frame (Data Frame 2) that comprised both automatically retrieved and manually identified studies. This data frame contains 108,207 rows (samples) coming from 250 different studies and a total of 61 columns containing the metadata.

The metadata associated with the datasets can be differentiated into three major categories: i) metadata related to dataset submission (obtained by the automatic search), ii) metadata associated with the laboratory procedures and bioinformatic pipelines (obtained by automatic and manual searches) and iii) manually collected context metadata describing other relevant aspects of the study (e.g. disease/condition investigated or sample origin).

The automatic search for metadata with SRAdb and EDirect was performed for all the datasets, both manually and automatically retrieved, to collect metadata related to dataset submission (i). After the curation step, we conserved a total of 37 metadata columns that were included in all three data frames.

These 37 columns contain information related to:

- the study with BioProject, Study_ID, Study_description and Study_abstract;
- the submission and its date with the Year_of_release, Release_Date and Load_Date;
- the experiment with the Library Strategy used (Library_Strategy), specification on if it was performed a pair-end or a single-end sequencing (Library_Layout) and the library Insert size (Insert_Size);
- the sequencing platform and the model used (Platform, Model);
- the run with the average sequence length (AvgLength), the spots, the bases, the size of the file (Size_MB) and the path for the download (Download_path);
- the experiment title (Experiment_title);
- a description of its design (Design_description);
- the name of the library (Library_name) and attributes of the experiment (Experiment_attribute);
- the sample with BioSample, Sample_ID, Sample_alias, Sex, Body_Site, Description and Sample_attribute and
- the associated Taxonomic ID with the scientific name (TaxID, Scientific_Name).

A comprehensive description of all the 37 metadata is available in Supplementary File 1.

In Data Frame 2 and 3 we also included 23 additional columns that contain metadata not available on INSDC and obtained from the manual inspection of the publication. These metadata were recovered only for the manually retrieved datasets and contained information on the laboratory procedures and bioinformatic pipelines (ii) together with other relevant metadata describing the context of the study (iii).
In the next sections, all the categories of metadata and their distribution are outlined. A full description of the metadata included in the data frames is given

in Supplementary File 1, also available in our Github repository (https://github.com/giuliaago/SKIOMEMetadataRetrieval).

6.3.2 Distribution of metadata related to dataset submission and library preparation

By comparing the distribution of the number of datasets released over the years among the three different data frames (Figure 2b), we observed that Data Frame 1 showed a peak in 2015 when 17,551 datasets were released. Differently, Data Frame 2 showed a peak in 2017 with 19,041 datasets released during that year. For Data Frame 3, we observed two peaks: one in 2013 with 4,841 datasets released and one in 2017 with 7,293 datasets released. However, if we look at the number of studies, the peak was reached in 2019 with 16 studies investigating the human skin microbiome (Figure 2a).

After removing datasets with a value equal to zero for the following metadata, we calculated the median number of spots (sequencing clusters that generated sequence), bases (nucleotides), average read length and insert size (size of the amplicon without sequencing adapters) for Data Frames 1, 2, and 3. The median number of spots were respectively 23,590, 24,564, 22,560.5 (Figure 2e), while the median number of bases were 4,114,610, 4,364,032 and 7,270,396 (Figure 2f). The mean of the datasets' average read length in Data Frame 1 is 227.0235 bp, while for Data Frame 2 is 254.0603 bp and for Data Frame 3 is 440.2783 bp. The median values are 150 bp for Data Frame 1 and 2, and 502 bp for Data Frame 3 (Figure 2g and Figure 2i). The median insert size is 500 in Data Frame 1 and 2 and 300 in Data Frame 3 (Figure 2h and Figure 2i). Mean values are 455.5963, 440.2783 and 349.0783, respectively.

Information about the sex of the individuals can be collected for 36,231 out of 97,182 samples in Data Frame 1 (20,011 females; 16,220 males), 37,340 out of 108,207 samples in Data Frame 2 (20,234 females; 17,106 males), and 3,461 out of 21,958 samples in Data Frame 3 (1,276 females; 2,185 males).

We recognized 66 different descriptions (more or less accurate), defining the sampled region of the body. However, metadata on the body site is absent in most of the datasets. In detail, a total of 42,489 empty metadata information were found for Data Frame 1, 52,972 for Data Frame 2 and 18,061 for Data Frame 3.

In our data frames, we have observed the use of different Taxon IDs to describe the samples. Data Frame 3, which contains only samples of human skin microbiome, presents 11 different taxon IDs, which correspond to the following scientific names: "human skin metagenome", "Homo sapiens", "metagenome", "metagenomes", "human metagenome", "skin metagenome", "Staphylococcus aureus", "clinical metagenome", "gut metagenome", "human gut metagenome" and "bacterium". The number of Taxon IDs increases in the other two data frames so that in Data Frame 2 we observe 173 different Taxon IDs.

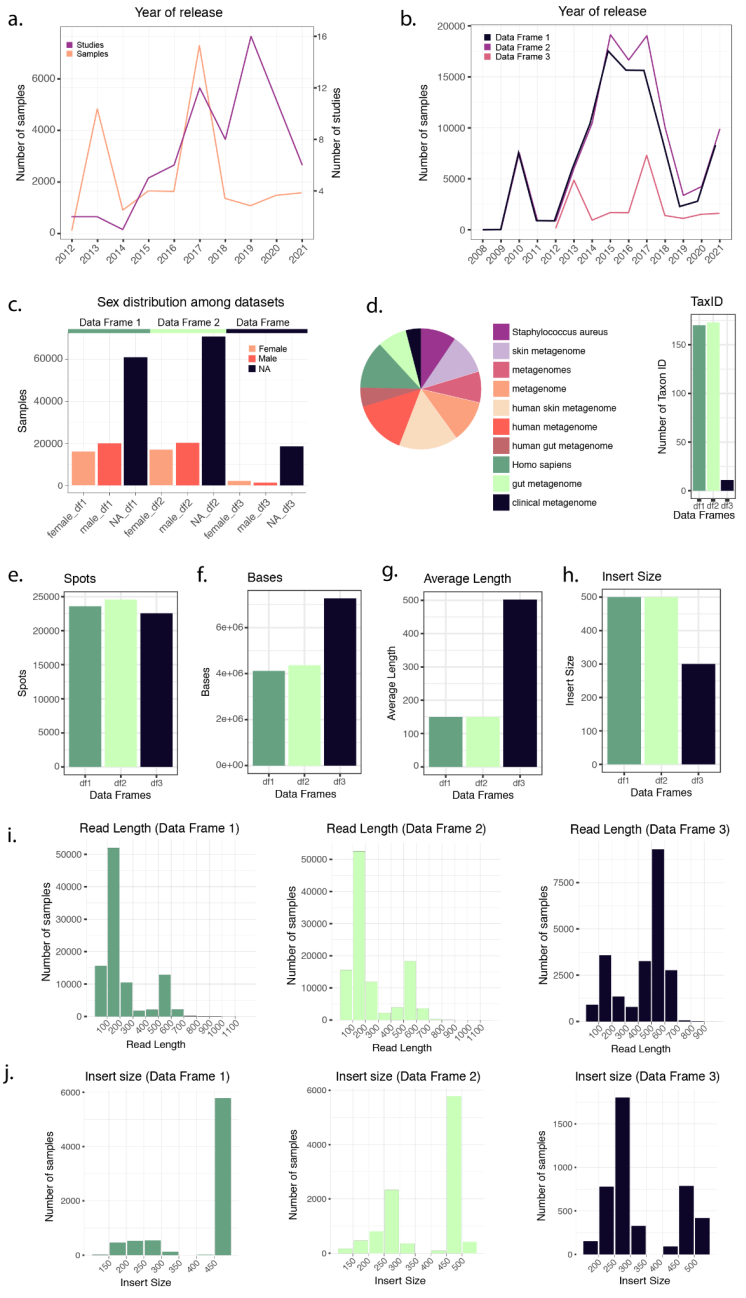

Figure 2. **a**) Number of studies and samples from Data Frame 3 released every year starting from 2012. **b**) Comparison of the number of samples released each year for the three Data Frames (Data Frame 1 in blue, Data Frame 2 in black and Data Frame 3 in red). Data Frames 1 and 2 contain

samples starting from 2008, while Data Frame 3 only from 2012. **c**) Distribution of the variable "sex" in the three Data Frames. In all three cases, the majority of the samples don't have such information reported. **d**) The number of Taxon ID/Scientific names used in the three Data Frames (barplot) and relative abundance (as a logarithm) of the Taxon ID/Scientific names used for the samples in Data Frame 3 (pie chart). **e-h**) Comparison of the median number of spots (**e**), bases (**f**), reads average length (**g**) and insert size (**h**) in the three Data Frames. **i**) Read length distribution in the three Data Frames. **j**) Distribution of the insert size in the three Data Frames.

6.3.3 Methodological pipeline insights and context-metadata of skin microbiome datasets

For the 68 manually retrieved studies we further collected other metadata from the associated publications. Based on these manually collected metadata, we observed that most of the studies had used swabs to collect samples (53 studies; 19,928 samples), with only a few relying on other methods like biopsies (5 studies; 257 samples), scrubs buffer washes (1 study; 1,358 samples) or a combination of swabs and other methods (7 studies; 311 samples).

Considering the marker gene used, the most commonly sequenced hypervariable regions of the 16S rRNA gene have been the V1-V3 (6,176 samples), followed by the V4 (5,694) (Figure 3a). However, if we consider the number of studies, we observed that most of them relied on the V1-V3 (24 studies) and V3-V4 (21 studies) regions (Figure 3a). The Illumina sequencing platforms were the most used (88,295 samples in Data Frame 2), particularly the Illumina Miseq platform (49,297 samples in Data Frame 2), followed by Roche 454 platform (19,777 samples in Data Frame 2). A total of 11,412 samples have no specific platform model assigned (Figure 3c).

Regarding the bioinformatic pipeline used, most of the manually inspected studies have clustered reads into Operational Taxonomic Units (OTUs) (56 studies), only a few (6 studies) relied on Amplicon Sequence Variants (ASVs) or

Ribosomal (35) Sequence Variants (RSVs). For 6 studies this information was not reported in the article methods (Figure 3d).
Taxonomy assignment was mainly performed with Greengenes database (McDonald et al., 2012) (29 studies), followed by SILVA (Quast et al., 2013) (15 studies). Other works relied on different databases, including RDP (Cole et al., 2014) (3 studies), EzTaxon-e (Kim et al., 2012) (3 studies), NCBI (1 study), and HOMD (1 study). Strikingly, many studies did not report this information in the articles' method section (16 studies) (Figure 3e).

Our analysis also comprehended a detailed inspection of skin and disease conditions related to the microbiome analysis. Among our list, we identify 42 studies investigating 26 different diseases/conditions of the skin (Figure 3b). The most commonly investigated disease in our curated dataset is atopic dermatitis (8 studies), followed by psoriasis and parapsoriasis (5 studies), while 7 studies investigated skin injuries of different kinds. Among the other diseases/conditions investigated, we observed acne (3 studies), skin pathogenic infections, such as bacterial, fungal and parasitic infection (3 studies), allergic traits and atopic individuals (3 studies), dandruff (2 studies), leprosy (2 studies), hidradenitis suppurativa condition (2 studies), autoimmune bullous disease (1 study), dystrophic epidermolysis bullosa (1 study), vitiligo (1 study), squamous cell carcinoma (1 study), filaggrin-deficient human skin (1 study), and other conditions such as obesity and low birth weight (2 studies). Overall, 26 studies collected samples from healthy human skin (in Data frame 3, column 43 'disease/condition').

Looking at the geographic distribution of the studies, we observed that most of them were conducted in the USA (22 studies), followed by European countries (19 studies) and China (11 studies). Other countries that featured more than one study were South Korea (4 studies), Brazil (3 studies) and India (2 studies) (in Data frame 3, column 44 'Location') (Figure 3.g).
Finally, the 68 manually retrieved studies were published in 40 different

scientific journals from 17 different WOS research areas. According to Scopus classification, 36 studies were published in medicine-related scientific journals (Research Subject = Medicine). Figure 3f shows how often specific WOS Research areas and Scopus Research Subjects are associated with the Scopus research subject “Medicine” in the present dataset.

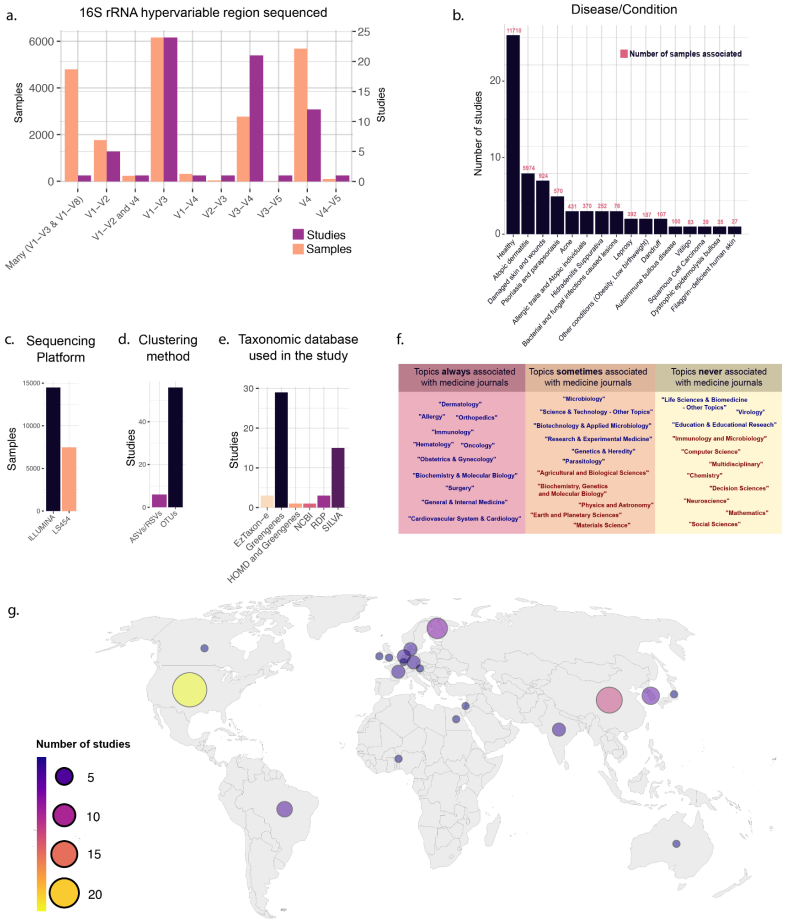

Figure 3. **a)** Number of samples (pink) and studies (purple) that used specific 16S rRNA hypervariable regions in Data Frame 3. **b)** The number of studies and samples for each disease/condition investigated in Data Frame 3. **c-e)** Frequency of use of the different sequencing platforms (**c**), clustering methods (**d**) and taxonomic databases (**e**) in Data Frame 3. **f)** Table showing the Web Of Science research areas (blue) and Scopus Research Subjects (red) that described the scientific journals in which the studies of Data Frame 3 have been published. The research areas/subjects are divided into three boxes depending on how often they were associated with the Scopus research subject "Medicine". Going from left to right are shown the research areas/subjects that were always (left), sometimes (center) and never (right) associated with the Scopus research subject "Medicine". **g)** Geographical distribution of the studies included in Data Frame 3.

6.4 Discussion

In this section we discuss the results obtained from our work, in particular focusing the attention on three main aspects: i) outcomes related to dataset collection, ii) caveats related to metadata retrieval and data reuse and, finally, iii) the importance of having a curated collection of a microbiome dataset for advancing the microbiome research field through data-driven approaches and powerful meta-analysis.

6.4.1 Skin microbiome data retrieval: dataset collection is not an easy task

The INSDCs databases are the source of an enormous amount of publicly available datasets which can be accessed and downloaded to perform powerful meta-analyses (Arita et a., 2021). The field of microbiome research can greatly benefit from the availability of this large amount of data (Wood-Charlson et al., 2020). However, the reusability of a dataset strictly depends on the possibility of retrieving it and on the amount of information (metadata) deposited by the authors at the time of submission (Gonçalves and Musen, 2019; Miron et al., 2020).

If the number of datasets available is limited (such as for poorly studied environments), a manual search will consent to gather all the studies available

in a relatively fast way. However, for well-studied environments, the number of datasets can be very large and it becomes more convenient to rely on automatic approaches (Baumgartner et al., 2007). The automatic approach allows for a fast and comprehensive search of datasets of interest, but at the same time, it lacks a curation step that validates the recovered datasets. Moreover, the automatic search does not permit the retrieval of important information that was not deposited in the INSDC databases together with the raw data. Conversely, the manual search is more accurate and allows a researcher to retrieve a well-validated list of studies together with other information by inspecting the associated publication. Its drawbacks are that it is time-consuming and presumably less comprehensive than the automatic search. Moreover, it does not consent to retrieve sample-specific information.

Our results showed that the automatic search did find a greater number of datasets than the manual (97,182 samples from 203 studies vs 21,958 samples from 68 studies). Many can be the reasons that explain this difference. First, the automatic search tends to be more exhaustive than a manual one if the number of available datasets is large. Second, the list of studies is not inspected to remove undesired studies that do not match some of the desired criteria but might be retrieved by the searching tool. Third, the manual search was limited to the dataset deposited in the last 10 years, starting from 2012, while the automatic search recovered studies starting from 2008. Indeed, 8,492 samples found by the automatic search were uploaded before 2012. Despite these observations, neither the manual nor the automatic search with SRAdb, were capable of recovering all the studies, highlighting the importance of combining the two approaches.

Together, our results indicated that SRAdb was not exhaustive in its search, and to maximize the number of datasets retrieved, a combination of manual and automated approaches might represent the optimal strategy. We observe that the larger the number of available datasets, the less feasible an extensive manual search, favoring an automated approach for the dataset retrieval step.

Conversely, for topics with a particularly small number of datasets available, the manual search still remains the most accurate way of recovering them.

6.4.2 Caveats of metadata retrieval and data reuse

Depending on the topic, a researcher interested in performing a meta-analysis can decide to rely on different approaches to retrieve metadata associated with the datasets of interest, both directly through the INSDC data portal (Arita et al., 2021) or with specific tools (Zhu et al., 2013; Kans, 2021; Eaton, 2020). In this work, we decided to combine three approaches, based on SRAdb (Zhu et al., 2013), Entrez (Kans, 2021) plus a manual search from the publication, with the aim of generating a comprehensive data frame containing all the datasets from the human skin microbiome amplicon sequencing available on INSDC databases. As for the search of the datasets, also for metadata retrieval, we observed that the combination of automatic and manual approaches is capable of gathering a larger amount of information than the two approaches alone.

However, while with a manual search it is possible to recover much information related to a dataset if a publication is available, this approach is not feasible if the number of datasets is high (Baumgartner et al., 2007). Moreover, sample-specific information for large datasets can only be collected using automatic approaches, making an automatic search a necessity.

Automatic approaches of metadata retrieval (such as those used in this study) collect the metadata deposited on the INSDC databases. As such, they are capable of accessing only the metadata that were made available by the researchers during the data submission. Failing in accessing specific metadata can aect the re-usability of a given dataset, highlighting the importance of proper and extensive metadata storage.

We recognized three major causes that aect the reusability of publicly available microbiome datasets: 1) **Missing metadata**. A lot of essential metadata are simply not available either because not included among the requested metadata or because not mandatory and hence not compiled by the submitter. One example is the absence of metadata specifying the 16S rRNA hypervariable

region amplified and sequenced for most of the studies, which seriously compromise data harmonization eorts. Another information that is often not reported is the presence of an associated publication. The availability of the raw reads on public databases is a requirement for publication in many scientific journals. During the raw reads submission, the researcher is required to provide metadata associated with the dataset, including the presence of a publication. As such, since this step predates the publication itself most of the datasets are uploaded without specifying this information. 2) **Metadata wrongly assigned**. Sometimes metadata can be wrongly assigned to the samples. This can also be the result of mandatory metadata fields that are ambiguous and can lead a researcher inexperienced in the submission process to compile the field in an incorrect way. Wrong metadata can cause the inclusion of wrong datasets into an analysis, potentially aecting the results and leading to incorrect biological conclusions, or, conversely, they can cause the exclusion of datasets from analyses in which they would have fitted. As an example, by comparing the metadata deposited on INSDC with what was reported in the publication we were able to identify studies that wrongly assigned the library strategy as “RNA-Seq” and “WGS” instead of “AMPLICON”.

3) **Inconsistency of the used terminology**. Some metadata fields can be filled with multiple correct metadata leading to inconsistency in the terminology used and aecting the possibility of automatizing the search and filtering of datasets based on these metadata. Good examples are the numerous Taxon ID and scientific names associated with the samples, which are not necessarily wrong, but the lack of consistency in the terms used compromises the usefulness and value of this metadata.

Different works demonstrated the caveats of metadata retrieval and its consequences (Gonçalves and Musen, 2019; Bernasconi, 2021; Jurburg et al., 2020). Researchers have undertaken different approaches to ameliorate this step, in particular using a manual or automated/semi-automated curation (Klie et al., 2021), or developing tools specific for the download of metadata information (Hoarfrost et al., 2019). Most of the automated or semi-automated

methods are based on Natural Language Processing (NLP) techniques, used to recognize predefined entities in unstructured text, in order to retrieve metadata from the text associated with the samples. Others try to normalize metadata information by grouping or mapping to ontologies (Bernstein et al., 2017; Hu et al, 2017; Martinez-Romero et al., 2019). These methods still need a revised step of manual curation and sometimes cannot reconstruct the totality of the metadata associated (Klie et al., 2021). As we demonstrated before, manual curation seems the most accurate solution (Klie et al., 2021; Wang et al., 2019) if data remains human-readable.

Considering the microbiome field, the INSDC significantly contributed with a recent perspective paper describing the steps that the microbiome research community should take to favor data FAIRification and metadata incorporation (Vangay et al., 2021). As microbiome samples are particularly related to the context in which they were collected, data describing measurements or variables related to the context are critical (Vangay et al., 2021). Two main subject areas were indicated by the INSDC to improve data standards: i) promote microbiome data sharing and ii) try to remove obstacles and diculties related to data and metadata submission. Some of their observations and proposals are currently applied by the research community, as for example the “Minimum Information about any (x) Sequence” (MIxS) packages (Yilmaz et al., 2011) or the incorporation of DOIs for datasets (Cousijn et al., 2018). Unfortunately, some work is still needed to establish standard procedures and a universal set of ontologies that are easily accessible by the entire community (Vangay et al., 2021; Buttigieg et al., 2016).

In this context, this work also wants to disclose the situation of a sub-field of the microbiome data world: the skin microbiome. The issues revealed by our results show that the search and secondary use of the datasets is still not easy to achieve. Since different studies can rely on different methodologies, different datasets might not be directly comparable and precautions must be taken before combining multiple datasets in a meta-analysis. Without some metadata, a potentially valid dataset can not be included in a meta-analysis. Therefore it is

essential for a researcher that wants to valorize a dataset to upload as much information as possible together with the raw reads so as to make the dataset reusable. To motivate researchers in uploading more information, the submission procedure should be made as simple and guided as possible, also to avoid misinterpretations and wrong metadata assignments. To reduce the missingness of metadata, more fields should be made mandatory, such as those referred to the 16S rRNA region sequenced, and new metadata should be included, such as a field that easily discriminates biological samples from negative controls. It also urges the need for standardization of the Taxon ID used in microbiome studies. Guidelines should be given to avoid the use of imprecise Taxon IDs. Efforts should also be made to associate a link to the publication whenever it becomes available, to allow for easier and straightforward access to this resource.

As we have stated, numerous are the aspects related to data and metadata submission that can be improved. Some relate to the submission process itself which can be refined to favor microbiome data reusability, while others strictly depend on the commitment of the researcher performing the submission, who should not overlook the relevance of this step and its importance for the whole scientific community.

6.4.3 The value of a curated skin microbiome collection

Over the past decade, researchers have explored the intricate ecosystem of the skin microbiome (Byrd et al., 2018), unveiling the interactions between the microbiome players (bacteria, archaea, fungi and viruses), the skin cells, and the host immune cells that act as barriers, constituting a defense against pathogens invasion and inflammation (Byrd et al., 2018; Prescott et al., 2017). Perturbations in the skin ecosystem can cause an unbalance that can even lead to the rising of immune disorders, like allergies, dermatitis or eczema, or chronic injuries, like ulcers. Determining the causes and eects of these processes is not an easy task. Traditional approaches to study skin microbiome mechanisms relies on culture-based techniques, leading to an underestimation of the actors

and a bottle-neck selection due to the strict range of cultivable species. The case of Staphylococcus genus can serve as an example. Being more easily cultivable than microorganisms belonging to Corynebacterium spp. or Propionibacterium spp., it would dominate a microbiome dataset, leading to an underestimation of the real biodiversity (Kong and Segre, 2021). It became obvious that to overcome culture-dependent bottlenecks and to explore the skin microbiome as a whole, a sequencing method must be applied (Byrd et al., 2018).

In this context, large-scale sequencing data enable microbiology researchers to obtain deep insights in genetic and functional profiling (Byrd et al., 2018) and, nowadays, grand challenges in microbiome science rely on large-scale data science approaches (Kyrpides et al., 2016). Secondary analysis can be full of potential and by-passing the need of generating new large datasets can enormously reduce the costs associated with this kind of study. Impactful meta-analyses have already contributed to advancing the microbiome field, as demonstrated by numerous studies (Duvallet et al., 2017; Bisanz et al., 2019; Kosti et al., 2020).

From the more applied and clinically relevant studies of skin health and disease to the more theoretical works investigating microbial ecology and the holobiont evolution, all these sub-fields of microbiome research will benefit from the adoption of data-driven approaches based on large-datasets integration (Ross et al., 2018). The availability of a curated collection of microbiome datasets represents the required starting point to make this transition possible and scalable (Wood-Charlson et al., 2020; Vangay et al., 2021).

Currently, numerous research teams around the world have put eorts in trying to collect and harmonize data from different microbiome fields and various curated collections of microbiome datasets have been published, like the TerrestrialMetagenomeDB (Correa et al., 2020), the HumanMetagenomeDB (Kasmanas et al., 2021), or the Planet Microbe (Ponsero et al., 2021). Each one of these collections is focused on a specific topic and sometimes on a specific

type of data and aims at providing each microbiome research sub-field with a valuable resource to perform data-driven meta-analyses.

Based on these premises and focusing on the skin microbiome sub-field, our work resulted in a comprehensive list of human skin microbiome datasets enriched with metadata information related to the methodological pipelines and the context of the dataset under study.

Skin research produces large quantities of data using a wide range of methods and equipment that require large collaborative eorts. These research endeavors span a broad range of disciplines and are critical to investigating the skin physiology, functions, interactions and health status, from a broad perspective. This can be seen in the bibliometric analysis of published papers related to the datasets retrieved. Research areas and categories from the Web of Science collection and Elsevier's Scopus classifications showed a scattered distribution of publications in different research areas, but with a higher proportion related to the medicine-related area. As the number of studies grows, it clearly appears that crossing the boundary between medicine and microbial ecology is the lynchpin for a deep understanding of skin health (Callewaert et al., 2020; Prescott et al., 2017). Indeed, a consistent proportion of the data collected is dedicated to disease conditions, providing valuable material for clinical researchers, but also for microbial ecologists and researchers from other fields of research interested in studying the microbial dynamics in the skin ecological niche. Moreover, taken together, more than half of the studies in our Data Frame 3 collected microbiome data from healthy subjects, providing an invaluable source of information. One of the main challenges for data harmonization is to link the phylogenetic diversity of host-associated microbes to their functional roles within the community and with the host. Much remains to be learned about us as holobionts and much of the information is still kept inside the data.

The curated list we generated can serve as a most comprehensive collection of datasets that can be searched and queried to identify datasets of interest.

Researchers interested in conducting meta-analyses with human skin microbiome datasets can use these data frames as a starting point to recover the dataset more suited for their analyses. As demonstrated by the presence of errors in the metadata, these data frames require a curation step. Here, we reported a curated data frame (Data Frame 3) in which we manually corrected errors in the metadata. We also reported two non-curated data frames obtained with the automatic search (Data Frame 1) and with a combination of manual and automatic search (Data Frame 2). These two data frames contain a greater number of studies and samples, however, a careful inspection of these datasets is advised before including any one of those into a meta-analysis.

6.5 Conclusions

The aim of our effort was to help accelerate human skin microbiome research by reducing the amount of time needed to search for datasets and metadata of interest and at the same time favoring data reuse by maximizing the amount of information associated with each dataset. Here we report three data frames containing a comprehensive collection of human skin microbiome datasets enriched with metadata recovered from different sources. The data frames are easily explorable and can be useful for researchers interested in conducting meta-analyses with human skin microbiome amplicon data.

Furthermore, we demonstrated that the reusability of a dataset depends on the amount of information that can be gathered on the dataset itself, that is the amount of metadata deposited by the authors at the time of submission. We are aware that data sharing is increasing throughout the microbiome community, but there are still barriers to making microbiome data truly FAIR. Metadata standards exist, but their proper adoption by the research community is still lagging, as also demonstrated by the NMDC community.

Skin microbiome sampling has the advantage of being non-invasive, easily accessible, and able to provide a huge amount of meaningful information. A

curated collection of skin microbiome datasets, enriched with study-related metadata, could be used to investigate health-related phenotypes, oering the potential for non-invasive diagnosis and condition monitoring. Our framework sets the stage for new analyses implementing AI approaches focused on understanding the complex relationships between microbial communities and phenotypes, to predict any condition from microbiome samples. Indeed, considering the skin microbiome topic, a few, very recent works included data integration strategies and AI applications (Marcos-Zambrano et al., 2021; Jaiswal et al., 2021; Carrieri et al., 2021), showing the potential held by these approaches in advancing skin microbiome research.

As the microbiome research field is headed to become a science founded on big-data, the necessity of developing standardized procedures to generate and analyze data acquires importance. The adoption of standard methodologies will help future data integration efforts for the benefit of the whole research community. For this reason, we advocate for a concerted effort to favor standardized microbiome research and exhaustive data sharing.

Further, with this work we want to build a foundation that places microbiome research at the nexus of many subdisciplines within and beyond biology, as for example dermatology, medicine and microbial ecology.

For this reason, this project has the potential to accelerate the development of microbiome-based personalized medicine and non-invasive diagnostics.

6.6 Supplementary

Github repository will be available upon request. The pipeline is available in my Github repository (https://github.com/giuliaago). In particular, here I provide the list of the files with a description:

- Supplementary_file_1.csv: Comprehensive list of the manually curated metadata with description

- README.txt: README of our Github repository
- skiome_workflow-01.png: Figure representing the metadata retrieval workflow
- SKIOME_pipeline.Rmd: Complete pipeline for dataset and metadata retrieval
- Human_Skin_Datasets_Manual_Search.csv: Result of the manual search for the datasets and metadata
- data_frames.7z: Compress file containing the three dataframes obtained from our work

7. Conclusions

DNA metabarcoding has great potential in several research fields of application. However, some steps must be disentangled to exploit it fully. As I described within **Chapter 1**, the reason beyond its success is the great compromise between costs and benefits: DNA metabarcoding is still the best approach to collect wide amount of data at several depth of investigation and transferability on lot of matrices, from food, water, pollens, besides of course to human samples. In this dissertation, I presented four main issues that can be ameliorated: i) use molecular information as a main source of information when non-bacterial molecular markers are used, ii) issues related to the taxonomy assignment and development of strategies to enhance it, iii) pattern reconstruction via data mining methods and iv) public data as a valuable resource for meta-analysis and data integration projects. All the work done keeps in mind an important message and a way of strategy: a data-centered perspective and application.

Considering the taxonomy assignment, NCBI is still the widest collector of molecular data, as also depicted in **Chapter 2**. Of course, specific reference databases exist, but are not always updated and easily available. With ExTaxsI (**Chapter 4**), we provide an easy-to-use standalone tool able to interact with NCBI databases and personal datasets, offering instruments to standardize taxonomy information and visualize vast quantities of data distributed on different taxonomic levels. Visualization plots are also included, easily shareable through HTML formats. ExTaxsI may help researchers involved in environmental genomics fields, from phylogeographic studies to DNA metabarcoding surveys, and also in projects related to human health, as demonstrated with SARS-CoV-2 case study.

In our opinion, ExTaxsI data management and its visual interactive exploration can really improve the experimental design phase and the awareness of the

information available, facilitating the work and incentivizing data exploration and sharing.

The issues related to taxonomy assignment were also depicted through the case study of EXPO2015 (**Chapter 3**). Limitations in managing data were considered and, in addition, both markers chosen and the molecular information registered into the reference databases were discussed. Despite these issues, I demonstrated that the power of DNA metabarcoding is related not only to the molecular fingerprint obtained with sequencing features, but also to the ability to collect large amounts of data, achieving a sort of freeze frame of the environment under study. Moreover, I firmly believe that collecting new information and submitting datasets to reference databases is mandatory to ameliorate the comprehension of biodiversity all around the world, implementing both our current knowledge and future research.

This last point introduces another topic of this dissertation: the data reuse strategy. With the SKIOME Project (**Chapter 6**), we demonstrated that the reusability of a dataset depends on the amount of information that can be gathered on the dataset itself. Considering microbiome data, specifically, data sharing is increasing throughout the microbiome community, making them reusable, ensuring meta-analysis and data integration practices. The main results of this work were three data frames containing a comprehensive collection of human skin microbiome datasets enriched with metadata recovered from different sources. Moreover, issues related to metadata enrichment were highlighted: we still observed pitfalls to make microbiome data truly FAIR. Metadata standards exist, but their proper adoption by the research community is still lagging, as also demonstrated by the NMDC community. Of course, in recent years the microbiome community made a great effort, and for some research areas great results can be observed. However, a lot of work must be done to fully organize and exploit microbiome data.

Besides this, microbiome data analysis has acquired more and more standardized procedures, so that researchers shifted the focus towards post-processing analysis, demonstrating the potentials that 16S rRNA metabarcoding data may offer to answer complex biological questions.
One of the areas most under investigation is the identification of species association and interaction. Moving towards the co-occurrence reconstruction, several tools and methods have been developed to study microbiome patterns.
Association Rule Mining is a famous and widely used supervised machine learning technique to calculate patterns and rule between them. With the work of microFIM (**Chapter 5**), I tried to disentangle this type of analysis for microbiome pattern exploration. In particular, I depicted issues that I think must be considered before using an ARM approach on 16S rRNA metabarcoding data, specifically. To report them briefly: i) consciously setting the input data to be analysed, ii) consider the use of metadata to filter the data before starting, iii) define what is interesting for the specific case study, iv) consider computational time, v) use suited tools and visualization strategies and, finally, vii) setting evaluation and benchmarking strategies to retrieve sound results. This last point is also strictly connected to make associations an evidence of causality, an important step forward for the microbiome pattern research field.

All the works presented in this dissertation explore different aspects of metabarcoding data. In this work, data can be used to create new knowledge, to perform new and meta-analyses and to define biological associations.
In the last years, metagenomic and sequencing approaches have drastically changed the way to study any type of biological problem. We collect any type of data and we integrate different information to answer biological questions. The future goals of these fields will not only be how to collect and analyse, but also how to extrapolate, to integrate, to enhance the value of the data that we have accumulated over the years.

To really exploit these potentials we will surely have to implement the following steps: i) the development of new strategies that ameliorate the visualization of high intricate data; ii) the creation of user-friendly interfaces as important resources to transfer new technologies to a wider audience; iii) the integration of metabarcoding data, in particular from microbiome, as it has the greater audience and the most standardized framework, to functional properties, in order to fully exploit the metabarcoding technique; iv) the organization of data to guarantee a fluid and direct investigation of data collections, in order to pave the way for reproducibility, re-analysis, meta-analysis and data integration projects; and v) great efforts to make also other molecular markers a standard to be fully exploited. In this PhD dissertation, some steps have been taken, depicting new solutions and potential obstacles that must be overcome. However, other efforts are needed: interdisciplinarity and cooperation will be fundamental to address the issues discussed and create new synergies.

One of the points that has not been addressed here but certainly of significant importance is to transfer this knowledge to citizens, patients and companies: the dissemination of the use of techniques such as DNA metabarcoding will be fundamental to expand their usage and, consequently, develop stronger frameworks that would ensure transferability and reproducibility.

Moreover, as also demonstrated by the COVID-19 pandemic that we have just faced, data has a great value in all the research contexts. With the work of ExTaxsI, in which we analysed the availability of information of genes related to SARS-CoV-2 virus, data may reveal new insights and the use of metadata associated with it may integrate important information to the data acquired. However, interpreting data with awareness is now fundamental, especially considering the way in which data are collected and submitted. The access and availability of data and metadata associated is fundamental to organize the information in our repositories. With the SKIOME Project, issues related to metadata accessibility were presented and, currently, only with a manual

curation approach the access was guaranteed. Of course, other automatic strategies can be implemented, but the path is not straightforward, as several tests and benchmarks must be done to reconstruct metadata associated when they are not available.

In general, as sequencing technologies have great potential in several research fields, showing and integrating a data-centered perspective in biological projects will be essential to make researchers and partners aware about the complexity of all the work behind the world of data. Thanks to dissemination and education, both citizens and researchers would foster future data collection, thus making the researchers' work fully useful and usable, expanding real and concrete applications.

References

Afshari, R., Pillidge, C. J., Read, E., Rochfort, S., Dias, D. A., Osborn, A. M., & Gill, H. (2020). New insights into cheddar cheese microbiota-metabolome relationships revealed by integrative analysis of multi-omics data. Scientific Reports, 10(1), 1–13.

Agapito, G., Guzzi, P. H., & Cannataro, M. (2015). DMET-Miner: Efficient discovery of association rules from pharmacogenomic data. Journal of Biomedical Informatics, 56, 273–283.

Agostinetto, G., Brusati, A., Sandionigi, A., Chahed Adam, Parladori Elena, Bachir, B., Antonia, B., Dario, P., & Maurizio, C. (2021). Supporting data for "ExTaxsI: An exploration tool of biodiversity molecular data" (p. 1 GB) [Data set]. GigaScience Database.

Agostinetto, G., Sandionigi, A., Chahed, A., Brusati, A., Parladori, E., Balech, B., Bruno, A., Pescini, D., & Casiraghi, M. (2020). ExTaxsI: an exploration tool of biodiversity molecular data. BioRxiv.

Agrawal, R., Imieliński, T., & Swami, A. (1993). Mining association rules between sets of items in large databases. Proceedings of the 1993 ACM SIGMOD International Conference on Management of Data - SIGMOD '93, 207–216.

Agrawal, R., Mannila, H., Srikant, R., Toivonen, H., & Verkamo, A. I. (1996). Fast discovery of association rules. Advances in Knowledge Discovery and Data Mining, 12(1), 307–328.

Akaike, M., Miyagawa, H., Kimura, Y., Terasaki, M., Kusaba, Y., Kitagaki, H., & Nishida, H. (2020). Chemical and bacterial components in sake and sake production process. Current Microbiology, 77(4), 632–637.

Alberdi, A., Aizpurua, O., Gilbert, M. T. P., & Bohmann, K. (2018b). Scrutinizing key steps for reliable metabarcoding of environmental samples. Methods in Ecology and Evolution, 9(1), 134–147.

Al-Juhani, W. S. (2019). Evaluation of the capacity of the DNA barcode ITS2 for identifying and discriminating dryland plants. Genetic Molecular Research, 18(1), 32–46.

Almeida, A., Mitchell, A. L., Boland, M., Forster, S. C., Gloor, G. B., Tarkowska, A., Lawley, T. D., & Finn, R. D. (2019). A new genomic blueprint of the human gut microbiota. Nature, 568(7753), 499–504.

Alves, R., Rodriguez-Baena, D. S., & Aguilar-Ruiz, J. S. (2010). Gene association analysis: A survey of frequent pattern mining from gene expression data. Briefings in Bioinformatics, 11(2), 210–224.

Amos, G. C. A., Logan, A., Anwar, S., Fritzsche, M., Mate, R., Bleazard, T., & Rijpkema, S. (2020). Developing standards for the microbiome field. Microbiome, 8(1), 98.

Anaconda Software Distribution. (2020). Anaconda Documentation. Anaconda Inc. Retrieved from https://docs.anaconda.com/

Andersen, K. G., Rambaut, A., Lipkin, W. I., Holmes, E. C., & Garry, R. F. (2020). The proximal origin of SARS-CoV-2. Nature Medicine, 26(4), 450–452.

Anderson, M. J. (2001a). A new method for non-parametric multivariate analysis of variance. Austral Ecology, 26(1), 32–46.

Anderson, M. J. (2001b). Permutation tests for univariate or multivariate analysis of variance and regression. Canadian Journal of Fisheries and Aquatic Sciences, 58(3), 626–639.

Ankenbrand, M. J., Keller, A., Wolf, M., Schultz, J., & Förster, F. (2015). ITS2 database V: Twice as much. Molecular Biology and Evolution, 32(11), 3030–3032.

ANSES OPINION of the French Agency for Food, Environmental and Occupational Health & Safety on “the use of insects as food and feed and the review of scientific knowledge on the health risks related to the consumption of insects” | Anses - Agence nationale de sécurité sanitaire de l’alimentation, de l’environnement et du travail. (n.d.). Retrieved April 22, 2020, from https://www.anses.fr/en/content/opinion-french-agency-food-environmental-and-occu pational-health-safety-use-insects-food-and.

Arita, M., Karsch-Mizrachi, I., Cochrane, G., & on behalf of the International Nucleotide Sequence Database Collaboration. (2021). The international nucleotide sequence database collaboration. Nucleic Acids Research, 49(D1), D121–D124.

Aslam, H., Green, J., Jacka, F. N., Collier, F., Berk, M., Pasco, J., & Dawson, S. L. (2020). Fermented foods, the gut and mental health: A mechanistic overview with implications for depression and anxiety. Nutritional Neuroscience, 23(9), 659–671.

Baiano, A. (2020). Edible insects: An overview on nutritional characteristics, safety, farming, production technologies, regulatory framework, and

socio-economic and ethical implications. Trends in Food Science & Technology, 100, 35–50.

Baird, D. J., & Hajibabaei, M. (2012). Biomonitoring 2.0: A new paradigm in ecosystem assessment made possible by next-generation DNA sequencing: NEWS AND VIEWS: OPINION. Molecular Ecology, 21(8), 2039–2044.

Balech, B., Sandionigi, A., Manzari, C., Trucchi, E., Tullo, A., Licciulli, F., Grillo, G., Sbisà, E., De Felici, S., & Saccone, C. (2018). Tackling critical parameters in metazoan meta-barcoding experiments: A preliminary study based on coxl DNA barcode. PeerJ, 6, e4845.

Bálint, M., Bahram, M., Eren, A. M., Faust, K., Fuhrman, J. A., Lindahl, B., O'Hara, R. B., Öpik, M., Sogin, M. L., Unterseher, M., & Tedersoo, L. (2016). Millions of reads, thousands of taxa: Microbial community structure and associations analyzed via marker genes. FEMS Microbiology Reviews, 40(5), 686–700.

Ballmer-Weber, B. K., Holzhauser, T., Scibilia, J., Mittag, D., Zisa, G., Ortolani, C., Oesterballe, M., Poulsen, L. K., Vieths, S., & Bindslev-Jensen, C. (2007). Clinical characteristics of soybean allergy in Europe: A double-blind, placebo-controlled food challenge study. Journal of Allergy and Clinical Immunology, 119(6), 1489–1496.

Banchi, E., Ametrano, C. G., Stanković, D., Verardo, P., Moretti, O., Gabrielli, F., Lazzarin, S., Borney, M. F., Tassan, F., & Tretiach, M. (2018). DNA metabarcoding uncovers fungal diversity of mixed airborne samples in Italy. PloS One, 13(3), e0194489.

Baumgartner, W. A., Cohen, K. B., Fox, L. M., Acquaah-Mensah, G., & Hunter, L. (2007). Manual curation is not sufficient for annotation of genomic databases. Bioinformatics (Oxford, England), 23(13), i41–i48.

Bayraktarov, E., Ehmke, G., O'connor, J., Burns, E. L., Nguyen, H. A., McRae, L., Possingham, H. P., & Lindenmayer, D. B. (2019). Do big unstructured biodiversity data mean more knowledge? Frontiers in Ecology and Evolution, 6, 239.

Bellati, A., Tiberti, R., Cocca, W., Galimberti, A., Casiraghi, M., Bogliani, G., & Galeotti, P. (2014). A dark shell hiding great variability: A molecular insight into the evolution and conservation of melanic Daphnia populations in the Alps. Zoological Journal of the Linnean Society, 171(4), 697–715.

Belluco, S., Losasso, C., Maggioletti, M., Alonzi, C., Ricci, A., & Paoletti, M. G. (2015). Edible insects: A food security solution or a food safety concern? Animal Frontiers, 5(2), 25–30.

Bengtsson-Palme, J., Hartmann, M., Eriksson, K. M., Pal, C., Thorell, K., Larsson, D. G. J., & Nilsson, R. H. (2015). metaxa 2: Improved identification and taxonomic classification of small and large subunit rRNA in metagenomic data. Molecular Ecology Resources, 15(6), 1403–1414.

Beninde, J., Veith, M., & Hochkirch, A. (2015). Biodiversity in cities needs space: A meta-analysis of factors determining intra-urban biodiversity variation. Ecology Letters, 18(6), 581–592.

Bernasconi, A. (2021). Data quality-aware genomic data integration. Computer Methods and Programs in Biomedicine Update, 1, 100009.

Bernstein, M. N., Doan, A., & Dewey, C. N. (2017). MetaSRA: Normalized human sample-specific metadata for the Sequence Read Archive. Bioinformatics, 33(18), 2914–2923.

Berry, D., & Widder, S. (2014). Deciphering microbial interactions and detecting keystone species with co-occurrence networks. Frontiers in Microbiology, 5.

Bharti, R., & Grimm, D. G. (2021). Current challenges and best-practice protocols for microbiome analysis. Briefings in Bioinformatics, 22(1), 178–193.

Biella, P., Tommasi, N., Akter, A., Guzzetti, L., Klecka, J., Sandionigi, A., Labra, M., & Galimberti, A. (2019). Foraging strategies are maintained despite workforce reduction: A multidisciplinary survey on the pollen collected by a social pollinator. PLOS ONE, 14(11), e0224037.

Bik, H. M., Porazinska, D. L., Creer, S., Caporaso, J. G., Knight, R., & Thomas, W. K. (2012). Sequencing our way towards understanding global eukaryotic biodiversity. Trends in Ecology & Evolution, 27(4), 233–243.

Bisanz, J. E., Upadhyay, V., Turnbaugh, J. A., Ly, K., & Turnbaugh, P. J. (2019). Meta-Analysis Reveals Reproducible Gut Microbiome Alterations in Response to a High-Fat Diet. Cell Host & Microbe, 26(2), 265-272.e4.

Bishop, A. H. (2019). The signatures of microorganisms and of human and environmental biomes can now be used to provide evidence in legal cases. FEMS Microbiology Letters, 366(3).

Bista, I., Carvalho, G. R., Tang, M., Walsh, K., Zhou, X., Hajibabaei, M., Shokralla, S., Seymour, M., Bradley, D., Liu, S., Christmas, M., & Creer, S. (2018). Performance of amplicon and shotgun sequencing for accurate biomass estimation in invertebrate community samples. Molecular Ecology Resources, 18(5), 1020–1034.

Blaalid, R., Kumar, S., Nilsson, R. H., Abarenkov, K., Kirk, P. M., & Kauserud, H. (2013). ITS 1 versus ITS 2 as DNA metabarcodes for fungi. Molecular Ecology Resources, 13(2), 218–224.

Blomberg, N., & Lauer, K. B. (2020). Connecting data, tools and people across Europe: ELIXIR's response to the COVID-19 pandemic. European Journal of Human Genetics, 28(6), 719–723.

Boeckhout, M., Zielhuis, G. A., & Bredenoord, A. L. (2018). The FAIR guiding principles for data stewardship: Fair enough? European Journal of Human Genetics, 26(7), 931–936.

Bogart, E., Creswell, R., & Gerber, G. K. (2019). MITRE: Inferring features from microbiota time-series data linked to host status. Genome Biology, 20(1), 186.

Bokulich, N. A., Chung, J., Battaglia, T., Henderson, N., Jay, M., Li, H., D. Lieber, A., Wu, F., Perez-Perez, G. I., Chen, Y., Schweizer, W., Zheng, X., Contreras, M., Dominguez-Bello, M. G., & Blaser, M. J. (2016). Antibiotics, birth mode, and diet shape microbiome maturation during early life. Science Translational Medicine, 8(343).

Bokulich, N. A., Collins, T. S., Masarweh, C., Allen, G., Heymann, H., Ebeler, S. E., & Mills, D. A. (2016). Associations among Wine Grape Microbiome, Metabolome, and Fermentation Behavior Suggest Microbial Contribution to Regional Wine Characteristics. MBio, 7(3).

Bokulich, N. A., Dillon, M. R., Bolyen, E., Kaehler, B. D., Huttley, G. A., & Caporaso, J. G. (2018). q2-sample-classifier: Machine-learning tools for microbiome classification and regression. Journal of Open Research Software, 3(30).

Bokulich, N. A., Lewis, Z. T., Boundy-Mills, K., & Mills, D. A. (2016). A new perspective on microbial landscapes within food production. Current Opinion in Biotechnology, 37, 182–189.

Bokulich, N. A., & Mills, D. A. (2013). Facility-Specific “House” Microbiome Drives Microbial Landscapes of Artisan Cheesemaking Plants. Applied and Environmental Microbiology, 79(17), 5214–5223.

Bokulich, N. A., Thorngate, J. H., Richardson, P. M., & Mills, D. A. (2014). Microbial biogeography of wine grapes is conditioned by cultivar, vintage, and climate. Proceedings of the National Academy of Sciences, 111(1), E139–E148.

Bokulich, N. A., Ziemski, M., Robeson, M. S., & Kaehler, B. D. (2020). Measuring the microbiome: Best practices for developing and benchmarking microbiomics methods. Computational and Structural Biotechnology Journal, 18, 4048–4062.

Bolyen, E., Rideout, J. R., Dillon, M. R., Bokulich, N. A., Abnet, C. C., Al-Ghalith, G. A., Alexander, H., Alm, E. J., Arumugam, M., Asnicar, F., Bai, Y., Bisanz, J. E., Bittinger, K., Brejnrod, A., Brislawn, C. J., Brown, C. T., Callahan, B. J., Caraballo-Rodríguez, A. M., Chase, J., … Caporaso, J. G. (2019). Reproducible, interactive, scalable and extensible microbiome data science using QIIME 2. Nature Biotechnology, 37(8), 852–857.

Bonada, N., Prat, N., Resh, V. H., & Statzner, B. (2006). Developments in aquatic insect biomonitoring: A comparative analysis of recent approaches. Annu. Rev. Entomol., 51, 495–523.

Bongaarts, J. (2019). IPBES, 2019. Summary for policymakers of the global assessment report on biodiversity and ecosystem services of the

Intergovernmental Science-Policy Platform on Biodiversity and Ecosystem Services. Population and Development Review, 45(3), 680–681.

Borgelt, C. (2012). Frequent item set mining: Frequent item set mining. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 2(6), 437–456.

Bouslimani, A., Porto, C., Rath, C. M., Wang, M., Guo, Y., Gonzalez, A., Berg-Lyon, D., Ackermann, G., Christensen, G. J. M., Nakatsuji, T., Zhang, L., Borkowski, A. W., Meehan, M. J., Dorrestein, K., Gallo, R. L., Bandeira, N., Knight, R., Alexandrov, T., & Dorrestein, P. C. (2015). Molecular cartography of the human skin surface in 3D. Proceedings of the National Academy of Sciences, 112(17), E2120–E2129.

Boutorh, A., & Guessoum, A. (2016). Complex diseases SNP selection and classification by hybrid Association Rule Mining and Artificial Neural Network—Based Evolutionary Algorithms. Engineering Applications of Artificial Intelligence, 51, 58–70.

Boyer, F., Mercier, C., Bonin, A., Le Bras, Y., Taberlet, P., & Coissac, E. (2016). obitools: A unix-inspired software package for DNA metabarcoding. Molecular Ecology Resources, 16(1), 176–182.

Bruno, A., Sandionigi, A., Agostinetto, G., Bernabovi, L., Frigerio, J., Casiraghi, M., & Labra, M. (2019). Food Tracking Perspective: DNA Metabarcoding to Identify Plant Composition in Complex and Processed Food Products. Genes, 10(3), 248.

Bruno, A., Sandionigi, A., Rizzi, E., Bernasconi, M., Vicario, S., Galimberti, A., Cocuzza, C., Labra, M., & Casiraghi, M. (2017). Exploring the

under-investigated “microbial dark matter” of drinking water treatment plants. Scientific Reports, 7(1), 44350.

Bush, A., Compson, Z. G., Monk, W. A., Porter, T. M., Steeves, R., Emilson, E., Gagne, N., Hajibabaei, M., Roy, M., & Baird, D. J. (2019). Studying ecosystems with DNA metabarcoding: Lessons from biomonitoring of aquatic macroinvertebrates. Frontiers in Ecology and Evolution, 7, 434.

Buttigieg, P. L., Pafilis, E., Lewis, S. E., Schildhauer, M. P., Walls, R. L., & Mungall, C. J. (2016). The environment ontology in 2016: Bridging domains with increased scope, semantic density, and interoperation. Journal of Biomedical Semantics, 7(1), 57.

Byrd, A. L., Belkaid, Y., & Segre, J. A. (2018). The human skin microbiome. Nature Reviews Microbiology, 16(3), 143–155.

Callahan, B. J., McMurdie, P. J., & Holmes, S. P. (2017). Exact sequence variants should replace operational taxonomic units in marker-gene data analysis. The ISME Journal, 11(12), 2639–2643.

Callahan, B. J., McMurdie, P. J., Rosen, M. J., Han, A. W., Johnson, A. J. A., & Holmes, S. P. (2016). DADA2: High-resolution sample inference from Illumina amplicon data. Nature Methods, 13(7), 581–583.

Callewaert, C., Ravard Helffer, K., & Lebaron, P. (2020). Skin Microbiome and its Interplay with the Environment. American Journal of Clinical Dermatology, 21(1), 4–11.

Camacho, C., Coulouris, G., Avagyan, V., Ma, N., Papadopoulos, J., Bealer, K., & Madden, T. L. (2009). BLAST+: Architecture and applications. BMC Bioinformatics, 10(1), 421.

Cambon, M., Ogier, J.-C., Lanois, A., Ferdy, J.-B., & Gaudriault, S. (2018). Changes in rearing conditions rapidly modify gut microbiota structure in Tenebrio molitor larvae [Preprint]. Microbiology.

Capo, E., Spong, G., Königsson, H., & Byström, P. (2020). Effects of filtration methods and water volume on the quantification of brown trout (Salmo trutta) and Arctic char (Salvelinus alpinus) eDNA concentrations via droplet digital PCR. Environmental DNA, 2(2), 152–160.

Capone, K. A., Dowd, S. E., Stamatas, G. N., & Nikolovski, J. (2011). Diversity of the Human Skin Microbiome Early in Life. Journal of Investigative Dermatology, 131(10), 2026–2032.

Caporaso, J. G., Lauber, C. L., Walters, W. A., Berg-Lyons, D., Huntley, J., Fierer, N., Owens, S. M., Betley, J., Fraser, L., Bauer, M., Gormley, N., Gilbert, J. A., Smith, G., & Knight, R. (2012). Ultra-high-throughput microbial community analysis on the Illumina HiSeq and MiSeq platforms. The ISME Journal, 6(8), 1621–1624.

Carmona-Saez, P., Chagoyen, M., Rodriguez, A., Trelles, O., Carazo, J. M., & Pascual-Montano, A. (2006). Integrated analysis of gene expression by association rules discovery. BMC Bioinformatics, 7(1), 54.

Carrieri, A. P., Haiminen, N., Maudsley-Barton, S., Gardiner, L.-J., Murphy, B., Mayes, A. E., Paterson, S., Grimshaw, S., Winn, M., Shand, C., Hadjidoukas, P., Rowe, W. P. M., Hawkins, S., MacGuire-Flanagan, A., Tazzioli, J., Kenny, J. G., Parida, L., Hoptroff, M., & Pyzer-Knapp, E. O. (2021). Explainable AI reveals changes in skin microbiome composition linked to phenotypic differences. Scientific Reports, 11(1), 4565.

Chaffron, S., Rehrauer, H., Pernthaler, J., & von Mering, C. (2010). A global network of coexisting microbes from environmental and whole-genome sequence data. Genome Research, 20(7), 947–959.

Chaix, E., Deléger, L., Bossy, R., & Nédellec, C. (2019). Text mining tools for extracting information about microbial biodiversity in food. Food Microbiology, 81, 63–75.

Chandra, R. V., & Varanasi, B. S. (2015). Python requests essentials. Packt Publishing Ltd.

Chariton, A. A., Stephenson, S., Morgan, M. J., Steven, A. D., Colloff, M. J., Court, L. N., & Hardy, C. M. (2015). Metabarcoding of benthic eukaryote communities predicts the ecological condition of estuaries. Environmental Pollution, 203, 165–174.

Chen, S., Yao, H., Han, J., Liu, C., Song, J., Shi, L., Zhu, Y., Ma, X., Gao, T., Pang, X., Luo, K., Li, Y., Li, X., Jia, X., Lin, Y., & Leon, C. (2010). Validation of the ITS2 Region as a Novel DNA Barcode for Identifying Medicinal Plant Species. PLoS ONE, 5(1), e8613.

Chen, T., Yu, W.-H., Izard, J., Baranova, O. V., Lakshmanan, A., & Dewhirst, F. E. (2010). The Human Oral Microbiome Database: A web accessible resource for investigating oral microbe taxonomic and genomic information. Database, 2010(baq013).

Cheng, S., Melkonian, M., Smith, S. A., Brockington, S., Archibald, J. M., Delaux, P.-M., Li, F.-W., Melkonian, B., Mavrodiev, E. V., & Sun, W. (2018). 10KP: A phylodiverse genome sequencing plan. Gigascience, 7(3), giy013.

Ching, T., Himmelstein, D. S., Beaulieu-Jones, B. K., Kalinin, A. A., Do, B. T., Way, G. P., Ferrero, E., Agapow, P.-M., Zietz, M., Hoffman, M. M., Xie, W., Rosen, G. L., Lengerich, B. J., Israeli, J., Lanchantin, J., Woloszynek, S., Carpenter, A. E., Shrikumar, A., Xu, J., … Greene, C. S. (2018). Opportunities and obstacles for deep learning in biology and medicine. Journal of The Royal Society Interface, 15(141), 20170387.

Cline, E. (2012). Marketplace substitution of Atlantic salmon for Pacific salmon in Washington State detected by DNA barcoding. Food Research International, 45(1), 388–393.

Cock, P. J. A., Antao, T., Chang, J. T., Chapman, B. A., Cox, C. J., Dalke, A., Friedberg, I., Hamelryck, T., Kauff, F., Wilczynski, B., & de Hoon, M. J. L. (2009). Biopython: Freely available Python tools for computational molecular biology and bioinformatics. Bioinformatics, 25(11), 1422–1423.

Cole, J. R., Wang, Q., Fish, J. A., Chai, B., McGarrell, D. M., Sun, Y., Brown, C. T., Porras-Alfaro, A., Kuske, C. R., & Tiedje, J. M. (2014). Ribosomal Database Project: Data and tools for high throughput rRNA analysis. Nucleic Acids Research, 42(D1), D633–D642.

Collins, F. S., Morgan, M., & Patrinos, A. (2003). The Human Genome Project: Lessons from large-scale biology. Science, 300(5617), 286–290.

Compson, Z. G., McClenaghan, B., Singer, G. A. C., Fahner, N. A., & Hajibabaei, M. (2020). Metabarcoding From Microbes to Mammals: Comprehensive Bioassessment on a Global Scale. Frontiers in Ecology and Evolution, 8, 581835.

Comtet, T., Sandionigi, A., Viard, F., & Casiraghi, M. (2015). DNA (meta) barcoding of biological invasions: A powerful tool to elucidate invasion processes and help managing aliens. Biological Invasions, 17(3), 905–922.

Cordier, T., Alonso-Sáez, L., Apothéloz-Perret-Gentil, L., Aylagas, E., Bohan, D. A., Bouchez, A., Chariton, A., Creer, S., Frühe, L., Keck, F., Keeley, N., Laroche, O., Leese, F., Pochon, X., Stoeck, T., Pawlowski, J., & Lanzén, A. (2021). Ecosystems monitoring powered by environmental genomics: A review of current strategies with an implementation roadmap. Molecular Ecology, 30(13), 2937–2958.

Cornelis, J., & Hermy, M. (2004). Biodiversity relationships in urban and suburban parks in Flanders. Landscape and Urban Planning, 69(4), 385–401.

Corrêa, F. B., Saraiva, J. P., Stadler, P. F., & da Rocha, U. N. (2020). TerrestrialMetagenomeDB: A public repository of curated and standardized metadata for terrestrial metagenomes. Nucleic Acids Research, 48(D1), D626–D632.

Costello, M. J., Bouchet, P., Boxshall, G., Fauchald, K., Gordon, D., Hoeksema, B. W., Poore, G. C. B., van Soest, R. W. M., Stöhr, S., Walter, T. C., Vanhoorne, B., Decock, W., & Appeltans, W. (2013). Global Coordination and Standardisation in Marine Biodiversity through the World Register of Marine Species (WoRMS) and Related Databases. PLoS ONE, 8(1), e51629.

Cousijn, H., Kenall, A., Ganley, E., Harrison, M., Kernohan, D., Lemberger, T., Murphy, F., Polischuk, P., Taylor, S., Martone, M., & Clark, T. (2018). A data citation roadmap for scientific publishers. Scientific Data, 5(1), 180259.

Cowart, D. A., Pinheiro, M., Mouchel, O., Maguer, M., Grall, J., Miné, J., & Arnaud-Haond, S. (2015). Metabarcoding is powerful yet still blind: A

comparative analysis of morphological and molecular surveys of seagrass communities. PloS One, 10(2), e0117562.

Cristescu, M. E., & Hebert, P. D. (2018). Uses and misuses of environmental DNA in biodiversity science and conservation. Annual Review of Ecology, Evolution, and Systematics, 49, 209–230.

Curry, C. J., Gibson, J. F., Shokralla, S., Hajibabaei, M., & Baird, D. J. (2018). Identifying North American freshwater invertebrates using DNA barcodes: Are existing COI sequence libraries fit for purpose? Freshwater Science, 37(1), 178–189.

Dabravolski, S. A., & Kavalionak, Y. K. (2020). SARS-CoV-2: Structural diversity, phylogeny, and potential animal host identification of spike glycoprotein. Journal of Medical Virology, 92(9), 1690–1694.

Database resources of the National Center for Biotechnology Information. (2016). Nucleic Acids Research, 44(D1), D7–D19.

Davidson, R. M., & Kurlansky, M. (1997). Cod: A Biography of the Fish That Changed the World.

De Barba, M., Miquel, C., Boyer, F., Mercier, C., Rioux, D., Coissac, E., & Taberlet, P. (2014). DNA metabarcoding multiplexing and validation of data accuracy for diet assessment: Application to omnivorous diet. Molecular Ecology Resources, 14(2), 306–323.

de Carvalho, N. M., Walton, G. E., Poveda, C. G., Silva, S. N., Amorim, M., Madureira, A. R., Pintado, M. E., Gibson, G. R., & Jauregi, P. (2019). Study of in vitro digestion of Tenebrio molitor flour for evaluation of its impact on the human gut microbiota. Journal of Functional Foods, 59, 101–109.

De Filippis, F., Parente, E., & Ercolini, D. (2018). Recent Past, Present, and Future of the Food Microbiome. Annual Review of Food Science and Technology, 9(1), 589–608.

Deagle, B. E., Thomas, A. C., McInnes, J. C., Clarke, L. J., Vesterinen, E. J., Clare, E. L., Kartzinel, T. R., & Eveson, J. P. (2019). Counting with DNA in metabarcoding studies: How should we convert sequence reads to dietary data? Molecular Ecology, 28(2), 391–406.

Deiner, K., Bik, H. M., Mächler, E., Seymour, M., Lacoursière-Roussel, A., Altermatt, F., Creer, S., Bista, I., Lodge, D. M., & De Vere, N. (2017). Environmental DNA metabarcoding: Transforming how we survey animal and plant communities. Molecular Ecology, 26(21), 5872–5895.

Dequiedt, S., Saby, N. P. A., Lelievre, M., Jolivet, C., Thioulouse, J., Toutain, B., Arrouays, D., Bispo, A., Lemanceau, P., & Ranjard, L. (2011). Biogeographical patterns of soil molecular microbial biomass as influenced by soil characteristics and management. Global Ecology and Biogeography, 20(4), 641–652.

DeSantis, T. Z., Hugenholtz, P., Larsen, N., Rojas, M., Brodie, E. L., Keller, K., Huber, T., Dalevi, D., Hu, P., & Andersen, G. L. (2006). Greengenes, a Chimera-Checked 16S rRNA Gene Database and Workbench Compatible with ARB. Applied and Environmental Microbiology, 72(7), 5069–5072.

Di Mattia, C., Battista, N., Sacchetti, G., & Serafini, M. (2019). Antioxidant Activities in vitro of Water and Liposoluble Extracts Obtained by Different Species of Edible Insects and Invertebrates. Frontiers in Nutrition, 6, 106.

Di Pinto, A., Di Pinto, P., Terio, V., Bozzo, G., Bonerba, E., Ceci, E., & Tantillo, G. (2013). DNA barcoding for detecting market substitution in salted cod fillets and battered cod chunks. Food Chemistry, 141(3), 1757–1762.

Dimitriu, P. A., Iker, B., Malik, K., Leung, H., Mohn, W. W., & Hillebrand, G. G. (n.d.). New Insights into the Intrinsic and Extrinsic Factors That Shape the Human Skin Microbiome. MBio, 10(4), e00839-19.

DiMucci, D., Kon, M., & Segrè, D. (2018). Machine Learning Reveals Missing Edges and Putative Interaction Mechanisms in Microbial Ecosystem Networks. MSystems, 3(5).

Dobermann, D., Swift, J. A., & Field, L. M. (2017). Opportunities and hurdles of edible insects for food and feed. Nutrition Bulletin, 42(4), 293–308.

Duvallet, C. (2020). Data detectives, self-love, and humility: A research parasite's perspective. GigaScience, 9(1).

Duvallet, C., Gibbons, S. M., Gurry, T., Irizarry, R. A., & Alm, E. J. (2017). Meta-analysis of gut microbiome studies identifies disease-specific and shared responses. Nature Communications, 8(1), 1784.

Eaton, K. (2020). NCBImeta: Efficient and comprehensive metadata retrieval from NCBI databases. Journal of Open Source Software, 5(46), 1990.

Edgcomb, V., Orsi, W., Bunge, J., Jeon, S., Christen, R., Leslin, C., Holder, M., Taylor, G. T., Suarez, P., & Varela, R. (2011). Protistan microbial observatory in the Cariaco Basin, Caribbean. I. Pyrosequencing vs Sanger insights into species richness. The ISME Journal, 5(8), 1344–1356.

EFSA (2015). Risk profile related to production and consumption of insects as food and feed. EFSA Journal, 13(10), 4257.

Elbrecht, V., Braukmann, T. W., Ivanova, N. V., Prosser, S. W., Hajibabaei, M., Wright, M., Zakharov, E. V., Hebert, P. D., & Steinke, D. (2019). Validation of COI metabarcoding primers for terrestrial arthropods. PeerJ, 7, e7745.

Fahner, N. A., Shokralla, S., Baird, D. J., & Hajibabaei, M. (2016). Large-scale monitoring of plants through environmental DNA metabarcoding of soil: Recovery, resolution, and annotation of four DNA markers. PloS One, 11(6), e0157505.

Faith, D. P., Minchin, P. R., & Belbin, L. (1987). Compositional dissimilarity as a robust measure of ecological distance. Vegetatio, 69(1–3), 57–68.

Fasolato, L., Cardazzo, B., Carraro, L., Fontana, F., Novelli, E., & Balzan, S. (2018). Edible processed insects from e-commerce: Food safety with a focus on the Bacillus cereus group. Food Microbiology, 76, 296–303.

Faust, K. (2021). Open challenges for microbial network construction and analysis. The ISME Journal, 15(11), 3111–3118.

Faust, K., & Raes, J. (2012). Microbial interactions: From networks to models. Nature Reviews Microbiology, 10(8), 538–550.

Federhen, S. (2012). The NCBI Taxonomy database. Nucleic Acids Research, 40(D1), D136–D143.

Fernandes, T. J. R., Costa, J., Oliveira, M. B. P. P., & Mafra, I. (2017). DNA barcoding coupled to HRM analysis as a new and simple tool for the authentication of Gadidae fish species. Food Chemistry, 230, 49–57.

Fernández-Álvarez, F. Á., Machordom, A., García-Jiménez, R., Salinas-Zavala, C. A., & Villanueva, R. (2018). Predatory flying squids are detritivores during their early planktonic life. Scientific Reports, 8(1), 1–12.

Field, S. A., Tyre, A. J., & Possingham, H. P. (2005). Optimizing allocation of monitoring effort under economic and observational constraints. The Journal of Wildlife Management, 69(2), 473–482.

Folmer, O., Black, M., Hoeh, W., Lutz, R., & Vrijenhoek, R. (1994). DNA primers for amplification of mitochondrial cytochrome c oxidase subunit I from diverse metazoan invertebrates. Molecular Marine Biology and Biotechnology, 3(5), 294–299.

Fox, M., Mitchell, M., Dean, M., Elliott, C., & Campbell, K. (2018). The seafood supply chain from a fraudulent perspective. Food Security, 10(4), 939–963.

Franceschini, A., Szklarczyk, D., Frankild, S., Kuhn, M., Simonovic, M., Roth, A., Lin, J., Minguez, P., Bork, P., von Mering, C., & Jensen, L. J. (2012). STRING v9.1: Protein-protein interaction networks, with increased coverage and integration. Nucleic Acids Research, 41(D1), D808–D815.

Frigerio, J., Agostinetto, G., Galimberti, A., De Mattia, F., Labra, M., & Bruno, A. (2020). Tasting the differences: Microbiota analysis of different insect-based novel food. Food Research International, 137, 109426.

Frigerio, J., Agostinetto, G., Mezzasalma, V., De Mattia, F., Labra, M., & Bruno, A. (2021). DNA-Based Herbal Teas' Authentication: An ITS2 and psbA-trnH Multi-Marker DNA Metabarcoding Approach. Plants, 10(10), 2120.

Frigerio, J., Agostinetto, G., Sandionigi, A., Mezzasalma, V., Berterame, N. M., Casiraghi, M., Labra, M., & Galimberti, A. (2020). The hidden 'plant side' of insect novel foods: A DNA-based assessment. Food Research International, 128, 108751.

Frigerio, J., Gorini, T., Galimberti, A., Bruni, I., Tommasi, N., Mezzasalma, V., & Labra, M. (2019). DNA barcoding to trace Medicinal and Aromatic Plants from the field to the food supplement. Journal of Applied Botany and Food Quality, 92, 33–38.

Frontalini, F., Greco, M., Di Bella, L., Lejzerowicz, F., Reo, E., Caruso, A., Cosentino, C., Maccotta, A., Scopelliti, G., & Nardelli, M. P. (2018). Assessing the effect of mercury pollution on cultured benthic foraminifera community using morphological and eDNA metabarcoding approaches. Marine Pollution Bulletin, 129(2), 512–524.

Galimberti, A., Bruno, A., Agostinetto, G., Casiraghi, M., Guzzetti, L., & Labra, M. (2021). Fermented food products in the era of globalization: Tradition meets biotechnology innovations. Current Opinion in Biotechnology, 70, 36–41.

Galimberti, A., Bruno, A., Mezzasalma, V., De Mattia, F., Bruni, I., & Labra, M. (2015). Emerging DNA-based technologies to characterize food ecosystems. Food Research International, 69, 424–433.

Galimberti, A., Casiraghi, M., Bruni, I., Guzzetti, L., Cortis, P., Berterame, N. M., & Labra, M. (2019). From DNA barcoding to personalized nutrition: The evolution of food traceability. Current Opinion in Food Science, 28, 41–48.

Galvin-King, P., Haughey, S. A., & Elliott, C. T. (2018). Herb and spice fraud; the drivers, challenges and detection. Food Control, 88, 85–97.

Ganda, H., Zannou-Boukari, E. T., Kenis, M., Chrysostome, C. A. A. M., & Mensah, G. A. (2019). Potentials of animal, crop and agri-food wastes for the production of fly larvae. Journal of Insects as Food and Feed, 5(2), 59–67.

Garofalo, C., Milanović, V., Cardinali, F., Aquilanti, L., Clementi, F., & Osimani, A. (2019). Current knowledge on the microbiota of edible insects intended for human consumption: A state-of-the-art review. Food Research International, 125, 108527.

Garofalo, C., Osimani, A., Milanović, V., Taccari, M., Cardinali, F., Aquilanti, L., Riolo, P., Ruschioni, S., Isidoro, N., & Clementi, F. (2017). The microbiota of marketed processed edible insects as revealed by high-throughput sequencing. Food Microbiology, 62, 15–22.

Geib, S. M., Hall, B., Derego, T., Bremer, F. T., Cannoles, K., & Sim, S. B. (2018). Genome Annotation Generator: A simple tool for generating and correcting WGS annotation tables for NCBI submission. GigaScience, 7(4).

Gevers, D., Knight, R., Petrosino, J. F., Huang, K., McGuire, A. L., Birren, B. W., Nelson, K. E., White, O., Methé, B. A., & Huttenhower, C. (2012). The Human Microbiome Project: A Community Resource for the Healthy Human Microbiome. PLOS Biology, 10(8), e1001377.

Ghannam, R. B., & Techtmann, S. M. (2021). Machine learning applications in microbial ecology, human microbiome studies, and environmental monitoring. Computational and Structural Biotechnology Journal, 19, 1092–1107.

Gilbert, J. A., Jansson, J. K., & Knight, R. (2014). The Earth Microbiome project: Successes and aspirations. BMC Biology, 12(1), 69.

Gloor, G. B., Macklaim, J. M., Pawlowsky-Glahn, V., & Egozcue, J. J. (2017). Microbiome Datasets Are Compositional: And This Is Not Optional. Frontiers in Microbiology, 8, 2224.

Gonçalves, R. S., & Musen, M. A. (2019). The variable quality of metadata about biological samples used in biomedical experiments. Scientific Data, 6(1), 190021.

Gonzalez, A., Navas-Molina, J. A., Kosciolek, T., McDonald, D., Vázquez-Baeza, Y., Ackermann, G., DeReus, J., Janssen, S., Swafford, A. D., Orchanian, S. B., Sanders, J. G., Shorenstein, J., Holste, H., Petrus, S., Robbins-Pianka, A., Brislawn, C. J., Wang, M., Rideout, J. R., Bolyen, E., … Knight, R. (2018). Qiita: Rapid, web-enabled microbiome meta-analysis. Nature Methods, 15(10), 796–798.

Gordon, D. E., Jang, G. M., Bouhaddou, M., Xu, J., Obernier, K., White, K. M., O'Meara, M. J., Rezelj, V. V., Guo, J. Z., Swaney, D. L., Tummino, T. A., Hüttenhain, R., Kaake, R. M., Richards, A. L., Tutuncuoglu, B., Foussard, H., Batra, J., Haas, K., Modak, M., … Krogan, N. J. (2020). A SARS-CoV-2 protein interaction map reveals targets for drug repurposing. Nature, 583(7816), 459–468.

Greathouse, K. L., Sinha, R., & Vogtmann, E. (2019). DNA extraction for human microbiome studies: The issue of standardization. Genome Biology, 20(1), 212.

Gupta, V. K., Paul, S., & Dutta, C. (2017). Geography, Ethnicity or Subsistence-Specific Variations in Human Microbiome Composition and Diversity. Frontiers in Microbiology, 0.

Hahsler, M., Chelluboina, S., Hornik, K., & Buchta, C. (2011). The arules R-package ecosystem: Analyzing interesting patterns from large transaction data sets. The Journal of Machine Learning Research, 12, 2021–2025.

Hamady, M., & Knight, R. (2009). Microbial community profiling for human microbiome projects: Tools, techniques, and challenges. Genome Research, 19(7), 1141–1152.

Hampton, S. E., Jones, M. B., Wasser, L. A., Schildhauer, M. P., Supp, S. R., Brun, J., Hernandez, R. R., Boettiger, C., Collins, S. L., Gross, L. J., Fernández, D. S., Budden, A., White, E. P., Teal, T. K., Labou, S. G., & Aukema, J. E. (2017). Skills and Knowledge for Data-Intensive Environmental Research. BioScience, 67(6), 546–557.

Han, J., Pei, J., Yin, Y., & Mao, R. (2004). Mining Frequent Patterns without Candidate Generation: A Frequent-Pattern Tree Approach. Data Mining and Knowledge Discovery, 8(1), 53–87.

Hardisty, A., Roberts, D., & The Biodiversity Informatics Community. (2013). A decadal view of biodiversity informatics: Challenges and priorities. BMC Ecology, 13(1), 16.

Harris, C. R., Millman, K. J., van der Walt, S. J., Gommers, R., Virtanen, P., Cournapeau, D., Wieser, E., Taylor, J., Berg, S., Smith, N. J., Kern, R., Picus, M., Hoyer, S., van Kerkwijk, M. H., Brett, M., Haldane, A., del Río, J. F., Wiebe, M., Peterson, P., … Oliphant, T. E. (2020). Array programming with NumPy. Nature, 585(7825), 357–362.

Harrison, J. P., Chronopoulou, M., Salonen, I., Jilbert, T., & Koho, K. (2021). 16S and 18S rRNA gene metabarcoding provide congruent information on the

responses of sediment communities to eutrophication. Frontiers in Marine Science, 8, 862.

Harrison, P. W., Alako, B., Amid, C., Cerdeño-Tárraga, A., Cleland, I., Holt, S., Hussein, A., Jayathilaka, S., Kay, S., Keane, T., Leinonen, R., Liu, X., Martínez-Villacorta, J., Milano, A., Pakseresht, N., Rajan, J., Reddy, K., Richards, E., Rosello, M., … Cochrane, G. (2019). The European Nucleotide Archive in 2018. Nucleic Acids Research, 47(D1), D84–D88.

Haynes, E., Jimenez, E., Pardo, M. A., & Helyar, S. J. (2019). The future of NGS (Next Generation Sequencing) analysis in testing food authenticity. Food Control, 101, 134–143.

Hebert, P. D. N., Ratnasingham, S., & de Waard, J. R. (2003). Barcoding animal life: Cytochrome c oxidase subunit 1 divergences among closely related species. Proceedings of the Royal Society of London. Series B: Biological Sciences, 270(suppl_1).

Hellberg, R. S., Kawalek, M. D., Van, K. T., Shen, Y., & Williams-Hill, D. M. (2014). Comparison of DNA Extraction and PCR Setup Methods for Use in High-Throughput DNA Barcoding of Fish Species. Food Analytical Methods, 7(10), 1950–1959.

Herbold, C. W., Pelikan, C., Kuzyk, O., Hausmann, B., Angel, R., Berry, D., & Loy, A. (2015). A flexible and economical barcoding approach for highly multiplexed amplicon sequencing of diverse target genes. Frontiers in Microbiology, 6.

Hirst, J. (1952). An automatic volumetric spore trap. Annals of Applied Biology, 39(2), 257–265.

Hoarfrost, A., Brown, N., Brown, C. T., & Arnosti, C. (2019). Sequencing data discovery with MetaSeek. Bioinformatics, 35(22), 4857–4859.

Hoffmann, M., Kleine-Weber, H., Schroeder, S., Krüger, N., Herrler, T., Erichsen, S., Schiergens, T. S., Herrler, G., Wu, N.-H., Nitsche, A., Müller, M. A., Drosten, C., & Pöhlmann, S. (2020). SARS-CoV-2 Cell Entry Depends on ACE2 and TMPRSS2 and Is Blocked by a Clinically Proven Protease Inhibitor. Cell, 181(2), 271-280.e8.

Hosoda, S., Nishijima, S., Fukunaga, T., Hattori, M., & Hamada, M. (2020). Revealing the microbial assemblage structure in the human gut microbiome using latent Dirichlet allocation. Microbiome, 8(1), 95.

House, J. (2016). Consumer acceptance of insect-based foods in the Netherlands: Academic and commercial implications. Appetite, 107, 47–58.

Hu, W., Zaveri, A., Qiu, H., & Dumontier, M. (2017). Cleaning by clustering: Methodology for addressing data quality issues in biomedical metadata. BMC Bioinformatics, 18(1), 415.

Huerta-Cepas, J., Serra, F., & Bork, P. (2016). ETE 3: Reconstruction, Analysis, and Visualization of Phylogenomic Data. Molecular Biology and Evolution, 33(6), 1635–1638.

Hugenholtz, P., Goebel, B. M., & Pace, N. R. (1998). Impact of Culture-Independent Studies on the Emerging Phylogenetic View of Bacterial Diversity. Journal of Bacteriology, 180(18), 4765–4774.

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

Hunter, J. D. (2007). Matplotlib: A 2D Graphics Environment. Computing in Science & Engineering, 9(3), 90–95.

Hussein, N., Alashqur, A., & Sowan, B. (2015). Using the interestingness measure lift to generate association rules. Journal of Advanced Computer Science & Technology, 4(1), 156.

Huttenhower, C., Gevers, D., Knight, R., Abubucker, S., Badger, J. H., Chinwalla, A. T., Creasy, H. H., Earl, A. M., FitzGerald, M. G., Fulton, R. S., Giglio, M. G., Hallsworth-Pepin, K., Lobos, E. A., Madupu, R., Magrini, V., Martin, J. C., Mitreva, M., Muzny, D. M., Sodergren, E. J., … The Human Microbiome Project Consortium. (2012). Structure, function and diversity of the healthy human microbiome. Nature, 486(7402), 207–214.

Integrative, H. M. P., Proctor, L. M., Creasy, H. H., Fettweis, J. M., Lloyd-Price, J., Mahurkar, A., ... & Huttenhower, C. (2019). The integrative human microbiome project. Nature, 569(7758), 641-648.

Jaiswal, S. K., Agarwal, S. M., Thodum, P., & Sharma, V. K. (2021). SkinBug: An artificial intelligence approach to predict human skin microbiome-mediated metabolism of biotics and xenobiotics. IScience, 24(1), 101925.

Jamwal, P. S., Bruno, A., Galimberti, A., Magnani, D., Krupa, H., Casiraghi, M., & Loy, A. (2021). First assessment of eDNA-based detection approach to monitor the presence of Eurasian otter in southern Italy. Hystrix, the Italian Journal of Mammalogy.

Johansen, P. G., Owusu-Kwarteng, J., Parkouda, C., Padonou, S. W., & Jespersen, L. (2019). Occurrence and Importance of Yeasts in Indigenous Fermented Food and Beverages Produced in Sub-Saharan Africa. Frontiers in Microbiology, 10, 1789.

Johansen, S. D., Coucheron, D. H., Andreassen, M., Karlsen, B. O., Furmanek, T., Jørgensen, T. E., Emblem, Å., Breines, R., Nordeide, J. T., Moum, T., Nederbragt, A. J., Stenseth, N. C., & Jakobsen, K. S. (2009). Large-scale sequence analyses of Atlantic cod. New Biotechnology, 25(5), 263–271.

Johnson, J. L., Moore, W. E. C., & Moore, L. V. H. (1986). Bacteroides caccae sp. Nov., Bacteroides merdae sp. Nov., and Bacteroides stercoris sp. Nov. Isolated from Human Feces. International Journal of Systematic Bacteriology, 36(4), 499–501.

Jordan, M. I., & Mitchell, T. M. (2015). Machine learning: Trends, perspectives, and prospects. Science, 349(6245), 255–260.

Jorde, P. E., Kleiven, A. R., Sodeland, M., Olsen, E. M., Ferter, K., Jentoft, S., & Knutsen, H. (2018). Who is fishing on what stock: Population-of-origin of individual cod (Gadus morhua) in commercial and recreational fisheries. ICES Journal of Marine Science, 75(6), 2153–2162.

Jurburg, S. D., Konzack, M., Eisenhauer, N., & Heintz-Buschart, A. (2020). The archives are half-empty: An assessment of the availability of microbial community sequencing data. Communications Biology, 3(1), 1–8.

Kamimura, B. A., Cabral, L., Noronha, M. F., Baptista, R. C., Nascimento, H. M., & Sant'Ana, A. S. (2020). Amplicon sequencing reveals the bacterial diversity in milk, dairy premises and Serra da Canastra artisanal cheeses produced by three different farms. Food Microbiology, 89, 103453.

Kans, J. (2021). Entrez Direct: E-utilities on the Unix Command Line. In Entrez Programming Utilities Help [Internet]. National Center for Biotechnology Information (US).

Kanz, C., Aldebert, P., Althorpe, N., Baker, W., Baldwin, A., Bates, K., Browne, P., van den Broek, A., Castro, M., & Cochrane, G. (2005). The EMBL nucleotide sequence database. Nucleic Acids Research, 33(suppl_1), D29–D33.

Karpinets, T. V., Park, B. H., & Uberbacher, E. C. (2012). Analyzing large biological datasets with association networks. Nucleic Acids Research, 40(17), e131–e131.

Kasmanas, J. C., Bartholomäus, A., Corrêa, F. B., Tal, T., Jehmlich, N., Herberth, G., von Bergen, M., Stadler, P. F., Carvalho, A. C. P. de L. F. de, & Nunes da Rocha, U. (2021). HumanMetagenomeDB: A public repository of curated and standardized metadata for human metagenomes. Nucleic Acids Research, 49(D1), D743–D750.

Kato, T., Fukuda, S., Fujiwara, A., Suda, W., Hattori, M., Kikuchi, J., & Ohno, H. (2014). Multiple Omics Uncovers Host–Gut Microbial Mutualism During Prebiotic Fructooligosaccharide Supplementation. DNA Research, 21(5), 469–480.

Kaur, P., Klan, F., & König-Ries, B. (2018, June). Issues and Suggestions for the Development of a Biodiversity Data Visualization Support Tool. In EuroVis (Short Papers) (pp. 73-77).

Keller, A., Hohlfeld, S., Kolter, A., Schultz, J., Gemeinholzer, B., & Ankenbrand, M. J. (2020). BCdatabaser: On-the-fly reference database creation for (meta-)barcoding. Bioinformatics, 36(8), 2630–2631.

Khailany, R. A., Safdar, M., & Ozaslan, M. (2020). Genomic characterization of a novel SARS-CoV-2. Gene Reports, 19, 100682.

Kim, D., Lee, J.-Y., Yang, J.-S., Kim, J. W., Kim, V. N., & Chang, H. (2020). The Architecture of SARS-CoV-2 Transcriptome. Cell, 181(4), 914-921.e10.

Kim, O.-S., Cho, Y.-J., Lee, K., Yoon, S.-H., Kim, M., Na, H., Park, S.-C., Jeon, Y. S., Lee, J.-H., Yi, H., Won, S., & Chun, J. 2012. (n.d.). Introducing EzTaxon-e: A prokaryotic 16S rRNA gene sequence database with phylotypes that represent uncultured species. International Journal of Systematic and Evolutionary Microbiology, 62(Pt_3), 716–721.

Klie, A., Tsui, B. Y., Mollah, S., Skola, D., Dow, M., Hsu, C.-N., & Carter, H. (2021). Increasing metadata coverage of SRA BioSample entries using deep learning–based named entity recognition. Database, 2021(baab021).

Klunder, H. C., Wolkers-Rooijackers, J., Korpela, J. M., & Nout, M. J. R. (2012). Microbiological aspects of processing and storage of edible insects. Food Control, 26(2), 628–631.

Knight, R., Vrbanac, A., Taylor, B. C., Aksenov, A., Callewaert, C., Debelius, J., Gonzalez, A., Kosciolek, T., McCall, L.-I., McDonald, D., Melnik, A. V., Morton, J. T., Navas, J., Quinn, R. A., Sanders, J. G., Swafford, A. D., Thompson, L. R., Tripathi, A., Xu, Z. Z., … Dorrestein, P. C. (2018). Best practices for analysing microbiomes. Nature Reviews Microbiology, 16(7), 410–422.

Knights, D., Kuczynski, J., Koren, O., Ley, R. E., Field, D., Knight, R., DeSantis, T. Z., & Kelley, S. T. (2011). Supervised classification of microbiota mitigates mislabeling errors. The ISME Journal, 5(4), 570–573.

Knudsen, S. W., Ebert, R. B., Hesselsøe, M., Kuntke, F., Hassingboe, J., Mortensen, P. B., Thomsen, P. F., Sigsgaard, E. E., Hansen, B. K., Nielsen, E. E., & Møller, P. R. (2019). Species-specific detection and quantification of

environmental DNA from marine fishes in the Baltic Sea. Journal of Experimental Marine Biology and Ecology, 510, 31–45.

Kong, H. H. (2011). Skin microbiome: Genomics-based insights into the diversity and role of skin microbes. Trends in Molecular Medicine, 17(6), 320–328.

Kong, H. H., & Segre, J. A. (2012). Skin Microbiome: Looking Back to Move Forward. Journal of Investigative Dermatology, 132(3), 933–939.

Kosti, I., Lyalina, S., Pollard, K. S., Butte, A. J., & Sirota, M. (2020). Meta-Analysis of Vaginal Microbiome Data Provides New Insights Into Preterm Birth. Frontiers in Microbiology, 0.

Koyutürk, M., Kim, Y., Subramaniam, S., Szpankowski, W., & Grama, A. (2006). Detecting Conserved Interaction Patterns in Biological Networks. Journal of Computational Biology, 13(7), 1299–1322.

Krehenwinkel, H., Fong, M., Kennedy, S., Huang, E. G., Noriyuki, S., Cayetano, L., & Gillespie, R. (2018). The effect of DNA degradation bias in passive sampling devices on metabarcoding studies of arthropod communities and their associated microbiota. PLoS One, 13(1), e0189188.

Krehenwinkel, H., Wolf, M., Lim, J. Y., Rominger, A. J., Simison, W. B., & Gillespie, R. G. (2017). Estimating and mitigating amplification bias in qualitative and quantitative arthropod metabarcoding. Scientific Reports, 7(1), 17668.

Kruskal, W. H., & Wallis, W. A. (1952). Use of ranks in one-criterion variance analysis. Journal of the American Statistical Association, 47(260), 583–621.

Kyrpides, N. C., Eloe-Fadrosh, E. A., & Ivanova, N. N. (2016). Microbiome Data Science: Understanding Our Microbial Planet. Trends in Microbiology, 24(6), 425–427.

Lagier, J.-C., Million, M., Hugon, P., Armougom, F., & Raoult, D. (2012). Human Gut Microbiota: Repertoire and Variations. Frontiers in Cellular and Infection Microbiology, 2.

Lallias, D., Hiddink, J. G., Fonseca, V. G., Gaspar, J. M., Sung, W., Neill, S. P., Barnes, N., Ferrero, T., Hall, N., & Lambshead, P. J. D. (2015). Environmental metabarcoding reveals heterogeneous drivers of microbial eukaryote diversity in contrasting estuarine ecosystems. The ISME Journal, 9(5), 1208–1221.

Lamb, P. D., Hunter, E., Pinnegar, J. K., Creer, S., Davies, R. G., & Taylor, M. I. (2019). How quantitative is metabarcoding: A meta-analytical approach. Molecular Ecology, 28(2), 420–430.

Langan, E. A., Griffiths, C. E. M., Solbach, W., Knobloch, J. K., Zillikens, D., & Thaçi, D. (2018). The role of the microbiome in psoriasis: Moving from disease description to treatment selection? British Journal of Dermatology, 178(5), 1020–1027.

Lanzén, A., Lekang, K., Jonassen, I., Thompson, E. M., & Troedsson, C. (2016). High-throughput metabarcoding of eukaryotic diversity for environmental monitoring of offshore oil-drilling activities. Molecular Ecology, 25(17), 4392–4406.

Layeghifard, M., Hwang, D. M., & Guttman, D. S. (2017). Disentangling Interactions in the Microbiome: A Network Perspective. Trends in Microbiology, 25(3), 217–228.

Lee, J.-Y., Sadler, N. C., Egbert, R. G., Anderton, C. R., Hofmockel, K. S., Jansson, J. K., & Song, H.-S. (2020). Deep learning predicts microbial interactions from self-organized spatiotemporal patterns. Computational and Structural Biotechnology Journal, 18, 1259–1269.

Letko, M., Marzi, A., & Munster, V. (2020). Functional assessment of cell entry and receptor usage for SARS-CoV-2 and other lineage B betacoronaviruses. Nature Microbiology, 5(4), 562–569.

Leyden, J. J., McGinley, K. J., Mills, O. H., & Kligman, A. M. (1975). Age-related changes in the resident bacterial flora of the human face. The Journal of Investigative Dermatology, 65(4), 379–381.

Li, L., Xie, B., Dong, C., Wang, M., & Liu, H. (2016). Can closed artificial ecosystem have an impact on insect microbial community? A case study of yellow mealworm (Tenebrio molitor L.). Ecological Engineering, 86, 183–189.

Lima-Mendez, G., Faust, K., Henry, N., Decelle, J., Colin, S., Carcillo, F., Chaffron, S., Ignacio-Espinosa, J. C., Roux, S., Vincent, F., Bittner, L., Darzi, Y., Wang, J., Audic, S., Berline, L., Bontempi, G., Cabello, A. M., Coppola, L., Cornejo-Castillo, F. M., … Raes, J. (2015). Determinants of community structure in the global plankton interactome. Science, 348(6237), 1262073.

Liu, M., Ye, Y., Jiang, J., & Yang, K. (2021). MANIEA: A microbial association network inference method based on improved Eclat association rule mining algorithm. Bioinformatics, 37(20), 3569–3578.

Liu, Y.-X., Qin, Y., Chen, T., Lu, M., Qian, X., Guo, X., & Bai, Y. (2021). A practical guide to amplicon and metagenomic analysis of microbiome data. Protein & Cell, 12(5), 315–330. https://doi.org/10.1007/s13238-020-00724-8

Longo, D. L., & Drazen, J. M. (2016). Data Sharing. New England Journal of Medicine, 374(3), 276–277.

Lozupone, C., Lladser, M. E., Knights, D., Stombaugh, J., & Knight, R. (2011). UniFrac: An effective distance metric for microbial community comparison. The ISME Journal, 5(2), 169–172.

Lu, R., Zhao, X., Li, J., Niu, P., Yang, B., Wu, H., Wang, W., Song, H., Huang, B., Zhu, N., Bi, Y., Ma, X., Zhan, F., Wang, L., Hu, T., Zhou, H., Hu, Z., Zhou, W., Zhao, L., … Tan, W. (2020). Genomic characterisation and epidemiology of 2019 novel coronavirus: Implications for virus origins and receptor binding. The Lancet, 395(10224), 565–574.

Luan, J., Jin, X., Lu, Y., & Zhang, L. (2020). SARS-CoV-2 spike protein favors ACE2 from Bovidae and Cricetidae. Journal of Medical Virology, 92(9), 1649–1656.

Luna, P. C. (2020). Skin Microbiome as Years Go By. American Journal of Clinical Dermatology, 21(1), 12–17.

Ma, B., Wang, Y., Ye, S., Liu, S., Stirling, E., Gilbert, J. A., Faust, K., Knight, R., Jansson, J. K., Cardona, C., Röttjers, L., & Xu, J. (2020). Earth microbial co-occurrence network reveals interconnection pattern across microbiomes. Microbiome, 8(1), 82.

Macher, T., Beermann, A. J., & Leese, F. (2021). TaxonTableTools: A comprehensive, platform-independent graphical user interface software to explore and visualise DNA metabarcoding data. Molecular Ecology Resources, 21(5), 1705–1714.

Magara, H. J. O., Tanga, C. M., Ayieko, M. A., Hugel, S., Mohamed, S. A., Khamis, F. M., Salifu, D., Niassy, S., Sevgan, S., Fiaboe, K. K. M., Roos, N., & Ekesi, S. (2019). Performance of Newly Described Native Edible Cricket Scapsipedus icipe (Orthoptera: Gryllidae) on Various Diets of Relevance for Farming. Journal of Economic Entomology, 112(2), 653–664.

Magurran, A. E., Baillie, S. R., Buckland, S. T., Dick, J. M., Elston, D. A., Scott, E. M., Smith, R. I., Somerfield, P. J., & Watt, A. D. (2010). Long-term datasets in biodiversity research and monitoring: Assessing change in ecological communities through time. Trends in Ecology & Evolution, 25(10), 574–582.

Mahé, F., Rognes, T., Quince, C., de Vargas, C., & Dunthorn, M. (2015). Swarm v2: Highly-scalable and high-resolution amplicon clustering. PeerJ, 3, e1420.

Makiola, A., Compson, Z. G., Baird, D. J., Barnes, M. A., Boerlijst, S. P., Bouchez, A., Brennan, G., Bush, A., Canard, E., Cordier, T., Creer, S., Curry, R. A., David, P., Dumbrell, A. J., Gravel, D., Hajibabaei, M., Hayden, B., van der Hoorn, B., Jarne, P., … Bohan, D. A. (2020). Key Questions for Next-Generation Biomonitoring. Frontiers in Environmental Science, 7, 197.

Manda, P. (2020). Data mining powered by the gene ontology. WIREs Data Mining and Knowledge Discovery, 10(3).

Manda, P., McCarthy, F., & Bridges, S. M. (2013). Interestingness measures and strategies for mining multi-ontology multi-level association rules from gene ontology annotations for the discovery of new GO relationships. Journal of Biomedical Informatics, 46(5), 849–856.

Manda, P., Ozkan, S., Wang, H., McCarthy, F., & Bridges, S. M. (2012). Cross-Ontology Multi-level Association Rule Mining in the Gene Ontology. PLoS ONE, 7(10), e47411.

Marco, M. L., Heeney, D., Binda, S., Cifelli, C. J., Cotter, P. D., Foligné, B., Gänzle, M., Kort, R., Pasin, G., Pihlanto, A., Smid, E. J., & Hutkins, R. (2017). Health benefits of fermented foods: Microbiota and beyond. Current Opinion in Biotechnology, 44, 94–102.

Marcos-Zambrano, L. J., Karaduzovic-Hadziabdic, K., Loncar Turukalo, T., Przymus, P., Trajkovik, V., Aasmets, O., Berland, M., Gruca, A., Hasic, J., Hron, K., Klammsteiner, T., Kolev, M., Lahti, L., Lopes, M. B., Moreno, V., Naskinova, I., Org, E., Paciência, I., Papoutsoglou, G., … Truu, J. (2021). Applications of Machine Learning in Human Microbiome Studies: A Review on Feature Selection, Biomarker Identification, Disease Prediction and Treatment. Frontiers in Microbiology, 0.

Marsland, R., Cui, W., Goldford, J., & Mehta, P. (2020). The Community Simulator: A Python package for microbial ecology. PLOS ONE, 15(3), e0230430.

Martínez-Romero, M., O'Connor, M. J., Egyedi, A. L., Willrett, D., Hardi, J., Graybeal, J., & Musen, M. A. (2019). Using association rule mining and ontologies to generate metadata recommendations from multiple biomedical databases. Database, 2019(baz059).

Mazzucchelli, G., Holzhauser, T., Cirkovic Velickovic, T., Diaz-Perales, A., Molina, E., Roncada, P., Rodrigues, P., Verhoeckx, K., & Hoffmann-Sommergruber, K. (2018). Current (Food) Allergenic Risk Assessment: Is It Fit for Novel Foods? Status Quo and Identification of Gaps. Molecular Nutrition & Food Research, 62(1), 1700278.

McDonald, D., Clemente, J. C., Kuczynski, J., Rideout, J. R., Stombaugh, J., Wendel, D., Wilke, A., Huse, S., Hufnagle, J., Meyer, F., Knight, R., & Caporaso,

J. G. (2012). The Biological Observation Matrix (BIOM) format or: How I learned to stop worrying and love the ome-ome. GigaScience, 1(1), 7.

McDonald, D., Price, M. N., Goodrich, J., Nawrocki, E. P., DeSantis, T. Z., Probst, A., Andersen, G. L., Knight, R., & Hugenholtz, P. (2012). An improved Greengenes taxonomy with explicit ranks for ecological and evolutionary analyses of bacteria and archaea. The ISME Journal, 6(3), 610–618.
McGee, K. M., Robinson, C. V., & Hajibabaei, M. (2019). Gaps in DNA-Based Biomonitoring Across the Globe. Frontiers in Ecology and Evolution, 7, 337.

McKinney, M. L. (2008). Effects of urbanization on species richness: A review of plants and animals. Urban Ecosystems, 11(2), 161–176.

McKinney, W. (2010, June). Data structures for statistical computing in python. In Proceedings of the 9th Python in Science Conference (Vol. 445, pp. 51-56). doi:10.25080/Majora- 92bf1922-00a

McKinney, W. (2010). Data Structures for Statistical Computing in Python. 56–61.

Menni, C., Jackson, M. A., Pallister, T., Steves, C. J., Spector, T. D., & Valdes, A. M. (2017). Gut microbiome diversity and high-fibre intake are related to lower long-term weight gain. International Journal of Obesity, 41(7), 1099–1105.

Mezzasalma, V., Sandionigi, A., Bruni, I., Bruno, A., Lovicu, G., Casiraghi, M., & Labra, M. (2017). Grape microbiome as a reliable and persistent signature of field origin and environmental conditions in Cannonau wine production. PLOS ONE, 12(9), e0184615.

Mezzasalma, V., Sandionigi, A., Guzzetti, L., Galimberti, A., Grando, M. S., Tardaguila, J., & Labra, M. (2018). Geographical and Cultivar Features

Differentiate Grape Microbiota in Northern Italy and Spain Vineyards. Frontiers in Microbiology, 9, 946.

Michener, W. K., & Jones, M. B. (2012). Ecoinformatics: Supporting ecology as a data-intensive science. Trends in Ecology & Evolution, 27(2), 85–93.

Miller, D. A., Pacifici, K., Sanderlin, J. S., & Reich, B. J. (2019). The recent past and promising future for data integration methods to estimate species' distributions. Methods in Ecology and Evolution, 10(1), 22–37.

Miller, D. D., & Mariani, S. (2010). Smoke, mirrors, and mislabeled cod: Poor transparency in the European seafood industry. Frontiers in Ecology and the Environment, 8(10), 517–521.

Miron, L., Gonçalves, R. S., & Musen, M. A. (2020). Obstacles to the reuse of study metadata in ClinicalTrials.gov. Scientific Data, 7(1), 443.

Mitchell, A. L., Almeida, A., Beracochea, M., Boland, M., Burgin, J., Cochrane, G., Crusoe, M. R., Kale, V., Potter, S. C., Richardson, L. J., Sakharova, E., Scheremetjew, M., Korobeynikov, A., Shlemov, A., Kunyavskaya, O., Lapidus, A., & Finn, R. D. (2020). MGnify: The microbiome analysis resource in 2020. Nucleic Acids Research, 48(D1), D570–D578.

Méndez, E., Vela, C., Immer, U., & Janssen, F. W. (2005). Report of a collaborative trial to investigate the performance of the R5 enzyme linked immunoassay to determine gliadin in gluten-free food: European Journal of Gastroenterology & Hepatology, 17(10), 1053–1063.

Mohajeri, M. H., Brummer, R. J. M., Rastall, R. A., Weersma, R. K., Harmsen, H. J. M., Faas, M., & Eggersdorfer, M. (2018). The role of the microbiome for

human health: From basic science to clinical applications. European Journal of Nutrition, 57(1), 1–14.

Montassier, E., Al-Ghalith, G. A., Ward, T., Corvec, S., Gastinne, T., Potel, G., Moreau, P., de la Cochetiere, M. F., Batard, E., & Knights, D. (2016). Pretreatment gut microbiome predicts chemotherapy-related bloodstream infection. Genome Medicine, 8(1), 49.

Mortazavi, S., Samih, M. A., Ghajarieh, H., & Jafari, A. (2015). Effect of some diets on demographic parameters of Ectomyelois ceratoniae (Zeller) (Lepidoptera: Pyralidae) in vitro. Journal of Plant Protection Research, 55(2), 212–219.

Mueller, S., Handy, S. M., Deeds, J. R., George, G. O., Broadhead, W. J., Pugh, S. E., & Garrett, S. D. (2015). Development of a COX 1 based PCR-RFLP method for fish species identification. Food Control, 55, 39–42.

Muino, D. P., & Borgelt, C. (2014). Frequent item set mining for sequential data: Synchrony in neuronal spike trains. Intelligent Data Analysis, 18(6), 997-1012.

Murali, A., Bhargava, A., & Wright, E. S. (2018). IDTAXA: A novel approach for accurate taxonomic classification of microbiome sequences. Microbiome, 6(1), 140.

Myers, G., & Pettigrew, S. (2018). A qualitative exploration of the factors underlying seniors' receptiveness to entomophagy. Food Research International, 103, 163–169.

Naulaerts, S., Meysman, P., Bittremieux, W., Vu, T. N., Vanden Berghe, W., Goethals, B., & Laukens, K. (2015). A primer to frequent itemset mining for bioinformatics. Briefings in Bioinformatics, 16(2), 216–231.

Naulaerts, S., Moens, S., Engelen, K., Berghe, W. V., Goethals, B., Laukens, K., & Meysman, P. (2016). Practical Approaches for Mining Frequent Patterns in Molecular Datasets. Bioinformatics and Biology Insights, 10, BBI.S38419.

Nelson, J. S., Grande, T. C., & Wilson, M. V. H. (2016). Fishes of the World: Nelson/Fishes of the World. John Wiley & Sons, Inc.

Nilsson, R. H., Anslan, S., Bahram, M., Wurzbacher, C., Baldrian, P., & Tedersoo, L. (2019). Mycobiome diversity: High-throughput sequencing and identification of fungi. Nature Reviews Microbiology, 17(2), 95–109.

Nilsson, R. H., Larsson, K.-H., Taylor, A. F. S., Bengtsson-Palme, J., Jeppesen, T. S., Schigel, D., Kennedy, P., Picard, K., Glöckner, F. O., Tedersoo, L., Saar, I., Kõljalg, U., & Abarenkov, K. (2019). The UNITE database for molecular identification of fungi: Handling dark taxa and parallel taxonomic classifications. Nucleic Acids Research, 47(D1), D259–D264.

Noor, E., Cherkaoui, S., & Sauer, U. (2019). Biological insights through omics data integration. Current Opinion in Systems Biology, 15, 39–47.

Núñez, A., Amo de Paz, G., Ferencova, Z., Rastrojo, A., Guantes, R., García, A. M., Alcamí, A., Gutiérrez-Bustillo, A. M., & Moreno, D. A. (2017). Validation of the hirst-type spore trap for simultaneous monitoring of prokaryotic and eukaryotic biodiversities in urban air samples by next-generation sequencing. Applied and Environmental Microbiology, 83(13), e00472-17.

O'Donnell, S. T., Ross, R. P., & Stanton, C. (2020). The Progress of Multi-Omics Technologies: Determining Function in Lactic Acid Bacteria Using a Systems Level Approach. Frontiers in Microbiology, 10, 3084.

Ogasawara, O., Kodama, Y., Mashima, J., Kosuge, T., & Fujisawa, T. (2020). DDBJ Database updates and computational infrastructure enhancement. Nucleic Acids Research, 48(D1), D45–D50.

Oliveira, F. S., Brestelli, J., Cade, S., Zheng, J., Iodice, J., Fischer, S., Aurrecoechea, C., Kissinger, J. C., Brunk, B. P., Stoeckert, C. J., Jr, Fernandes, G. R., Roos, D. S., & Beiting, D. P. (2018). MicrobiomeDB: A systems biology platform for integrating, mining and analyzing microbiome experiments. Nucleic Acids Research, 46(D1), D684–D691.

Omiecinski, E. R. (2003). Alternative interest measures for mining associations in databases. IEEE Transactions on Knowledge and Data Engineering, 15(1), 57–69.

Ong, H. F., Mustapha, N., Hamdan, H., Rosli, R., & Mustapha, A. (2020). Informative top-k class associative rule for cancer biomarker discovery on microarray data. Expert Systems with Applications, 146, 113169.

Oonincx, D. G. A. B., & de Boer, I. J. M. (2012). Environmental Impact of the Production of Mealworms as a Protein Source for Humans – A Life Cycle Assessment. PLoS ONE, 7(12), e51145.

Oonincx, D. G. A. B., Laurent, S., Veenenbos, M. E., & Loon, J. J. A. (2020). Dietary enrichment of edible insects with omega 3 fatty acids. Insect Science, 27(3), 500–509.

Osimani, A., Milanović, V., Cardinali, F., Roncolini, A., Garofalo, C., Clementi, F., Pasquini, M., Mozzon, M., Foligni, R., Raffaelli, N., Zamporlini, F., & Aquilanti, L. (2018). Bread enriched with cricket powder (Acheta domesticus): A technological, microbiological and nutritional evaluation. Innovative Food Science & Emerging Technologies, 48, 150–163.

Pali-Schöll, I., Verhoeckx, K., Mafra, I., Bavaro, S. L., Clare Mills, E. N., & Monaci, L. (2019). Allergenic and novel food proteins: State of the art and challenges in the allergenicity assessment. Trends in Food Science & Technology, 84, 45–48.

Parente, E., De Filippis, F., Ercolini, D., Ricciardi, A., & Zotta, T. (2019). Advancing integration of data on food microbiome studies: FoodMicrobionet 3.1, a major upgrade of the FoodMicrobionet database. International Journal of Food Microbiology, 305, 108249.

Parente, E., Zotta, T., Faust, K., De Filippis, F., & Ercolini, D. (2018). Structure of association networks in food bacterial communities. Food Microbiology, 73, 49–60.

Pasolli, E., De Filippis, F., Mauriello, I. E., Cumbo, F., Walsh, A. M., Leech, J., Cotter, P. D., Segata, N., & Ercolini, D. (2020). Large-scale genome-wide analysis links lactic acid bacteria from food with the gut microbiome. Nature Communications, 11(1), 2610.

Pasolli, E., Schiffer, L., Manghi, P., Renson, A., Obenchain, V., Truong, D. T., Beghini, F., Malik, F., Ramos, M., Dowd, J. B., Huttenhower, C., Morgan, M., Segata, N., & Waldron, L. (2017). Accessible, curated metagenomic data through ExperimentHub. Nature Methods, 14(11), 1023–1024.

Pasolli, E., Truong, D. T., Malik, F., Waldron, L., & Segata, N. (2016). Machine Learning Meta-analysis of Large Metagenomic Datasets: Tools and Biological Insights. PLOS Computational Biology, 12(7), e1004977.

Patel, S., Suleria, H. A. R., & Rauf, A. (2019). Edible insects as innovative foods: Nutritional and functional assessments. Trends in Food Science & Technology, 86, 352–359.

Pawlowski, J., Kelly-Quinn, M., Altermatt, F., Apothéloz-Perret-Gentil, L., Beja, P., Boggero, A., Borja, A., Bouchez, A., Cordier, T., & Domaizon, I. (2018). The future of biotic indices in the ecogenomic era: Integrating (e) DNA metabarcoding in biological assessment of aquatic ecosystems. Science of the Total Environment, 637, 1295–1310.

Payne, C. L. R., Scarborough, P., Rayner, M., & Nonaka, K. (2016). Are edible insects more or less 'healthy' than commonly consumed meats? A comparison using two nutrient profiling models developed to combat over- and undernutrition. European Journal of Clinical Nutrition, 70(3), 285–291.

Perez, F., & Granger, B. E. (2007). IPython: A System for Interactive Scientific Computing. Computing in Science & Engineering, 9(3), 21–29.

Petrakis, E. A., Cagliani, L. R., Polissiou, M. G., & Consonni, R. (2015). Evaluation of saffron (Crocus sativus L.) adulteration with plant adulterants by 1H NMR metabolite fingerprinting. Food Chemistry, 173, 890–896.

Pillay, T. S. (2020). Gene of the month: The 2019-nCoV/SARS-CoV-2 novel coronavirus spike protein. Journal of Clinical Pathology, 73(7), 366–369.

Pimm, S. L., Alibhai, S., Bergl, R., Dehgan, A., Giri, C., Jewell, Z., Joppa, L., Kays, R., & Loarie, S. (2015). Emerging Technologies to Conserve Biodiversity. Trends in Ecology & Evolution, 30(11), 685–696.

Piñol, J., Senar, M. A., & Symondson, W. O. C. (2019). The choice of universal primers and the characteristics of the species mixture determine when DNA metabarcoding can be quantitative. Molecular Ecology, 28(2), 407–419.

Piper, A. M., Batovska, J., Cogan, N. O., Weiss, J., Cunningham, J. P., Rodoni, B. C., & Blacket, M. J. (2019). Prospects and challenges of implementing DNA metabarcoding for high-throughput insect surveillance. GigaScience, 8(8), giz092.

Pirovano, W., Boetzer, M., Derks, M. F. L., & Smit, S. (2015). NCBI-compliant genome submissions: Tips and tricks to save time and money: Table 1. Briefings in Bioinformatics, bbv104.

Pollock, J., Glendinning, L., Wisedchanwet, T., & Watson, M. (n.d.). The Madness of Microbiome: Attempting To Find Consensus “Best Practice” for 16S Microbiome Studies. Applied and Environmental Microbiology, 84(7), e02627-17.

Ponsero, A. J., Bomhoff, M., Blumberg, K., Youens-Clark, K., Herz, N. M., Wood-Charlson, E. M., Delong, E. F., & Hurwitz, B. L. (2021). Planet Microbe: A platform for marine microbiology to discover and analyze interconnected ’omics and environmental data. Nucleic Acids Research, 49(D1), D792–D802.

Poore, G. D., Kopylova, E., Zhu, Q., Carpenter, C., Fraraccio, S., Wandro, S., Kosciolek, T., Janssen, S., Metcalf, J., Song, S. J., Kanbar, J., Miller-Montgomery, S., Heaton, R., Mckay, R., Patel, S. P., Swafford, A. D., & Knight, R. (2020). Microbiome analyses of blood and tissues suggest cancer diagnostic approach. Nature, 579(7800), 567–574.

Porter, T. M., & Hajibabaei, M. (2018). Scaling up: A guide to high-throughput genomic approaches for biodiversity analysis. Molecular Ecology, 27(2), 313–338.

Prescott, S. L., Larcombe, D.-L., Logan, A. C., West, C., Burks, W., Caraballo, L., Levin, M., Etten, E. V., Horwitz, P., Kozyrskyj, A., & Campbell, D. E. (2017). The skin microbiome: Impact of modern environments on skin ecology, barrier integrity, and systemic immune programming. World Allergy Organization Journal, 10, 29.

Proctor, L. M., Creasy, H. H., Fettweis, J. M., Lloyd-Price, J., Mahurkar, A., Zhou, W., Buck, G. A., Snyder, M. P., Strauss, J. F., Weinstock, G. M., White, O., Huttenhower, C., & The Integrative HMP (iHMP) Research Network Consortium. (2019). The Integrative Human Microbiome Project. Nature, 569(7758), 641–648.

Pruesse, E., Quast, C., Knittel, K., Fuchs, B. M., Ludwig, W., Peplies, J., & Glockner, F. O. (2007). SILVA: A comprehensive online resource for quality checked and aligned ribosomal RNA sequence data compatible with ARB. Nucleic Acids Research, 35(21), 7188–7196.

Qu, K., Guo, F., Liu, X., Lin, Y., & Zou, Q. (2019). Application of Machine Learning in Microbiology. Frontiers in Microbiology, 10, 827.

Quast, C., Pruesse, E., Yilmaz, P., Gerken, J., Schweer, T., Yarza, P., Peplies, J., & Glöckner, F. O. (2013). The SILVA ribosomal RNA gene database project: Improved data processing and web-based tools. Nucleic Acids Research, 41(D1), D590–D596.

Quéméré, E., Hibert, F., Miquel, C., Lhuillier, E., Rasolondraibe, E., Champeau, J., Rabarivola, C., Nusbaumer, L., Chatelain, C., & Gautier, L. (2013). A DNA

metabarcoding study of a primate dietary diversity and plasticity across its entire fragmented range. PloS One, 8(3), e58971.

Quince, C., Walker, A. W., Simpson, J. T., Loman, N. J., & Segata, N. (2017). Shotgun metagenomics, from sampling to analysis. Nature Biotechnology, 35(9), 833–844.

Raschka, S. (2018). MLxtend: Providing machine learning and data science utilities and extensions to Python's scientific computing stack. Journal of Open Source Software, 3(24), 638.

Rasmussen, R. S., & Morrissey, M. T. (2008). DNA-Based Methods for the Identification of Commercial Fish and Seafood Species. Comprehensive Reviews in Food Science and Food Safety, 7(3), 280–295.

Ratnasingham, S., & Hebert, P. D. N. (2007). BARCODING: Bold: The Barcode of Life Data System (http://www.barcodinglife.org): BARCODING. Molecular Ecology Notes, 7(3), 355–364.

Ravel, J., Gajer, P., Abdo, Z., Schneider, G. M., Koenig, S. S. K., McCulle, S. L., Karlebach, S., Gorle, R., Russell, J., Tacket, C. O., Brotman, R. M., Davis, C. C., Ault, K., Peralta, L., & Forney, L. J. (2011). Vaginal microbiome of reproductive-age women. Proceedings of the National Academy of Sciences, 108(Supplement_1), 4680–4687.

Reavie, E. D., Jicha, T. M., Angradi, T. R., Bolgrien, D. W., & Hill, B. H. (2010). Algal assemblages for large river monitoring: Comparison among biovolume, absolute and relative abundance metrics. Ecological Indicators, 10(2), 167–177.

Reback, J., McKinney, W., Den Van Bossche, J., Augspurger, T., Cloud, P., Klein, A., Roeschke, M., Hawkins, S., Tratner, J., & She, C. (2020). Pandas-dev/pandas: Pandas 1.0. 3. Zenodo.

Reese, A. T., Madden, A. A., Joossens, M., Lacaze, G., & Dunn, R. R. (2020). Influences of Ingredients and Bakers on the Bacteria and Fungi in Sourdough Starters and Bread. MSphere, 5(1).

Regulation (Ec) No 1069/2009 of the European Parliament and of the council of 11 December 2015. Available online https://eur-lex.europa.eu/legal-content/EN/TXT/ PDF/?uri=OJ:L:2015:327:FULL&from=EN (accessed on 19 April 2019).

REGULATION (EU) 2015/ 2283 OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL - of 25 November 2015 - on novel foods, amending Regulation (EU) No 1169/ 2011 of the European Parliament and of the Council and repealing Regulation (EC) No 258/ 97 of the European Parliament and of the Council and Commission Regulation (EC) No 1852/ 2001. Available from: https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:32015R2283&from=IT.

Regulation (EU) 2017/893 of the European Parliament and of the council of 24 May 2017. Available online https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri= CELEX:32017R0893&from=en (accessed on 19 April 2019).

Regulation (EU) No 828/2014 of 30 July 2014. Available online https://eur-lex.europa. eu/legal-content/EN/TXT/PDF/?uri=CELEX:32014R0828&from=IT (accessed on 19 April 2019).

Rhie, A., McCarthy, S. A., Fedrigo, O., Damas, J., Formenti, G., Koren, S., Uliano-Silva, M., Chow, W., Fungtammasan, A., & Kim, J. (2021). Towards complete and error-free genome assemblies of all vertebrate species. Nature, 592(7856), 737–746.

Ribeiro, J. C., Cunha, L. M., Sousa-Pinto, B., & Fonseca, J. (2018). Allergic risks of consuming edible insects: A systematic review. Molecular Nutrition & Food Research, 62(1), 1700030.

Richardson, R. T., Curtis, H. R., Matcham, E. G., Lin, C., Suresh, S., Sponsler, D. B., Hearon, L. E., & Johnson, R. M. (2019). Quantitative multi-locus metabarcoding and waggle dance interpretation reveal honey bee spring foraging patterns in Midwest agroecosystems. Molecular Ecology, 28(3), 686–697.

Risk profile related to production and consumption of insects as food and feed. (n.d.). EFSA Journal, 2015;13(10):4257.

Rognes, T., Flouri, T., Nichols, B., Quince, C., & Mahé, F. (2016). VSEARCH: A versatile open source tool for metagenomics. PeerJ, 4, e2584.

Roncolini, A., Milanović, V., Aquilanti, L., Cardinali, F., Garofalo, C., Sabbatini, R., Clementi, F., Belleggia, L., Pasquini, M., Mozzon, M., Foligni, R., Federica Trombetta, M., Haouet, M. N., Serena Altissimi, M., Di Bella, S., Piersanti, A., Griffoni, F., Reale, A., Niro, S., & Osimani, A. (2020). Lesser mealworm (Alphitobius diaperinus) powder as a novel baking ingredient for manufacturing high-protein, mineral-dense snacks. Food Research International, 131, 109031.

Roncolini, A., Milanović, V., Cardinali, F., Osimani, A., Garofalo, C., Sabbatini, R., Clementi, F., Pasquini, M., Mozzon, M., Foligni, R., Raffaelli, N., Zamporlini, F., Minazzato, G., Trombetta, M. F., Van Buitenen, A., Van Campenhout, L., &

Aquilanti, L. (2019). Protein fortification with mealworm (Tenebrio molitor L.) powder: Effect on textural, microbiological, nutritional and sensory features of bread. PLOS ONE, 14(2), e0211747.

Roses, A. D. (2003). The genome era begins... Nature Genetics, 33(S3), 217–217.

Ross, A. A., Müller, K. M., Weese, J. S., & Neufeld, J. D. (2018). Comprehensive skin microbiome analysis reveals the uniqueness of human skin and evidence for phylosymbiosis within the class Mammalia. Proceedings of the National Academy of Sciences, 115(25), E5786–E5795.

Rumpold, B. A., & Schlüter, O. K. (2013). Nutritional composition and safety aspects of edible insects. Molecular Nutrition & Food Research, 57(5), 802–823.

Ruppert, K. M., Kline, R. J., & Rahman, M. S. (2019). Past, present, and future perspectives of environmental DNA (eDNA) metabarcoding: A systematic review in methods, monitoring, and applications of global eDNA. Global Ecology and Conservation, 17, e00547.

Ryan, S. F., Adamson, N. L., Aktipis, A., Andersen, L. K., Austin, R., Barnes, L., Beasley, M. R., Bedell, K. D., Briggs, S., Chapman, B., Cooper, C. B., Corn, J. O., Creamer, N. G., Delborne, J. A., Domenico, P., Driscoll, E., Goodwin, J., Hjarding, A., Hulbert, J. M., … Dunn, R. R. (2018). The role of citizen science in addressing grand challenges in food and agriculture research. Proceedings of the Royal Society B: Biological Sciences, 285(1891), 20181977.

Sa, D., Sriharsha, M., Rk, N., Sn, K., & Trk, S. (2015). Role of Diet in Dermatological Conditions.

Salvatori, G., Luberto, L., Maffei, M., Aurisicchio, L., Roscilli, G., Palombo, F., & Marra, E. (2020). SARS-CoV-2 SPIKE PROTEIN: An optimal immunological target for vaccines. Journal of Translational Medicine, 18(1), 222.

Sankaran, K., & Holmes, S. P. (2019). Multitable Methods for Microbiome Data Integration. Frontiers in Genetics, 10, 627.

Sayers, E. W., Agarwala, R., Bolton, E. E., Brister, J. R., Canese, K., Clark, K., Connor, R., Fiorini, N., Funk, K., Hefferon, T., Holmes, J. B., Kim, S., Kimchi, A., Kitts, P. A., Lathrop, S., Lu, Z., Madden, T. L., Marchler-Bauer, A., Phan, L., … Ostell, J. (2019). Database resources of the National Center for Biotechnology Information. Nucleic Acids Research, 47(D1), D23–D28.

Schloss, P. D., & Westcott, S. L. (2011). Assessing and improving methods used in operational taxonomic unit-based approaches for 16S rRNA gene sequence analysis. Applied and Environmental Microbiology, 77(10), 3219–3226.

Schlüter, O., Rumpold, B., Holzhauser, T., Roth, A., Vogel, R. F., Quasigroch, W., Vogel, S., Heinz, V., Jäger, H., Bandick, N., Kulling, S., Knorr, D., Steinberg, P., & Engel, K.-H. (2017). Safety aspects of the production of foods and food ingredients from insects. Molecular Nutrition & Food Research, 61(6), 1600520.

Seekatz, A. M., Theriot, C. M., Rao, K., Chang, Y.-M., Freeman, A. E., Kao, J. Y., & Young, V. B. (2018). Restoration of short chain fatty acid and bile acid metabolism following fecal microbiota transplantation in patients with recurrent Clostridium difficile infection. Anaerobe, 53, 64–73.

Segata, N., Haake, S., Mannon, P., Lemon, K. P., Waldron, L., Gevers, D., Huttenhower, C., & Izard, J. (2012). Composition of the adult digestive tract

bacterial microbiome based on seven mouth surfaces, tonsils, throat and stool samples. Genome Biology, 13(6), R42.

Shakya, M., Lo, C.-C., & Chain, P. S. G. (2019). Advances and Challenges in Metatranscriptomic Analysis. Frontiers in Genetics, 10, 904.

Shetty, S. A., Smidt, H., & de Vos, W. M. (2019). Reconstructing functional networks in the human intestinal tract using synthetic microbiomes. Current Opinion in Biotechnology, 58, 146–154.

Shi, W., Qi, H., Sun, Q., Fan, G., Liu, S., Wang, J., Zhu, B., Liu, H., Zhao, F., Wang, X., Hu, X., Li, W., Liu, J., Tian, Y., Wu, L., & Ma, J. (2019). gcMeta: A Global Catalogue of Metagenomics platform to support the archiving, standardization and analysis of microbiome data. Nucleic Acids Research, 47(D1), D637–D648.

Shokralla, S., Spall, J. L., Gibson, J. F., & Hajibabaei, M. (2012). Next-generation sequencing technologies for environmental DNA research. Molecular Ecology, 21(8), 1794–1805.

Sirén, K., Mak, S. S. T., Fischer, U., Hansen, L. H., & Gilbert, M. T. P. (2019). Multi-omics and potential applications in wine production. Current Opinion in Biotechnology, 56, 172–178.

Smid, E. J., & Lacroix, C. (2013). Microbe–microbe interactions in mixed culture food fermentations. Current Opinion in Biotechnology, 24(2), 148–154.

Solden, L., Lloyd, K., & Wrighton, K. (2016). The bright side of microbial dark matter: Lessons learned from the uncultivated majority. Current Opinion in Microbiology, 31, 217–226.

Song, Y., Garg, S., Girotra, M., Maddox, C., von Rosenvinge, E. C., Dutta, A., Dutta, S., & Fricke, W. F. (2013). Microbiota Dynamics in Patients Treated with Fecal Microbiota Transplantation for Recurrent Clostridium difficile Infection. PLoS ONE, 8(11), e81330.

Srivastava, D., Baksi, K. D., Kuntal, B. K., & Mande, S. S. (2019). “EviMass”: A Literature Evidence-Based Miner for Human Microbial Associations. Frontiers in Genetics, 10, 849.

Star, B., Nederbragt, A. J., Jentoft, S., Grimholt, U., Malmstrøm, M., Gregers, T. F., Rounge, T. B., Paulsen, J., Solbakken, M. H., Sharma, A., Wetten, O. F., Lanzén, A., Winer, R., Knight, J., Vogel, J.-H., Aken, B., Andersen, Ø., Lagesen, K., Tooming-Klunderud, A., … Jakobsen, K. S. (2011). The genome sequence of Atlantic cod reveals a unique immune system. Nature, 477(7363), 207–210.

Stoops, J., Crauwels, S., Waud, M., Claes, J., Lievens, B., & Van Campenhout, L. (2016). Microbial community assessment of mealworm larvae (Tenebrio molitor) and grasshoppers (Locusta migratoria migratorioides) sold for human consumption. Food Microbiology, 53, 122–127.

Stoops, J., Vandeweyer, D., Crauwels, S., Verreth, C., Boeckx, H., Van Der Borght, M., Claes, J., Lievens, B., & Van Campenhout, L. (2017). Minced meat-like products from mealworm larvae (Tenebrio molitor and Alphitobius diaperinus): Microbial dynamics during production and storage. Innovative Food Science & Emerging Technologies, 41, 1–9.

Stull, V. J., Finer, E., Bergmans, R. S., Febvre, H. P., Longhurst, C., Manter, D. K., Patz, J. A., & Weir, T. L. (2018). Impact of Edible Cricket Consumption on Gut Microbiota in Healthy Adults, a Double-blind, Randomized Crossover Trial. Scientific Reports, 8(1), 10762.

Su, X., Jing, G., Zhang, Y., & Wu, S. (2020). Method development for cross-study microbiome data mining: Challenges and opportunities. Computational and Structural Biotechnology Journal, 18, 2075–2080.

Sun-Waterhouse, D., Waterhouse, G. I. N., You, L., Zhang, J., Liu, Y., Ma, L., Gao, J., & Dong, Y. (2016). Transforming insect biomass into consumer wellness foods: A review. Food Research International, 89, 129–151.

Swaney, M. H., & Kalan, L. R. (n.d.). Living in Your Skin: Microbes, Molecules, and Mechanisms. Infection and Immunity, 89(4), e00695-20.

Sze, M. A., & Schloss, P. D. (n.d.). Looking for a Signal in the Noise: Revisiting Obesity and the Microbiome. MBio, 7(4), e01018-16.

Taberlet, P., Gielly, L., Pautou, G., & Bouvet, J. (1991). Universal primers for amplification of three non-coding regions of chloroplast DNA. Plant Molecular Biology, 17(5), 1105–1109.

Tamang, J. P., Shin, D.-H., Jung, S.-J., & Chae, S.-W. (2016). Functional Properties of Microorganisms in Fermented Foods. Frontiers in Microbiology, 7.

Tan, P.-N., Kumar, V., & Srivastava, J. (2002). Selecting the right interestingness measure for association patterns. Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining - KDD '02, 32.

Tandon, D., Haque, M. M., & Mande, S. S. (2016). Inferring Intra-Community Microbial Interaction Patterns from Metagenomic Datasets Using Associative Rule Mining Techniques. PLOS ONE, 11(4), e0154493.

Tang, L., Zhang, L., Luo, P., & Wang, M. (2012). Incorporating occupancy into frequent pattern mining for high quality pattern recommendation. Proceedings of the 21st ACM International Conference on Information and Knowledge Management - CIKM ’12, 75.

Tatti, N., & Mampaey, M. (2010). Using background knowledge to rank itemsets. Data Mining and Knowledge Discovery, 21(2), 293–309.

Tauber, H. (1974). A static non-overload pollen collector. New Phytologist, 73(2), 359–369.

Taylor, B. C., Lejzerowicz, F., Poirel, M., Shaffer, J. P., Jiang, L., Aksenov, A., Litwin, N., Humphrey, G., Martino, C., Miller-Montgomery, S., Dorrestein, P. C., Veiga, P., Song, S. J., McDonald, D., Derrien, M., & Knight, R. (2020). Consumption of Fermented Foods Is Associated with Systematic Differences in the Gut Microbiome and Metabolome. MSystems, 5(2).

Taylor, H. R., & Gemmell, N. J. (2016). Emerging Technologies to Conserve Biodiversity: Further Opportunities via Genomics. Response to Pimm et al. Trends in Ecology & Evolution, 31(3), 171–172.

Thompson, J., Johansen, R., Dunbar, J., & Munsky, B. (2019). Machine learning to predict microbial community functions: An analysis of dissolved organic carbon from litter decomposition. PLOS ONE, 14(7), e0215502.

Thomsen, P. F., & Willerslev, E. (2015). Environmental DNA–An emerging tool in conservation for monitoring past and present biodiversity. Biological Conservation, 183, 4–18.

Thudi, M., Li, Y., Jackson, S. A., May, G. D., & Varshney, R. K. (2012). Current state-of-art of sequencing technologies for plant genomics research. Briefings in Functional Genomics, 11(1), 3–11.

Tinker, K. A., & Ottesen, E. A. (2018). The hindgut microbiota of praying mantids is highly variable and includes both prey-associated and host-specific microbes. PLOS ONE, 13(12), e0208917.

Toju, H., Tanabe, A. S., Yamamoto, S., & Sato, H. (2012). High-coverage ITS primers for the DNA-based identification of ascomycetes and basidiomycetes in environmental samples. PloS One, 7(7), e40863.

Tommasi, N., Biella, P., Guzzetti, L., Lasway, J. V., Njovu, H. K., Tapparo, A., Agostinetto, G., Peters, M. K., Steffan-Dewenter, I., & Labra, M. (2021). Impact of land use intensification and local features on plants and pollinators in Sub-Saharan smallholder farms. Agriculture, Ecosystems & Environment, 319, 107560.

Tramuta, C., Gallina, S., Bellio, A., Bianchi, D. M., Chiesa, F., Rubiola, S., Romano, A., & Decastelli, L. (2018). A Set of Multiplex Polymerase Chain Reactions for Genomic Detection of Nine Edible Insect Species in Foods. Journal of Insect Science, 18(5).

Trebitz, A. S., Hoffman, J. C., Darling, J. A., Pilgrim, E. M., Kelly, J. R., Brown, E. A., Chadderton, W. L., Egan, S. P., Grey, E. K., & Hashsham, S. A. (2017). Early detection monitoring for aquatic non-indigenous species: Optimizing surveillance, incorporating advanced technologies, and identifying research needs. Journal of Environmental Management, 202, 299–310.

Turnbaugh, P. J., Ley, R. E., Hamady, M., Fraser-Liggett, C. M., Knight, R., & Gordon, J. I. (2007). The Human Microbiome Project. Nature, 449(7164), 804–810.

Valdés, I., García, E., Llorente, M., & Méndez, E. (2003). Innovative approach to low-level gluten determination in foods using a novel sandwich enzyme-linked immunosorbent assay protocol: European Journal of Gastroenterology & Hepatology, 15(5), 465–747.

Valsecchi, E., Coppola, E., Pires, R., Parmegiani, A., Casiraghi, M., Galli, P., & Bruno, A. (2021). Newly developed ad hoc molecular assays shows how eDNA can witness and anticipate the monk seal recolonization of central Mediterranean. BioRxiv.

van Broekhoven, S., Oonincx, D. G. A. B., van Huis, A., & van Loon, J. J. A. (2015). Growth performance and feed conversion efficiency of three edible mealworm species (Coleoptera: Tenebrionidae) on diets composed of organic by-products. Journal of Insect Physiology, 73, 1–10.

van der Fels-Klerx, H. J., Adamse, P., Punt, A., & van Asselt, E. (2018). Data Analyses and Modelling for Risk Based Monitoring of Mycotoxins in Animal Feed. Toxins, 10(2), 54.

van der Heyde, M., Bunce, M., Wardell-Johnson, G., Fernandes, K., White, N. E., & Nevill, P. (2020). Testing multiple substrates for terrestrial biodiversity monitoring using environmental DNA metabarcoding. Molecular Ecology Resources, 20(3), 732–745.

van der Spiegel, M., Noordam, M. Y., & van der Fels-Klerx, H. J. (2013). Safety of Novel Protein Sources (Insects, Microalgae, Seaweed, Duckweed, and Rapeseed) and Legislative Aspects for Their Application in Food and Feed

Production: Safety aspects of novel protein sources…. Comprehensive Reviews in Food Science and Food Safety, 12(6), 662–678.

van Dorst, J., Bissett, A., Palmer, A. S., Brown, M., Snape, I., Stark, J. S., Raymond, B., McKinlay, J., Ji, M., & Winsley, T. (2014). Community fingerprinting in a sequencing world. FEMS Microbiology Ecology, 89(2), 316–330.

van Huis, A. (2013). Potential of Insects as Food and Feed in Assuring Food Security. Annual Review of Entomology, 58(1), 563–583.

Van Reckem, E., Geeraerts, W., Charmpi, C., Van der Veken, D., De Vuyst, L., & Leroy, F. (2019). Exploring the Link Between the Geographical Origin of European Fermented Foods and the Diversity of Their Bacterial Communities: The Case of Fermented Meats. Frontiers in Microbiology, 10, 2302.

Vandeweyer, D., Crauwels, S., Lievens, B., & Van Campenhout, L. (2017). Metagenetic analysis of the bacterial communities of edible insects from diverse production cycles at industrial rearing companies. International Journal of Food Microbiology, 261, 11–18.

Vangay, P., Burgin, J., Johnston, A., Beck, K. L., Berrios, D. C., Blumberg, K., Canon, S., Chain, P., Chandonia, J. M., Christianson, D., Costes, S. V., Damerow, J., Duncan, W. D., Dundore-Arias, J. P., Fagnan, K., Galazka, J. M., Gibbons, S. M., Hays, D., Hervey, J., … Eloe-Fadrosh, E. A. (2021). Microbiome metadata standards: Report of the national microbiome data collaborative's workshop and follow-on activities. MSystems, 6(1), e01194.

Vangay, P., Hillmann, B. M., & Knights, D. (2019). Microbiome Learning Repo (ML Repo): A public repository of microbiome regression and classification tasks. GigaScience, 8(5).

Vasselon, V., Rimet, F., Tapolczai, K., & Bouchez, A. (2017). Assessing ecological status with diatoms DNA metabarcoding: Scaling-up on a WFD monitoring network (Mayotte island, France). Ecological Indicators, 82, 1–12.

Virtanen, P., Gommers, R., Oliphant, T. E., Haberland, M., Reddy, T., Cournapeau, D., Burovski, E., Peterson, P., Weckesser, W., Bright, J., van der Walt, S. J., Brett, M., Wilson, J., Millman, K. J., Mayorov, N., Nelson, A. R. J., Jones, E., Kern, R., Larson, E., … Vázquez-Baeza, Y. (2020). SciPy 1.0: Fundamental algorithms for scientific computing in Python. Nature Methods, 17(3), 261–272.

Walia, K., Kapoor, A., & Farber, J. M. (2018). Qualitative risk assessment of cricket powder to be used to treat undernutrition in infants and children in Cambodia. Food Control, 92, 169–182.

Wan, Y., Shang, J., Graham, R., Baric, R. S., & Li, F. (2020). Receptor Recognition by the Novel Coronavirus from Wuhan: An Analysis Based on Decade-Long Structural Studies of SARS Coronavirus. Journal of Virology, 94(7).

Wang, Q., Garrity, G. M., Tiedje, J. M., & Cole, J. R. (2007). Naïve Bayesian Classifier for Rapid Assignment of rRNA Sequences into the New Bacterial Taxonomy. Applied and Environmental Microbiology, 73(16), 5261–5267.

Wang, Z., Lachmann, A., & Ma'ayan, A. (2019). Mining data and metadata from the gene expression omnibus. Biophysical Reviews, 11(1), 103–110.

Waterhouse, R. M., Adam-Blondon, A.-F., Agosti, D., Baldrian, P., Balech, B., Corre, E., Davey, R. P., Lantz, H., Pesole, G., & Quast, C. (2021).

Recommendations for connecting molecular sequence and biodiversity research infrastructures through ELIXIR. F1000Research, 10(1238), 1238.

Watson, J. D. (1990). The human genome project: Past, present, and future. Science, 248(4951), 44–49.

Weigand, H., Beermann, A. J., Čiampor, F., Costa, F. O., Csabai, Z., Duarte, S., Geiger, M. F., Grabowski, M., Rimet, F., & Rulik, B. (2019). DNA barcode reference libraries for the monitoring of aquatic biota in Europe: Gap-analysis and recommendations for future work. Science of the Total Environment, 678, 499–524.

Weiss, S., Van Treuren, W., Lozupone, C., Faust, K., Friedman, J., Deng, Y., Xia, L. C., Xu, Z. Z., Ursell, L., Alm, E. J., Birmingham, A., Cram, J. A., Fuhrman, J. A., Raes, J., Sun, F., Zhou, J., & Knight, R. (2016). Correlation detection strategies in microbial data sets vary widely in sensitivity and precision. The ISME Journal, 10(7), 1669–1681.

Wen, C., Wu, L., Qin, Y., Van Nostrand, J. D., Ning, D., Sun, B., Xue, K., Liu, F., Deng, Y., Liang, Y., & Zhou, J. (2017). Evaluation of the reproducibility of amplicon sequencing with Illumina MiSeq platform. PLOS ONE, 12(4), e0176716.

Westfall, K. M., Therriault, T. W., & Abbott, C. L. (2020). A new approach to molecular biosurveillance of invasive species using DNA metabarcoding. Global Change Biology, 26(2), 1012–1022.

White, T. J., Bruns, T., Lee, S., & Taylor, J. (1990). Amplification and direct sequencing of fungal ribosomal RNA genes for phylogenetics. PCR Protocols: A Guide to Methods and Applications, 18(1), 315–322.

White, E. P., Baldridge, E., Brym, Z. T., Locey, K. J., McGlinn, D. J., & Supp, S. R. (2013). Nine simple ways to make it easier to (re) use your data. Ideas in Ecology and Evolution, 6(2).

Wilke, A., Bischof, J., Gerlach, W., Glass, E., Harrison, T., Keegan, K. P., Paczian, T., Trimble, W. L., Bagchi, S., Grama, A., Chaterji, S., & Meyer, F. (2016). The MG-RAST metagenomics database and portal in 2015. Nucleic Acids Research, 44(D1), D590–D594.

Wilkinson, M. D., Dumontier, M., Aalbersberg, Ij. J., Appleton, G., Axton, M., Baak, A., Blomberg, N., Boiten, J.-W., da Silva Santos, L. B., Bourne, P. E., Bouwman, J., Brookes, A. J., Clark, T., Crosas, M., Dillo, I., Dumon, O., Edmunds, S., Evelo, C. T., Finkers, R., … Mons, B. (2016). The FAIR Guiding Principles for scientific data management and stewardship. Scientific Data, 3(1), 160018.

Williams, M. R., & Gallo, R. L. (2015). The Role of the Skin Microbiome in Atopic Dermatitis. Current Allergy and Asthma Reports, 15(11), 65.

Wolfe, B. E., & Dutton, R. J. (2013). Towards an Ecosystem Approach to Cheese Microbiology. Microbiology Spectrum, 1(1).

Wong, E. H.-K., & Hanner, R. H. (2008). DNA barcoding detects market substitution in North American seafood. Food Research International, 41(8), 828–837.

Wood-Charlson, E. M., Anubhav, Auberry, D., Blanco, H., Borkum, M. I., Corilo, Y. E., Davenport, K. W., Deshpande, S., Devarakonda, R., Drake, M., Duncan, W. D., Flynn, M. C., Hays, D., Hu, B., Huntemann, M., Li, P.-E., Lipton, M., Lo, C.-C., Millard, D., … Eloe-Fadrosh, E. A. (2020). The National Microbiome Data

Collaborative: Enabling microbiome science. Nature Reviews Microbiology, 18(6), 313–314.

Wu, F., Zhao, S., Yu, B., Chen, Y.-M., Wang, W., Song, Z.-G., Hu, Y., Tao, Z.-W., Tian, J.-H., Pei, Y.-Y., Yuan, M.-L., Zhang, Y.-L., Dai, F.-H., Liu, Y., Wang, Q.-M., Zheng, J.-J., Xu, L., Holmes, E. C., & Zhang, Y.-Z. (2020). A new coronavirus associated with human respiratory disease in China. Nature, 579(7798), 265–269.

Wu, S., Sun, C., Li, Y., Wang, T., Jia, L., Lai, S., Yang, Y., Luo, P., Dai, D., Yang, Y.-Q., Luo, Q., Gao, N. L., Ning, K., He, L., Zhao, X.-M., & Chen, W.-H. (2020). GMrepo: A database of curated and consistently annotated human gut metagenomes. Nucleic Acids Research, 48(D1), D545–D553.

Wynants, E., Crauwels, S., Lievens, B., Luca, S., Claes, J., Borremans, A., Bruyninckx, L., & Van Campenhout, L. (2017). Effect of post-harvest starvation and rinsing on the microbial numbers and the bacterial community composition of mealworm larvae (Tenebrio molitor). Innovative Food Science & Emerging Technologies, 42, 8–15.

Xiao, S., Fei, N., Pang, X., Shen, J., Wang, L., Zhang, B., Zhang, M., Zhang, X., Zhang, C., Li, M., Sun, L., Xue, Z., Wang, J., Feng, J., Yan, F., Zhao, N., Liu, J., Long, W., & Zhao, L. (2014). A gut microbiota-targeted dietary intervention for amelioration of chronic inflammation underlying metabolic syndrome. FEMS Microbiology Ecology, 87(2), 357–367.

Xie, M., An, F., Wu, J., Liu, Y., Shi, H., & Wu, R. (2019). Meta-omics reveal microbial assortments and key enzymes in bean sauce mash, a traditional fermented soybean product. Journal of the Science of Food and Agriculture, 99(14), 6522–6534.

Xie, M., Wu, J., An, F., Yue, X., Tao, D., Wu, R., & Lee, Y. (2019). An integrated metagenomic/metaproteomic investigation of microbiota in dajiang-meju, a traditional fermented soybean product in Northeast China. Food Research International, 115, 414–424.

Xiong, H., Tan, P.-N., & Kumar, V. (2006). Hyperclique pattern discovery. Data Mining and Knowledge Discovery, 13(2), 219–242.

Yancy, H. F., Zemlak, T. S., Mason, J. A., Washington, J. D., Tenge, B. J., Nguyen, N.-L. T., Barnett, J. D., Savary, W. E., Hill, W. E., Moore, M. M., Fry, F. S., Randolph, S. C., Rogers, P. L., & Hebert, P. D. N. (2008). Potential Use of DNA Barcodes in Regulatory Science: Applications of the Regulatory Fish Encyclopedia. Journal of Food Protection, 71(1), 210–217.

Yilmaz, P., Kottmann, R., Field, D., Knight, R., Cole, J. R., Amaral-Zettler, L., Gilbert, J. A., Karsch-Mizrachi, I., Johnston, A., Cochrane, G., Vaughan, R., Hunter, C., Park, J., Morrison, N., Rocca-Serra, P., Sterk, P., Arumugam, M., Bailey, M., Baumgartner, L., … Glöckner, F. O. (2011). Minimum information about a marker gene sequence (MIMARKS) and minimum information about any (x) sequence (MIxS) specifications. Nature Biotechnology, 29(5), 415–420.

Yoon, Y., & Lee, G. G. (2012). Subcellular Localization Prediction through Boosting Association Rules. IEEE/ACM Transactions on Computational Biology and Bioinformatics, 9(2), 609–618.

Yu, N., Wei, Y., Zhang, X., Zhu, N., Wang, Y., Zhu, Y., Zhang, H., Li, F., Yang, L., Sun, J., & Sun, A. (2017). Barcode ITS2: A useful tool for identifying Trachelospermum jasminoides and a good monitor for medicine market. Scientific Reports, 7(1), 5037.

Zafeiropoulos, H., Viet, H. Q., Vasileiadou, K., Potirakis, A., Arvanitidis, C., Topalis, P., Pavloudi, C., & Pafilis, E. (2020). PEMA: a flexible Pipeline for Environmental DNA Metabarcoding Analysis of the 16S/18S ribosomal RNA, ITS, and COI marker genes. GigaScience, 9(3), giaa022.

Zakrzewski, M., Proietti, C., Ellis, J. J., Hasan, S., Brion, M.-J., Berger, B., & Krause, L. (2016). Calypso: A user-friendly web-server for mining and visualizing microbiome–environment interactions. Bioinformatics, btw725.

Zhang, G. K., Chain, F. J. J., Abbott, C. L., & Cristescu, M. E. (2018). Metabarcoding using multiplexed markers increases species detection in complex zooplankton communities. Evolutionary Applications, 11(10), 1901–1914.

Zhang, J., Kobert, K., Flouri, T., & Stamatakis, A. (2014). PEAR: A fast and accurate Illumina Paired-End reAd mergeR. Bioinformatics, 30(5), 614–620.

Zhang, S., Zhao, J., & Yao, M. (2020). A comprehensive and comparative evaluation of primers for metabarcoding eDNA from fish. Methods in Ecology and Evolution, 11(12), 1609–1625.

Zhao, M., Su, X. Q., Nian, B., Chen, L. J., Zhang, D. L., Duan, S. M., Wang, L. Y., Shi, X. Y., Jiang, B., Jiang, W. W., Lv, C. Y., Wang, D. P., Shi, Y., Xiao, Y., Wu, J.-L., Pan, Y. H., & Ma, Y. (2019). Integrated Meta-omics Approaches To Understand the Microbiome of Spontaneous Fermentation of Traditional Chinese Pu-erh Tea. MSystems, 4(6).

Zhou, C., Meysman, P., Cule, B., Laukens, K., & Goethals, B. (2013). Mining spatially cohesive itemsets in protein molecular structures. Proceedings of the 12th International Workshop on Data Mining in Bioinformatics - BioKDD ’13, 42–50.

Zhu, Y., Stephens, R. M., Meltzer, P. S., & Davis, S. R. (2013). SRAdb: Query and use public next-generation sequencing data from within R. BMC Bioinformatics, 14(1), 19.

Zimmermann, J., Glöckner, G., Jahn, R., Enke, N., & Gemeinholzer, B. (2015). Metabarcoding vs. Morphological identification to assess diatom diversity in environmental studies. Molecular Ecology Resources, 15(3), 526–542.

Acknowledgments

Three years have gone by quickly, but the journey was intense. Firstly, I would like to express my sincere gratitude to my tutor Prof. Maurizio Casiraghi for his continuous support, for his advice, motivation, and immense knowledge. From the first day in class, I learned to be open minded, happy and proud to be part of the research world. Without his lessons and qualities, my path would have been different. I also would like to thank my co-tutor, Prof. Dario Pescini, for his support and the hard question which encouraged me to widen my research from various perspectives. With his help, I was able to approach new things, learning to interface with people with a different background. In addition, I have had the opportunity to participate in different projects, such as the last two editions of the BioHackathon Europe, getting to know new people and enjoying beautiful moments of both work and fun.

My sincere thanks also goes to Dr. Anna Sandionigi, who provided me an opportunity to join her activities also during my master degree. She helped me approach the world of bioinformatics and research for the first time, and she allowed me to meet new and pleasant people. In general, I would like to thank Maurizio, Dario and Anna for giving me the chance to be myself and dedicate myself to the things I like. I would also like to extend my deepest gratitude to Dr. Antonia Bruno, without her precious support it would not be possible to conduct this research. Her guidance helped me in the very last time of research and writing of this thesis, teaching me how to be persistent and constant as a researcher.

My sincere thanks also goes to Dr. Karoline Faust, who gave me important tips to develop some of the work I presented here. The COVID-19 interrupted us, but I hope to be able to start again soon.

I had great pleasure of working with ZooPlantLab colleagues and the FEM2-Ambiente group. In particular, I would like to thank Prof. Massimo Labra,

who taught me to confront any challenges in the future and be open minded to new perspectives, and Dr. Andrea Galimberti, for his constant support and advice, even in the most difficult moments.

Any of the work I discussed above would never have been accomplished without the expertise contributed by everyone involved. A special thanks to my co-authors Jessica Frigerio, Valerio Mezzasalma, Nicola Tommasi and Bachir Balech.

I cannot conclude without mentioning the first bioinfo group I was part of, even if only for a short time: Alberto Brusati, his invaluable help is unquantifiable; Adam Chahed and Elena Parladori, to whom I wish all the best.

I would like to thank my brand new colleagues: Sara, who has recently started her PhD journey; Chiara, Giulia and Davide. I wish everyone the best.

Thanks to all my colleagues and friends at TeCSBi and at the Department. Last year would not have been the same without you. Any moment we had together was wonderful and allowed me to make new and beautiful friends. A special thanks to Prof. Paola Branduardi, for her constant support and hard work, and Isabella Mauri. Together, they put their heart into this doctoral project every day, allowing us to find our place in the world. And I also thank the Pippo Band people: Erika, Alessandra and Carlo, that every morning readily prepared me for the day.

I am also grateful to Dott. Rosanna, her extraordinary support allowed me to face the different obstacles I encountered, and all of my friends, especially Giulia D., Giulia R., Irene, Karin and Carla. I'll never forget their precious support.

Of course, I would like to thank Simone, the person who endures my chaotic life with deadlines every day, and my big family, especially mom, dad, my brother Gabriele, Omar and Nives. Everyone supports me and bears with me everyday, encouraging me to never give up, even more in the face of difficulties. Without their help, I would not be the person I am now, proud to have started this path and happy and determined to pursue my dreams.

And finally, dear Lorenzo, I would like to dedicate this hard work to you. You will always be in my heart and I hope that a part of me will be able to do you justice and honor, for all the good things you have done. I love you and I will always remember you for your smile.
Thank you, Lorenzo. Thank you all.