UNIVERSITY OF VERONA

DEPARTMENT OF
NEUROSCIENCES, BIOMEDICINE AND MOVEMENT SCIENCES

GRADUATE SCHOOL OF
LIFE AND HEALTH SCIENCES

PhD PROGRAM IN
APPLIED LIFE AND HEALTH SCIENCES

With funding by
Department of Excellence

Pitfalls in bacteria recognition: a 16S rRNA gene sequencing-based simulation study on the oral microbiome environment

Cycle XXXV / Year 2019
S.S.D. MED/03

Coordinator: Prof. Simone Accordini
Supervisor: Prof. Giovanni Malerba

Candidate: Dott.ssa Elena Locatelli

CONTENTS

Riassunto

La ricerca sul microbioma si è evoluta rapidamente ed è diventata un argomento importante nel corso degli anni. Lo studio del microbioma consente di indagare le comunità batteriche che risiedono in specifici distretti corporei e di comprendere il loro ruolo in condizioni di salute o malattia dell'ospite, osservandone la struttura, la funzione e l'interazione tra questi.

Le tecniche di *amplicon sequencing* di un gene target (16S rRNA gene) vengono comunemente utilizzate per conoscere la composizione batterica del microbiota dell'ospite umano, raccogliendo informazioni sull'abbondanza dei diversi batteri.

In questo elaborato abbiamo proposto un'analisi bioinformatica del gene rRNA 16S, eseguendo una classificazione tassonomica dei batteri che risiedono nel cavo orale umano, e successivamente simulando un microbioma, dissezionando, e sottolineando quali sono i punti deboli di un'analisi effettuata dal sequenziamento tramite ampliconi.

Per rilevare la composizione del microbiota orale di 4 individui sani, l'analisi è stata eseguita utilizzando tutte le regioni del gene rRNA 16S, sequenziando 6 ampliconi comprendenti 2 o 3 regioni ipervariabili consecutive. Abbiamo utilizzato il *Divisive Amplicon Denoising Algorithm 2* (DADA2) in grado di dedurre la composizione batterica dei campioni mediante il rilevamento di *Amplicon Sequence Variants* (ASVs), offrendo uno studio con risoluzione a singolo nucleotide. Il database pubblico *Human Oral Microbiota Database* (eHOMD) è stato utilizzato come riferimento per assegnare la tassonomia alle sequenze batteriche. È stato eseguito un confronto tra tutte le regioni del gene rRNA 16S per identificare quale tra queste regioni identificasse al meglio la composizione batterica.

La seconda parte dello studio si è basata sui risultati ottenuti dall'analisi del gene rRNA 16S utilizzati come pannello di riferimento. Abbiamo sviluppato un simulatore basato sul gene rRNA 16S per mimare un processo di sequenziamento tramite un gene marcatore, evidenziando le possibili problematiche che potrebbero insorgere durante il sequenziamento di ampliconi, seguito poi dalla loro successiva analisi. Abbiamo selezionato le sequenze dell'intero gene rRNA 16S dal database eHOMD e le abbiamo suddivise in ampliconi. Quindi, è stata effettuata una classificazione per assegnare un livello tassonomico a ciascuna sequenza simulata. Le sequenze batteriche a cui non è

stata assegnata una tassonomia sono state sottoposte ad allineamento multiplo combinato con a un albero filogenetico. Per fare ciò, abbiamo utilizzato tutte le sequenze di riferimento dell'intero gene rRNA 16S provenienti da eHOMD e appartenenti alle possibili specie batteriche coinvolte nella soluzione offerta dal classificatore in modo tale da indagare la loro similarità di sequenza.

I nostri dati hanno riportato migliori risultati svolgendo l'analisi sull'intero set di ampliconi del 16S identificando 204 diverse specie batteriche. Complessivamente, l'analisi sui conteggi delle ASV ha evidenziato che circa il 44% degli ASV totali ha raggiunto la classificazione più dettagliata a livello di specie. Tuttavia, l'amplicone V2V3 è risultata la porzione del gene rRNA 16S che ha riconosciuto il maggior numero di batteri a livello di specie (135) attraverso l'analisi del singolo amplicone.

Il processo di simulazione, che prevedeva l'estrazione degli ampliconi dalle sequenze batteriche simulate dell'intero gene rRNA 16S, ha mostrato che non tutti gli ampliconi sono stati identificati, suggerendo che alcuni geni 16S rRNA potrebbero non venire amplificati, come spesso accade nel mondo reale. Inoltre, diverse specie batteriche risultano avere un alto grado di similarità di sequenza in una data regione del gene 16S rRNA, rendendo in alcuni casi la classificazione incompleta.

L'utilizzo del simulatore sarà poi utile per comprendere la sensibilità delle diverse regioni del gene 16S rRNA nel riconoscere la presenza di specie batteriche potenzialmente patogene.

Il nostro studio verrà applicato inoltre su un campione più ampio di centinaia di individui, migliorando le prestazioni di classificazione della tassonomia.

Abstract

Microbiome research has evolved rapidly and became an important topic over the years. The study of the microbiome permits us to investigate bacteria living in specific body sites and to understand its role in host health and disease conditions, observing the structure, function, and interaction between the different compounds.

The amplicon sequencing is the most widely used technique to get the diversity composition of the microbiota. In particular, the 16S rRNA gene is the commonly adopted marker gene to identify microbial communities within the human host, collecting information on their relative abundance.

In this elaborate we proposed a bioinformatic analysis of the 16S rRNA gene, performing a taxonomical classification of bacteria residing in the human oral cavity, and simulating a microbiome environment, dissecting, and pointing out what are the weak points of an analysis carried out by amplicons sequencing.

To detect the microbial composition in 4 oral healthy individuals, the analysis was performed across the 16S rRNA gene amplicon regions by sequencing 6 amplicons involving 2 or 3 consecutive 16S rRNA hypervariable regions. We used the *Divisive Amplicon Denoising Algorithm 2* (DADA2) able to infer the bacteria composition of samples by the detection of Amplicon Sequence Variants (ASVs), offering a single nucleotide resolution study. The public *extended Human Oral Microbiota Database* (eHOMD) was adopted as reference database to assign the taxonomy to the bacteria sequences. A comparison between all the 16S rRNA gene amplicon region was performed to identify which amplicon region better identified microbial communities.

The second and most relevant part of the study was based on the results obtained from the 16S rRNA gene amplicon regions analysis used as reference panel. We developed a 16S rRNA gene-based simulator to mimic a marker gene sequencing process, evaluating what might be the pitfalls that could be encountered by carrying out a targeted metagenomic analysis. We selected the full-length 16S rRNA gene sequences from the eHOMD database and we subdivided the sequences into different amplicon regions. Then, a taxonomic classification was carried out to assign a taxonomic rank to the simulated amplicons. Taxonomically unassigned bacterial species underwent to multiple sequence alignment combined with a phylogenetic tree construction against

all full-length rRNA gene reference sequences of the eHOMD belonging to the possible Species solutions to investigate the sequence identity between them.

Our data reported an improved 16S analysis run with DADA2 using the pooled marker gene amplicon regions in which we identified 204 different bacterial species. Overall, the analysis on the unique ASV counts of the pooled 16S rRNA gene reported that approximately 44% of total ASVs achieved the most detailed species-level classification. However, the amplicon region V2V3 proved to be the portion of the 16S rRNA that recognized the highest number of bacteria at the species taxonomic level (135) through the single amplicon analysis.

The simulation process, involving the extraction of amplicon sequences from the full-length 16S rRNA gene sequences of simulated bacterial sequences, showed that not all the amplicons can be extracted, suggesting, several 16S rRNA gene amplicon sequences were not amplified, as often happens in the real world. Furthermore, several bacterial species appear to have a high degree of sequence similarity in a given region of the 16S rRNA gene, making classification incomplete in some cases.

The use of the simulator will then be useful to understand the sensitivity of the different regions of the 16S rRNA gene in recognizing the presence of pathogenic bacterial species.

Our study can then be adopted to build up an accurate catalog of oral ASVs to essay large sample sets (e.g., hundreds of individuals), aiming at improving the taxonomy classification performances.

1 INTRODUCTION

Microbiome research has rapidly evolved over the years and with the advent of NGS technologies the field of metagenomics has begun[1,2].

The set of prokaryotic organisms, including bacteria and archaea, fungi, viruses, and eukaryotes, live in specific environments and colonize different body sites[3,4], maintaining an equilibrium with the host is called microbiota[3]. More in detail, the term microbiome refers to the catalog of microbes and their genes[5], living in the same habitat[1,4].

New sequencing technologies permitted us to study and characterize the microbial communities of specific environments, like in humans, animals, and plants[1,2]. Moreover, current methods pointed out beneficial and critical associations between microbiome and the human host, that is largely studied[1,2].

1.1 The Human Microbiome

Microbiome research connects many research fields, especially human medicine[1].

The human microbiome is defined as the community of microorganisms inhabiting the human host[2], recognized as the human "last organ"[1]. Microorganisms could be found on the surface and inside the human host, and their heterogeneity depends on the body site that they colonize (e.g., skin, tooth, respiratory tracts, gut, mucosal surface) and the human genetics[2].

The METAgenomics of the Human Intestinal Tract (Meta-HIT)[6] consortium reported that the human microbiota consists of 10^{14} bacterial cells living within the human body, first among all in the gut[5]. Humans contain ~ 23.000 protein-coding genes[7,4], so ten times smaller than the bacterial cells.

The human microbiota can be found in different human body sites, and it is different according to the site where it is located, for example in skin, mouth, urogenital tract, respiratory system, and the digestive tract that contains the largest amount of bacteria in the human body[4].

The human microbiome plays a key role in immune system and in metabolic functions[8].

Many microbiome projects aim to understand how the role of microbiota impacts on human health and disease[5]. Different studies revealed that healthy individuals present different microbes occupying the different body sites, but most of this observed diversity remains unexplained[9]. In addition, host genetics, diet, environmental factors are also implicated[9].

The Microbiome Project (HMP)[10] collect many projects with the aim to identify and understand how healthy conditions or disease predispositions are determined, studying the diversity of the human genetics and individual's physiology, and looking at human microbiome and different factors that influence the distribution of the microorganisms[7].

One of the questions to answer is whether a "core microbiome" is identifiable[7], detecting the temporal spatial dynamics of microbiomes[1], that may help to determine which members of a microbiome are permanent, thus identifying temporal and variable bacteria, ideally associated with specific conditions[1].

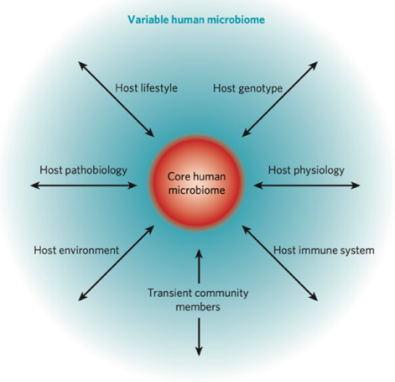


Figure 1. The core human microbiome involves a set of genes present in a given habitat (entire body, or a specific surface area) and linked to a variable part depending on the studied subject. Image from Turnbaugh et al., 2007[7].

Many microbiota analyses identified common and rare taxa in several body habitats[9]. As an example, in a healthy human oral cavity many studies reported a prevalence of Firmicutes, Actinobacteria, Proteobacteria, Fusobacteria, Bacteroidetes and Spirochaetes[8,11], that constitute ~96% of the total oral bacteria community[11].

In the human gut Firmicutes and Bacteroides phyla are the most prevalent bacteria in healthy status[8].

Microbes biodiversity is defined as the abundance distribution of different types of organisms living in a given body habitat[9].

Moreover, many microbiome analyses distinguished beneficial and pathogenic microorganisms, and also some neutral microorganisms[1], looking at their interaction with the human host. Indeed, the beneficial bacteria seem to maintain the equilibrium of a healthy state (healthy microbiota)[1], while the pathogenic microorganisms lead the host to a dysbiosis status[1].

Unfortunately, less than 1% of microbes can be cultivated, and with the next generation sequencing technologies and the decreasing cost of DNA sequencing, many computational methods have been developed for helping the microbiome studies and bacteria identification[8].

1.1.1 The Human Oral Microbiota

The human oral microbiota is one of the largest microbial communities (after gut) and it's defined as the set of microorganisms residing in the human oral cavity[12]. Aerodigestive tract includes the oral cavity, pharynx, esophagus, nasal passages, and sinuses[13], and the human oral cavity contains different habitats colonized by bacteria (the teeth, gingival sulcus, tongue, cheeks, hard and soft palates, and tonsils)[12].

The human aerodigestive tract harbors beneficial and pathogenic bacterial species of the same genus[13]. So, understanding the microbiome composition and its function is important to understand the healthy conditions and also the disease conditions associated mostly with bacterial pathogens that characterize dysbiosis[13].

The human oral microbiome is one of the most studied human microflora[14], and it is a major gateway to the human body[12], constantly exposed to exogenous foreign substances[15].

Oral cavity-associated microbes can influence immune responses[15]. A lot of studies associated the oral bacteria to a number of systemic diseases, like cardiovascular diseases, stroke, and diabetes[12], so oral cavity-associated microbes have been associated in many distant organ sites, including small intestine, lungs, heart, brain,

…[15] These kinds of associations could be implicated in many infectious diseases, like periodontitis, and chronic conditions, such as cardiovascular diseases[12,14,15], and other infections recognized to be caused by group of organisms rather than a single pathogen[12]. For this reason, we need to identify most of the human microorganisms in order to understand the healthy and diseased human conditions and to carry out meaningful clinical research[12].

It is estimated that over 700 prevalent taxa have been identified at the strain-level in human oral cavity[12,14], representing the second most diverse microbiota environment of human body[15].

Unfortunately, 53% of the species have been not yet named and 35% remain uncultivable[14], and the unnamed oral taxa are referenced by clone numbers of 16S rRNA gene that could not have a taxonomic assignment[12]. Optimizing the identification of the aerodigestive microbial sequences at species level or at least at genus layer could help in clinical relevance of the microbiome studies[13].

The oral cavity is constantly exposed to external environment[16], and it is the first organ to encounter foodstuff, exogenous microbes, and allergens, before the gastrointestinal and/or the respiratory tract, so the identification of the exact composition of the oral microbiota is difficult[15]. Oral microbiome composition changes throughout life by many factors that also include the host genetics, as well as environmental factors; the microbial composition depends also on individual's habits (e.g., oral hygiene)[16]. Diet also is able to enrich the oral microbiota community. Commensal oral microbiota bacteria are important in maintaining a healthy status: as an example, the Haemophius parainfluenzae is able to decrease the adhesion ability of Porphyromonas gingivalis, involved in etiology of dental caries[15].

1.2 16S rRNA gene

Many computational methods have been developed, in order to understand the structure and the composition of a microbial community[8,17]. Sequence-based approaches are generally designed to assess microbial composition[15] by sequencing technologies like PCR using DNA and RNA-based approaches[1], identifying the

community members, exploiting sequence variabilities[15], focusing also on a specific microbial gene or region.

The introduction of marker genes like the prokaryotic 16S ribosomal RNA (16S rRNA) gene by Carl Woese and George E. Fox in 1977[1] permitted the investigation of microbial community[15] and the identification of bacterial species performing taxonomic assignments and phylogenetic analysis of microbes[18,19].

The 16S rRNA gene is mostly used gene, because it encodes prokaryotic small 30S subunit of the 70S ribosomal complex[20] and it is present in all microorganisms[7,20]. The classification of organisms into taxa is the aim of the taxonomy[17]. Since the 16S rRNA gene it is conserved among bacteria, the classification can occur for a broad range of prokaryotic organisms[21].

The 16S rRNA gene is ~1500 bp length and it consists of nine hypervariable regions, named from V1 to V9, interspersed between conserved primer binding sites regions[20], that enabling PCR amplification of target sequences using universal primers[18].

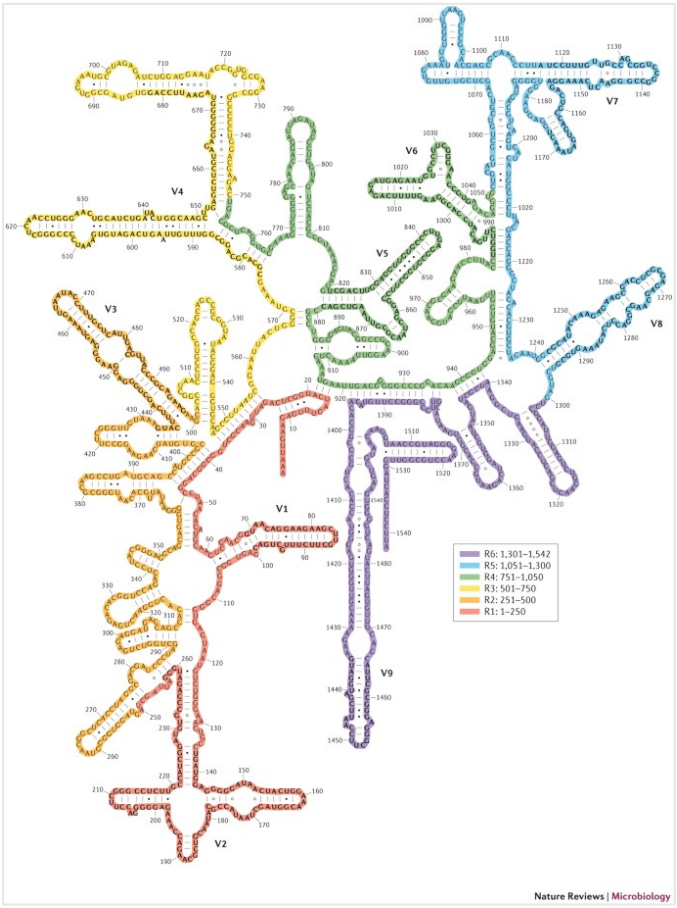


Figure 2. Secondary structure of the 16S rRNA. Image from Yarza P et al., 2014[19].

The hypervariable regions can distinguish and highlight sequence diversity among different bacteria[18,21], but the different degree of sequence diversity of hypervariable regions limits the ability of a single region to distinguish between all bacteria[18]. Indeed, some hypervariable regions of the gene appear to be more reliable in identifying some bacterial species than others. As an example, the hypervariable region V1 is used in differentiation of oral colonizers like Streptococcus[11] and it's also able to distinguish among Staphylococcus aureus and coagulase negative Staphylococcus sp.[18]; then, hypervariable region V3 seems to be specialized in distinguish among Haemophilus species[18]. V1-V3 amplicon sequences are reported to better differentiate among Prevotella, Porphyromonas, Fusobacterium, and Bacteroides[11]. Many 16S rRNA gene sequencing protocols choose two or more hypervariable regions for an accurate analysis of the microbial profiles[20].

The principle of 16S rRNA gene-based approach is that bacterial-specific 16S gene sequences are useful for phylogenetic purpose[2]. The identification of microbial community members through the hypervariable regions is generally not so straight [15] to identify bacteria at strain taxonomic levels[15], due to the sequence conservation of the 16S rRNA gene[21].

Community profiling of the 16S rRNA gene is currently conducted by short-read sequencing technologies using only small fragments of the entire gene[1,2,13].

Using marker genes, it is necessary to predict the taxonomy of most microorganisms known only from the environmental sequencing by inferring a phylogenetic tree correlated with taxonomy[17].

Identifying bacterial sequences at the species level is not entirely simple as errors can arise during PCR amplification in which artifacts are then amplified confounding analyses of the real microbial sequences. Thus, during the PCR cycles processes, it may happen that some errors occur during the amplification of the sequence. These errors lead to the formation of artifacts, called chimeras, coming from two or more biological sequences not properly joined together[22]. For example, an incomplete (or a partial) fragment extension during a PCR cycle could act as a primer for the next cycle, amplifying a fragment of another sequence. This artifact fragment can in turn be amplified, generating many copies of itself. Studies have estimated ~30% of the sequences may be chimeric (Wang and Wang 1996, 1997). Many factors could lead to the formation of these chimeric sequences, such as the number of PCR cycles[22].

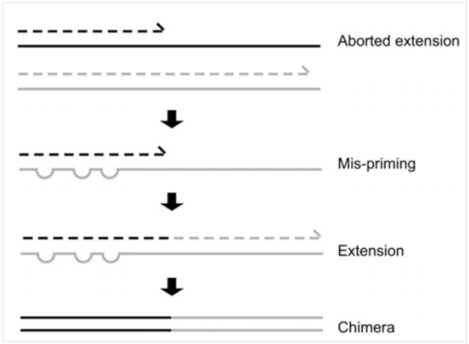


Figure 3. Representation of a chimera formation. Image from Haas et al., 2011[22].

Identified chimeric sequences must be removed in order to cluster 16S rRNA gene sequenced and build a phylogenetic tree[5] according to various 16S identification tools[21], like Ribosomal Database Project[23], Greengenes[24], SILVA[25], eHOMD[26]. Despite this, technical issues exist related to a necessity of database updates, and specific tools[2].

There are two different approach able to cluster the 16S rRNA sequences: one is a homology-based approach that exploits available databases, while the second method clusters the sequences based on their similarity between a representative sequence of the cluster (centroid) and any sequence in the cluster that have a similarity threshold higher than desired[8]. The homology-based approach assigns to the sequence the closest microbial species, comparing target sequences with specific databases[2,8]. Unfortunately, this approach is limited in distinguishing between closely related bacteria sequences, so it is difficult to find novel species[8,21] and not aligned bacteria's reads are so not identified[2].

1.2.1 Taxonomic Databases

Sequencing the ribosomal RNA gene allows to identify and detect the microbial composition of a specific environment. The taxonomic assignment of microbial sequences can be performed using specialized databases like SILVA, Greengenes, RDP, and HOMD[25]. RDP and Greengenes cover Archaea and Bacteria domains for small subunit rRNA gene (SSU) sequences, whereas SILVA database also includes Eukaryota domain[25]. Indeed, SILVA (from Latin silva, forest, *http://www.arb-silva.de*)

is a manually curated database for all three domains of life (Bacteria, Archaea and Eukaryota domains) based on the phylogenesis for the small and the large subunit rRNA genes[27]. This database contains the full-length SSU gene sequences (>1200 bp), currently containing more than 1,719,541 bacterial 16S rRNA sequences, clustered at 99% identity level[27].

Other 16S rRNA gene databases are the Ribosomal Database Project (RDP) [23] and Greengenes[24] database. The last one was specifically created for full-length 16S rRNA genes, allowing a de novo tree inferenced taxonomic classification[2], but many reads and the experimentally collected 16S sequences are not classified[2,28].

The Human Oral Microbiome Database (HOMD)[26] is an important 16S rRNA database containing information relative to more than 800 microbial taxa of the human aerodigestive tract[13], with the aim to provide a tool for the identification of the 16S rRNA gene oral taxa[14], describing information linked to oral species.

Molecular methods are used to identify novel species and can cluster the sequences into phylotypes, or taxa based on their 16S rRNA sequence[14]. A phylotype is defined by the sequence similarity using 98.5% of similarity threshold[14], that defines almost species level clusters for oral bacteria, assigning then a Human Oral Taxon (HOT) ID number[14].

HOMD database was then expanded to collect bacterial communities present in the oral cavity, pharynx, nasal passages, esophagus[11].

The expanded Human Oral Microbiome Database (eHOMD)[26] collects information from the entire human aerodigestive tract, containing 16S rRNA gene reference sequences and allowing species-level taxonomy assignment to most of the sequences derived from different sites in the aerodigestive tract[13].

In addition to the complexity in assigning taxonomy to 16S rRNA gene sequences, the choice of the variable region in microbiome study may impact on the phylogenetic resolution[13].

1.2.2 Operational Taxonomic Units and Amplicon Sequence Variants

Analyses of marker-gene, like 16S rRNA gene, are based on the construction of Operational Taxonomic Units (OTUs) that cluster the molecular sequence of the target gene by sequence similarity, given an identity threshold of 97%[29].

The OTU clustering compares sequences using different alignment algorithms against a reference database for the taxonomic assignment[2].

There are at least three different strategies to build OTU clusters: de novo OTU picking, closed-reference OTU picking and open-reference OTU picking[28].

In the de novo clustering approach sequences are clustered according to their sequence similarity, allowing ~3% of sequence dissimilarity. Here, no reference databases is used. In reference-based OTU clustering approach a reference database is exploited to associate taxonomic labels to the target sequences looking at known available taxa in the database[28,29].

There is also another method to assign the taxonomy to sequenced amplicons, that resolves the exact sequence of a sample recording the time in which the sequence was read[29]. This approach uses the amplicon sequence variants (ASVs) that do not require a dissimilarity threshold, resolve a sequence down to the level of single-nucleotide differences over the sequenced gene region[1].

ASVs could replace OTU approaches inferring the exact sample sequence providing more specificity and sensitivity in identification of microbes and pointing out biodiversity among individuals[1].

1.2.3 Divisive Amplicon Denoising Algorithm

Amplicon sequencing of marker genes, like 18S rRNA gene, 16S rRNA gene or the internal transcribed spacer (ITS), region is the most common method providing information regarding microbial communities[30,31].

There are many bioinformatics tools able to perform data evaluation of 16S rRNA gene sequences. Some of these are Mothur, QIIME2, and DADA2[1].

QIIME2, for example, is one of the most widely used tool in community recognition that is able to filter out errors in Illumina-sequenced amplicon data and cluster amplicon sequences into operational taxonomic units (OTUs)[30].

Another widely used tool for 16S analysis is the Divisive Amplicon Denoising Algorithm 2 (DADA2)[30], an open-source R package built to infer and correct Illumina amplicon errors constructing ASVs[2,30,31], and recognize the "true" (or real) sequences of a sample. DADA2 models the relationship between the error probability and the quality score for each single base of the read. Indeed, the strength of this tool is that it's able to evaluate and correct sequences with a single nucleotide resolution.

There are many steps involved in the analysis of 16S sequences as filtering and trimming sequences at certain length and removing sequence bases carrying a quality lower than a threshold. Dereplication step that grouped amplicons with the same nucleotide sequence into unique amplicon sequences, memorizing the number of times (abundance) the sequence appears in the analyzed sample[30]. The Denoising step is used to build and infer a model to correct Illumina sequence errors, and the chimera identification by performing a global alignment of each sequence against the more abundant sequences in the set of sequences. DADA2 looks for half of the left sequence and half of the right sequence from which the chimeric sequence was generated[30,31].

This method offers the advantage of being very sensitive and precise, and it is also a widely used tool given the low costs of amplicon sequences. However, amplicon sequencing provides information to a single genetic locus since amplification of sequences reaches 100-500 nucleotides in length (short-reads sequencing)[31] but gives more accurate bacteria community profiles[2].

Aim of the project

In this elaborate we presented a bioinformatic analysis of the 16S rRNA gene performed in two different ways: investigating the pooled 16S rRNA gene amplicon regions, then analyzing 16S amplicon regions singularly.

The aim of the work was to understand what the advantages in sequencing were and analyzing a pool set of sequences belonging to 6 overlapping amplicon regions of 16S rRNA gene rather than working with some selected 16S amplicon regions.

On these analyzed data we then created a panel of reference data, on which we have developed a simulator to mimic an oral microbiota environment reproducing a 16S rRNA gene-based amplicon sequencing, pointing out critical issues that could be encountered during an amplicon analysis, since only certain amplicon regions were selected to perform a taxonomical analysis. Moreover, we investigated those cases where taxonomy encountered difficulties in assigning detailed taxa to some simulated amplicon regions.

2 METHODS AND MATERIALS

2.1 Setting up reference panel of oral microbiome

Oral swabs species were sampled from 4 healthy individuals. Molecular analysis was conducted by sequencing the 9 hypervariable regions (6 overlapping amplicons) of the 16S rRNA gene[21]. Through bioinformatics, the 16S rRNA gene variants were screened to target the largest number of bacterial species in the oral cavity of the 4 individuals. This was done to get a realistic picture for the bacterial distribution that was used as reference panels for the following simulation study.

We developed an amplicon simulator, based on a bacterial reference panel, to mimic the microbiome composition in the oral environment at molecular level. This was used to test the effectiveness of set of standard primers in binding the target sequence to be amplified, and to test the overall accuracy of the pre-trained databases to assign taxonomic profiles.

2.1.1 Oral microbiota typesetting of 4 healthy individuals

This elaborate deals with already available sequence data of the oral microbiome that were prepared (collection of oral samples, DNA extraction, library preparation and the sequencing) in the laboratories of the University of Verona. Briefly, oral swab samples from 4 healthy individuals were collected (Prof. De Santis and Dr. Luciano, Department of Surgery, Dentistry, Paediatrics and Gynaecology, University of Verona) and DNA was extracted using standard methods (Dr. Patuzzo, Department of Neurosciences, Biomedicine and Movement Sciences, University of Verona). DNA paired-end (PE) libraries of the 16S rRNA gene were prepared according to the QIAseq 16S ITS panel handbook (www.qiagen.com, QIAseq® 16S/ITS Panel Handbook, 2021) and sequenced on an Illumina MiSeq NGS platform (Technology Platform Centre (CPT) of the University of Verona; Personal Genomics Srl, Verona).

The 4 oral samples were named ID1, ID2, ID3, ID4. The sample ID4 was handled and sequenced 2 times (ID4a and ID4bc, respectively). The protocol leads to the

production of 6 consecutive 16S rRNA amplicon regions. Amplicons were prepared into 3 distinct pools. The "pool 1" contained amplicons V1V2, V4V5, and fungal Internal Transcribed Spacer (ITS) region, "pool 2" contained V2V3 and V5V7 amplicons, and "pool 3" contained amplicons V3V4 and V7V9. In this thesis ITS regions were not included in the study. Figure 4 shows a schematic view of the library preparation.

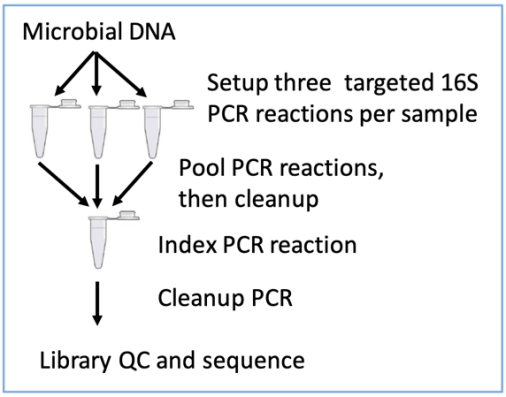


Figure 4. Library preparation of 16S rRNA gene regions (Image from *QIAseq ® 16S/ITS Panel Handbook*, 2021)

2.1.2 Preparation of fastQ sequence files for every individual and 16S rRNA amplicon region

Sequencing data underwent a demultiplexing process, in which the data are split into several fastQ files. For each sample, 2 fastQ files were generated, containing the sequences of the pooled amplicon regions. FastQ files having only sequences belonging to the same amplicon region were prepared from the fastQ files with the pooled amplicon regions. Bioinformatic pipeline (figure 5) consisted of the following steps: 1) primer-based amplicon sequence extraction from the pooled amplicon sequences, 2) the removal of primer sequences using cutadapt program able to snip the primer sequences both at 5' and 3' terminals, 3) calling of amplicon sequence variants (ASV) through the DADA2 R package (details are reported below), 4) arrangement of ASVs into an OTU table data structure, reporting the abundance and the sequence of each analyzed amplicon sequence variant.

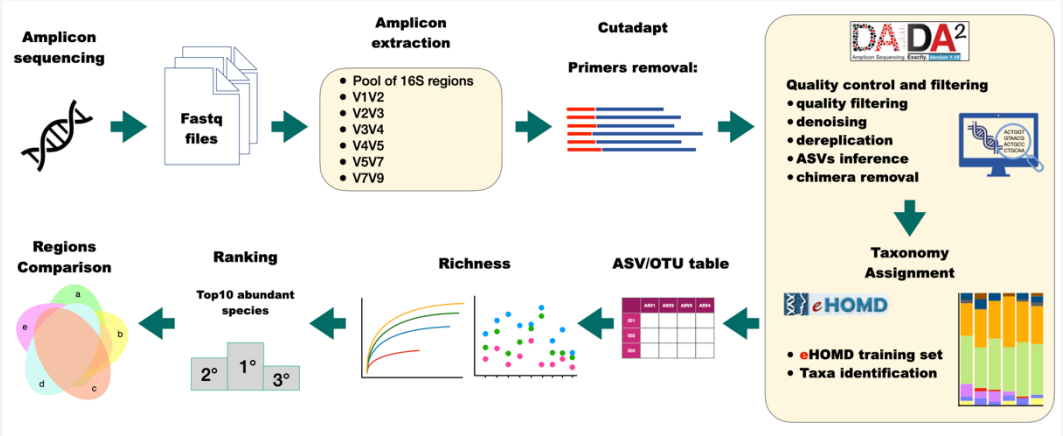


Figure 5. 16S rRNA gene amplicon analysis: a human oral microbiota workflow. 16S rRNA gene amplicon sequencing returned fastQ files containing the sequenced reads and the associated quality for each single nucleotide. Then the bioinformatic pipeline started. We subdivided pooled 16S rRNA gene regions contained in fastQ file into distinct fastQ files; then, we removed the primer sequences from both the sequence ends. DADA2 pipeline started involving many different steps (yellow box). At the end, we built an OTU table and evaluated the individual's bacterial richness. To conclude, we ranked the first top 10 species among all the study conditions, and we compared the efficacy of bacteria detection in different datasets (pooled-Vs versus single amplicon regions).

To capture most of 16S gene variability, the study workflow provides either the analysis of the pooled amplicon regions or the analysis of the single individual amplicon regions. In this way, it was possible to compare the two methods and to understand if the analysis of the pooled 16S rRNA gene amplicon is advantageous in identifying a greater amount of the oral bacteria than the single amplicon region analysis.

2.1.3 DADA2 analysis

Data sequences of each sample were stored into 2 fastQ files: one containing all the forward reads (e.g., the head part of fragment is read on forward strand), and the other one containing the reverse reads (e.g., the tail part of fragment is read on forward strand). FastQ files containing the pooled sequences of all the 6 amplicon regions were named "pooled-Vs", indicating that samples were analyzed with all the amplicon at the same time. From every "pooled-Vs" fastQ files (forward and reverse), 6 fastQ files, each reporting the sequences according to sequence of the primer regions, were prepared, clustering the sequences according to sequence of the primer regions.
The 6x2 fastQ files containing sequences of individual amplicon regions (for every sample) were prepared using a bash script specifically designed for this purpose.

The "degenerate" sequences of the primers were identified and used to catalog the amplicon sequences. Overall, 7x2 fastQ files for each of the 5 oral samples were analyzed, 6 for the individual amplicons, and 1 for the pooled 16S regions (pooled-Vs).

Cutadapt program, using a pattern matching search for degenerate sequence, was used to remove primer sequences from DNA fragments. Primer sequences were searched only at the head terminal (first 14 nt) or at the tail terminal (last 14 nt) of forward or reverse mate sequences, respectively. Below, there is a short example of cutadapt syntax usage.

```
cutadapt -g XN{14} primerForward -G XN{14} primerReverse
```

The pre-processed reads were analyzed using the Divisive Amplicon Denoising Algorithm (DADA2)[30], identifying for each sample a set of Amplicon Sequence Variants (ASVs). We analyzed data in two different ways: first we investigated the microbiome composition among the pooled amplicon regions (pooled-Vs), then we examined the microbial community of amplicon regions singularly.

The first step was removing the low quality sequenced based, keeping the mean average almost to 30Q (Phred Score). We performed the truncation of the reads at 245bp for all the individual's forward reads and at 190bp for all the individual's reverse reads. In order to have reads long enough to perform the analyses, the minimum length parameter was set to 160 bp; reads shorter than this value was discarded. These settings were maintained for the pooled-Vs analysis and for the analysis of each amplicon region, except for V3V4 amplicon regions for which truncation occurred at 245bp for both forward and reverse reads. This step is one of the critical parts of the entire DADA2 pipeline, because we have to maintain reads long enough and with a good quality score so as not to lose the overlap between pairs.

The denoising step followed quality control. DADA2 infers samples composition, applying a statistical sequencing error correction. Then, the dereplication step can occur. Since the microbiome contains many copies of the same 16S rRNA gene sequence belonging to a specific bacterium, DADA2 collapses the data into unique sequences in a process called dereplication, memorizing the number of times a sequence is seen for a given sample, reducing sequence redundancy, and saving

computational memory for later steps. Ideally, the collection of unique sequences (ASVs) represents the number of bacterial species residing in the oral samples.

The subsequent merging step of the forward and reverse unique reads provided the formation of the ASVs. DADA2 merging function converts reverse sequences into proper complement reverse strands, to exactly overlap the forward 3' end. As general rule, at least 12 bases need perfectly match without any errors; if not, the merging doesn't occur, discarding a potential amplicon to analyze, therefore a bacterium to identify. Figure 6 shows an example of overlap by aligning two reads using Clustal Omega[34].

```
CLUSTAL O(1.2.4) multiple sequence alignment

F           CTATTGTTAGTTGCCATCATTCAGTTGGGCACTCTAGCGAGACTGCCGGTAATAAACCGG        60
merged      CTATTGTTAGTTGCCATCATTCAGTTGGGCACTCTAGCGAGACTGCCGGTAATAAACCGG        60
R           ------------------------------------------------------------        0

F           AGGAAGGTGGGGATGACGTCAAATCATCATGCCCCTTATGACCTGGGCTACACACGTGCT        120
merged      AGGAAGGTGGGGATGACGTCAAATCATCATGCCCCTTATGACCTGGGCTACACACGTGCT        120
R           ------------------------------------------------------------        0

F           ACAATGGCTGGTACAACGAGTCGCAAGCCGGTGACGGCAAGCTAATCTCTTAAAGCCAGT        180
merged      ACAATGGCTGGTACAACGAGTCGCAAGCCGGTGACGGCAAGCTAATCTCTTAAAGCCAGT        180
R           ------------------------------------------------------------        0

F           CTCAGTTCGGATTGTAGGCTGCAACTCGCCTACATGAAGTCGGAATCGCTAGTAATCGCG        240
merged      CTCAGTTCGGATTGTAGGCTGCAACTCGCCTACATGAAGTCGGAATCGCTAGTAATCGCG        240
R           ------TCGGATTGTAGGCTGCAACTCGCCTACATGAAGTCGGAATCGCTAGTAATCGCG        54
                  ****************************************************

F           GATCA-------------------------------------------------------        245
merged      GATCAGCACGCCGCGGTGAATACGTTCCCGGGCCTTGTACACACCGCCCGTCACACCACG        300
R           GATCAGCACGCCGCGGTGAATACGTTCCCGGGCCTTGTACACACCGCCCGTCACACCACG        114
            *****

F           ------------------------------------------------------------        245
merged      AGAGTTTGTAACACCCGAAGTCGGTGAGGTAACCGTAAGGAGCCAGCCGCCTAAGGTGGG        360
R           AGAGTTTGTAACACCCGAAGTCGGTGAGGTAACCGTAAGGAGCCAGCCGCCTAAGGTGGG        174

F           ----------------    245
merged      ATAGATGATTGGGGTG    376
R           ATAGATGATTGGGGTG    190
```

Figure 6. Example of overlapping read amplicons. Figure shows the perfect overlap between read forward and read reverse resulting in a merged final sequence, thus an amplicon sequence variant (ASV).

All the steps described above were carried out for all 7 datasets that we produced (pooled-Vs, V1V2, V2V3, V3V4, V4V5, V5V7, V7V9).

2.1.4 Chimera detection and removal

Collected information were joined into a single table, containing the identified unique ASVs belonging to all samples, linked to the information of the absolute abundance for each individual sample, so the number of times that a sequence was seen. This table is generally called OTU table.
Once the OTU table was built, DADA2 must look for sequence artifacts generated during the PCR amplification and distinguish them from real biological sequences.
The "*removeBimeraDenovo*" function can detect and remove chimeras as soon as it identifies a sequence that presents the left segment and the right segment as identical parts of two or more abundant sequences, called parent sequences. We adopted the "consensus" option of this function in which the samples are independently checked for bimeras, and then the system makes a consensus decision.

Once the data were correctly filtered, a rarefaction curve was then performed to evaluate the sequencing performance in relation to the amount of ASVs found per sample. In other words, we could see roughly whether the library size and the sequencing coverage per sample was sufficient to detect all (or most) of the ASVs in a sample. The rarefaction curve is followed by the alpha diversity analysis to investigate the individual bacterial richness.

2.1.5 Taxonomic classification

Once we removed unreal sequences (chimeras) from all the datasets, we assigned identification labels to recognize the bacterial sequences.
The taxonomic classification labeled each ASV in its proper taxa. The classification is assigned according to 7 ranks (Kingdom, Phylum, Class, Order, Family, Genus, and Species). DADA2 tool analyzed the sequences at single nucleotide resolution, in order to assign the species label to an ASV with a high confidence. So, the assignment algorithm classified sequences according to a Bayesian Classifier, trying to reach an identity greater than 97% or 100%, thus, assigning the "Species" to the ASVs under examination.

The classification step was performed using a human oral 16S rRNA gene database[26]. The extended Human Oral Microbiota Database[26] (*http://www.ehomd.org*) is a database that is useful to study the microbiome of aerodigestive tract body sites in human in health and disease. These sequences have a unique assigned number (human microbial taxon), that can be used for searching sequences in other databases[13]. For our work we used the pre-trained classifier eHOMD v15.1 (described in F Escapa et al., 2020)[32] contained into the extended Human Oral Microbiota Database (eHOMD). This training set is a multi-fasta files containing ~223144 sequences of full length 16S rRNA gene. For the taxonomic assignment two files were used: the first one (eHOMDv15.1_FL_Compilation_TS.fa.gz) was used to assign the taxonomy down to the "Genus" rank, the second pre-trained set (eHOMDv15.1_FL_Compilation_TS_ID_GB.fa.gz) contained a unique ID associated with the "Species", used to assign the most detailed rank to the ASVs sequences if 100% sequence identity is achieved.

The taxonomic classification step returned an R object called phyloseq[33], which links all the information obtained from the steps above. In this object we could explore a taxonomy table composed by 7 columns, corresponding to the 7 taxonomic ranks, and N rows, corresponding to the number of unique ASVs identified in a dataset; then, we could explore the OTU table that present the absolute abundance of the ASVs associated with each sample; the metadata table was also available, containing information regarding the individuals; finally, a list containing the ASVs sequences.

2.1.6 Classification on the amplicon regions at the "Genus" and "Species" rank

We assigned the taxonomic labels to all the 7 datasets analyzed so far. To bring out the most abundant species of each studied dataset (pooled-Vs and all single amplicon regions) we made a ranking of the identified species, from the most abundant to the least abundant. The ranking step allowed to evaluate the ability of the method to have a good detectability accuracy. We evaluated the first 10 most abundant species belonging to the analysis performed with the pooled 16S rRNA gene amplicon regions (pooled-Vs), getting the average of the bacteria species across the 5 samples. We performed the same analysis for the single amplicon regions checking whether the top 10 detected species reflected the same trend as the pooled-Vs dataset.

Each amplicon was compared to the others to investigate how many species were identified by each 16S amplicon region, and to identify species detected by only a single amplicon region. Moreover, we searched whether some bacterial species were identified only in the pooled-Vs and not in single amplicon regions, or vice versa.

Moreover, a heatmap was built depicting all the genera found in the analyses of the single amplicons. The heatmap described the presence or absence of a certain bacterial "Genus" in the given amplicon regions. These comparisons were performed using R programming language, developing some script able to fulfill the request.

2.2 Simulation of an oral microbiota environment based on eHOMD 16S rRNA gene sequences

We proposed a method to simulate a microbiome environment to bring out critical issues from the analysis of 16S rRNA gene sequencing, and to test the capability of the current database to assign the correct taxonomic label to the sequenced 16S rRNA gene fragment.

Figure 7 shows the adopted workflow to perform the 16S simulation.

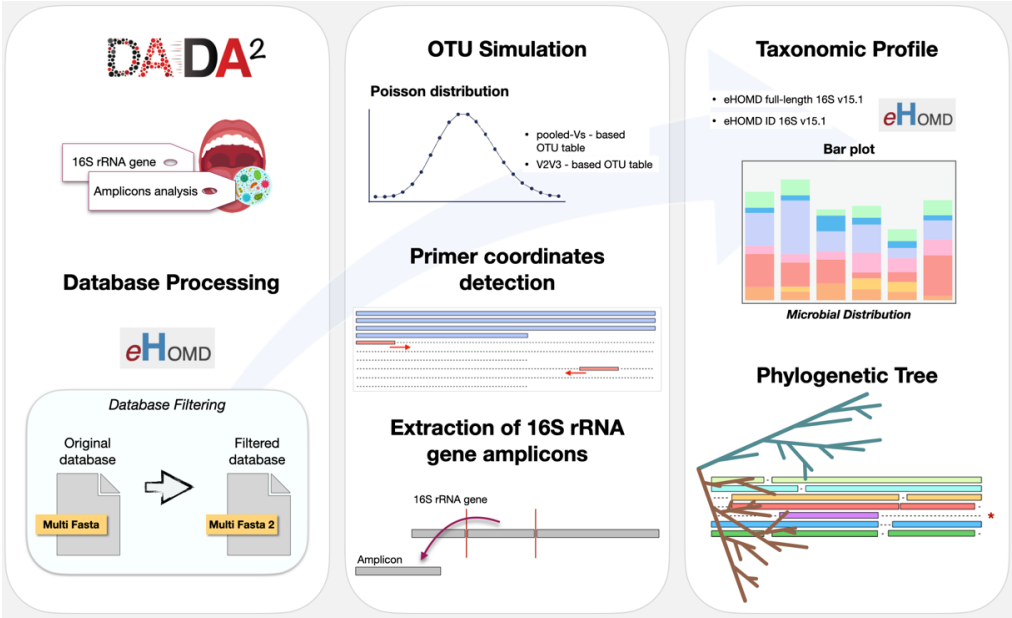


Figure 7. 16S rRNA gene simulation workflow. Starting point of the simulation (left box): microbiome analysis of oral samples using DADA2, and eHOMD database processing, in which duplicated sequences were removed. OTU table simulation and amplicon extraction (central box): two simulated OTU tables were created using counts from pooled-Vs DADA2 analysis and from V2V3 amplicon. Then, a multiple alignment was performed to identify the amplicon sequences to extract from the 16S full-length sequence. At the end, the amplicon extraction takes place. Taxonomy and phylogenesis (right box): taxonomic classification of simulated amplicons and phylogenetic tree construction to identify taxonomic mis-assignment.

We started the simulation of the microbiome environment by taking as reference the count table (OTU table) resulted from the analysis of the pooled 16S rRNA gene amplicon regions (pooled-Vs dataset), and subsequently, we fixed the process performing another simulation taking as reference the OTU table of the V2V3 amplicon region of the DADA2 analysis, the best 16S fragment in the identification of bacterial species in our study.

2.2.1 Removal of redundant 16S rRNA gene sequences from eHOMD

The eHOMD v15.1 training set was adopted to start the simulation. We chose training sets formatted according to DADA2 requirements for the taxonomic assignment function usage, containing 223144 full-length 16S rRNA gene sequences as described in F Escapa et al., 2020[32]. Two file formats were available: the first one (eHOMDv15.1_FL_Compilation_TS_16SrRNA_full-length_trainingset.fa) is generally used to assign the taxonomy according to the Bayesian Classifier algorithm along multiple taxa ranks, from Kingdom to Genus. For each bacterial sequence the multi-fasta file presents a header line, containing taxonomical information divided into 7 fields (Kingdom, Phylum, Class, Order, Family, Genus, Species) followed by the sequence line. The second file is used for the species taxonomical assignment (eHOMDv15.1_FL_Compilation_TS_ID_GB.fa) that is thought to assign species labels at or near the resolution limit of the adopted marker gene. This file presents a line code for each header sequence line (e.g., TS106522), followed by a NCBI identifier, used to look for bacteria in other databases; the header line continues with the same string as the previous file with 7 taxonomic ranks. Some of the 16S rRNA gene reference sequences bear the same NCBI ID and therefore the same taxonomic classification, as reported in figure 8.

```
>TS106522|DQ346435.2 Bacteria;Firmicutes;Bacilli;Lactobacillales;Streptococcaceae;Streptococcus;sp.HMT423

CTCAGGACGAACGCTGGCGGCGTGCCTAATACATGCAAGTAGAACGCTGAAGGAGGAGCTTGCTTCTCTGGATGAGTTGCGAACGGGTGAGTAACGCGTAGGTAACCTGCCTGGTAGCGGG
GGATAACTATTGGAAACGATAGCTAATACCGCATAACAGTAGATGTTGCATGACATTTGCTTAAAAGGTGCAATTGCATCACTACCAGATGGACCTGCGTTGTATTAGCTAGTTGGTGGGG
TAACGGCTCACCAAGGCGACGATACATAGCCGACCTGAGAGGGTGATCGGCCACACTGGGACTGAGACACGGCCCAGACTCCTACGGGAGGCAGCAGTAGGGAATCTTCGGCAATGGACGG
AAGTCTGACCGAGCAACGCCGCGTGAGTGAAGAAGGTTTTCGGATCGTAAAGCTCTGTTGTAAGAGAAGAACGAGTGTGAGAGTGGAAAGTTCACACTGTGACGGTATCTTACCAGAAAGG
GACGGCTAACTACGTGCCAGCAGCCGCGGTAATACGTAGGTCCCGAGCGTTGTCCGGATTTATTGGGCGTAAAGCGAGCGCAGGCGGTTAGATAAGTCTGAAGTTAAAGGCTGTGGCTTAA
CCATAGTACGCTTTGGAAACTGTTTAACTTGAGTGCAAGAGGGGAGAGTGGAATTCCATGTGTAGCGGTGAAATGCGTAGATATATGGAGGAACACCGGTGGCGAAAGCGGCTCTCTGGCT
TGTAACTGACGCTGAGGCTCGAAAGCGTGGGGAGCAAACAGGATTAGATACCCTGGTAGTCCACGCCGTAAACGATGAGTGCTAGGTGTTAGACCCTTTCCGGGGTTTAGTGCCGCAGCTA
ACGCATTAAGCACTCCGCCTGGGGAGTACGACCGCAAGGTTGAAACTCAAAGGAATTGACGGGGGCCCGCACAAGCGGTGGAGCATGTGGTTTAATTCGAAGCAACGCGAAGAACCTTACC
AGGTCTTGACATCCCTCTGACCTCTCTAGAGATAGAGATTTCCTTCGGGACAGAGGTGACAGGTGGTGCATGGTTGTCGTCAGCTCGTGTCGTGAGATGTTGGGTTAAGTCCCGCAACGAG
CGCAACCCCTATTGTTAGTTGCCATCATTCAGTTGGGCACTCTAGCGAGACTGCCGGTAATAAACCGGAGGAAGGTGGGGATGACGTCAAATCATCATGCCCCTTATGACCTGGGCTACAC
ACGTGCTACAATGGCTGGTACAACGAGTCGCAAGCCGGTGACGGCAAGCTAATCTCTTAAAGCCAGTCTCAGTTCGGATTGTAGGCTGCAACTCGCCTACATGAAGTCGGAATCGCTAGTA
ATCGCGGATCAGCACGCCGCGGTGAATACGTTCCCGGGCCTTGTACACACCGCCCGTCACACCACGAGAGTTTGTAACACCCGAAGTCGGTGAGGTAACCGTAAGGAGCCAGCCGCCTAAG
GTGGGATAGATGA


>TS106571|DQ346435.2 Bacteria;Firmicutes;Bacilli;Lactobacillales;Streptococcaceae;Streptococcus;sp.HMT423

AGAGTTTGATCCTGGCTCAGGACGAACGCTGGCGGCGTGCCTAATACATGCAAGTAGAACGCTGAAGGAGGAGCTTGCTTCTCTGGATGAGTTGCGAACGGGTGAGTAACGCGTAGGTAAC
CTGCCTGGTAGCGGGGGATAACTATTGGAAACGATAGCTAATACCGCATAACAGTAGATGTTGCATGACATTTGCTTAAAAGGTGCAATTGCATCACTACCAGATGGACCTGCGTTGTATT
AGCTAGTTGGTGGGGTAACGGCTCACCAAGGCGACGATACATAGCCGACCTGAGAGGGTGATCGGCCACACTGGGACTGAGACACGGCCCAGACTCCTACGGGAGGCAGCAGTAGGGAATC
TTCGGCAATGGACGGAAGTCTGACCGAGCAACGCCGCGTGAGTGAAGAAGGTTTTCGGATCGTAAAGCTCTGTTGTAAGAGAAGAACGAGTGTGAGAGTGGAAAGTTCACACTGTGACGGT
ATCTTACCAGAAAGGGACGGCTAACTACGTGCCAGCAGCCGCGGTAATACGTAGGTCCCGAGCGTTGTCCGGATTTATTGGGCGTAAAGCGAGCGCAGGCGGTTAGATAAGTCTGAAGTTA
AAGGCTGTGGCTTAACCATAGTACGCTTTGGAAACTGTTTAACTTGAGTGCAAGAGGGGAGAGTGGAATTCCATGTGTAGCGGTGAAATGCGTAGATATATGGAGGAACACCGGTGGCGAA
AGCGGCTCTCTGGCTTGTAACTGACGCTGAGGCTCGAAAGCGTGGGGAGCAAACAGGATTAGATACCCTGGTAGTCCACGCCGTAAACGATGAGTGCTAGGTGTTAGACCCTTTCCGGGGT
TTAGTGCCGCAGCTAACGCATTAAGCACTCCGCCTGGGGAGTACGACCGCAAGGTTGAAACTCAAAGGAATTGACGGGGGCCCGCACAAGCGGTGGAGCATGTGGTTTAATTCGAAGCAAC
GCGAAGAACCTTACCAGGTCTTGACATCCCTCTGACCTCTCTAGAGATAGAGATTTCCTTCGGGACAGAGGTGACAGGTGGTGCATGGTTGTCGTCAGCTCGTGTCGTGAGATGTTGGGTT
AAGTCCCGCAACGAGCGCAACCCCTATTGTTAGTTGCCATCATTCAGTTGGGCACTCTAGCGAGACTGCCGGTAATAAACCGGAGGAAGGTGGGGATGACGTCAAATCATCATGCCCCTTA
TGACCTGGGCTACACACGTGCTACAATGGCTGGTACAACGAGTCGCAAGCCGGTGACGGCAAGCTAATCTCTTAAAGCCAGTCTCAGTTCGGATTGTAGGCTGCAACTCGCCTACATGAAG
TCGGAATCGCTAGTAATCGCGGATCAGCACGCCGCGGTGAATACGTTCCCGGGCCTTGTACACACCGCCCGTCACACCACGAGAGTTTGTAACACCCGAAGTCGGTGAGGTAACCGTAAGG
AGCCAGCCGCCTAAGGTGGGATAGATGATTGGGGTGAAGTCGTAACAAGGTAGCCGTAAGCTTGGATCCCGGG
```

Figure 8. Example of duplicated reference IDs. The code "DQ346435.2" is repeated for these sequences, consequently the taxonomic label. The difference of the two compared sequences stays in the different length of the sequences.

In this example (figure 8) we proposed two sequences carrying the same NCBI ID (in the eHOMD database 5 sequences represented this species). It is to note that, except for the line number, the NCBI code and the taxonomy are identical. The presence of multiple sequences, leading to the same bacterial species, is probably due to their different sequence length. Indeed, these sequences seemed to have the same middle sequence part, but the 5' and/or 3' ends of the sequences present different base pairs in length.
So, we reduced the eHOMD species ID file according to the NCBI identifier, choosing among identical sequences the one with greater length.
We obtained a database of unique full-length 16S sequences, from which to start the simulation.

2.2.2 Random selection of 16S rRNA gene sequences based on taxa frequencies from reference panels

To mimic an oral microbial environment, we performed an in-silico simulation based on count tables obtained from the previously analyzed oral samples. We exploited the OTU tables as reference, choosing two count tables derived from the computational analysis of the pooled 16S rRNA gene amplicon regions (pooled-Vs) and the computational analysis coming from the single V2V3 amplicon region. In both cases, the methods adopted were the same.
The designed method could be imagined as a "backward approach" since the selection of the sequences started from the lowest taxonomic rank (Species) and ended at the first taxonomic level (Kingdom). The real selected count tables, or ASV tables, are enclosed in R phyloseq[33], containing the information of the taxonomy, the metadata, and the ASVs themselves. All information are then tied together by an ASV identifier to track the sequence choice during our simulation.

We created 7 phyloseq objects as the number of the 7 taxonomic ranks. According to the "backward" procedure, we started at the lowest taxonomic rank. At the "Species" rank we created a subset of the original count table with the ASVs taxonomically assigned to this lowest layer. Meanwhile, we've kept track of just the isolated ASVs.

Then, we proceeded with the higher level (Genus), and we selected those sequences that reached this taxonomic rank, excluding the previously selected sequences, since they had reached a more classification detailed solution. This procedure continued up to the first taxonomic rank, the "Kingdom" level. For all the taxonomic levels, the signed "NA" sequences were presumably recognized by the higher levels, pointing out that some sequences did not reach a given taxonomic rank.

At the end of the subsampling, 7 table were created for each sample, containing sequences and the associated absolute abundance. For each sample, we worked with unique ASVs.

Following the Poisson distribution, we generated random frequencies (given a known average rate) to model the expected ASV absolute abundances, using the rpois() R function. This step was performed for each of the 7 taxonomic subsampled objects of each sample individually. The sum of the expected ASV absolute abundances should be near the observed ASV absolute abundances.

So, we built a simulated OTU table for each taxonomic level singularly, starting from the just subsampled data.

The simulation was performed by extracting the known archived full-length 16S sequences available in the reduced eHOMD training set, to test the primers binding ability in attaching to and amplifying the sampled sequence during the amplification and also to test the precision of the taxonomic assignment.

We developed an R script able to select the sequences from the eHOMD database and connect the simulated absolute abundances to the selected sequences.

How does the sequence selection work? An empty vector is generated to collect all the sequences chosen during the steps. The script read the list of bacteria subsampled during the previous steps, searching for all the respective sequences for each taxon contained in the reduced eHOMD database. So, starting from the first listed bacterium at species taxonomic level, the algorithm looked for all the archived sequences that bear that species label. If the reduced eHOMD training set presented more than one request sequence, the script draws out one random sequence between these and associates the Poisson distributed number to the selected sequence. Before continuing, the system memorized these selected sequences. The script goes ahead to the next listed bacterium. If the next taxon was the same species just examined, the selection

controlled the previously selected 16S sequences, so you don't extract the same sequences again.

This step was performed for all taxonomic levels backwards, looking to the taxa label within the database. At the end of this process, we obtained 7 table-like formatted files containing the selected full-length 16S rRNA gene sequences with the relative absolute abundances for each sample. This part was replicated both in the case of the amplicons pooled together (pooled-Vs) and in the case of the V2V3 amplicon counts alone, used as reference panels.

2.2.3 Primer-based multiple alignment of selected sequences and extraction of amplicon regions

We performed a primer-based multiple alignment to get the amplicon regions coordinates. Ideally, for each 16S rRNA gene sequence the alignment result should appear like figure 9, in which the primers covered the whole target gene.

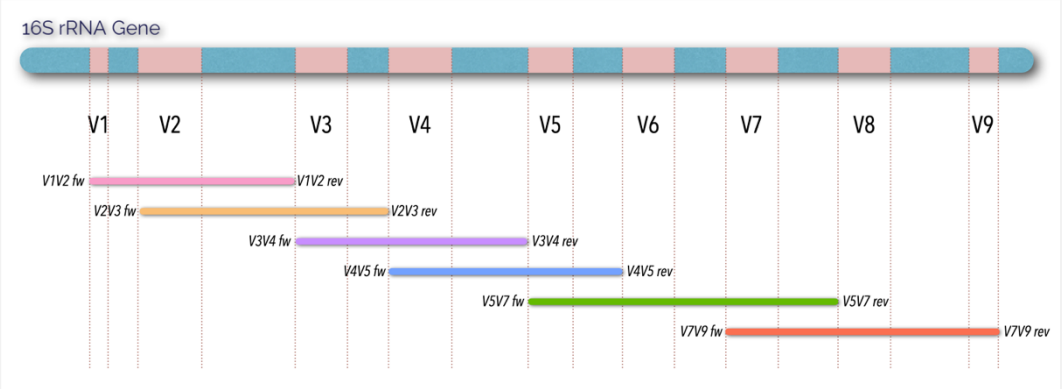


Figure 9. Representation of primers overlap. Each primer is designed in such a way as to be able to cover the entire 16S rRNA gene during sequencing.

The previously generated files were converted into fasta files. The header of these fasta files contained a label, indicating the taxon from which the sequence was taken and the relative number of that sequence of the taxon. In other words, the number reported in these files corresponded to the number of the 16S sequences extracted from the eHOMD database at the corresponding taxa.

Multiple alignment was carried out for each sequence of the fasta files, giving as output a number of alignments equal to the number of sequences for each taxon. For this purpose, we used the Clustal Omega[34] aligner from the command line, and we gave as input one sequence at a time and all the degenerated primer sequences.

To simulate a 16S rRNA gene sequencing by amplicons strategy, we extracted sequence fragments (simulated amplicon regions) outlined by the start and end coordinates of the primers aligned with Clustal Omega.
Firstly, we developed a function that could lead the degenerate nucleotide of the primer sequence back into the correct nucleotide. During this step also the simulated absolute abundance was considered.
The header of each target sequence collected in the fasta aligned file were compared to the list of headers containing only the absolute abundance relative to the sequence under examination, in order to link the abundance to the respective simulated amplicon. For each aligned sequence, we search for the coordinates of the theoretical amplification primer starting point and the ending point. Since we did not keep the primer sequence, for all primer pairs, we considered 5 nucleotides close to the sequence that should be amplified by PCR. As soon as the primer matches the target sequence, amplification can begin. In our case, we investigated whether the last 5 nucleotide bases of the forward primer and the first 5 nucleotide bases of the reverse primer align perfectly with the sequence of interest during the multiple alignment. If this happened, it was possible to derive the positions of these bases and extract the amplicon of the 16S sequence.
An exception was made for amplicons V1V2 and V7V9. During the alignment with Clustal Omega, for these two amplicons the primers might map just before region V1, or just after in the case of region V9. This leads to the failure of coordinates identification for the extraction of amplicon sequences. Thus, here we considered only one primer of the pair, the one inside the target sequence, in order to pick the sequence from that point up to the beginning of the sequence if we were looking for amplicon V1V2, or from the primer coordinate until the end of the sequence if we were looking for amplicon V7V9.

If the 5-base alignment shows mismatches, the primer does not bind the target sequence correctly; therefore, amplification cannot take place, therefore in-silico extraction also cannot happen.

Once the amplicons were created, we discarded the sequences that did not reach 200 bp of length. The filtering was not applied to the sequences of V1V2 amplicon and to the sequences of the V7V9 amplicon.

All the deleted sequences and the sequences that failed the amplicon selection were collected for statistical analyses.

For each taxonomic level we obtained 6 files as the number of 16S rRNA gene amplicons. All this procedure was adopted for each sample to be simulated.

2.2.4 Taxonomic classification

What was expected was to have unique ASVs to which to assign the taxonomy. However, during an analysis of 16S amplicons some bacterial sequences share identical portions of sequences, therefore during DADA2 process these sequences are collapsed into a unique sequence.

Taxonomy assignment analysis was performed in an R environment. We worked with one amplicon at a time for all individuals. For each amplicon, we merged all the taxa separated sequences into a single file to which we assign the taxonomy.

The sequences of each group of amplicons were dereplicated and pooled in a new count table (OTU table). In addition, we created a control table containing the dereplicated sequences that had to be taxonomically assigned and the details of the dereplication step. That is, we had the information of which sequences have merged in a single sequence.

The taxonomic assignment was performed using the pre-trained eHOMD v15.1 database through two steps: 1) we assigned the taxonomy to the sequences down to the genus level (6^{th} level), using dada2 function "*assignTaxonomy()*"; 2) we assigned the species label (7^{th} level) using dada2 function "*addSpecies()*".

The DADA2 classifier can reach the lowest taxonomic level if the target sequence perfectly matches (100% identity) only one sequence template of the database. If this condition doesn't occur, the species level cannot be reached. Moreover, the classifier

won't be able to give a solution, if more than one sequence matches perfectly with the target sequence.
Therefore, we performed again the taxonomic assignment, allowing the system to assign the label of the species level even in uncertain situations.
In a second time, we merged all the samples into a single count table, and we assigned again the taxonomy to have a general overview of the total ASVs we worked with.

We assigned multiple species labels to those sequences that did not reach the Species classification layer also taking the information about those ASVs that underwent the dereplication step, pointing out original starting sequences. We collected into a multi-fasta file all the full-length 16S rRNA gene sequences from eHOMD that corresponded to the multiple solutions given by the classifier, queuing the unsolved sequence. We performed a multiple alignment and then a phylogenetic tree based on DNA distance to identify sequence similarities between the examined sequences and to investigate how phylogenetically close our unsolved sequence was to other sequences. We used Clustal Omega[34] aligner to carry out multiple sequence alignment, then we followed a short part of the workflow suggested by Toparslan et al., 2020[35], using R.

3 RESULTS

3.1 Setting up reference panels of oral microbiome

Individual reference panels were prepared by sequencing 6 amplicon regions of the 16S rRNA gene. Sequencing was performed on DNA from oral swabs of 4 healthy individuals. Bioinformatic analyses were computed using the DADA2 R package on either pooled or individual amplicon regions. The reference panels were used to drive the simulation study and the following downstream taxonomic assignment for each *in-silico* amplicon sequence variant (ASV).

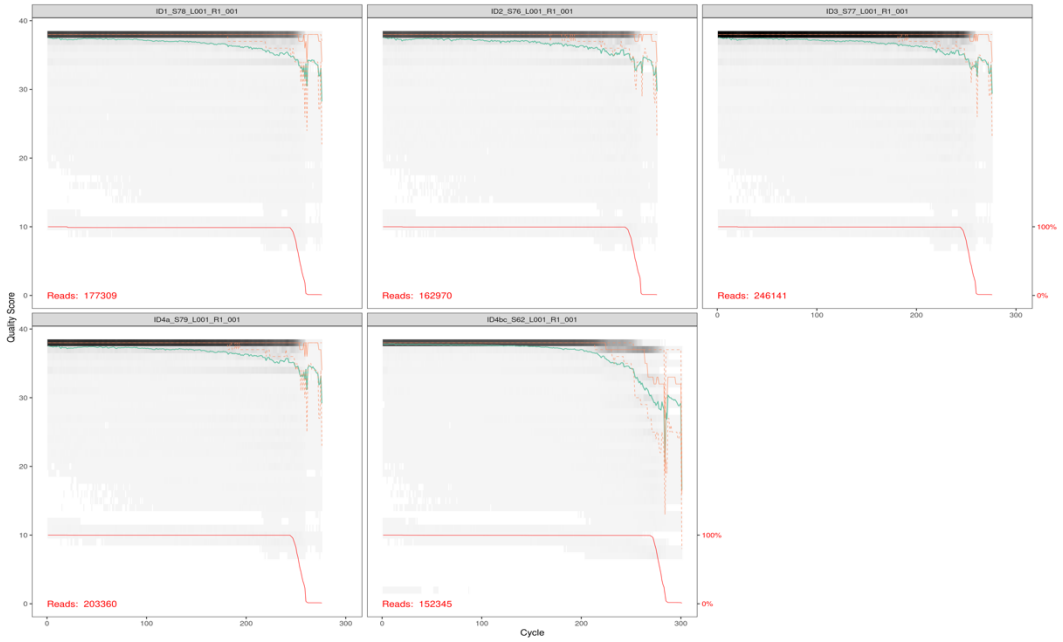
3.1.1 Preparation of fastQ sequence files for every individual and 16S rRNA amplicon region

Overall, 14 sequence fastQ files were prepared: 7x2 PE (pooled-Vs V1V2, V2V3, V3V4, V4V5, V5V7, V7V9) for each sample. Analyses were conducted on either the pooled amplicon sequences (pooled ASV analysis – pooled-Vs) or the 6 single amplicon regions (single ASV analysis for V1V2, V2V3, V3V4, V4V5, V5V7 or V7V9 amplicon regions).

3.1.2 Filtering good quality sequences for pooled and single 16S rRNA region amplicons

DADA2 removes low quality sequenced bases keeping the mean average almost to 30Q (Phred Score). Figure 10a shows the forward reads trend of the sequenced reads for each oral sample individual, whereas figure 10b shows the trend of reverse reads.

a)



b)

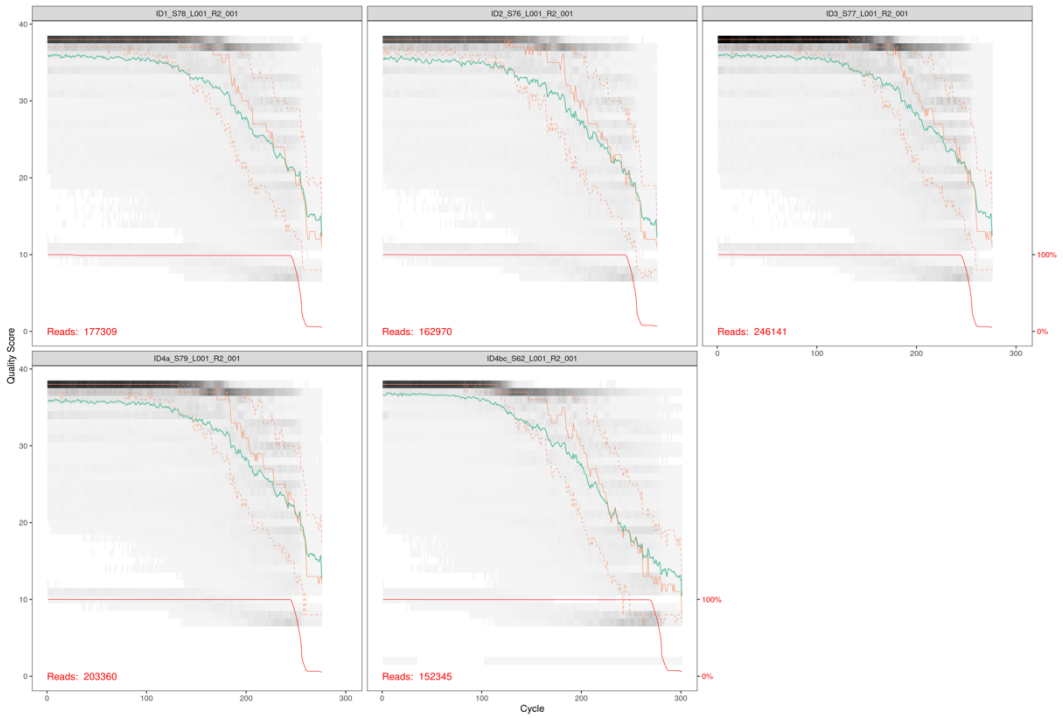


Figure 10. DADA2 nucleotide quality by position. The first part (a) shows forward read qualities in mean for each sample. The green line marks the quality median of each base of the read. In the second part of the picture (b) there is the reverse read qualities in mean per sample.

The quality control step produced reads truncated at the set parameters (245bp forward, 190bp reverse; 245bp forward and 245 bp reverse only for V3V4 amplicon regions) without performing the trimming passage. Figure 11 shows forward (a) and reverse (b) reads trend after length truncation. From the image it can be observed that at the 3' end the average quality per base remained approximately at a level of 30Q.

a)

b)

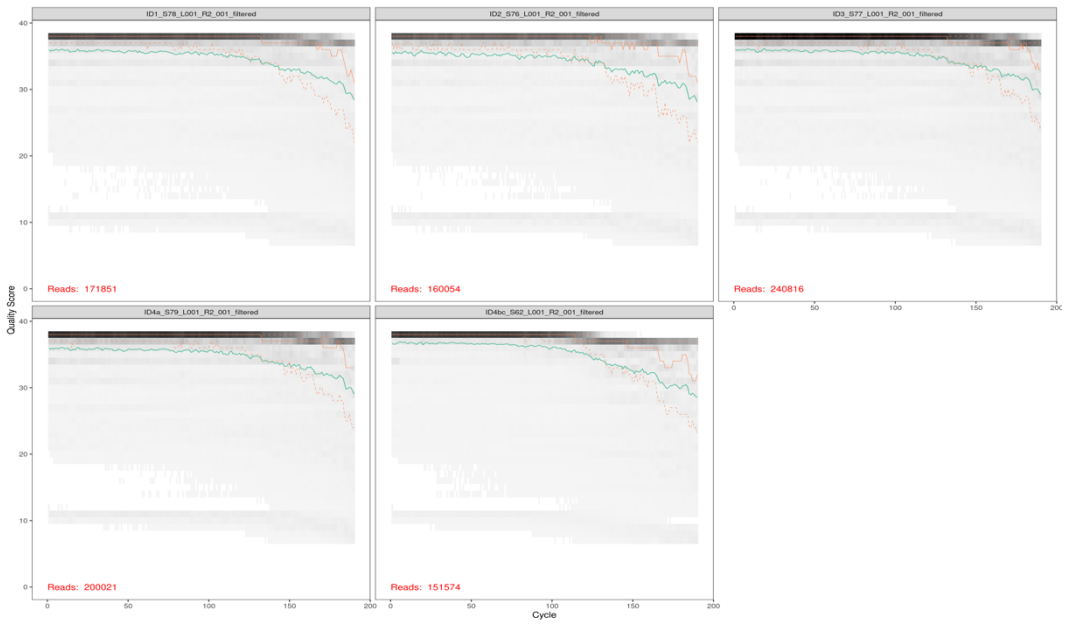


Figure 11. DADA2 truncation during quality control step. The first part (a) shows forward read qualities in mean for each sample. The green line marks the quality median of each base of the read. In the second part of the picture (b) there is the reverse read qualities in mean per sample.

3.1.3 DADA2 Analysis

DADA2 pipeline produced for each analyzed dataset a collection of unique sequences (ASVs).

Table 1 reports the number of retained reads along the DADA2 steps. Moreover, in the "truncation" column we applied the same filtering parameters for what concerns the pooled-Vs dataset and amplicon regions like V1V2, V2V3, V3V4, V4V5, V5V7, V7V9, whereas amplicon region V3V4 resulted from a different parameter setting (results in table shown as V3V4*).

Table 1. Number of retained sequences along each step of DADA2 analysis. For pooled-Vs and the V1V2, V2V3, V4V5, V5V7, and V7V9 amplicon regions, reads were truncated after 245bp (forward strand) or 190 bp (reverse strand); (*) for the V3V4 amplicon region, reads were truncated after 245bp for both forward and reverse strands.

pooled-Vs	**input**	**truncation**	**denoisedF**	**denoisedR**	**Merged**	**(%)**
ID1	177309	171851	169270	165434	140308	(79.1%)
ID2	162970	160054	158187	153647	122939	(75.4%)
ID3	246141	240816	237858	232809	196887	(80.0%)
ID4a	203360	200021	197507	193668	158740	(78.0%)
ID4bc	152345	151574	148483	145941	119887	(78.7%)
V1V2	**input**	**truncation**	**denoisedF**	**denoisedR**	**Merged**	**(%)**
ID1	20207	20195	19833	19525	18598	(92.0%)
ID2	15061	15058	14721	14531	13662	(90.7%)
ID3	26923	26919	26507	26205	25227	(93.7%)
ID4a	22555	22552	22158	22069	21140	(93.7%)
ID4bc	34034	34008	33158	32949	31504	(92.5%)
V2V3	**input**	**truncation**	**denoisedF**	**denoisedR**	**Merged**	**(%)**
ID1	36433	36419	35853	35200	33498	(91.9%)
ID2	31261	31259	30810	30103	29395	(94.0%)
ID3	52047	52039	51377	50526	49490	(95.0%)
ID4a	40608	40608	40052	39538	38106	(93.8%)
ID4bc	17022	16994	16575	16256	15598	(91.6%)
V3V4 *	**input ***	**truncation ***	**denoisedF ***	**denoisedR ***	**Merged**	**(%) ***
ID1	25066	24803	24432	23222	22506	(89.7%)
ID2	29747	29676	29402	28103	27593	(92.7%)
ID3	36562	36410	36005	34906	33546	(91.7%)
ID4a	34352	34306	33982	33040	32537	(94.7%)
ID4bc	23993	23944	23539	22774	22306	(92.9%)
V4V5	**input**	**truncation**	**denoisedF**	**denoisedR**	**Merged**	**(%)**
ID1	18447	18426	18276	17786	16532	(89.6%)
ID2	16722	16715	16495	16039	14947	(89.3%)
ID3	25195	25175	24986	24329	22526	(89.4%)
ID4a	22505	22497	22357	21880	21142	(93.9%)
ID4bc	36624	36597	36228	35403	33931	(92.6%)

V5V7	input	truncation	denoisedF	denoisedR	Merged	(%)
ID1	43844	40512	39983	38985	37961	(86.5%)
ID2	29010	26797	26549	25481	24904	(85.8%)
ID3	56534	52376	51866	50092	48919	(86.5%)
IDa	38002	35067	34705	33355	32561	(85.6%)
ID4bc	17821	17773	17391	17060	16428	(92.1%)
V7V9	**input**	**truncation**	**denoisedF**	**denoisedR**	**Merged**	**(%)**
ID1	30576	30573	30254	29597	29111	(95.2%)
ID2	39584	39582	39365	38490	38257	(96.6%)
ID3	46496	46485	46018	45356	44763	(96.2%)
ID4a	43715	43715	43385	42795	42415	(97.0%)
ID4bc	20913	20901	20509	20202	19811	(94.7%)

As we can see from the table, the remaining merged sequences were more than 85% of the sequences given in input.

3.1.4 Chimera detection and removal

All the data have been collapsed into a unique OTU table including information on chimera sequences and count of each detected ASVs, for every sample.
Table 2 shows the number of sequences (non-unique) after chimera removal.

Table 2. Number of sequences after chimera removal. The table shows for each sample the number of sequences for pooled-Vs and each single amplicon region.

	pooled-Vs	V1V2	V2V3	V3V4	V4V5	V5V7	V7V9
ID1	133224	18265	31229	22359	15995	36905	26350
ID2	114462	13619	24314	26920	14708	24020	36495
ID3	180931	24751	44699	31756	21784	46946	36923
ID4a	153780	21040	36438	31390	20585	32185	40189
ID4bc	116279	31249	15214	21739	32842	16262	18868

Overall, nearly 75% and 87% of total number of sequenced reads were retained after chimeral removal from the pooled-Vs and single amplicon regions, respectively.

In figure 12, we can see a graphical representation of the reads filtering trend for the pooled-Vs and for each 16S single amplicon regions.

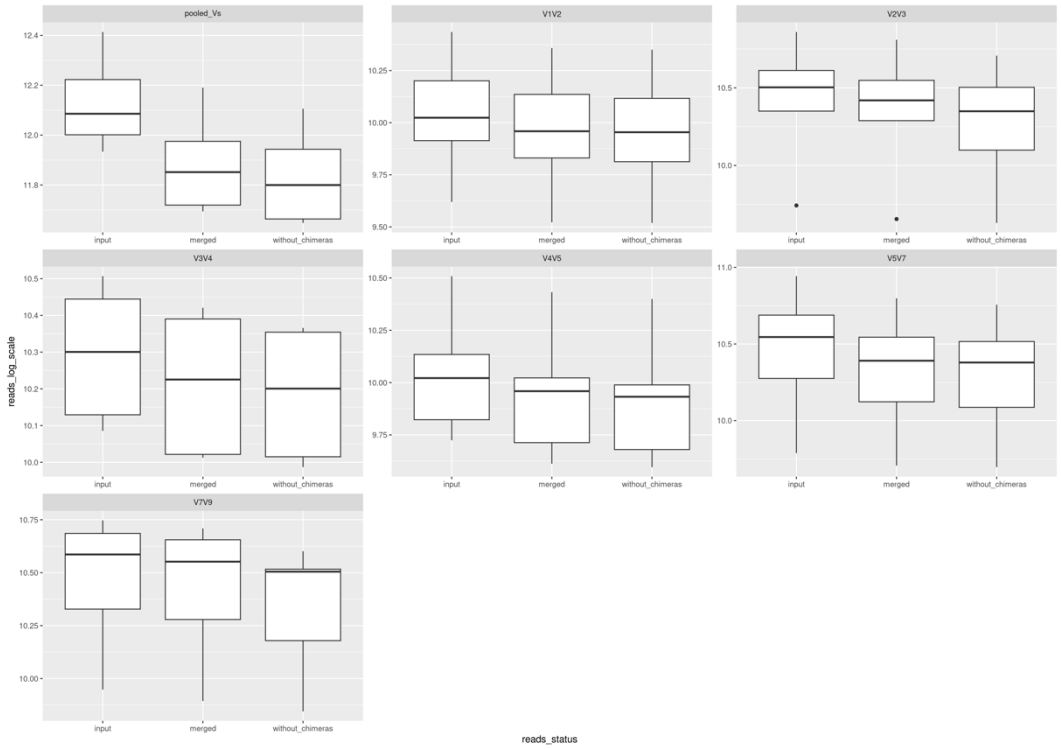


Figure 12. Average number of reads after two DADA2 read filtering steps for each amplicon region. Each frame shows three boxplots reporting the average numbers of reads and their variability before any preprocessing steps (input), during merging step (merged), and after chimera removal (without_chimeras), respectively. Y-axis reports the number of reads on logarithmic scale.

On filtered results, we investigated the trend of data rarefaction curve and then we measured the alpha diversity to calculate the individual bacterial richness based on the unique number of ASVs.

The rarefaction curve was performed to evaluate the sequencing performance in relation to the amount of ASVs found per sample (figure 13). Thus, we investigated whether the library size and the sequencing coverage per sample was sufficient to detect all (or most) of the ASVs in a sample.

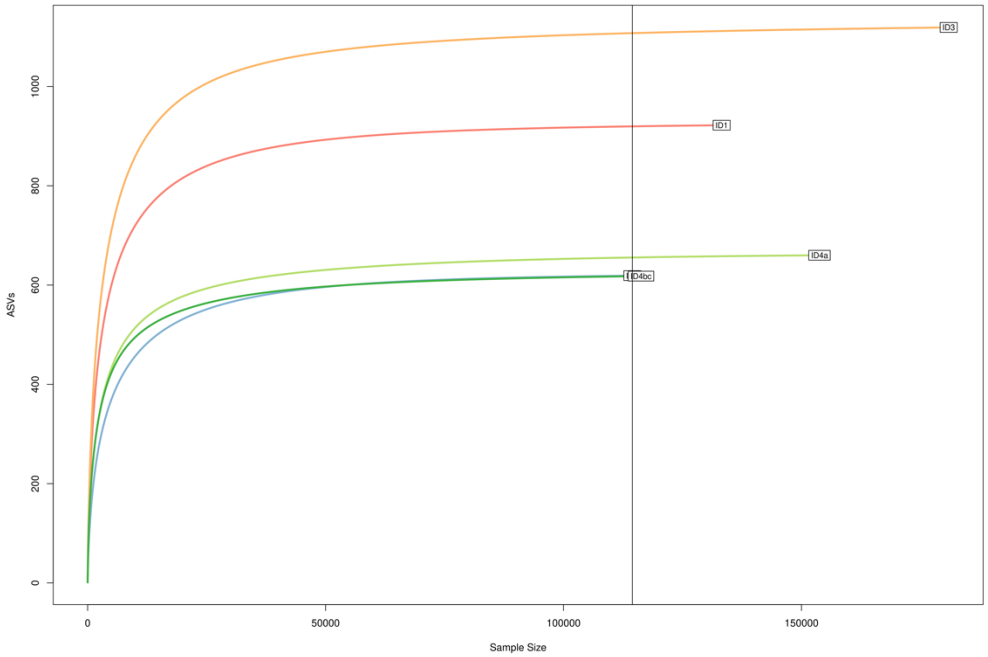


Figure 13. Rarefaction curve. Here, a representation of the alpha diversity is reported. In the x-axis the number of sequenced reads is reported, whereas the number of detected ASVs is reported in the y-axis. This picture shows how effective the sequencing was to allow the identification of the largest number of ASVs per sample. When the curve reaches a plateau, the sequenced reads (ideally) are enough to find all amplicons in that sample.

The plot shows the number of sequenced reads (x-axis) per sample related to the number of detected ASVs (y-axis) per sample. If the curve reaches the plateau, we have ideally reached the maximum number of reads in order to have the maximum read of ASVs capacity.

Then, we applied the Shannon Index metric on the unique ASVs to measure the individual's bacterial richness. We calculate the alpha diversity using the number of unique ASVs of the pooled-Vs dataset, and then on the number of unique ASVs detected in every single amplicon region separately (figure 14).

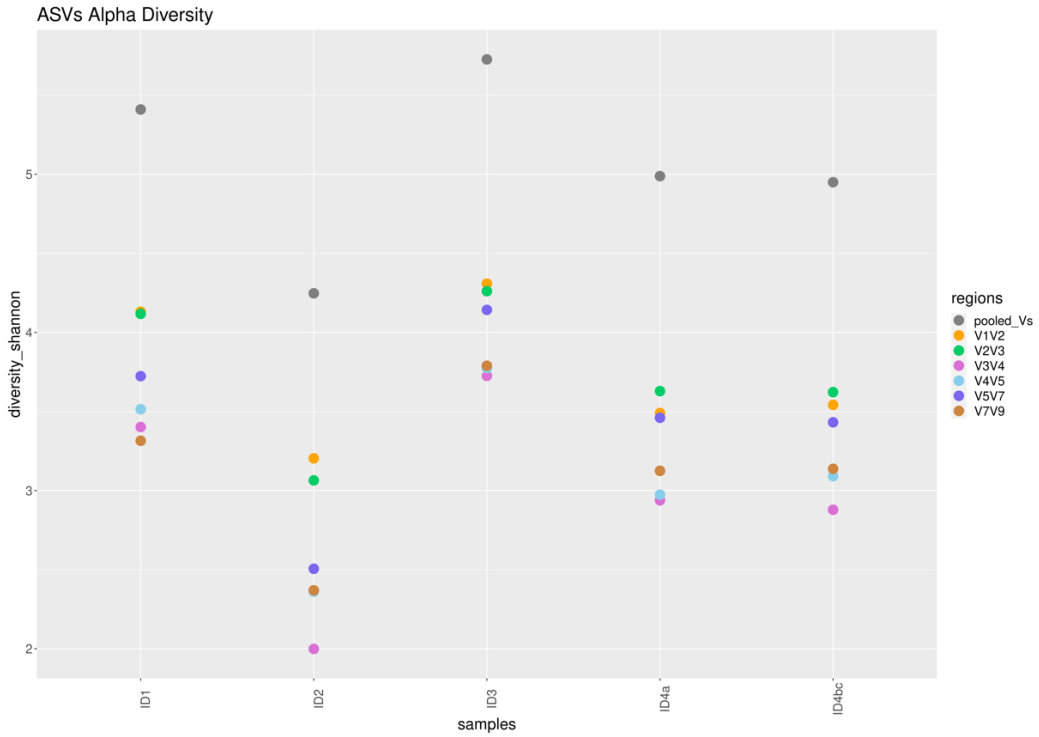


Figure 14. ASVs Alpha Diversity. The figure shows the result of the alpha diversity metric applying the Shannon Index on the number of unique ASVs using the pooled-Vs dataset and every single amplicon region. In the x-axis the 5 samples were reported, whereas in the y-axis the Shannon Index is shown. The pooled-Vs (grey dots) show the greatest biodiversity among all samples compared with the other single amplicon regions.

Figure 14 shows that for each sample the pooled-Vs (grey dots) has a higher bacterial richness (in term of number of ASVs) than the other amplicon regions considered separately.

3.1.5 Taxonomic classification

The taxonomic classification parsed the whole set of ASVs, and provided a taxonomic assignment, for all possible ranks, to every ASVs. Not all the sequences reached the lowest taxonomic rank (e.g., "species"). The information was stored into a single R object called "phyloseq".

The taxonomic classification of the 5 samples reported 2600 unique ASVs when classifying the pooled-Vs. The same analysis reported 2708 unique ASVs when summing the ASV from all the single amplicon regions.

Table 3 shows the number of successfully classified ASVs for every taxonomic level (5 oral samples).

Table 3. Number of successfully classified unique ASV at every taxon rank. The table reports the overall number of ASVs identified across the 5 samples. As an example, of the 2600 ASVs (pooled-Vs) classified at "Kingdom" rank, 1147 ASVs were classified at "Species" rank.

	Kingdom	Phylum	Class	Order	Family	Genus	Species
pooled-Vs	2600	2570	2522	2510	2492	2422	1147
V1V2	542	531	518	517	516	511	232
V2V3	668	661	655	652	647	638	308
V3V4	383	378	371	369	367	365	197
V4V5	307	307	300	300	299	288	133
V5V7	434	431	422	420	410	384	196
V7V9	374	372	366	361	359	343	124

The average number of ASVs reaching the "Species" taxonomic rank ranges between 33% (V7V9) and 51% (V3V4).

Table 4 shows how much the classification becomes less precise as we go in detail. More precisely, the numbers shown here represent the amount of ASVs that stop their classification at certain taxonomic level, up to the number of classified ASVs at species layer (corresponding to table 4).

Table 4. Number of ASVs without a taxonomic assignment. The table reports the number of ASVs not fulfilling a taxonomic classification further than the given taxonomic rank. As an example, 30 ASVs (pooled-Vs) were classified at the "Kingdom" rank but not at the other more detailed taxonomic ranks.

	Kingdom	Phylum	Class	Order	Family	Genus	Species
pooled-Vs	30	48	12	18	70	1275	1147
V1V2	11	13	1	1	5	279	232
V2V3	1	6	3	5	9	330	308
V3V4	5	7	2	2	2	168	197
V4V5	-	7	-	1	11	155	133
V5V7	3	9	2	10	26	188	196
V7V9	2	6	5	2	16	219	124

Table 5 shows the number of different taxa that were revealed by the ASV taxonomic assignment, at every classification rank.

Table 5. Number of recognized taxa at every taxonomic rank. The table reports the overall number of taxa identified across the 5 samples, for all the taxonomic ranks and amplicon sequence region. As an example, 204 "Species" were recognized when working with pooled-Vs.

	Kingdom	Phylum	Class	Order	Family	Genus	Species
pooled-Vs	1	9	21	35	53	87	204
V1V2	1	9	19	31	45	69	134
V2V3	1	8	16	25	39	63	135
V3V4	1	9	20	27	40	63	118
V4V5	1	7	16	27	40	63	90
V5V7	1	9	18	26	41	68	107
V7V9	1	8	16	25	39	61	94

The analysis of the pooled-Vs detected 204 species and 87 genera. The amplicon region V2V3 showed the highest rate of detection with 135 Species and 63 Genera. Combining the number of Species detected by all the single amplicon regions, 206 unique Species (90 unique Genera) were reported (data not shown).

Table 6 reports the overall abundance of the different detected "Phyla" in the oral microbiome of the 5 samples.

Table 6. The Phyla detected in the oral microbiome (5 samples). The table reports the "Phyla" and their overall estimated abundance conditional on the amplicon regions. For a given region, the relative abundance (%) was estimated by counting the number of amplicons calling a given taxon over the total number of amplicons and averaging this value over the 5 samples.

	pooled-Vs (%)	V1V2 (%)	V2V3 (%)	V3V4 (%)	V4V5 (%)	V5V7 (%)	V7V9 (%)
Absconditabacteria (SR1)	0.0008	0.02	-	0.0002	-	0.0005	-
Actinobacteria	4.6	5.2	5.4	3.8	2.9	5.7	1.2
Bacteroidetes	5.8	6.1	3	4.7	6	6	4.7
Firmicutes	29.1	30	26.2	27	27.7	32	33.4
Fusobacteria	3.5	4.4	4.1	2.8	3.2	1.9	2.2
Gracilibacteria (GN02)	0.03	0.003	0.02	0.03	0.02	0.02	0.04
Proteobacteria	56.4	53.2	60.7	61.2	59.8	53.8	58
Saccharibacteria (TM7)	0.1	0.2	0.2	0.1	-	0.1	0.1
Spirochaetes	0.07	0.1	0.03	0.05	0.1	0.07	0.07
Unclassified	0.05	0.04	0.05	0.04	-	0.0008	0.0005

The Proteobacteria and Firmicutes resulted to be the most present "Phyla", showing an estimated abundance between 53.2% (V1V2) and 61.2% (V3V4), and between 26.2% (V2V3) and 33.4% (V7V9), respectively. Some amplicons were not able to detect rare phyla. For instance, the region V4V5 did not detect several rare "Phyla" (Absconditabacteria-SR1, Saccharibacteria-TM7), that were instead detected by other amplicon regions.

3.1.6 Classification on the amplicon regions at the "Genus" and "Species" rank

In this paragraph we report the results of the analyses the were employed to investigate the ability to call the taxonomy at "Genus" or "Species" rank, by the different amplicon regions. Figure 15 reports the list of all the 90 Genera that were detected by one or more amplicon regions in at least one of the 5 samples.

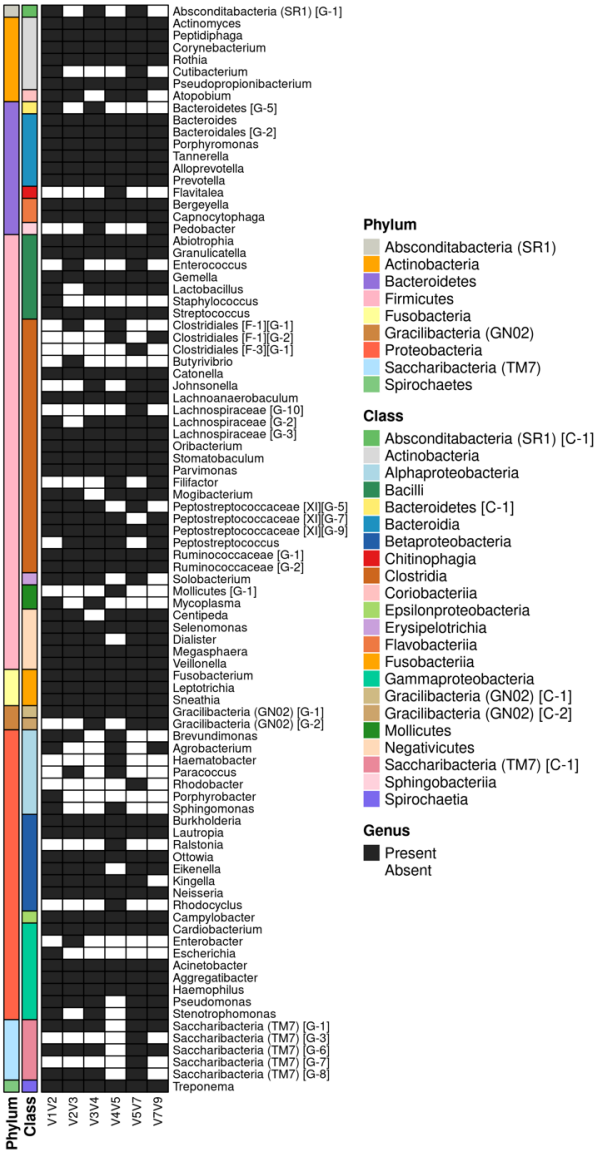


Figure 15. Heatmap of bacterial Genera identified in different amplicons. The presence or absence of a genus is highlighted by the black or white cell, respectively. Of note, most genera were detected by all the amplicons. Some genera are detected by a subset of amplicon regions. As an example, the genus Saccharibacteria (TM7)[G-8] was detected by V1V2, V2V3, V3V4, and V5V7 amplicon regions, but not by V4V5 and V7V9.

The table 7 reports the estimated relative abundance of the first ten species when using pooled-Vs over the 5 samples.

Table 7. Top 10 abundant Species (pooled-Vs). The table reports the frequency of the most abundant "Species" detected using the pooled-Vs over the 5 oral samples. The Haemophilus parainfluenzae resulted to be the most abundant species (the result holds for all the individual amplicon regions, see supplementary table 20).

Species	Relative Abundance (%)
Haemophilus parainfluenzae	34.6
Lautropia mirabilis	4.01
Neisseria oralis	1.72
Haemophilus paraphrohaemolyticus	1.63
Streptococcus mitis	1.24
Peptidiphaga gingivicola	0.93
Streptococcus sanguinis	0.73
Abiotrophia defectiva	0.69
Prevotella melaninogenica	0.63
Veillonella sp.HMT780	0.60

We performed a comparison of all the species found during the 16S analyses. Identified species of amplicon sets were compared to each other to investigate which amplicons shared species with other 16S hypervariable regions, and to found species specifically found in one single region.

Figure 16 shows all the comparisons between amplicon sets at species taxonomic rank.

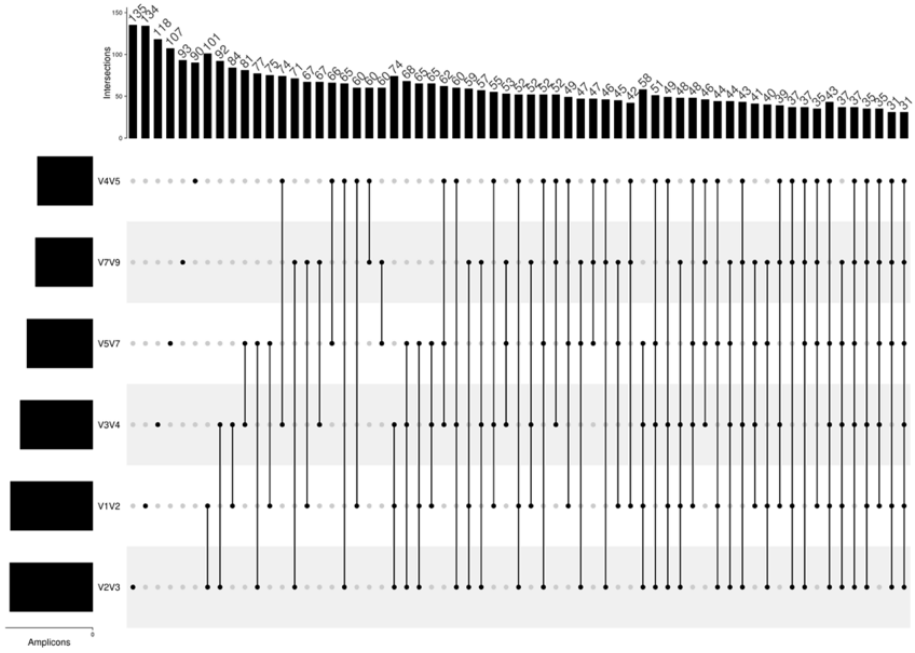


Figure 16. Species detected across the amplicon regions. The figure shows the number of species identified by every single amplicon region or a combination of them. The bar plot (top) reports the number of "Species" detected by the different single amplicon regions or a combination of two or more. Combinations of amplicon regions are indicated by the vertical bars (middle) connected by dots indicating the included amplicon regions. Horizonal bars (left) reports the numbers of "Species" detected by the single amplicons. The region V2V3 detected the highest number of "Species" (135). As additional observation, V1V2 amplicon region identified 134 different species, whereas V2V3 amplicon region detected 135 different species: the comparison between these amplicons reported a total of 101 common species; therefore, of the overall 134 species, 33 species were recognized only by V1V2 amplicon region, and 34 species were recognized only by V2V3 amplicon region.

3.2 Simulation of oral microbiota environment based on eHOMD 16S rRNA gene sequences

We performed two simulations of the oral microbiome profile of 5 samples based on the frequency of bacterial taxa at the different taxonomic rank estimated from either the pooled-Vs or the V2V3 amplicon region analysis. Both the analyses used the eHOMD v15.1 to retrieve the 16S rRNA gene sequence data used in the simulated profiles. eHOMD v15.1 was cleaned of redundant data before any simulation.

3.2.1 Removal of redundant 16S rRNA gene sequences from eHOMD

We used as reference database the eHOMD v15.1 training set, since we were working with oral microbiome data. eHOMD training set contained 223144 full-length 16S rRNA gene sequence within two formats files as described in the methods chapter. We reduced the eHOMD dataset to remove duplicate sequences that differed only in length. So, we looked for redundant sequences having the same species NCBI code, choosing the longest sequence. In this way, we obtained a new training set of eHOMD of 199970 bacterial sequences, so in general we maintained ~89.6% of the whole dataset. Using this new database, we could perform the microbiome simulation.

3.2.2 Primer-based alignment of randomly selected sequences and extraction of amplicon regions

The oral microbiota environment was *in-silico* simulated according to the estimated absolute abundances of microbe species estimated in 3.1.5. The simulation approach is described in section 2.2.2 of "Material and Methods".

Absolute abundance of simulated data was based on a random selection of sequences based on a Poisson Distribution of ASV frequency of the observed bacterial taxa (for all the ranks) among the 5 samples.

For each simulated sample, one list of sequences with their frequency at every taxonomic rank was produced.

Primers sequences (e.g., degenerate sequences and therefore multiple sequences) of all the amplicon regions were aligned against all the selected 16S rRNA gene sequences, one at a time. This step revealed the sequence coordinates to extract the amplicon sequence for each 16S rRNA gene, as it would have been done by a PCR amplification. Figure 17 shows an example of multiple alignment (Clustal Omega) of primers against one 16S sequence.

Figure 17. Example of the alignment between a selected 16S eHOMD sequence and primer sequences for amplicon regions: (Clustal Omega output). The figure reports the primers sequences used in the study.

The absolute abundance of each aligned sequence was extracted and saved into a separated file containing the list of all the ASV identifiers.

Figure 18 reports an example of some amplicon sequences and their simulated abundances that were generated in the simulation process.

Figure 18. Product section of the amplicon extraction. In more detail, the figure shows the selection of amplicon V1V2 (here called amplicon_1). Amplicon sequence of different 16S genes may have different lengths, as shown in section "a" of the figure. Figure section "b" shows the absolute abundance to apply to the extracted amplicons.

The simulation of a 16S rRNA gene sequencing by amplicon strategy was performed using an R script. The program extracted the amplicon sequences when the gene sequence did fully matched both the primers sequences (forward and reverse) in at least the very first 5 nucleotides at the 3' end.

It is noteworthy that it is known, according to the simulation steps, the taxonomic classification (at each of the ranks) of every randomly selected sequence.

3.2.2.1 Pooled-Vs reference panel results

Table 8 reports the number of extracted amplicons (successfully extracted) for each taxonomic level for each sample from each simulated data. Between brackets the number of expected amplicons are reported.

Table 8. Number of simulated (in-silico amplified) 16S amplicons. Number of ASVs (from simulated reads and observed reads) that achieve classification up to a given specific taxonomy level. Starting from a taxonomic level of interest (e.g., Order), by adding the ASVs of the columns to the right of the subsequent taxonomic layers, it is possible to know the number of ASVs classified at the taxonomic level of interest. In the table the expected number of amplicons are described in parentheses.

Sample_ID Amplicon_region	Kingdom Sim. (Real)	Phylum Sim. (Real)	Class Sim. (Real)	Order Sim. (Real)	Family Sim. (Real)	Genus Sim. (Real)	Species Sim. (Real)
S1 V1V2	5 (5)	12 (12)	6 (6)	4 (4)	20 (20)	387 (421)	414 (451)
S1 V2V3	4 (5)	10 (12)	5 (6)	3 (4)	18 (20)	321 (421)	372 (451)
S1 V3V4	5 (5)	12 (12)	6 (6)	4 (4)	20 (20)	409 (421)	440 (451)
S1 V4V5	5 (5)	11 (12)	6 (6)	4 (4)	20 (20)	412 (421)	436 (451)
S1 V5V7	5 (5)	11 (12)	6 (6)	4 (4)	20 (20)	416 (421)	444 (451)
S1 V7V9	3 (5)	12 (12)	4 (6)	4 (4)	18 (20)	364 (421)	402 (451)
S2 V1V2	11 (15)	19 (19)	2 (4)	12 (12)	24 (24)	253 (266)	244 (272)
S2 V2V3	15 (15)	17 (19)	4 (4)	11 (12)	23 (24)	238 (266)	240 (272)
S2 V3V4	14 (15)	19 (19)	4 (4)	10 (12)	24 (24)	264 (266)	269 (272)
S2 V4V5	15 (15)	19 (19)	4 (4)	11 (12)	24 (24)	264 (266)	268 (272)
S2 V5V7	15 (15)	19 (19)	4 (4)	12 (12)	24 (24)	261 (266)	270 (272)
S2 V7V9	11 (15)	17 (19)	3 (4)	11 (12)	23 (24)	236 (266)	223 (272)
S3 V1V2	7 (7)	8 (8)	3 (3)	1 (1)	27 (27)	476 (506)	525 (565)
S3 V2V3	7 (7)	8 (8)	3 (3)	0 (1)	26 (27)	400 (506)	458 (565)
S3 V3V4	7 (7)	8 (8)	3 (3)	1 (1)	27 (27)	497 (506)	549 (565)
S3 V4V5	7 (7)	8 (8)	3 (3)	1 (1)	27 (27)	494 (506)	542 (565)
S3 V5V7	6 (7)	8 (8)	2 (3)	1 (1)	26 (27)	496 (506)	556 (565)
S3 V7V9	5 (7)	8 (8)	2 (3)	1 (1)	24 (27)	458 (506)	509 (565)
S4 V1V2	7 (9)	8 (8)	3 (3)	3 (3)	11 (11)	288 (303)	304 (325)
S4 V2V3	8 (9)	8 (8)	3 (3)	2 (3)	9 (11)	248 (303)	276 (325)
S4 V3V4	9 (9)	8 (8)	3 (3)	3 (3)	10 (11)	299 (303)	319 (325)
S4 V4V5	9 (9)	8 (8)	3 (3)	3 (3)	10 (11)	292 (303)	314 (325)
S4 V5V7	9 (9)	8 (8)	2 (3)	3 (3)	11 (11)	298 (303)	322 (325)
S4 V7V9	7 (9)	7 (8)	2 (3)	3 (3)	5 (11)	274 (303)	279 (325)
S5 V1V2	7 (7)	9 (10)	2 (2)	1 (1)	7 (7)	271 (290)	282 (301)
S5 V2V3	7 (7)	9 (10)	2 (2)	1 (1)	4 (7)	244 (290)	254 (301)
S5 V3V4	7 (7)	10 (10)	2 (2)	1 (1)	7 (7)	282 (290)	300 (301)
S5 V4V5	7 (7)	10 (10)	2 (2)	1 (1)	7 (7)	279 (290)	295 (301)
S5 V5V7	7 (7)	10 (10)	2 (2)	1 (1)	7 (7)	286 (290)	294 (301)
S5 V7V9	5 (7)	9 (10)	2 (2)	1 (1)	4 (7)	261 (290)	267 (301)

We observed that a large number of amplicon sequences were not extracted due to the lack of a full nucleotide matching between target and the primer sequences.
In general, the sequences that failed the extraction process were nearly 40-50% of the total of the sequences for all the samples. In particular, 46% (of the 919), 51% (of the 612), 42% (of the 1117), 41% (of the 662), and 39% (of the 618) of the unique sequences were not extracted for the sample S1, S2, S3, S4 and S5, respectively.

Moreover, the extraction failed for a large number of amplicon sequences belonging to V2V3 region (see table 9).

Table 9. Percentage of sequences not passing the amplicon extraction for each 16S rRNA gene amplicon region.

	S1	S2	S3	S4	S5
V1V2	7.7%	7.6%	6.2%	5.7%	6.3%
V2V3	20.2%	10.4%	19.2%	16.3%	15.7%
V3V4	2.5%	1.3%	2.2%	1.6%	1.4%
V4V5	2.7%	1.1%	3.1%	3.4%	2.7%
V5V7	1.4%	1.1%	1.7%	1.3%	1.7%
V7V9	12.1%	14.3%	9.8%	12.8%	11.1%

Before taxonomy calling process amplicon sequences shorter than 200bp were filtered out. This filtering approach was not applied to V1V2 and V7V9 regions since in many cases the available 16S bacterial sequence did not encompass the primer region. No amplicon sequences resulted to be filtered out.

3.2.2.2 Taxonomic classification

Table 10 reports the number of ASVs that underwent the classification process.

Table 10. Number of ASVs within the Control Table to be submitted to the taxonomy classifier. Table shows the number of ASVs derived from the simulation process that underwent the taxonomic classification (at each rank) for each simulated sample.

Control Table	S1	S2	S3	S4	S5
V1V2	742	493	870	538	513
V2V3	591	448	695	452	425
V3V4	666	436	768	456	448
V4V5	585	401	686	404	413
V5V7	622	436	733	479	428
V7V9	633	428	769	448	444

We assigned the taxonomy to the simulated amplicons using the pre-trained set eHOMD v15.1 through two steps: the first one using a training set for assigning taxonomy up to the genus level, then another training set to add species labels.

Table 11 shows the number of unique ASVs classified at each taxonomic level.

Table 11. Number of unique ASVs that reach a taxonomic level during the classification step. The Species level was able to classify a high percentage of the ASVs on the total.

	Kingdom	Phylum	Class	Order	Family	Genus	Species
V1V2	2568	2568	2568	2564	2563	2550	2402
V2V3	2035	2035	2035	2035	2035	2033	1928
V3V4	2124	2124	2124	2124	2124	2119	1957
V4V5	1880	1880	1880	1880	1880	1855	1628
V5V7	1998	1998	1998	1998	1998	1961	1756
V7V9	2157	2157	2157	2156	2156	2125	1797

As an overall result, ~90% of the ASV were classified at the "Species" taxonomic rank.

Table 12 showed the number of unique taxa estimated in the different taxonomic ranks.

Table 12. Number of bacterial taxa that reach a taxonomic level during the classification assignment.

	Kingdom	**Phylum**	**Class**	**Order**	**Family**	**Genus**	**Species**
V1V2	1	9	21	38	58	107	346
V2V3	1	9	19	34	54	103	317
V3V4	1	9	21	37	58	107	350
V4V5	1	9	20	36	57	106	325
V5V7	1	9	21	38	59	109	344
V7V9	1	9	21	36	55	101	301

As we can see, the number of Species were almost comparable; amplicon V3V4 was the best Species identifier than the others.

3.2.2.3 V2V3 reference panel results

The number of ASVs simulated in the previous part of the elaborate overestimated the number of sequences per 16S fragment, moving away from what would be a real case. The resulting products of the simulation program were 7 fasta-like files belonging each to one taxonomic rank and containing the ASV sequences which were successfully extracted in this step.

The subdivision for taxa layers served us only for the purpose of monitoring the different steps of the simulation.

Below (table 13) we reported the table of the number of extracted amplicons per taxonomic level for each sample (see method section 2.2.3).

Table 13. Number of ASVs (from simulated reads and observed reads) that achieve classification up to a given specific taxonomy level. Starting from a taxonomic level of interest (e.g., Order), by adding the ASVs of the columns to the right of the subsequent taxonomic layers, it is possible to know the number of ASVs classified at the taxonomic level of interest. In the table the expected number of amplicons are described in parentheses.

Sample_ID Amplicon_region	Kingdom Sim. (Real)	Phylum Sim. (Real)	Class Sim. (Real)	Order Sim. (Real)	Family Sim. (Real)	Genus Sim. (Real)	Species Sim. (Real)
S1 V1V2	1 (1)	2 (3)	1 (1)	1 (1)	2 (2)	97 (103)	109 (122)
S1 V2V3	1 (1)	3 (3)	1 (1)	1 (1)	2 (2)	85 (103)	111 (122)
S1 V3V4	1 (1)	3 (3)	1 (1)	1 (1)	2 (2)	102 (103)	120 (122)
S1 V4V5	1 (1)	3 (3)	1 (1)	1 (1)	2 (2)	102 (103)	119 (122)
S1 V5V7	1 (1)	3 (3)	1 (1)	1 (1)	2 (2)	103 (103)	122 (122)
S1 V7V9	1 (1)	3 (3)	1 (1)	1 (1)	2 (2)	94 (103)	106 (122)
S2 V1V2	3 (3)	3 (3)	1 (1)	4 (4)	6 (6)	68 (71)	56 (66)
S2 V2V3	2 (3)	3 (3)	1 (1)	4 (4)	6 (6)	63 (71)	66 (66)
S2 V3V4	3 (3)	3 (3)	1 (1)	4 (4)	6 (6)	71 (71)	66 (66)
S2 V4V5	3 (3)	3 (3)	1 (1)	4 (4)	6 (6)	71 (71)	63 (66)
S2 V5V7	3 (3)	3 (3)	1 (1)	4 (4)	6 (6)	71 (71)	66 (66)
S2 V7V9	2 (3)	3 (3)	1 (1)	3 (4)	5 (6)	63 (71)	53 (66)
S3 V1V2	3 (3)	2 (2)	1 (1)	4 (4)	6 (6)	117 (127)	124 (138)
S3 V2V3	2 (3)	2 (2)	1 (1)	4 (4)	6 (6)	104 (127)	125 (138)
S3 V3V4	3 (3)	2 (2)	1 (1)	4 (4)	6 (6)	124 (127)	136 (138)
S3 V4V5	3 (3)	2 (2)	1 (1)	4 (4)	6 (6)	125 (127)	133 (138)
S3 V5V7	3 (3)	2 (2)	1 (1)	4 (4)	6 (6)	125 (127)	134 (138)
S3 V7V9	2 (3)	2 (2)	1 (1)	3 (4)	5 (6)	115 (127)	127 (138)
S4 V1V2	2 (2)	2 (2)	1 (1)	4 (4)	0 (1)	68 (71)	80 (87)
S4 V2V3	2 (2)	2 (2)	1 (1)	4 (4)	1 (1)	60 (71)	81 (87)
S4 V3V4	2 (2)	2 (2)	1 (1)	4 (4)	1 (1)	69 (71)	84 (87)
S4 V4V5	2 (2)	2 (2)	1 (1)	4 (4)	1 (1)	69 (71)	85 (87)
S4 V5V7	2 (2)	2 (2)	1 (1)	4 (4)	1 (1)	71 (71)	86 (87)
S4 V7V9	1 (2)	2 (2)	0 (1)	3 (4)	1 (1)	67 (71)	74 (87)

S5 V1V2	2 (2)	2 (2)	1 (1)	4 (4)	0 (1)	49 (55)	56 (65)
S5 V2V3	2 (2)	2 (2)	1 (1)	4 (4)	1 (1)	48 (55)	59 (65)
S5 V3V4	2 (2)	2 (2)	1 (1)	4 (4)	1 (1)	51 (55)	63 (65)
S5 V4V5	2 (2)	2 (2)	1 (1)	4 (4)	1 (1)	52 (55)	65 (65)
S5 V5V7	2 (2)	2 (2)	1 (1)	4 (4)	1 (1)	55 (55)	65 (65)
S5 V7V9	1 (2)	2 (2)	0 (1)	3 (4)	1 (1)	55 (55)	55 (65)

In addition, the percentage of sequences that did not pass the amplicon extraction are shown in table 14.

Table 14. Percentage of sequences not passing the amplicon extraction for each 16S rRNA gene amplicon region.

ID	S1	S2	S3	S4	S5
V1V2	8.5%	8.4%	8.5%	6.5%	12.3%
V2V3	12.4%	5.8%	13.17%	10.1%	10%
V3V4	1.3%	0.0%	1.7%	3%	4.6%
V4V5	1.7%	1.9%	2.5%	2.3%	2..3%
V5V7	0.0%	0.0%	2.1%	0.6%	0.0%
V7V9	10.7%	15.6%	9.2%	12%	10%

Sequences that failed the simulation of amplicon extraction were nearly 30-40% of the total of the sequences among the different samples. In particular, ~34% (of the 233), ~31% (of the 154), ~37% (of the 281), ~34% (of the 168), and ~39% (of the 130) of the unique sequences were not extracted from the S1, S2, S3, S4, and S5 samples , respectively. As the previously simulated data, the amplicon V2V3 lost the higher percentage of sequence during amplicon extraction.

3.2.2.4 Taxonomic classification

All the file generates until here, were concatenated to proceed with the final taxonomic step, in order to obtain for each sample a single set of ASVs to classify.

Also here, as in the previously case, some bacteria could share an identical sequence fragment that could fit exactly into the portion of the amplicon we are going to test, resulting in duplicated sequences. We then collapsed identical amplicon sequence fragments, reproducing in this way the dereplication DADA2 step. Moreover, a control table was built to keep track of all sequences that underwent the classification, especially the collapsed ones. The total numbers of ASVs to be classified are reported in table 15.

Table 15. Number of ASVs within the Control Table to be submitted to the taxonomy classifier.

Control Table	S1	S2	S3	S4	S5
V1V2	204	130	240	149	108
V2V3	192	132	225	143	108
V3V4	200	133	238	147	108
V4V5	185	122	221	138	103
V5V7	196	135	224	140	118
V7V9	192	119	231	138	110

The classification part was carried out one amplicon at a time.

We assigned the taxonomy to the simulated amplicons using the pre-trained set eHOMD v15.1 through two steps: the first one using a training set for assigning taxonomy up to the genus level, then another training set to add species labels.

In table 16 we showed the number of unique ASVs classified at each taxonomic level among all samples combined.

Table 16. Number of unique ASVs that reach a taxonomic level during the classification step. The Species level was able to classify a high percentage of the ASVs on the total.

	Kingdom	Phylum	Class	Order	Family	Genus	Species
V1V2	710	710	710	709	709	705	652
V2V3	643	643	643	643	643	642	596
V3V4	650	650	650	650	650	650	578
V4V5	571	571	571	571	571	564	467
V5V7	613	613	613	613	613	597	509
V7V9	632	632	632	632	632	621	483

The difference with the first simulation lies in the fact that here the number of simulated ASVs approached what might be the amplicon size resulting from a 16S sequencing.
Here, almost 85% of the ASVs are classified at Species rank.

Table 17 shows the comparison of the distribution of phyla between the real and the simulated data, for all the samples and amplicon sequences.
Overall, the results of the simulated data reflect the real values although in some cases, the frequency of some phylum resulted to be higher in the simulated data.

Table 17. Classified bacteria at Phylum taxonomic rank of real and simulated data. For each simulated sample we observed that the distribution of the phyla reflects what was observed in the real data.

Classified bacteria at Phylum Level								
		Real data Observed frequency						
S1		**N of sequences**	**V1V2**	**V2V3**	**V3V4**	**V4V5**	**V5V7**	**V7V9**
Total sequences		234	204	192	200	185	196	192
	actinobacteria	11.5%	8.3%	13.0%	12.0%	9.7%	11.7%	8.3%
	bacteroidetes	13.6%	14.7%	4.1%	15.0%	15.1%	14.2%	14.5%
	firmicutes	32.0%	31.8%	35.4%	29.5%	28.1%	28.5%	31.2%
	fusobacteria	11.5%	12.2%	13.5%	10.5%	12.9%	13.2%	13.0%
	gracilibacteria (GN02)	0.43%	0.49%	/	0.5%	0.5%	0.5%	0.5%
	proteobacteria	27.3%	28.9%	30.7%	29.0%	29.7%	28.0%	28.6%
	Saccharibacteria (TM7)	2.5%	2.94%	3.1%	3.0%	3.2%	3.0%	3.1%
	spirochaetes	0.43%	0.49%	/	0.5%	0.5%	0.5%	0.5%
	unknown	0.43%	/	/	/	/	/	/
S2		**N of sequences**	**V1V2**	**V2V3**	**V3V4**	**V4V5**	**V5V7**	**V7V9**
Total sequences		154	130	132	133	122	135	119
	actinobacteria	5.8%	5.3%	6.0%	6.7%	6.5%	5.1%	2.5%
	bacteroidetes	16.2%	19.2%	12.8%	17.2%	19.6%	17.0%	18.4.%
	firmicutes	43.5%	40.7%	45.4%	43.6%	41.8%	43.7%	42.8%
	fusobacteria	5.8%	6.9%	6.8%	6.7%	6.5%	5.9%	6.7%
	proteobacteria	24.0%	25.3%	26.5%	22.5%	22.1%	25.1%	26.8%
	Saccharibacteria (TM7)	0.6%	0.7%	0.7%	0.7%	0.8%	0.7%	0.8%
	spirochaetes	1.9%	1.5%	1.5%	2.2%	2.4%	2.2%	1.6%
	unknown	1.9%	/	/	/	/	/	/

S3		N of sequences	V1V2	V2V3	V3V4	V4V5	V5V7	V7V9
Total sequences		271	240	225	238	221	224	231
	actinobacteria	10.3%	5.8%	11.5%	10.9%	10.8%	10.7%	6.9%
	bacteroidetes	18.4%	20.8%	8.4%	21.0%	21.7%	22.3%	22.0%
	firmicutes	24.7%	26.2%	32.8%	28.5%	28.0%	26.7%	26.4%
	fusobacteria	13.6%	14.1%	12.0%	10.5%	10.4%	10.7%	14.2%
	gracilibacteria (GN02)	0.3%	0.4%	0.4%	0.4%	0.4%	0.4%	0.4%
	proteobacteria	28.7%	29.5%	32.0%	26.0%	25.7%	25.8%	26.8%
	saccharibacteria (TM7)	2.2%	2.5%	2.2%	2.1%	2.2%	2.6%	2.6%
	spirochaetes	0.3%	0.4%	0.4%	0.4%	0.4%	0.4%	0.4%
	unknown	0.7%	/	/	/	/	/	/
S4		N of sequences	V1V2	V2V3	V3V4	V4V5	V5V7	V7V9
Total sequences		163	149	143	147	138	140	138
	actinobacteria	6.7%	5.3%	7.6%	6.8%	7.2%	7.8%	5.0%
	bacteroidetes	22.0%	24.8%	15.3%	22.4%	23.9%	25.0%	24.6%
	firmicutes	31.9%	33.5%	38.4%	35.3%	31.8%	32.1%	31.8%
	fusobacteria	7.9%	8.7%	8.3%	7.4%	6.5%	5.0%	9.4%
	proteobacteria	30.0%	27.5%	30.0%	27.8%	30.4%	30.0%	28.9%
	unknown	1.2%	/	/	/	/	/	/
S5		N of sequences	V1V2	V2V3	V3V4	V4V5	V5V7	V7V9
Total sequences		122	108	108	108	103	118	110
	actinobacteria	7.3%	6.4%	8.3%	7.4%	7.7%	8.4%	4.5%
	bacteroidetes	20.4%	24.0%	13.8%	22.2%	19.4%	22.0%	23.6%
	firmicutes	34.4%	32.4%	42.5%	35.1%	36.8%	33.9%	35.4%
	fusobacteria	6.5%	7.4%	5.5%	5.5%	6.8%	6.7%	6.3%
	proteobacteria	31.1%	29.6%	29.6%	29.6%	29.1%	28.8%	30.0%

3.2.3 Simulator performances

The 16S sequences extraction of simulated data followed the method described in section 2.2.2. This in-silico simulation used real data from V2V3 count tables as reference panel for taxa abundance. Here, in the tables below we report the absolute numbers of real observed sequences and the absolute abundance of the simulated sequences, to verify the concordance rate of taxa between real and simulated eHOMD sequence data.

Table 18 shows the number of 16S rRNA gene sequences extracted for each identified phylum (table 18a), while table 18b shows the percentage of these selected sequences.

Table 18. Comparison between real number of observed sequences and the simulated sequences. In the first table (a) the number of real and simulated sequences are reported and grouped by the Phylum taxonomic level. The number of sequences of the real analyzed samples are represented in columns showing "ID" labels, instead simulated samples are shown with an "S". The second table (b) shows the percentage of the sequences described in table 18a.

a)

Phylum	ID1	S1	ID2	S2	ID3	S3	ID4a	S4	ID4bc	S5
	n reads	n reads	n reads	n reads	n reads	n reads	n reads	n reads	n reads	n reads
Actinobacteria	27	27	9	9	28	28	11	12	9	9
Bacteroidetes	32	32	25	26	50	51	36	36	25	25
Firmicutes	75	76	67	68	67	69	52	52	42	42
Fusobacteria	27	27	9	9	37	37	13	13	8	8
Gracilibacteria (GN02)	1	1	-	-	1	1	-	-	-	-
Proteobacteria	64	63	37	37	78	78	49	49	38	38
Saccharibacteria (TM7)	6	6	1	1	6	6	-	-	-	-
Spirochaetes	1	1	3	3	1	1	-	-	-	-
Unclassified	1	0	3	0	3	0	2	0	-	-

b)

Phylum	ID1	S1	ID2	S2	ID3	S3	ID4a	S4	ID4bc	S5
	%	%	%	%	%	%	%	%	%	%
Actinobacteria	11.5	11.6	5.8	5.8	10.3	10.3	6.7	7.4	7.3	7.3
Bacteroidetes	13.6	13.7	16.2	16.9	18.4	18.8	22.1	22.2	20.4	20.5
Firmicutes	32.1	32.6	43.5	44.4	24.7	25.4	31.9	32.1	34.4	34.4
Fusobacteria	11.5	11.6	5.8	5.8	13.6	13.6	7.9	8.1	6.5	6.5
Gracilibacteria (GN02)	0.4	0.4	-	-	0.3	0.3	-	-	-	-
Proteobacteria	27.3	27.1	24.1	24.1	28.7	28.7	30.1	30.2	31.1	31.1
Saccharibacteria (TM7)	2.5	2.5	0.6	0.6	2.2	2.2	-	-	-	-
Spirochaetes	0.4	0.4	1.9	1.9	0.3	0.3	-	-	-	-
Unclassified	0.4	0.0	1.9	0.0	1.1	0.0	1.2	0.0	-	-

We can observe that selection of 16S rRNA gene selected sequences from the eHOMD database respected the observed real number of sequences.
In figure 19, we reported a plot that compares the real percentage of sequences grouped by the Phylum taxonomic level and the simulated percentage of the sequences.

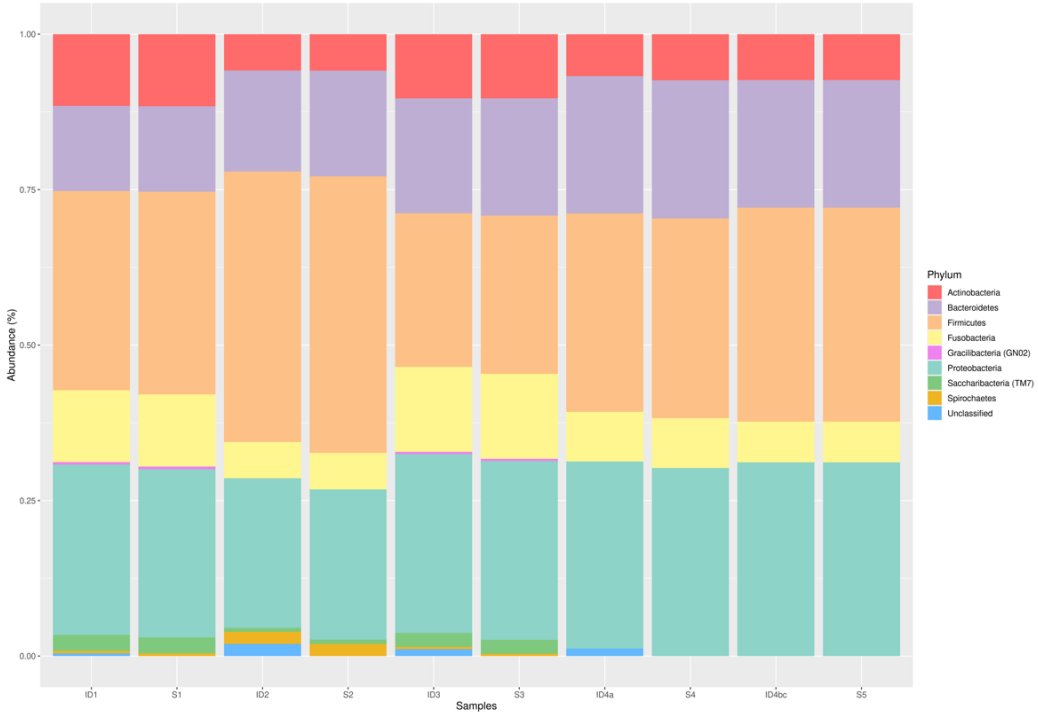


Figure 19. Comparison between real 16S rRNA gene sequences and simulated 16S rRNA gene sequences. The figure shows for each sample ("ID", the real samples; "S", the simulated ones) represented in the x-axis, the percentage of each sequence grouped by the Phylum taxonomic level (y-axis). The distribution of the sequences are quite the same for each real and simulated sample.

Moreover, the 16S sequences extraction method continued with the assignment of an absolute abundance for each chosen sequence, following the reference panel and applying the Poisson distribution (section 2.2.2).

The table 19 shows the resulting and simulated relative abundances for each Phylum and for each simulated samples, compared with the real relative abundances grouped also for the Phylum taxonomic level.

Table 19. Comparison between real relative sequence abundances and simulated sequence abundances. The table shows for each real sample ("ID") and simulated sample ("S") the relative abundance of sequences grouped by Phylum taxonomic level.

Phylum	ID1	S1	ID2	S2	ID3	S3	ID4a	S4	ID4bc	S5
	%	%	%	%	%	%	%	%	%	%
Actinobacteria	16.6	12.5	1.9	2.4	6.6	11	0.9	1.4	1.2	2.5
Bacteroidetes	2.3	5.1	1.07	5.3	3.2	7.4	4.1	6.7	4.4	7.5
Firmicutes	37.9	41.7	9.9	53.1	15.7	27.6	31.1	36.3	36.3	45.6
Fusobacteria	6	9	2	2.8	6.6	9.2	2.9	5.7	3.3	4.7
Gracilibacteria (GN02)	0.01	0.0	-	-	0.1	0.01	-	-	-	-
Proteobacteria	36.4	29.6	84.7	34.6	67.3	44.3	60.7	49,93	54.7	39.6
Saccharibacteria (TM7)	0.7	1.7	0.11	0.7	0.3	0.3	-	-	-	-
Spirochaetes	0.01	0.3	0.2	0.9	0.01	0.06	-	-	-	-
Unclassified	0.0	0.0	0.17	0.0	0.02	0.0	0.07	0.0	-	-

In figure 20 relative abundances of each Phylum of the real sequences and the simulated sequences are reported.

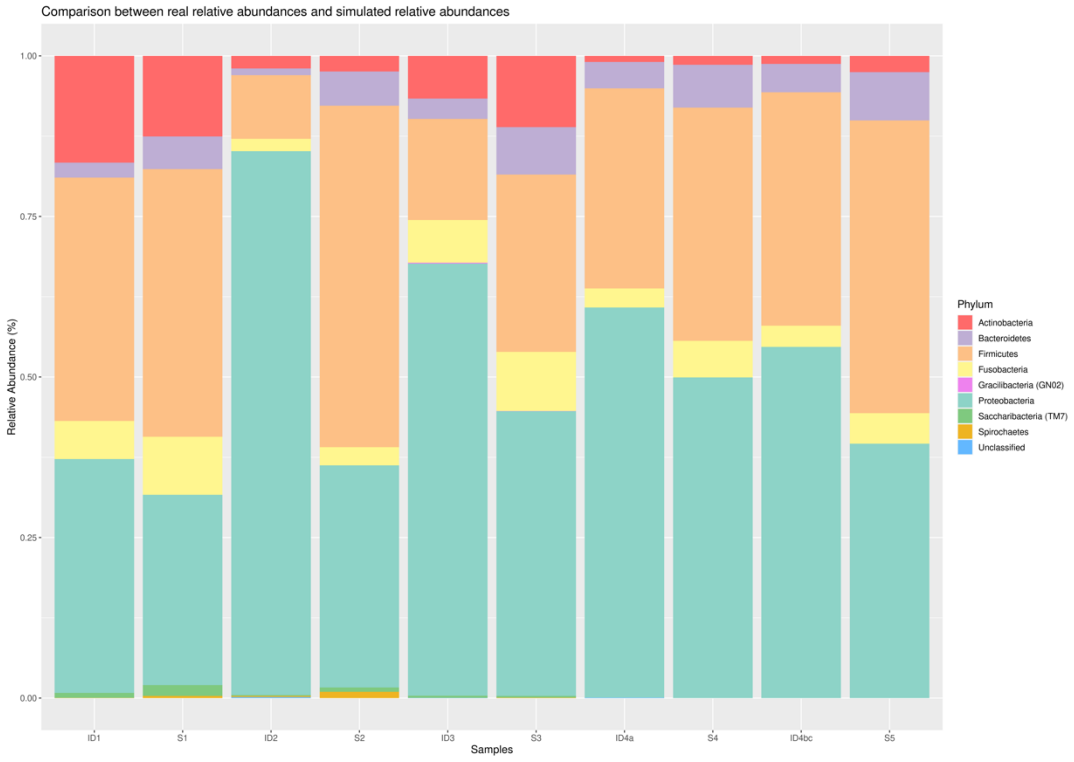


Figure 20. Comparison between real relative sequence abundances and simulated sequence abundances. For each sample represented in the x-axis ("ID", the real samples; "S", the simulated ones) the relative of each Phylum is showed in the y-axis.

3.2.4 Searching for pitfalls in Species classification

When focused on the classification at Species rank, we observed that ~80% of the ASVs (all the amplicon regions, all simulated samples) were classified (data not shown). We then aimed to understand why 20% of ASVs failed the classification.
We went into the detail of this classification process and looked at the ASVs that did not classify. To do this, the classification procedure was re-run allowing for multi-species assignment (in the case the same ASV overlaps the sequence of 16S rRNA gene of more than one different bacterium). All the unclassified ASVs showed match to more than one different bacterial species, without any exceptions.

To further study the ambiguous classification of some ASVs (whose true classification is known because they were selected from 16S rRNA gene sequences of the eHOMD, according to the simulation procedure) we selected some ASVs for a detailed analysis. For every selected ASVs, we built a taxonomic tree based on the multiple alignment of the eHOMD sequences of the bacterial species sharing the sequence of the ASV under investigation. Given the two or more Species as probable solution, we selected all the eHOMD 16S sequences belonging to such Species to be submitted to the multiple alignment.
Figure 21 shows an example of multiple sequence alignment linked to a phylogenetic tree. We aligned an ASV (called ASV172) belonging to the V1V2 amplicon fragment against the full-length 16S rRNA gene reference sequences of all "Prevotella sp.HMT472" and "Prevotella loescheii" Species classified into the eHOMD database. The multiple alignment of ASV172 resulted to be exactly matched the first 16S fragment (V1V2 amplicon) of the two species just mentioned, and let's see how the three sequences are placed close together in the branches of the tree.

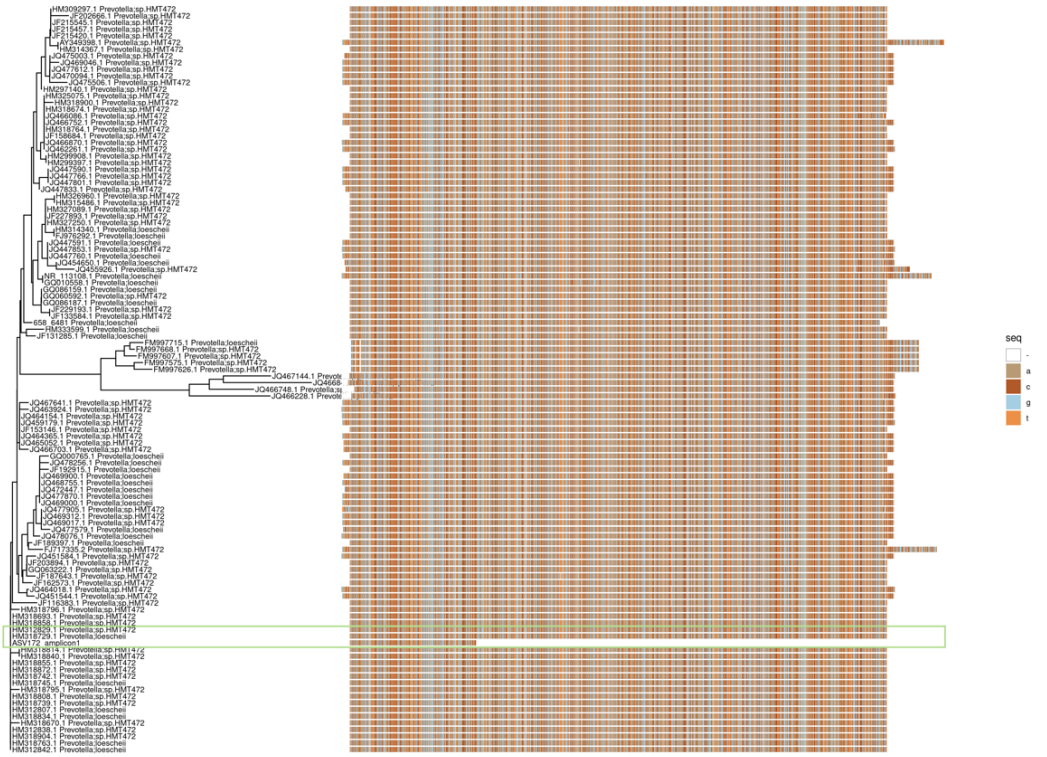


Figure 21. Phylogenetic tree and MSA of ASV172 belong to V1V2 amplicon (amplicon1). This sequence was a dereplication product of two distinct sequences coming originally from "HM318729.1 Prevotella loescheii", and "HM312829.1 Prevotella sp.HMT472". Each color of the MSA represents a nucleotide base, whereas the alignment gaps were represented by white spaces.

The ASV172 and the sequences of the 2 bacterial Species suggested by the classifier are shown in the green box.

In the figure below (figure 22), we reported another example. We performed the multiple sequence alignment using the ASV6 from the simulated V3V4 amplicon region against all full-length 16S rRNA gene sequences of eHOMD belonging to the genus "Campylobacter"; in particular we extracted the species "Campylobacter rectus" and "Campylobacter showae".

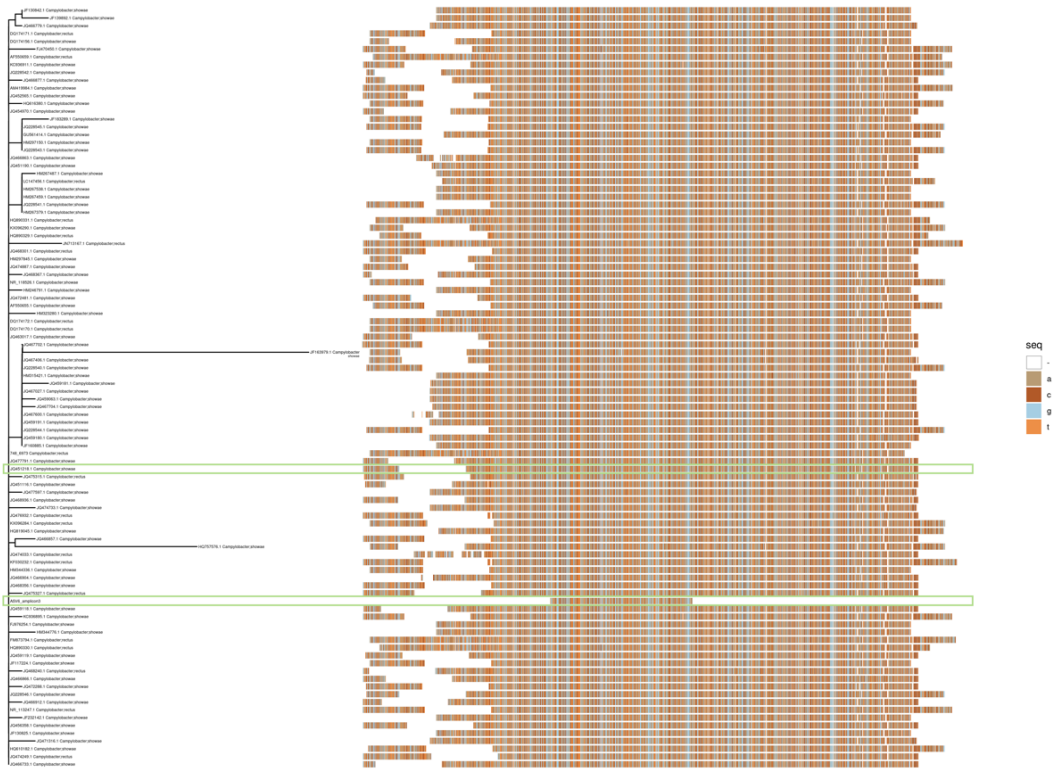


Figure 22. Phylogenetic tree and MSA of ASV6 belong to V3V4 amplicon (amplicon3). This sequence originally came from "JQ451218.1 Campylobacter showae", placed far apart from the target ASV. Each color of the MSA represents a nucleotide base, whereas the alignment gaps were represented by white spaces.

In this case the ASV6 sequence did not undergo the dereplication step, but also had difficulties in assigning the species label. We traced the original sequence of the amplicon sequence variant, picking out the species "JQ451218.1 Campylobacter showae". In the tree, the sequences under consideration were placed far apart from each other. The ASV6_amplicon3 and the template are highlighted in green boxes.
In some cases, the unsolved ASVs showed greater complexity, as some of these had been merged with many different ASVs (e.g., 7 ASVs), resulting in 14 different species.

4 DISCUSSION AND CONCLUSIONS

The human microbiota consists of about 10^{14} bacterial cells and ~3 million different genes, compared to the human genome that contains 23,000 protein-coding genes[4,5,7]. The interaction between these bacteria and the human body takes place daily at the level of the skin, gastrointestinal tract, respiratory system, and other.

After the human gut microbiota, the oral microbiota is considered the second most populated bacterial environment in human body, and its composition is constantly changing during life, making the microbiota a dynamic ecosystem. Oral microbiota differs between individuals and its composition depends on many different factors, including the host genetics, the maternal transmission, and the environmental factors[11,15]. An altered oral microflora could be associated with several diseases.

The 16S rRNA gene is widely used to profile the bacterial abundance of human body site of interest, assigning taxonomy to the bacterial sequences. In the 16S rRNA gene sequence, nine hypervariable regions show a low degree of similarity/conservation among bacterial species (V1 to V9). These regions can be successfully sequenced and analyzed for taxonomic purposes and the most actually used are the regions from V1 to V4[11].

The Divisive Amplicon Denoising Algorithm 2 (DADA2) was used to investigate the microbial composition of 4 oral samples derived from the 16S rRNA gene sequencing of oral swabs. It allowed an analysis at the single nucleotide resolution. Since the estimate of the abundance of the various taxa is based on the similarity degree among the detected sequences of the 16S rRNA gene, it is pivotal that sequences undergoing the downstream analysis do present a very low number of errors. In this project, 16S rRNA gene analyses were re-run multiple times in order to tune the parameters of the analyses regarding the removal of low-quality bases, the mate-pairs merging, and chimeras' detection. The analyses were performed leveraging on either the pooled sequences for all the amplicon regions or the sequences of the single amplicon regions. It is noteworthy that the here reported microbial results have been achieved by the sequencing of more than 1 or 2 amplicon sequence regions, as instead commonly performed in the standard microbial studies.

During the analysis processes, nearly 20% of sequences (ASVs) was discarded because sequences were not reliable. In the analyses performed on the single amplicon regions,

we observed that some bacterial Species were not detected by all amplicons, highlighting that the different 16S regions can specifically be able to recognize specific bacteria taxa. This result was well depicted at the taxonomic rank of "Genus", in which most of the ASVs were successfully classified (figure 15*: Heatmap of bacterial Genera identified in different amplicons*).

When looking at the amplicon regions that recognized the higher number of oral bacterial species, we observed that the V2V3 and V1V2 amplicons showed the best performances. Of course, the best result was obtained by analyzing the pooled 16S rRNA gene amplicon regions, where we detected total of 204 different species. This suggests that methods based on the sequencing of the whole 16S rRNA gene sequence are likely to perform better than methods based on single amplicon region sequencing although a satisfying classification at the resolution of the "Species" rank does not appear possible when studying one single gene only.

The analyses were conducted without normalizing the ASV counts among the samples. This was done to let the analyses able to detect as many bacterial species as possible although for some authors this could affect the performance of a direct analysis of differential bacterial abundances among the amplicon sequence regions and samples (this was out of the goals of the present work)[37].

The results of microbial profiles were then used to drive the simulation study that aimed to investigate possible pitfalls of the 16S rRNA gene-based bacterial taxonomy of oral microbiota. According to the corresponding real data, each simulated sample was profiled with a similar number of ASVs, to mimic possible troubles in the sequencing process.

Our designed random sampling method of whole 16S rRNA gene sequences (method described in section 2.2.2) based on reference panel seemed to respect the frequency of the observed bacteria for each of the real samples (table 18 shows the Phylum taxonomic layer). However, the method of randomizing absolute abundances for each simulated sequence using a bootstrap method is being developed and improved, also paying attention to those sequences whose taxonomy is unknown, however presenting a certain absolute abundance in the analyzed samples.

The, the simulation process implied the extraction of amplicon sequences from the whole 16S rRNA gene sequences of simulated bacteria in the oral microbiome. This resulted in being a key task. We report that not all the amplicons can be successfully "amplified" (approximately 40-50% across all the simulated amplicons for each sample) from present targets, suggesting that, also in the real world, many 16S rRNA genes cannot be amplified and therefore detected. This would strongly impact many analyses including the detection of a large amount of bacterial species (nearly 20% of the *in-silico* ASV failed the amplification) and the comparison among groups of samples of the abundance of specific bacterial taxa. We also observed that 16S rRNA genes sequences of different bacterial species may share identical amplicon sequences. In such a case, the taxonomic classification for these ASVs did not take place at the species rank or even at higher ranks.

It is known that the 16S rRNA gene analysis is a good molecular-based method to investigate the microbiome composition within specific human body districts. However, there are several limitations in assigning the species taxonomic label. These are due to many factors, such as the novel bacteria sequences not yet archived in public databases, or annotation mistakes in certain databases[17]. In our case, the classifier did not assign a taxonomic label to the several sequences because of their sequence identity with many 16S rRNA gene sequences. The classifier, therefore, rather than reporting the possible Species solutions, left the sequence unassigned.

We then decided to compare the unclassified sequence (target ASV) with the sequences of the "possible Species solutions".

A multi alignment sequence approach combined with a phylogenetic tree construction showed that the target ASV shared identical sequence with the whole 16 rRNA gene sequences of the "possible Species solutions". In some cases, the target ASV, resulted to be either closely or distantly located to 16S rRNA gene sequence from which it was extracted during the simulation process. This suggested that several different bacterial species show a high degree of similarity in a given 16S rRNA gene sequence region making the classification not reliable. We argue that this could be solved by investigating other gene regions, or better a list of several genes other than the 16S rRNA.

The results of simulation showed a higher rate of classification (80%) than the real data (50%). A possible explanation is that the simulation selected sequences came from a

well annotated database, each having the full taxonomic classification, so this knowledge somehow biased the downstream analysis towards a higher rate of classification.

We might ask whether the analysis of one or more 16S rRNA hypervariable region could be a good approach to provide a complete picture of the bacterial composition of an individual. The 16S rRNA gene is the most used gene in taxonomy profile studies, but it could not be enough in detecting all the species residing in a human body site, knowing the difficulties in reaching the species layer. Thus, studying only a part of this gene could not be sufficient.

Currently, community profiling of the 16S rRNA gene is conducted using short-read sequencing technologies of two or three consecutive 16S hypervariable regions, most commonly V1V3 amplicon, investigating only a portion of the complete gene largely responsible for the well-known difficulty in detecting "Species"[31]. On the other hand, analyses from whole 16S rRNA gene sequencing are currently being evaluated, producing long read to potentially achieve the species taxonomic resolution[36].

In order to have a more robust reference panel, we could design the experiments for the in-vitro analysis of the microbiota, providing the use of technical replicates for each sample. Moreover, the samples should be analyzed at the same time to avoid a greater number of bias effects.

Furthermore, the reference panel has been built through the analysis of a limited number of individuals, which would certainly be increased if we want to build a more accurate catalog of bacterial types. The catalog could be validated by verifying with alternative methods (e.g., real-time) the presence of some bacterial strains identified by bioinformatic analysis only.

It should be noted that the thesis work was addressed to the realization of the simulator which, once developed, requires input data describing the abundance of the different species to be included in the simulation. Having a good reference is important for the specific evaluation of the quality of the protocol used, but not interesting in the activity of the simulator itself.

The use of the simulator will be useful to understand the sensitivity of the different regions of the 16S rRNA gene in a targeted 16S analysis in recognizing the presence of possible known pathogenic bacteria in the context of a study sample affected by a certain disease, such as the periodontitis.

We are also tuning our simulation method performing a bootstrap-like resampling of individual's 16S amplicon regions, also for a bigger sample set of hundreds of individuals, trying to improve bacteria detection and taxonomy identification.

Bibliography

[1] Berg, G., Rybakova, D., Fischer, D. *et al.* Microbiome definition re-visited: old concepts and new challenges.*Microbiome* **8**, 103 (2020). https://doi.org/10.1186/s40168-020-00875-0

[2] D'Argenio V. Human Microbiome Acquisition and Bioinformatic Challenges in Metagenomic Studies. Int J Mol Sci. 2018 Jan 27;19(2):383. doi: 10.3390/ijms19020383. PMID: 29382070; PMCID: PMC5855605.

[3] Sekirov, I., Russell, S. L., Antunes, L. C. M., & Finlay, B. B. (2010). Gut microbiota in health and disease. Physiological reviews, 90(3), 859-904.

[4] Malard, F., Dore, J., Gaugler, B. *et al.* Introduction to host microbiome symbiosis in health and disease. *Mucosal Immunol* **14**, 547–554 (2021). https://doi.org/10.1038/s41385-020-00365-4

[5] Ursell LK, Metcalf JL, Parfrey LW, Knight R. Defining the human microbiome. Nutr Rev. 2012 Aug;70 Suppl 1(Suppl 1):S38-44. doi: 10.1111/j.1753-4887.2012.00493.x. PMID: 22861806; PMCID: PMC3426293.

[6] Qin, J. et al. A human gut microbial gene catalogue established by metagenomic sequencing. *Nature* **464**, 59–65 (2010)

[7] Turnbaugh, P., Ley, R., Hamady, M. *et al.* The Human Microbiome Project. *Nature* **449**, 804–810 (2007). https://doi.org/10.1038/nature06244

[8] Kim Y, Koh I, Rho M. Deciphering the human microbiome using next-generation sequencing data and bioinformatics approaches. Methods. 2015 Jun;79-80:52-9. doi: 10.1016/j.ymeth.2014.10.022. Epub 2014 Oct 28. PMID: 25448477.

[9] The Human Microbiome Project Consortium. Structure, function and diversity of the healthy human microbiome. *Nature* **486**, 207–214 (2012). https://doi.org/10.1038/nature11234

[10] The Human Microbiome Project Consortium. A framework for human microbiome research. Nature 486, 215–221 (2012). https://doi.org/10.1038/nature11209

[11] Verma D, Garg PK, Dubey AK. Insights into the human oral microbiome. Arch Microbiol. 2018 May;200(4):525-540. doi: 10.1007/s00203-018-1505-3. Epub 2018 Mar 23. PMID: 29572583.

[12] Dewhirst FE, Chen T, Izard J, Paster BJ, Tanner AC, Yu WH, Lakshmanan A, Wade WG. The human oral microbiome. J Bacteriol. 2010 Oct;192(19):5002-17. doi: 10.1128/JB.00542-10. Epub 2010 Jul 23. PMID: 20656903; PMCID: PMC2944498.

[13] Escapa IF, Chen T, Huang Y, Gajare P, Dewhirst FE, Lemon KP. New Insights into Human Nostril Microbiome from the Expanded Human Oral Microbiome Database (eHOMD): a Resource for the Microbiome of the Human Aerodigestive Tract. mSystems. 2018 Dec 4;3(6):e00187-18. doi: 10.1128/mSystems.00187-18. PMID: 30534599; PMCID: PMC6280432.

[14] Chen T, Yu WH, Izard J, Baranova OV, Lakshmanan A, Dewhirst FE. The Human Oral Microbiome Database: a web accessible resource for investigating oral microbe taxonomic and genomic information. Database (Oxford). 2010 Jul 6;2010:baq013. doi: 10.1093/database/baq013. PMID: 20624719; PMCID: PMC2911848.

[15] Sedghi L, DiMassa V, Harrington A, Lynch SV, Kapila YL. The oral microbiome: Role of key organisms and complex networks in oral health and disease. Periodontol 2000. 2021 Oct;87(1):107-131. doi: 10.1111/prd.12393. PMID: 34463991; PMCID: PMC8457218.

[16] Frank A. Scannapieco, The oral microbiome: Its role in health and in oral and systemic infections, Clinical Microbiology Newsletter, Volume 35, Issue 20, 2013,Pages 163-169, ISSN 0196-4399

[17] Edgar R. Taxonomy annotation and guide tree errors in 16S rRNA databases. PeerJ. 2018 Jun 12;6:e5030. doi: 10.7717/peerj.5030. PMID: 29910992; PMCID: PMC6003391.

[18] Chakravorty S, Helb D, Burday M, Connell N, Alland D. A detailed analysis of 16S ribosomal RNA gene segments for the diagnosis of pathogenic bacteria. J Microbiol Methods. 2007 May;69(2):330-9. doi: 10.1016/j.mimet.2007.02.005. Epub 2007 Feb 22. PMID: 17391789; PMCID: PMC2562909.

[19] Yarza, P., Yilmaz, P., Pruesse, E. *et al.* Uniting the classification of cultured and uncultured bacteria and archaea using 16S rRNA gene sequences. *Nat Rev Microbiol* **12**, 635–645 (2014). https://doi.org/10.1038/nrmicro3330

[20] Bharti R, Grimm DG. Current challenges and best-practice protocols for microbiome analysis. Brief Bioinform. 2021 Jan 18;22(1):178-193. doi: 10.1093/bib/bbz155. PMID: 31848574; PMCID: PMC7820839.

[21] Lan Y, Rosen G, Hershberg R. Marker genes that are less conserved in their sequences are useful for predicting genome-wide similarity levels between closely related prokaryotic strains. Microbiome. 2016 May 3;4(1):18. doi: 10.1186/s40168-016-0162-5. PMID: 27138046; PMCID: PMC4853863.

[22] Haas BJ, Gevers D, Earl AM, Feldgarden M, Ward DV, Giannoukos G, Ciulla D, Tabbaa D, Highlander SK, Sodergren E, Methé B, DeSantis TZ; Human Microbiome Consortium; Petrosino JF, Knight R, Birren BW. Chimeric 16S rRNA sequence formation and detection in Sanger and 454-pyrosequenced PCR amplicons. Genome Res. 2011 Mar;21(3):494-504. doi: 10.1101/gr.112730.110. Epub 2011 Jan 6. PMID: 21212162; PMCID: PMC3044863.

[23] Cole JR, Wang Q, Fish JA, Chai B, McGarrell DM, Sun Y, Brown CT, Porras-Alfaro A, Kuske CR, Tiedje JM. Ribosomal Database Project: data and tools for high throughput rRNA analysis. Nucleic Acids Res. 2014 Jan;42(Database issue):D633-42. doi: 10.1093/nar/gkt1244. Epub 2013 Nov 27. PMID: 24288368; PMCID: PMC3965039.

[24] DeSantis TZ, Hugenholtz P, Larsen N, Rojas M, Brodie EL, Keller K, Huber T, Dalevi D, Hu P, Andersen GL. Greengenes, a chimera-checked 16S rRNA gene database and workbench compatible with ARB. Appl Environ Microbiol. 2006 Jul;72(7):5069-72. doi: 10.1128/AEM.03006-05. PMID: 16820507; PMCID: PMC1489311.

[25] Quast C, Pruesse E, Yilmaz P, Gerken J, Schweer T, Yarza P, Peplies J, Glöckner FO. The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. Nucleic Acids Res. 2013 Jan;41(Database issue):D590-6. doi: 10.1093/nar/gks1219. Epub 2012 Nov 28. PMID: 23193283; PMCID: PMC3531112.

[26] Isabel Fernàndez Escapa, Tsute Chen, Yanmei Huang, Prasad Gajare, Floyd E Dewhirst, and Katherine P Lemon. New insight into human nostril microbiome from the expanded Human Oral Microbiome Database (eHOMD): a resource for species-level identification of microbiome data from the aerodigestive tract. DOI: 10.1128/mSystems.00187-18. Online Open Access: https://msystems.asm.org/content/3/6/e00187-18

[27] Yilmaz P, Parfrey LW, Yarza P, Gerken J, Pruesse E, Quast C, Schweer T, Peplies J, Ludwig W, Glöckner FO. The SILVA and "All-species Living Tree Project (LTP)" taxonomic frameworks. Nucleic Acids Res. 2014 Jan;42(Database issue):D643-8. doi: 10.1093/nar/gkt1209. Epub 2013 Nov 28. PMID: 24293649; PMCID: PMC3965112.

[28] Yuhan Hao, Zhiheng Pei, Stuart M. Brown, Chapter 1 - Bioinformatics in Microbiome Analysis, Editor(s): Colin Harwood, Methods in Microbiology, Academic Press, Volume 44, 2017, Pages 1-18, ISSN 0580-9517, ISBN 9780128137147,

https://doi.org/10.1016/bs.mim.2017.08.002.
(https://www.sciencedirect.com/science/article/pii/S0580951717300028)

[29] Callahan, B., McMurdie, P. & Holmes, S. Exact sequence variants should replace operational taxonomic units in marker-gene data analysis. *ISME J* **11**, 2639–2643 (2017). https://doi.org/10.1038/ismej.2017.119

[30] Callahan BJ, McMurdie PJ, Rosen MJ, Han AW, Johnson AJ, Holmes SP. DADA2: High-resolution sample inference from Illumina amplicon data. Nat Methods. 2016 Jul;13(7):581-3. doi: 10.1038/nmeth.3869. Epub 2016 May 23. PMID: 27214047; PMCID: PMC4927377.

[31] Callahan BJ, Wong J, Heiner C, Oh S, Theriot CM, Gulati AS, McGill SK, Dougherty MK. High-throughput amplicon sequencing of the full-length 16S rRNA gene with single-nucleotide resolution. Nucleic Acids Res. 2019 Oct 10;47(18):e103. doi: 10.1093/nar/gkz569. PMID: 31269198; PMCID: PMC6765137.

[32] F. Escapa, I., Huang, Y., Chen, T. *et al.* Construction of habitat-specific training sets to achieve species-level assignment in 16S rRNA gene datasets. *Microbiome* **8**, 65 (2020). https://doi.org/10.1186/s40168-020-00841-w.

[33] McMurdie PJ, Holmes S. phyloseq: an R package for reproducible interactive analysis and graphics of microbiome census data. PLoS One. 2013 Apr 22;8(4):e61217. doi: 10.1371/journal.pone.0061217. PMID: 23630581; PMCID: PMC3632530

[34] Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega Sievers F, Wilm A, Dineen DG, Gibson TJ, Karplus K, Li W, Lopez R, McWilliam H, Remmert M, Söding J, Thompson JD, Higgins D Molecular Systems Biology 7 Article number: 539 doi:10.1038/msb.2011.75

[35] Toparslan E, Karabag K, Bilge U. A workflow with R: Phylogenetic analyses and visualizations using mitochondrial cytochrome b gene sequences. PLoS One. 2020 Dec 15;15(12):e0243927. doi: 10.1371/journal.pone.0243927. PMID: 33320915; PMCID: PMC7737995.

[36] Johnson, J.S., Spakowicz, D.J., Hong, BY. et al. Evaluation of 16S rRNA gene sequencing for species and strain-level microbiome analysis. Nat Commun 10, 5029 (2019). https://doi.org/10.1038/s41467-019-13036-1

[37] Lin, H., Peddada, S.D. Analysis of microbial compositions: a review of normalization and differential abundance analysis. *npj Biofilms Microbiomes* **6**, 60 (2020). https://doi.org/10.1038/s41522-020-00160-w

Supplementary Materials

Supplementary Table 20. Average relative abundance of the top 10 Species detected within the study of all amplicons pooled together. The Haemophilus parainfluenzae is always the most abundant species among samples. All tables shared almost the same bacterial species. The "rank" column represents the order of which these species were identified according to their relative abundance and the top 10 species identified in pooled-Vs.

a)

pooled-Vs	**Top10 (%)**	**Rank**
Haemophilus_parainfluenzae	34.6%	1
Lautropia_mirabilis	4.01%	2
Neisseria_oralis	1.71%	3
Haemophilus_paraphrohaemolyticus	1.63%	4
Streptococcus_mitis	1.24%	5
Peptidiphaga_gingivicola	0.93%	6
Streptococcus_sanguinis	0.72%	7
Abiotrophia_defectiva	0.69%	8
Prevotella_melaninogenica	0.63%	9
Veillonella_sp.HMT780	0.60%	10

b)

V1V2	**Top10 (%)**	**Rank**
Haemophilus_parainfluenzae	32.3%	1
Streptococcus_mitis	8.07%	2
Lautropia_mirabilis	3.79%	3
Neisseria_oralis	1.69%	4
Streptococcus_sanguinis	1.34%	6
Peptidiphaga_gingivicola	1.08%	9
Prevotella_melaninogenica	0.98%	10
Veillonella_sp.HMT780	0.4%	23
Abiotrophia_defectiva	0.07%	55

c)

V2V3	Top10 (%)	Rank
Haemophilus_parainfluenzae	36.7%	1
Lautropia_mirabilis	3.24%	2
Neisseria_oralis	2.09%	3
Haemophilus_paraphrohaemolyticus	1.67%	4
Streptococcus_sanguinis	1.39%	5
Peptidiphaga_gingivicola	1.02%	8
Abiotrophia_defectiva	0.93%	9
Prevotella_melaninogenica	0.87%	10
Veillonella_sp.HMT780	0.64%	14
Streptococcus_mitis	0.14%	43

d)

V3V4	Top10 (%)	Rank
Haemophilus_parainfluenzae	39.4%	1
Lautropia_mirabilis	4.4%	2
Neisseria_oralis	1.89%	3
Haemophilus_paraphrohaemolyticus	1.42%	4
Streptococcus_sanguinis	1.33%	5
Abiotrophia_defectiva	0.96%	6
Peptidiphaga_gingivicola	0.76%	7
Prevotella_melaninogenica	0.76%	8
Veillonella_sp.HMT780	0.61%	9

e)

V4V5	Top10 (%)	Rank
Haemophilus_parainfluenzae	36.1%	1
Lautropia_mirabilis	5.78%	2
Neisseria_oralis	2.37%	3
Haemophilus_paraphrohaemolyticus	1.24%	4
Veillonella_sp.HMT780	0.98%	5
Peptidiphaga_gingivicola	0.76%	6
Abiotrophia_defectiva	0.63%	7
Streptococcus_sanguinis	0.12%	29

f)

V5V7	Top10 (%)	Rank
Haemophilus_parainfluenzae	29.8%	1
Lautropia_mirabilis	6.08%	2
Neisseria_oralis	2.61%	3
Haemophilus_paraphrohaemolyticus	2.25%	4
Peptidiphaga_gingivicola	1.14%	5
Abiotrophia_defectiva	1%	6
Streptococcus_sanguinis	0.39%	15
Veillonella_sp.HMT780	0.35%	17

g)

V7V9	Top10 (%)	Rank
Haemophilus_parainfluenzae	38.2%	1
Lautropia_mirabilis	2.3%	2
Haemophilus_paraphrohaemolyticus	2.33%	3
Veillonella_sp.HMT780	0.9%	5
Abiotrophia_defectiva	0.89%	6
Prevotella_melaninogenica	0.59%	7
Streptococcus_sanguinis	0.54%	8