Università degli Studi di Padova

Università degli Studi di Padova

Department of Information Engineering

Master of Science in Bioengineering

Microbiota analysis in Eosinophilic Esophagitis

A comparison of bioinformatic pipelines and case-control group design

Supervisor:
Prof. Enrico Lavezzo

Student:
Mirco Stradiotto
ID: 2062246

Advisor:
PhD candidate, Laura Manuto

Accademic Year 2023/2024

Abstract

Eosinophilic esophagitis (EoE) is a rare condition characterized by eosinophilic infiltration of the esophagus, leading to a chronic inflammatory process that results in swallowing difficulties. Diagnosing and treating EoE can be challenging due to the lack of a reliable non-invasive biomarker. Recent advancements in sequencing technologies have highlighted the role of the human microbiota in diseases exhibiting inflammatory patterns, suggesting its potential for providing new diagnostic and therapeutic insights.

Sequencing data preprocessing is a crucial step in microbiome studies, yet it often lacks standardization, which can introduce biases and hinder the comparability of studies' results. This study compared two processing pipelines: a custom-built pipeline that integrates various tools and an automated pipeline that utilizes *KneadData*, a wrapper tool that simplifies the process.

After demonstrating the excellent trade-off achieved with KneadData, this research focused on the development of *BioDonut*, a straightforward pipeline designed to analyze paired-end shotgun metagenomics data from human microbiota studies, specifically based on fecal and saliva samples. BioDonut covers a comprehensive workflow, from initial preprocessing phases (such as quality filtering and decontamination) to several first-line downstream analyses.

Finally, since BioDonut is intended for studies comparing microbiota composition between healthy and diseased individuals, a *propensity score matching* algorithm was developed to reduce biases introduced by confounders when designing case-control groups from volunteer participants.

BioDonut is publicly available at github.com/strmrc/BioDonut

Ai miei genitori
Claudia e Silvano

Contents

Chapter 1

Introduction

1.1 Human microbiota

The term *microbiota* refers to the unique combination of microorganisms (such as fungi, bacteria and viruses) that exists in a particular environment. Despite being used as its synonymous, the term *microbiome* is related to a slightly different concept, referring to the collection of genomes of these microorganisms including their whole "theatre of activity" (structural elements, metabolites and the environmental conditions) [1], as summarized in Figure 1.1.

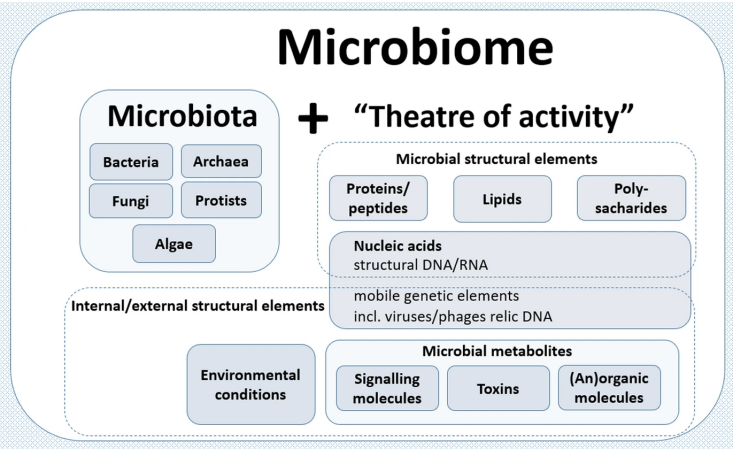


Figure 1.1: A schematic representation of microbiome definition. Adapted from [1].

A lot of different environments have a microbiota: both natural and artificial but also organisms themselves can be home to other organisms. They are defined *host environments*. A perfect example is the human body, one of the most complex microbial ecosystems on earth, capable to outnumber the host itself. Numbers help grasp the concept: considering cell count, the human body

is composed of approximately 30 trillion human cells versus 39 trillion of microbial cells [2]. When it comes to gene count, comparison is even more outstanding: 20.000 human genes versus an approximation of 2 to 20 million microbial genes [3], [4], something that can be referred to as "our second genome" [5]. The composition of human microbiome is unique in each individual and far more personal than the human genome. Indeed, if we consider two different non-related humans, their genome will be identical for the vast majority (99.5%). Instead, their microbiome could be totally unrelated, with almost no overlapping, reaching a degree of personalization useful even for forensic applications [4].

It is clear that microbiota dimension cannot be ignored. The picture being painted suggests that these living communities somehow, at some level, must be involved in the host biological processes. The nature of this impact has been object of controversy since Pasteur and Metchnikoff days, when the idea of a "normal flora" beneficial for life was totally opposed to a flora antagonists of the host, in competition for essential resources [6]. To date, we know that both points of view are true and that's all about balance. A perfectly balanced situation, where good microorganisms, which are present in greater percentage, live in mutual harmony with potentially bad ones is known as *eubiotic status*. On the opposite, large shifts in phyla ratios or the expansion of new bacterial groups lead to a disease-promoting imbalance, which is often referred to as *dysbiosis* [7]. An eubiotic status is not only desirable but needed in every respect since these commensal organisms deploy some basic functions that we did not evolve on our own [5] as a result of a long coexistence. Increasing evidence shows that microbiota contribution is affecting many, if not most, pathways that are crucial in most, if not all, biological processes that constitute human health and disease [8].

Even though the idea that organisms we carry must be important seems obvious nowadays, this was not the direction in which medical science headed in the past. Awareness about the importance of microbiota role is relatively young, and so are the studies that focuses on this complex topic. The advanced technologies that facilitate our investigations into the microbiome, like the improvements in throughput and accuracy of DNA sequencing platforms, have only emerged in the past decade [4], [5], [8]. Scientists are now starting to address issues related to microbiome variation, stability, and development, as well as the impacts of disturbances and the resulting interactions with host

physiology and pathophysiology. Additionally, epidemiological research is beginning to explore the consequences of microbiome alterations [8] . Therefore, characterizing the healthy human microbiota represents a fundamental preliminary step to allow for the identification of meaningful differences. This was one of the major goals of first phase of the Human Microbiome Project (HMP), a five-year program supported by the National Institutes of Health (NIH) Common Fund. The overall mission was to generate resources analyzing multiple body sites in a large cohort, aiming to facilitate characterization of human microbiota. Over 32 terabytes of data were produced, all publicly available through their portal [9].

Microorganisms are present throughout all human body, but each location shows its peculiarities, as stated by the HMP and various site-specific studies [5]. Each body site functions as a unique niche, defined by its distinct microbial communities, community dynamics, and interactions with host tissues [5]. To date, five major regions have been well-characterized: gut, oral, respiratory, skin, and vaginal microbiota [5], [10]. These regions are specialized at the point that is possible to state that human microbiota is more similar across individuals than across body sites [11], which is outstanding if we consider that, as mentioned before, the microbiome could be serve as a secondary fingerprint.

It is important to specify that, in the majority of research work available to date, the characterization is mostly focused on bacteria, leaving out all of the remaining biota [12]. This phenomenon can be explained by the assumption that bacteria component represents the greatest percentage of the total composition of the microbiota and thus having a major impact on the human biological processes. However, even the smallest fraction may play a key role [12]. Actually, this is more likely to be about technological limitations, since widespread and less expensive sequencing techniques (i.e., amplicon sequencing) are well-established for bacteria profiling rather than virus, fungi and yeast [13]. Thanks to technological advance, future work should focus on the complete and full analysis of microbiome, including eukaryotic, prokaryotic and viral interactions [12].

The gut houses bulk of human microbiota and is one of the most extensively studied sites [5], [10]. Although not as well-characterized as the previous one, the oral site is recognized as the second largest microbial community in humans [5], [10]. Their easily accessible location, along with their involvement in existing routine clinical tests due to the potential large amount of biomass obtainable

from samples, are two of the reasons why they are considered ideal candidate sites for research works and analyses.

1.1.1 Gut microbiota

Extensive studies revealed gut microbiota involvement in basic biological processes. Thanks to the versatile metabolic genes providing independent unique enzymes and biochemical pathways, gut microbiota is essential for energy and nutrient extractions from food, which is reduced into small and easily absorbable units [10], [12], [14]. It also accounts for identification of potentially hazardous elements that we could ingest, providing to their eventual neutralization [14]. Aside from digestion purposes, it is also essential for its cooperation with the host immune system, playing a fundamental role in protecting host from external pathogens by producing antimicrobial substances, helping in intestinal mucosa ad immune system development [10]. It is also involved in the production of vitamins, amino acids and other bioactive molecules [10], [12]. Finally, the role of the gut microbiota extends far beyond the intestine to the metabolism of systemic drugs and diseases manifestation in other organs systems [5].

It is unlikely that this significant ensemble of activities is carried out by a uniform niche of organisms. Actually, it is a heterogeneous collection of distinct habitats that take place along all the Gastrointestinal (GI) tract [14]. Even considering only the gut, which represents the last portion of the GI tract, it's still possible to describe multiple sub-niches. For instance, Proteobacteria like *Enterobacteriaceae* are present in the small intestine but absent from the colon. In contrast, Bacteroidetes families such as *Bacteroidaceae*, *Prevotellaceae*, and *Rikenellaceae* are commonly found in the colon [15]. Regardless many attempts trying to characterize healthy gut microbiota, it's still difficult to list a precise composition at any deep level of taxonomic resolution due to the unique distribution of each individual influenced by microbial growth rates, structural variants within microbial genes, environmental exposures and host genetics [16]. Nonetheless, variations are retained to be caused also by a distinctive combination of forces that have not purely deterministic but also stochastic nature [14]. Still, it is possible to state that gut microbiota is quite limited in diversity at higher level of taxonomic resolution showing six typically primary phyla: *Firmicutes*, *Bacteroidetes*, *Actinobacteria*, *Proteobacteria*, *Fusobacteria*, and *Verrucomicrobia*, with *Firmicutes* and *Bacteroidetes* being the predominant

types [10]. Instead, considering deeper phylogenetic levels (i.e., species and strains) it exhibits significant diversity, possibly reaching over 5,000 different bacterial taxa [5], [10]. Under healthy conditions, the gut microbiota demonstrates stability and resilience, engaging in a symbiotic relationship with the host, being characterized by high taxonomic diversity, substantial microbial gene richness, and a stable core [16]. Especially the assumption on the existence of a "stable core" seems to be inconsistent with the extreme variability and difficulty to define how a healthy gut should be. As a matter of fact, it can be said that human microbiome has a dichotomous essence: it's able to mutate exhibiting remarkable adaptability while maintaining exceptional durability over extended periods and in the face of various changes [4]. This peculiarity can be appreciated over periods of days or months, thanks to longitudinal studies where subjects provide multiple samples at different timepoints [17]. Due to its nature, future works should focus on the definition of healthy microbiome that takes account for other characteristics rather than the single taxonomic profile, such as functional potential, stability over time and perturbation resilience [14]. Each of us builds its own microbiota drawing consortium of microbes from the broader pool of microbes available on nearby hosts and environments [14]. The most significant shaping phase is considered to be during first three years of life [3]. We vertically inherit our mother microbial communities at birth, depending on delivery mode. Neonates that are vaginally delivered show an initial microbiota similar to vaginal flora (e.g., *Lactobacillus* and *Prevotella*), instead cesarean section tends to transmit skin-associated communities (e.g., *Propionibacterium*, *Staphylococcus*, and *Corynebacterium*) [3], [5], [10]. Significant differences also exist between the gut microbiota of formula-fed and breast-fed infants. Formula-fed infants typically have bacteria linked to increased antibiotic use, hospitalization, and prematurity [5]. Also weaning period, with the introduction of solid food, impacts the maturation of the microbiome [3]. Over the first year of life, the gut microbiota progressively develops to resemble that of an adult and these differences tend to diminish as children grow [10]. Gut microbiota composition continues to change according to the age, increasing in diversity between childhood and adulthood, and decreasing at older ages [18].

1.1.2 Oral microbiota

The human GI tract is a complex system that begins at the mouth, passes through the stomach and intestines, and concludes at the anus [12]. Therefore, even if it's considered a whole separate niche, microbial communities in oral cavity are somehow related to the ones in the gut, and they are involved in a lot of common processes (e.g., digestion). This is one of the reasons why, while most of the studies are focusing on the gut, also oral microbiota is gaining attention with the characterization of the *oral-gut axis* [10].

Unlike the gut environment, the oral cavity features both the hard surfaces of teeth and the epithelial surfaces of the mucosal membrane, being home to approximately 50 species and 1000 sub-species of microorganisms. This diversity is manly attributed to different anatomical and functional structures present. Consequently, we are able to identify multiple subniches: saliva, tongue, tooth surfaces, gums, buccal mucosa, palate, and subgingival/supragingival plaque [10]. These habitats can undergo significant and rapid shifts in both composition and activity due to factors such as pH changes, genetic mutations, and bacterial interactions [19]. Due to their constant exposure to saliva, these microbes have developed a strong adherence ability ensuring resistance to fluid forces [20]. In fact, both commensal and transient species resort to the creation of biofilms [12].

While it is known that minor differences are present, the overall microbiota composition across these seven sites is similar. Communities of a healthy mouth are predominantly streptococcal species, with common representation also from *Actinomyces*, *Veillonella*, *Fusobacterium*, *Porphromonas*, *Prevotella*, *Treponema*, *Nisseria*, *Haemophilus*, *Eubacteria*, *Lactobacterium*, *Capnocytophaga*, *Eikenella*, *Leptotrichia*, *Peptostreptococcus*, *Staphylococcus*, and *Propionibacterium* [5], [21]. Up-to-date information on taxa of the oral microbiome may be found in the Human Oral Microbiome Database (HOMD, homd.org) [12].

Note that, being first point of contact during ingestion, the oral cavity is frequently exposed to various pathogens, with the ongoing basal activation of the immune system serving as a key selective force in shaping the microbial community [14]. Some symbiotic inhabitants of the oral microbiome (e.g., *Streptococcus*, *Veillonella* as mentioned above) have been found to play a role in defensive processes by enhancing the production of immune effectors such as antimicrobial peptides (AMPs) and pro-inflammatory cytokines, as well as strengthening epithelial barrier function and increasing mucosal thickness [14].

1.1.3 Additional factors affecting the microbiota

Apart from the specific body region, host genetics, birth and after-birth procedures and other elements already discussed, there is a lot of other factors being able to affect the microbiota's composition and characteristics. Diet, lifestyle and use of antibiotics are everyday life key variables.

Diet

Evidence shows diet plays a significant role in shaping the gut microbiome, both in the shortand long-term periods [3], [4]. This relationship is quite straightforward given that the microorganisms rely on host nutrient supplies to live. Studies comparing different communities or individuals with distinct dietary patterns provide evidence of diet's long-term impact on the microbiome, reflecting habits that span many years, if not entire lifetimes [22]. In the short term, substantial dietary changes can also lead to notable shifts in microbiome composition (e.g., reducing fiber intake, eliminating gluten, significantly increasing protein consumption) [22]. It is a matter of fact that nutritional food content influences which microbial species will thrive or die [3]. While all mammalian gut microbiomes share a fundamental set of genes responsible for key metabolic functions, the distribution of these genes and the particular taxa that possess them vary significantly among carnivores, omnivores, and herbivores [3]. Another study confirmed the extent of diet impact, especially on the gut microbiome, highlighting how even a short-term consumption of an entirely animal-based or plant-based diet can significantly alter the structure of the microbial community, emphasizing the usual differences in microbial gene expression between individuals [23]. Other findings support microbial communities being more similar in subjects with common diets, including the influences of ethnicity and geography [24]. Indeed, diet indirectly conveys differences due to culture identity and physical location. For instance, differences in Western and non-Western populations are significant and are likely influenced, at least in part, by diet [3]. Emerging evidence also highlights the effects of specific dietary items. For instance, dietary emulsifiers and artificial sweeteners have demonstrated comparable impacts in both human and animal studies [3]. A final note must be done in regard of probiotics and prebiotics, which tend to be popular discussion topics in microbiota research since they are often used as dietary supplement for clinical intervention aiming to microbiota modulation by oral administration, but evidence on this seems to

have not reached a consensus yet. Despite significant interest in the potential of probiotics to alter the microbiome, their effects are generally modest and may primarily influence gene expression rather than leading to substantial changes in the microbial community [3]. Discovering probiotics that can produce significant, lasting impacts on the microbiome remains a key objective for future research [3].

Lifestyle

Exercise, infections, stress, sleep patterns, living with pets, having housemates are all examples of lifestyle traits that have the ability to influence the microbiome (even if the effect sizes are typically small) [3], [4]. Exercise seems to affect microbiome structure by reducing inflammation [25]. Sleep deprivation is linked to alterations in the gut microbiome, such as an increased ratio of *Firmicutes* to *Bacteroidetes* and higher levels of *Coriobacteriaceae* and *Erysipelotrichaceae* [26]. Stress increases intestinal permeability concomitant with changes in *Bacteroidetes* and *Actinobacteria* and inflammatory markers [27]. Environmental exposure changes when individuals must physically move to different locations or building, for example due to workplace. This study, for example, proved how skin-associated bacterial community structure and composition could predict whether a sample came from an urban or a rural resident [28]. Relationships are also relevant: sexual activity between heterosexual partners tends to increase the similarity between penile and vaginal microbiota [29]. Also, couples who engage in physical interactions exhibit more similar microbiota than those who share a living space without physical contact, emphasizing the role of physical interaction in microbial sharing and microbiome similarity [30].

Antibiotics

The germ theory was unquestionably proven by Louis Pasteur and Robert Koch in the mid 19th century and, since then, hygiene habits, infection control and healthcare in general have come a long way. Antibiotics development, started with the discovery of penicillin in 1928, changed forever our society. It's not difficult to find them referenced as "wonder drugs" in a lot of official documents of 1950s, conveying the excitement of patients, healthcare professionals, and policymakers for medications that turned previously life-threatening bacterial infections into treatable conditions [3], [31]. However, this unrestrained enthusiasm led to their

abuse. Indeed, antibiotics and antimicrobials have a non-selective effect, causing two main issues: first, the killing of both pathogens and commensal bacteria, which causes ecological disorders [10]; and second, as a consequence, the selection for those mutants and strains with the capacity to survive large doses of antibiotic drugs [5], resulting in what is known as *expansion of resistance* [32]. Expansion of resistance is a well-known problem to specialists and authorities, that causes a reduction in the effectiveness of drugs, leading to a paradox: "the miracle drugs are destroying the miracle" [33]. The rise of multidrug-resistant bacterial strains has highlighted the need to identify and monitor reservoirs of antibiotic resistance genes that could potentially be transferred to clinically significant pathogens [5].

On the other hand, ecological disorders interfere with already mentioned activities of commensal organisms. In fact, antibiotics are retained to be "one of the most dramatic ways to influence the microbiome" [3]. Even brief courses of antibiotics prescribed for acute infections can disrupt gut microbial composition for extended periods (even months and years) [34], [35]. While the extent and nature of dysbiosis due to antibiotic treatment differ among individuals [3], [4], some common patterns emerge. For instance, bacterial diversity typically decreases in the week following antibiotic exposure and then starts to recover, though the original microbial state often isn't fully restored (not only in terms of microbial diversity, but also in terms of microbial gene expression, protein activity, and overall metabolic function) [3]. Interestingly, some studies have highlighted how short time frames (e.g., one week) are often not able to fully capture these state changes [4], emphasizing once again the need for larger scale longitudinal studies of diverse cohorts.

The overuse of antibiotics is becoming a major public health problem especially in children: as already mentioned, the first years of life are crucial for microbiome maturation and antibiotic treatments during this period could potentially lead to serious long-term effects (increasing evidence are associating early antibiotic treatment with obesity, inflammatory bowel disease and other disorders already linked to dysbiosis) [3], [4]. Also, it must be considered the cumulative effects of antibiotic treatments both in children and adults, with their effect becoming more pronounced with additional courses [3], [4].

1.1.4 Microbiota and diseases

External changes can disrupt the balance of the microbiota community, potentially leading to dysregulation of bodily functions and the development of

diseases [10] as summarized in Figure 1.2. Increasing evidence supports the association between microbiota and a range of well-known conditions, like dental caries and bacterial vaginosis, as well as other chronic and more complex conditions including cardiovascular diseases, cancer, respiratory illnesses, diabetes, chronic kidney diseases, and liver diseases [3], [10]. Additionally, the microbiome may be connected to conditions in which their involvement may not seem so straightforward, such as Parkinson's disease, autism, and depression [3].

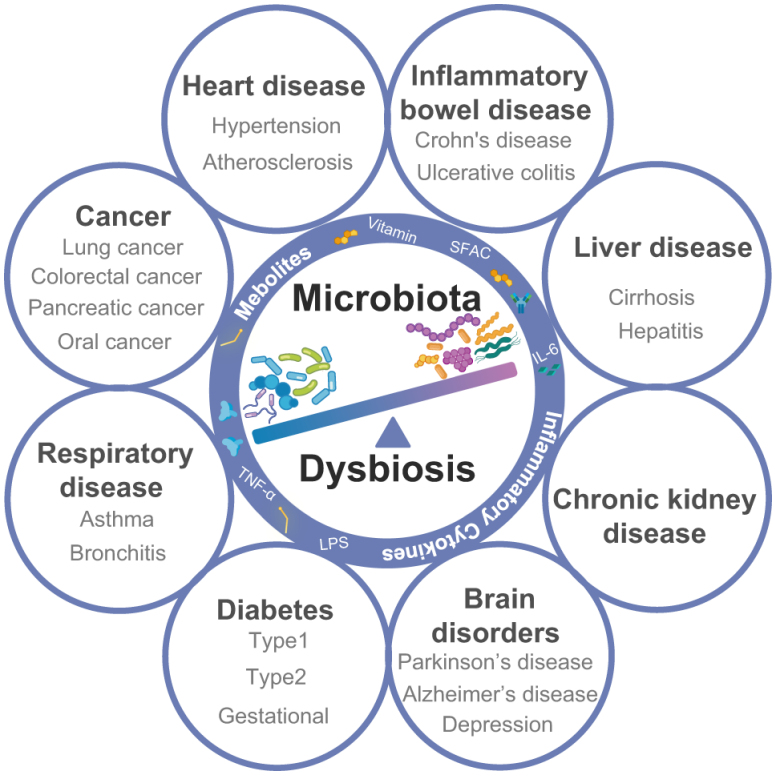


Figure 1.2: Human microbiota dysbiosis contributes to various diseases. Adapted from [10].

Of particular interest due to objective of this research work are progress made in research field of microbiome dysbiosis linked to chronic disorders, which are characterized by inflammation processes. These diseases are one of the most challenging and persistent diseases affecting modern population [8]. Beyond Eosinophilic Esophagitis (EoE), which will be discussed later in details, another prominent example is Inflammatory Bowel Disease (IBD). This is one of the most thoroughly researched human conditions linked to the gut microbiota, which has been shown to differ significantly between healthy individuals and those with IBD [10], both in species richness (the number of bacterial species present) and species abundance (the population size of each species) [36]. As it will be detailed later, in microbiota research, bacteria are identified through

sequencing rather than traditional culturing techniques, leading to the classification of bacterial species or genera as operational taxonomic units. Research has shown that individuals with IBD typically exhibit decreased bacterial diversity, with lower levels of *Firmicutes* and *Bacteroidetes* and higher levels of *Proteobacteria* compared to healthy individuals [3].

Apart from IBD, of interest are all diseases related to inflammatory processes gathered interest due to their strong relationship with the gut microbiome status and its interaction with the host's immune system, affecting both innate and adaptive immune functions [3]. The innate immune response involves cells such as dendritic cells, neutrophils, and natural killer cells, whereas the adaptive immune response includes the activation of T and B cells [3]. Certain microbiomes are linked to the development of specific T cell subtypes, and changes in the gut microbiome can lead to both beneficial and harmful outcomes through the regulation of CD4+ T cell subtypes [3]. Nonetheless, in complex diseases like IBD and Parkinson's disease, the microbiome plays a role in disease etiology but is not the sole causative factor [3]. Genetic predisposition, epigenetic regulation, and environmental factors all contribute to this intricate interactome [3].

Another category worth mentioning is atopic diseases. In general, atopy is defined as the predisposition of an individual to produce an exaggerated immune response (IgE-mediated) to common environmental allergens [37]. Atopic dermatitis, allergic rhinitis, and asthma impact around 20% of people's lives worldwide and have increased in the last decades [38]. Such trend has often been associated with the hygiene hypothesis [36], which suggests that a reduced exposure to microbial antigens in early childhood (i.e., due, for example, to highly sanitized environments) may alter the composition of the infant gut microbiota, thereby impairing immune system development and promoting the onset of allergic diseases [3]. This hypothesis is further supported by epidemiological studies, which reported higher rates of atopic diseases among infants born via cesarean section, those who were formula-fed, and those who were exposed to antibiotics [3]. Interestingly, for some diseases, the microbiota has been reported to have a greater impact on the disease phenotype respect to the host genetic factors [4]. However, the exact relationship between microbiome and disease development remains unclear in most, if not all, cases [36]. The primary challenge is determining whether the dysbiosis is a contributing factor to the disease or a mere consequence of it [36]. In other

words, the problem is being able to distinguish between simple correlation and true causation. Nevertheless, even if the microbiome was not the direct cause of a condition, it could still serve as a valuable tool to diagnose and possibly stratify complex or not fully characterized diseases, such as Crohn's disease [3], potentially becoming a source for novel biomarkers [4].

1.2 Eosinophilic Esophagitis

1.2.1 Definition

Eosinophilic Esophagitis (EoE) is a chronic inflammatory disease of the esophagus, presenting esophageal dysfunction symptoms. It is mainly characterized by infiltration of eosinophiles in the esophageal mucosa, although several different other type 2 inflammation mediators are involved in the pathogenesis [39].

EoE was once considered a rare disease, but is now increasingly common, being "one of the most common conditions diagnosed during the assessment of feeding problem in children and during the evaluation of dysphagia and food impaction in adults" [40]. The earliest reports of EoE emerged in the 1970s [41], but first formal characterization does not come up until 1990s, thanks to Attwood and Straumann's work [42]. Initially, esophageal eosinophilia was thought to be exclusively related to gastroesophageal reflux disease (GERD). However, its recognition in both adults and children with symptoms that did not improve with acid suppression or anti-reflux surgery was clearly pointing out a distinct disorder [40]. To date, we know EoE and GERD represent two distinct clinical entities that may coexist in the same patient and interact [39]. GERD and EoE are respectively the first and second-most prevalent cause of chronic esophagitis [43]. Nevertheless, nearly five decades since its first characterization, many aspects of EoE are still unclear [41].

1.2.2 Epidemiology

To date, it is possible to outline an overall increasing trend: recent reports states that previous estimates of 5 to 10 new cases per 100.000 inhabitants annually are moving up to 20 considering some countries, with the highest prevalence being reported for Europe and North America, with one case each 1.000 people [44]. However, even if it can be assumed that the frequency of EoE is higher in Western countries than in the East, predominantly affecting Caucasians [41], EoE incidence is increasing also in Asia [42]. As a result there has been a corresponding growth in emergency room visits, with nearly half of these patients requiring an endoscopy and the 40% needing the removal of the impacted food [42]. Therefore, it is crucial for healthcare institutions and insurance providers to assess disease's burden, including both direct and indirect costs, to inform economic planning and

develop effective management strategies [42]. Besides the increase in the diseases rates, it must be noted that this trend can be related also to an increased general awareness of EoE's clinical features and to improved diagnostic guidelines [39], [40]. Evidence of an increased awareness can be observed, for example, in the reduction of diagnostic delays, which were approximately 7 years before 2007 and dropped to 0.7 years after 2018 [45].

Considering demographical factors, EoE has been described in all age groups, involving both children and adults [39], [40], [41], [42]. The disease incidence increases with age and peaks in early adulthood [39]. When it comes to sex, it occurs three times more often in men than woman [39], [40], [41] even if there is no difference concerning disease severity [41].

While numerous studies in the literature discuss the rising prevalence and incidence of EoE, there appears to be a moderate level of evidence addressing mortality related to the condition [41]. According to the available studies, no links were identified between EoE and increased mortality from cancer or cardiovascular diseases [40], [41], which makes EoE a nonpremalignant condition [39]. These findings may offer some reassurance, but additional largescale population studies focusing on particular conditions (e.g., Barrett's esophagus) are needed to confirm these conclusions [39].

An interesting issue is the seasonality of morbidity and symptom severity, which suggests a possible role of the climate. Nearly half of new diagnoses in a U.S. study on children occurred in spring, with symptom exacerbation during fall and summer [41]. Other studies suggest a correlation with dry climates, which typically result in a prolonged pollen season [41].

1.2.3 Etiology

Etiology refers to the cause or origin of a disease, identifying factors or agents (such as bacteria, viruses, genetic mutations, environmental exposures, etc.) that lead to the onset of a disease. To date, EoE etiology is still unclear with different hypothesis proposed about its triggering factors. What is certain is that, although eosinophils are present in various tissues, they are normally absent in the esophagus [41]. When stimulated by factors that are not yet fully understood (e.g., food allergens), epithelial and dendritic cells release cytokines such as IL-25, IL-33, and thymic stromal lymphopoietin (TSLP) [41]. These cytokines activate immune cells, predominantly initiating a T helper 2 (Th-2) immune response. This response involves the secretion of Th-2 type cytokines,

including IL-4, IL-5, IL-13, eotaxin-3, and periostin [41]. IL-5 plays a key role by stimulating eosinophil proliferation and migration from the bone marrow to the bloodstream and subsequently into all layers of the esophagus [41].

One of the most studied triggers are food antigens, which seem to have a key role in stimulating the immune response. A lot of studies proved that, when specific food is eliminated, a significant percentage of patients is able to reach remission and, when the same food is reinserted, a relapse is observed [41]. The predominant mechanism of food allergy in EoE appears to be a non-IgE-mediated process, even if it has been linked to atopy [41]. Therefore, routine IgE skin testing does not consistently identify food antigen triggers [40]. A high level of evidence considers EoE as the last step of progression of the atopic march, which is the description of the temporal trajectory of development of type 2 inflammation (from atopic dermatitis, IgE-mediated food allergy, allergic asthma to EoE) [39], [42] explaining why such diseases are often overlapping in symptoms and manifestation. Indeed, a personal or family history of atopic disorders, (i.e., asthma, eczema, rhinitis, and anaphylactic food allergy) is common in EoE patients [40].

As already anticipated, an emerging theory referred to as "hygiene hypothesis" supports the link between EoE and immune system abnormalities due to an "excessive asepticism" of the environment. As previously mentioned, asepticism condition can interfere with the correct maturation of the microbiome (especially in children), which is necessary for the correct development of the immune system [41]. Most research has focused on the gut microbiota, leaving the role of the esophageal microbiome relatively unexplored [41]. However, a study using a mouse model of EoE found that therapeutic supplementation with the probiotic *Lactococcus lactis NCC 2287* significantly reduced eosinophil counts in esophageal tissue. In contrast, another probiotic, *Bifidobacterium lactis NCC 2818*, did not show a notable effect on esophageal eosinophilia [41]. While the human microbiome is increasingly studied in relation to inflammatory and autoimmune diseases, as well as the gut-brain axis, its role in the development of EoE remains unclear [41] and should be the object of further evaluations.

1.2.4 Risk factors

At this point, it is clear why EoE can be considered a multifactorial disease (i.e., no single specific etiological factor) [41] and, as mentioned, it shares several

risk factors with other atopic disorders [42]. Therefore, we are able to draw risk factors associated to EoE and to divide them into three different categories: *genetic*, *environmental* and *biological factors*.

Genetic factors

Genome-Wide Association Studies (GWAS) look at the whole genomes to identify singlenucleotide polymorphisms (SNPs) that might be linked to specific conditions or diseases, suggesting possible risk loci. In Table 1.1 are summarized the 42 published risk loci associated to EoE to 2020 [46].

Table 1.1: Reported EoE risk loci. Adapted from [46].

EoE risk locus	**Tag genetic variant**	**PMID**	**Genes at and near risk variants**	**Risk allele frequency**	**P value**	**Odds ratio**
1p13.3	rs2000260	25017104	SLC25A24	0.57	7 × 10-7	1.32
1p36.13	rs28530674	25017104	KIF17	0.04	3 × 10-7	1.83
	rs2296225	25017104		0.08	1 × 10-7	1.63
1p32.2	rs11206830	25017104	AC119674.2	0.02	8 × 10-8	2.16
2p23.1	rs149864795	25407941	CAPN14	0.052	5 × 10-10	2.22
	rs77569859	25017104		0.05	3 × 10-10	1.98
3q26.32	rs6799767	20208534		0.58	4 × 10-7	1.49
4q21.1	rs13106227	20208534	SHROOM3	0.62	4 × 10-6	1.52
	rs1986734	20208534		0.49	1 × 10-6	1.54
5q22.1	rs3806932	20208534	WDR36, TSLP	0.54	3 × 10-9	1.85
	rs3806933	25017104		0.56	2 × 10-8	1.37
	rs252716	25407941		0.447	4 × 10-14	1.52
5q23.1	rs2055376	25017104	SEMA6A	0.02	7 × 10-8	2.3
5q14.2	rs1032757	20208534		0.07	2 × 10-6	1.96
6p11.2	rs9500256	20208534	AL445250.1	0.58	5 × 10-6	2.04
8p23.1	rs2898261	25017104	XKR6	0.58	5 × 10-8	1.35

Continued on next page

Table 1.1: Reported EoE risk loci. Adapted from [46]. (Continued)

8q24.12	rs11989782	20208534	SNTB1	0.23	7 × 10-6	1.53
8q22.2	rs13278732	20208534	ERICH5	0.27	6 × 10-6	1.31
10p12.31	rs11819199	25017104	MIR4675	0.06	3 × 10-7	1.62
10q23.1	rs2224865	20208534	MARK2P15-LINC02650	0.31	9 × 10-6	1.44
11q13.5	rs61894547	25407941	LRRC32, EMSY, CAPN5	0.043	4 × 10-11	2.44
	rs2155219	25017104		0.51	4 × 10-7	1.37
	rs77301713	25017104		0.02	1 × 10-7	2.22
11q14.2	rs118086209	25017104	CCDC81	0.02	2 × 10-7	2.19
11q21	rs1939875	20208534	NR	0.26	3 × 10-6	1.54
12q13.3	rs167769	20208534	STAT6	0.37	2 × 10-6	1.36
	rs167769	25407941	STAT6	0.377	2 × 10-7	1.35
14q12	rs8008716	25407941	NOVA1	0.087	7 × 10-8	1.71
15q13.3	rs8041227	25017104	LOC283710, KLF13	0.72	6 × 10-10	1.52
16p13	rs12924112	29904099	CLEC16A	0.301	2 × 10-9	0.76
16q24.1	rs371915	20208534	MEAK7	0.87	2 × 10-8	1.9
17q24.3	rs6501384	20208534	CALM2P1-AC011990.1	0.33	6 × 10-6	1.41
17q25.3	rs3744790	20208534	TIMP2, CEP295NL	0.8	8 × 10-7	1.54
18q12.1	rs7236477	20208534	DSG1, DCC	0.03	7 × 10-6	2.22
	rs9956738	25407941		0.01	4 × 10-7	2.47
19q13.11	rs3815700	20208534	ANKRD27	0.14	2 × 10-9	1.62
21q22.3	rs17004598	25017104	HSF2BP	0.01	1 × 10-7	2.57
22q11.21	rs2075277	25017104	P2RX6	0.09	9 × 10-7	1.54

The majority of genetic variants associated with EoE are found either between genes (intergenic, 36.7%) or within gene introns (intronic, 42.4%) [46]. Only a

small portion of these variants alter the amino acid sequence of genes (coding, 2.2%), with just 3 out of 31 risk loci containing such a variant [46]. Therefore, the majority of EoE-associated risk loci are located outside the coding regions of genes, highlighting the importance of genotype-dependent gene regulation in EoE patients, which aligns with patterns seen in other complex diseases [46]. In particular, genetic variants at four loci have been consistently found at genome-wide significance among three reported GWAS. Of particular interest are 5q22 [TSLP/WDR36] and 2p23 [CAPN14] [46]. The thymic stromal lymphopoietin (TSLP) is a cytokine released by epithelial and dendritic cells, that promotes the Th2-mediated immune response [41]. It seems that the gene coding for its receptor is located in a chromosomic region where genetic changes are associated with a higher risk of EoE, particularly in men [47]. CAPN14 encodes a protein (calpain 14) that is part of the calpain family of calcium-dependent, non-lysosomal cysteine proteases [48]. It is overexpressed in the esophagus of EoE patients [41] and is thought to influence epithelial barrier function, with the induction of disruptive changes of esophageal epithelium [39].

A family history of eosinophilic esophagitis is commonly observed, and the estimated heritability risk for the condition is around 2% [42]. However, it is worth noting that study focused on twin heritability suggest that the tendency of EoE to recur frequently within the same family is primarily attributed to the shared family environment, rather than to genetic factors [49], suggesting a stronger impact of the environment rather than genetics. Therefore, the risk of disease in families with a genetic predisposition may be amplified by factors encountered early in life. In particular, the colonization of the gut and esophagus by commensal microbiota that influence the immune system could play a crucial role as an environmental risk factor [49].

Biological factors

Even if the mechanisms underlying the association of EoE and *Helicobacter pylori* infection has yet to be established, some studies suggest that this bacterium could protect from EoE development since a significant association between *H. pylori* exposure and reduced odds of EoE in Western countries was shown [50]. Other studies have focused on the relationship between EoE and HIV infection. A recent study found that individuals with HIV are "twice as likely to have EoE compared to those without HIV" [51]. This relationship is clinically relevant also for treatment choices since that oral and esophageal candidiasis, frequent in

HIV patients, can be exacerbated by the use of topical steroids for EoE [41]. A diagnosis of celiac disease is also retained to increases the risk of EoE, even if there is no evidence that EoE is associated with the human leukocyte antigen (HLA) genetic locus [41].

Environmental factors

Different factors, the majority of which related to early life days, have been associated with EoE like cesarean section, premature delivery, antibiotic exposure, non-exclusive breastfeeding and living in an area of lower population density [40], [41]. It is hypothesized that early-life exposures may produce an epigenetic imprint that heightens the risk of developing eosinophilic esophagitis [40].

All these factors are known to impact the correct development and maturation of human microbiota, supporting the validity of the hygiene hypothesis and clearly pointing out to the existence of a strict relationship between the microbiome, its status and EoE, which is the general underlying topic of this research work. Indeed, previous studies about other atopic diseases (such as asthma and atopic dermatitis) have already suggested that a lack of early exposure to microbes and (as a result) an altered microbiome may play a role in their mechanisms [40].

Finally, it is worth of note that to date there's no evidence linking EoE with stimulants like alcohol or cigarettes, as research on this topic is limited and only a few studies have been performed [41].

1.2.5 Symptoms

EoE shows different symptoms depending on the age. Ideally, we can divide patients in three major age categories: children ($< 11y$), adolescents ($11 - 18y$) and adults ($> 18y$) (precise age range may vary across different studies). Children have a wide variety of nonspecific heterogenous symptoms beyond dysphagia and food impaction, like nausea, vomiting, dyspepsia, acid regurgitation, abdominal pain, heartburn and failure to thrive [40], [41]. Children have higher probability to suffer also from atopic diseases [41] (e.g., asthma, atopic dermatitis). In adolescents and adults, symptoms start to act more similar, and they tend to diminish in number, with the majority of patients reporting dysphagia, food impaction, general discomfort in swallowing and heartburn [40], [41], [42]. It must be noted that these symptoms may disappear between periods of exacerbation [52]. Moreover, as already

mentioned, GERD act as a confounder when it comes to EoE symptoms since they have overlapping characteristics. Also, it has been shown how GERD may predispose patient to EoE by impacting the integrity of esophageal mucosa [53].

Patients may present not only general digestive system discomfort, but also some respiratory fatigue manifesting cough, croup, hoarseness and throat clearing [41]. It is still not clear if these symptoms are associated to or simply coexisting with EoE [54]. Finally, probably the most dangerous symptom that requires immediate endoscopic intervention, is esophageal food impaction (EFI) which may lead to esophageal perforation [41].

Furthermore, EoE patients typically show abnormalities visible through esophagus endoscopy. Alterations prevalence varies with age and diseases duration and progression [42] and in up to one fifth of patients (especially children) they could be absent. Possible endoscopic abnormalities are presence of exudates (i.e., whitish plaques in the esophagus, often misinterpret as candidiasis), edema, mucosal fragility, furrows, plaques, strictures, rings of different thickness and still present even with air being blown into the esophagus (i.e., their eventual collapse suggests the presence of other gastrointestinal diseases) [41], [42]. To effectively categorize endoscopic findings, a classification system known as *The EoE Endoscopic Reference Score (EREFS)* was developed. EREFS's key advantage lies in its clear and straightforward scoring of specific changes. While research indicates that this system is effective for quantifying changes observed during endoscopy (demonstrating a high prediction accuracy in both adults and children) [42], it doesn't always align closely with the clinical and histological aspects of the disease, which may limit its utility in assessing therapeutic effectiveness [41].

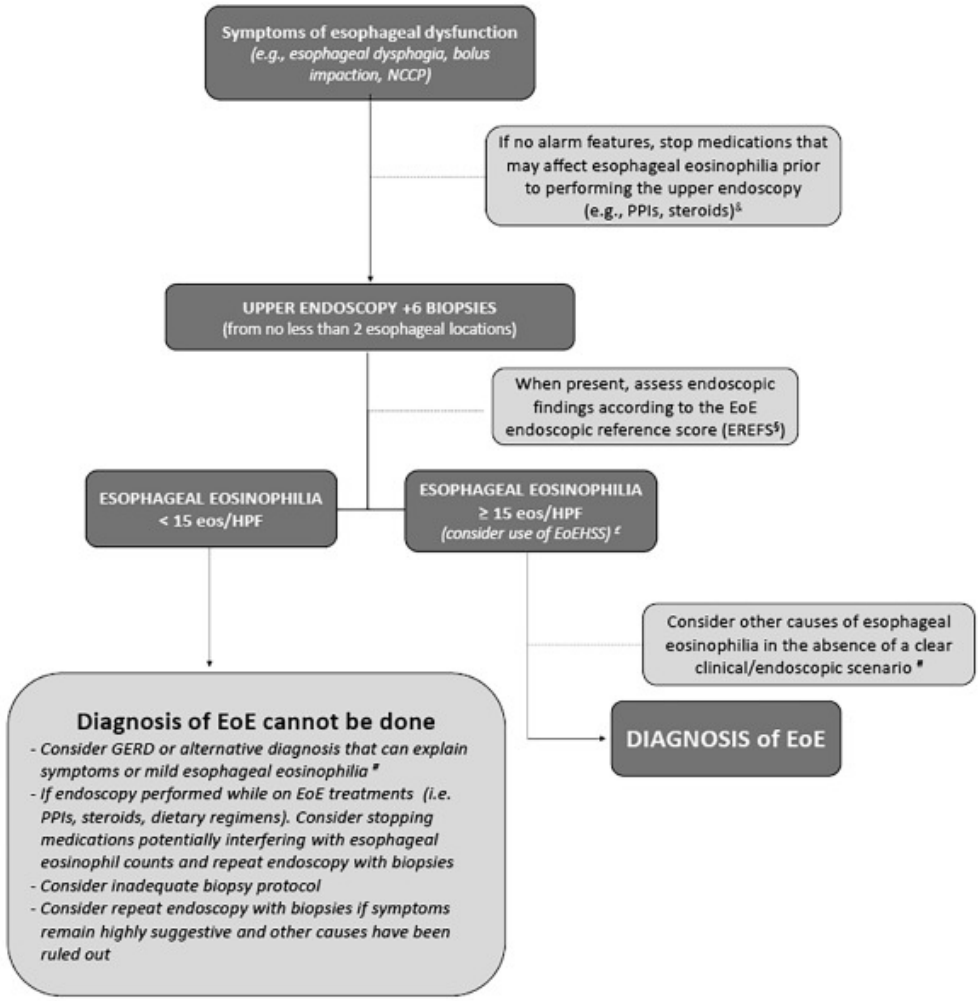
It is worth noting that in both children and adults, symptoms can often be underestimated or even totally hidden because they are unintentionally masked by changes in behavior. Eating slowly, chewing carefully, cutting food into small pieces, lubricating food with sauce, dietary changes preferring liquid meals to solid, are all examples of habits that could lead to an important diagnostic delay [40], [41], [42]. The question is whether adults with eosinophilic esophagitis have a disease that went undetected for years due to very mild "silent" inflammation in childhood, or if their condition represents a genuinely late onset, involving a distinct pathogenesis or phenotype [40]. A natural history study revealed that 85% of adults who had untreated symptoms for 20 years eventually developed esophageal strictures, supporting the idea that undiagnosed subclinical disease

or silent inflammation from childhood precedes the adult presentation [40]. In any case, it is certainly clear that an early diagnosis can help prevent disease-related complications. As already discussed, evidence suggests a general increased EoE awareness and a consequently reduced diagnosis delay [42] which is a sign that the right direction has been taken. However, EoE can still be diagnosed late, highlighting the importance of finding a reliable noninvasive EoE-specific diagnostic biomarker since guidelines being used today still relies on invasive procedures (i.e., endoscopy, histology) [55].

1.2.6 Diagnosis

Unfortunately, a reliable noninvasive biomarker for diagnosis EoE still doesn't exist. Indeed, the only presence of endoscopic abnormalities, even if it is considered a strong indicator, is not enough to make a diagnosis official [56]. To date, a formal EoE diagnosis follows a specific algorithm presented in Figure 1.3, which requires meeting three criteria: 1) presence of esophageal dysfunction symptoms, 2) eosinophils number in the esophageal tract above the reference cutoff, 3) exclusion of other diseases causing esophageal eosinophilia (i.e., GERD, achalasia) [41].

The second one is the most important condition to be satisfied. Infiltration is typically assessed by counting the peak number of eosinophils per high-power field (HPF) [42]. For the sake of clarity, an HPF is an area of a sample that is visible under a microscope when using a high-power objective lens, typically with a magnification of 400x [57]. A diagnostic cutoff of 15 eosinophils per HPF (around 60 eos/mm2) in at least one high-power field on esophageal biopsy [39] is highly accurate in distinguishing EoE from GERD and is also used to define histological remission, which is indicated by fewer than 15 eosinophils per HPF [42]. Due to the uneven distribution of eosinophil infiltration in the epithelium in EoE, multiple tissue samples should be taken, typically six biopsies from different sections of the esophagus, with a focus on areas showing mucosal alterations on endoscopy [39], [42]. If no endoscopic signs of EoE are visible, but clinical suspicion remains high, biopsies should be randomly obtained from the upper, middle, and lower segments of the esophagus [42]. Additionally, to rule out involvement of the stomach and duodenum, biopsies from these areas should be taken, particularly in children [42]. Also, drug assumption has to be assessed: proton pump inhibitors should be withdrawn at least 3-4 weeks prior to biopsy collection to achieve an accurate diagnosis [39]. Unfortunately, the percentage of



NCCP: non-cardiac chest pain; PPIs: proton pump inhibitors; EREFS: EoE Endoscopic Reference Score; EoE: eosinophilic esophagitis; EoEHSS: EoE histologic scoring system; GERD: gastroesophageal reflux disease; HPF: high-power field

δ Endoscopy should not be delayed when dysphagia, anemia, and/or weight loss are present. In case of non-diagnostic esophageal biopsies, repeat endoscopy off medications affecting esophageal eosinophil counts should be considered.
* Endoscopic features are not required for the diagnosis but can help in identifying EoE patients.
§ The EREFS is useful to standardize description and severity assessment of endoscopic findings.
£ The EoEHSS is useful to recognize doubtful cases and assess severity and extent of EoE.
Consider other causes of secondary esophageal eosinophilia: GERD, esophageal candidiasis, esophageal motor disorders, systemic disorders, autoimmune disorders, Crohn's disease.

Figure 1.3: Diagnostic algorithm of eosinophilic esophagitis. Adapted from [39].

European gastroenterologists following these recommendations is low [41].

Other diagnostic methods include esophageal manometry, functional luminal imaging probe (FLIP), and barium esophagrams obtained via barium X-ray [41]. Of particular interest is the increasing focus on developing non-invasive markers. For instance, ongoing research is exploring the detection of miRNA-4668 in salivary and esophageal secretions, with studies showing that patients with EoE exhibit elevated levels of miRNA-4668 in their saliva [41].

1.2.7 Treatment

When referring to EoE treatment strategy, the term 3 "D"'s therapy is often used because *Diet*, *Drugs* and *Dilations* are its cornerstones [58]. All strategies, except for dilation which is more an emergency procedure, aim to reduce inflammation trying to reach the remission state (i.e., eosinophils count < 15 per 0.3 mm2) [41]. Diet and drugs, which include the use of proton pump inhibitors, corticosteroids, immunosuppressants, and dietary interventions are the most effective in the early stages of treatment [58]. Histological evaluations should also be conducted after every therapeutic intervention, usually within 6 to 12 weeks, to assess treatment effectiveness, as clinical symptoms alone are not reliable indicators of disease activity [42].

Diet

Identification of food allergens as one of the primary antigenic triggers of EoE leads to the exploration of eliminating six major food groups (milk, wheat, egg, soy, nuts, and seafood) [41]. A less restrictive alternative is the four-food elimination diet (4-FED), which excludes cow's milk, wheat, egg, and soy. The two-food elimination diet (2-FED), which removes only cow's milk and wheat, is the least restrictive and most patient-friendly option [41]. Research indicates that the highest remission rates (50–71%) are achieved with the six-food elimination diet (6-FED) [41]. Remission rates diminish with four-food and two food elimination to 46–54% and 43–44% of cases, respectively [41]. Therefore, a step-up empiric approach to food elimination is recommended, starting with the 2-FED. If remission is not observed within 8–12 weeks, the 4-FED is introduced, and if that also fails, the 6-FED is implemented [41].

Drugs

Proton pump inhibitors (PPI) were the main drug class being used in drug therapy. They reduce gastric acid secretion and expression of eotaxin-3 (i.e., a Th-2 cytokine involved in inflammation). Unfortunately, they could contribute to inhibition of protein digestion and development of IgE antibodies in response to those proteins [41].

In the past two years, significant therapeutic advancements have been made. The approval of swallowed topical corticosteroids (STCs), such as *Jorveza*®(BOT, i.e., orodispersible budesonide tablets), has been a

cornerstone, becoming the first drug for EoE treatment in adult patients in Europe, as it can be read in the European Medicine Agency (EMA) site [59].

Jorveza was shown to be effective in two key studies involving a total of 292 adults with eosinophilic esophagitis. The first study, which included 88 patients with active EoE, compared the effects of 1 mg of Jorveza twice daily with a placebo. The primary outcomes measured were eosinophil levels in the esophagus and symptoms improvement. After 6 weeks, approximately 58% of patients treated with Jorveza experienced reduced eosinophil levels and had no or minimal symptoms, while none of the patients receiving the placebo showed these effects. The second study involved 204 patients whose eosinophilic esophagitis symptoms were already under control. This study compared the effectiveness of 0.5 mg or 1 mg of Jorveza taken twice daily against a placebo over 48 weeks. Results showed that around 74% of patients taking 0.5 mg and 75% of those taking 1 mg of Jorveza twice daily maintained satisfactory symptoms control, compared to only 4% of patients receiving the placebo [59]. There are also other STCs available being used off-label for EoE treatment in the form of spray nebulized liquid (fluticasone), powder (fluticasone) or viscous preparation of liquid budesonide [41]. Regardless of their formulation, STCs modulate the immune system by suppressing inflammatory response stimulators [41]. Unfortunately, even if STCs appears to have a favorable safety profile with no serious effects in the short term [60], the long-term effects of swallowed topical corticosteroids remain unclear, as do the optimal dosage and follow-up strategies [41]. To date, evidence suggests the need for maintenance therapy (usually at reduced dosage) since rapid recurrence of illness is observed when drug administration is stopped [40]. Usually, for BOT (in adults) the induction dosing is 2 mg/day, while maintenance dosing is 1 mg/day [60].

The next frontier in EoE therapy aim to reach specific immune targets, through the so-called biological treatment. Several biological therapies, primarily used for treating severe eosinophilic asthma, have been tested in clinical trials for EoE. These include monoclonal antibodies targeting IgE and eosinophils, specifically through the inhibition of interleukin (IL)-5, the IL-5 receptor, and more recently, IL-4/IL-13 [41], [42]. In particular, *dupilumab* is a fully human monoclonal antibody, approved by FDA to treat a variety of atopic disorders, including EoE [61]. It acts on the IL-4α subunit shared by both IL-4 and IL-13 receptors. By blocking these receptors, dupilumab inhibits their signaling, suppressing Th2-mediated proinflammatory cytokines, which play a

key role in pathways involved in atopic conditions [61]. However, to date, dupilumab is the only treatment belonging to this class that has been approved due to its success in achieving the co-primary endpoints of clinical and histological response in phase 2 and 3 of randomized clinical trials [42]. Drugs that directly target eosinophils (i.e., anti-IL-5 agents like *mepolizumab*, *reslizumab*, *benralizumab*, *lirentelimab*) have not been successful. Although these treatments effectively reduced tissue eosinophilia, they did not result in significant clinical improvement, particularly in reducing dysphagia, as measured by assessment scales [42]. While new and powerful drugs offer the potential to greatly improve the natural course of EoE and enhance patients' quality of life, the exact place of biologics in the treatment protocol is still not well defined. It remains uncertain whether biologics should be used as a first-line treatment for most patients or reserved for those who do not respond to initial therapies, such as PPIs, STCs, or allergen-free diets, or for those who experience side effects from these treatments [42].

Dilation

In more severe cases where esophageal strictures significantly impact the patient's functioning, esophageal dilation may be considered. Although this procedure can provide immediate relief, it is important to note that the effect is only symptomatic [41].

1.3 Microbiome analysis

Since the importance of the human microbiome has been clearly highlighted, tools and instruments that allow its analysis are becoming crucial while exploring this relatively new research field. Development of Next Generation Sequencing (NGS) has been the real game changer allowing for precise identification and analysis of complex microbial communities that live around and with us, leading to the era of *meta-omics*. Indeed, a meta-omic study typically aims to identify a panel of microbial organisms, genes, variants, pathways or metabolic functions characterizing the microbial community populating an uncultured sample [62]. This research field got even more boost thanks to important and collective projects like already mentioned Human Microbiome Project (HMP) and others like the European MetaHIT and the Integrative Human Microbiome Project (iHMP) [13], [63].

1.3.1 Meta-omics

In the vast majority of cases, typical workflows for microbiome analysis projects try to address two key questions: *which microbes are present in the sample* and *what are the microbes doing*. To answer them in a complete way, at least two *-omics* techniques are required, going for the so-called *multi-omics* studies. This approach heavily relies on the availability of large amounts of data, their management, statistical analysis and, therefore, requires skilled personnel, investment of time and a lot of resources [64]. Choice of the specific -omic technique depends on research aims and available resources. For this specific research work, meta-omics information provided by *metagenomic* (i.e., direct genome analysis) and *metatranscriptomic* (i.e., direct transcriptome analysis) have been chosen.

For the sake of clarity, the term *genome* refers to the complete set of DNA molecules within an organism, encoding for all the fundamental information that defines traits and characteristics of an individual. In eukaryotic organisms, like humans, genome is organized into chromosomes, each one containing a portion of the total genetic material. In prokaryotic organisms, like bacteria, the genome is typically a single circular DNA molecule. The genome encompasses both coding (i.e., genes) and non-coding regions, so it represents the starting point for protein synthesis but also a lot of other instructions that do not directly code for proteins, but that can account for structural or regulatory

functions, such as gene-expression control. The term *transcriptome*, instead, refers to the complete set of RNA molecules that are transcribed from DNA at a specific time point and/or under particular conditions. It includes all types of RNA: mRNA but also rRNA, tRNA and various non-coding RNAs. The transcriptome provides valuable information about which genes are actively being expressed in a cell, tissue or an organism at given time. It can vary significantly depending on factors like developmental stage, environmental conditions and health state. Therefore, they complement each other giving the full picture about what's going on in a sample when combined.

However, this research field have been traditionally polarized towards the analysis of genome instead of transcriptome due to different reasons. Firstly, the dynamic nature of the transcriptome, which reflects active gene expression that can fluctuate based on numerous conditions, makes the analysis more context-dependent and complex [65], often requiring high number of samples which have non-negligible costs [66]. This is contrasted with the more stable nature of the metagenome, which represents the potential genetic activity of the microbiome [65] which is still a valuable starting point, especially for firstly exploratory studies of rare diseases and atypical conditions. Then there are sampling-related challenges, in particular the ability to acquire enough high-quality RNA [67], and technology limitations related to available sequencing platforms. Therefore, following description will address protocols for metagenomics studies.

1.3.2 Metagenomics sequencing techniques

The two primary methodologies for microbial identification and genotyping are *amplicon sequencing* (or marker gene sequencing) and *shotgun metagenomics*.

Gene amplicon

For more than two decades, gene amplicon has been the unique option for the analysis of complex microbial communities. It uses specific conserved regions of a gene of interest, known as "marker genes" [68], in order to determine microbial phylogenies of a sample. The selected region generally includes a highly variable segment useful for precise identification (i.e., taxa fingerprint [68]), surrounded by highly conserved regions that can act as binding sites or PCR primers [69]. The gold standard involves the use of 16S rRNA (or 16S

rDNA), which is a gene encoding prokaryotic small 30S subunit of the 70S ribosomal complex in most bacteria and archaea [13]. The fact that this region has been conserved highlights its key role in cellular function, allowing for precise genomic classification of microbial taxa [13]. Another reason of its success as marker gene, is the possibility to be sequenced even in large sample sizes, due to its relatively short base pair number ($\sim$ 1542 bp) [13]. When it comes to fungi and yeasts, internal transcribed spacer (ITS) regions have been the most used target for their genotyping [63], [68]. Early studies utilized this method to identify hundreds or even thousands of 16S amplicons, which were cloned into plasmids and subsequently sequenced using Sanger sequencing [68]. With the development of higher throughput sequencing approaches, allowing for the generation of millions of reads, the number of identifiable amplicons (and consequently the number of communities) increased, leading to higher quality results [68]. 16S rRNA (and ITS) sequencing are well-established, rapid, and cost-effective techniques for obtaining a broad overview of a microbial community. However, because DNA sequences differ in the regions amplified by primers [63], these primers do not have the same affinity for all possible DNA sequences, leading to bias during PCR amplification [63], [69]. Additional sources of bias in marker gene sequencing include the choice of variable region, amplicon size, and the number of PCR cycles [69]. While optimizing primer selection can help reduce bias, this approach requires prior knowledge of the microbial community to evaluate the taxonomic resolution and coverage of the target [69]. Despite optimization, primers often achieve only genus-level taxonomic resolution [69].

Shotgun metagenomics

Whole-genome shotgun metagenomics, which sequences all the DNA in a sample rather than focusing solely on amplified marker genes, allows for higher-resolution profiling of microbial communities, reaching strain level and enabling the study of gene content, function, and genomic variation [68]. Techniques like high-throughput 16S rRNA gene sequencing, which focus on specific organisms or single marker genes, are often labeled as metagenomics. However, this term is inaccurate because these methods do not analyze the complete genomic content of a sample [70]. Capturing the full repertoire of genetic information allows the study of all microorganisms (i.e., bacteria, fungi, DNA viruses, and others) in a culture-free manner [68], but reference genomes

and scientific knowledge are required [63]. Especially the possibility to reach a strain level analysis makes this approach particularly powerful. Bacterial strains can differ significantly in traits like pathogenicity, immune evasion, antibiotic resistance, and metabolic capabilities [68]. Although environments such as the human gut may host organisms from the same species, there is considerable inter-individual variation in the strains present [68]. Strain-level sequencing was proven to discriminate significant differences in microbial activity that may be linked to disease, revealing important connections that would be missed with higher-level taxonomic profiling [4]. There are various experimental and computational methods available for each stage of a typical shotgun metagenomics study, presenting researchers with a wide and potentially insidious range of options. While the technique is generally reliable and appears straightforward, it does have limitations, including potential experimental biases and the complexities involved in computational analyses and their interpretations [70]. In Figure 1.4 are summarized various experimental and computational challenges associated with both 16S rRNA-based and shotgun metagenomic sequencing.

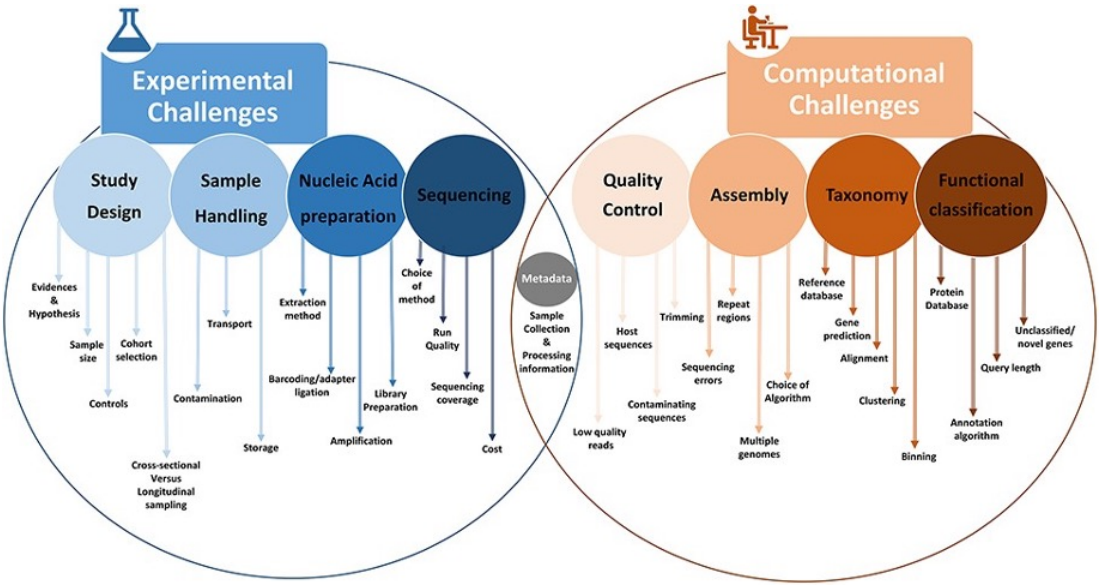


Figure 1.4: A schematic overview of experimental and computational challenges associated with both amplicon and shotgun metagenomic sequencing. Adapted from [13].

Among all of them, the one that is likely the main reason why this technique has not been widely adopted is cost. Advantages coming from shotgun metagenomics are directly related to the amount of available data or, in other words, to the number of sequenced reads per sample (i.e., sequencing depth). It is a crucial factor that impacts the accuracy, resolution, and

interpretability of metagenomic data. Indeed, species/strain level and the assembly of whole microbial genomes from short DNA sequences reads are possible only if an adequate sequencing depth is given [69]. Unfortunately, a higher resolution leads to increased expenses [68], making this approach unavailable for small-funded research groups. A possible work around for this could be decreasing sequencing depth, adopting the so-called low pass or shallow sequencing, searching for a trade-off between resolution (still trying to reach higher resolution than 16S amplicon sequencing) and costs [68].

1.3.3 Typical microbiome acquisition protocol

Experimental design choice

A well-thought study design is a crucial starting point in microbiome analysis, requiring a careful planning to enhance data processing steps and mitigate confounding effects, ultimately leading to more reliable and robust findings [13], [69]. The scope of experiment and research questions should be clearly defined, as it allows for the decision of an appropriate study design. For example, a *cross-sectional study* is a good choice when it comes to compare different populations, like patients versus healthy controls [13], [69]. These studies are generally simpler to design and execute, with no need for extended followups [13]. However, observed differences are not easily attributed to a single effect or treatment, since they could result from various additive or multiplicative effects [13], [69]. Indeed, the large variation in the microbiome between individuals and the profound impact of environmental factors (i.e., lifestyle, diet, etc.) must be considered, because they could be the origin of observed differences [13], [69]. *Longitudinal studies* that collect a baseline and then observe populations across multiple time points are better from a statistical point of view [13], but they are not free from potential bias (e.g., during sample collection) and they are more expensive [69]. For identifying specific effects of a course of treatment on the microbiome and disease state, an *interventional study* (i.e., double blind randomized control studies) should be considered.

Regardless the study type, key aspects must be addressed. In order to be able to discern technical variability and real biological results, statistical power must be assessed [69]. Generally, findings from small samples may not accurately reflect broader populations [13]. Choosing an adequate sample size based on statistical principles is particularly vital when the results are intended for clinical use and

interpretation, and avoiding adjustments on the go is highly recommended [13]. Defining clear inclusion and exclusion criteria for subjects involved help limiting confounding covariates [13], [69]. An example of exclusion criteria might include exposure to drugs that are known to affect the microbiome, such as antibiotics [70]. Once the right experimental design has been chosen and all details have been clearly stated, an acquisition protocol must be defined. Protocols usually share some fundamental key steps that will be described below considering a shotgun metagenomic approach.

Sample and metadata collection

Sample collection could be a significant confounding factor that might affect the results and interpretations of a study [13], [69], [70]. During this phase, three key aspects should be addressed: *contamination*, *transportation* and *storage* [13]. Proper sample environment control during collection is vital to prevent contamination and immediate freezing and aseptic handling minimize the risk of cross-contamination and sample degradation. Even if short-term storage temperature has minimal impact on microbiome structure, consistent storage conditions are necessary for optimal nucleic acid integrity and yield before sequencing. It must be noted that nucleic acid extraction methods can introduce biases, therefore the chosen protocol should be described in research papers in order to allow for study comparability [13].

Along with biological samples, also metadata should be carefully collected. They are essentially "data about data" [71], a catalogue containing all the samples' details used in the experiment. In many studies, this is structured as a matrix with samples as rows and metadata categories (e.g., age, sex, BMI, disease state, etc.) as columns [69]. They can include sample's owner information, details about preliminary steps, extraction protocol parameters and other type of knowledge that could be relevant for the analysis's purposes. Information about subjects involved in the study are typically collected using paper or digital questionnaires and they are extremely useful for later downstream analyses [13], since they allow for confounders management. For example, in case-control experimental designs, metadata categories are typically used as control criteria during matching process [69]. Moreover, many contemporary statistical comparison tools require metadata input [13].

Sequencing platform and reads preprocessing

Next crucial step is the choice of sequencing platform, which aim is to produce raw reads as its output. In bioinformatics, in the context of DNA sequencing, a *read* refers to the sequence of base pairs (or base pair probabilities) derived from a single DNA fragment [72]. In fact, a typical sequencing experiment requires the preliminary fragmentation of starting genomic material. Fragments are then size-selected and ligated to adapters, in order to build the so-called *sequencing library*. *Adapters* are short, synthetic DNA sequences that are attached to both ends of the DNA fragments during library preparation. They provide a uniform and standardized sequence at the ends of DNA fragments which is necessary for at least two reasons: amplification, since they serve as binding sites for primers used in the PCR amplification step, and parallelization (i.e., simultaneous sequencing of multiple samples). Indeed, adapters often contain unique index sequences, known as *barcodes*, that allow for the identification of different samples within a single sequencing run (i.e., *multiplexing*).

The number of bases that are read at one time (i.e., the number of letters that will appear in each read) is known as *read length* [73]. Read length depends upon the sequencing platform of choice. Current sequencing platforms are broadly classified as either short-read (250–300 bp; Illumina) offering higher sequencing depths or long-read (500–4000 bp; PacBio and Oxford Nanopore) offering better contig assembly [13]. Long-read sequencing does not need initial fragmentation of starting genetic material.

Once raw reads have been collected, a *quality filtering step* is mandatory due to many reasons. First, as mentioned, high-throughput sequencing technologies relies on adapters usage that do not contain any biological information. Therefore, they need to be removed prior the execution of analysis. Second, each read is associated to a *quality score*. Sequencing quality scores measure the probability that a base is called incorrectly. With sequencing by synthesis (SBS) technology, as the one implemented by Illumina sequencer, each base in a read is assigned a quality score by a phred-like algorithm, similar to that originally developed for Sanger sequencing experiments [74]. The sequencing quality score of a given base (Q) is computed as: $Q = -10log_{10}(e)$, where e is the estimated probability of the base call being wrong [74]. Lower Q scores may lead to increased false-positive variant calls, resulting in inaccurate conclusions. In Table 1.2 relationship between sequencing quality score and base call accuracy provided by Illumina is reported [74]. As it can be seen, $Q30$ sequencing quality correspond to a situation

where virtually all reads will be perfect. Indeed, $Q30$ is considered a benchmark for quality in NGS [74].

Quality score	Probability of incorrect base call	Inferred base call accuracy
$Q10$	1 in 10	90%
$Q20$	1 in 100	99%
$Q30$	1 in 1000	99.9%

Table 1.2: Relationship between sequencing quality score and base call accuracy. Adapted from [74].

Clearly, high quality reads are desirable, since they indicate a high level of confidence in the accuracy of each base call, ensuring a more correct representation of samples' genetic material. Even if advancement in technology made it possible to have the vast majority of bases scoring high quality scores, still there could be some of them that needs to be discarded during preprocessing steps. Last part of preprocessing needs to remove human genomic sequences, since only microorganisms' genetic material is of interest for microbiota research. This part is known as *decontamination* and is carried out by discarding any reads that align to the human genome reference.

In Table 1.3 several adapter trimming and quality control tools are listed. In particular, *Tirmmomatic* is designed for Illumina sequencing platform, and it is capable of manage both single-end (i.e., each DNA fragment is read from one end of the fragment to the other, producing a single continuous sequence read) and paired-end reads (i.e., both ends of a DNA fragment are sequenced, producing two reads that are typically of equal length) [13]. *FastQC* is a useful tool to assess quality of reads before and after trimming. *MultiQC* complement FastQC action, summarizing and reporting results from multiple samples within a single file [75]. *BWA* and *Bowtie* are among the most widely used tools for decontamination, capable of aligning short sequencing reads to reference genomes [75]. *KneadData* is known as a wrapper tool, since it incorporates all of the above functions including quality control, adapter trimming and host genomic sequences removal, remarkably simplifying preprocessing steps [75]. All these tools work with FASTQ input files, which is a common file format for sharing sequencing read data, combining both sequence and associated per base quality score [76]. Once preprocessing is complete, reads' analysis can begin using either a reference-based or assembly-based approach [13].

Software	**Major functions**	**Description and advantages**
FastQC	Quality assessment.	Identifies potential problems; Summary report production.
Cutadapt	Quality control, adapter trimming.	Removes unwanted sequence.
Trimmomatic	Quality control, adapter trimming	Good performance in trimming poor-quality data; Contains a library of Illumina adapters and primer sequences; Used for Illumina platform sequence data.
Trim galore	Quality control, adapter trimming,	A wrapper tool for FastQC and Cutadapt; Automatic identification of adapters.
fastp	Quality control, adapter trimming.	Fast; Automatic identification of adapters; Summary report production.
KneadData	Quality control, adapter trimming, removing host genomic sequence.	A pipeline that integrates multiple tools, including Trimmomatic, FastQC, and Bowtie2; Designed for metagenomic and metatranscriptomic sequencing data.

Table 1.3: Software programs used for preprocessing raw sequences reads. Adapted from [75].

Reference-based analysis

Reference-based approach relies on the availability of ever-growing reference databases. There are essentially two class of tools to obtain compositional profiling of communities from metagenomic sequencing data that needs an existent reference:

- *Taxonomic profilers* depending on a small set of curated genomic markers that can be either *universal* (e.g., MIDAS) or *clade-specific* (e.g., MetaPhlAn) [13], [68], [77];
- *Taxonomic binners* which compare metagenomic reads against reference utilizing *kmers based* (e.g., Kraken and One Codex) or *alignment-based* binning methods (e.g., MEGAN, MALT and Sigma) [13], [68], [77].

Taxonomic profilers use a limited set of reference sequences or known marker genes to profile microbial communities, offer high computational efficiency and accuracy [68], [77]. On the other hand, taxonomic binners use whole-genome or whole-genome fragments. Nevertheless, taxonomic profilers often exhibit lower resolution in species-level identification compared to whole taxonomic binners, particularly when a species is present in low abundance within the sample. This can result in only a limited number of reads aligning to a narrow set of markers [68], [77]. Whole genome mapping instead provides a more comprehensive understanding of the microbial community, including details on gene content, coverage, and variants [68]. They have enhanced sensitivity benefiting from much larger databases, being able to usually detect more microbial species at very low abundance [77]. However, these methods often suffer from inaccuracies, such as false positives, because of the incomplete nature of these databases [77]. This can lead to reads from unrepresented taxa being incorrectly assigned to various related organisms [77]. While marker-based taxonomic profilers struggle with identifying microbial species of low abundance, taxonomic binners can present challenges in interpretation due to their lower precision [77]. Moreover, k-mers requires substantial memory to handle larger databases and they do not indicate the specific location of the match within the reference genome [68].

Identification of genes and functions through read-based analysis is crucial for understanding the functional potential of microbial communities. This process involves quantifying gene and pathway abundances in shotgun

metagenomes, employing techniques similar to those used for taxonomic profiling. Specifically, these techniques classify, or map reads directly to pangenomes (i.e., entire set of non-redundant genes across all genomes within a specific clade) and non-redundant protein databases [68]. These tasks can be effectively carried out using tools like *HUMAnN* or *MEGAN* [13], [68].

HUMAnN is a widely used tool that utilizes two-step alignment process. It requires taxonomic information as input (if it won't be provided, as it was developed by the same team, it natively runs MetaPhlAn as a preliminary step [78]) in order to narrow the search scope for the subsequent pangenome search, where reads are matched to genes in the identified taxa and gene abundance is measured. In the final step, reads that are either ambiguously mapped or unmapped are aligned to a non-redundant protein database using a translated search, where each read is translated in all six reading frames to identify potential protein sequences [68].

MEGAN, on the other hand, annotates sequences using KEGG orthology and COG/NOG classifications based on SEED classifications. For long-read sequences, the DIAMOND aligner can be used either independently or in conjunction with MEGAN to perform both pairwise and frameshift alignments [13].

Assembly-based analysis

Many microorganisms within natural microbial communities lack available genomes, because many taxa are difficult, or even impossible, to be cultivated in laboratories. Even considering significant efforts and technological advancement like single-cell genome sequencing, the need for special equipment and technical challenges are still open problems [75]. This led to assembly-based approaches development, which focus on utilizing sequencing reads to create genome assemblies for individual microorganisms within a community, eliminating the need to isolate and culture organisms to generate microbial genomes. This approach has also greatly expanded our understanding of the diversity among uncultivated species across various environments [68]. This process begins by assembling contigs through the overlapping sequences between reads [75]. Following this, sequence-intrinsic characteristics like tetranucleotide frequency and coverage are used to "bin" contigs from the same microbial genome, resulting in metagenome-assembled genomes (MAGs) [68].

As it can be imagined, even this approach is far from being challenges-free.

There are lot of open questions and problems. For example, MAGs vary significantly due to the absence of uniform bioinformatics procedures and quality standards [75]. Additionally, essential metadata, such as sample sources and sequencing details, are often missing, complicating data reuse and integration [75]. Moreover, MAGs are derived from bacterial sequences, with few archaeal, viral, or fungal genomes recovered due to varying abundances and the difficulty in assembling genomes from low-abundance species, creating a bias that limits our understanding of the full microbial community [75]. Looking ahead, advancements in machine learning are anticipated to enhance metagenome assembly processes, either by organizing reads prior to assembly or by refining assembly graphs [68].

Hybrid analysis

Reference-based methods are limited since they are able to detect only microbial species annotated in available databases (typically a small portion of the community object of interest), but they offer great sensitivity and resolution for known organisms. They require lower computational power and storage, and their results are more straightforward to interpret. On the other hand, assembly approaches allow for detection of microbial sequences that have not yet been isolated, but they are not able to perform well in complex environments due, for example, to varying abundances. At this point, it is clear why future work should be focused on understanding how to achieve "the best of both worlds".

Some tools that have been heading in this direction are available. In particular, the last available version of *MetaPhlAn*, *MetaPhlAn 4*, aims to (at least, partially) combine these two approaches. This release utilizes an integrated and extensive collection of microbial genomes and Metagenome Assembled Genomes (MAGs) to establish a broader set of species-level genome bins (SGBs) and precisely assess their presence and abundance in metagenomes [79]. SGBs encompass both known species (kSGBs) and previously uncharacterized species (uSGBs), which are defined solely through MAGs. This expanded dataset enhances the MetaPhlAn algorithm, allowing for more in-depth and accurate quantitative taxonomic analyses of human, host-associated, and environmental microbiomes, and offers valuable insights into studies linking the microbiome with host conditions [79].

Postprocessing analyses

At this point of data processing, results will typically consist of data matrices representing samples versus microbial features, such as species, taxa, genes, and pathways [70]. While this output may appear straightforward, it is actually quite complex, often involving thousands of distinct features and matrices characterized by sparsity [69]. Indeed, large multivariate structure of these data makes retrieving meaningful biological information significantly challenging, requiring meticulous statistical handling to interpret and understand possible existing relationships between findings and sample metadata [70].

Identifying a universally applicable postprocessing pipeline for statistical analysis is generally difficult, as it heavily relies on the specific objectives and hypotheses of each study [13]. However, in most cases, the analysis typically focuses on identifying differences in microbial diversity, taxa abundance, or functional components (i.e., genes or pathways) between comparison groups (e.g., patients vs healthy controls, treatment vs placebo) [63]. Statistical tests must follow regardless of the particular method or index applied, in order to assess if differences or trend observed are meaningful. Statistical methods used for microbiome analysis are constantly evolving due to the inherent complexity of these datasets [63] and many of the statistical techniques applied in this context are not unique to metagenomics [70].

Typically, a first line analysis consists in *alpha* and *beta diversity* calculation, in order to highlight possible patterns across microbiome variations [69]. They are two key concepts in ecology and microbial community analysis used to quantify and compare the diversity of species or taxa within and between communities or samples. More specifically, alpha diversity metrics provide a summary of a microbial community's structure in terms of richness (i.e., the number of taxonomic groups), evenness (i.e., the distribution of abundances among those groups), or a combination of both [80]. Commonly used alpha diversity metrics include *phylogenetic diversity (PD)*, *Chao1*, *Simpson's index*, and *Shannon's index* [80]. Instead, beta diversity metrics provide a summary of differences between samples by accounting for sequence abundances or simply their presence and absence [80]. Frequently used beta diversity metrics can be divided into two major categories: quantitative metrics (e.g., *Bray–Curtis*, *Canberra* and *weighted UniFrac*) that use features abundance data, and qualitative metrics (e.g., *binary Jaccard index* and *unweighted UniFrac*) that only consider the presence or absence of features [69].

Difference between these two types of indexes is crucial: alpha diversity measures the diversity of features *within* individual samples and allows for comparisons across different sample groups (e.g., researchers can use alpha diversity indices to compare the average species diversity between a sample from a diseased individual and a healthy control) [69]. In contrast, beta diversity calculates a *distance matrix* that quantifies the differences in microbial composition *between* all pairs of samples [69]. There are different solutions available for these indexes' computation such as software like *QIIME* and *Mothur*, R packages like *DESeq2* and *vegan* [69], [70] or python packages like *scikit-bio* [81].

As a following step, most researchers start digging into deeper observations by *visualizing* their data to search for potential associations or markers, which can then be examined using more robust statistical techniques [63]. Given the complexity of microbiome data, visualization techniques often utilize dimension-reduction methods such as principal coordinate analysis (PCoA) or principal component analysis (PCA) [63], [69], [70]. These methods condense distance matrices into two- or three-dimensional visualizations, making it easier to represent sample distances (i.e., beta diversity) [63], [69]. Samples can then be labeled with different categories (e.g., using color or shape) to overlay relevant clinical metadata, enables researchers to visually explore potential clustering patterns based on clinical variables in an unsupervised way [63], [69], [70]. More sophisticated statistical evaluations can be employed to assess whether observed clustering patterns are biologically significant. For examining differences in overall community composition, ANOSIM (Analysis of Similarities) is used to determine significant clustering by comparing the similarity within groups versus between groups using distance metrics [63]. ANOSIM tests the null hypothesis that the average similarity within one group is the same as the average similarity between different groups [63]. PERMANOVA (Permutational Multivariate Analysis of Variance) is another nonparametric permutation test that evaluates the overall difference in microbiome community structure between different clusters or groups based on distance matrices [63], [69]. Heat maps are another widely used method for visualizing microbiome data, helping to identify potential clusters or differences between group, typically displaying taxa abundance across samples or the presence/absence of specific gene families [63], [70]. Also, machine learning methods (e.g., classifiers like *random forests* or *support vector machines*) could

represent a viable approach [70] but it has to be considered they typically require a substantial sample size, and they should always be complemented with cross-validation, independent test sets, or other experimental and biological validations to ensure robust and reliable findings [69].

Need for standardized pipelines

Despite significant advancements in the field, one of the main challenges remains the selection of appropriate software or pipeline from available bioinformatic tools [75]. This difficulty is compounded by the absence of standardized laboratory and computational protocols, which raises concerns about the reproducibility of published microbial sequencing studies [13]. The lack of standardization often leads to the introduction of biases and results that cannot be easily compared across different studies [13]. To address these issues, there is a growing need to integrate bioinformatics pipelines and controls into standardization efforts [69]. This includes using possibly cloud-based, but above all reproducible computing resources that rely on open-source code and publicly available data to validate scientific findings, enhancing consistency and comparability across the microbiome research field [69]. By establishing and disseminating best practices, the field can better translate findings from controlled laboratory clinical settings, ultimately fulfilling its promise of broad applicability and impact [69].

1.4 Research aim and objectives

The rising awareness about the role played by the microbiota in human health, its already demonstrated involvement in a lot of different conditions, especially those with inflammatory patterns, the advances in sequencing technologies and the availability of powerful analysis tool clearly suggest that this dimension is holding the potential to provide soon novel and useful biomarkers for diseases diagnosis and monitoring. The increasing trend of Eosinophilic Esophagitis (EoE), its significant burden and impact both on patients' quality of life and national health system and, above all, the non-availability of a reliable non-invasive diagnostic biomarker represent an interesting opportunity to pursue the ongoing "microbiome revolution". Indeed, the aim of this research work is beginning to investigate across different available tools for microbiome analysis, laying the groundwork to establish more clearly the potential existing relationship between human microbiome and EoE.

In deeper details, objectives of this research work are:

i) to evaluate and compare some of the most widespread bioinformatic tools for microbiome sample analysis;

ii) to establish a reliable pipeline for sequencing data preprocessing and subsequent taxonomic/functional annotation;

iii) to define a valuable metadata management strategy, in order to aid retrospective case-control group design and consequently downstream analysis.

Chapter 2

Materials and methods

2.1 Partecipant recruitment

2.1.1 Recruitment of EoE Patients

Patients' recruitment was made in the context of a collaboration with the gastroenterology unit of the University Hospital of Padova. Already diagnosed patients and potential ones, aged 18 years or older, who had an endoscopy scheduled from October 2022 onwards, were invited to participate in the study and provided with informed consent forms. Participants were sent a survey in order to collect *clinical* and *lifestyle* information, along with two kits for self-collection of saliva and stool samples, approximately a week before the scheduled endoscopy. Participants were instructed to collect these samples within 48 hours before their appointment and complete the survey on the same day. Both kits contained a DNA/RNA preservative to maintain nucleic acid integrity until delivery to the University of Padova, where the samples were stored at -80°C.

2.1.2 Recruitment of Healthy Volunteers

A flyer outlining the project goals and eligibility criteria for healthy volunteers was distributed across various university buildings. Interested individuals completed a digital form to provide clinical and lifestyle metadata, which was then processed using an in-house algorithm implementing a quasi-experimental method of *propensity score matching*, in order to match volunteers with patients based on age, sex, BMI, and diet. Selected volunteers were given the same self-collection kits for saliva and stool samples, along with an additional survey

to gather further metadata. Samples were required to be delivered to the University of Padova within 48 hours, where they were subsequently stored at -80°C.

2.2 Sample Collection

2.2.1 Saliva Sample Collection

Participants were advised not to consume food or beverages for at least 30 minutes prior to saliva collection. Saliva samples were self-collected using the DNA/RNA Shield Saliva Collection Kit (Zymo Research), following the manufacturer's guidelines. The samples were kept at temperatures below 24°C until delivery to the University of Padova and then stored at 80°C until nucleic acid extraction.

2.2.2 Stool Sample Collection

Participants self-collected stool samples at home using the DNA/RNA Shield - Fecal Collection Tube (Zymo Research) according to the manufacturer's instructions. Samples were maintained below 24°C during transportation to the University of Padova and were stored at -80°C upon arrival until nucleic acid extraction.

2.3 Extraction protocol and sequencing

2.3.1 Nucleic Acid Extraction

Nucleic acids were isolated from saliva and stool samples using the ZymoBIOMICS™ DNA/RNA Miniprep Kit (Zymo Research), with protocol optimization for each sample type. Mechanical lysis was achieved using a vortex with a specialized adaptor, and the duration of 10 and 40 minutes was tested. The Zymo Research mock community was also processed to evaluate the efficiency of the protocol. Additionally, the ZymoBIOMICS™ Spike-in Control I (High Microbial Load) was included in each sample for quality control purposes. To assess the nucleic acid extraction protocol's effectiveness in uniformly extracting nucleic acids from Gram-negative and Gram-positive bacteria, as well as yeasts, the ZymoBIOMICS Microbial Community Standard was processed in

the same way as other samples. Furthermore, saliva and stool samples were processed under two different lysis conditions, and two different spikein concentrations were tested to further refine the protocol.

2.3.2 Pilot study

To evaluate the effectiveness of the nucleic acid extraction protocol in uniformly extracting nucleic acids from gram-negative bacteria, gram-positive bacteria, and yeasts, the ZymoBIOMICS Microbial Community Standard was processed under the same conditions as the other samples. Additionally, saliva and stool samples were subjected to two different lysis conditions, and two different spike-in doses were tested to further refine the protocol. This led to the origin of first pilot study samples that have been subsequently sequenced and used as initial test samples for developed pipelines. Pilot study samples consist of 2 saliva (ID: 118500, 118501) and 3 feces samples (ID: 118502, 118503, 118504) coming from the same subject. In Table 2.1 extraction protocol parameters are detailed.

Sample ID	**Patient ID**	**Class**	**Type**	**Spike-In (uL)**	**Vortex (min)**	**V (uL)**	**Concentration (ng/uL)**	**ug**
118500	P10PRM	Control	Saliva	-	10	45	580	26.1
118501			Saliva	20 (I)	40	45	900	40.5
118502			Saliva	20 (II)	40	45	600	27
118503			Feces	-	10	45	510	22.95
118504			Feces	100 (II)	40	45	520	23.4
118506	-	-	Mock	-	10	45	42.9	1.9

Table 2.1: Extraction protocol parameters of first pilot study samples.

Once the extraction protocol was assessed, additional samples were processed following the selected protocol, introducing both patients and controls to the subjects pool, leading to second pilot study samples, detailed in Table 2.2.

2.3.3 Shotgun metagenomic sequencing

Library preparation was conducted using the Celero™ DNA-Seq kit (NuGEN, San Carlos, CA) according to the manufacturer's instructions. Both input DNA and the final library were quantified using a Qubit 2.0 Fluorometer (Invitrogen, Carlsbad, CA) and assessed for quality with an Agilent 2100 Bioanalyzer High Sensitivity DNA assay (Agilent Technologies, Santa Clara, CA). Sequencing libraries were prepared and sequenced on the NovaSeq6000 in pairedend 150 bp

Sample ID	Patient ID	Class	Type	Spike-In (uL)	Vortex (min)	V (uL)	Concentration (ng/uL)	ug
132659	C57TM	Patient	Saliva	5 (l)	20	50	162.2	8.11
132660	C45MF	Patient					143.1	7.16
132661	C54SN	Patient					356.3	17.82
132662	C55SF	Control					90.2	4.51
132665	C52MP	Patient					211.5	10.58
132667	C47BL	Patient					18.8	0.93
132668	C49BC	Patient					258.9	12.95
132663	C45MF	Patient	Feces	50 (l)	20	50	30.9	2.32
132664	C53GG	Patient					127.8	9.59
132666	C49BC	Patient					248.2	18.62
132669	C48GG	Patient					100.9	7.57

Table 2.2: Extraction protocol parameters of second pilot study samples.

mode. Reads were trimmed to a length of 150 bp. Base calling, demultiplexing, and adapter masking were performed using Illumina BCL Convert v3.9.315. During demultiplexing, adapter sequences were converted to 'N' characters, and the quality was adjusted to 2 to facilitate downstream trimming.

2.4 Preprocessing pipelines comparison

Performance of two different preprocessing pipelines has been assessed, using as test-input paired end sequences coming from first pilot study samples, stored as FASTQ files. First pipeline was a *custom-built* one, originating from the combination of multiple tools for quality filtering and decontamination. The other pipeline was an *automated* one, completely relying on the use of a wrapper tool known as *KneadData* from *biobakery*, a meta'omic analysis environment [82]. Pipelines performances evaluation focused around two factors: reads quality pre and post run and the number of reads survived after each step (i.e., available for subsequent analyses). For the sake of simplicity, only sequences surviving in pairs have been considered (i.e., since they are paired ends, if one sequence failed to survive, even the other one was discarded) for factors computation and even for the rest of all subsequent analyses. Reads quality was provided by *FastQC (v0.11.9)* [83], which reports were merged by *MultiQC v1.21* [84]. Reads count was provided by in-house bash script, that simply counts and divides lines number of fastq files (available in the project repository).

Custom-bult pipeline run *Trimmomatic v0.39* [85], with the following set of

parameters:

ILLUMINACLIP:TruSeq3-PE.fa:2:30:10:2:True
SLIDINGWIDOW:4:28 LEADING:20 TRAILING:20 MINLEN:50

Then, it run *Bowtie v2.4.2* [86] with default set of parameters and in *–end-to-end / –sensitive mode*. As human genome reference, the GRCh38 (Genome Reference Consortium Human Build 38) was used [87].

Automated pipeline run *KneadData (v0.12)* with default parameters and *-p 4 -t 16* for computational efficiency. Since it's a wrapper, it requires different tools pre-installed on machine, including *Trimmomatic* and *Bowtie* also used for custom pipeline. Therefore, their versions and human reference were the same, Also, *Tandem Repeat Finder (TRF) (v4.09.1)* [88] has been enabled in this configuration. Before running KneadData, a preliminary correction step was needed since original sequences' headers were not natively compatible. An in-house code to automate the correction process was developed, based on the suggestions provided in the biobakery forum [89].

2.5 Taxonomic profiling and functional annotation

For taxonomic profiling, *MetaPhlAn (v.4.1.0)* [79] was installed in a dedicated conda environment and run with default parameters, with *–nproc 4* for computational efficiency. In order to reduce run time, previous intermediate bowtie output was provided in input, using the *–bowtie2out* flag. The marker gene database version used was the vJun23_202307. To merge all the taxonomic profiles obtained from each sample, *merge_metaphlan_tables.py* utility provided by biobakery was used.

For functional annotation, *HUMAnN (v3.9)* [90] run with default parameters and taxonomic profile previously produced by MetaPhlAn provided in input. To reduce required runtime during development, –bypass-translated-search flag has been enabled. Therefore, only nucleotide search has been performed, with the ChocoPhlAn database at version v201901_v31. HUMAnN is not able to manage paired-ends information, therefore all reads have been joined together in a single FASTQ file for each sample by an in-house script, available at the project repository. Original quantification of genes and pathways are provided in units of RPKs (reads per kilobase), accounting for gene length but not sample sequencing depth. For

downstream analyses, both gene families and pathways abundances results have been normalized to CPM (copies per million) units using *human_renorm_table* utility included in the HUMAnN package. Also here, pathways and gene families' tables across different samples have been merged in a unique table using *humann_join_tables* utility.

2.6 Downstream analyses

For downstream and statistical analyses, a Python script was custom-built, tested on a restricted selection of samples coming from both first and second pilot study (only one sample per type per patient has been considered, discarding multiple ones and the mock reference community). Pandas [91] data frames have been used as main data structure. It also requires NumPy [92], SciPy [93], Scikit-bio [81], Seaborn [94], and Matplotlib [95] packages.

2.7 Propensity Score Matching

For Propensity Score Matching (PSM) computation, an in-house algorithm (available in the project's repository) has been developed using Python coding language, requiring Pandas [91], NumPy [92], Scikit-learn [81], Matplotlib [95], and Seaborn [94] packages. Metadata used for algorithm testing have been extracted from the above-mentioned surveys provided to both patients and healthy controls through Google Forms. To be processed by PSM algorithm, unstructured data have been modeled, resulting in only continuous, binary or categorical variables. Information gathered from multiple questionnaire items (e.g., related to diet and lifestyle) has been consolidated into a single variable using a in-house developed scoring system, where each response was assigned a specific score, and then either summed or averaged, depending on the context.

Chapter 3

Results

3.1 Custom-built vs automatic preprocessing pipeline

Quality filtering and decontamination are essential preprocessing steps that must be considered in any genomic or metagenomic analysis. Therefore, this research work started with a preliminary analysis of potential preprocessing pipelines, aiming to explore available tools and assess which would represent the most suitable fit for this work's specific purposes. Two potential pipelines have been tested: a *custom-built* one, manually assembled implementing the use of trimmomatic for quality filtering and bowtie for decontamination, and an *automated one*, implementing the use of a wrapper tool known as KneadData that largely relies on the use of the same tools with slightly different parameters and adjustments. The two different pipelines are summarized in Figure 3.1 and 3.2.

First pilot study samples have been used as a test dataset to conduct all these preliminary analyses, composed by 2 saliva (IDs: 11500, 118501) and 3 feces (IDs: 118502, 118503, 118504) samples. Before carrying out any operation, a starting quality baseline has been defined using *FastQC*, a tool designed to perform quality control on raw sequencing data generated by high-throughput sequencing workflows. It provides a collection of insights and statistics that allow for a rapid evaluation, helping identify potential issues in the data before proceeding with more in-depth analysis. Specifically, it reports useful basic information like sequences length, GC content percentage, along with some detailed graphs about, for example, per-base sequence quality, average

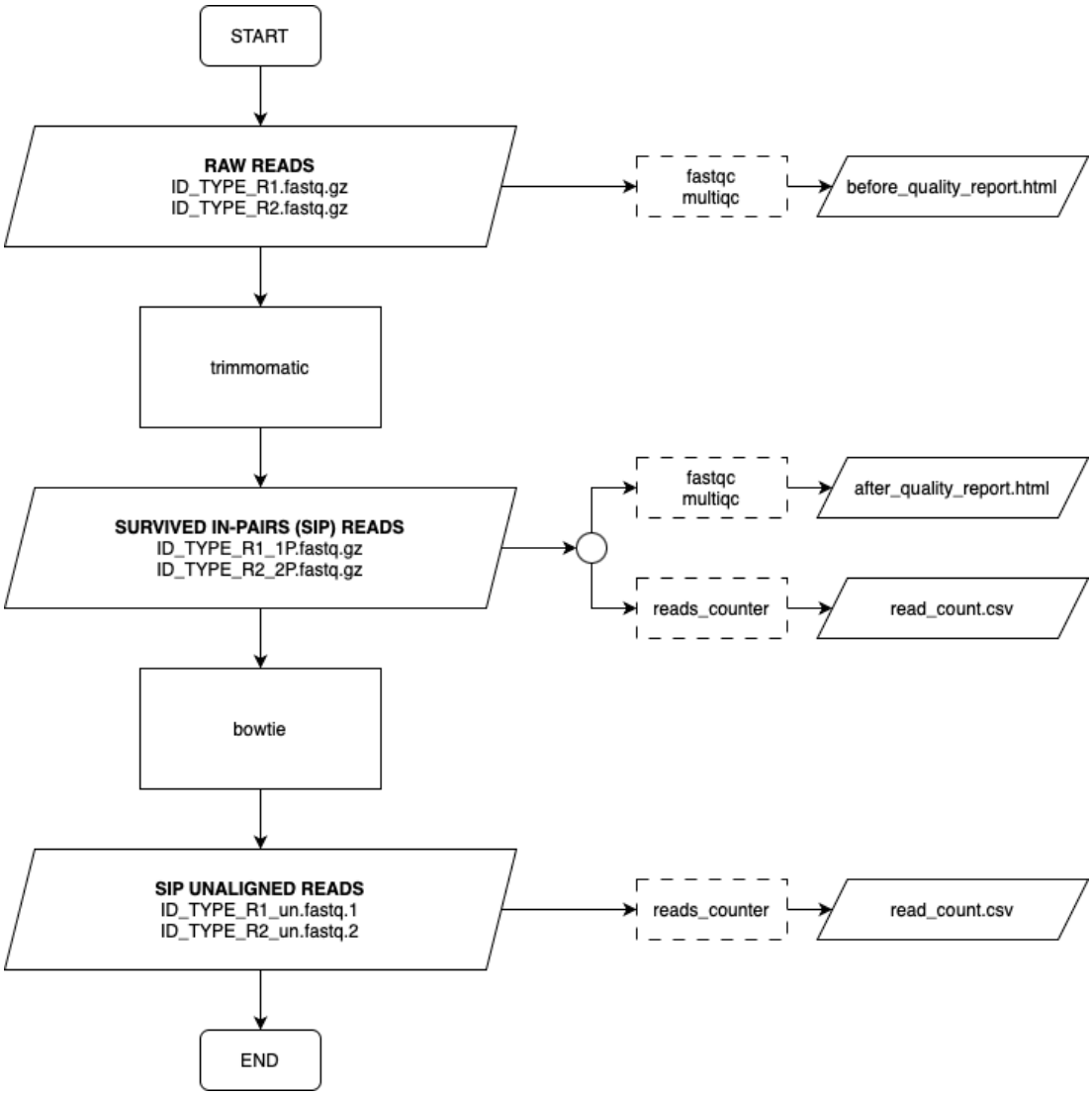


Figure 3.1: Schematic representation of custom-built pipeline.

per-sequence quality and per-base sequence content. All these data are provided in output in the form of html report file. Given that an html report is produced for each FASTQ file, the use of another tool, *MultiQC*, allows the creation of a unique summary from which the following figures have been taken.

In Figure 3.3, mean quality scores are reported: they represent the mean quality value at each base position across all the reads. On the x-axis, base position is considered. Instead, on the y-axis, Phred quality scores are represented. Every line represents a FASTQ file: since the considered 5 samples have been sequenced in paired ends mode, there are two files for each sample (forward and reverse reads), resulting in a total of 10 lines plotted. The plot area is divided into three color coded sections. At the top, in green, there are

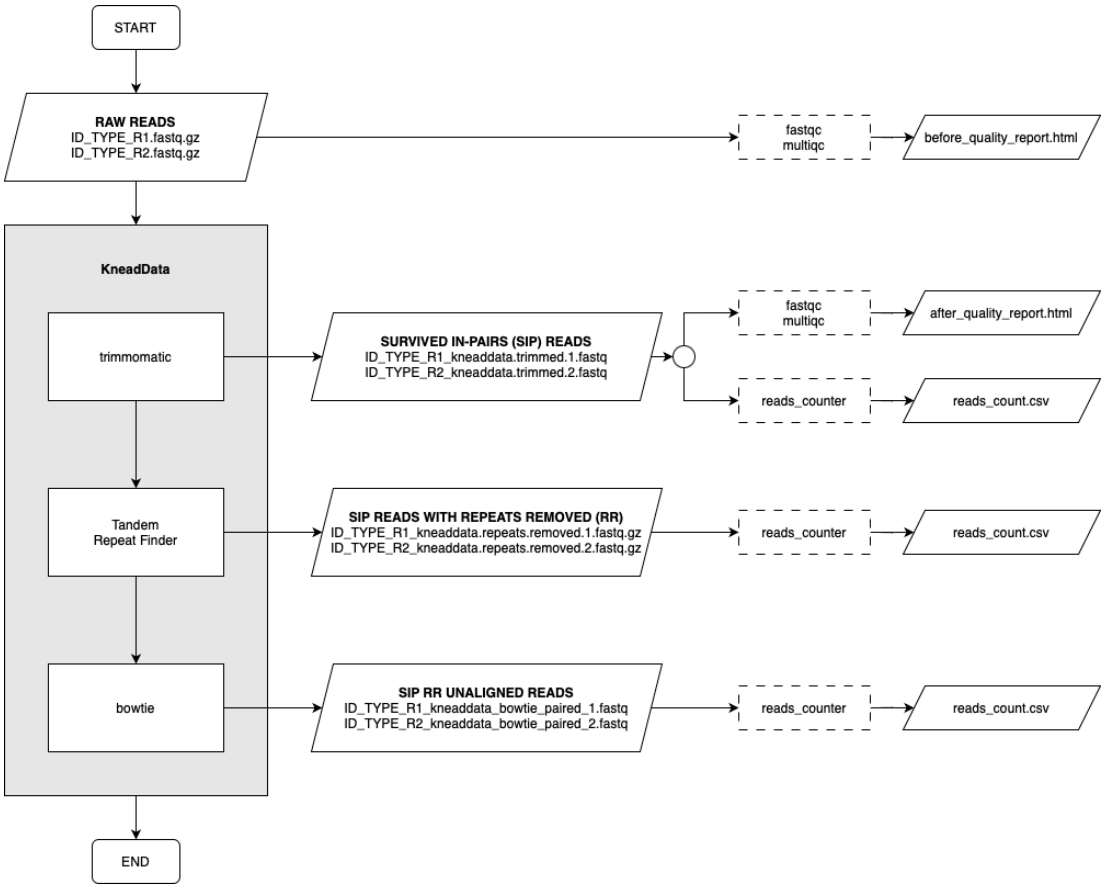


Figure 3.2: Schematic representation of automated pipeline.

desirable quality scores. Instead, at the bottom, in red, it houses undesirable and insufficient quality scores that are typically discarded. These two sections are separated by a middle-one, in yellow, where quality is questionable. As it can be seen, first pilot study samples had a general good quality: for each sample, almost all base pairs are contained in the green area, especially the ones in first positions. At the end of the sequences, a reduction in base call accuracy is observed, resulting in mean lower quality scores. This is a well-documented phenomenon, that can be attributed to various technical factors that affect sequencing accuracy as the sequencing reaction progresses. In the later cycles of the read, the quality of the signal tends to degrade, leading to a reduction in base calling accuracy [96], [97]. This is one of the reasons why quality filtering is needed.

Another useful insight is provided by per-base N content, illustrated in Figure 3.4. When a sequencer cannot determine a base with enough confidence, it typically replaces it with an 'N' instead of assigning a standard base. Therefore, the smaller the N content, the better the sequencing experiment has

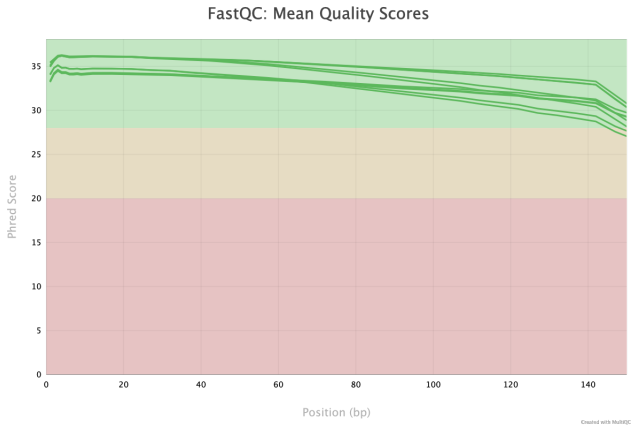


Figure 3.3: Raw reads' per-base mean quality scores.

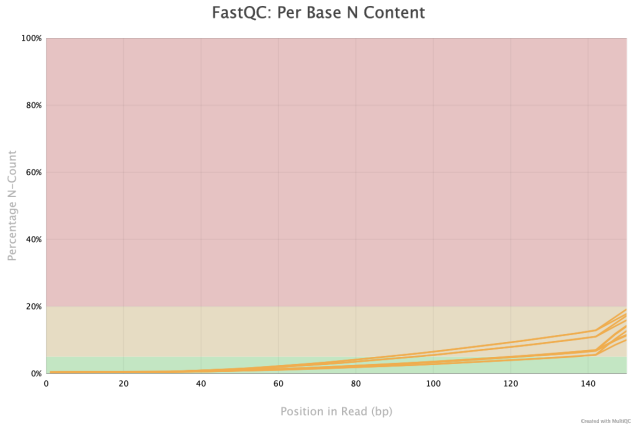


Figure 3.4: Raw reads' per-base N content.

performed. Due to already mentioned factors, a reduction in base call accuracy towards reads' final ends is typically observed, resulting in higher percentage of N content. However, this particular bases' character has a double meaning since 'N' is also used to substitute adapter sequences during demultiplexing sequencing phase. Even if adapters removal is usually performed during data preprocessing, some residue could persist at the read's ends. Also N content graph is colored-coded, in order to provide an idea about overall quality at a glance. As it can be seen, going towards final base pairs, N content percentage increase, almost reaching plot's red portion.

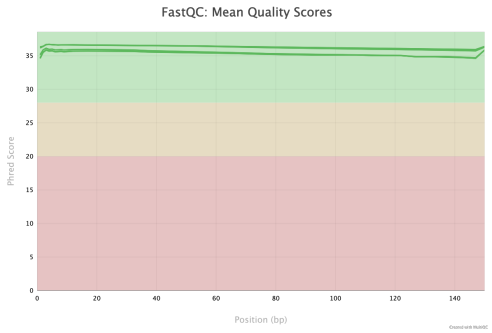
After defining a starting quality baseline, another report has been generated after running both custom-built and automated pipeline on the same test samples, to observe the impact on above mentioned quality statistics. Both pipelines incorporate *Trimmomatic*, with slightly different parameters, detailed in Table 3.1. Comaprison between mean quality scores and per-base N content achieved by each pipeline trimming step are reported in Figure 3.5.

Parameter	Custom-built pipeline	Automated pipeline
SLIDING WINDOW	4:28	4:20
LEADING	20	—
TRAILING	20	—
MINLEN	50	75

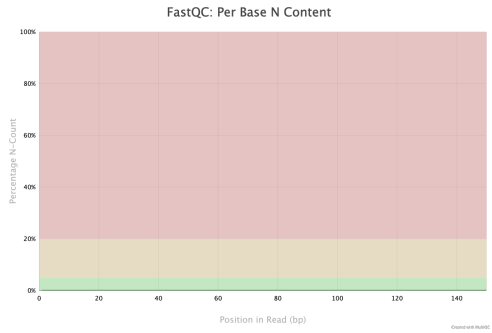
Table 3.1: Trimmomatic parameters comaprison between custom-built and automated pipeline.

Qualitatively, both pipelines have achieved an excellent result in retaining only reads with an overall good mean quality across all reads length and in significantly reducing the N content, since all plotted lines are now entirely contained in the green area, with excellent mean quality scores even in final positions and an approximately null N content. It can be observed that, although the difference is very small, the custom-built pipeline seems to achieve better mean quality scores at final base positions.

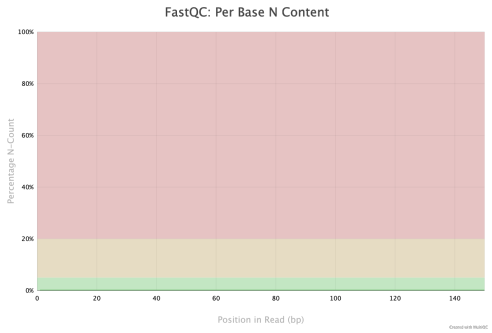
(a) Mean Quality Scores (Custom-built)



(b) Mean Quality Scores (Automated)



(c) N Content (Custom-built)



(d) N Content (Automated)

Figure 3.5: Comparison of mean quality scores and N content between custom-built and automated preprocessing pipelines.

Beyond quality pre and post trimming, the other benchmark considered to evaluate pipelines' performance was the number of reads survived after each intermediate step. Therefore, the exact numbers of reads survived after each

step of both pipelines have been analyzed and reported in Figure **??**, along with percentage in respect to raw starting number. As specified in previous section, only reads *surviving in-pairs (SIP)* have been considered across all this research work. This means that for each read discarded (e.g., due to poor quality), even if its matching one in the opposite verse was good enough, both have been be no longer considered. As a result, number of reads after each step will be identical in both forward and reverse. The only step after which the number of reads per verse might differ is the TRF step. A *tandem repeat* in DNA is two (or more) adjacent, approximate copies of a pattern of nucleotides. TRF is able to locate and report tandem repeats in DNA in order to allow filtering and removing reads that contains them. Their removal is important for bias reduction, because repeated regions could interfere with subsequent alignment accuracy (i.e., tandem repeats are among the human genetic sequences that may not be represented in the human reference because they are source of extreme variability). As it can be assumed in Figure 3.2, it's included only in the automated pipeline. Since the number of reads survived after TRF only slightly differs across verses ($< 0.2\%$), only number of forward reads is reported. In Figure 3.6a and 3.6b, percentages contained in Table 3.2 have been plotted using histograms, allowing for an immediate visual comparison.

					SIP READS					
	SAMPLE ID	TYPE	RAW READS		AFTER TRIMMING		AFTER TRF		AFTER DECONTAMINATION	
			absolute	%	*absolute*	%	*absolute*	%	*absolute*	%
CUSTOM BUILT PIPELINE	118500	S	53891310	100	31596746	58.6	*SKIPPED*		8195453	15.2
	118501	S	60795378	100	35504023	58.4			9543500	15.7
	118502	F	56388336	100	29483969	52.3			29436578	52.2
	118503	F	64344308	100	34452454	53.5			34404991	53.5
	118504	F	46736138	100	25000517	53.5			24969149	53.4
AUTOMATED PIPELINE	118500	S	53891310	100	43009245	79.8	40835629	75.8	4002466	7.4
	118501	S	60795378	100	47921045	78.8	45520491	74.9	5506130	9.1
	118502	F	56388336	100	43853231	77.8	43580623	77.3	43335329	76.9
	118503	F	64344308	100	50713557	78.8	50396429	78.3	50105012	77.9
	118504	F	46736138	100	36983313	79.1	36762005	78.7	36555429	78.2

Table 3.2: SIP reads count after each step of both custom-built and automated pipelines.

Both numbers in Table 3.2 and bars of histograms plot in Figures 3.6a and 3.6b suggest two different approaches in filtering operations performed by these workflows. Custom-built pipeline seems to be more aggressive in quality trimming phase and more conservative during the alignment performed for decontamination. On the other hand, KneadData seems to apply a milder trimming in favor of a more subsequent sensitive alignment. It's interesting to

note how the sample type seems to influence the final number of reads available: custom built workflow allows to obtain almost twice the number of saliva reads obtained with the automated one ($\sim$ 8% vs 15%). On the contrary, for feces samples KneadData saves approximately 25% more reads when compared to custom one ($\sim$ 50% vs 75%). In order to test how this gap in reads number impacts the resolution of subsequent taxonomic profiling and functional annotation, *MetaPhlAn* and *HUMAnN* have been run on both of sets of preprocessed reads. The number of strains identified by *MetaPhlAn* and the number of gene families and pathways detected by *HUMANnN* have been compared for each workflow and results are reported below in Tables 3.3, 3.4 and in Figure 3.7.

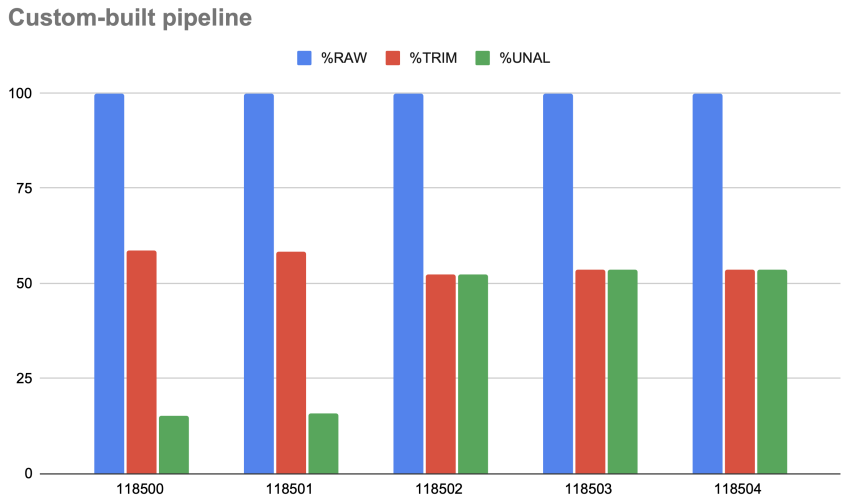
As it can be seen, *KneadData* preprocessing generally allows *MetaPhlAn* and *HUMAnN* to find more features. Indeed, as reported in Table 3.4, a positive increment in the number of strains, gene families and pathways have been find across all samples types, even for saliva samples where the available starting number of reads was smaller compared to custom-built. Difference in quality achieved do not justify the read loss in feces sample. Also, loss of saliva reads seems not to impact features identification. This highlights the validity of the automated workflow, clearly pointing out at a better trade-off implemented by *KneadData*. For this reason, this has been the preprocessing workflow of choice that has been implemented in the following steps.

	SAMPLE ID	TYPE	NUMBER OF		
			STRAINS	GENE FAMILIES	PATHWAYS
CUSTOM BUILT PIPELINE	118500	S	191	74871	267
	118501	S	211	100621	289
	118502	F	228	256638	290
	118503	F	227	263547	301
	118504	F	227	252467	294
AUTOMATED PIPELINE	118500	S	210	93093	271
	118501	S	224	120018	293
	118502	F	229	272104	300
	118503	F	231	277412	314
	118504	F	234	269716	309

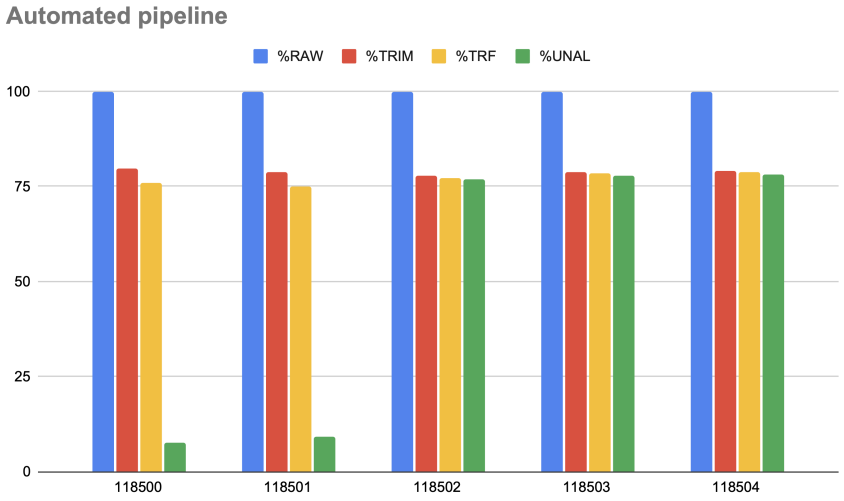
Table 3.3: Number of annotated features after each pipeline run.

	SAMPLE ID	TYPE	Δ OF					
			STRAINS		GENE FAMILIES		PATHWAYS	
			absolute	%	absolute	%	absolute	%
Δ#Features [Automated – Custom]	118500	S	19	+9,9	18222	+24,3	4	+1,5
	118501	S	13	+6,2	19397	+19,3	4	+1,4
	118502	F	1	+0,4	15466	+6,0	10	+3,4
	118503	F	4	+1,8	12365	+5,3	13	+4,3
	118504	F	7	+3,1	17249	+6,8	15	+5,1

Table 3.4: Difference in number of features annotated between automated and custom-built pipeline.

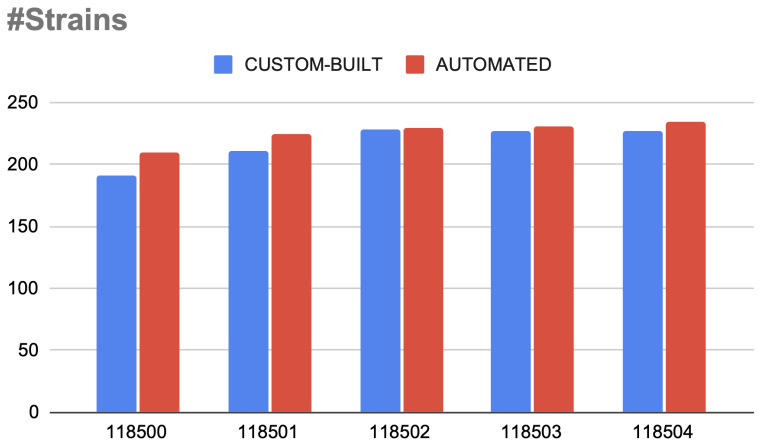


(a) Reads count after each step of the custom-built pipeline.

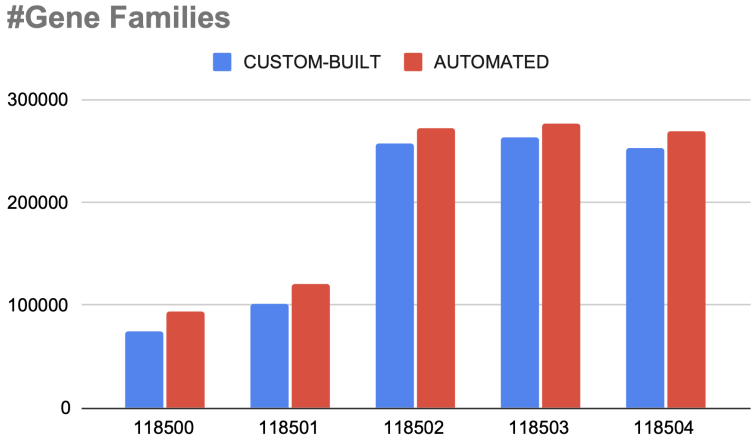


(b) Reads count after each step of the automated pipeline.

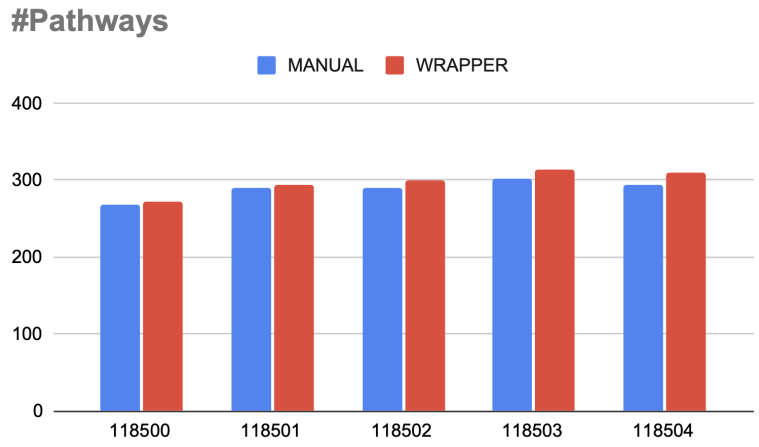
Figure 3.6: Histograms of SIP reads count after each step of both pipelines.



(a) Number of strains.



(b) Number of gene families.



(c) Number of pathways.

Figure 3.7: Comparison between the number of features annotated by each pipeline.

3.2 BioDonut 1.0

3.2.1 What is BioDonut?

BioDonut is a simple pipeline born from the need of a reliable workflow that addresses shotgun metagenomic paired-ends data coming from human microbiota studies. More specifically, BioDonut addresses analysis of microbiota communities contained in biological samples of feces and saliva, where the aim is searching for differences in microbiome features between two distinct groups (e.g., conditions vs control). It has been originally created for analyses in the context of Eosinophilic Esophagitis (EoE) and then generalized to be useful even in multiple studies with the same design, promoting standardization and trying to ease study comparison.

The name BioDonut has been chosen for two main reasons. First, it wants to highlight the circular nature of the workflow implemented: a single tool that goes from very first preprocessing phases like quality filtering and decontamination, till some first-line downstream analyses useful to start exploring the characteristics and potential trends present within the dataset under observation. Second, it wants to be an ironic way to reference the *bioBakery* platform, since BioDonut heavily relies on most of its tools (i.e., *KneadData*, *MetaPhlAn* and *HUMAnN*), addressing their specific output files' structure.

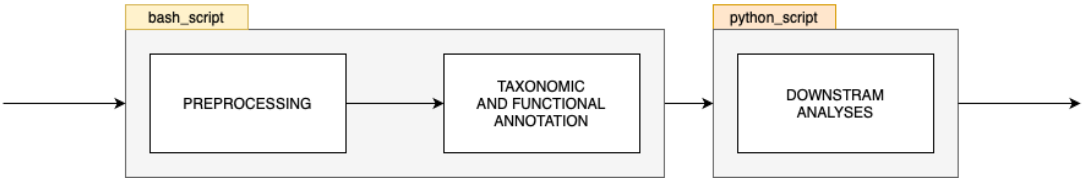


Figure 3.8: Biodonut's blocks schematic representation.

BioDonut is composed by 3 fundamental blocks, as represented in Figure 3.8: preprocessing, taxonomic/functional annotation and downstream analyses. First two blocks are implemented in a single bash script. Then, bash script output is used as starting point for a python script, that implements the last block. Following tests and images are coming from the execution of BioDonut on samples coming from both first and second pilot studies (i.e., 16 samples in total) in order to consider as many data as possible since a real case scenario usually involves a huge number of samples to be tested and analyzed. All scripts are publicly available at github.com/strmrc/BioDonut.

3.2.2 Preprocessing block

Along with other mandatory options specified in the repository documentation (i.e., various tools paths, reference databases paths, etc.), raw sequences data must be provided in input in FASTQ format. File names must follow a specific pattern:

SampleID_SampleTYPE_R1,2.fastq[.gz]

- *SampleID* can be a free numeric code meant for uniquely identifying a specific sample, that can be chosen by the pipeline user, that must be consistent across pipeline inputs. It will be used to link reads to metadata and as label in most of the pipeline output files;
- *SampleTYPE* must be a single letter meant for identifying the nature of the sample. Since this pipeline is born to manage saliva and feces samples, this part of the name must contain either the letter 'F' or 'S';
- *R1,2* specifies reads verse; it must be composed by the letter R followed by number 1 for forward reads, or 2 for reverse reads;
- FASTQ format extension have to be specified at the end (*.fastq*);
- Compressed files in *.gz* format are accepted, but extension must be present at the end.

An example of a file correctly named is *118500_S_R1.fastq* or *118500_S_R1.fastq.gz* in case of a compressed archive. In this preprocessing phase, the tool of choice that has been implemented is *KneadData*, since it has been proved to offer a good tradeoff between quality filtering and decontamination. Workflow reported in the previous paragraph for testing phase has remained largely unchanged, with the addition of little intermediate corrections. For example, a header-correction step has been added since *bowtie* as run by KneadData does not manage correctly paired-ends information as provided in files coming from current Illumina sequencing platforms. Somehow, identifiers contained in reads' header are not correctly recognized and all the reads are then classified as "unmatched", losing verse correspondence. As suggested from users and bioBakery lab members in the bioBakery forum [89], manually manipulating the header is possible to restore the correct pairing. While waiting for an official patch provided by bioBakery developers, this

simple correction procedure has been integrated in BioDonut, and it can be either enabled or disabled through the dedicated flag.

This pipeline block provides some intermediate outputs, in order to allow for some flexibility and personalization of personal user workflow. Indeed, it is possible to access to fastq files decompressed and header-corrected, as long as whole intermediate KneadData outputs (that includes both surviving and non-surviving in pairs reads coming from trimming, TRF and bowties procedures).

3.2.3 Taxonomic/functional annotation block

The same bash script carries out also the taxonomic and functional annotation phase. MetaPhlAn uses the marker gene database provided in input to annotate species and strains of microbial communities, merging all single profiles in a single text file. Also in this case, intermediate single taxonomic profiles are stored and freely accessible for the user, along with intermediate bowtie files useful for repeating the annotation without having to completely rerun the time-consuming alignment step. It is worth noting that, if used, BioDonut has been designed with the possibility to ignore the spike-in microbial community during taxonomic annotation. In microbiota studies, a "spike-in" refers to a known quantity of DNA (or RNA) coming from an exogenous organism (i.e., an organism that is external to the biological sample object of the analysis). Its main purpose is to serve as an internal control or reference measure for normalization, quality control and absolute quantification purposes. Due to preliminary and exploratory nature of this research, these procedures have not been addressed and ZymoBIOMICS^{TM} spike-in included in each sample have been removed to avoid interference with subsequent analyses. Indeed, providing a text file, within the specific flag, containing the names of marker gene that are related to the specific spike-in, they will be ignored by MetaPhlAn annotation process, preserving the true count of species and strains found, along with their relative abundances. In this way, the spike-in removal made at taxonomic profiling level will reflects also in functional annotation, avoiding annotate for gene families and pathways related to exogenous organisms, because HUMAnN relies on the MetaPhlAn output.

Before running HUMAnN, BioDonut implements a read-joining step, because paired ends information is not considered by this tool. Therefore, for each sample, both forward and reverse reads are merged into a single file (e.g., *118500_S_R1.fastq* and *118500_S_R2.fastq* become *118500.fastq*). Once the

merge is completed, HUMAnN can start its processing following the setting provided by the user. Indeed, it is possible to provide both a nucleotide database and a protein database or only one of them. BioDonut is able to recognize what's provided in input and, if one the required databases is not specified, it sets HUMAnN to bypass the nucleotide or the translated search. Nucleotide search relies on the use of *ChocoPhlAn* database to align reads for identifying microbial genes at DNA level. Instead, translated search is made at protein level. This is a deeper level of analysis, useful when nucleotide search fails to map certain reads to a known pangenome. Since it requires the translation of nucleotides in proteins sequences for their alignment with the database provided (e.g., UniRef), it is more time consuming but it offers the possibility to capture functions that may not be detectable at nucleotide level. Being able to bypass part of the analysis can be useful during tests, or to avoid redundant steps if some part of it has already been done, saving time and computational resources. As for MetaPhlAn outputs, also HUMAnN generates annotation for each sample individually and then it merges them all into a single file. Even if the pipeline relies on the merged version, individual annotations are accessible to the user. Both gene families and pathways abundances are provided in two different versions. HUMAnN natively calculates gene and pathway abundances using *RPKs (reads per kilobase)*, which adjusts for gene length but does not account for sequencing depth in the sample. However, it is possible to normalize abundances into *copies per million (CPM)*, ore more correctly CoPM, which is a *total sum scaling (TSS)* normalization approach, meaning that abundances in each sample sum to a total of 1 million. This is analogous to the *TPM (transcripts per million)* used in RNA-seq and it is different from the counts per million normalization that accounts only for sequencing depth and not for gene length, that unfortunately have the same acronym. To avoid confusion, the abbreviation *CoPM (copies per million)* should be used. BioDonut makes available to the user both versions, the original and the CoPM normalized one. Since CoPM normalization accounts for both sequencing depth and gene length, is highly recommended instead of RPKs for samples comparison and to perform downstream analyses [78]. Therefore, in the following steps, BioDonut relies on CoPM normalized functional annotation.

3.2.4 Downstream analyses block

The third block of BioDonut aims to perform some first-line downstream analyses in order to start exploring dataset characteristics and to start searching for differences and trends upon the composition of the microbiome across different groups of individuals. Analyses carried out consist in a general basic framework that repeats across multiple subsets of the original one, each one analyzed at every level of detail that shotgun metagenomic allows to obtain. More specifically, the original dataset is supposed to be made of multiple samples, from both feces and saliva, coming from two distinct group of patients (i.e., patients and healthy controls). From this dataset, two subsets are then created grouping samples by their type. This means that, available for analyses, there will be a *whole mix dataset*, along with an *only-saliva* and an *only-feces* datasets. Dataset splitting operation is schematized in Figure 3.9 . Creation of sub-datasets by sample type aims to avoid obscuring trends that are group-specific (and therefore possibly related to the condition object of the study) by inherent differences between samples. Indeed, saliva and feces represent two distinct environments with different microbial compositions. This extreme and already-expected variation could mask the more subtle or specific differences related to disease within a single sample type. Having separate datasets should ease the observation of differences that can emerge between patients and healthy controls, without being confused by the variability related to difference between saliva and feces. BioDonut then starts to analyze each dataset individually, exploring them at all annotation levels, as visible in Figure 3.9. It starts analyzing taxonomic composition at both species and strains levels, to switch immediately after at functional level, considering gene families and pathways.

Base analysis framework

The base analysis framework that repeats across all levels, mainly considers samples *alpha* and *beta diversity*. Indeed, for each sample multiple alpha diversity metrics are computed upon the indication provided by the user through the dedicated option. Since their computation relies on the use of *scikit-bio.diversity* package, desired metrics have to be chosen among the ones made available by the package. In addition to them, a simple features count has been added by default. Once alpha diversity metrics have been obtained, they are added to a data frame where also all the subject metadata is reported.

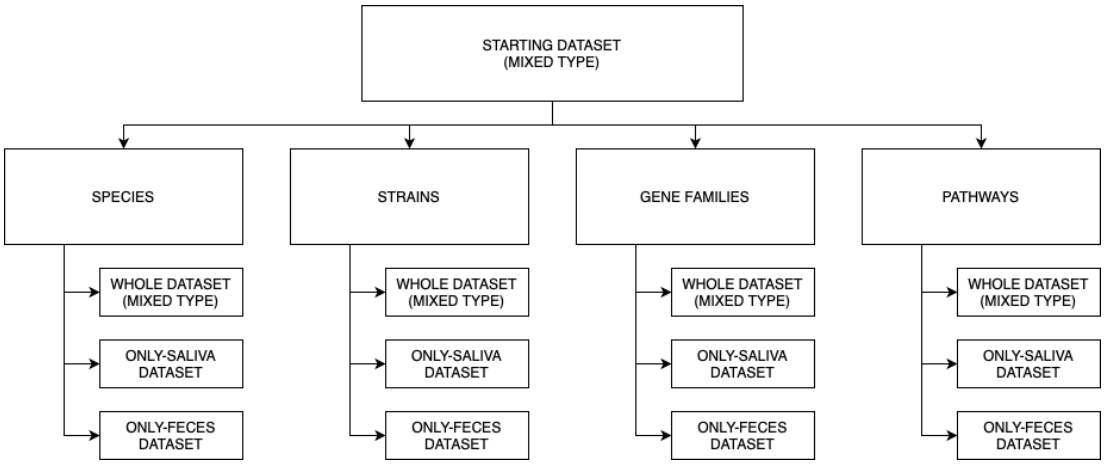
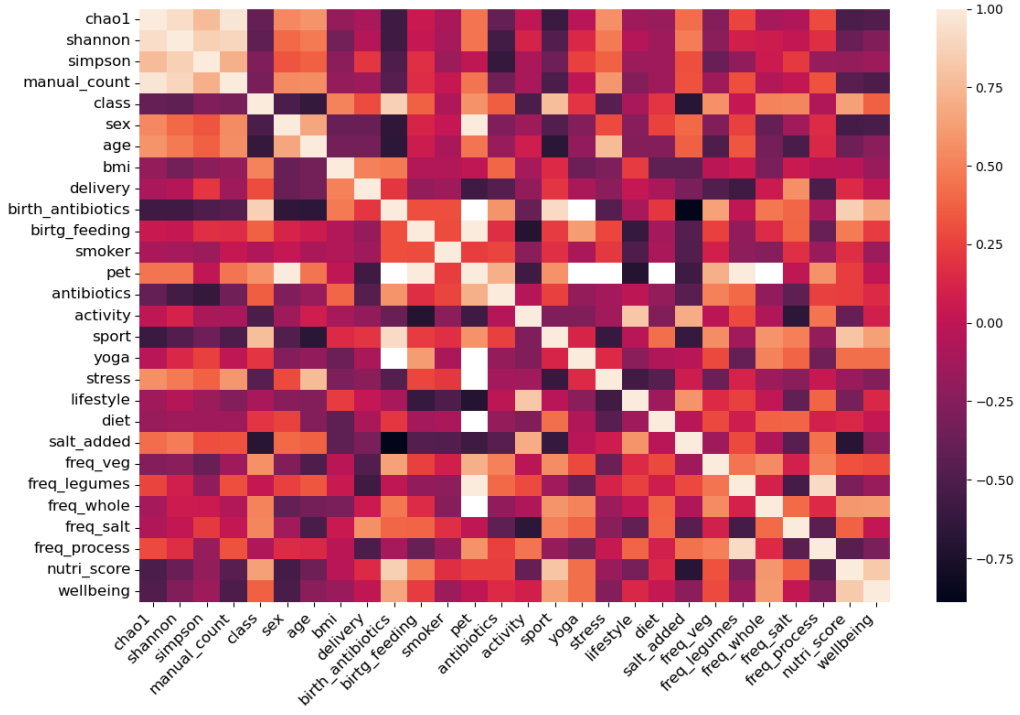


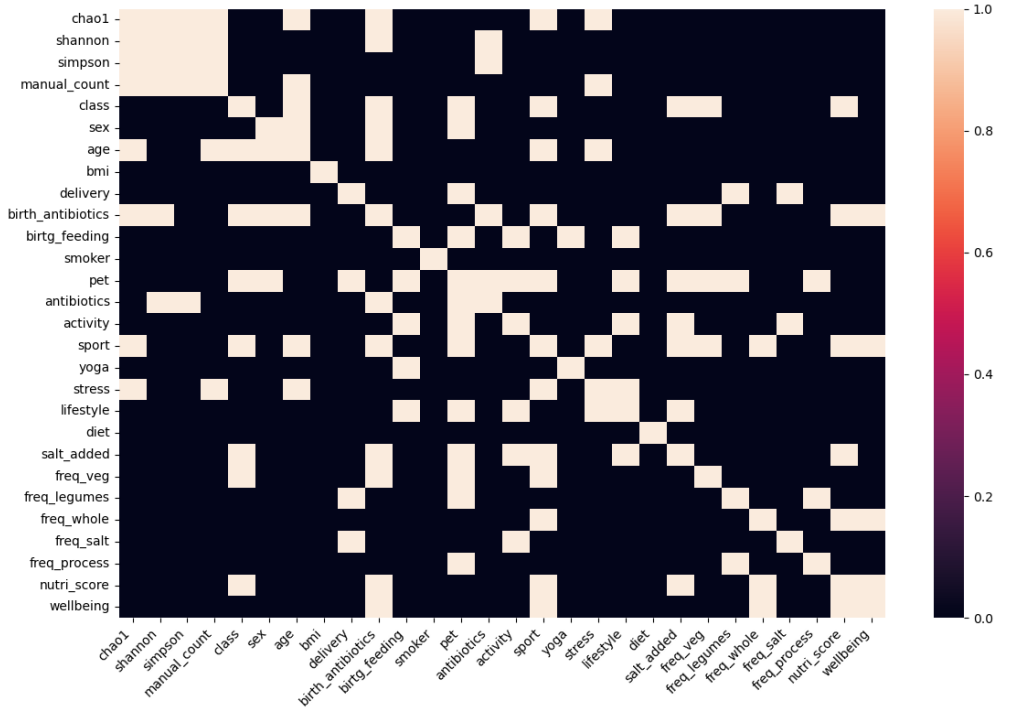
Figure 3.9: Schematic representation of Biodonut's dataset splitting.

Based on this specific data frame, *Spearman's correlation* matrix is computed and made available to the user as an *.xlsx* file, useful for further analyses. The same matrix is also used for plotting a heatmap, useful for visually searching high values of correlation that could represent the presence of a potential trend. Depending on the number of metadata variables provided, the heatmap could result in a very packed and crowded plot, where it's difficult to distinguish across all correlation values. Therefore, a quick check heatmap is provided along with the classic one. The quick check heatmap plot only two levels, identifying pairs of variables for which the correlation level is above a certain threshold (that can be eventually personalized modifying the code). An example of these heatmaps can be observed in Figure 3.10.

Beta diversity aims to analyze samples dissimilarity, quantifying differences overall taxonomic composition and functional profile between two samples. One of the most used metrics is the *Bray-Curtis dissimilarity*. BioDonut use a specific function of *scikit-bio.diversity* package for its calculation. The obtained dissimilarity matrix is then visualized thanks to *Principal Coordinates Analysis (PCoA)*, a multidimensional scaling technique used to visualize and explore relationships between samples in a dataset based on a distance or dissimilarity matrix. It aims to represent the samples in lower-dimensional space, usually 2D or 3D, where the distances between points approximate the dissimilarities between samples. Therefore, in this particular case, similar samples in terms of taxonomic profile and functional annotation should be visually clustering together in PCoA plots. BioDonuts provides in output multiple plots, each one considering a different variable, using a dedicated color to represent points according to its value. This should ease visual cluster recognition. An example



(a) Spearman's correlation heatmap.



(b) Spearman's correlation quick check heatmap.

Figure 3.10: Examples of heatmaps provided by BioDonut.

is showed in Figure 3.11, where we can see a dissimilarity plot obtained on whole dataset of mix samples type, considering strains profile. The considered variable for dot coloring was the sample type. As expected, saliva and feces samples are distant ones from the others, forming clearly visible colored clusters.

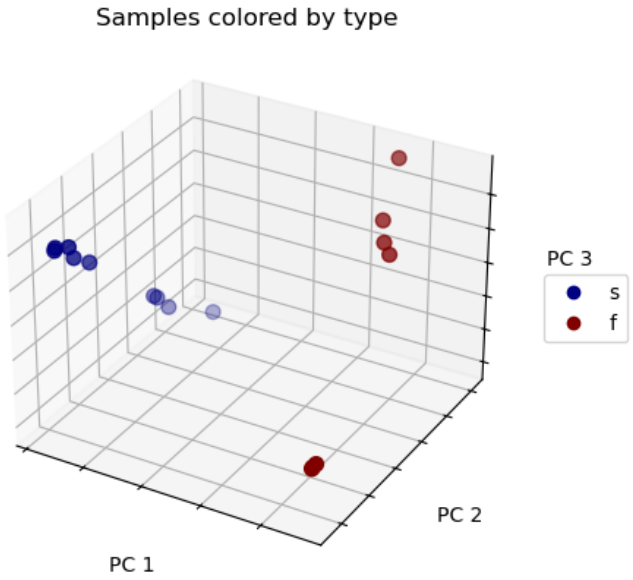
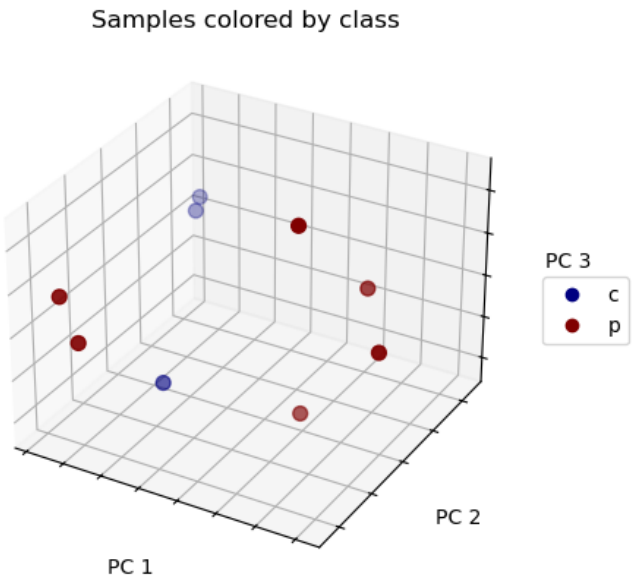
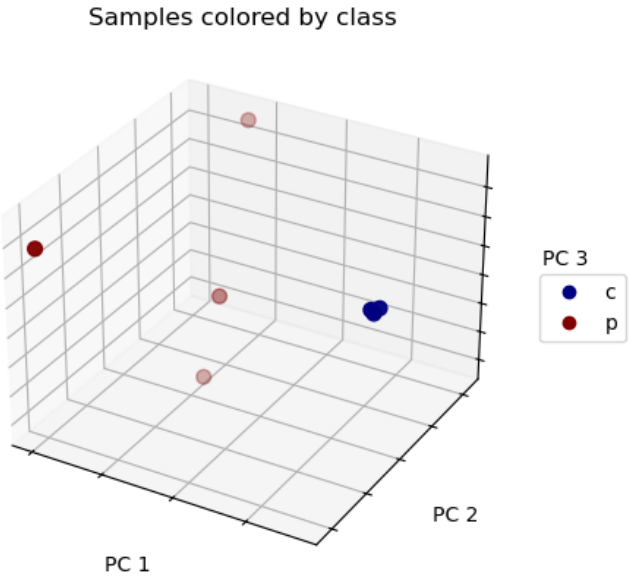


Figure 3.11: PCoA plot colored by samples type.

This is a clear example where difference masking phenomenon has occurred. In Figure 3.11, we are considering *mix dataset*, composed by both feces and saliva. The huge distance that we can observe between dots is due to the extremely different microbiota strains composition, that is the direct reflection of the different samples' origin. This distance is making difficult for smaller distances within cluster to be clearly represented. In Figure 3.12, examples of PCoA plots coming from *only-saliva* and *only-feces* dataset are presented. In these plots smaller distances, that can be caused by possible different reasons, are better visible. In these cases, coloring dots by sample type wouldn't make sense, so by-type coloring is not performed. All the other PCoA variables provided in input by the user are considered. In these examples, dots are coloured by class.



(a) PCoA plot colored by class (Saliva).



(b) PCoA plot colored by class (Feces).

Figure 3.12: Comparison of PCoA plots colored by class for saliva and feces samples.

Clearly defined clusters are typically the desired outcome of this analysis, as they help identify ways to separate different groups. When no clear separation is present, the resulting plot resembles the one shown in Figure 3.12a, where purple dots represent saliva samples from controls, and yellow dots represent those from patients. The dots do not form distinct color-based clusters,

indicating that microbiota composition does not follow a clear pattern based on class (i.e., patient/control). In contrast, Figure 3.12b shows purple dots (representing fecal samples from controls) forming a distinct cluster, indicating a high similarity in microbiota composition within the control group. This cluster is also well separated from the yellow dots (representing patients' fecal samples), suggesting that the microbiota compositions of controls and patients are different. This scenario is ideal in microbiota studies that aim to identify novel biomarkers associated with specific conditions, as it suggests a strong relationship between the fecal microbial community and the condition in question, even if further investigation would be needed to confirm this relationship. Unfortunately, this is not the case in the current example. As shown in Tables 4 and 5, the fecal samples from controls all belong to the same individual, which explains the very tight clustering. However, this still supports the validity of the clustering approach.

Searching for significant differences between classes (i.e., patients vs healthy controls) is the main goal in this type of research work. BioDonut implements the execution of *PERMANOVA*, that is a nonparametric permutation test for evaluating the presence of an overall difference between different groups. Base analysis framework implements a *PERMANOVA by condition*, searching for significant difference between patients and controls, saving the result in an output text file. Considering the limited pool of samples available for test, also a *PERMANOVA by type* has been implemented only for whole-dataset analyses, searching for statistically significant differences between samples type. This type of assessment is purely for validation purposes, as these differences are already expected. Examples of PERMANOVA output are reported below, in Figures 3.13 and 3.14, coming from the analysis of whole mix dataset at strains level.

```
method name                PERMANOVA (by condition)
test statistic name         pseudo-F
sample size                       16
number of groups                   2
test statistic              1.437737
p-value                        0.202
number of permutations           999
Name: PERMANOVA results, dtype: object
```

Figure 3.13: Example output of PERMANOVA by condition.

```
method name                PERMANOVA (by type)
test statistic name         pseudo-F
sample size                       16
number of groups                   2
test statistic             12.345843
p-value                        0.001
number of permutations           999
Name: PERMANOVA results, dtype: object
```

Figure 3.14: Example output of PERMANOVA by type.

3.2.5 Additional analyses for single type datasets

In addition to the above mentioned analyses, the base framework adds some additional exploratory procedures that are conducted only for single type datasets (i.e., *only-saliva* and *only-feces* datasets). Previous analyses focused on overall features' number and composition. In order to enter in a deeper level of detail, investigating the presence or absence of specific features, a *Venn diagram* is provided.

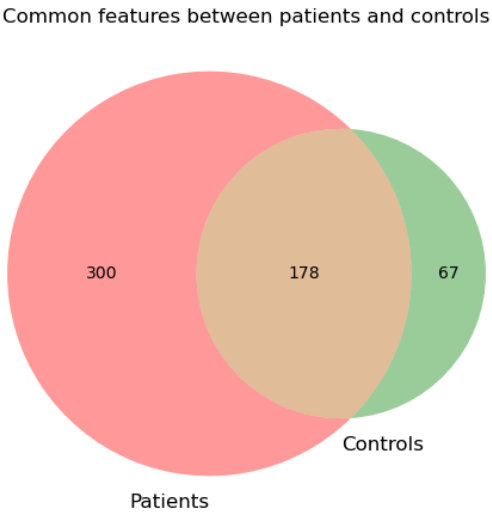


Figure 3.15: Example of exploratory Venn's diagram.

Figure 3.15 is an example of Venn diagram computed for strains of only-feces dataset. It gives the user information about quantity of both shared and exclusive features among subjects' classes. Also, a name list of exclusive features of both classes is provided in output as *.txt (tab-delimited text file)*, readily available for the user. An example of the list provided containing the name of strains common to all patients feces samples is represented in Figure 3.16. This type

of information can be useful to extrapolate biological meaning, for example, of a particular strain or pathway that is found to be common only to one class of subjects and absent in the other.

```
strain  132663  132664  132666  132669
t__SGB14853 0.0105  0.00151 0.00358 0.05202
t__SGB6140  0.00369 0.06938 0.03304 0.08228
...
t__SGB14807 0.00263 0.0187  0.00053 0.03237
t__SGB5089  0.00236 0.05599 0.02723 0.28734
```

Figure 3.16: Example output of common strains found for patients' feces samples.

This pipeline block is carried out by a Python script, that needs multiple input files in order to properly function. Obviously, text files containing taxonomic profile and functional annotation of both genes and pathways are needed. If this third block is executed after running first two blocks of BioDonut, there's a dedicated option to specify the output folder produced by the bash script used. In this way, since the structure of directory tree is known, the python script is able to automatically retrieve all needed files coming from the previous elaboration. If preprocessing and taxonomic/functional profiling have been carried out without relying on BioDonut, there's the possibility to specify each file path individually, allowing for flexibility and modularity. For example, the user can decide to personalize preliminary steps and relying on BioDonuts only for downstream analyses. However, output files compatibility should be assessed and is not granted. Besides all annotations required, there also other required fundamental inputs. In addition to list of alpha metrics and variable to be considered during PCoA plots coloring process, also *subjects information* must be provided. Indeed, both controls and patients' data need to be specified in a *tab delimited text file (.txt)*, following a specific structure. Along with subjects' metadata, another additional file that contains the information required to link each *Sample ID* to its *Patient ID* must be provided. Both files' structures are specified in the readme file included in the BioDonut repository.

3.3 Propensity Score Matching

Propensity Score Matching (PSM) is a *quasi-experimental* technique used to design observational studies in a way that simulates aspects of a randomized controlled trial [98], trying to reduce the risk of confounding. The idea is to ensure that groups being compared (i.e., case and controls) are similar in terms of their characteristics at the start, using a balancing score. This technique assumes that subjects with the same score have a similar baseline characteristic, making it easier to assess the effect of a specific variable without bias from the others [98]. The propensity score exists in both randomized experiments and observational studies. However, in randomized trials, the actual propensity score is defined by the study's design and is known. In contrast, observational studies usually do not have a known true score, but it can be estimated using the collected data.

Among the various existent PSM methods, *one-to-one matching* is the most frequently employed [98] and the one also implemented in this research work. Indeed, in the BioDonut repository there is an additional python script available, that implements PSM technique designed for selecting controls among a pool of volunteers in order to be subsequently matched with recruited patients according to their baseline characteristics, allowing for retrospective design of case-control groups. The estimation of each subject's propensity score is carried out by a *binary logistic regression model*, implemented through the *scikit-learn* python package. *Logistic regression* is a statistical model that models the *log-odds* of an event as a linear combination of one or more independent variables. In more simple terms, the idea is to summarize all subject relevant information and possible confounding factors into a single number, ranging from 0 to 1. This number is then used to perform one-to-one matching, meaning that each patient will have its own matched control. In this way, the resulting casecontrol group should have on average the same baseline characteristics.

The resulting script is made up of 3 main sections:

1. The first one is responsible for the *data import* and *conversion*;
2. The second one builds the model and computes *propensity scores*;
3. The third one performs a pairwise distance matrix based *matching*.

In order to start the elaboration, the *first section* captures the subjects' information provided in input, previously retrieved through questionnaires. These are essentially the information that have been already used for downstream analyses in the third block of BioDonuts pipeline. Data must be modeled in order to be composed only by binary, categorial or continuous variables, since these are the only data types that logistic regression model is able to manage and describe. This means that unstructured information, retrieved through open and/or multiple questions in the questionnaire, must undergo a proprietary preprocessing and modeling phase. An example of how the information dataset should be composed is available in the BioDonut repository, in the PSM section. Along with both patients and healthy controls information, also the variables' name to be considered for propensity score calculation must be specified, since they can be a subset of all the variables available in the provided dataset. Features selection is an important step, that will be discussed in the following chapter. This block also implements a section where data, if needed, is correctly converted.

The *second section* builds the model and creates *dummy variables* for each categorical variable that is considered for the PSM calculation. *Dummy variables* are needed since logistic regression model is able to natively manage only continuous and binary variables. Dummies creation transforms each categorical variable in an additional separate binary variable. A *k-categories* variable will always need *k-1 dummy variables* to be defined. For example, a categorical variable that accounts for smoking history of a subject, that can have 3 different categories (i.e., non-smoker, ex-smoker and actual smoker), will need 2 separate new binary variables. However, this script is able to automatically recognize a categorical variable and to create needed dummies without any action required by the user. Once the model is built, propensity scores are then computed and stored in a data frame.

The *third section* is responsible of matching operations. As already mentioned, this implementation of PSM aims to create case control groups retrospectively, prioritizing the inclusion of each patient. Since EoE is a rare condition, all eligible subjects should be considered aiming to retrieve as many subjects as possible. Therefore, the selection is performed among the pool of controls, selecting those that achieve the best matches with enrolled patients. Best matches are selected thanks to a *pairwise distance matrix*: for each patient, the difference between its scores and the ones of all the candidate

controls are performed. An in-house algorithm then sorts the matrix to select the matches that scored the minimum distances. The algorithm is designed in order to avoid *reinsertions*: this means that each control can be included in the study only once. If a control matches with more than one patient, the best match is selected and for the other patients the second-best match will be chosen. Matching results are then exported to a *.csv file* that the user can access, containing for each *Patient ID* the associated *Healthy Control ID*, along with the value of distance and a briefly recap of the considered variables values.

In addition to matching result, the algorithm provides in output a plot of the so-called *common support*. Matching operation will always be performed, regardless the quality of effective pairs computed. A successful matching needs an overlapping distribution of propensity score across the patients group and the healthy controls group. The overlap section between the two propensity scores distribution is known as *common support*. If a common support is present, it is possible to effectively recreate a similar condition that we have with groups created in RCTs. In Figure 3.17, an example of the plot returned by the script is shown. As it can be seen, two different versions of propensity score are presented (i.e., l*ogit propensity* and *propensity*) and we can clearly see an overlapping portion between the two distributions. The greater the common support, the better the matching. If the common support is poorly present, the matching operation will not be qualitatively successful.

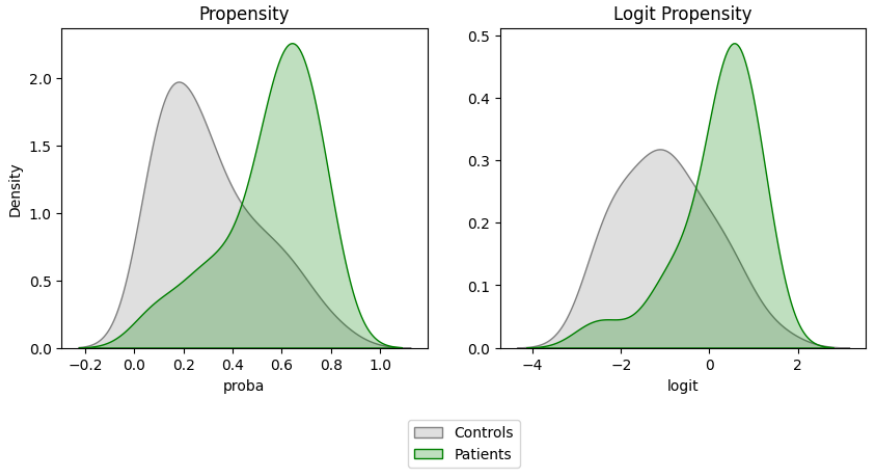


Figure 3.17: Example output of propensity scores distributions and common support plots.

Chapter 4

Discussion

This research work aims at exploring potential existent relationship between EoE and the human microbiome characteristics and composition. An ambitious goal, made possible by recent advances in sequencing techniques and by the availability of powerful bioinformatic tools. However, this relatively new field of study still presents some challenges, as the lack of standardization and the absence of clearly defined protocols, especially when it comes to sequencing data processing. Results and drawn conclusions heavily depend upon the specific designed processing flow, therefore study comparability is seriously undermined. One of the main achievements of this work is the development of *BioDonut*: a simple but powerful pipeline, able to cover all the main steps of preprocessing required for metagenomic studies, and able to produce also first-line insights due to implementation of some general downstream analyses. It has been designed to be run by two modular scripts, aiming to be accessible also to users without specific bioinformatic skills. Although it was specifically designed to investigate Eosinophilic Esophagitis patients' microbiome, it has been developed in order to be as generalizable as possible, capable of addressing almost all types of shotguns metagenomic studies that involve feces and saliva samples.

Preprocessing pipelines comparison have proved the reliability of the workflow implemented by *KneadData*. Sequencing depth is a critical factor impacting microbial communities' studies. Reads quality is important, but also is the number of available reads for subsequent analyses. In the name of a perfect quality achievement, an aggressive trimming could potentially discard more reads than necessary, sacrificing resolution and accuracy of the following taxonomic profiling and functional annotation. A reasonable trade-off between quantity and quality must be reached to benefit from shotgun metagenomic

advantages. The custom-built pipeline was intentionally set to a high level of aggressiveness in order to evaluate the case where reads quality is weighted more than reads number. A possible further analysis should focus on the test of a various range of *trimmomatic* parameters, considering a boarder spectrum of trimming aggressiveness. The same parameters should be also tested within the use of *KenadData*, in order to determine the contribution of TRF additional step and different bowtie sensitivity settings, since tandem repeats are known to be source of noise that hinders the alignment performed for decontamination. In this analysis, TRF has been observed to be responsible of a relatively small percentage of reads discarded ($\sim$ 4-5% in saliva and $\sim$ 0.5% in feces samples), but further in-depth evaluation should be conducted. This pipeline addresses data coming from paired-end sequencing, therefore as input are always required both forward and reverse files. Tools on which BioDonut relies are able to manage paired ends reads, except for *HUMAnN*, that requires a preliminary merging step. When using paired-end reads, both ends of each DNA fragment are sequenced and additional complementary information (e.g., orientation and distance between fragments) are retrieved. This typically improves accuracy, contributes to noise reduction and results in a better mapping. If tools are used in single-end mode, ideally merging all reads into a single file, they process all reads independently, potentially lowering the annotation quality due to loss of this additional information. On the other hand, running single-end mode would allow to consider all reads, not only the ones *surviving in-pairs (SIP)*. Indeed, single-end mode would allow for considering even reads discarded because their paired read has been judged qualitatively insufficient. Considering also non-SIP reads could significantly increase the number of reads available, virtually increasing the annotation resolution. Further analyses should assess the contribution of paired-ends vs single-end information, and the impact on the number of features being annotated. A possible intermediate approach could be the introduction of non-SIP reads only in the merging phase before running functional annotation. Considering this specific step performed by *HUMAnN*, also the impact of translated search on the recognition of gene families and pathways has been left unaddressed and should be object of further investigation to determine whether the additional time and computational resources are worth the potential gain in resolution. In the work of Franzosa et al., published in 2018, the *runtime variability* of HUMAnN was already stated to "vary inversely with the (a priori unknown) fraction of sample reads

explained before the translated search tier" [99]. In the research work where *HUMAnN 3* is presented [90], its runtime is compared with performances of a different competing tool. It seems that by bypassing the translated search step, HUMAnN 3 is still able to annotate the majority of the reads annotated by the competing tool (70.9 ± 9.6% per sample) but with a 9x speed up process, although this is generally only appropriate for communities known to be well-covered by related reference sequences.

Indeed, it is worth remembering that BioDonut is implementing a completely *reference-based analysis*. Reference-based tools, on which it relies on, are faster, computationally efficient, and provide consistent taxonomic and functional profiles for well-characterized communities. However, they completely rely on database availability, being limited to known species and may introducing bias toward well-studied organisms. Further versions of BioDonut should consider the implementation of *de novo assembly techniques*, which enables the discovery of novel organisms and strain-level resolution, offering a more comprehensive view of microbial diversity. It is particularly useful for exploring uncharacterized environments, but it requires significant computational resources and can result in fragmented genomes in highly complex communities. Ideally, a *hybrid approach* may combine the efficiency of reference-based tools with the exploratory power of de novo assembly.

Considering downstream analyses performed, the *base framework* that repeats across all subdatasets tries to consider trends among the overall features recognized. *Alpha* and *beta diversity* metrics are one of the first calculation being performed in microbiota studies, in order to evaluate diversity within and between samples. BioDonut allows the user for selection of desired alpha diversity metrics, and provides additional graphical outputs. Three-dimensional PCoA plots are provided in order to visually represent *Bray Curtis* dissimilarity matrix obtained. Utilizing three dimensions allows capturing more variance in the data: while the first two dimensions may explain a significant portion of the variance, the third dimension can provide additional context, potentially revealing patterns that would be missed in a two-dimensional plot. If data includes distinct clusters, visualizing them in three dimensions may help to illustrate their separation more effectively. Clusters that might seem close or overlapping in two dimensions can be more distinguishable when viewed from different angles in three dimensions, allowing for recognition of complex relationships and structures.

Despite these clear advantages, the already mentioned analyses still need to be object of some sort of validation process. When BioDonut development started, only first and second pilot study samples sequencing data were available. Therefore, tests have been conducted on a limited pool of samples, resulting in the inability to achieve any biologically meaningful results. It has been possible to obtain only already expected conclusions, implementing steps only for demonstration purposes (e.g., *PERMANOVA by sample type*). BioDonut has been laying the groundwork for the subsequent analysis of a proper full dataset of EoE subejects and matched controls. Indeed, during this research work, the pools of patients and controls were significantly expanded, with a total of 37 enrolled patients, each having provided both fecal and saliva samples. Considering healthy controls, 46 subjects were recruited, of which 22 have provided both sample types to date. All samples are waiting to be sequenced, processed and analyzed. Also, patients and controls metadata have been collected for propensity score matching purposes.

BioDonut is a useful tool especially in the context of case-control observational studies, which aims to define the importance of predictor variable in relation to the presence or absence of the disease [100]. Major known problems with this type of studies are *selection bias* and *confounding variables management* [100]. *Confounding* occurs when an external variable, referred to as confounder, influences both the effect object of examination and the outcome, creating a false or distorted association. This problem is extremely relevant when it comes to validity of final results and drawn conclusion. Usually, *Randomized Control Trials (RCTs)* is the gold standard approach in trying to avoid confounding [98], but it's rarely feasible and accessible mainly due to resources availability. Therefore, some techniques have been developed that help to reduce its impact, especially during the case control group design phase. For this research work purposes, *Propensity Score Matching (PSM)* has been implemented and fully tested with already available metadata. As already presented, *one-to-one matching* has been the method of choice implemented, allowing to obtain groups with the same number of subjects. One-to-one is an example of matching across all the available different ones. Even one-to-one matching itself can present some variations. For example, matching with and without reinsertion. In the script provided in the project repository, *reinsertion* has not been considered. This means that when a control subject is matched, it is removed from the pool of available controls for subsequent matching operations. This allows for the creation of a more diverse

and representative control group, with a more straightforward and interpretable possible comparison between patients and controls. However, reinsertion could be useful to be considered when few controls subjects are available, preventing unmatched patients. Also, there could be more than one patient that achieve a small distance with the same control. Allowing for reinsertions, overall closer matches can be achieved. Reinsertion should be carefully evaluated since it could lead to overrepresentation due to extreme use of a single subject, leading to a control group that does not reassemble the variability of a true population. Another crucial aspect of PSM techniques is *feature selection*. Since logistic regression is known to be an automated *supervised machine learning* algorithm, its performance is strictly dependent on quantity of data available. With a limited set of data, it is possible to account only for few possible confounders. A general guideline is provided by the so-called *rule of thumb* (Eq. 4.1). It states that the number of subjects in the minority class divided by the number of features (i.e., variables) considered should return a number greater than ten.

$$\frac{\#\text{subjects in the minority class}}{\#\text{variables considered (including dummy variables)}} \geq 10 \tag{4.1}$$

In this research work, more than 25 variables have been collected from questionnaires given to both patients and controls, addressing the main aspects of subjects' routine and characteristics that could be factors influencing the microbiota composition and, therefore, representing possible cofounders. *Age*, *sex*, *BMI*, *antibiotics therapy*, *level of activity*, *diet* and generally *lifestyle* are only few examples of information that have been modeled and made available for any possible analyses, including propensity score matching, trying to build case control groups where the baseline covariates are *on average* equally distributed. However, if we apply the above-mentioned rule, it is immediate to realize that to account for all these possible confounders, a lot more subjects must be enrolled. Indeed, if we perform this simple calculation, it turns out that to address 25 variables in matching operations, the number of subjects in the minority class should be at least 250, without even considering possible categorical variables that would lead to even more variables due to dummies creation. A number of subjects difficult to recruit with the effort of a single research group, considering sequencing costs per sample and rarity of Eosinophilic Esophagitis condition. For this reason, case-control group design for this research work have involved a small subset of all variables available,

composed by 4 main variables: *sex*, *age*, *BMI*, *nutri score* (i.e., 4 variables should involve at least 40 subjects in the minority class, that is approximately the number of patients enrolled in the study). Future projects should involve the joint effort of multiple research groups in working towards the creation of a single dataset of sequencing data from both patients with eosinophilic esophagitis and controls, in order to make potential discoveries and results as meaningful as possible. This context would be the ideal case scenario for the use of a customized pipeline for data processing like BioDonut, ensuring consistency and reproducibility across studies, facilitating more robust comparisons and the identification of meaningful insights from the combined data. *Shotgun metagenomics* is clearly a powerful technique, that seems to be holding the key to reveal the microbiome potential, offering a way to achieve a complete overview of a sample's microbial community. It allows obtaining both taxonomic and functional insights at high resolution, from a quantitative point of view. It's a great starting point when it comes to assess complex microbial communities like the ones composing the human microbiota. Beyond the implementation of eventual additional analyses and some refinements, next versions of BioDonut should be focusing on the implementation of *metatranscriptomics*. While shotgun metagenomics is a powerful tool, it is limited to assess the genetic potential of the microbiome or, in other words, what the microbes could do based on their genome. Instead, metatranscriptomics assesses *what microbes are actually doing* in real-time by analyzing the RNA transcripts, providing a snapshot of actual gene expression. These approaches are complementary in decoding microbial functions, particularly in microbiome-related diseases. While shotgun metagenomics is invaluable for identifying the microbial players and their genetic potential, metatranscriptomics provides an additional layer of insight by revealing the actual functions microbes are contributing to the host's health or disease.

Chapter 5

References

[1] G. Berg, D. Rybakova, D. Fischer, *et al.*, "Microbiome definition re-visited: Old concepts and new challenges," *Microbiome*, vol. 8, pp. 1–22, 1 Jun. 2020, ISSN: 20492618. DOI: 10.1186/S40168-020-00875-0/FIGURES/7. [Online]. Available: https://microbiomejournal.biomedcentral.com/articles/10.1186/s40168-020-00875-0http://creativecommons.org/publicdomain/zero/1.0/.

[2] R. Sender, S. Fuchs, and R. Milo, "Revised estimates for the number of human and bacteria cells in the body," 2016. DOI: 10.1371/journal.pbio.1002533. [Online]. Available: https://erc.europa.eu/funding-and-grants.

[3] R. Knight, C. Callewaert, C. Marotz, *et al.*, "The microbiome and human biology," *Annual Review of Genomics and Human Genetics*, 2017. DOI: 10.1146/annurev-genom-083115. [Online]. Available: https://doi.org/10.1146/annurev-genom-083115-.

[4] J. A. Gilbert, M. J. Blaser, J. G. Caporaso, J. K. Jansson, S. V. Lynch, and R. Knight, "Current understanding of the human microbiome," *Nature Medicine 2018 24:4*, vol. 24, pp. 392–400, 4 Apr. 2018, ISSN: 1546-170X. DOI: 10.1038/nm.4517. [Online]. Available: https://www.nature.com/articles/nm.4517.

[5] E. A. Grice and J. A. Segre, "The human microbiome: Our second genome *," 2012. DOI: 10.1146/annurev-genom-090711-163814. [Online]. Available: www.annualreviews.org.

[6] P. A. Mackowiak, "The normal microbial flora," *New England Journal of Medicine*, vol. 307, pp. 83–93, 2 Jul. 1982, ISSN: 0028-4793. DOI: 10.1056/NEJM198207083070203.

[7] G. A. Weiss and T. Hennet, "Mechanisms and consequences of intestinal dysbiosis," *Cellular and Molecular Life Sciences: CMLS*, vol. 74, p. 2959, 16 Aug. 2017, ISSN: 14209071. DOI: 10.1007/S00018-017-2509-X. [Online]. Available: /pmc/articles/PMC11107543/ /pmc/articles/PMC11107543 /?report=abstracthttps://www.ncbi.nlm.nih.gov/pmc/articles/PMC11107543/.

[8] M. J. Blaser, "The microbiome revolution," *The Journal of Clinical Investigation*, vol. 124, p. 4162, 10 Oct. 2014, ISSN: 15588238. DOI: 10.1172/JCI78366. [Online]. Available: /pmc/articles/PMC4191014/ /pmc/articles/PMC4191014/?report=abstracthttps://www.ncbi.nlm.nih.gov/pmc/articles/PMC4191014/.

[9] *Nih human microbiome project - home*. [Online]. Available: https://www.hmpdacc.org/hmp/.

[10] K. Hou, Z. X. Wu, X. Y. Chen, *et al.*, "Microbiota in health and diseases," *Signal Transduction and Targeted Therapy 2022 7:1*, vol. 7, pp. 1–28, 1 Apr. 2022, ISSN: 2059-3635. DOI: 10.1038/s41392-022-00974-4. [Online]. Available: https://www.nature.com/articles/s41392-022-00974-4.

[11] C. Huttenhower, D. Gevers, R. Knight, *et al.*, "Structure, function and diversity of the healthy human microbiome," *Nature 2012 486:7402*, vol. 486, pp. 207–214, 7402 Jun. 2012, ISSN: 1476-4687. DOI: 10.1038/nature11234. [Online]. Available: https://www.nature.com/articles/nature11234.

[12] E. T. Hillman, H. Lu, T. Yao, and C. H. Nakatsu, "Microbial ecology along the gastrointestinal tract," *Microbes and Environments*, vol. 32, p. 300, 4 2017, ISSN: 13474405. DOI: 10.1264/JSME2.ME17017. [Online]. Available: /pmc/articles/PMC5745014//pmc/articles/PMC5745014/?report=abstracthttps://www.ncbi.nlm.nih.gov/pmc/articles/PMC5745014/.

[13] R. Bharti and D. G. Grimm, "Current challenges and best-practice protocols for microbiome analysis," *Briefings in Bioinformatics*, vol. 22, pp. 178–193, 1 Jan. 2021, ISSN: 14774054. DOI: 10.1093/BIB/BBZ155. [Online]. Available: https://dx.doi.org/10.1093/bib/bbz155.

[14] M. S. Kennedy and E. B. Chang, "The microbiome: Composition and locations," DOI: 10.1016/bs.pmbts.2020.08.013.

[15] H. J. Flint, K. P. Scott, P. Louis, and S. H. Duncan, "The role of the gut microbiota in nutrition and health," *Nature Reviews Gastroenterology Hepatology 2012 9:10*, vol. 9, pp. 577–589, 10 Sep. 2012, ISSN: 1759-5053. DOI: 10.1038/nrgastro.2012.156. [Online]. Available: https://www.nature.com/articles/nrgastro.2012.156.

[16] Y. Fan and O. Pedersen, "Gut microbiota in human metabolic health and disease," *Nature Reviews Microbiology 2020 19:1*, vol. 19, pp. 55–71, 1 Sep. 2020, ISSN: 1740-1534. DOI: 10.1038/s41579-020-0433-9. [Online]. Available: https://www.nature.com/articles/s41579-020-0433-9.

[17] J. G. Caporaso, C. L. Lauber, E. K. Costello, *et al.*, "Moving pictures of the human microbiome," *Genome Biology*, vol. 12, pp. 1–8, 5 May 2011, ISSN: 1474760X. DOI: 10.1186/GB-2011-12-5-R50/FIGURES/3. [Online]. Available: https://genomebiology.biomedcentral.com/articles/10.1186/gb-2011-12-5-r50.

[18] E. Rinninella, P. Raoul, M. Cintoni, *et al.*, "What is the healthy gut microbiota composition? a changing ecosystem across age, environment, diet, and diseases," *Microorganisms*, vol. 7, p. 14, 1 Jan. 2019, ISSN: 20762607. DOI: 10.3390/MICROORGANISMS7010014. [Online]. Available: /pmc/articles/PMC6351938//pmc/articles/PMC6351938/?report=abstracthttps://www.ncbi.nlm.nih.gov/pmc/articles/PMC6351938/.

[19] N. Segata, S. K. Haake, P. Mannon, *et al.*, "Composition of the adult digestive tract bacterial microbiome based on seven mouth surfaces, tonsils, throat and stool samples," *Genome biology*, vol. 13, 6 2012, ISSN: 1474-760X. DOI: 10.1186/GB-2012-13-6-R42. [Online]. Available: https://pubmed.ncbi.nlm.nih.gov/22698087/.

[20] R. J. Lamont, H. Koo, and G. Hajishengallis, "The oral microbiota: Dynamic communities and host interactions," *Nature Reviews Microbiology 2018 16:12*, vol. 16, pp. 745–759, 12 Oct. 2018, ISSN: 1740-1534. DOI: 10.1038/s41579-018-0089-x. [Online]. Available: https://www.nature.com/articles/s41579-018-0089-x.

[21] M. Avila, D. M. Ojcius, and Özlem Yilmaz, "The oral microbiota: Living with a permanent guest," *DNA and Cell Biology*, vol. 28, p. 405, 8 Aug. 2009, ISSN: 10445498. DOI: 10.1089/DNA.2009.0874. [Online]. Available: /pmc/articles/PMC2768665//pmc/articles/PMC2768665/?report=abstracthttps://www.ncbi.nlm.nih.gov/pmc/articles/PMC2768665/.

[22] E. M. Quigley and P. Gajula, "Recent advances in modulating the microbiome," *F1000Research*, vol. 9, p. 46, 2020, ISSN: 1759796X. DOI: 10.12688/F1000RESEARCH.20204.1. [Online]. Available: /pmc/articles/PMC6993818//pmc/articles/PMC6993818/?report=abstracthttps://www.ncbi.nlm.nih.gov/pmc/articles/PMC6993818/.

[23] L. A. David, C. F. Maurice, R. N. Carmody, *et al.*, "Diet rapidly and reproducibly alters the human gut microbiome," *Nature 2013 505:7484*, vol. 505, pp. 559–563, 7484 Dec. 2013, ISSN: 1476-4687. DOI: 10.1038/nature12820. [Online]. Available: https://www.nature.com/articles/nature12820.

[24] H. J. Flint, S. H. Duncan, K. P. Scott, and P. Louis, "Interactions and competition within the microbial community of the human colon: Links between diet and health," *Environmental microbiology*, vol. 9, pp. 1101–1111, 5 May 2007, ISSN: 1462-2912. DOI: 10.1111/J.1462-2920.2007.01281.X. [Online]. Available: https://pubmed.ncbi.nlm.nih.gov/17472627/.

[25] M. D. Cook, J. M. Allen, B. D. Pence, *et al.*, "Exercise and gut immune function: Evidence of alterations in colon immune cell homeostasis and microbiome characteristics with exercise training," *Immunology and cell biology*, vol. 94, pp. 158–163, 2 Feb. 2016, ISSN: 1440-1711. DOI: 10.1038/ICB.2015.108. [Online]. Available: https://pubmed.ncbi.nlm.nih.gov/26626721/.

[26] C. Benedict, H. Vogel, W. Jonas, *et al.*, "Gut microbiota and glucometabolic alterations in response to recurrent partial sleep deprivation in normal-weight young individuals," *Molecular metabolism*, vol. 5, pp. 1175–1186, 12 Dec. 2016, ISSN: 2212-8778. DOI: 10.1016/J.MOLMET.2016.10.003. [Online]. Available: https://pubmed.ncbi.nlm.nih.gov/27900260/.

[27] J. P. Karl, L. M. Margolis, E. H. Madslien, *et al.*, "Changes in intestinal microbiota composition and metabolism coincide with increased intestinal permeability in young adults under prolonged physiological stress," *American journal of physiology. Gastrointestinal and liver physiology*, vol. 312, G559–G571, 6 Jun. 2017, ISSN: 1522-1547. DOI: 10.1152/AJPGI.00066.2017. [Online]. Available: https://pubmed.ncbi.nlm.nih.gov/28336545/.

[28] S. Ying, D. N. Zeng, L. Chi, *et al.*, "The influence of age and gender on skin-associated microbial communities in urban and rural human populations," *PloS one*, vol. 10, 10 Oct. 2015, ISSN: 1932-6203. DOI: 10.1371/JOURNAL.PONE.0141842. [Online]. Available: https://pubmed.ncbi.nlm.nih.gov/26510185/.

[29] M. Zozaya, M. J. Ferris, J. D. Siren, *et al.*, "Bacterial communities in penile skin, male urethra, and vaginas of heterosexual couples with and without bacterial vaginosis," *Microbiome*, vol. 4, 2016, ISSN: 20492618. DOI: 10.1186/S40168-016-0161-6. [Online]. Available: /pmc/articles/PMC4835890//pmc/articles/PMC4835890/?report=abstracthttps://www.ncbi.nlm.nih.gov/pmc/articles/PMC4835890/.

[30] S. Lax, D. P. Smith, J. Hampton-Marcell, *et al.*, "Longitudinal analysis of microbial interaction between humans and the indoor environment," *Science (New York, N.Y.)*, vol. 345, pp. 1048–1052, 6200 Aug. 2014, ISSN: 1095-9203. DOI: 10.1126/SCIENCE.1254529. [Online]. Available: https://pubmed.ncbi.nlm.nih.gov/25170151/.

[31] R. Bud, "Antibiotics: The epitome of a wonder drug," *BMJ*, vol. 334, s6–s6, suppl 1 Jan. 2007, ISSN: 0959-8138. DOI: 10.1136/BMJ.39021.640255.94. [Online]. Available: https://www.bmj.com/content/334/suppl_1/s6https://www.bmj.com/content/334/suppl_1/s6.abstract.

[32] R. Polk, "Optimal use of modern antibiotics: Emerging trends," *Clinical Infectious Diseases*, vol. 29, pp. 264–274, 2 Jul. 1999, ISSN: 1058-4838. DOI: 10.1086/520196. [Online]. Available: https://dx.doi.org/10.1086/520196.

[33] S. B. Levy, "The antibiotic paradox," *The Antibiotic Paradox*, 1992. DOI: 10.1007/978-1-4899-6042-9.

[34] H. E. Jakobsson, C. Jernberg, A. F. Andersson, M. Sjölund-Karlsson, J. K. Jansson, and L. Engstrand, "Short-term antibiotic treatment has differing long-term impacts on the human throat and gut microbiome," *PloS one*, vol. 5, 3 2010, ISSN: 1932-6203. DOI: 10.1371/JOURNAL.PONE.0009836. [Online]. Available: https://pubmed.ncbi.nlm.nih.gov/20352091/.

[35] L. Dethlefsen, S. Huse, M. L. Sogin, and D. A. Relman, "The pervasive effects of an antibiotic on the human gut microbiota, as revealed by deep 16s rrna sequencing," *PLoS biology*, vol. 6, pp. 2383–2400, 11 Nov. 2008, ISSN: 1545-7885. DOI: 10.1371/JOURNAL.PBIO.0060280. [Online]. Available: https://pubmed.ncbi.nlm.nih.gov/19018661/.

[36] P. Amon and I. Sanderson, "What is the microbiome?" *Archives of Disease in Childhood - Education and Practice*, vol. 102, pp. 257–260, 5 Oct. 2017, ISSN: 1743-0585. DOI: 10.1136/ARCHDISCHILD-2016-311643. [Online]. Available: https://ep.bmj.com/content/102/5/257https://ep.bmj.com/content/102/5/257.abstract.

[37] *Atopy - statpearls - ncbi bookshelf*. [Online]. Available: https://www.ncbi.nlm.nih.gov/books/NBK542187/.

[38] B. SK, Z. Z, and Z. T, "The atopic march: Progression from atopic dermatitis to allergic rhinitis and asthma," *Journal of clinical cellular immunology*, vol. 5, 2 2014, ISSN: 2155-9899. DOI: 10.4172/2155-9899.1000202. [Online]. Available: https://pubmed.ncbi.nlm.nih.gov/25419479/.

[39] N. de Bortoli, P. Visaggi, R. Penagini, *et al.*, "The 1st eoetaly consensus on the diagnosis and management of eosinophilic esophagitis - definition, clinical presentation and diagnosis," *Digestive and liver disease : official journal of the Italian Society of Gastroenterology and the Italian Association for the Study of the Liver*, vol. 56, pp. 951–963, 6 Jun. 2024, ISSN: 1878-3562. DOI: 10.1016/J.DLD.2024.02.005. [Online]. Available: https://pubmed.ncbi.nlm.nih.gov/38423918/.

[40] G. T. Furuta and D. A. Katzka, "Eosinophilic esophagitis," *New England Journal of Medicine*, vol. 373, J. R. Ingelfinger, Ed., pp. 1640–1648, 17 Oct. 2015, ISSN: 0028-4793. DOI: 10.1056/NEJMRA1502863. [Online]. Available: https://www.nejm.org/doi/full/10.1056/NEJMra1502863.

[41] J. M. Wojciesko, E Eosinophilic, J. W. Asik, and E. Małecka-Wojciesko, "Eosinophilic esophagitis—what do we know so far?" *Journal of Clinical Medicine 2023, Vol. 12, Page 2259*, vol. 12, p. 2259, 6 Mar. 2023, ISSN: 2077-0383. DOI: 10.3390/JCM12062259. [Online]. Available: https://www.mdpi.com/2077-0383/12/6/2259/htmhttps://www.mdpi.com/2077-0383/12/6/2259.

[42] C. M. Rossi, G. Santacroce, M. V. Lenti, and A. di Sabatino, "Eosinophilic esophagitis in the era of biologics," *Expert review of gastroenterology hepatology*, vol. 18, pp. 271–281, 6 2024, ISSN: 1747-4132. DOI: 10.1080/17474124.2024.2374471. [Online]. Available: https://pubmed.ncbi.nlm.nih.gov/38940016/.

[43] N. Ishimura, E. Okimoto, K. Shibagaki, N. Nagano, and S. Ishihara, "Similarity and difference in the characteristics of eosinophilic esophagitis between western countries and japan," *Digestive Endoscopy*, vol. 33, pp. 708–719, 5 Jul. 2021, ISSN: 1443-1661. DOI: 10.1111/DEN.13786. [Online]. Available: https://onlinelibrary.wiley.com/doi/full/10.1111/den.13786https://onlinelibrary.wiley.com/doi/abs/10.1111/den.13786https://onlinelibrary.wiley.com/doi/10.1111/den.13786.

[44] Ángel Arias and A. J. Lucendo, "Epidemiology and risk factors for eosinophilic esophagitis: Lessons for clinicians," *Expert Review of Gastroenterology Hepatology*, vol. 14, pp. 1069–1082, 11 Nov. 2020, ISSN: 17474132. DOI: 10.1080/17474124.2020.1806054. [Online]. Available: https://www.tandfonline.com/doi/abs/10.1080/17474124.2020.1806054.

[45] P. Navarro, E. J. Laserna-Mendieta, S. Casabona, *et al.*, "Accurate and timely diagnosis of eosinophilic esophagitis improves over time in europe. an analysis of the eoe connect registry," *United European gastroenterology journal*, vol. 10, pp. 507–517, 5 Jun. 2022, ISSN: 2050-6414. DOI: 10.1002/UEG2.12240. [Online]. Available: https://pubmed.ncbi.nlm.nih.gov/35578565/.

[46] L. C. Kottyan, S. Parameswaran, M. T. Weirauch, M. E. Rothenberg, and L. J. Martin, "The genetic etiology of eosinophilic esophagitis," *Journal of Allergy and Clinical Immunology*, vol. 145, pp. 9–15, 1 Jan. 2020, ISSN: 10976825. DOI: 10.1016/j.jaci.2019.11.013. [Online].

Available: http://www.jacionline.org/article/S0091674919315489/fulltexthttp://www.jacionline.org/article/S0091674919315489/abstracthttps://www.jacionline.org/article/S0091-6749(19)31548-9/abstract.

[47] H. Philpott, S. Nandurkar, S. G. Royce, F. Thien, and P. R. Gibson, "Risk factors for eosinophilic esophagitis," *Clinical and Experimental Allergy*, vol. 44, pp. 1012–1019, 8 2014, ISSN: 13652222. DOI: 10.1111/CEA.12363.

[48] V. A. Litosh, M. Rochman, J. K. Rymer, A. Porollo, L. C. Kottyan, and M. E. Rothenberg, "Calpain-14 and its association with eosinophilic esophagitis," *The Journal of allergy and clinical immunology*, vol. 139, p. 1762, 6 Jun. 2017, ISSN: 10976825. DOI: 10.1016/J.JACI.2016.09.027. [Online]. Available: /pmc/articles/PMC5461191//pmc/articles/PMC5461191/?report=abstracthttps://www.ncbi.nlm.nih.gov/pmc/articles/PMC5461191/.

[49] E. S. Alexander, L. J. Martin, M. H. Collins, *et al.*, "Twin and family studies reveal strong environmental and weaker genetic cues explaining heritability of eosinophilic esophagitis," *The Journal of allergy and clinical immunology*, vol. 134, 1084–1092.e1, 5 Nov. 2014, ISSN: 1097-6825. DOI: 10.1016/J.JACI.2014.07.021. [Online]. Available: https://pubmed.ncbi.nlm.nih.gov/25258143/.

[50] S. C. Shah, A. Tepler, R. M. Peek, J. F. Colombel, I. Hirano, and N. Narula, "Association between helicobacter pylori exposure and decreased odds of eosinophilic esophagitis-a systematic review and meta-analysis," *Clinical gastroenterology and hepatology : the official clinical practice journal of the American Gastroenterological Association*, vol. 17, 2185–2198.e3, 11 Oct. 2019, ISSN: 1542-7714. DOI: 10.1016/J.CGH.2019.01.013. [Online]. Available: https://pubmed.ncbi.nlm.nih.gov/30659992/.

[51] M. M. Asfari, K. Kendrick, M. T. Sarmini, P. Uy, and K. J. Vega, "Association of eosinophilic esophagitis and human immunodeficiency virus," *Digestive Diseases and Sciences*, vol. 66, pp. 2669–2673, 8 Aug. 2021, ISSN: 15732568. DOI: 10.1007/S10620-020-06566-Y/METRICS. [Online]. Available: https://link.springer.com/article/10.1007/s10620-020-06566-y.

[52] Y. Suzuki, T. Iizuka, A. Hosoi, *et al.*, "Clinicopathological differences between eosinophilic esophagitis and asymptomatic esophageal eosinophilia," *Internal Medicine*, vol. 61, pp. 1319–1327, 9 May 2022, ISSN: 0918-2918. DOI: 10.2169/INTERNALMEDICINE.8241-21. [Online]. Available: http://internmed.jp.

[53] P. Visaggi, E. Savarino, G. Sciume, *et al.*, "Eosinophilic esophagitis: Clinical, endoscopic, histologic and therapeutic differences and similarities between children and adults," *Therapeutic Advances in Gastroenterology*, vol. 14, Jan. 2021, ISSN: 17562848. DOI: 10.1177/1756284820980860/ASSET/IMAGES/LARGE/10.1177_1756284820980860-FIG1.JPEG. [Online]. Available: https://journals.sagepub.com/doi/10.1177/1756284820980860.

[54] E. Rubinstein and R. L. Rosen, "Respiratory symptoms associated with eosinophilic esophagitis," *Pediatric Pulmonology*, vol. 53, pp. 1587–1591, 11 Nov. 2018, ISSN: 1099-0496. DOI: 10.1002/PPUL.24168. [Online]. Available: https://onlinelibrary.wiley.com/doi/full/10.1002/ppul.24168https://onlinelibrary.wiley.com/doi/abs/10.1002/ppul.24168https://onlinelibrary.wiley.com/doi/10.1002/ppul.24168.

[55] E. S. Dellon, C. A. Liacouras, J. Molina-Infante, *et al.*, "Updated international consensus diagnostic criteria for eosinophilic esophagitis: Proceedings of the agree conference," *Gastroenterology*, vol. 155, 1022–1033.e10, 4 Oct. 2018, ISSN: 1528-0012. DOI: 10.1053/J.GASTRO.2018.07.009. [Online]. Available: https://pubmed.ncbi.nlm.nih.gov/30009819/.

[56] J. A. Alexander, "Endoscopic and radiologic findings in eosinophilic esophagitis," *Gastrointestinal Endoscopy Clinics of North America*, vol. 28, pp. 47–57, 1 Jan. 2018, ISSN: 1052-5157. DOI: 10.1016/J.GIEC.2017.07.003.

[57] D. Kim, L. Pantanowitz, P. Schüffler, *et al.*, "(re) defining the high-power field for digital pathology," *Journal of Pathology Informatics*, vol. 11, p. 33, 1 Jan. 2020, ISSN: 21533539. DOI: 10.4103/JPI.JPI_48_20. [Online]. Available: /pmc/articles/PMC7737490//pmc/articles/PMC7737490/?report=abstracthttps://www.ncbi.nlm.nih.gov/pmc/articles/PMC7737490/.

[58] A. Gómez-Aldana, M. Jaramillo-Santos, A. Delgado, C. Jaramillo, and A. Lúquez-Mindiola, "Eosinophilic esophagitis: Current concepts in diagnosis and treatment," *http://www.wjgnet.com/*, vol. 25, pp. 4598–4613, 32 Aug. 2019, ISSN: 22192840. DOI: 10.3748/WJG.V25.I32.4598. [Online]. Available: https://www.wjgnet.com/1007-9327/full/v25/i32/4598.htm.

[59] *Jorveza — european medicines agency (ema)*. [Online]. Available: https://www.ema.europa.eu/en/medicines/human/EPAR/jorveza.

[60] S. Feo-Ortega and A. J. Lucendo, "Evidence-based treatments for eosinophilic esophagitis: Insights for the clinician," *Therapeutic advances in gastroenterology*, vol. 15, Jan. 2022, ISSN: 1756-283X. DOI: 10.1177/17562848211068665. [Online]. Available: https://pubmed.ncbi.nlm.nih.gov/35069803/.

[61] E. P. Syverson, E. Rubinstein, J. J. Lee, D. R. McDonald, and E. Hait, "The role of dupilumab in the treatment of eosinophilic esophagitis," *Immunotherapy*, Jul. 2024, ISSN: 17507448. DOI: 10.1080/1750743X.2024.2377060. [Online]. Available: https://www.tandfonline.com/doi/abs/10.1080/1750743X.2024.2377060.

[62] N. Segata, D. Boernigen, T. L. Tickle, X. C. Morgan, W. S. Garrett, and C. Huttenhower, "Computational meta'omics for microbial community studies," *Molecular Systems Biology*, vol. 9, p. 666, 1 2013, ISSN: 17444292. DOI: 10.1038/MSB.2013.22. [Online]. Available: /pmc/articles/PMC4039370//pmc/articles/PMC4039370/?report=abstracthttps://www.ncbi.nlm.nih.gov/pmc/articles/PMC4039370/.

[63] J. Galloway-Peña and B. Hanson, "Tools for analysis of the microbiome," *Digestive diseases and sciences*, vol. 65, p. 674, 3 Mar. 2020, ISSN: 15732568. DOI: 10.1007/S10620-020-06091-Y. [Online]. Available: /pmc/articles/PMC7598837//pmc/articles/PMC7598837/?report=abstracthttps://www.ncbi.nlm.nih.gov/pmc/articles/PMC7598837/.

[64] T. M. Sirangelo, "Human gut microbiome analysis and multi-omics approach," *International Journal of Pharma Medicine and Biological Sciences*, vol. 7, pp. 52–57, 3 Jul. 2018, ISSN: 22785221. DOI: 10.18178/IJPMBS.7.3.52-57. [Online]. Available:

https://www.researchgate.net/publication/335069570_Human_Gut_Microbiome_Analysis_and_Multi-omics_Approach.

[65] E. A. Franzosa, X. C. Morgan, N. Segata, *et al.*, "Relating the metatranscriptome and metagenome of the human gut," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 111, 22 Jun. 2014, ISSN: 1091-6490. DOI: 10.1073/PNAS.1319284111. [Online]. Available: https://pubmed.ncbi.nlm.nih.gov/24843156/.

[66] M. Shakya, C. C. Lo, and P. S. Chain, "Advances and challenges in metatranscriptomic analysis," *Frontiers in genetics*, vol. 10, SEP Sep. 2019, ISSN: 1664-8021. DOI: 10.3389/FGENE.2019.00904. [Online]. Available: https://pubmed.ncbi.nlm.nih.gov/31608125/.

[67] T. Ojala, E. Kankuri, and M. Kankainen, "Understanding human health through metatranscriptomics," *Trends in Molecular Medicine*, vol. 29, pp. 376–389, 5 May 2023, ISSN: 1471499X. DOI: 10.1016/J.MOLMED.2023.02.002/ASSET/8D11FE2A-00AD-4A67-A7E2-9E6964E621EA/MAIN.ASSETS/GR2.JPG. [Online]. Available: http://www.cell.com/article/S1471491423000345/fulltexthttp://www.cell.com/article/S1471491423000345/abstracthttps://www.cell.com/trends/molecular-medicine/abstract/S1471-4914(23)00034-5.

[68] Y. Pinto and A. S. Bhatt, "Sequencing-based analysis of microbiomes," *Nature Reviews Genetics 2024*, pp. 1–17, Jun. 2024, ISSN: 1471-0064. DOI: 10.1038/s41576-024-00746-6. [Online]. Available: https://www.nature.com/articles/s41576-024-00746-6.

[69] R. Knight, A. Vrbanac, B. C. Taylor, *et al.*, "Best practices for analysing microbiomes," *Nature Reviews Microbiology 2018 16:7*, vol. 16, pp. 410–422, 7 May 2018, ISSN: 1740-1534. DOI: 10.1038/s41579-018-0029-9. [Online]. Available: https://www.nature.com/articles/s41579-018-0029-9.

[70] C. Quince, A. W. Walker, J. T. Simpson, N. J. Loman, and N. Segata, "Shotgun metagenomics, from sampling to analysis," *Nature Biotechnology 2017 35:9*, vol. 35, pp. 833–844, 9 Sep. 2017, ISSN: 1546-1696. DOI: 10.1038/nbt.3935. [Online]. Available: https://www.nature.com/articles/nbt.3935.

[71] T. Cernava, D. Rybakova, F. Buscot, *et al.*, "Metadata harmonization–standards are the key for a better usage of omics data for integrative microbiome analysis," *Environmental Microbiomes*, vol. 17, pp. 1–10, 1 Dec. 2022, ISSN: 25246372. DOI: 10.1186/S40793-022-00425-1/FIGURES/2. [Online]. Available: https://link.springer.com/articles/10.1186/s40793-022-00425-1https://link.springer.com/article/10.1186/s40793-022-00425-1.

[72] *Read - genomics education programme*. [Online]. Available: https://www.genomicseducation.hee.nhs.uk/glossary/read/.

[73] *Read length - genomics education programme*. [Online]. Available: https://www.genomicseducation.hee.nhs.uk/glossary/read-length/.

[74] *Sequencing quality scores*. [Online]. Available: https://emea.illumina.com/science/technology/next-generation-sequencing/plan-experiments/quality-scores.html.

[75] Y. Zhou, M. Liu, and J. Yang, "Recovering metagenome-assembled genomes from shotgun metagenomic sequencing data: Methods, applications, challenges, and opportunities," *Microbiological Research*, vol. 260, p. 127 023, Jul. 2022, ISSN: 0944-5013. DOI: 10.1016/J.MICRES.2022.127023.

[76] P. J. Cock, C. J. Fields, N. Goto, M. L. Heuer, and P. M. Rice, "The sanger fastq file format for sequences with , and the solexa/illumina fastq variants," *Nucleic Acids Research*, vol. 38, p. 1767, 6 Dec. 2010, ISSN: 03051048. DOI: 10.1093/NAR/GKP1137. [Online]. Available: /pmc/articles/PMC2847217//pmc/articles/PMC2847217/?report=abstracthttps://www.ncbi.nlm.nih.gov/pmc/articles/PMC2847217/.

[77] Z. Zhou, N. Luhmann, N.-F. Alikhan, C. Quince, and M. Achtman, "Accurate reconstruction of microbial strains from metagenomic sequencing using representative reference genomes," *bioRxiv*, p. 215 707, Nov. 2017. DOI: 10.1101/215707. [Online]. Available: https://www.biorxiv.org/content/10.1101/215707v2https://www.biorxiv.org/content/10.1101/215707v2.abstract.

[78] *Biobakery/humann: Humann is the next generation of humann 1.0 (hmp unified metabolic analysis network).* [Online]. Available: https://github.com/biobakery/humann?tab=readme-ov-file#download-a-translated-search-database.

[79] A. Blanco-Míguez, F. Beghini, F. Cumbo, *et al.*, "Extending and improving metagenomic taxonomic profiling with uncharacterized species using metaphlan 4," *Nature Biotechnology 2023 41:11*, vol. 41, pp. 1633–1644, 11 Feb. 2023, ISSN: 1546-1696. DOI: 10.1038/s41587-023-01688-w. [Online]. Available: https://www.nature.com/articles/s41587-023-01688-w.

[80] J. G. Kers and E. Saccenti, "The power of microbiome studies: Some considerations on which alpha and beta metrics to use and how to report results," *Frontiers in Microbiology*, vol. 12, Mar. 2022, ISSN: 1664302X. DOI: 10.3389/fmicb.2021.796025.

[81] J. R. Rideout, G. Caporaso, E. Bolyen, *et al.*, *Biocore/scikit-bio: Scikit-bio 0.5.9: Maintenance release*, Aug. 2023. DOI: 10.5281/zenodo.8209901. [Online]. Available: https://doi.org/10.5281/zenodo.8209901.

[82] L. J. McIver, G. Abu-Ali, E. A. Franzosa, *et al.*, "Biobakery: A meta'omic analysis environment," *Bioinformatics (Oxford, England)*, vol. 34, pp. 1235–1237, 7 Apr. 2018, ISSN: 1367-4811. DOI: 10.1093/BIOINFORMATICS/BTX754. [Online]. Available: https://pubmed.ncbi.nlm.nih.gov/29194469/.

[83] *Babraham bioinformatics - fastqc a quality control tool for high throughput sequence data.* [Online]. Available: https://www.bioinformatics.babraham.ac.uk/projects/fastqc/.

[84] P. Ewels, M. Magnusson, S. Lundin, and M. Käller, "Multiqc: Summarize analysis results for multiple tools and samples in a single report," *Bioinformatics*, vol. 32, pp. 3047–3048, 19 Oct. 2016, ISSN: 14602059. DOI: 10.1093/BIOINFORMATICS/BTW354. [Online]. Available: https://multiqc.info/.

[85] A. M. Bolger, M. Lohse, and B. Usadel, "Trimmomatic: A flexible trimmer for illumina sequence data," *Bioinformatics*, vol. 30, pp. 2114–2120, 15 Aug. 2014, ISSN: 1367-4803. DOI: 10.1093/BIOINFORMATICS/BTU170. [Online]. Available: https://dx.doi.org/10.1093/bioinformatics/btu170.

[86] B. Langmead and S. L. Salzberg, "Fast gapped-read alignment with bowtie 2," *Nature Methods 2012 9:4*, vol. 9, pp. 357–359, 4 Mar. 2012, ISSN: 1548-7105. DOI: 10.1038/nmeth.1923. [Online]. Available: https://www.nature.com/articles/nmeth.1923.

[87] *Homo sapiens genome assembly grch38 - ncbi - nlm*. [Online]. Available: https://www.ncbi.nlm.nih.gov/datasets/genome/GCF_000001405.26/.

[88] G. Benson, "Tandem repeats finder: A program to analyze dna sequences," *Nucleic Acids Research*, vol. 27, pp. 573–580, 2 Jan. 1999, ISSN: 0305-1048. DOI: 10.1093/NAR/27.2.573. [Online]. Available: https://dx.doi.org/10.1093/nar/27.2.573.

[89] *Updated kneaddata to fix issue with paired-end reads? - infrastructure and utilities / kneaddata - the biobakery help forum*. [Online]. Available: https://forum.biobakery.org/t/updated-kneaddata-to-fix-issue-with-paired-end-reads/5719/3.

[90] F. Beghini, L. J. McIver, A. Blanco-Míguez, *et al.*, "Integrating taxonomic, functional, and strain-level profiling of diverse microbial communities with bioBakery 3," *eLife*, vol. 10, May 2021, ISSN: 2050084X. DOI: 10.7554/ELIFE.65088.

[91] W. McKinney, "Data structures for statistical computing in python," *Proceedings of the 9th Python in Science Conference*, pp. 56–61, 2010. DOI: 10.25080/MAJORA-92BF1922-00A. [Online]. Available: https://www.researchgate.net/publication/340177686_Data_Structures_for_Statistical_Computing_in_Python.

[92] C. R. Harris, K. J. Millman, S. J. van der Walt, *et al.*, "Array programming with numpy," *Nature 2020 585:7825*, vol. 585, pp. 357–362, 7825 Sep. 2020, ISSN: 1476-4687. DOI: 10.1038/s41586-020-2649-2. [Online]. Available: https://www.nature.com/articles/s41586-020-2649-2.

[93] P. Virtanen, R. Gommers, T. E. Oliphant, *et al.*, "Scipy 1.0: Fundamental algorithms for scientific computing in python," *Nature Methods 2020 17:3*, vol. 17, pp. 261–272, 3 Feb. 2020, ISSN: 1548-7105. DOI: 10.1038/s41592-019-0686-2. [Online]. Available: https://www.nature.com/articles/s41592-019-0686-2.

[94] M. L. Waskom, "Seaborn: Statistical data visualization," *Journal of Open Source Software*, vol. 6, p. 3021, 60 Apr. 2021, ISSN: 2475-9066. DOI: 10.21105/JOSS.03021. [Online]. Available: https://joss.theoj.org/papers/10.21105/joss.03021.

[95] J. D. Hunter, "Matplotlib: A 2d graphics environment," *Computing in Science and Engineering*, vol. 9, pp. 90–95, 3 2007, ISSN: 15219615. DOI: 10.1109/MCSE.2007.55.

[96] J. R. Miller, S. Koren, and G. Sutton, "Assembly algorithms for next-generation sequencing data," *Genomics*, vol. 95, pp. 315–327, 6 Jun. 2010, ISSN: 0888-7543. DOI: 10.1016/J.YGENO.2010.03.001.

[97] "Sequence-specific error profile of illumina sequencers," *Nucleic Acids Research*, vol. 39, e90–e90, 13 Jul. 2011, ISSN: 0305-1048. DOI: 10.1093/NAR/GKR344. [Online]. Available: https://dx.doi.org/10.1093/nar/gkr344.

[98] P. C. Austin, "An introduction to propensity score methods for reducing the effects of confounding in observational studies," *Multivariate Behavioral Research*, vol. 46, p. 399, 3 May 2011, ISSN: 00273171. DOI: 10.1080/00273171.2011.568786. [Online]. Available: /pmc/articles/PMC3144483//pmc/articles/PMC3144483/?report=abstracthttps://www.ncbi.nlm.nih.gov/pmc/articles/PMC3144483/.

[99] "Species-level functional profiling of metagenomes and metatranscriptomes," *Nature Methods 2018 15:11*, vol. 15, pp. 962–968, 11 Oct. 2018, ISSN: 1548-7105. DOI: 10.1038/s41592-018-0176-y. [Online]. Available: https://www.nature.com/articles/s41592-018-0176-y.

[100] C. J. Mann, "Observational research methods. research design ii: Cohort, cross sectional, and case-control studies," *Emergency Medicine Journal*, vol. 20, pp. 54–60, 1 Jan. 2003, ISSN: 1472-0205. DOI: 10.1136/EMJ.20.1.54. [Online]. Available: https://emj.bmj.com/content/20/1/54https://emj.bmj.com/content/20/1/54.abstract.