INTERNATIONAL DOCTORAL
SCHOOL OF THE USC

Alba
Regueira Iglesias

PhD Thesis

# Limitations of 16S rRNA gene as phylogenetic marker: a large-scale meta-omics analysis of plaque microbiota in periodontal diseases

Santiago de Compostela, 2022

**Doctoral Programme in Dental Science**

DOCTORAL THESIS

# Limitations of 16S rRNA gene as phylogenetic marker: a large-scale meta-omics analysis of plaque microbiota in periodontal diseases

Alba Regueira Iglesias

**INTERNATIONAL PHD SCHOOL OF THE UNIVERSITY OF SANTIAGO DE COMPOSTELA**

**PHD PROGRAMME IN DENTAL SCIENCE**

SANTIAGO DE COMPOSTELA

2022

**DECLARACIÓN DEL AUTOR/A DE LA TESIS**

D./Dña. **Alba Regueira Iglesias**

Título de la tesis: **Limitations of 16S rRNA gene as phylogenetic marker: a large-scale meta-omics analysis of plaque microbiota in periodontal diseases**

Presento mi tesis, siguiendo el procedimiento adecuado al Reglamento y declaro que:

1) La tesis abarca los resultados de la elaboración de mi trabajo.
2) De ser el caso, en la tesis se hace referencia a las colaboraciones que tuvo este trabajo.
3) Confirmo que la tesis no incurre en ningún tipo de plagio de otros autores ni de trabajos presentados por mí para la obtención de otros títulos.
4) La tesis es la versión definitiva presentada para su defensa y coincide la versión impresa con la presentada en formato electrónico.

Y me comprometo a presentar el Compromiso Documental de Supervisión en el caso que el original no esté depositado en la Escuela.

En **Santiago de Compostela**, **02 de febrero de 2022**.

**Firma electrónica**

# AUTORIZACIÓN DE LOS DIRECTORES DE LA TESIS

**Limitations of 16S rRNA gene as phylogenetic marker: a large-scale meta-omic analysis of plaque microbiota in periodontal diseases**

Dª. Inmaculada Tomás Carmona
D. Javier Tamames De la Huerta
D. Víctor Manuel Arce Vázquez

INFORMA/N:

Que la presente tesis, se corresponde con el trabajo realizado por Dª. Alba Regueira Iglesias, bajo nuestra dirección, y autorizamos su presentación, considerando que reúne los requisitos exigidos en el Reglamento de Estudios de Doctorado de la USC, y que como directores de esta no incurre en las causas de abstención establecidas en la Ley 40/2015.

De acuerdo con lo indicado en el Reglamento de Estudios de Doctorado, declaramos también que la presente tesis doctoral es idónea para ser defendida en base a la modalidad Monográfica con reproducción de publicaciones, en los que la participación de la doctoranda fue decisiva para su elaboración y las publicaciones se ajustan al Plan de Investigación.

En Santiago de Compostela, 2 de Febrero de 2022

"Nothing in life is to be feared, it is only to be understood. Now is the time to understand more, so that we may fear less."

**Marie Curie**

## Acknowledgments

A mi tutora la Dra. Inmaculada Tomás, por la dedicación, minuciosidad, y pasión puestas en cada trabajo realizado, que han sido una fuente de inspiración y motivación constantes para querer dar lo mejor de mí. Poder contar con la capacidad y experiencia de una investigadora excepcional, en un ámbito de confianza y trabajo en equipo; ha sido un privilegio. Las enseñanzas académicas y humanas transmitidas a lo largo de estos años son incalculables. Gracias.

A mis directores el Dr. Javier Tamames y el Dr. Víctor Manuel Arce, grandes investigadores, por su participación en el desarrollo de este trabajo.

Al equipo de la Dra. María José Carrera, perteneciente al Centro de Investigación Singular en Tecnologías Inteligentes la Universidad de Santiago de Compostela (CiTIUS); especialmente a Carlos Balsa y a Lara Vázquez, por los conocimientos bioinformáticos y estadísticos, y por el tiempo aportados para que esta Tesis haya salido adelante.

A la Dra. Marta Relvas por la aportación de muestras orales, y a la Dra. Manuela Alonso por llevar a cabo los análisis de laboratorio realizados en esta Tesis.

A todas mis compañeras de la Unidad de Pacientes con Necesidades Especiales coordinada por la Dra. Tomás, por todos los momentos compartidos; y en especial, a Triana Blanco por su labor en la selección de los pacientes y en la recogida de muestras. Es un placer formar parte un equipo con personas tan maravillosas.

A mi pilar fundamental: mi familia; a mis amigas, y a José, por haberme acompañado durante este largo y, a veces complicado, camino. Gracias por apoyarme cuando decidí seguir esta ruta y compartir conmigo la felicidad de cada pequeño paso hacia delante. Pero, sobre todo, gracias por estar ahí para levantarme cuando el sendero se acomplejaba y el entusiasmo desaparecía, y por siempre recordarme que puedo hacer lo que me proponga.

# RESUMO DA TESE

# Resumo da Tese

## “Limitacións do xene ARNr 16S como marcador filoxenético: unha análise meta-ómica a gran escala da microbiota da placa nas enfermidades periodontais”

### INTRODUCIÓN

O termo “enfermidades periodontais” refírese a unha serie de condición diferentes iniciadas pola biopelícula dental que afectan ós tecidos que rodean e soportan os dentes (1). A xenxivite é unha inflamación localizada que non se estende ó aparello de unión periodontal e é reversíbel reducindo os niveis de placa (2). Sen embargo, en individuos susceptíbeis, se non é tratada pode progresar a periodontite, a cal se caracteriza pola destrución gradual do aparello de unión do dente (1). Estas patoloxías xorden cando se perde o equilibrio entre a biopelícula microbiana e o sistema inmune debido ben a unha disbiosis ou a unha reacción esaxerada do hóspede ós microbios. Ademais, atópanse entre as enfermidades orais máis prevalecentes e con maiores consecuencias en todo o mundo (3), cunha prevalencia global da periodontite severa en 2015 do 7.4% (4). No mesmo ano, as estimacións para España revelaron que un 5% da poboación adulta de entre 34 e 44 anos e un 10% entre 65 e 74, tiveron bolsas periodontais profundas (≥6 mm) (5).

A periodontite exerce un efecto negativo sobre a calidade de vida das persoas que a padecen, especialmente naquelas con periodontite severa, comprometendo aspectos relacionados tanto coa función como coa estética (6). Especificamente, se non se aplica ningún tratamento, os resultados poden levar á perda de dentes, un deterioro do rendemento mastigatorio, un peor estado nutricional, unha menor autoestima e, incluso, pode ter efectos negativos sobre a saúde xeral (7). En relación con isto, dende hai anos observouse que unha serie de enfermidades e afeccións sistémicas poden afectar o aparello de inserción periodontal. Na actualidade, unha gran cantidade de investigacións apoian esta relación bidireccional entre a periodontite e as enfermidades cardiovasculares (8), a diabetes mellitus (9), as afeccións

respiratorias (10), a artrite reumatoide (11), a enfermidade de Alzheimer (12) e os resultados adversos do embarazo (13).

O inicio e progreso da periodontite están relacionados con múltiples factores etiolóxicos e factores de risco modificables e non modificables (7,14), sendo de especial importancia a interacción entre os microorganismos locais e a resposta inmune do hóspede. Como consecuencia, o desenvolvemento de enfoques terapéuticos efectivos require a identificación dos principais microbios e marcadores do hóspede asociados á periodontite.

Poñendo o foco sobre o compoñente microbiano, o noso coñecemento da etioloxía e patoxénese das enfermidades periodontais cambiou ó longo do tempo grazas 1) á hipótese de que estas condicións están causadas por biopelículas e non por microbios nun estado planctónico, 2) ó emprego de conceptos ecolóxicos para estudar a microbiota oral (a poboación de microorganismos que coloniza unha parte do copo) e, especialmente, 3) ás melloras tecnolóxicas nos métodos empregados para estudar os microbios presentes nas mostras orais (15).

## XUSTIFICACIÓN E OBXECTIVOS

Os enormes avances producidos no campo da microbioloxía nas últimas décadas debido ó desenvolvemento e implantación das tecnoloxías de secuenciación de próxima xeración son innegables (16). De xeito específico, a secuenciación do xene ARN ribosomal (ARNr) 16S, considerado por moitos como o marcador filoxenético definitivo grazas principalmente á súa presenza ubicua en bacterias e arqueas e á intercalación de zonas conservadas e variábeis (17); permitiu estudar as comunidades microbianas complexas como a oral a profundidades sen precedentes (18).

Non obstante, os resultados de investigación poden verse afectados por múltiples fontes de posibles nesgos durante cada paso do fluxo de traballo da secuenciación do xene ARNr 16S (19-21). A selección do par de cebadores é un deses pasos (19-21). Os cebadores constrúense en base a secuencias de consenso, pero poden presentar discordancias con algúns taxons que poden levar á sobre- ou infrarepresentación dun grupo microbiano concreto (22). En consecuencia, o uso dun cebador non apropiado podería dar como resultado conclusións

biolóxicas cuestionables sobre o nicho a estudar (22). Sen embargo, non hai ningunha análise sobre a cobertura dos cebadores empregados para detectar os microorganismos procariotas que habitan na boca humana, entendendo como cobertura a porcentaxe de coincidencias para un determinado grupo de secuencias/rango taxonómico.

Por outro lado, algúns nesgos asociados co fluxo de traballo da secuenciación son limitacións inherentes do propio xene (23). De feito, nun primeiro exemplo, varios autores demostraron a existencia de múltiples copias do xene ARNr 16S nos xenomas procariotas (24-29), o que afecta as estimanzas de abundancia baseadas en contas do xene de xeito que taxons cun menor número de xenes tenden a ser infraestimados e aqueles cun maior número son sobreestimados (24,27). En segundo lugar, as nove rexións variábeis do xene teñen diferentes graos de heteroxeneidade de secuencia (26,30). Ademais, algunhas especies diferentes poder ten amplicóns coincidentes (en inglés matching amplicons, MAs), definidos como aqueles cunha similitude do 100% e o mesmo número de nucleótidos. Outras especies, pola contra, comparten secuencias altamente similares, incluso por riba do comunmente empregado limiar do 97% para construír unidades operacionais taxonómicas (operational taxonomic units, OTUs) (27,31), o que significa que poden ser agrupadas de forma errónea no mesmo OTU. Este agrupamento afecta á construción das táboas de OTUs e, por extensión, ás asignacións taxonómicas e ós resultados de diversidade. Con todo, a pesar das cuestións mencionadas anteriormente, non se realizaron investigacións exhaustivas sobre a mellor maneira de avaliar o número de xenes intraxenómicos do ARNr 16S nas bacterias e arqueas que ocupan a cavidade oral, nin sobre o impacto que ten o cebador elixido para as diferentes rexións na detección de MAs ou de amplicóns moi similares de taxons distintos. Ademais, segue sendo necesario avaliar a cuestión de cantos taxons orais diferentes e que especies orais específicas poden agruparse erroneamente no mesmo OTU, en función do cebador utilizado.

A comparación dos estudos do microbioma periodontal baseados na secuenciación é controvertida polas significativas diferenzas metodolóxicas en pasos relevantes dentro do fluxo de traballo típico. É amplamente coñecido que cada tecnoloxía de secuenciación funciona de xeito diferente na relación entre a lonxitude da lectura, o rendemento da secuencia e a taxa de error (21), sendo Illumina preferible a Roche 454 e Ion Torrent (19). Tamén, como mencionamos anteriormente, as diferentes rexións variábeis e, en consecuencia, os amplicóns

derivados delas teñen distintos graos de heteroxeneidade de secuencia (26,30). En consecuencia, parece cuestionable a comparación das secuencias de mostras orais obtidas mediante tecnoloxías de secuenciación e rexións xenéticas distintas. Con todo, ata a data ningunha investigación avaliou as secuencias obtidas en estudos sobre o microbioma periodontal presente en diferentes condicións de saúde, en particular as xeradas mediante a plataforma de alto rendemento Illumina e distinguidas pola rexión do xene ARNr 16S máis amplificada.

Debido á falta de evidencia sobre as cuestións sinaladas anteriormente, esta Tese tiña os seguintes obxectivos:

1) Analizar *in silico* a cobertura dos pares de cebadores empregados nos estudos baseados na secuenciación da microbiota oral; para iso utilizaranse dúas bases de datos específicas para a boca que conteñen secuencias do xene ARNr 16S de especies bacterianas e de arqueas.

2) Analizar *in silico* o número de xenes ARNr 16S nos xenomas completos das especies bacterianas e de arqueas que habitan na boca humana. Ademais, avaliar como o uso de diferentes pares de cebadores dirixidos a rexións distintas afecta á detección de MAs de taxons diferentes, identificando así as especies orais que teñen MAs.

3) Analizar *in silico* o rendemento de diferentes pares de cebadores de distintas rexións para identificar distintas especies procariotas orais con valores de similitude do amplicón do xene ARNr 16S ≥97%, establecendo así as especies orais que poden agruparse erroneamente no mesmo OTU.

4) Analizar os perfís da comunidade microbiana na placa supraxinxival e subxinxival de 2045 doentes con diferentes condicións periodontais (sans, xenxivite, periodontite e periodontite tratada) en relación coa diversidade bacteriana, os patróns de redes de co-ocorrencia e os modelos preditivos; sendo as secuencias utilizadas da plataforma Illumina, cun enfoque na rexión 3-4, e tratadas co mesmo protocolo bioinformático.

## OBXECTIVO 1. Avaliación *in silico* e selección dos mellores cebadores do xene ARNr 16S para o seu uso na secuenciación de próxima xeración para detectar bacterias e arqueas orais

### 1.1 MATERIAL E MÉTODOS

A través da base de datos PubMed e utilizando o software estatístico R (32) e o paquete RISmed (33), realizáronse procuras automáticas para elaborar unha listaxe de: 1) cebadores dos xenes ARNr 16S empregados para detectar e amplificar bacterias e arqueas en mostras orais antes da secuenciación masiva; e 2) especies de arqueas que habitan na boca humana. Tras aplicar técnicas avanzadas de análise de texto a tódolos resumos descargados utilizando o paquete tm de R (34), quedamos con 129 estudos sobre bacterias e 16 sobre arqueas que implicaban o uso de polo menos un cebador diferente do xene ARNr 16S, e con 53 artigos que contiñan información sobre especies orais de arqueas.

Identificáronse un total de 444 cebadores do xene ARNr 16S: 204 directos (forward, F), 230 reversos (reverse, R) e 12 non identificados (unidentified, UI). Deles, 278 obtivéronse das procuras en PubMed e 166 extraéronse do artigo de Klindworth et al. (35). A tódolos cebadores asignóuselles un identificador único baseado na súa procedencia -"OP" para os cebadores orais e "KP" para os de Klindworth (35)- e a súa dirección (F, R ou UI), seguido dun número de tres díxitos. Trala comparación das secuencias 5'-3' de tódolos cebadores, identificáronse 75 coas mesmas secuencias; o que nos deixou con 369 cebadores do xene do ARNr 16S diferentes. Por outra banda, obtivemos 177 nomes diferentes de especies de arqueas orais.

A base de datos de Escapa et al. (36), que inclúe 223.143 variantes da secuencia do amplicón (amplicon sequence variants, ASVs) das secuencias do xene ARNr 16S, contén erros de anotación que fan imposible calcular a posición correcta dos cebadores dentro de cada secuencia en caso de coincidencia. Co obxectivo de mellorala, desenvolvemos scripts en Python (37) e Bash (38). Primeiro, as secuencias do xene do ARNr 16S das ASV do mesmo nivel xerárquico separáronse en 769 arquivos fasta diferentes. Despois, inseriuse un identificador de especie a tódalas secuencias antes da xerarquía taxonómica. As secuencias da mesma xerarquía aliñáronse simultaneamente utilizando Clustal Omega (39) contra un conxunto de secuencias do xene ARNr 16S de *Escherichia coli*. Tódolos ocos creados por Clustal Omega (39) foron eliminados, salvo os inseridos dende o inicio ata o primeiro nucleótido de cada secuencia. Os

arquivos fasta aliñados combináronse nun único arquivo para crear unha base de datos de ASVs completamente aliñadas, sendo a posición un o primeiro nucleótido de *E. coli* J01859.1. Por último, recortáronse as secuencias aliñadas con bases nunha posición inferior á do primeiro nucleótido de J01859.1, e aquelas con nucleótidos por enriba da posición 2000.

A continuación, na base de datos de nucleótidos non redundantes do Centro Nacional de Información Biotecnolóxica (National Centre for Biotechnology Information, NCBI) (40), procuramos os xenomas completos das 177 especies de arqueas orais. A través dun script en Python (37) e cos identificadores, puidemos descargar 193 xenomas de RefSeq (41) e oito de GenBank (42). O script completouse co módulo search_16S.py (43), baseado no algoritmo de Edgar (44), que nos permitiu: detectar e extraer as secuencias do xene ARNr 16S dos xenomas completos descargados, eliminar tódalas secuencias repetidas e almacenar as variantes identificadas nun arquivo fasta. O módulo e a integración da ferramenta "The Entrez Programming Utilities (E-utilities)" (45) en Biopython (46) permitiu obter e asignar fácil e automaticamente o rango taxonómico completo ós xenes. Ademais, buscáronse as secuencias do xene do ARNr 16S das especies sen identificadores xenómicos completos (47-49). Finalmente, tódalas secuencias do xene agrupáronse nun único arquivo fasta.

As secuencias neste arquivo foron empregadas para facer BLASTN (50,51) contra a base de datos de nucleótidos non redundantes do NCBI (40). A continuación, utilizando unha cobertura de busca ≥98% e unha porcentaxe de identidade ≥99%, descargáronse as secuencias do xene ARNr 16S e as rexións aliñadas cos xenomas completos; e ambos tipos de secuencias foron tratadas como ASVs. A base de datos de arqueas orais creouse utilizando outro script, e contén 2842 ASVs. As secuencias da base de datos foron aliñadas e melloradas seguindo os mesmos pasos que na base de bacterias.

Para levar a cabo a análise *in silico* dos cebadores, definíronse as coberturas a nivel de variante (variant coverage, VC): porcentaxe de coincidencias dun cebador concreto en relación co total de secuencias da base de datos; e a nivel de especie (species coverage, SC): porcentaxe de especies con coincidencias en polo menos unha das súas variantes de secuencia cando se utiliza un cebador concreto. As coincidencias entre cebadores e secuencias das bases de datos avaliáronse aplicando as expresións regulares do módulo regex (52) de Python (37).

Os cebadores individuais cunha SC ≥75,00% foron escollidos e tódalas combinacións posibles entre F e R foron identificadas. Estimouse a lonxitude media entre estas dúas posicións para clasificar os pares de cebadores nunha das tres categorías de lonxitude media dos amplicóns: 1) curta (short, S)= 100 a 300 pares de bases (base pairs, bps); 2) media (medium, M)= 301 a 600; e 3) longa (long, L)= >600. Tódolos pares de cebadores obtidos foron avaliados con respecto ás bases de datos de bacterias e arqueas, o que nos permitiu determinar se eran específicos de dominio ou para ambos.

### 1.2 RESULTADOS

Un total de 148 e 65 cebadores individuais tiveron valores de SC de bacterias e arqueas ≥75,00%, respectivamente. Tras aplicar os criterios de formación de pares de cebadores, 3993 combinacións bacterianas e 645 de arqueas foron posibles. Delas, 156 estiveron repetidas para ambos dominios, e o resto foron específicas de dominio.

*Pares de cebadores específicos de bacterias*

Na categoría de lonxitude S, 139 pares tiñan valores de SC bacteriana ≥95,00% (rango= 99,09% - 95,19%), mentres que 33 tamén tiñan unha SC de arqueas de 0,00%. Estes últimos amplificaban as rexións xénicas 3-4 ou 5-7 e tiveron valores de SC bacteriana que oscilaron entre o 97,92% e o 95,58%, o que significou que non se cubriron entre 16 e 34 especies de bacterias orais. Para a maioría deles, a lonxitude media dos seus amplicóns foi de ó redor de 186 (rango= 189 - 182). Destaca o par OP_F009-OP_R030 da rexión 5-7, cunha lonxitude media de 297 e un valor de SC bacteriano do 96,88%, polo que só 24 especies de bacterias orais non foron cubertas por este par.

Na categoría de lonxitude M, 68 pares de cebadores tiñan valores de SC bacteriana ≥95,00% (rango= 98,83% - 95,06%), dos cales 45 tiñan unha SC de arqueas de 0,00%. Os seus valores de SC bacteriana tamén oscilaron no rango anterior o que significou que entre nove e 38 especies non foron cubertas. Ademais, estes pares dirixíanse ás rexións xénicas 3-5, 3-6 ou 4-7, e tiñan lonxitudes de lectura medias entre 566 e 454. Dos pares coas maiores lonxitudes medias de amplicón, os que proporcionaron a mellor cobertura foron, por orde: KP_F051-OP_R030; OP_F021-OP_R030; KP_F048-OP_R073; KP_F051-KP_R053; OP_F021-KP_R053; e OP_F050-OP_R073 (rango de SC bacteriana= 98,83% - 96,23%; rango de

lonxitude de lectura media= 566 - 546). Estes amplificaron as rexións 3-6 ou 4-7 e non cubriron entre nove e 29 especies de bacterias.

Na categoría de lonxitude L, 20 pares de tiñan valores de SC bacteriana ≥95,00% (rango= 97,14% - 95,06%), e 17 tamén tiñan un valor de SC de arqueas de 0,00%. Estes últimos pares tiñan o mesmo rango de SC bacteriano e deixaban entre 22 e 38 especies sen cubrir. Todos eles dirixíanse á rexión xénica 3-7 e tiñan unha lonxitude media de lectura de entre 772 e 732. Os cebadores co mellor equilibrio entre a lonxitude media de lectura e a cobertura foron KP_F048-KP_R074 (SC bacteriana= 97,01%; lonxitude media de lectura= 767); e OP_F050-KP_R074 (96,36%; 766). Con todo, houbo opcións interesantes con lonxitudes >1000 bps e valores de SC bacteriana ≥90,00% (rango de SC bacteriana= 93,37% - 90,64%; rango de lonxitude media de lectura= 1066 - 1059). Neste sentido, KP_F048-KP_R060, KP_F048-KP_R076 e KP_F048-OP_R121 da rexión 3-9 tiñan lonxitudes de lectura medias de 1061, 1060 e 1060, respectivamente, e valores de SC bacteriana do 93,37%; deixando sen cubrir 51 especies de bacterias orais.

*Pares de cebadores específicos de arqueas*

Na categoría de lonxitude S, 12 cebadores tiñan valores de SC de arqueas ≥95,00% (rango= 98,45% - 95,36%). Deles, oito tiñan valores de SC bacterianas do 0,00%: OP_F066-KP_R013; KP_F059-KP_R013; KP_F016-KP_R002; KP_F018-KP_R003; OP_F066-KP_R006; KP_F018-OP_R102; KP_F059-KP_R006; e KP_F018-KP_R002. A súa SC de arqueas oscilou entre o 95,88% e o 95,36%, tiveron lonxitudes de lectura medias de 275 a 144, amplificaron as rexións xénicas 3 ou 5-6 e, por último, non cubriron entre oito e nove especies de arqueas orais.

Dezanove pares de cebadores na categoría de lonxitude M tiñan valores de SC de arqueas ≥95,00% (rango= 97,42% - 95,36%). Entre eles, nove tiñan tamén un valor de SC bacteriana do 0,00%: KP_F018-KP_R031; KP_F018-KP_R032; KP_F018-KP_R035; KP_F018-OP_R020; KP_F018-OP_R070; KP_F020-KP_R006; KP_F020-KP_R013; KP_F016-KP_R032; e OP_F114-KP_R006. Estes amplificaban as rexións 3-5 ou 3-6 e tiveron unha lonxitude media de amplicón de 551 a 414. Os pares cubriron entre 95,88% e 95,36% das especies de arqueas, deixando entre oito e nove sen cubrir.

Só un par de cebadores na categoría de >600 bps tiña un valor SC ≥95,00% na base de datos de arqueas: OP_F114-KP_R013; o cal tamén tiña un valor de SC bacteriana do 0,00%. Este par amplificaba a rexión 3-6, tivo unha lonxitude media de 679 e non detectou oito especies de arqueas. Vinte e sete pares de cebadores tiñan unha SC de arqueas ≥90,00%, unha SC bacteriana de 0,00% e unha lonxitude media >679, e 10 dos cales eran superiores a 1100 (rango de lonxitude media= 1131 - 681). Deles, o mellor equilibrio entre a cobertura e a lonxitude media do amplicón atopouse en: KP_F016-KP_R066; KP_F016-KP_R063; KP_F018-KP_R066; e KP_F018-KP_R063. A SC de arqueas foi do 92,78% para os dous primeiros pares e do 93,81% para os dous segundos, deixando 14 ou 12 especies, respectivamente, sen cubrir. Todos estes dirixíanse á rexión 3-9 e tiñan, por orde, lonxitudes medias de amplicón de 1129, 1128, 1119 e 1118.

*Pares de cebadores de bacterias e arqueas*

Dez pares da categoría S tiñan valores de SC bacteriana e de arqueas ≥95,00% (rango= 95,97% - 95,32%; e 99,48% - 97,94%, respectivamente). A súa lonxitude media oscilou entre 288 e 284 e todos amplificaron a rexión 4-5: KP_F020-KP_R031; KP_F020-OP_R070; KP_F020-KP_R032; KP_F020-KP_R035; KP_F020-OP_R020; KP_F020-KP_R038; KP_F020-OP_R010; KP_F020-OP_R014; KP_F020-OP_R036; e KP_F020-OP_R048. O número de especies bacterianas e de arqueas non cubertas por estes pares oscilou entre 31 e 36; e entre unha e catro, respectivamente.

Na categoría M, dous cebadores tiñan valores de SC bacteriana e de arqueas ≥95,00%: OP_F114-OP_R070 (SC bacteriana= 95,58%; SC de arquea= 98,45%); e OP_F114-KP_R031 (95,71%; 98,45%). Ambos amplificaron a rexión 3-5 e tiveron lonxitudes medias de 460 e 457, respectivamente. Non cubriron 33 (OP_F114-KP_R031) ou 34 (OP_F114-OP_R070) especies bacterianas e tres de arqueas. Ó baixar o corte a SC ≥90,00%, atopamos seis pares cunha secuencia media máis longa. Entre eles destacou OP_F114-OP_R073, cunha lonxitude media de 549, dirixíase á rexión xénica 3-6 e presentaba uns valores de SC de bacterias e arqueas do 94,80% e o 93,30%, respectivamente. Non cubriu 40 especies de bacterias e 13 de arqueas.



Ningún par de cebadores da categoría L tivo valores SC ≥95,00% en ningunha das bases de datos. Pola contra, 28 tiñan SC bacterianas e de arqueas ≥90,00% (rango= 94,54% - 90,64%; e 96,91% - 96,39%, respectivamente). Estes amplificaron as rexións 3-9, 4-9 ou 5-9, tiñan

lonxitudes medias entre 622 e 1063, e non cubrían de 42 a 72 especies bacterianas e de seis a sete de arqueas. A combinación de OP_F066 con KP_R060, KP_R076 e OP_R121 produciu os valores de cobertura máis altos, e todos eles dirixíronse á rexión 5-9. Estes cebadores non cubriron 42 especies de bacterias e 6 de arqueas. Con todo, as súas lonxitudes medias foron 623, 622 e 622, respectivamente. Os pares de cebadores formados por OP_F114 con KP_R060, KP_R076 ou OP_R121, da rexión 3-9, presentaban un mellor equilibrio entre os resultados de cobertura (SC bacteriana= 91,42%, SC de arqueas= 96,91%) e as lonxitudes medias das secuencias (1063, 1062 e 1062). Sesenta e seis bacterias e seis arqueas non foron detectadas.

### 1.3 CONCLUSIÓNS

Tendo en conta as tres categorías de lonxitude media do amplicón, os pares de cebadores coa mellor cobertura estimada para detectar as bacterias orais dirixíronse ás rexións 3-4, 4-7 e 3-7, e foron: KP_F048-OP_R043 (posición do par de cebadores para *E. coli* J01859.1: 342-529), KP_F051-OP_R030 (514-1079) e KP_F048-OP_R030 (342-1079). Para a detección de arqueas orais, os pares con mellor cobertura amplificaron as rexións 5-6, 3-6 e 3-6, e foron: OP_F066-KP_R013 (784-indefinido), KP_F020-KP_R013 (518-indefinido) e OP_F114-KP_R013 (340-indefinido). Os pares coa mellor cobertura dos dominios de bacterias e arqueas conxuntamente atopáronse nas rexións 4-5, 3-5 e 5-9, e foron: KP_F020-KP_R032 (518-801), OP_F114-KP_R031 (340-801) e OP_F066-OP_R121 (784-1405). Os pares de cebadores coa mellor cobertura identificados neste estudo non se atopan entre os máis empregados na literatura sobre o microbioma oral.

**OBXECTIVO 2. Impacto da redundancia do xene ARNr 16S e da selección dos pares de cebadores na cuantificación e clasificación da microbiota oral na secuenciación de próxima xeración**

### 2.1 MATERIAL E MÉTODOS

Un total de 518 xenomas co estado de secuenciación completo indicado polo sitio web da base de datos ampliada do microbioma oral humano (expanded human oral microbiome database, eHOMD) (53) foron escollidos entre os 2074 dispoñibles. Estes xenomas teñen un ou máis identificadores Genbank (42), e usáronse para obter as secuencias completas almacenadas na base de datos do NCBI (41,42,54). Ademais, a listaxe de 177 especies de arqueas orais cos seus identificadores do GenBank (42), obtidos no obxectivo 1, permitíronnos acceder ós seus xenomas completos no NCBI (40).

A integración da ferramenta E-utilities (45) no script de Python (37) permitiunos adquirir as URLs necesarias para recuperar a información de interese de varias bases de datos do NCBI (41,42,54). Os xenomas completos das bacterias e arqueas orais foron descargados e, finalmente, obtívose a taxonomía de cada unha delas.

A continuación, desenvolvemos un script en Python (37) para detectar os nucleótidos non específicos da Unión Internacional de Química Pura e Aplicada (International Union of Pure and Applied Chemistry, IUPAC) distribuídos ó longo dalgúns xenomas, e substituílos aleatoriamente por un dos nucleótidos específicos equivalentes. Outros xenomas foron excluídos porque tiñan un exceso de nucleótidos IUPAC.

Co módulo search_16S.py (43) integrado no script, detectáronse e extraéronse as secuencias do xene ARNr 16S dos xenomas completos, e almacenáronse as variantes nun arquivo fasta. A tódalas variantes do xene designóuselles unha taxonomía a nivel cepa ou especie se non existía un nome de cepa. Isto deixounos con 518 xenomas de bacterias orais, correspondentes a 186 especies, e 191 xenomas de arqueas orais de 135 especies.



Para cada xenoma avaliado, calculamos: o seu tamaño; os tamaños dos xenes ARNr 16S; o número total de xenes ARNr 16S; o número de variantes diferentes; e o número de xenes de ARNr 16S en cada cadea. As medias dos datos obtidos determináronse posteriormente

utilizando os módulos NumPy (55) e pandas (56) de Python (37) para os niveis xerárquicos superiores á cepa.

Despois, escollemos os pares de cebadores cos mellores valores de cobertura *in silico* no obxectivo 1, así como os máis utilizados na literatura sobre o microbioma oral. Isto deixounos con 33 e 6 pares de cebadores, respectivamente, que se clasificaron segundo a lonxitude media dos amplicóns nas tres categorías descritas no obxectivo 1.

As secuencias F e R de cada par de cebadores utilizáronse en combinación co módulo regex (52) de Python (37) para obter os amplicóns *in silico* dos xenes ARNr 16S identificados nos xenomas descargados. Para cada par de cebadores, determinouse: o tamaño medio e o número dos amplicóns do xene ARNr 16S; o número de variantes do xene; o número de xenomas e especies detectadas; e a porcentaxe de cobertura a nivel de especie sen amplicóns coincidentes (species coverage with no MA, SC-NMA):

SC-NMA (%)= [(Número de especies detectadas – Número de especies con MA) /Total de especies avaliadas] x 100

Ademais, tamén se calculou a sobreestimación da abundancia a nivel de especie (overestimation factor, OF), a cal representou para cada especie a combinación do número de copias dos amplicóns do xene ARNr 16S e o número de MA. Para eliminar a sobreestimación derivada da redundancia xénica intraxenómica, dividiuse o OF de cada especie polo número de copias do xene, dando como resultado o OF causado pola presenza de MA (OF-MA). As especies con valores iguais a 1,00 non tiñan amplicóns que coincidisen con outras especies para o par de cebadores correspondente, mentres que as que tiñan estimacións superiores a 1,00 si os tiñan. Para cada par de cebadores, ambos parámetros expresáronse de forma acumulada e como media. Os mellores foron aqueles cun valor SC-NMA máis alto e, destes, os que tiñan o valor OF-MA máis baixo. Os peores pares de cebadores foron os que tiñan o menor SC-NMA e o maior OF-MA.



**2.2 RESULTADOS**

Os 518 xenomas de bacterias orais tiñan un tamaño medio de 2.933.660,68 bps e un número medio de 4,55 xenes ARNr 16S intraxenómicos que, á súa vez, tiñan un tamaño medio de

1.501,32 e unha media de 2,60 variantes. Once especies bacterianas (5,91%) tiñan un xene por xenoma, 159 (85,49%) tiveron unha media de entre dous e seis, e 16 (8,60%) valores medios de sete ou máis xenes. En canto ó número medio de variantes xénicas intraxenómicas, 63 especies bacterianas (33,87%) presentaron unha variante por xenoma, 118 (63,44%) entre dúas e seis, e cinco especies (2,69%) sete ou máis.

Pola contra, os 191 xenomas de arqueas orais tiñan un tamaño medio de 2.545.441,40 bps e unha media de 1,95 xenes ARNr 16S intraxenómicos que, á súa vez, tiñan un tamaño medio de 1.471,25 e unha media de 1,44 variantes. Sesenta e catro especies de arqueas (47,41%) tiñan unha media dun xene por xenoma, 67 (49,63%) mostraban entre dous e tres xenes e catro especies (2,96%) valores medios superiores a tres. En canto ó número medio de variantes xénicas intraxenómicas, 93 especies (68,89%) tiñan un número medio dunha variante por xenoma e 42 (31,11%) tiñan entre dúas e tres.

A análise dos pares de cebadores revelou que as combinacións seleccionadas por nós detectaron amplicóns nun rango do 99,46% ao 88,71% das especies bacterianas e do 99,26% ao 90,37% das especies de arqueas. Con todo, estas porcentaxes foron menores para os pares de cebadores máis empregados na literatura sobre o microbioma oral (95,16% - 74,19% para bacterias, e 63,70% e 30,37% para arqueas).

En xeral, excluíndo os pares de cebadores máis empregados na literatura, a diferenza dos valores de cobertura, os valores de SC-NMA aumentan a medida que aumenta a lonxitude media dos amplicóns. Se contrastamos as porcentaxes de especies detectadas coa súa SC-NMA, os pares de cebadores curtos analizados mostraron as maiores diferenzas entre ambos os parámetros (diferenza media= 21,34% para bacterias e 23,70% para arqueas), seguidos dos de lonxitude media (7,30% e 13,75%, respectivamente). Os pares de cebadores longos presentaron as menores diferenzas entre os valores de cobertura e SC-NMA (4,30% e 5,82%, respectivamente).



Segundo os valores SC-NMA, os tres mellores pares de cebadores específicos para bacterias foron: KP_F048-OP_R030 e KP_F048-KP_R060 (lonxitude longa, SC-NMA= 93,55%, seis MA, OF-MA= 1,06 para ambos) e OP_F053-KP_R020 (M, 93,01%, seis, 1,06).

Pola contra, o peor foi OP_F066-KP_R040 (S, 47,31%, 77, 2,78). Os pares de cebadores máis empregados na literatura non destacaron polos seus valores SC-NMA entre os da súa categoría.

Considerando as tres categorías de lonxitudes de amplicón, os valores SC-NMA dos pares de cebadores específicos para arqueas oscilaron entre 89,63% con KP_F018-KP_R063 (L, seis MA, OF-MA= 1,11), 85,93% con KP_F022-KP_R063 (M, oito, 1,14) e 69,63% con OP_F066-KP_R013 (S, 35, 1,99). Curiosamente, o par longo KP_F014-KP_R011, que é o máis utilizado na literatura para detectar arqueas orais, só puido identificar o 30,37% das especies avaliadas neste estudo, o que deu lugar ao valor SC-NMA máis baixo (26,67%, cinco MA, OF-MA= 1,14).

En relación cos pares de cebadores de bacterias e arqueas, os valores SC-NMA globais oscilaron entre o 92,52% de OP_F114_OP_R121 (L, 12 MA, OF-MA= 1,08), o 88,79% de OP_F114-KP_R031 (M, 29, 1,26) e o 54,21% de OP_F066-OP_R073 (S, 134, 3,45). En termos de SC-NMA global, o segundo peor foi KP_F078-OP_R010 (S, 66,67%, 48 MA, OF-MA= 1,68), principalmente debido á súa baixa capacidade para detectar arqueas (63,70%), o que afectou directamente ó valor SC-NMA para arqueas (48,89%). Con todo, este cebador é o máis empregado na literatura para detectar ambos dominios.

Dependendo do par de cebadores avaliado, entre o 46,70% e o 1,29% das especies bacterianas e entre o 38,89% e o 4,65%. das especies de arqueas tiñan MA. Ademais, houbo especies con MA obtidas con polo menos 10 pares de cebadores diferentes. No caso das bacterias, estas especies pertencían ós xéneros: *Actinomyces*, *Cronobacter*, *Fusobacterium*, *Klebsiella*, *Lacticaseibacillus*, *Lactobacillus*, *Staphylococcus* e *Streptococcus*; e para arqueas a: *Haloarcula*, *Halomicrobium*, *Methanosarcina*, *Nitrososphaera*, *Pyrococcus* e *Thermococcus*.

### 2.3 CONCLUSIÓNS

O funcionamento dos pares de cebadores para detectar especies sen MAs aumenta a medida que aumenta a lonxitude media dos amplicóns; ningún deles sendo os pares de cebadores máis utilizados na literatura sobre o microbioma oral. Os mellores pares de cebadores foron KP_F048-OP_R030 (para bacterias; rexión 3-7; posición do par de cebadores con respecto a *E.*

*coli* J01859.1: 342-1079), KP_F018-KP_R063 (para arqueas; 3-9; indefinido-1506), e OP_F114_OP_R121 (para ambos dominios; 3-9; 340-1405).

Ademais da redundancia do xene ARNr 16S, a considerable presenza de amplicóns coincidentes debe controlarse para garantir a interpretación precisa dos datos de diversidade microbiana. De feito, dependendo do par de cebadores utilizado, ata case a metade das especies orais tiñan MA, o que afectaba a xéneros relevantes presentes na cavidade bucal como *Actinomyces*, *Fusobacterium*, *Lactobacillus*, *Methanosarcina*, *Staphylococcus* ou *Streptococcus*.

O SC-NMA é un parámetro máis útil que a porcentaxe de cobertura convencional para seleccionar os mellores pares de cebadores. A elección do par de cebadores afecta de xeito significativo ás estimacións de diversidade e á clasificación taxonómica, condicionando a comparativa dos estudos da microbiota oral que utilizan diferentes pares de cebadores.

## OBXECTIVO 3. Detección *in silico* de especies procariotas orais con segmentos de secuencia do ARN 16S altamente similares usando diferentes pares de cebadores

### 3.1 MATERIAL E MÉTODOS

Os 518 xenomas cun estado de secuenciación completo indicado polo sitio web da eHOMD (53) foron seleccionados. Estes teñen un ou máis identificadores do GenBank (42), que se empregaron para acceder ás secuencias completas almacenadas na base de datos do NCBI (40). Ademais, a lista de 177 arqueas orais diferentes e os seus correspondentes identificadores de GenBank (42), obtida no obxectivo 1, permitiunos acceder ás súas secuencias completas e anotacións na base de datos do NCBI (40). A integración da ferramenta E-utilities (45) no noso script de Python (37) permitiunos obter as URLs necesarias para recuperar a información de interese de varias bases de datos do NCBI (41,42,54). A xerarquía taxonómica completa de tódolos xenomas completos descargados foi designada polo identificador taxonómico incluído na información anotada na base de datos do NCBI (39).

Un total de 33 pares de cebadores coa mellor cobertura *in silico*, segundo o obtido no obxectivo 1, foron escollidos xunto cos seis pares de cebadores máis utilizados na literatura sobre o microbioma oral. Estes clasificáronse segundo a lonxitude media dos amplicóns nas tres categorías de lonxitude media antes mencionadas.

Aplicando o noso script en combinación co módulo regex (52) de Python (37), utilizáronse as secuencias F e R de cada par de cebadores para obter segmentos de secuencia *in silico* dos xenomas analizados (en diante, amplicóns *in silico*). Tódolos amplicóns *in silico* da mesma especie, aínda que fosen de diferentes cepas, consideráronse para a análise. Para cada par de cebadores, creouse un arquivo fasta no que se almacenaron tódolos amplicóns *in silico* atopados. As secuencias almacenadas identificáronse coa mesma xerarquía taxonómica que os xenomas dos que se detectaron. Como se detectaron moitos amplicóns *in silico* dentro da mesma especie (que difiren en polo menos un nucleótido), as variantes da secuencia identificáronse cunha numeración correlativa nun novo nivel xerárquico por baixo do nome da especie.

Nos arquivos fasta, tódalas secuencias incluían un identificador de especie (SPn) e un identificador de variante (Vn) na súa cabeceira e, tamén, a cabeceira de cada secuencia incluía a xerarquía taxonómica ata o nivel da variante dentro de cada especie. Finalmente obtivéronse os amplicóns *in silico* de 186 especies de bacterias orais e 135 de arqueas.

Por outro lado, desenvolveuse un script co wrapper NcbiblastnCommandline de Biopython (46) para manexar BLAST+ 2.11 (57) en modo local dende Biopython. Isto permitiu transferir facilmente os datos obtidos nos aliñamentos para a súa posterior análise en Python (37). Os parámetros de aliñación configuráronse para que fosen os mesmos que os predeterminados en MegaBLAST (58). Tódalas secuencias pertencentes ó mesmo arquivo fasta para o mesmo par de cebadores aliñáronse entre si; para iso, cada arquivo fasta inseriuse como suxeito e consulta en BLASTN (51) para obter a porcentaxe de similitude entre amplicóns *in silico* pertencentes a diferentes especies orais.

Dos resultados obtidos, seleccionáronse os amplicóns *in silico* cunha cobertura de aliñación do 100% das secuencias de consulta e cun valor de similitude ≥97%. Dos aliñamentos obtidos, descartáronse os seguintes por non ser de interese: 1) amplicóns *in silico* co mesmo identificador único (SPn + Vn); 2) amplicóns *in silico* co mesmo identificador de especie; e 3) aliñacións duplicadas.

Se dúas especies diferentes tiñan máis dun valor de similitude de amplicón *in silico* ≥97% (amplicon similarity value ≥97%, ASI97) entre elas, elixíase un ó azar. Os resultados dos pares de especies altamente similares almacenáronse empregando os módulos de Python (37) pandas (56) e xlsxwriter (59) .

A continuación, creouse unha matriz de similitude para cada par de cebadores. Mediante un script en R (32) calculamos para cada par de cebadores 1) o número de especies con polo menos un ASI97 con outras especies; 2) o número total de ASI97 entre especies diferentes; 3) o número medio e máximo de ASI97 por especie.



Tamén se estimou para cada par de cebadores a porcentaxe de especies detectadas (SC) e a porcentaxe de especies detectadas sen ASI97 (species coverage with no ASI97, SC-NASI97).

Este último parámetro empregouse como criterio para seleccionar aqueles cebadores asociados a un menor número de especies orais que puidesen estar agrupadas de forma errónea.

Por último, describíronse os pares de especies que mostraban un ASI97 e avaliouse se pertencían a xéneros ou rangos taxonómicos superiores diferentes.

### 3.2 RESULTADOS

Os pares de cebadores dirixidos ás bacterias obtiveron unha media de 91,88 (49,40%) especies bacterianas cun ASI97 e unha media de 153,46 ASI97 con especies distintas. No caso dos dirixidos a arqueas, estas cifras foron de 65,60 (48,59%) e 162,26, respectivamente. Se se exclúen os cebadores máis utilizados na literatura sobre o microbioma oral, os de lonxitudes de amplicón curtas (a diferenza das porcentaxes de SC) tiveron os valores máis baixos de SC-NASI97 tanto para as bacterias (S= 39,54%) como para arqueas (S= 40,44%) en comparación cos cebadores de lonxitude media e longa (M= 45,82% e 46,35%, respectivamente; L= 48,39% e 44,32%, respectivamente).

Polo que respecta ós pares de cebadores específicos para bacterias, o número de especies bacterianas cun ASI97 e o número total de ASI97 oscilou entre 37 e 32 cun dos cebadores máis utilizados, KP_F031-KP_R021 (M; SC-NASI97= 54,30%), e 120 e 277 con OP_F066-KP_R040 (S; SC-NASI97= 24,19%), respectivamente. Este último cebador tamén tivo o valor SC-NASI97 máis baixo, mentres que OP_F053-KP_R020 detectou o maior número de especies sen ASI97 (M; SC-NASI97= 65,05%). Ademais, excepto OP_F053-KP_R020, tódolos cebadores específicos de bacterias tiñan un número máximo de ASI97/especies superior a cinco (rango= 15 - 4 ASI97/especies).

En canto ós pares de cebadores específicos para arqueas, o número de especies de arqueas cun ASI97 e o número total de ASI97 oscilou entre 24 e 96 co amplamente utilizado KP_F014-KP_R011 (L; SC- NASI97= 12,59%) e 89 e 240 con OP_F066-KP_R013 (S; SC-NASI97= 29,63%), respectivamente. O primeiro cebador detectou o menor número de especies sen ASI97, e KP_F018-KP_R002 o maior (S; SC- NASI97= 51,11%). Ademais, tódolos cebadores específicos para arqueas tiñan un número máximo de ASI97/especie ≥10 (rango= 13 - 10 ASI97/especie).

Empregando os pares de cebadores para bacterias e arqueas, o número de especies bacterianas e de arqueas cun ASI97 e o número total de ASI97 oscilou entre 84 e 60 e 118 e 126, respectivamente, con OP_F114-KP_R002 (S; SC-NASI≥97= 47. 31% para bacterias e 54,81% para arqueas) a 124 e 95 e 239 e 286, respectivamente, con OP_F066-OP_R073 (S; SC-NASI≥97= 31,18% para bacterias e 22,96% para arqueas). Este último cebador tamén detectou o menor número de especies sen ASI97 e OP_F114-KP_R031 o maior (M; SC-NASI97= 51,08% para bacterias e 53,33% para arqueas). A maioría das combinacións de cebadores de bacterias e arqueas tiñan un número máximo de ASI97/especies ≥10 (rango= 14 - 9 ASI/especies e 14 - 11 ASI/especies para ámbolos dous dominios, respectivamente).

Doutra banda, 149 (80,11%) das especies de bacterias orais e 108 (80,00%) das especies de arqueas orais avaliadas tiñan un ASI97 con polo menos unha especie distinta. Entre elas destacan pola súa relevancia na cavidade oral as bacterias: *Campylobacter concisus, Campylobacter curvus, Rothia dentocariosa, Streptococcus mitis, Streptococcus mutans, Streptococcus oralis,* e *Tannerella forsythia*; e as arqueas: *Halovivax ruber, Methanosalsum zhiliniae, Methanosarcina barkeri, Methanosarcina mazei,* e *Methanosarcina vacuolata*; entre outras. Ademais, houbo 30 especies bacterianas e 27 de arqueas que puideron agruparse cun máximo de ≥10 especies diferentes cando se empregaron tódolos pares de cebadores analizados. A maioría destas pertencían ós xéneros bacterianos: *Streptococcus* and *Staphylococcus*; e de arqueas: *Methanosarcina*, *Thermococcus,* and *Pyrococcus*. Pola contra, 37 (19,89%) especies bacterianas e 27 (20,00%) de arqueas non tiñan ASI≥97% con outros taxons orais.

Tódolos primers dirixidos a bacterias permitíronnos detectar 4450 relacións dous a dous entre 408 pares de especies bacterianas diferentes cun ASI97. Dezaoito destes pares de especies foron obtidos cos 29 pares de cebadores avaliados (frecuencia= 29; número de veces que un par de taxons teñen un ASI97 nos diferentes pares de cebadores), e pertencían ós xéneros *Actinomyces*, *Lactobacillus*, *Neisseria*, *Staphylococcus*, e *Streptococcus*. Sen embargo, 50 pares de especies con ASI97 só se detectaron cun primer (frecuencia= 1). Aínda que as relacións dous a dous implicaban maioritariamente a especies dos mesmos xéneros (3641; 81,82%), 809 relacións (18,18%) estaban constituídas por taxons de xéneros diferentes. A combinación de especies de *Klebsiella* con outras de *Cronobacter* foi a máis frecuente (frecuencia= 99), seguida de *Klebsiella-Serratia*, *Escherichia-Klebsiella*, *Cronobacter-Escherichia* e *Aggregatibacter-*

*Haemophilus* (frecuencias= 67 - 28). En canto ós rangos taxonómicos superiores, 293 (6,58%) relacións déronse entre pares de especies cun ASI97 pertencentes a familias distintas e mesmo houbo 26 (0,58%) relacións entre pares de especies cun ASI97 de ordes distintas.

Os cebadores dirixidos a arqueas permitiron detectar 3232 relacións dous a dous entre 340 pares de especies de arqueas diferentes cun ASI97. Tódolos pares de cebadores analizados identificaron sete pares de especies (frecuencia= 20), que pertencían ós xéneros *Methanobrevibacter* e *Methanocaldococcus*. Houbo 66 pares de especies detectadas unha soa vez por un só par de cebadores (frecuencia= 1). Unha vez máis, a maioría das relacións de dous a dous foron entre especies arqueas do mesmo xénero (2359, 72,99%), pero 873 (27,01%) relacións implicaban pares de taxons cun ASI97 de xéneros distintos. A combinación de especies de *Pyrococcus* e *Thermococcus* foi a máis frecuente (frecuencia= 428), seguida de *Palaeococcus* e *Thermococcus* (frecuencia= 109). Para os rangos taxonómicos superiores, 35 (1,08%) relacións eran pares de especies cun ASI97 de familias distintas, tres (0,09%) de ordes, e unha de clases (0,03%) distintas.

### 3.3 CONCLUSIÓNS

Os pares de cebadores avaliados dirixidos a bacterias e/ou arquexas detectaron unha media de máis de 150 OTU potenciais que poderían conter especies diferentes, cando se utilizou o limiar de similitude do ≥97%. Segundo o parámetro SC-NASI97, os mellores pares de cebadores foron OP_F053-KP_R020 para bacterias (rexión 1-3; posición para *E. coli* J01859.1: 9-356); KP_F018-KP_R002 para arqueas (4; indefinido-532); e OP_F114-KP_R031 para ambas (3-5; 340-801).

Ó redor do 80% das especies de bacterias e arqueas orais analizadas tiñan un ASI97 con polo menos outra especie distinta. Estas especies altamente similares desempeñan distintas funcións na microbiota oral e pertencen a xéneros bacterianos como *Campylobacter*, *Rothia*, *Streptococcus* e *Tannerella*, e a xéneros de arqueas como *Halovivax*, *Methanosalsum* e *Methanosarcina*. Ademais, o 20% e o 30% das relacións de similitude dous a dous establecéronse entre especies de diferentes xéneros bacterianos e de arqueas, respectivamente. Mesmo taxons de familias, ordes e clases distintas puideron agruparse no mesmo OTU potencial.

Independentemente do par de cebadores empregado, a agrupación de secuencias cunha similitude ≥97% proporciona unha descrición inexacta das especies orais bacterianas e de arqueas, o que pode afectar en gran medida ós parámetros de diversidade microbiana. Como resultado, a agrupación de OTUs condiciona a credibilidade das asociacións entre algunhas especies orais e certas condicións de saúde e enfermidade. Isto limita de xeito significativo a comparativa dos resultados da diversidade microbiana reportados na literatura do microbioma oral.

## OBXECTIVO 4. Unha análise meta-ómica a gran escala da microbiota da placa nas enfermidades periodontais

### 4.1 MATERIAL E MÉTODOS

Un total de 120 participantes: 55 periodontalmente sans e 65 afectados por periodontite non tratada; que cumprían os criterios de inclusión preestablecidos foron recrutados. Os diagnósticos periodontais foron realizados por dous dentistas experimentados. A presenza de saúde periodontal ou de periodontite crónica xeneralizada de moderada a grave estableceuse segundo información clínica e radiográfica, aplicando criterios previamente publicados (60,61). Unha ou dúas semanas despois do exame inicial, recolléronse mostras de placa subxinxival de tódolos participantes.

O ADN total das mostras foi extraído e illado para avaliar a súa calidade e concentración. Neste punto, excluíronse dúas mostras subxinxivales do grupo san por non cumprir os requisitos de calidade e concentración. A continuación, realizouse unha amplificación por reacción en cadea da polimerasa (PCR) da rexión 3-4 do xene ARNr 16S (62) A secuenciación levouse a cabo na plataforma Illumina MiSeq con lecturas de 2x300 bps. As secuencias obtidas depositáronse no arquivo de lecturas de secuencias (sequence read archive, SRA) (63) co número de acceso PRJNA773202.

Doutra banda, tamén se incluíron na nosa investigación estudos previos sobre a diversidade microbiana na placa supraxinxival e subxinxival en individuos adultos con diferentes condicións periodontais. Incorporáronse tódolos estudos que empregaron cebadores da rexión 3-4, a tecnoloxía de secuenciación Illumina, e tiñan acceso ao repositorio das secuencias. O estándar de referencia para o diagnóstico dunha afección periodontal podía basearse unicamente en parámetros clínicos ou clínicos e radiográficos, independentemente dos criterios de referencia diagnósticos aplicados; pero tiña que ser reportado. Ademais, seleccionáronse aquelas investigacións nas que os metadatos de interese por mostra estaban correctamente asignados no repositorio; e para os que as secuencias almacenadas cumprían criterios adicionais de inclusión e exclusión.

En xullo de 2021 realizáronse 120 buscas en cada unha das bases de datos electrónicas PubMed, Scopus e Embase para identificar estudos de secuenciación mediante Illumina sobre

o microbioma periodontal utilizando dous conxuntos de termos relacionados con: 1) condicións de saúde periodontal, nichos orais e microbiota; e 2) a tecnoloxía de secuenciación do xene ARNr 16S. Realizáronse procuras adicionais na base de datos SRA (63) para garantir que examinaramos tódolos posibles bioproxectos de interese.

A manipulación dos datos identificados nas procuras realizouse utilizando o software R (32), e os resultados de cada unha almacenáronse individualmente nun arquivo txt. (PubMed) ou csv. (Scopus e Embase). Os duplicados foron detectados e eliminados. A continuación, os resumos analizáronse de xeito computacional mediante sete conxuntos de termos positivos utilizando os paquetes tm e tratamento da linguaxe natural (natural language processing, NLP) (34,64) e cada un recibiu 100 ou un punto por cada termo presente. As publicacións con puntuacións predefinidas seleccionáronse para as posteriores avaliacións manuais dos seus resumos e texto completo. O proceso automatizado de extracción de datos validouse previamente nunha serie limitada de artigos que cumprían os criterios de inclusión.

Os identificadores dos bioproxectos dos artigos seleccionados empregáronse para acceder á base de datos do SRA (63) e ó selector de execucións "SRA run selector" (https://www.ncbi.nlm.nih.gov/traces/study/). Entón, descargáronse e avaliáronse as táboas de metadatos alí depositadas. Os autores foron contactados nos casos necesarios para obter os metadatos ou con fins de aclaración. Neste punto, o noso bioproxecto, PRJNA773202, engadiuse ó total.

Coa información da base de datos SRA (63) construíuse manualmente unha táboa de metadatos para cada un dos bioproxectos, que incluía información sobre aspectos relacionados tanto co bioproxecto como coas variables demográficas e clínicas dos doentes ós que pertencía cada mostra.

En relación coas secuencias almacenadas, para cada bioproxecto descargouse a lista de identificadores (listas de acceso) correspondente ás mostras de interese en formato txt. Para descargar e almacenar as secuencias coas mencionadas listas de acceso, instalouse o software gratuíto SRA Toolkit (65) en modo local. A continuación desenvolveuse un script en Bash (38) en combinación cos comandos prefetch e fastq-dump do SRA Toolkit. As mostras de placas de

cada bioproxecto almacenáronse en arquivos fastq individuais para a súa posterior manipulación.

O preprocesamento e a avaliación da calidade das secuencias de cada arquivo fastq realizáronse con USEARCH (66). As secuencias foron aliñadas e ensambladas, aceptándose un máximo de cinco desaxustes e unha porcentaxe mínima de similitude do 90% para os 2x250 bps, e 10 bps e o 80% para os 2x300 bps. Permitíronse un máximo de dous desaxustes na secuencia de cada cebador individual e catro nun par. Por último, descartáronse as secuencias cun erro máximo esperado >1 ou cunha lonxitude mínima <300 bps.

Tódolos arquivos fasta dun determinado bioproxecto fusionáronse usando Bash (38). Isto, mais un script desenvolvido en R (32), permitiunos crear un arquivo de grupo para cada bioproxecto. Executáronse os comandos screen. seqs e unique. seqs de mothur (67) para obter o arquivo de nomes de cada bioproxecto. A continuación, tódolos arquivos fasta de tódolos bioproxectos fusionáronse nun único arquivo, xunto cos arquivos "grupos" e "nomes" para formar os arquivos “meta-omics”.

A continuación, realizouse un filtrado de baixa abundancia, no que se eliminaron as secuencias das mostras globais con valores de abundancia <500 recontos. Aplicouse o pipeline de mothur (67) para ASVs con lixeiras modificacións, incluíndo a aplicación da base de datos específica oral para a clasificación taxonómica de ASVs descrita por Escapa et al. (36). Permitíronse as secuencias cunha lonxitude >400; e eliminároronse as que tiñan máis de oito homopolímeros, as que se consideraron quimeras tras aplicar o algoritmo VSEARCH de mothur (67,68) e as clasificadas como taxons descoñecidos no nivel xerárquico máis alto. As secuencias non se agruparon en ningún nivel xa que o noso obxectivo era identificar e clasificar o maior número posible de secuencias no nivel ASV.

Unha vez completado o pipeline de mothur (67), exportáronse os seguintes arquivos meta-ómicos a R-bioconductor (69) para a súa posterior análise: a táboa de reconto, a xerarquía taxonómica a nivel ASV, a árbore filoxenética e a táboa de metadatos.

Ademais, dous autores avaliaron de xeito independente a calidade dos metadatos dos bioproxectos incluídos na nosa investigación empregando unha lista de comprobación que deseñamos para este fin, a cal contiña 19 variables relacionadas cos datos dispoñibles sobre os suxeitos da mostra. Segundo a puntuación obtida, os bioproxectos clasificáronse como: baixa calidade= 0,00 - 0,33; calidade media= 0,34 - 0,66; e alta calidade= 0,67 - 1,00.

Tamén se avaliou o número de mostras por bioproxecto e o número medio de secuencias por mostra en cada proxecto como parámetro de calidade. O número medio de secuencias dividiuse por 10.000 e este parámetro denominouse puntuación media da secuencia (average sequence score, ASS). Os valores de ASS interpretáronse como representativos de 1) <0,25, secuencias de moi baixa cantidade; 2) 0,25 - 0,75, - secuencias de baixa cantidade; 3) 0,75 - 1,0 - secuencias de cantidade aceptable; 4) 1,0 - 2,0 - secuencias de alta cantidade; e 5) >2,0 - secuencias de moi alta cantidade.

Por último, a análise estatística dos datos de secuenciación do ARNr 16S a nivel ASV realizouse segundo o protocolo proposto por McMurdie e Holmes (70), utilizando implementacións en R que incluían os paquetes phyloseq, DESeq2 e microbiome (71-73).

Para eliminar as mostras cun baixo número de secuencias, excluíronse aquelas con menos de 2500 (n= 62), polo que quedaron 2124 mostras. A continuación, creáronse grupos segundo o tipo de placa dental e o estado de saúde periodontal dos participantes, obténdose un total de 11 grupos (ordenados alfabeticamente):

1) Placa supraxinxival, saúde periodontal, sitios sans (Sup_x0HHx; n= 210).
2) Placa supraxinxival, xenxivite, sitios enfermos (Sup_x0GDx; n= 79).
3) Placa supraxinxival, periodontite, sitios enfermos (Sup_x0PDx; n= 493).
4) Placa supraxinxival, periodontite, sitios enfermos tratados (Sup_x1PDx; n= 81).
5) Placa subxinxival, saúde periodontal, sitios sans (Sub_x0HHx; n= 155).
6) Placa subxinxival, xenxivite, sitios enfermos (Sub_x0GDx; n= 20).
7) Placa subxinxival, periodontite, sitios sans (Sub_x0PHx; n= 62).
8) Placa subxinxival, periodontite, zonas enfermas (Sub_x0PDx; n= 768).
9) Placa subxinxival, periodontite, sitios enfermos tratados (Sub_x1PDx; n= 197).

10) Placa submucosa, peri-implantite, sitios sans (Imp_x0IHx; n= 18).
11) Placa submucosa, peri-implantite, sitios enfermos ( Imp_x0IDx; n= 41).

Os grupos Sub_x0GDx, Imp_x0IHx e Imp_x0IDx eliminánronse debido ó seu baixo tamaño de mostra (n= <50), deixando un total de 2045 mostras para analizar.

A relación entre as diferentes condicións de saúde periodontal e a microbiota da placa investigouse dende varias perspectivas. En primeiro lugar, utilizáronse os paquetes phyloseq e microbiome para obter os datos de diversidade alfa (71,73). Como indicadores da riqueza de taxons, calculáronse o reconto absoluto de ASVs e o índice de cobertura do 95%. Determináronse os índices de Shannon e Pielou como indicadores de diversidade e uniformidade (74,75). Utilizouse a U de Mann-Whitney para as análises comparativas.

Segundo, empregouse unha análise de compoñentes principais (principal component analysis, PCA) para visualizar a agrupación das mostras de placas en relación co seu estado de saúde. O paquete mixOmics (76) empregouse para obter os gráficos de dispersión das dúas compoñentes principais baseadas na abundancia relativa das ASVs, amosando os centroides de cada grupo clínico e as elipses que representan o intervalo de confianza do 96%. Utilizouse unha análise non paramétrica multivariante permutada da varianza (permutational multivariate analysis of variance, PERMANOVA) (77) para medir as diferenzas a nivel de comunidade entre os grupos. Estas análises realizáronse co paquete vegan (78).

Terceiro, utilizouse o paquete microbiome (73) para identificar as ASVs centrais ou do núcleo presentes cunha taxa de prevalencia de ≥75% en cada tipo de placa e cada condición periodontal.

En cuarto lugar, utilizouse o paquete DESeq2 (72) para identificar as ASV cos cambios máis significativos na abundancia diferencial para as distintas condicións periodontais. As abundancias diferenciais medíronse co valor log2foldchange (log2 FC), e as diferentes condicións compáraronse utilizando a proba de Wald coa corrección de Benjamini-Hochberg. As medidas foron estatisticamente significativas se o p axustado era <0,01.

En quinto lugar, realizouse a análise de redes de co-ocorrencia nos grupos clínicos con máis de 100 mostras, filtrando as ASV cunha abundancia do 0,01%. Utilizáronse os parámetros por defecto e o paquete SpiecEasi (79) para executar SparCC, e a matriz de correlación obtida filtrouse por unha puntuación de correlación absoluta maior ou igual a 0,5. A continuación, as redes visualizáronse co paquete igraph (80), onde cada nodo representa unha ASV, e cada aresta representa as correlacións entre as abundancias das ASV. Ademais, calculouse un conxunto de medidas para describir a topoloxía das redes resultantes, e en liña con Banerjee et al. (81), utilizouse unha puntuación combinada baseada nun valor de grao alto e un valor de centralidade da interrelación (betweenness centrality, BC) alto como limiar para definir as ASVs clave.

Por último, realizamos unha clasificación supervisada en forma de análise discriminante de mínimos cadrados parciais (sparse partial least-squares discriminant analysis, sPLS-DA) (82) para facilitar a categorización dos diferentes grupos clínicos e identificar as ASV que mellor distinguían dous grupos dentro de cada nicho. O sPLS-DA realizouse utilizando o paquete mixOmics (76). As ASVs cunha abundancia relativa <0,1% no total das mostras excluíronse previamente do desenvolvemento dos modelos preditivos. Construíronse curvas características operativas do receptor (receiver operating characteristic curve, ROC) coa taxa de positividade verdadeira (sensibilidade) en función da taxa de falsa positividade (1-especificidade), mentres que os valores da área baixo a curva (area under the curve, AUC) utilizáronse para distinguir entre cada grupo clínico na placa supraxinxival e subxinxival.

### 4.2 RESULTADOS

En primeiro lugar, os suxeitos recrutados por nós con periodontite tiñan uns valores de niveis de placa en toda a boca así como de sangrado ó sondaxe, profundidade de bolsa e perda de inserción clínica nos rexistros de toda a boca e nos dos sitios de mostra; significativamente máis altos que os suxeitos do grupo san ( $p<0,001$).

Por outra banda, avaliáronse 1159 artigos obtidos a través das bases de datos electrónicas e 39 bioproxectos do SRA. Un total de 32 artigos (32 bioproxectos) cumpriron os criterios de inclusión e, neste punto, engadimos o noso propio bioproxecto. Trala avaliación dos metadatos cinco artigos foron excluídos, e catro eliminánronse trala avaliación de mostras e secuencias.

Así, quedamos con 23 artigos (25 bioproxectos) para a súa inclusión na meta-análise, involucrando 2045 mostras distribuídas en oito grupos.

Tres dos 25 bioproxectos incluídos tiñan metadatos de alta calidade (rango= 0,95 - 0,77), catro eran de calidade media (rango= 0,45 - 0,34) e 18 de baixa (rango= 0,30 - 0,15). En xeral, aqueles con calidade media e baixa non incluían información sobre o tipo e a severidade da periodontite, a etnia ou os parámetros clínicos periodontais. Ademais, a maioría dos de baixa calidade non proporcionaban a idade ou o sexo dos participantes.

Dende o punto de vista do tamaño da mostra, 10 bioproxectos tiñan <50 mostras (40%), nove entre 50 e 100 (36%) e seis >100 (24%). Ningún tivo un ASS de <0,25. Catro bioproxectos cun total de 167 mostras (8,17% de tódalas analizadas), tiñan valores de ASS de 0,75 - 1,0; polo que eran dunha cantidade aceptable, con máis de 7500 secuencias por mostra. Oito bioproxectos e 422 mostras (20,63%) tiñan valores de ASS de 1,0 - 2,0; polo que se consideraron de cantidade alta, con 10.000 - 20.000 secuencias por mostra. Por último, 13 bioproxectos, cun total de 1.456 mostras (71,20%) tiveron un ASS >2,0, polo que se consideraron de moi alta cantidade, con máis de 20.000 secuencias por mostra.

En canto ás características dos estudos incluídos na presente análise, menos dun terzo deles (7/23 artigos + 1 bioproxecto propio non publicado; 29,17%) estableceron o diagnóstico periodontal empregando a nova Clasificación de Enfermidades e Condicións Periodontais e Peri-implantarias (83). En 13/24 investigacións (54,17%) os autores compararon os perfís microbianos en estados de saúde e enfermidade periodontal, mentres que en 10/24 (41,66%) só se avaliou a periodontite e nun (4,17%) só se puideron seleccionar cinco mostras sas. A placa dental que máis se recolleu foi a subxinxival (16/24; 66,66%), seguida da supraxinxival (4/24; 16,67%) ou de ambas (4/24; 16,67%). Ademais, 4/24 estudos (16,67%) avaliaron os cambios producidos na microbiota trala terapia periodontal non cirúrxica.

Con respecto ós resultados da alfa-diversidade, a riqueza da placa supraxinxival diminuíu significativamente do grupo san ata o de xenxivite e logo aumentou fortemente no de periodontite (mediana do número de ASVs observadas= 610,50, 474,00 e 892,00, respectivamente; índice de cobertura do 95%= 220,00, 130,00 e 288,00, respectivamente). Estes

dous parámetros diminuíron significativamente nas mostras despois da terapia en comparación coas recollidas antes (781,00 e 263,00), pero non alcanzaron os niveis do grupo san. Pola contra, os índices de diversidade e uniformidade mostraron unha tendencia ascendente continua dende a saúde ata a enfermidade, que continuou tralo tratamento (rango índice de Shannon= 4,75 - 4,07; rango índice de Pielou= 0,70 - 0,62).

Na microbiota subxinxival, detectáronse valores significativamente máis baixos no número de ASVs e no índice de cobertura do 95% nas zonas enfermas de periodontite con respecto á saúde (417,50 e 142,00 *vs.* 478,00 e 171,00). Trala terapia, produciuse un aumento significativo de ASVs observadas, superando incluso os niveis sans (507,00). Os índices de diversidade e uniformidade tenderon a aumentar nos sitios enfermos, aínda que só as comparacións do índice Pielou foron significativas. Así mesmo, non houbo variacións significativas entre os grupos de periodontite para a alfa-diversidade, excepto Pielou en Sub_x0PHx (0,65) *vs.* Sub_x0PDx (0,68) e *vs.* Sub_x1PDx (0,69), e o Shannon en Sub_x0PHx *vs.* Sub_x1PDx (4.05 *vs.* 4.20).

Ao contrastar ambas placas nos suxeitos co mesmo estado de saúde, observamos que a riqueza e a diversidade foron significativamente maiores no nicho supraxinxival que no subxinxival. A excepción foi o índice de Shannon dos grupos sans de ambas placas. Pola contra, os valores de uniformidade foron significativamente maiores na placa subxinxival, excepto no caso de Sup_x1PDx.

En canto á estrutura da comunidade bacteriana, as PCAs revelaron unha agrupación das mostras "supra" e "sub" en función do estado de saúde do suxeito e do sitio da mostra. As observacións visuais foron confirmadas polo PERMANOVA, con resultados significativos para tódalas comparacións.

Na comparación entre os dous nichos no mesmo estado de saúde periodontal, a PCA revelou unha agrupación das mostras segundo o tipo de placa recollida. Outra vez, as observacións visuais foron confirmadas polo PERMANOVA.



Con respecto ós resultados da microbiota do núcleo, 37 ASVs de 28 especies diferentes (0,48% e 5,31% dos taxons detectados, respectivamente) formaron o núcleo supraxinxival independentemente da condición periodontal, e representaron o 36,44% da abundancia total.

Considerando os diferentes grupos, o número de ASVs oscilou entre 41 - 76 (0,69% - 1,85%) e as especies entre 27 - 46 (5,76% - 10,02%). A abundancia relativa de ASVs do núcleo variou entre o 28,34 e o 48,08%.

A microbiota do núcleo subxinxival estaba formada por 24 ASVs de 18 especies (0,28% e 3,27%, respectivamente) e representaban o 23,70% da abundancia total. Considerando os diferentes grupos clínicos, o número de ASVs e especies oscilou entre 27 - 44 (0,34% - 0,85%) e 18 - 29 (3,78% - 5,57%), e a súa respectiva abundancia relativa, entre 23,50% - 32,30%.

Se se consideraban as dúas placas xuntas e independentemente do estado de saúde periodontal, a microbiota central estaba formada por 26 ASVs de 20 especies diferentes (0,29% e 3,62% dos taxons detectados, respectivamente), o que representaba o 27,50% da abundancia total.

Na análise das abundancias diferenciais, os números máis altos de ASVs e especies con abundancias diferenciais na placa "supra" obtivéronse para: Sup_x0GDx *vs.* Sup_x0PDx, cun total de 1290 ASVs e 243 especies con abundancias diferenciais (16,96% e 45,42% do total detectado polos dous grupos, respectivamente); Sup_x0HHx *vs.* Sup_x0GDx con 945 ASVs e 198 especies (15.09% e 39,52%); Sup_x0HHx *vs.* Sup_x1PDx con 926 ASVs e 210 especies (13,12% e 41,02%); e Sup_x0HHx *vs.* Sup_x0PDx, con 918 ASVs e 272 especies (12,17% e 51,52%). Pola contra, os números relativos máis baixos de ASVs e especies con abundancias diferenciais observáronse na análise de Sup_x0PDx *vs.* Sup_x1PDx (total= 660 ASVs, 8,95%; 145 especies, 27,62%).

Na placa "sub", os números relativos máis altos de ASVs e especies con abundancias diferenciais obtivéronse ao comparar: Sub_x0HHx *vs.* Sub_x0PDx cun total de 1074 ASVs de 273 especies (12,64% e 48,75%, respectivamente), e Sub_x0HHx *vs.* Sub_x1PDx con 1015 ASVs de 225 especies (14,45% e 41,67%). Pola contra, os números relativos máis baixos de ASVs e especies con abundancias diferenciais observáronse na análise de Sub_x0PHx *vs.* Sub_x0PDx (total= 364 ASVs; 4,28%; 156 especies; 27,81%) e Sub_x1PDx (total= 339 ASVs; 5,64%; 117 especies; 22,20%).

Ademais, na comparación entre placas no mesmo estado de saúde periodontal, a comparativa Sup_x0PDx *vs.* Sub_x0PDx revelou os maiores números relativos de ASVs e especies con abundancias diferenciais (total= 2367 ASVs, 27,34%; 349 especies, 62,21%). Pola contra, as estimacións relativas máis baixas observáronse para Sup_x1PDx *vs.* Sub_x1PDx (total= 198 ASVs, 3,85%; 72 especies, 14,55%).

En canto ós resultados das redes de co-ocorrencia, a cobertura da rede e o número de nodos en Sup_x0PDx foron lixeiramente superiores ós de Sup_x0HHx (2,54% e 187 fronte a 2,26% e 136, respectivamente). Ademais, o número de correlacións foi máis de tres veces maior no grupo enfermo (959 fronte a 290, respectivamente). Practicamente tódalas correlacións foron positivas en ambos os grupos, excepto unha en Sup_x0PDx. A rede enferma tiña menos subredes e un maior número de módulos (12 e 55 *vs.* 18 e 25), pero ambos grupos tiñan o mesmo número de módulos con máis de tres nodos. Na rede Sup_x0HHx, as tres principais ASVs clave foron *Streptococcus* unclassified ASV90, *R. dentocariosa* ASV2, e *S. oralis* subsp. *dentisani* clado 058 ASV1. As principais ASVs clave na rede Sup_x0PDx foron: *Streptococcus* unclassified ASV85, *Streptococcus sanguinis* ASV228, e *Streptococcus* unclassified ASV121.

Ó contrario que na placa "supra", a rede Sub_x0PDx mostrou a menor cobertura e número de nodos e correlacións, seguida de Sub_x1PDx con respecto á rede Sub_x0HHx (cobertura= 0,63% e 1,54% *vs.* 2,75%; nodos= 53 e 78 *vs.* 163; correlacións= 80 e 111 *vs.* 387). Tódalas correlacións no nicho subxinxival foron positivas. Aínda que houbo un número similar de subredes nos tres grupos (12 e 10), os grupos con periodontite tiñan menos módulos e menos módulos con máis de tres nodos co san (11 e 13 *vs.* 25, respectivamente; 4 e 6 *vs.* 11, respectivamente). Os principais taxons clave na rede Sub_x0HHx foron *Streptococcus* unclassified ASV85, *Fusobacterium* unclassified ASV14 e *Streptococcus* unclassified ASV90. Na rede Sub_x0PDx, foron: *Streptococcus* unclassified ASV121, *T. forsythia* ASV15, e *Streptococcus* unclassified ASV85. Por último, os principais taxons clave na rede Sub_x1PDx foron: *T. forsythia* ASV15, *Fusobacterium nucleatum* subsp. *vincentii* ASV10, e *S. oralis* subsp. *dentisani* clado 058 ASV1.



Por último, os modelos preditivos da placa supraxinxival que requiriron un menor número de ASVs foron os que distinguían a saúde periodontal da periodontite despois da terapia, e a xenxivite da periodontite tanto non tratada como tratada (rango=30 - 20), asociados a valores

de AUC >0,970. Pola contra, os modelos para diferenciar a saúde periodontal tanto da xenxivite como da periodontite estaban compostos por un maior número de ASVs (150 e 70, respectivamente) cos que alcanzaron valores de AUC ao redor de 0,895.

Entre as ASVs preditoras da saúde máis importantes, en orde de abundancia relativa atopábanse: *Leptotrichia hongkongensis* ASV24, *Neisseria macacae* ASV9, *Neisseria. bacilliformis* ASV281, *Capnocytophaga sputigena* ASV56 e *Capnocytophaga gingivalis* ASV93.

Para as preditoras da enfermidade, centrándonos nos valores de abundancia relativa en Sup_x0GDx, as ASVs máis relevantes foron *Campylobacter rectus* ASV20 e *Parvimonas* HMT110 ASV21, con capacidade para discriminar a xenxivite e a periodontite. *P.* HMT110 ASV21 tamén predixo a periodontite tratada.

As ASVs da placa supraxinxival dos xéneros *Corynebacterium*, *Eikenella*, *Streptococcus* e *Veillonella* xurdiron como preditoras tanto da saúde periodontal como da xenxivite ou a periodontite. Por exemplo, *S. oralis* subsp. *dentisani* clado 058, tivo catro ASVs que discriminaban a saúde; e tres que eran preditoras da periodontite. Ademais, detectouse un único ASV de *S. sanguinis* que predixo tanto saúde como periodontite.

Na microbiota subxinxival, os modelos preditivos que requiriron un menor número de ASVs foron Sub_x0PHx *vs.* Sub_x0PDx (20 ASVs) e Sub_x0HHx *vs.* Sub_x0PHx (50 ASVs), asociados a valores de AUC de 0,885 e 0,902, respectivamente. Os modelos restantes mostraron un maior número de ASVs preditoras, que van dende 80 ASVs para discriminar entre Sub_x0PDx *vs.* Sub_x1PDx (AUC= 0,796) ata 200 ASVs para Sub_x0HHx *vs.* Sub_x1PDx (AUC= 0,888).

Entre as ASVs preditoras de saúde máis importantes en orde de abundancia relativa atopábanse: *Granullicatella adiacens* ASV13, *Gemella haemolysans* ASV26, *Capnocytophaga leadbetteri* ASV126, *Aggregatibacter* HMT458 ASV145 e *Prevotella melaninogenica* ASV7. Ademais, *G. adiacens* ASV13 e *C. leadbetteri* ASV126 foron preditoras dos grupos Sub_x0PHx e Sub_x1PDx, respectivamente.

En canto ás preditoras da enfermidade, centrándonos nos valores de abundancia relativa observados en Sub_x0PDx, as ASVs máis relevantes foron *Peptostreptococcaceae* [XI][G-5] *saphenum* ASV129, *Dialister pneumosintes* ASV194, *Desulfobulbus* HMT041 ASV149 e *Mogibacterium timidum* ASV640. A primeira ASV tamén predixo sitios tratados e sans na periodontite ó comparar ambos modelos coa saúde periodontal, mentres que a segundo e a terceira tamén predixeron a periodontite tratada.

Dos resultados derivados dos modelos preditivos sobre a placa subxinxival, *F. nucleatum* subsp. *vicentii* foi a principal especie que mostrou un rendemento preditivo oposto. A ASV do núcleo e altamente abundante ASV10 foi unha forte preditora da periodontite detectado en varios modelos. Con todo, houbo outras ASVs menos abundantes que predixeron simultaneamente tanto a saúde como a periodontite.

En canto ás ASVs preditoras da saúde na placa "supra" e "sub", destacan pola súa abundancia as seguintes: *R. dentocariosa* ASV2, *Haemophilus parainfluenzae* ASV3, ASV78, ASV45, e ASV46, *Kingella oralis* ASV66, *Streptococcus vestibularis* ASV27 e *Actinomyces* HMT170 ASV119. Algunhas destas ASVs tamén se comportaron como fortes discriminantes de sitios sans na periodontite.

En canto ás preditoras da periodontite en ambas placas, atribúese especial atención polos seus valores de abundancia a: *T. forsythia* ASV15, *Filifactor alocis* ASV19, *Treponema denticola* ASV38 e ASV150, *Fretibacterium fastidiosum* ASV97, *Peptostreptococcaceae* [XI][G-4] HMT369 ASV124, *Streptococcus anginosus* ASV142, e *Peptostreptococcaceae* [XI][G-6] *nodatum* ASV189. Tamén, *F. fastidiosum* ASV97 e *S. anginosus* ASV142 foron preditoras de xenxivite na placa supraxinxival. Todos estes taxons predixeron a periodontite tratada na placa subxinxival, e *T. forsythia* ASV15 mesmo na supraxinxival.

As ASVs dos xéneros *Alloprevotella*, *Fusobacterium*, *Gemella*, *Granulicatella*, *Lachnoanaerobaculum* e *Ruminococcaceae* [G-1] discriminaron diferentes condicións clínicas entre ambas placas, emerxendo como preditoras da xenxivite na placa supraxinxival e da saúde na subxinxival. Exemplos disto foron: *Fusobacterium periodonticum* ASV11, *Lachnoanaerobaculum umeaense* ASV152 e *Granulicatella elegans* ASV207.

Centrándonos na correspondencia dos resultados dos modelos preditivos das principais ASVs con respecto ós seus respectivos resultados na abundancia diferencial, detectáronse numerosas inconsistencias biolóxicas nesta última. Na placa supraxinxival, 17 das 21 ASVs preditoras de saúde presentaban abundancias diferenciais, e o 88,23% (15/17) delas amosaban abundancias significativamente elevadas tanto en saúde como en periodontite. Das 25 ASVs preditoras de enfermidade, 22 eran diferencialmente abundantes e, delas, catro (18,18%) tiñan una tendencia a niveis elevados nas dúas condicións opostas.

Na placa subxinxival, 29 das 31 ASVs preditoras de saúde foron diferencialmente abundantes, cun 31,03% (9/29) amosando abundancias significativamente elevadas tanto en saúde como en periodontite. Das 31 ASVs preditoras de enfermidade, 30 amosaron abundancias diferenciais e, delas, tres (10,00%) presentaron niveis significativamente elevados nas dúas condicións opostas.

En ámbolos dous tipos de placa, 28 das 32 ASVs preditoras de saúde amosaron abundancia diferencial e, delas, o 71,42% (20/28) foron diferencialmente abundantes tanto para a saúde como para a periodontite. As 15 ASVs preditoras de enfermidade en ambas placas tamén foron diferencialmente abundantes e, delas, tres (20,0%) tiñan niveis significativamente elevados en ámbalas dúas condicións clínicas.

### 4.3 CONCLUSIÓNS

A riqueza bacteriana asociada á periodontite é maior que na saúde periodontal na placa supraxinxival e menor na subxinxival; a uniformidade é maior na enfermidade que na saúde en ambos os nichos. A microbiota supraxinxival é máis rica e diversa que a súa homóloga subxinxival para o mesmo estado de saúde periodontal. A estrutura da comunidade bacteriana é diferente para as distintas condicións periodontais na placa supraxinxival e subxinxival, así como para o mesmo estado de saúde entre os dous nichos.

O núcleo da microbiota da placa supraxinxival e subxinxival non permite caracterizar a saúde e a enfermidade periodontal, o que revela a gran heteroxeneidade da microbiota oral. A porcentaxe da comunidade bacteriana da placa dental que se organiza en redes de co-ocorrencia a nivel de ASV é moi pequena; a rede da periodontite non tratada da placa supraxinxival é máis

extensa e contén máis nodos, interconexións e grupos bacterianos interconectados que a súa homóloga subxinxival. As principais ASV clave nas redes de saúde periodontal da placa supraxinxival son *R. dentocariosa* ASV2 e *S. oralis* subsp. *dentisani* clade 058 ASV1. O eixo principal nas redes de periodontite non tratadas da placa supraxinxival é *S. sanguinis* ASV228; e na placa subxinxival, é *T. forsythia* ASV15. Os principais taxons clave na rede de periodontite tratada do nicho subxinxival son *T. forsythia* ASV15, *F. nucleatum* subsp. *vincentii* ASV10 e *S. oralis* subsp. *dentisani* clade 058 ASV1.

Unha pequena proporción dos taxons supra e subxinxivales teñen unha capacidade destacada para distinguir entre as condicións periodontais, e unha porcentaxe relevante son membros da microbiota central. Dende o punto de vista da metaxenómica clínica, a placa supraxinxival é un mellor biomarcador bacteriano que o seu homólogo subxinxival para diferenciar a saúde periodontal da periodontite non tratada e tratada.

As principais ASVs preditoras da saúde periodontal na placa supraxinxival e subxinxival son *R. dentocariosa* ASV2; *H. parainfluenzae* ASV3, ASV78, ASV45 e ASV46; *K. oralis* ASV66; *S. vestibularis* ASV27; e *A.* HMT170 ASV119. Pola contra, as principais ASVs preditoras da periodontite en ámbolos tipos de placa son: *T. forsythia* ASV15; *F. alocis* ASV19; *T. denticola* ASV38 e ASV150; *F. fastidiosum* ASV97; P. HMT369 ASV124; *S. anginosus* ASV142; e *P. nodatum* ASV189. Destas, *F. fastidiosum* ASV97 e *S. anginosus* ASV142 tamén actuaron como preditoras de xenxivite no nicho supraxinxival.

**REFERENCIAS**

(1) Kinane DF, Stathopoulou PG, Papapanou PN. Periodontal diseases. Nat Rev Dis Primers. 2017 Jun;3:17038. doi: 10.1038/nrdp.2017.38.

(2) Chapple ILC, Mealey BL, Van Dyke TE, Bartold PM, Dommisch H, Eickholz P, et al. Periodontal health and gingival diseases and conditions on an intact and a reduced periodontium: consensus report of workgroup 1 of the 2017 World Workshop on the Classification of Periodontal and Peri-Implant Diseases and Conditions. J Periodontol. 2018 Jun;89 Suppl 1:S74-84. doi: 10.1002/JPER.17-0719.

(3) Peres MA, Macpherson LMD, Weyant RJ, Daly B, Venturelli R, Mathur MR, et al. Oral diseases: a global public health challenge. Lancet. 2019 Jul;394(10194):249-60.

(4) Kassebaum NJ, Smith AGC, Bernabé E, Fleming TD, Reynolds AE, Vos T, et al. Global, regional, and national prevalence, incidence, and disability-adjusted life years for oral conditions for 195 countries, 1990–2015: a systematic analysis for the global burden of diseases, injuries, and risk factors. J Dent Res. 2017 Apr;96(4):380-7.

(5) Bravo-Pérez M, Almerich-Silla J, Ausina-Márquez V, Avilés-Gutiérrez P, Blanco-González J, Canorea-Díaz E, et al. Oral health survey in Spain 2015. RCOE. 2016 Jun;21(Suppl 1):8-48. Spanish.

(6) Ferreira MC, Dias-Pereira A, Branco-de-Almeida LS, Martins CC, Paiva SM. Impact of periodontal disease on quality of life: a systematic review. J Periodont Res. 2017 Aug;52(4):651-65.

(7) Chapple ILC, Bouchard P, Cagetti MG, Campus G, Carra M, Cocco F, et al. Interaction of lifestyle, behaviour or systemic diseases with dental caries and periodontal diseases: consensus report of group 2 of the joint EFP/ORCA workshop on the boundaries between caries and periodontal diseases. J Clin Periodontol. 2017 Mar;44 Suppl 18:S39-51. doi: 10.1111/jcpe.12685.

(8) Carrizales-Sepúlveda EF, Ordaz-Farías A, Vera-Pineda R, Flores-Ramírez R. Periodontal disease, systemic inflammation and the risk of cardiovascular disease. Heart Lung Circ. 2018 Nov;27(11):1327-34.

(9) Nascimento G, Leite F, Vestergaard P, Scheutz F, López R. Does diabetes increase the risk of periodontitis? A systematic review and meta-regression analysis of longitudinal prospective studies. Acta Diabetol. 2018 Jul;55(7):653-67.

(10) Gomes-Filho I, Cruz SSD, Trindade SC, Passos-Soares J, Carvalho-Filho P, Figueiredo ACMG, et al. Periodontitis and respiratory diseases: a systematic review with meta-analysis. Oral Dis. 2020 Mar;26(2):439-46.

(11) Fuggle NR, Smith TO, Kaul A, Sofat N. Hand to mouth: a systematic review and meta-analysis of the association between rheumatoid arthritis and periodontitis. Front Immunol. 2016 Mar;7:80. doi: 10.3389/fimmu.2016.00080.

(12) Leira Y, Domínguez C, Seoane J, Seoane-Romero J, Pías-Peleteiro JM, Takkouche B, et al. Is periodontal disease associated with alzheimer's disease? A systematic review with meta-analysis. Neuroepidemiology. 2017;48(1-2):21-31.

(13) Ide M, Papapanou PN. Epidemiology of association between maternal periodontal disease and adverse pregnancy outcomes – systematic review. J Clin Periodontol. 2013 Apr;40 Suppl 14:S181-94. doi: 10.1111/jcpe.12063.

(14) AlJehani YA. Risk factors of periodontal disease: review of the literature. Int J Dent. 2021 Feb;2021:8735071. doi: 10.1155/2021/8735071.

(15) Teles R, Teles F, Frias-Lopez J, Paster B, Haffajee A. Lessons learned and unlearned in periodontal microbiology. Periodontol 2000. 2013 Jun;62(1):95-162.



(16) A genomic approach to microbiology. Nat Rev Genet. 2019 Jun;20(6):311. doi: 10.1038/s41576-019-0131-5.

(17) del Rosario-Rodicio M, del Carmen-Mendoza M. Bacterial identification by 16S rRNA sequencing: rationale, methodology and applications in clinical microbiology. Enferm Infecc Microbiol Clin. 2004 Apr;22(4):238-45. Spanish.

(18) Zaura E. Next-generation sequencing approaches to understanding the oral microbiome. Adv Dent Res. 2012 Sep;24(2):81-5.

(19) Nearing JT, Comeau AM, Langille MGI. Identifying biases and their potential solutions in human microbiome studies. Microbiome. 2021 May;9(1):113. doi: 10.1186/s40168-021-01059-0.

(20) Robinson CK, Brotman RM, Ravel J. Intricacies of assessing the human microbiome in epidemiologic studies. Ann Epidemiol. 2016 May;26(5):311-21.

(21) de la Cuesta-Zuluaga J, Escobar JS. Considerations for optimizing microbiome analysis using a marker gene. Front Nutr. 2016 Aug;3:26. doi: 10.3389/fnut.2016.00026.

(22) Hamady M, Knight R. Microbial community profiling for human microbiome projects: tools, techniques, and challenges. Genome Res. 2009 Jul;19(7):1141-52.

(23) Rajendhran J, Gunasekaran P. Microbial phylogeny and diversity: small subunit ribosomal RNA sequence analysis and beyond. Microbiol Res. 2011 Feb;166(2):99-110.

(24) Acinas SG, Marcelino LA, Klepac-Ceraj V, Polz MF. Divergence and redundancy of 16S rRNA sequences in genomes with multiple rrn operons. J Bacteriol. 2004 May;186(9):2629-35.

(25) Pei AY, Oberdorf WE, Nossa CW, Agarwal A, Chokshi P, Gerz EA, et al. Diversity of 16S rRNA genes within individual prokaryotic genomes. Appl Environ Microbiol. 2010 Jun;76(12):3886-97.

(26) Sun D, Jiang X, Wu QL, Zhou N. Intragenomic heterogeneity of 16S rRNA genes causes overestimation of prokaryotic diversity. Appl Environ Microbiol. 2013 Oct;79(19):5962-9.

(27) Větrovský T, Baldrian P. The variability of the 16S rRNA gene in bacterial genomes and its consequences for bacterial community analyses. PLoS One. 2013;8(2):e57923. doi: 10.1371/journal.pone.0057923.

(28) Case RJ, Boucher Y, Dahllöf I, Holmström C, Doolittle WF, Kjelleberg S. Use of 16S rRNA and rpoB genes as molecular markers for microbial ecology studies. Appl Environ Microbiol. 2007 Jan;73(1):278-88.

(29) Lee ZM, Bussema C 3rd, Schmidt TM. rrnDB: documenting the number of rRNA and tRNA genes in bacteria and archaea. Nucleic Acids Res. 2009 Jan;37(Database issue):D489-93. doi: 10.1093/nar/gkn689.

(30) Johnson JS, Spakowicz DJ, Hong BY, Petersen LM, Demkowicz P, Chen L, et al. Evaluation of 16S rRNA gene sequencing for species and strain-level microbiome analysis. Nat Commun. 2019 Nov;10(1):5029. doi: 10.1038/s41467-019-13036-1.

(31) Schloss PD. Amplicon sequence variants artificially split bacterial genomes into separate clusters. mSphere. 2021 Aug;6(4):e0019121. doi: 10.1128/mSphere.00191-21.

(32) R Core Team. R: a language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing; 2021; Available at: https://www.R-project.org/.

(33) Kovalchik S. RISmed: download content from NCBI databases. 2017; Available at: http://www.CRAN.R-project.org/.

(34) Feinerer, I., Hornik, K., Meyer, D. Text mining infrastructure in R. J Stat Softw. 2008 Mar;25(5):1-54. doi: 10.18637/jss.v025.i05.

(35) Klindworth A, Pruesse E, Schweer T, Peplies J, Quast C, Horn M, et al. Evaluation of general 16S ribosomal RNA gene PCR primers for classical and next-generation sequencing-based diversity studies. Nucleic Acids Res. 2013 Jan;41(1):e1. doi: 10.1093/nar/gks808.

(36) F Escapa I, Huang Y, Chen T, Lin M, Kokaras A, Dewhirst FE, et al. Construction of habitat-specific training sets to achieve species-level assignment in 16S rRNA gene datasets. Microbiome. 2020 May;8(1):65. doi: 10.1186/s40168-020-00841-w.

(37) Python Software Foundation. Python. 2020; Available at: http://www.python.org/.

(38) GNU P. Free Software Foundation. Bash. 2020; Available at: http://www.gnu.org/.

(39) Sievers F, Wilm A, Dineen D, Gibson TJ, Karplus K, Li W, et al. Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. Mol Syst Biol. 2011 Oct;7:539. doi: 10.1038/msb.2011.75.

(40) NCBI Resource Coordinators. Database resources of the National Center for Biotechnology Information. Nucleic Acids Res. 2016 Jan;44(D1):D7-19. doi: 10.1093/nar/gkv1290.

(41) O'Leary NA, Wright MW, Brister JR, Ciufo S, Haddad D, McVeigh R, et al. Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. Nucleic Acids Res. 2016 Jan;44(D1): D733-45. doi: 10.1093/nar/gkv1189.

(42) Clark K, Karsch-Mizrachi I, Lipman DJ, Ostell J, Sayers EW. GenBank. Nucleic Acids Res. 2016 Jan;44:D67-72. doi: 10.1093/nar/gkv1276.

(43) Lyalina S. Search 16S py algorithm. 2019; Available at: https://github.com/slyalina/search_16S_py.

(44) Edgar R. SEARCH_16S: a new algorithm for identifying 16S ribosomal RNA genes in contigs and chromosomes. Preprint at bioRxiv 2017:124131. doi: 10.1101/124131.

(45) National Center for Biotechnology Information. Entrez programming utilities help. 2010; Available at: https://www.ncbi.nlm.nih.gov/books/NBK25501/.

(46) Cock PJ, Antao T, Chang JT, Chapman BA, Cox CJ, Dalke A, et al. Biopython: freely available Python tools for computational molecular biology and bioinformatics. Bioinformatics. 2009 Jun;25(11):1422-23.

(47) National Center for Biotechnology Information. NCBI RefSeq targeted loci project. Archaea FTP. 2008. ftp://ftp.ncbi.nlm.nih.gov/refseq/TargetedLoci/Archaea/.

(48) Quast C, Pruesse E, Yilmaz P, Gerken J, Schweer T, Yarza P, et al. 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.

(49) Parks DH, Chuvochina M, Chaumeil P, Rinke C, Mussig AJ, Hugenholtz P. A complete domain-to-species taxonomy for bacteria and archaea. Nat Biotechnol. 2020 Sep;38(9):1079-86.

(50) Morgulis A, Coulouris G, Raytselis Y, Madden TL, Agarwala R, Schäffer AA. Database indexing for production MegaBLAST searches. Bioinformatics. 2008 Aug;24(16):1757-64.

(51) Chen Y, Ye W, Zhang Y, Xu Y. High speed BLASTN: an accelerated MegaBLAST search tool. Nucleic Acids Res. 2015 Sep;43(16):7762-8.

(52) Barnett M. regex. 2020; Available at: https://pypi.org/.

(53) F Escapa I, 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;3(6):e00187-18. doi: 10.1128/mSystems.00187-18.

(54) Schoch CL, Ciufo S, Domrachev M, Hotton CL, Kannan S, Khovanskaya R, et al. NCBI Taxonomy: a comprehensive update on curation, resources and tools. Database (Oxford). 2020 Jan;2020:baaa062. doi: 10.1093/database/baaa062.

(55) Harris CR, Millman KJ, van der Walt, Stéfan J, Gommers R, Virtanen P, Cournapeau D, et al. Array programming with NumPy. Nature. 2020 Sep;585(7825):357-62.

(56) McKinney W. Data Structures for statistical computing in Python. In: van der Walt S, Millman J, editors. Proceedings of the 9th Python in Science Conference; 2010; Austin. Texas: SciPy; 2010. doi: 10.25080/Majora-92bf1922-00a.

(57) Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, et al. BLAST+: architecture and applications. BMC Bioinformatics. 2009 Dec;10:421. doi: 10.1186/1471-2105-10-421.

(58) Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. J Mol Biol. 1990 Oct;215(3):403-10.

(59) McNamara J. xlsxwriter. 2013; Available at: https://xlsxwriter.readthedocs.io/.

(60) Armitage GC. Development of a classification system for periodontal diseases and conditions. Ann Periodontol. 1999 Dec;4(1):1-6.

(61) Page RC, Eke PI. Case definitions for use in population-based surveillance of periodontitis. J Periodontol. 2007 Jul;78(7 Suppl):1387-99.

(62) Willis JR, González-Torres P, Pittis AA, Bejarano LA, Cozzuto L, Andreu-Somavilla N, et al. Citizen science charts two major "stomatotypes" in the oral microbiome of adolescents and reveals links with habits and drinking water composition. Microbiome. 2018 Dec;6(1):218. doi: 10.1186/s40168-018-0592-3.

(63) Leinonen R, Sugawara H, Shumway M, International Nucleotide Sequence, Database Collaboration. The sequence read archive. Nucleic Acids Res. 2011 Jan;39(Database issue):D19-21. doi: 10.1093/nar/gkq1019.

(64) Hornik, K. NLP: natural language processing infrastructure. 2020; Available at: https://CRAN.R-project.org/package=NLP.

(65) SRA Toolkit Development Team. Sequence read archive toolkit. Available at: http://www.ncbi.nlm.nih.gov/Traces/sra/sra.cgi?cmd=show&f=software&m=software&s=software.

(66) Edgar R. Search and clustering orders of magnitude faster than BLAST. Bioinformatics. 2010 Oct;26(19):2460-1.

(67) Schloss PD, Westcott SL, Ryabin T, Hall JR, Hartmann M, Hollister EB, et al. Introducing mothur: open-source, platform-independent, community-supported software for describing and comparing microbial communities. Appl Environ Microbiol. 2009 Dec;75(23):7537-41.

(68) Rognes T, Flouri T, Nichols B, Quince C, Mahe F. VSEARCH: a versatile open source tool for metagenomics. PeerJ. 2016 Oct;4:e2584. doi: 10.7717/peerj.2584.

(69) Gentleman RC, Carey VJ, Bates DM, Bolstad B, Dettling M, Dudoit S, et al. Bioconductor: open software development for computational biology and bioinformatics. Genome Biol. 2004 Sep;5(10):R80. doi: 10.1186/gb-2004-5-10-r80.

(70) McMurdie PJ, Holmes S. Waste not, want not: why rarefying microbiome data is inadmissible. PLoS Comput Biol. 2014 Apr;10(4):e1003531. doi: 10.1371/journal.pcbi.1003531.



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

(72) Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol. 2014 Dec;15(12):550. doi: 10.1186/s13059-014-0550-8.

(73) Lahti L, Shetty S. Tools for microbiome analysis in R. Microbiome package. 2017; Available at: http://microbiome.github.com/microbiome.

(74) Spellerberg IF, Fedor PJ. A tribute to Claude Shannon (1916–2001) and a plea for more rigorous use of species richness, species diversity and the 'Shannon–Wiener' index. Glob Ecol Biogeogr. 2003 Apr;12(3):177-9.

(75) Pielou EC. The measurement of diversity in different types of biological collections. J Theor Biol. 1966 Dec;13:131-44.

(76) Rohart F, Gautier B, Singh A, Lê Cao K. mixOmics: an R package for 'omics feature selection and multiple data integration. PLoS Comput Biol. 2017 Nov;13(11):e1005752. doi: 10.1371/journal.pcbi.1005752.

(77) Anderson M. A new method for non-parametric multivariate analysis of variance. Austral Ecol. 2001 Feb;26(1):32-46.

(78) Oksanen J, Blanchet FG, Friendly M, Kindt R, Legendre P, McGlinn D, et al. vegan: community ecology package. 2020; Available at: https://cran.r-project.org, https://github.com/vegandevs/vegan.

(79) Kurtz Z, Mueller C, Miraldi E, Bonneau R. SpiecEasi: sparse inverse covariance for ecological statistical inference. 2021; Available at: https://github.com/zdk123/SpiecEasi.

(80) Csardi G and Nepusz T. The Igraph software package for complex network research. InterJournal, Complex Systems. 2006; 1695; Available at: http://igraph.org.

(81) Banerjee S, Schlaeppi K, van der Heijden MGA. Keystone taxa as drivers of microbiome structure and functioning. Nat Rev Microbiol. 2018 Sep;16(9):567-76.

(82) Lê Cao KA, Boitard S, Besse P. Sparse PLS discriminant analysis: biologically relevant feature selection and graphical displays for multiclass problems. BMC Bioinformatics. 2011 Jun;12:253. doi: 10.1186/1471-2105-12-253.

# THESIS SUMMARY

# Thesis summary

## "Limitations of 16S rRNA gene as phylogenetic marker: a large-scale meta-omics analysis of plaque microbiota in periodontal diseases"

The first objective of the present Thesis was to analyse *in silico* the coverage of the 16S ribosomal RNA (rRNA) gene primers used to study the composition of the oral microbiota, using two databases containing 16S rRNA sequences from oral bacteria and archaea; and to describe the best primer pairs for each domain.

Searches were conducted in PubMed to create a list of 1) 16S rRNA gene primers used in sequencing-based studies of the oral microbiome, and 2) oral-archaea species inhabiting the human mouth. The individual primers found were evaluated against a previously reported database of 16S rRNA sequences from oral bacteria, which was modified by our group; and a self-created oral-archaea database, constructed based on the list of oral-archaeal species. Both databases contained the genomic variants detected for each included species. Primers were evaluated at the variant and species levels, and those with a species coverage (SC) ≥75.00% were selected for the pair analyses. All possible combinations of forward and reverse primers were identified and evaluated against the two databases.

A total of 369 distinct individual primers were found in the literature. After applying the primer-pair formation criteria, 4638 primer pairs were identified. The best bacteria-specific pairs targeted the 3-4, 4-7, and 3-7 16S rRNA gene regions, with SC levels of 98.83% - 97.14%; meanwhile, the optimum archaea-specific primer pairs amplified regions 5-6, 3-6, and 3-6, with SC estimates of 95.88%. Finally, the best pairs for detecting both domains targeted regions 4-5, 3-5, and 5-9, and produced SC values of 95.71% - 94.54% and 99.48% - 96.91% for bacteria and archaea, respectively.

Given the three amplicon length categories (100-300, 301-600, and >600 base pairs), the primer pairs with the best coverage values for detecting oral bacteria were: KP_F048-OP_R043 (region 3-4; primer pair position for *Escherichia coli* J01859.1: 342-529), KP_F051-OP_R030 (4-7; 514-1079), and KP_F048-OP_R030 (3-7; 342-1079). For detecting oral archaea, these were: OP_F066-KP_R013 (5-6; 784-undefined), KP_F020-KP_R013 (3-6; 518-undefined), and OP_F114-KP_R013 (3-6; 340-undefined). Lastly, for detecting both domains they were: KP_F020-KP_R032 (4-5; 518-801), OP_F114-KP_R031 (3-5; 340-801), and OP_F066-OP_R121 (5-9; 784-1405). The primer pairs with the best coverage identified herein are not among those described most widely in the oral microbiome literature.

The purpose of the second study was to evaluate the number of 16S rRNA genes in the complete genomes of all the bacterial and archaeal species ever detected in the human oral cavity; and to assess how the use of different primer pairs would affect the detection and classification of redundant amplicons and matching amplicons (MAs) from different taxa.

A total of 709 complete genomes (518 oral bacteria, 191 oral archaea) were downloaded from the NCBI database, and their complete 16S rRNA genes were extracted. The total number of genes and variants per genome were calculated. Next, 33 primer pairs selected from objective 1 and 6 commonly employed in the oral literature were used against all the genomes to obtain amplicons. For each primer pair, we calculated the number of 16S rRNA gene amplicons, variants, genomes, and species detected, as well as the percentage of coverage at the species level with no matching amplicons (SC-NMA).

In total, 94.1% of oral bacteria and 52.59% of oral archaea had more than one 16S rRNA gene in their respective genomes. Between 46.70% - 1.29% of the bacterial species and between 38.89% - 4.65% of the archaeal species detected by the evaluated primer pairs had MAs, affecting relevant genera present in the oral environment such as *Actinomyces, Fusobacterium, Lactobacillus, Methanosarcina, Staphylococcus,* and *Streptococcus.* The best primer pairs were (SC-NMA; region; primer pair position for *Escherichia coli* J01859.1): KP_F048-OP_R030 for bacteria (93.55%; 3-7; 342-1079), KP_F018-KP_R063 for archaea (89.63%; 3-9; undefined-1506), and OP_F114_OP_R121 for both bacteria and archaea (92.52%; 3-9; 340-1405).

In addition to the 16S rRNA gene redundancy, the considerable presence of MAs must be controlled to ensure the accurate interpretation of microbial diversity data. The SC-NMA is a more useful parameter than the conventional coverage percentage for selecting the best primer pairs. The performance of the primer pairs to detect non-MA species increases as the average length of the amplicons increases; none of these being the most widely used primer pairs in the oral literature. The choice of primer pair significantly affects diversity estimates and taxonomic classification, conditioning the comparability of oral microbiome studies using different primer pairs.

The aims of the third study were to evaluate *in silico* the coverage of a set of previously selected primer pairs to detect oral species having 16S rRNA sequence segments with ≥97% similarity; and to describe oral species with highly similar sequence segments and determine whether they belong to distinct genera or other higher taxonomic ranks.

Thirty-nine primer pairs were employed to obtain the *in silico* amplicons from the complete genomes of 186 bacterial and 135 archaeal species. Each fasta file for the same primer pair was inserted as subject and query in BLASTN for obtaining the similarity percentage between amplicons belonging to different oral species. Amplicons with 100% alignment coverage of the query sequences and with a similarity value ≥97% (ASI97) were selected. For each primer, the species coverage with no ASI97 (SC-NASI97) was calculated.

Based on the SC-NASI97 parameter, the best primer pairs were OP_F053-KP_R020 for bacteria (region 1-3; primer pair position for *Escherichia coli* J01859.1: 9-356); KP_F018-KP_R002 for archaea (4; undefined-532); and OP_F114-KP_R031 for both (3-5; 340-801). Around 80% of the oral-bacteria and oral-archaea species analysed had an ASI97 with at least one other species. These very similar species play different roles in the oral microbiota and belong to bacterial genera such as *Campylobacter*, *Rothia*, *Streptococcus*, and *Tannerella*, and archaeal genera such as *Halovivax*, *Methanosalsum*, and *Methanosarcina*. Moreover, ~20% and ~30% of these two-by-two similarity relationships were established between species from different bacterial and archaeal genera, respectively. Even taxa from distinct families, orders, and classes could be grouped in the same possible operational taxonomic unit (OTU).

Regardless of the primer pair used, sequence clustering with a 97% similarity provides an inaccurate description of oral-bacterial and oral-archaeal species, which can greatly affect microbial diversity parameters. As a result, OTU clustering conditions the credibility of associations between some oral species and certain health and disease conditions. This significantly limits the comparability of the microbial diversity findings reported in oral microbiome literature.

Lastly, the fourth objective was to analyse the supragingival and subgingival plaque microbiota at amplicon sequence variant (ASV) level of different periodontal conditions (periodontal health, gingivitis, and untreated and treated periodontitis) in terms of bacterial diversity, co-occurrence networks, and predictive models.

A total of 120 patients (55 controls, 65 periodontitis) were selected for subgingival plaque collection. Sequencing of the 3-4 16S rRNA gene region was performed in Illumina MiSeq. The obtained sequences and metadata were uploaded to the sequence read archive (SRA). Searches were performed in PubMed, Scopus, Embase, and the SRA to identify previously published Illumina 3-4 sequencing studies on the supragingival and subgingival plaque microbiome in distinct periodontal conditions. Research that met the criteria for sequences and metadata were included in the meta-omics analysis, comprising a total of 2045 samples. Sequences were processed under the same bioinformatics protocol, which included the ASV-level classification and the use of an oral-specific database for taxonomic classification. The statistical analysis was conducted using the phyloseq, DESeq2, microbiome, mixOmics vegan, SpiecEasi, and igraph packages.

Bacterial richness associated with periodontitis was higher than in health in supragingival plaque and lower in subgingival, but evenness was higher in disease in both niches. The supragingival microbiota was richer and more diverse than the subgingival for the same periodontal condition. The structure of the bacterial community differed among conditions in the supra- and subgingival plaque, as well as for the same health status between the two niches. In addition, the core microbiota of dental plaque did not allow the characterisation of periodontal health and disease; and the proportion of the bacterial community organised in co-occurrence networks at the ASV level was very small. However, a small proportion of supra-

and subgingival taxa had outstanding ability to distinguish between periodontal conditions, and a relevant percentage of them were core members. Supragingival plaque was a better bacterial biomarker than subgingival for discriminating periodontal health from untreated and treated periodontitis. The main health-predictor ASVs in supragingival and subgingival plaque were: *Rothia dentocariosa* ASV2, *Haemophilus parainfluenzae* ASV3, ASV78, ASV45, and ASV46, *Kingella oralis* ASV66, *Streptococcus vestibularis* ASV27, and *Actinomyces* HMT170 ASV119. The main predictor ASVs of periodontitis in dental plaque were: *Tannerella forsythia* ASV15, *Filifactor alocis* ASV19, *Treponema denticola* ASV38 and ASV150, *Fretibacterium fastidiosum* ASV97, *Peptostreptococcaceae* [XI][G-4] HMT369 ASV124, *Streptococcus anginosus* ASV142, and *Peptostreptococcaceae* [XI][G-6] *nodatum* ASV189.

# INTRODUCTION

# Introduction

## I.1. Periodontitis: epidemiology, diagnosis and classification

The term "periodontal diseases" refers to several different chronic inflammatory conditions that affect the tissue surrounding and supporting the teeth (1). Pathologies arise when the balance between the microbial biofilm and the immune system is lost, whether due to dysbiosis (biofilm imbalance) or an overreaction of the host to the microbes present (1). This is a matter of concern since the 2016 estimates of the World Health Organisation (WHO) report that oral conditions, including periodontal diseases, are the 10$^{th}$ cause of years of healthy life lost due to disability (YLDs) globally (2).

Gingivitis is a localised inflammation of the gums that originates from the bacteria present in the dental plaque deposited on the teeth and the gingiva (1). The condition manifests clinically with swelling, redness, bleeding on probing (BOP), and discomfort during the probing process. Patients also usually have symptoms like bleeding and swollen red gums, pain, halitosis, and difficulties when eating (3). Gingivitis does not, however, extend to the periodontal attachment apparatus (cementum, periodontal ligament, and alveolar bone) and is reversible by reducing plaque levels (3).

In susceptible individuals, untreated gingivitis can progress to periodontitis (1), which is characterised by the gradual destruction of the tooth-supporting apparatus. Considerable damage to both the connective tissue fibres and the apical extension of the junctional epithelium in response to the accumulation of plaque is evident in advanced periodontal lesions. The bone destruction produced at this advanced stage creates the periodontal pockets that are the hallmark of the disease (1). In a clinical exploration, periodontitis manifests with redness, a changed texture and swelling of the margin of the gums, BOP, the increased depth of the periodontal pockets, the destruction of the ligament and alveolar bone, the recession of the marginal gingiva, increased tooth mobility, and, eventually, tooth loss (4).

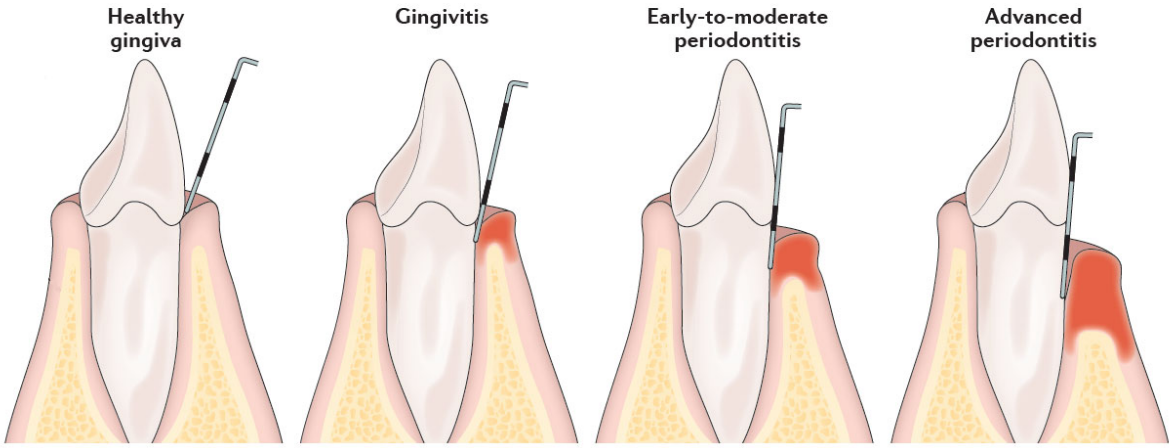



**Figure 1. Schematic representation of healthy gingiva, gingivitis, early-to-moderate periodontitis, and advanced periodontitis. The image was taken from Kinane et al. (1) with the permission of Springer Nature.**

### I.1.1. Epidemiology

Periodontal diseases are a significant public health concern, being among the most prevalent and consequential oral conditions worldwide (5). In 2015, the global prevalence of severe periodontitis was estimated to be 7.4%, with 538 million cases (6). In the US, 42.2% of dentate adults ≥30 years old had some category of periodontitis, with 7.8% affected by the severe disease (7). In those ≥65 years, these figures increased to 68% and 11%, respectively (8).

According to data from the 2015 Oral Health Survey, 5% of the adult population in Spain aged from 34 to 44 years old and 10% of those between 65 and 74 had deep periodontal pockets (≥6 mm) (9). The prevalences of moderate pockets (4-5 mm) in these groups were 18.5% and 27.0%, respectively. Although these outcomes had not changed substantially from the previous survey in 2010 (10), there was an increase in the presence of moderate pockets, which may signify a likely worsening periodontal status in these age groups in subsequent years (9).

### I.1.2. Impact on quality of life

Periodontitis has been associated with a negative effect on the quality of life, especially in patients with severe periodontitis, compromising aspects related to both function and aesthetics (11). Specifically, if no treatment is provided, outcomes can include tooth loss, impaired masticatory performance, a poorer nutritional status, lower self-esteem and quality of life, and negative effects on general health (12). The disease may also be a source of social inequality (4).

Treating periodontal diseases places an enormous economic burden on families and healthcare systems (5). Nonetheless, non-surgical periodontal therapy is worthwhile, since it quickly improves the oral health-related quality of life of adult patients, whose condition has been found to remain stable at three months after the treatment (13).

### I.1.3. Risk factors

Periodontitis is a biofilm-initiated inflammatory condition with multifactorial origins, including non-modifiable (inherent) and modifiable (acquired) risk factors. Genetics, age, and sex are among the former type of characteristics associated with the disease (14). Meanwhile, acquired exposures can include factors relating to the local environment (e.g., biofilm load or composition), as well as lifestyle, social, educational, and economic elements, and the presence of other diseases (12).

The genetic predisposition of the host is also regarded as relevant for the onset and progression of the periodontitis, with heritability estimates in family studies of 0.15% and twin research of 0.38% (15). According to a systematic review carried out by Nibali et al. (16), there is strong evidence of an association between periodontitis and the genes for the vitamin D receptor, the Fc gamma receptor IIA (Fc-cRIIA) and interleukin (IL) 10. There is also moderate evidence in distinct ethnic populations of a link between the disease and other genes and single-nucleotide polymorphisms (SNPs) (16). Although several studies have described a familial aggregation of periodontitis, consideration must also be given to the possibility that these patterns may simply reflect frequent exposure to environmental factors within such families (14).

An increase in the prevalence and severity of periodontitis occurs with age, while more significant periodontal destruction is observed in men (14). These sex-related differences may be due to the effects of the sex-specific genetic architecture and the circulating levels of sex-steroid hormones (17). The steroid reductions associated with ageing could also account for the increases in the susceptibility to periodontitis over time, especially in women (17). Other periods of hormonal changes experienced by women, including during puberty, the menstrual cycle, pregnancy, or when taking oral contraceptives, may also have an impact (14).

Gingivitis is a major risk factor and a necessary pre-requisite for periodontitis, meaning that managing the former is a primary prevention strategy for preventing the latter (12). As both diseases are plaque-induced, routinely performed oral hygiene is the most critical of the behavioural factors that affect periodontal conditions (18). Indeed, infrequent toothbrushing has been linked to the most severe types of periodontitis (19).

Nevertheless, gingivitis does not always progress to periodontitis, and various authors have demonstrated that individuals with similar plaque levels can have different inflammatory responses (20,21). Consequently, referring back to the point mentioned above, a susceptibility to periodontitis may depend on the genetic factors of the host (1).

Smoking is acknowledged to be an important modifiable risk factor (1,12,14,18,22), with both the active and passive forms of smoking being able to aggravate the development of the disease (22). A recently published systematic review and meta-analysis has confirmed the association between tobacco consumption and the incidence and progression of periodontitis, with smoking increasing the chances of suffering from the disease by 85% (23). Conversely, when these patients stopped consuming tobacco, the risk of the onset and progression of the condition was reduced and the outcomes of basic periodontal therapy improved (24).

Vasoconstriction and augmented gingival keratinisation mean that signs of gingival inflammation in smokers can be less evident than in non-smokers (1). The damage caused by tobacco is, nevertheless, clear, and includes a reduced immune response, an aggravated inflammatory reaction, and alveolar bone loss (23). Smoking is also a modulating factor for the metabolisation of bacteria and their survival, stimulating pathogenic microbial invasions (23). In this sense, the *in vitro* exposure of periodontal pathogens to nicotine and cigarette smoke results in the promotion of biofilm formation, colonisation, and infection, as well as modifications to bacterial functioning (e.g., in *Porphyromonas gingivalis*) (25). Moreover, dysbiosis in the periodontal microbiome has been observed in individuals who smoke, regardless of the state of their periodontal health (healthy, gingivitis, or periodontitis) (25).



Other unhealthy habits, including alcohol consumption, have also been associated with a greater risk of periodontitis (26). When combined, smoking and excessive alcohol consumption

can have a synergistic effect that increases the likelihood of this oral pathology (27). Furthermore, there is evidence in the literature that periodontitis is influenced by diet, with micronutrient deficiencies (e.g., vitamin C or calcium) inversely associated with periodontal health (12). At the macronutrient level, carbohydrates might similarly play a role in the development of periodontitis (12,28). Indeed, it has been demonstrated that a carbohydrate-rich diet increases the risk of inflammation and gingival bleeding (29). In contrast, gingival bleeding scores were reduced with an anti-inflammatory diet, i.e., one that is low in processed carbohydrates and animal proteins and rich in vitamin C and other micronutrients (30).

Even though it has been investigated less, stress has also been proposed as a modifiable risk factor for periodontitis (14). In fact, a systematic review identified a positive qualitative correlation between stress-related biomarkers and the severity of the disease (31). Economic status and education level are likewise thought to have an impact, with better gingival health linked to higher education and a more secure income stream (14). Finally, as explained in detail in Section I.2, there is growing scientific evidence of the negative effects of various systemic diseases and conditions on the periodontal attachment apparatus (18,32).

### I.1.4. Diagnosis and classifications of periodontal diseases

Traditional clinical measures are the best currently available for diagnosing periodontitis (1). Among the evaluable parameters in a clinical exploration are the: probing pocket depth (PPD); clinical attachment level (CAL); BOP; amount of plaque and calculus; presence/absence of oedema; grade of tooth mobility; existence of furcation defects; and radiographic evidence of bone loss. However, to achieve an accurate diagnosis, the clinician has to record multiple measurements from six sites per tooth, which is a laborious process and has outcomes that depend on the experience of the examiner (1).

Recent decades have seen the development of several classification systems for periodontal diseases and conditions. Some authors have evaluated a combination of clinical parameters, while others use a single element to define periodontal health (33). The threshold values of these classifications also vary (33). Consequently, the definitions and criteria used to diagnose periodontitis globally are inconsistent, which can affect the accuracy of any attempts to compare two investigations (33).

I.1.4.1. Classifications of periodontal diseases

Several authors have developed a new periodontitis case-definition system within the context of the 2017 World Workshop on the Classification of Periodontal and Peri-Implant Diseases and Conditions (4). As seen in table 1, this new approach differentiates between three forms of periodontitis based on their pathophysiology: necrotising periodontitis; periodontitis as a direct manifestation of systemic disease; and periodontitis. The concepts of "aggressive" and "chronic" have therefore disappeared, and are now encompassed in the general "periodontitis" category because of a lack of evidence that they are distinct pathophysiologically (34).

Differential diagnoses are based on patient history and specific signs and symptoms of necrotising periodontitis or the existence of an uncommon systemic disease that alters the host's immune response. Any remaining cases are diagnosed as "periodontitis" and then further categorised using a staging and grading system (4). The stage is based on the severity and extent of the disease and the complexity of managing the case (34). Conversely, the grade provides information about the rate of progression, the response to therapy, and individual risk factors (i.e., smoking or diabetes), which are employed as grade modifiers (34).

**Table 1. Summary of the definitions of periodontitis developed during the 2017 World Workshop on the Classification of Periodontal and Peri-Implant Diseases and Conditions (4).**

<table>
<tr><th colspan="2">New case-definitions for periodontitis (2017)</th></tr>
<tr><td colspan="2">I. Necrotising periodontitis</td></tr>
<tr><td colspan="2">Clinically characterised by a history of pain, ulceration of the gingival margin and/or fibrin deposits at sites with decapitated gingival papillae, and, sometimes, exposure of the marginal alveolar bone</td></tr>
<tr><td colspan="2">II. Periodontitis as a manifestation of systemic diseases</td></tr>
<tr><td colspan="2">Classification of these conditions should be made following the classification of the primary disease according to the International Statistical Classification of Diseases and Related Health Problems (ICD) codes</td></tr>
<tr><td colspan="2">III. Periodontitis</td></tr>
<tr><th>Stages</th><th>Grades</th></tr>
<tr><td>These are based on severity (interdental attachment loss at a site with most loss, radiographic bone loss, and tooth loss) and the complexity of any management strategies (PPD, type of bone loss, furcation lesions, tooth mobility, number of remaining teeth, bite collapse and residual ridge defect size):<br>• Stage I: initial periodontitis<br>• Stage II: moderate periodontitis<br>• Stage III: severe periodontitis with the potential for additional tooth loss<br>• Stage IV: severe periodontitis with the potential for the loss of the dentition<br>Extent and distribution are added as stage descriptors: localised (<30% of sites are involved); generalised; or molar-incisor pattern</td><td>These are based on direct and indirect evidence of disease progression and are modified by the presence of risk factors:<br>• Grade A: slow rate of progression<br>• Grade B: moderate rate of progression<br>• Grade C: rapid rate of progression</td></tr>
</table>

An agreed classification system opens the door to the future inclusion of information about the systemic impact of periodontitis, with the inflammatory burden assessed via levels of the C-reactive protein (CRP); the integration of validated biomarkers into the system can likewise be considered (34).

It should be noted that the 2017 workshop also led to new classification schemes for necrotising periodontal diseases (necrotising gingivitis, periodontitis, and stomatitis), endodontic-periodontal lesions, and periodontal abscesses (4).

### I.1.4.2. Host biomarkers for the diagnosis of periodontitis

As explained earlier, clinical measures are still essential for diagnosing periodontal diseases and, until now, have not been improved. Nonetheless, they can only assess the current extent and severity of the condition and cannot provide information about future disease activity. As a consequence, there is hope that emerging evidence on host biomarkers in oral fluids may be valuable for screening, diagnosing, and predicting disease progression (1).

In this regard, matrix metallopeptidase (MMP) 8 has been identified as clinically informative for diagnosing the presence of periodontitis in both gingival crevicular fluid (GCF) and saliva (35,36). This discovery arose from two different meta-analyses conducted by our research group, which aimed to assess the diagnostic accuracy of single molecular biomarkers in GCF and saliva for detecting periodontitis in systemically healthy subjects. Of the GCF biomarkers used for the meta-analyses, the best median sensitivity and specificity values were identified for MMP8 (76% and 93%) and elastase (78% and 76%) (35). The poorest performers were cathepsin and trypsin (35). In saliva, the highest sensitivity values for diagnosing periodontitis were obtained for IL1beta (79%) and MMP8 (73%) (36). These were followed, in order, by IL6, haemoglobin, and MMP9 (36). In terms of specificity, MMP9 produced the best results (81%), followed by IL1beta (78%) (36). Haemoglobin, IL6, and MMP8 had the lowest specificity values (36).

The sensitivity and specificity values derived from the meta-analyses were then used to calculate negative and positive predictive scores for the two most-studied biomarkers in both GCF and saliva. For a 45% prevalence of periodontitis, testing for the presence of MMP8 in GCF detected the disease in 90% of subjects (35); with elastase, this figure was 81% (35). Conversely, salivary IL1beta and MMP8 tests identified periodontitis in 67% and 63% of diseased patients, respectively (36).

## I.2. PERIODONTITIS AND ITS IMPLICATIONS FOR GENERAL HEALTH

Several systemic diseases and conditions, both inherent and acquired, can affect the periodontal attachment apparatus, causing the loss of periodontal tissue (32). This damage can either 1) influence the course of periodontitis, or 2) affect periodontal-supportive tissue, irrespective of dental plaque biofilm-induced inflammation (18). The first case includes both rare disorders, in which periodontitis is a manifestation of the systemic condition itself (e.g., genetic disorders), as well as more common diseases (e.g., diabetes mellitus). The damage in the second case arises from very rare conditions, many of which are neoplasms (32). In this section of the thesis, however, the focus is on the relationship between periodontitis and common systemic pathologies.

As long ago as 2000, Williams and Offenbacher (37) used the term "Periodontal Medicine" to define a then rapidly emerging branch of periodontology based on data that established a strong relationship between periodontal and systemic health or disease. Today, there is a convincing body of scientific evidence that is supportive of this notion of a two-way relationship between periodontitis and cardiovascular diseases (38), diabetes mellitus (39), respiratory conditions (40), rheumatoid arthritis (41), Alzheimer's disease (42) and adverse pregnancy outcomes (43). Furthermore, recent years have seen an increase in research linking periodontitis to other disorders such as metabolic syndrome (44), obesity (45), chronic kidney disease (46), and orodigestive cancers, including those of the oral cavity, gastrointestinal tract, and pancreas (47).

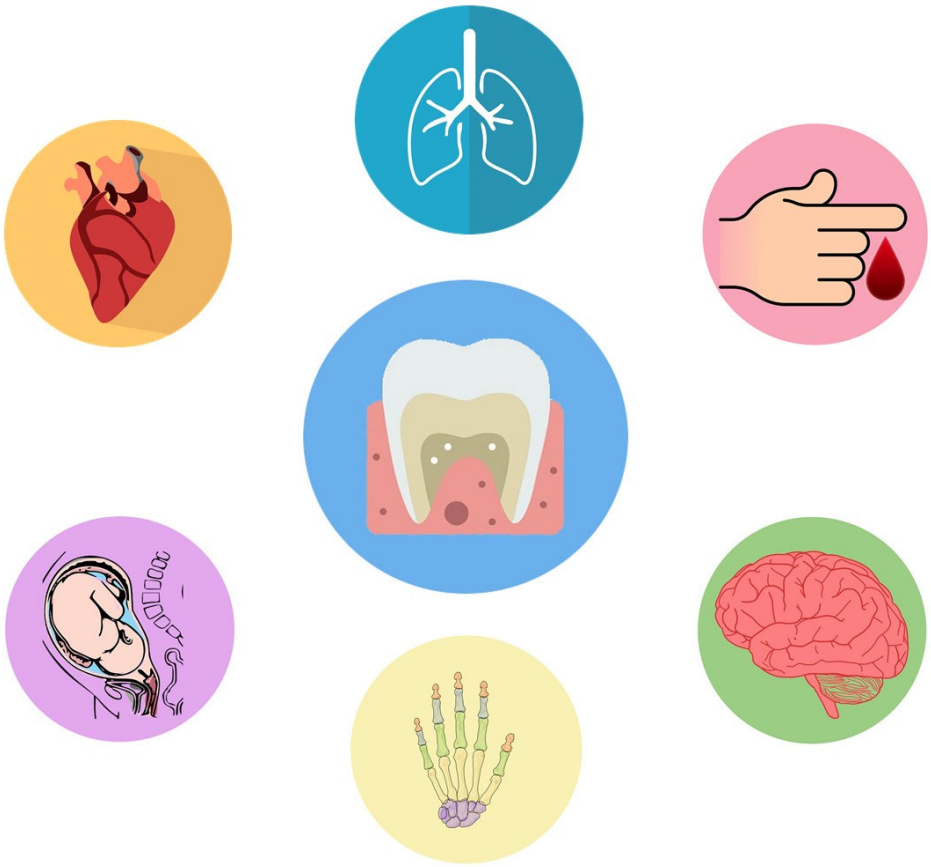


**Figure 2. Representation of the systemic conditions associated with periodontitis. From left to right: cardiovascular diseases, respiratory disorders, diabetes, Alzheimer's disease, rheumatoid arthritis, and adverse pregnancy outcomes.**

Three underlying mechanisms through which periodontitis may play a role in general health have been hypothesised (48,49):

- Metastatic infection - an infectious disease caused by microorganisms from a distant part of the body (48). Microbes can spread in three different ways: 1) direct propagation through contiguous spaces or venous or lymphatic drainage, 2) aspiration, or 3) bacteraemia or the access of oral bacteria to the bloodstream.
- Inflammation and inflammatory injury - the indirect damage caused to tissue and organs by microbes via the dissemination of bacterial exotoxins and endotoxins.
- Adaptative immunity - the immune response of the host to oral microorganisms and their virulence factors. Soluble antigens can enter the bloodstream, bond to a specific circulating antibody, and form a macromolecular immunocomplex, with the latter potentially leading to multiple acute and chronic inflammatory reactions at the deposition sites (49).

## I.3. AETIOLOGY OF PERIODONTITIS: MICROBIOTA AND HOST RESPONSE

As explained previously, both the initiation and progression of periodontitis are related to multiple aetiologies and risk factors, with the interaction between local microorganisms and the host's immune response being particularly relevant. Consequently, the development of effective therapeutic approaches requires the identification of the main periodontitis-associated microbes and host biomarkers.

### I.3.1. Microbiota

The human body is an ecosystem formed not only by human eukaryotic cells but also by a tremendous diversity of bacteria, archaea, fungi, and viruses. Such populations of microbes colonise the gastrointestinal and genitourinary tracts, the oral cavity, the nasopharynx, the respiratory tract, and the skin (50). Indeed, highlighting their importance, microorganisms are responsible for more than 200 grams of the total body weight of an average human (70 kilograms), while approximately $3.8 \times 10^3$ of the body's cells are contributed by bacteria and other microorganisms (51). The oral cavity in particular has a high abundance of microbes, exceeded only by the numbers present in the gastrointestinal tract (52).

Two imprecise terms are usually employed to designate this group of microorganisms: the "microbiota" or the "microbiome". However, there is a critical difference between them in that the former includes the population of microorganisms that colonises a body part, while the latter refers to the setting formed by microbes, their genes, and their metabolites in an ecological niche (53). As the focus of this Thesis is on the detection and identification of bacteria in different oral niches, the term microbiota is used throughout.

Our knowledge of the aetiology and pathogenesis of periodontal diseases has changed over time for four different reasons: 1) technological advances in the methods used to study the microbes present; 2) the current assumption that these conditions are caused by biofilms, and not by bacteria in a planktonic state, and the adoption of ecological concepts for studying the oral microbiota; 3) the discovery of the impact of genetic and environmental factors on the initiation and progression of these diseases; and 4) our understanding of the role played by immune mechanisms (54). As a consequence, several microbial theories have been proposed

for the aetiology of periodontitis, ranging from the "specific plaque hypothesis" to the "polymicrobial synergy and dysbiosis model" (PSD model) (54,55).

The study of the composition of the oral microbiota dates back to 1683, when Antony van Leeuwenhoek observed the microorganisms present in human dental plaque using the first prototype of a microscope. Almost two centuries later, Robert Koch, the father of modern microbiology, developed techniques for producing bacterial cultures, which allowed him to scrutinise any changes in the bacteria present over time (56). Since then, as in other microbiology disciplines, oral bacteria have been detected using culture-dependent methods (57). The first studies of the composition of dental plaque had used such techniques to identify important organisms like *Fusobacterium* spp., *Neisseria* spp., *Streptococcus* spp., and *Veillonella* spp. (58). Nonetheless, there are several issues with these traditional methods when it comes to cultivating species that require rigorous growth conditions. In fact, only 50% of oral bacteria are cultivable (59). Additionally, culture-dependent techniques are expensive and laborious, requiring experienced staff and sufficient time for their execution (57).

The goal of overcoming the latter issue led to the development of several molecular deoxyribonucleic acid (DNA)-based technologies, including DNA microarrays and the polymerase chain reaction (PCR) test. These enabled researchers to analyse oral-microbiota communities more comprehensively and perform large-scale studies (57,58). In 1998, Socransky et al. (60) used checker-board DNA-DNA-hybridisation techniques to identify five different bacterial complexes that had distinct levels of association with health and the severity of periodontitis. This discovery was revolutionary because, until then, periodontitis, like other infectious diseases, was thought to be caused by a single pathogen rather than a series of organisms working with each other (58). Three species in particular - *P. gingivalis*, *Tannerella forsythia,* and *Treponema denticola* - were closely associated with the clinical parameters for periodontitis and together constitute the so-called "Red Complex". This complex has also been related to other bacteria, including *Campylobacter gracilis*, *Campylobacter rectus*, *Campylobacter showae*, *Fusobacterium nucleatum*, *Fusobacterium periodonticum*, *Peptostreptococcus micros*, *Prevotella intermedia*, *Prevotella nigrescens,* and *Streptococcus constellatus*, which collectively form the "Orange Complex". In contrast, members of the "Yellow Complex" (*Streptococcus gordonii*, *Streptococcus intermedius*, *Streptococcus mitis*,

*Streptococcus oralis,* and *Streptococcus sanguis*) and the "Purple Complex" (*Actinomyces odontolyticus* and *Veillonella parvula*) are associated with healthy states.

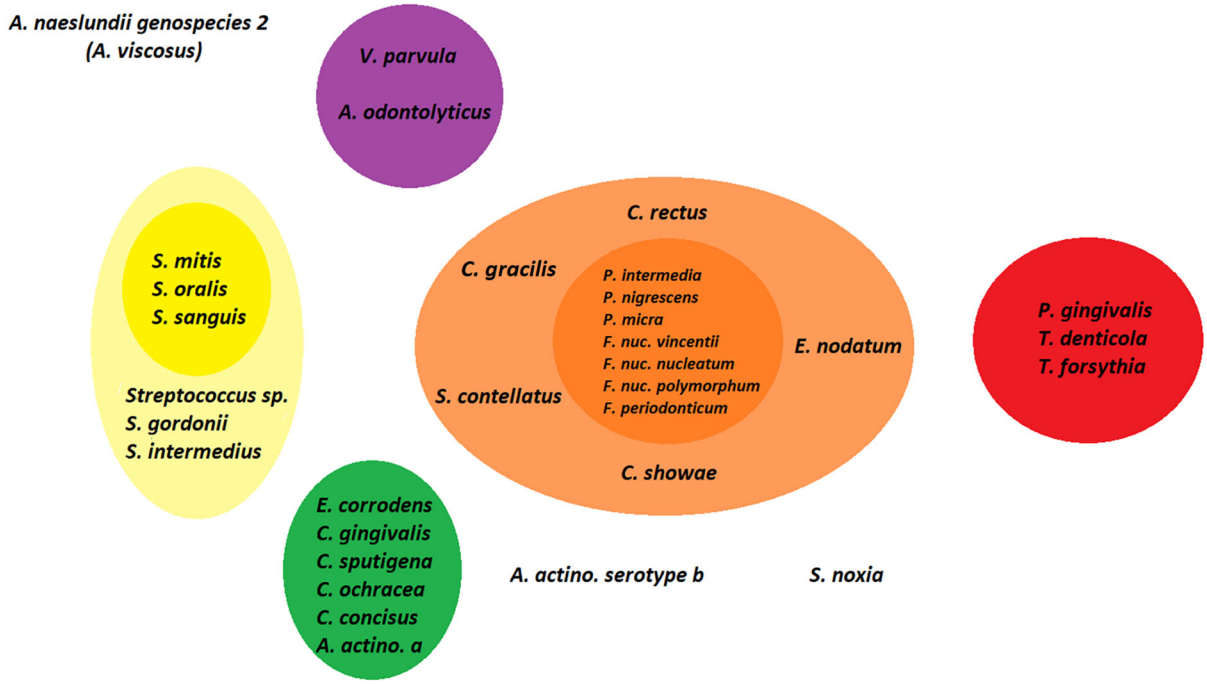



**Figure 3. Socransky's microbial complexes in subgingival plaque.**

A PCR test is an *in vitro* molecular technique that amplifies a gene or DNA fragment directly or a ribonucleic acid (RNA) indirectly. A variant of the PCR, known as a quantitative PCR (qPCR) or real-time PCR (RT-PCR), makes it possible to amplify, and simultaneously quantify, the amplification product obtained from a sample. In comparison to traditional cultures, this technology has been found to have high diagnostic accuracy when it comes to detecting *A. actinomycetemcomitans* and *P. gingivalis* (61). Furthermore, a qPCR enabled our research group to obtain eight bacterial cluster-based models with good predictive accuracy at identifying a site with periodontal destruction in a periodontitis patient (62). All of the models used by our team had an area under the curve (AUC) of ≥0.760 and sensitivity and specificity scores of ≥75.0%, with the best values for the cluster formed by *A. actinomycetemcomitans*, *F. nucleatum*, *Parvimonas micra*, *P. intermedia*, *T. forsythia* and *T. denticola* (AUC= 0.789; sensitivity and specificity= 77.5%). Overall, we concluded that clusters formed by species that had different etiopathogenic roles, i.e., those belonging to distinct Socransky complexes, had good predictive accuracy for diagnosing periodontitis (62).

The human oral microbe identification microarray (HOMIM) is another microarray-based platform that has been utilised to detect and identify both cultivated and not-yet-cultivated oral bacteria (58). Employment of this tool has enabled species like *Filifactor alocis* and *P. micra* to be observed more frequently and in higher numbers in periodontitis samples than in healthy specimens (63). However, this technique is currently no longer available (http://homings.forsyth.org/index2.html).

These DNA microarray methods (checkerboard-DNA, DNA-hybridisation, and HOMIM) are not without their limitations, since only a fixed number of species can be detected in a panel and a specific quantity of DNA is required to identify a microorganism (58). The sequence analysis of the 16S ribosomal RNA (rRNA) bacterial gene became the method of choice to overcome these shortcomings. The amplification of this gene, which is present in all prokaryotic organisms, was achieved through the use of universal primers followed by a subsequent sequencing step. This enabled the species present in a sample to be distinguished, even if they had not been identified previously (58). The characteristics of this gene will be explained in detail in Section I.4.

Initial research on sequence analyses of the 16S rRNA gene was based on the Sanger method, which is one of the so-called "first-generation sequencing" technologies (64). This technique used universal primers for the 16S rRNA gene to amplify the DNA isolated from a specimen. The resulting amplicons were cloned into *Escherichia coli*, and the inserts obtained were subsequently sequenced to determine the identities of the species present (56). New periodontitis-associated genera were discovered with this technology, including *Desulfobulbus* spp., *Eubacterium saphenum, F. alocis*, *Megasphaera* spp., and *Peptostreptococcus* spp. (65).

The years that followed saw the development of "second-generation sequencing" techniques, commonly known as "next-generation sequencing" (NGS). These enabled massive parallelisation and improved automation and speed, and were also less expensive (64). It thus became possible to complete large-scale sequencing projects in just a few days or sometimes even hours (57). The two NGS techniques employed the most - 454 pyrosequencing and Illumina - are described in depth in Section I.4.

Advances in molecular techniques and the subsequent development of NGS tools led to an entirely new field of research: omics. Whole-genome shotgun metagenomics, commonly known as just metagenomics, allowed researchers to sequence the complete DNA (genome) of a single microbial culture or a complex microbial population, enabling the generation of reference genomes (57). This provided information not only on phylogenetic compositions but also on the genetic potential of a community to carry out distinct functional activities (58). Nevertheless, this technique does not produce data on the fraction of the metagenome being expressed. which can instead be obtained using techniques like metatranscriptomics, metaproteomics, and metabolomics that, respectively, assess the synthesis of transcripts, proteins, or the metabolic products of a specific set of activities (58). The human microbiome project (HMP) (2008-2012) later emerged as an initiative of the United States (US) National Institutes of Health (NIH) and used 16S rRNA gene sequencing and metagenomics' techniques to initially identify and characterise the microorganisms associated with human health in different body parts, including the oral cavity (66). The research found that microbial profiles varied significantly, even between healthy subjects, meaning that taxonomic characterisation alone is not enough to uncover the relationship between the microbiota and the healthy state (67). In addition, metabolic pathway reconstructions of metagenomic data found that several pathways were ubiquitous in subjects and body habitats (67). Consequently, the second phase of the HMP (2013-2016), known as the Integrative HMP, comprised studies on dynamic changes in the microbiota and host as a result of physiological (pregnancy and preterm birth) or pathological (inflammatory bowel disease and pre-diabetes) conditions (68). This research has recently completed its first phase (69-71).

Although each of these integrative investigations revealed new biology within their respective areas of health and disease, a surprising range of immune and ecological features of the host microbiota were commonplace (72). The combination of shotgun metagenomics, untargeted metabolomics, and immuno-profiling measurements have efficiently captured the host and microbial properties linked to disease. As in most studies of the microbiota, changes that occurred within individuals, populations, or phenotypes were often much smaller than the baseline variations between them. Accordingly, it is clear from this research that health-associated microbiota interactions in individuals can manifest in extremely diverse ways (72).

Each of the integrative HMP studies found that other aspects of these interactions are highly localised and subject-specific. Microbial changes and associated host responses in the three conditions (individuals, populations, or phenotypes) were strongest when captured at the time the changes occurred and, often, within the tissue of origin. It is evident from these and other investigations that host-microbiota interactions have both localised and systemic effects. The NIH's HMP has now come to an end but has revealed multiple new avenues of research and technologies for studies in the future (72).

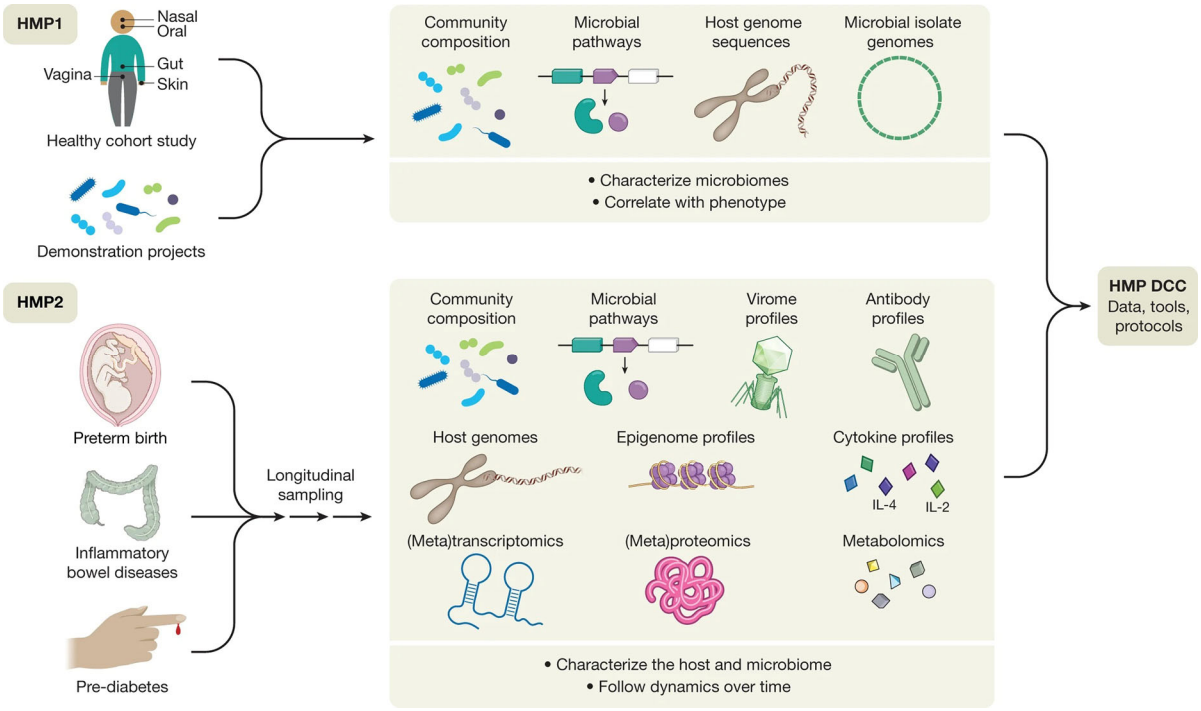



**Figure 4. The first and second phases of the Human Microbiome Project. The image was taken from The Integrative HMP Research Network Consortium (72), an open-access article distributed under a Creative Commons Attribution 4.0 International (CC BY 4.0) license (https://creativecommons.org/licenses/by/4.0/).**

According to the traditional Socransky viewpoint, a set of Red Complex bacteria is thought to be the causative agent behind periodontitis (60). Now, however, NGS and omics' technologies and techniques have detected the presence of bacteria like *P. gingivalis* in the absence of disease (73,74). It has also been found that the periodontal microbiota is more heterogeneous and diverse than previously thought, with new component organisms identified (75,76). These results have confirmed the hypothesis that periodontitis is initiated by the PSD of the entire microbial community (55). The PSD model states that different members or specific gene combinations perform distinct roles to shape and stabilise disease-provoking microbiota. Of particular relevance are the so-called "keystone pathogens", which impair the

host's immune response and elevate the virulence of the entire community through interactive communication with accessory pathogens. This dysbiotic microbial community is mainly composed of anaerobic genera from the phyla *Bacteroidetes*, *Firmicutes*, *Proteobacteria*, *Spirochaetes*, and *Synergistetes* (77). Consequently, while *P. gingivalis* is regarded as a keystone pathogen, *S. gordonii* and *T. forsythia* are viewed as accessories (77).

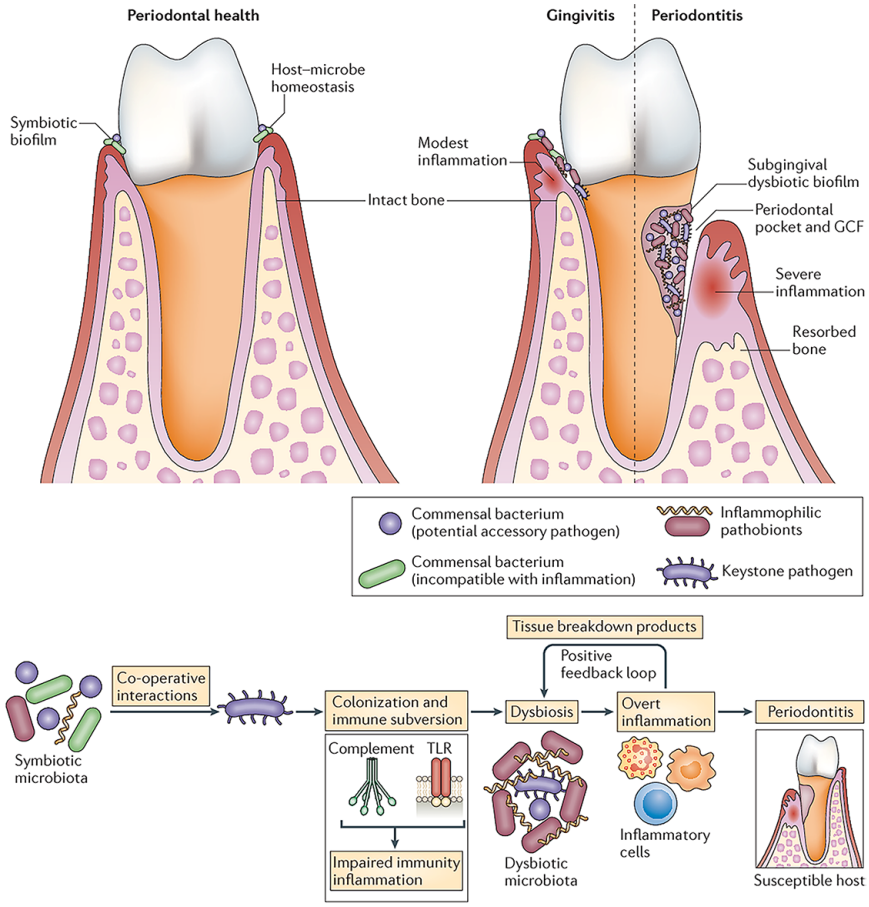



**Figure 5. Polymicrobial synergy and dysbiosis in periodontitis. The image was taken from Hajishengallis (77) with the permission of Springer Nature.**

### I.3.2. Host response

As noted previously, periodontitis is considered to be an inflammatory disease initiated by bacteria. However, despite the advances made in both periodontal microbiology and pathobiology, the issue of which comes first - the inflammatory response or the change to a dysbiotic subgingival microbiota - is still a matter of debate (78).

In physiological conditions, there is a balance between the local immune response and the microbiota. Bacteria are undoubtedly the principal cause of gingivitis, but it is the uncontrolled host-inflammatory and host-immune responses that largely drive tissue destruction, i.e., the progression of the disease (79).

It is important in studies of the pathogenesis of periodontitis to consider the temporal sequence of microbiota changes on the way to periodontal inflammation (78). The microbial shifts induced by inflammation go beyond the overgrowth of certain species. Indeed, growth conditions also provide an environment that changes the physiology, pathogenicity, and expression of the virulence factors of the polymicrobial biofilm community (79). Consequently, the inflammatory response and the resident microbiota are linked in a bi-directional balance state in health and an imbalance state in disease (78). If the constant interplay between microbes and the host inflammatory response is viewed as a continuum, it is thus evident that specific bacteria cannot be regarded as initial causal agents in the pathogenesis of periodontitis (78).

In a recently published review, Van Dyke et al. (78) presented a new model for the pathogenesis of periodontitis. Scientific evidence led them to conclude that chronic inflammation enables the development of a periodontal pocket that changes the redox and nutrient environment, thereby increasing the diversity and species richness of the biofilm. This results in dysbiosis, which reinforces and exacerbates inflammation as a way to initiate bone resorption. In light of both how inflammation mediates dysbiosis and the associated exacerbation of periodontal damage, the "inflammation-mediated polymicrobial-emergence and dysbiotic-exacerbation" (IMPEDE) model was, therefore, proposed by Van Dyke's team (78). This is designed to complement the current classification of periodontal diseases (CPD) approach (4) and suggests that inflammation can be present for each classification stage as a principal driver of the clinical condition.

The IMPEDE model recognises five stages (0–4) through which health, gingivitis, and periodontitis may develop, be contained, or progress. These stages are as follows: 0) periodontal health; I) gingivitis; II) initiation of or early periodontitis; III) inflammation-mediated dysbiosis and opportunistic infection; and IV) late-stage periodontitis. As set out in figure 6.B, inflammation-mediated polymicrobial dysbiosis and tissue damage can be exacerbated if no

treatment is provided. Conversely, the resolution of inflammation and tissue repair/regeneration can occur if treatment is initiated.

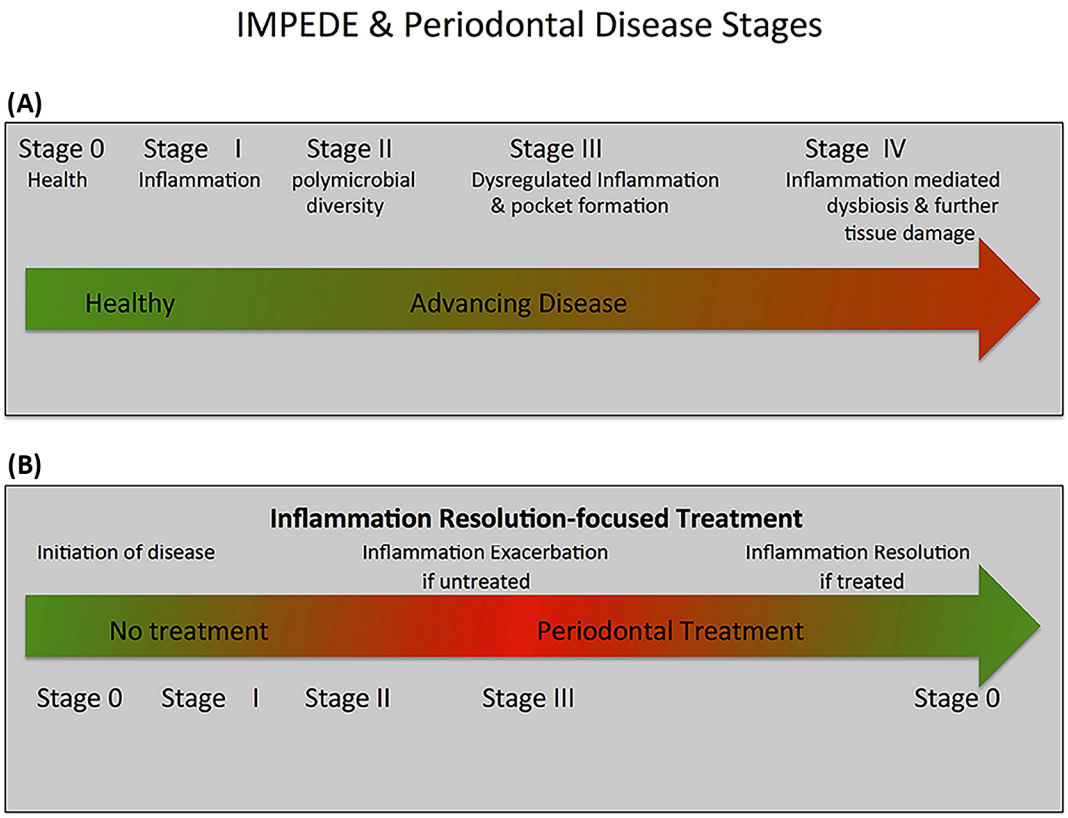



**Figure 6. The inflammation-mediated polymicrobial-emergence and dysbiotic-exacerbation (IMPEDE) model: classification of periodontal disease (CPD) stages and treatment. The image was taken from Van Dyke et al. (78), an open-access article distributed under a Creative Commons Attribution 4.0 International (CC BY 4.0) license (https://creativecommons.org/licenses/by/4.0/).**

The immune-inflammatory response in periodontitis is complex, involving both innate and adaptive immunity, which must be coordinated to return the injured tissue to homeostasis (80,81). The triggering of appropriate inherent mechanisms guarantees an effective acquired-immune response which, in turn, potentiates the innate functions against invading pathogens (80). This innate immunity is considered to be the first line of defence, with microbes identified as "strangers" and responses triggered to eliminate them (80). This mainly involves epithelial and myeloid cells (phagocytes), the complement cascade, and neuropeptides. The physiological response that occurs in this acute phase of inflammation is rapid and brief, pursuing the recruitment of enough cells at the injured site via the production of cytokines. Macrophages and neutrophils recognise and bind to the bacteria and chemokines are then secreted to attract more phagocytes. The complementary system also generates proteins that attract distinct

immune cells like monocytes, lymphocytes, and neutrophils, and is capable of killing bacteria directly. Moreover, mast cells induce vasodilatation that increases blood flow and phagocyte recruitment. At this point, clinical signs of inflammation, such as bleeding and an increase in GCF levels, can be observed (81).

Although inflammation begins as a protective response, it fails to be resolved in susceptible individuals because the elimination of inflammatory cells, leukocytes, and neutrophils does not occur. The chronic nature of the inflammatory state activates the specific adaptive or acquired response (81). An established periodontal lesion can be observed during the period of transition from innate to acquired immunity. Here, blood flow is affected, collagenolytic activity increases, and the collagen produced by fibroblasts is augmented. Clinically, this manifests as moderate to severe gingivitis, with gingival bleeding and colour and contour changes (81). First, bacterial antigens are presented and processed by lymphocytes, neutrophils, and dendritic cells (81). Of the two lymphocyte types recognised, i.e., B-cells and T-cells, the latter are particularly relevant in periodontitis. Typically, these T-cells are further classified into three subsets depending on the cytokines they produce: T-helper (Th) 1; Th 2; and Th 17 (80). Th1 cytokines have been linked to infectious inflammatory bone destruction, while the Th2 versions have been shown to minimise bone loss (80). Meanwhile, Th17 cells and their related cytokines play an essential role in periodontal tissue-specific immunity, with inflammatory properties involved in infectious, autoimmune, and osteolytic processes (80,82). Additionally, a further Th subpopulation, known as T-regulatory cells, controls the activation, proliferation, and effector functions of the conventional activated T-cells (80). Two classes of major histocompatibility complex (MHC) molecules are required to trigger the distinct T-cell subsets (81).

The transition from an established lesion to the advanced injury of periodontitis, with an irreversible loss of attachment and bone, is not well understood (81). What is clear is that T-lymphocytes and some innate immune cells communicate perfectly through cytokine networks (81). As noted previously, cytokines are produced by epithelial cells, fibroblasts, and phagocytes in the acute inflammation phase, and by lymphocytes in the established and advanced stages (81). On the one hand, the tumour necrosis factor alpha (TNFalpha), IL1beta, and IL6 are three well-established pro-inflammatory cytokines (82), with both of these interleukins associated with inflammatory cell migration and osteoclastogenesis (81).

TNFalpha, meanwhile, has many functions, including cell migration, the stimulation of chemokine production, the upregulation of IL1beta and IL6, and extracellular matrix degradation and bone resorption (81). In contrast, other cytokines, including interferon gamma (IFNgamma), IL11, IL12, IL35, IL37, and probably, IL27, have been described as having protective effects in the immune response to periodontitis; it is for this reason that they are regarded as anti-inflammatory cytokines (82).

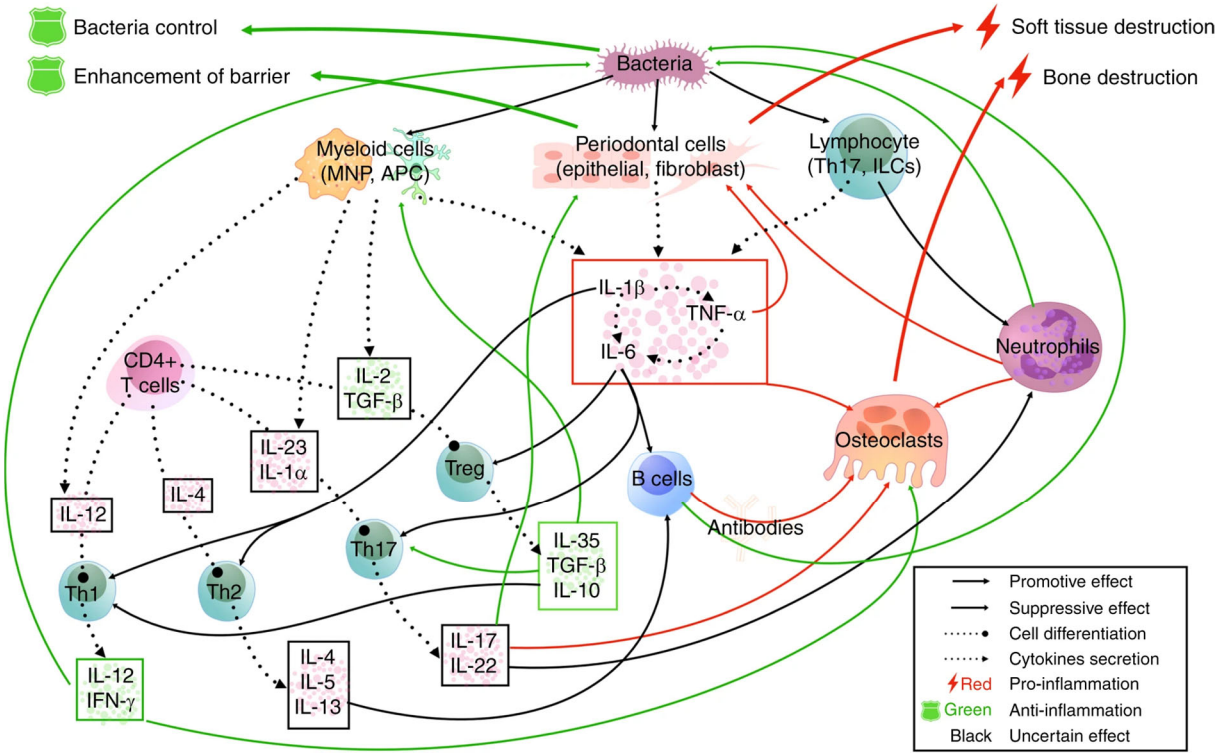



**Figure 7. The cytokine network involved in the pathogenesis of periodontitis. The image was taken from Pan et al. (82), an open-access article distributed under a Creative Commons Attribution 4.0 International (CC BY 4.0) license (https://creativecommons.org/licenses/by/4.0/).**

Skeletal homeostasis depends on the dynamic balance between the activities of bone-forming osteoblasts (OBLs) and bone-resorbing osteoclasts (OCLs) (80). Although several regulatory schemes, including the endocrine system, control this balance, the immune system also influences its regulation. In this sense, it is widely accepted that the perturbation of the balance by bacterial products and inflammatory cytokines in favour of OCLs is the main underlying cause of inflammation-induced bone loss (81). In fact, research characterising the receptor activator of the nuclear factor-k B (RANK), its ligand (RANKL), and the soluble decoy receptor of RANKL, known as osteoprotegerin, has contributed to the development of osteoimmunology. Specifically, RANKL is a protein that participates as a key modulator of

both physiological and pathological bone resorption, while osteoprotegerin inhibits osteoclastogenesis and bone resorption (80).

During the inflammatory process that occurs in periodontitis, pro-inflammatory cytokines like IL1beta, IL6, IL17, INFgamma, and TNFalpha can stimulate the periodontal OBLs to express the membrane-bound RANKL (80). Other cell types, mostly Th17 lymphocytes, also express this molecule (80). As evidenced by several studies, the ratio of RANKL/osteoprotegerin tends to increase when a patient moves from periodontal health to periodontitis and to decrease after periodontal treatment (80).

Finally, cytokines and other inflammatory mediators like reactive oxygen species and proteases can stimulate periodontal breakdown and collagen destruction via tissue-derived MMPs (80). These are a family of enzymes that degrade the extracellular matrix and basement membrane components; they are also involved in physiological processes such as tissue development, remodelling, and wound healing (81). The activity of MMPs is regulated by gene expression, proenzyme activation, and inhibition by endogenous inhibitors like the tissue inhibitors of MMPs (TIMPs) (80). Today, it is clear that MMPs are up-regulated in periodontal inflammation (81). In this context, higher messenger RNA (mRNA) expression levels of the MMP1, MMP2, and MMP9/TIMP ratios, as well as that of RANKL/osteoprotegerin, have been detected in gingival tissue taken from periodontitis patients. This has not been observed in healthy controls, whose ratios are lower (83). Furthermore, some periodontal pathogens, including *F. nucleatum*, *Fusobacterium necrophorum*, *Porphyromonas endodontalis*, and *Prevotella denticola*, have been found to induce the expression of MMPs (81).

## I.4. ORAL MICROBIOTA: BACTERIAL DIVERSITY

Human beings are ecosystems that not only carry their genes inside each cell but also play host to millions of microorganisms. The human body harbours approximately the same number of human and bacterial cells (51), while the mouth of an average adult contains about 50-100 billion bacteria, 687 of which, collectively, correspond to predominant species (57). All of these microbes have their own genetic content. The total genomic DNA of the organisms within a community is known as the "metagenome" (84). The human metagenome is thought to have at least 100 times more genes than the body's cells, which is why it is known as the "second genome" (85).

Moreover, the host immune system is involved in determining the microbial metagenome and the expression of its genes is regulated by the microbiota (85). Consequently, the importance of the role that the microbiota plays in the well-being and health of humans should not be underestimated. The amplification and subsequent sequencing of the 16S rRNA bacterial gene have enabled the analysis of the diversity, structure, and composition of the bacterial communities living in different body parts, as well as the description of the distinct bacterial profiles associated with states of health or disease.

### I.4.1. 16S rRNA gene: phylogenetic marker

Proteins and nucleic acids are macromolecules present in all living organisms. Over time, these undergo changes that occur randomly and increase linearly. The differences in the sequence of the monomers that comprise the homologous macromolecules, which are present in two distinct forms, reflect the evolutionary distance between them, enabling us to establish their phylogenetic relationships (86).

The 16S rRNA gene is a component of the minor subunit (30S) of the prokaryotic ribosomes and is the most widely used macromolecule in bacterial phylogeny and taxonomy investigations (86). At present, the identification of this molecule is carried out by sequencing the gene that encodes it: the aforementioned 16S rRNA gene. This gene alternates areas common to all organisms about which the sequence is known (named: conserved) with regions that undergo variations or changes over time (named: variable). The conserved zones are useful for designing universal primers that permit the amplification of the hypervariable zones.

Conversely, the nine variable regions (1-9) provide the most useful information for phylogeny and taxonomy studies. The following figure is a representation of the secondary structure of the 16S rRNA gene.

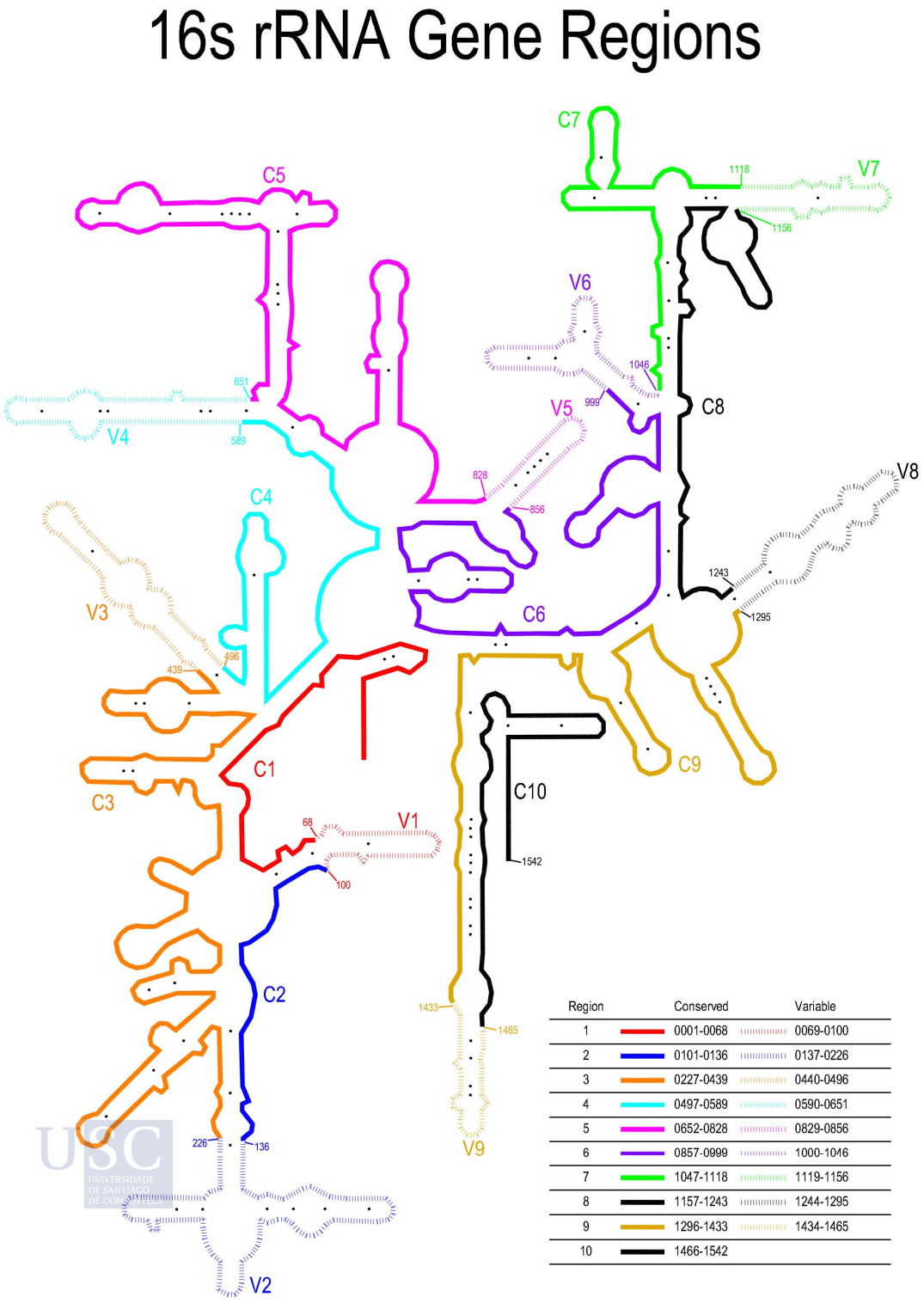



**Figure 8. Secondary structure of the 16S rRNA gene.**

Although other molecular markers are available, there are several reasons why the 16S rRNA gene has been regarded as definitive (86). First, it is present in all bacteria. Moreover, its structure and function have remained constant over time, suggesting that sequence alterations reflect random changes. These changes occur slowly enough that researchers can obtain data on all the prokaryotes. Moreover, the variability is such that both distant and close organisms can be distinguished. In addition, the relatively large size (1500 base pairs -bps-) of the gene makes it suitable for informatic purposes, and the conservation in its secondary structures favours accurate alignment. Finally, the ease with which the gene can be sequenced means that extensive, and constantly expanding, databases are available.

Nonetheless, the employment of this gene as a phylogenetic marker has its limitations. One of the most important is the presence of variations in the 16S rRNA operon copy numbers per bacterial genome, with values ranging from 1 to 15 (87,88) or 1 to 17 (89). A higher 16S rRNA operon copy number per genome for a specific taxon will overestimate its relative and absolute abundance values (89). Although the number of copies appears to be taxon-specific, there are also variations among strains of the same species (87). Furthermore, the 16S sequences obtained from the same species or within the same genome are often different. It was generally accepted that two gene sequences differing by 1 to 1.3% or more represented two distinct species (90). Nevertheless, Pei et al. (91) observed that 24 of the 586 species they analysed had an intragenomic diversity higher than 1-1.3%. Another investigation evaluated the complete genomes of 2013 bacteria and archaea and revealed that there was intragenomic heterogeneity in 952 of them (88). Even though the majority of the divergence was below 1%, 119 genomes presented with higher values (88). As intragenomic heterogeneity is believed to overestimate microbial diversity, these authors recommended using primers targeting regions 4 and 5, which had the fewest variations (88).

### I.4.2. DNA Sequencing techniques

DNA sequencing is the process that allows the nucleotide sequence of a DNA sample to be determined. Although the double helix structure of DNA was discovered in 1953, it was not until the 1970s that the first sequence of the human genome was obtained using first-generation sequencing techniques (64). The first two methods for DNA sequencing were reported in 1997: “chemical cleavage sequencing”, developed by Maxam and Gilbert (92), and “chain terminator

or Sanger sequencing", which was discovered by Sanger and colleagues (93). Due to its simplicity and reliability, the latter approach has been the gold standard over the last three decades (94,95).

The Sanger technique belongs to the sequencing-by-synthesis methods, which means that it uses the DNA synthesised by the DNA polymerase to identify the nitrogenous bases present in a DNA sequence. In addition to the four deoxyribonucleotide-triphosphates (dNTPs: dATP, dCTP, dGTP and dTTP), the sequencing reaction employs specific chain-terminator dideoxynucleotides (ddNTPs), i.e., nucleotides that lack a 3'-OH group (64). The incorporation of a ddNTP into a growing DNA molecule hampers the integration of a new nucleotide, as no phosphodiester bond can be formed because of the absence of the 3'-OH group. This means that the DNA synthesis is interrupted in that position (64,95). After several repetitions, this technique involves the products of the reactions being loaded in an agarose gel and subjected to electrophoresis. The ordered banding pattern thus obtained enables the sequence of the DNA template to be determined.

Over time, the Sanger method has incorporated a series of innovations involving the automation of the process using fluorescent terminator dyes linked to the ddNTPs and the development of software to interpret and analyse the sequences (95). Consequently, the method is still valuable when high-throughput is not required. The leader in the field of automated Sanger-sequencing is Applied Biosystems, whose current commercial sequencers can generate 600-1000 bps of a proper sequence (95).

The contribution of the Sanger approach to scientific advances in diverse areas has been invaluable. In 2001, the field of periodontal microbiology saw Paster et al. (96) use the method in the first comprehensive characterisation of the subgingival microbiota. The authors found that the subgingival niche harboured 347 species, 215 of which were novel phylotypes. They also estimated the number of unseen species in the population and determined that there were 68 additional taxa, accounting for a total number of 415 subgingival species.

As is already known, the years that followed saw the development of second-generation technologies, revolutionising the study of microbial diversity. Described below are the two NGS techniques used the most to examine the human microbiota.

#### I.4.2.1. 454 pyrosequencing

The first NGS application, known as 454, was introduced in 2005 by the biotechnology firm 454 Life Sciences, which was bought by Roche two years later (58,64). The technology comprises an initial emulsion PCR step, followed by subsequent pyrosequencing, which is a method of DNA-sequencing-by-synthesis based on the generation of light after nucleotides are incorporated in a growing DNA chain (64,94). The process of 454 pyrosequencing is as follows (94):

- The DNA is isolated, fragmented, ligated to special adapters, and separated into single strands.
- The DNA is amplified using emulsion PCR. The emulsion contains: the PCR reagents, the DNA template to be sequenced, the capture beads with the primers attached to them (complementing one of the adaptors), and another primer for the PCR. Emulsification takes place after controlled and vigorous agitation of the oil-water system. Millions of aqueous droplets are formed, with the amplification occurring inside them. Optimisation of the process guarantees that there is only one template and one bead in each droplet, meaning that millions of copies of the template are generated on each bead.
- The DNA is denatured and the beads containing single strands are transferred to the wells of a picotiter plate. Only one bead is deposited in each of the several hundred thousand wells.
- The DNA is sequenced through synthesis. This step requires: a single-stranded DNA sample, the sequencing primer, and the enzymes DNA-polymerase, adenosine triphosphate (ATP) sulfurylase, luciferase, and apyrase. Two substrates are also included in the reaction: adenosine 5’ phosphosulfate (APS) and luciferin. Cycles are then performed where each well receives, sequentially, one dNTP at a time. If there is complementarity, the DNA polymerase catalyses the incorporation into the DNA strand. The polymerase introduces a dNTP into a DNA nascent molecule, releasing pyrophosphate (PPi) in an amount equivalent to the quantity of the incorporated

nucleotide. The ATP sulfurylase converts the PPi into ATP in the presence of APS. Such an ATP is utilised by the luciferase to turn luciferin into oxyluciferin. This step produces light at an intensity proportional to the amount of ATP used. The light is detected by a camera and registered as a peak in a pyrogram, with the height of the peak being proportional to the number of incorporated nucleotides.

- The system is regenerated by the enzyme apyrase, which degrades the ATP and the unincorporated dNTPs. The next nucleotide is then added. As the process advances, the complementary DNA strand grows and the nucleotide sequence is determined according to the pyrogram values.

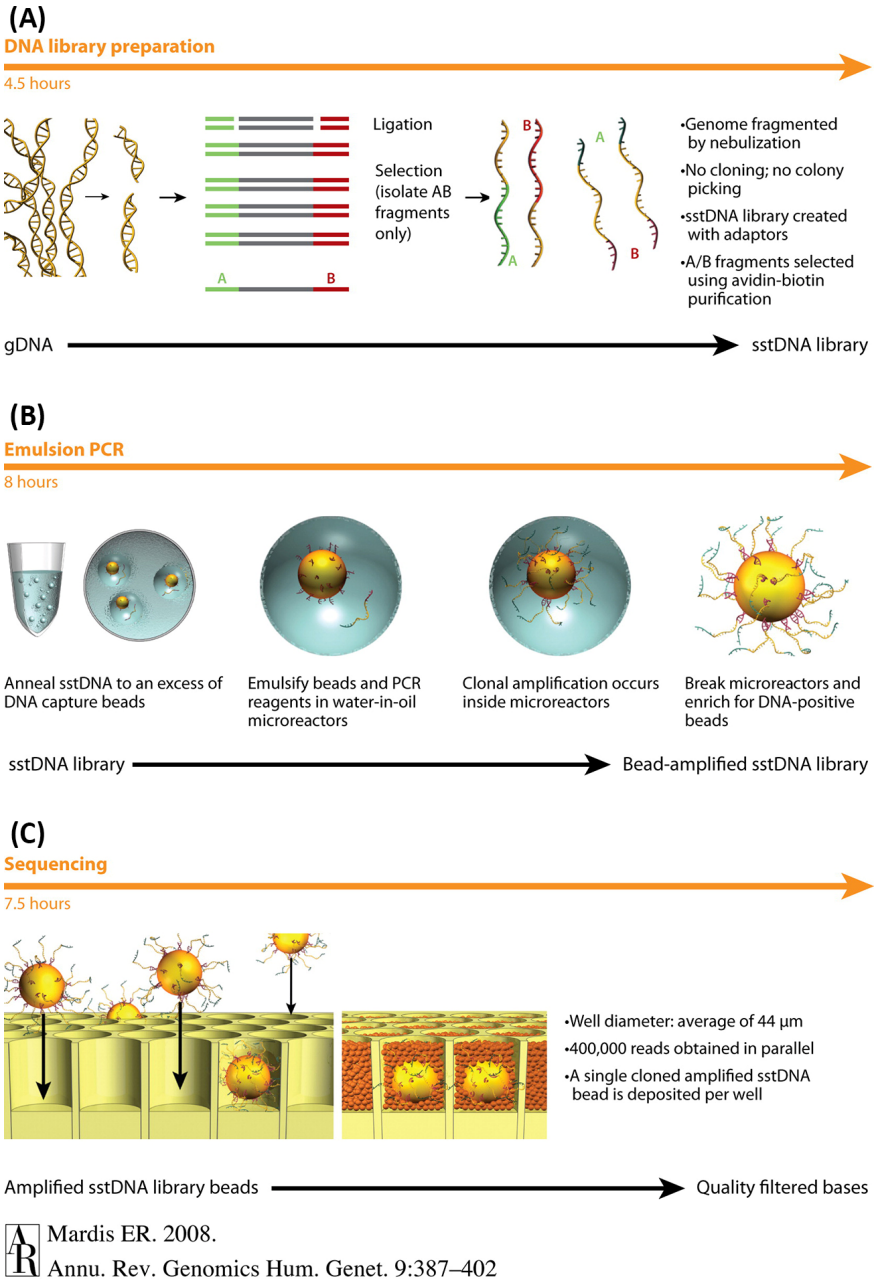



**Figure 9. The 454-pyrosequencing approach to amplifying single-stranded DNA copies from a fragment library on agarose beads. The image was taken from Mardis et al. (97) with the permission of Annual Reviews, Inc.**

Over the years, 454 Life Sciences has increased the length of its sequence reads and the number of bases per run. Its initial tools yielded sequence reads of 100 bps and up to 60 million bases per run. Later on, these sequence-read and bases-per-run figures reached 400 bps and approximately 500 million, respectively, on the company's well-known Genome Sequencer (GS) FLX Titanium platform (94); subsequently, read lengths up to 700 bps and an output of 700 mega bps (Mbps) were available with the GS FLX + series (98).

The capacity to obtain a high number of reads in a single run is one of the main advantages of the 454 technology over the Sanger technique, as it enables the acquisition of much more sequence data (94). As more reads are generated in a single run, the cost per base is much lower than with the Sanger method. Furthermore, as the cloning step is not required with the 454 technology, the bias inherent in that procedure are avoided (94). Pyrosequencing's use of "barcodes" (sequences introduced into the PCR primers) that work as unique sample identifiers enables sequences from different samples to be mixed in the same run. This increases efficiency and the number of outputs and also reduces costs (94).

The high-throughput of the 454 technology affects both the depth (number of sequences per sample) and breadth (number of samples evaluated) of the sampling (94). A greater sampling depth increases the opportunities available for detecting low-abundance or rare community members, while more breadth allows additional samples to be analysed. This means that the results are more robust for comparison purposes (94). If the goal of a study is to determine the composition of a community at distinct sites, such as the skin *vs.* the oral-cheek mucosa, more samples should be studied than more sequences per sample (99). In contrast, if specimens from the same site or one close-by (i.e., the tooth surface *vs.* the gingival crevice) are being compared, deeper sequencing is required to identify minor differences in the community's composition (100).

Pyrosequencing also has limitations. One of the most important is related to the detection of long homopolymers, which can lead to sequencing bias and an artificial increase in the richness estimators (94). The reagent cost is another notable drawback (64).

In 2013, Roche made the decision to close 454 Life Sciences because its technology was no longer competitive. Reagents are, however, still available from several suppliers (95).

#### I.4.2.2. Illumina

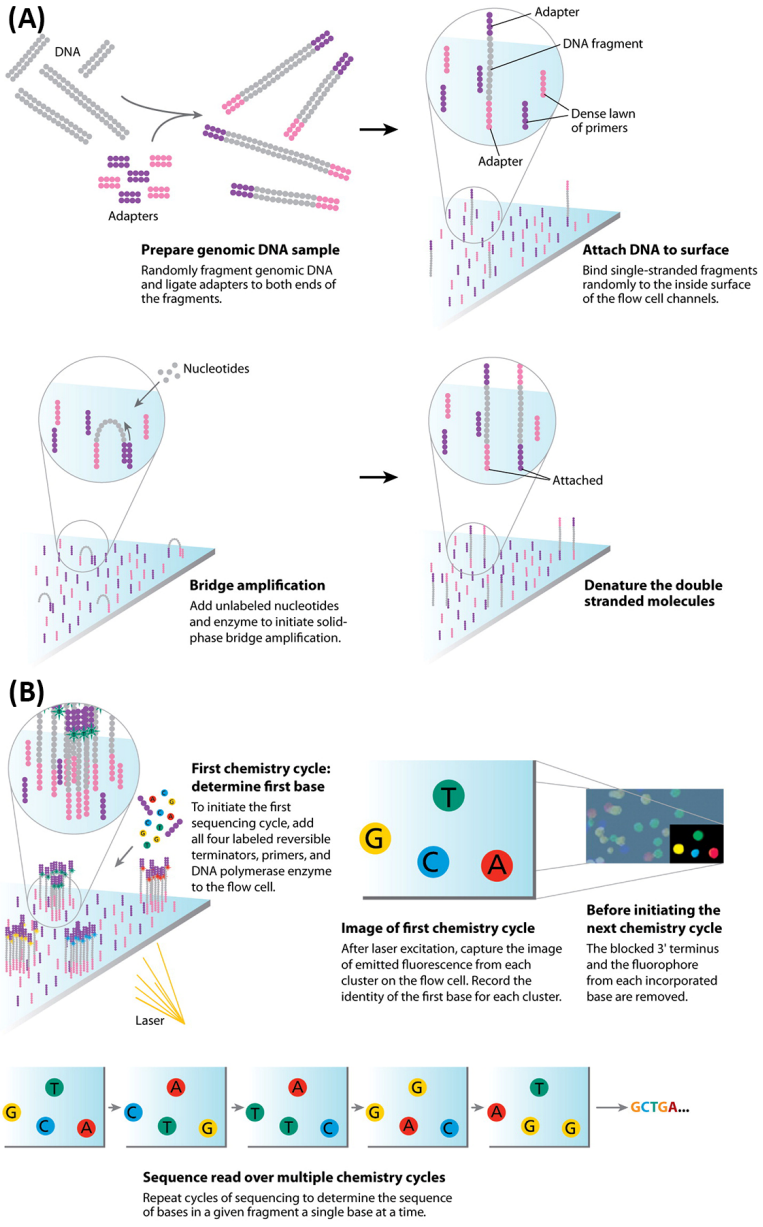
The first Solexa sequencer, named Genome Analyzer, was launched in 2006. Illumina acquired the company a year later and, since then, has brought a number of different sequencers to the market, with data output rates more than doubling annually (101): the initial Genome Analyzer could sequence 1 giga bps (Gbps) of data in a single run; by 2014, this figure had

increased to 1.8 tera bps (Tbps) with the HiSeqX Ten sequencer (101) or 6 Tbps with the NovaSeq6000 (98).

Currently, Illumina is undoubtedly the NGS technique used the most (58,95). Like the 454 technology, it is based on the sequencing-by-synthesis principle (64,94), and resembles the Sanger method in that it also relies on the incorporation of dye terminator nucleotides into the sequence (94). However, the Illumina terminators are reversible, meaning that the polymerisation can continue even after the fluorophore detection. The workflow of this technology is as follows (64,101):

- Library preparation: the DNA or cDNA sample is fragmented randomly and then ligated to two distinct types of adaptor at the 5' and 3' ends. Alternatively, the fragmentation and ligation reactions can be combined in a single step ("tagmentation"), which increases the efficiency of the process. The next stages, amplification and sequencing, take place in a solid surface called "Flow cell".
- Cluster generation employed to amplify DNA using bridge PCR. The fragments of the DNA template are denatured and the single strands are deposited in the flow cell, the surface of which is covered with primers that complement the adaptors. This complementarity enables the creation of bridges by joining the adapters to the primers (Figure 10.B). The PCR reagents are then incorporated (i.e., the nucleotides and the DNA polymerase) and each fragment is amplified into distinct clonal clusters via bridge amplification. When the second strand is formed, the DNA is denatured again so that new amplification cycles can be performed. It has been calculated that clonal clusters comprising about 1000 copies of each DNA fragment can be obtained; meanwhile, each flow cell can support millions of parallel cluster reactions (95).
- Reversible termination sequencing: when the cluster generation is complete, templates are then ready for sequencing using reversible terminator nucleotides. Each dNTP is marked with a different fluorescent molecule, allowing all four of them to be added at the same time. The complementary nucleotides are then incorporated and those that are not are eliminated. After laser excitation, the emitted fluorescence is recorded by a four-channel fluorescent channel and the first base is identified. A surface chemical treatment removes the fluorescent dye from the incorporated nucleotides. This unlocks the 3' carbon, enabling a new version to be used to continue the sequencing reaction.

The reactions are repeated for 300 or more rounds (95). Stacking and overlying the images obtained from all the cycles enables software to reconstruct the sequence of the DNA template fragment.




Mardis ER. 2008.
Annu. Rev. Genomics Hum. Genet. 9:387–402

**Figure 10. The Illumina sequencing process. The image was taken from Mardis et al. (97) with the permission of Annual Reviews, Inc.**

Illumina produces a variety of sequencing tools for different applications. These include genomic sequencing, targeted sequencing, metagenomics, chromatin immunoprecipitation

(ChiP) sequencing, RNA sequencing, and methylation sequencing (95,101). As stated above, the distinct platforms have varying throughput levels: MiniSeq -7.5 Gbps of data with 25 million reads/run and read lengths of 2x150 bps; MiSeq - 15 Gbps with 25 million reads/run and read lengths of 2x300 bps; and NextSeq - 120 Gbps with 400 million reads/run and read lengths of 2x150 bps (95).

Unlike the 454 technology, which sequences in one direction, the Illumina tools rely on "paired-end sequencing", where both ends of the DNA fragments are sequenced and the forward and reverse reads are aligned as read pairs. So, using the MiSeq platform, a final length of 600 bps would be obtained after joining two 300-bp strands, which is represented by 2x300 bps. As well as producing double the number of reads in the same time and for the same effort concerning the library preparation, sequences arranged as read pairs enable more accurate alignment and the ability to detect insertion-deletion variants (101). Finally, paired-end sequencing also facilitates the detection of genomic rearrangements and repetitive sequence elements like gene fusions and novel transcripts (https://emea.illumina.com/science/technology/next-generation-sequencing/plan-experiments/paired-end-vs-single-read.html?langsel=/es/).

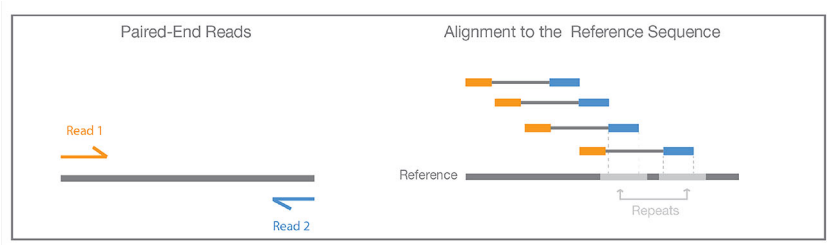



**Figure 11. Paired-end sequencing and alignment. The image was taken from Illumina Inc. (101) with the permission of Illumina, Inc.**

A further advantage of the Illumina technology is that the four dNTPs are present during each sequencing cycle, meaning that natural competition minimises incorporation bias and reduces raw error rates. As a result, it is possible to achieve highly accurate base-by-base sequencing that almost eliminates sequences' context-specific errors, even in repetitive regions and homopolymers (101). Moreover, due to its direct rather than camera-based imaging, fluorescent recording improves the detection speed (95).

Nevertheless, issues may arise with Illumina sequencing. First, extreme base compositions, i.e., guanine-cytosine (GC)-poor or GC-rich sequences, lead to an uneven coverage or even no coverage of the reads across the genome (102). Furthermore, the amount of template-DNA employed has to be quantified accurately to prevent the "overclustering" of the system. An analysis of the sequencing errors generated during the process can, however, be performed to identify the real sequence variants and the protocol-induced artefacts (95).

A common disadvantage of NGS techniques is that the length of the reads produced might not be as long as required, hampering the identification of bacteria (94). Indeed, genomes often contain numerous repeated sequences that are longer than the reads obtained with the technology, which may lead to misassemblies and gaps (103). However, although it is necessary to sequence the entire 16S rRNA gene to reliably identify some species and describe new ones, it is accepted that just sequencing the initial 500-bp region is sufficient for distinguishing a high number of bacteria (104). What is more, changes in community composition can be assessed using gene fragments as small as 100 bps (105). As most of the NGS platforms that are currently employed generate short reads, bacterial identification using these methods has focused on the hypervariable regions of the 16S gene which are, as is already known, extremely informative (94). Nonetheless, NGS tools like GS FLX Titanium or GS FLX + can achieve read lengths of more than 400 bps and up to 700 bps, respectively.

A final issue with short reads relates to the detection and characterisation of large structural variations (SVs), which can be challenging. Smaller variants, like single-nucleotide variations (SNVs) and short indels, can however be identified with greater accuracy (103).

In conclusion, each of the NGS techniques has its particularities, and getting the most out of them requires researchers to strike a balance between target size, read length, depth, sequence accuracy, usability, and cost (105).

### I.4.3. Third-generation sequencing

Although NGS methods have revolutionised biology, there was nevertheless a need to develop new techniques capable of overcoming the drawbacks described above (103). The shift from "long read" (e.g., automated Sanger) to "short read" (e.g., Illumina) technologies has led

to the development of third-generation sequencing (TGS) tools capable of generating longer primary read lengths and maintaining massive parallelisation (95). Globally, these new technologies promise advantages over earlier versions in addition to longer reads, such as higher throughput, faster turnaround times, the need for only small amounts of starting materials, and lower costs (106).

The distinguishing features of TGS technologies are single-molecule sequencing (SMS) and sequencing-in-real-time; this is in contrast to NGS, where the process is paused after each base incorporation (103). SMS refers to techniques that can read the base sequence directly from individual strands of DNA or RNA present in a sample of interest. Consequently, no PCR is required before sequencing, reducing the inherent bias associated with this step, and the DNA preparation time is shorter (107).

In 2011, the company Pacific Biosciences (PacBio) released the first TGS instrument: the PacBio RS sequencer (103). This relies on "single-molecule real-time" (SMRT) sequencing (108), which uses a modified enzyme and allows the enzymatic reaction to be observed directly as it occurs (107). As the DNA polymerase can be viewed as it synthesises the DNA strand, it is possible to leverage the speed and processivity of this enzyme to address many of the shortcomings of NGS (106).

Briefly, the DNA is isolated and the library is created by ligating adaptors to the double-stranded DNA, producing a circular template known as SMRT®bell. A primer and a polymerase are then annealed to the adapter. The SMRT cell, which consists of millions of zero-mode waveguides (ZMWs), is at the core of the sequencer. A single molecule of DNA is immobilised at the bottom of each ZMW, where the replication takes place (103). During the reaction, the enzyme not only incorporates the nucleotide into the complementary strand, but also cleaves off the fluorescent dye previously attached to it (107). Each of the four nitrogenous bases in the DNA is marked with a different fluorophore. The camera inside the machine captures the signal in a movie format in real-time. This provides information about both the fluorescent signal and the signal difference over time, which may be useful for predicting SVs in the sequence (107).

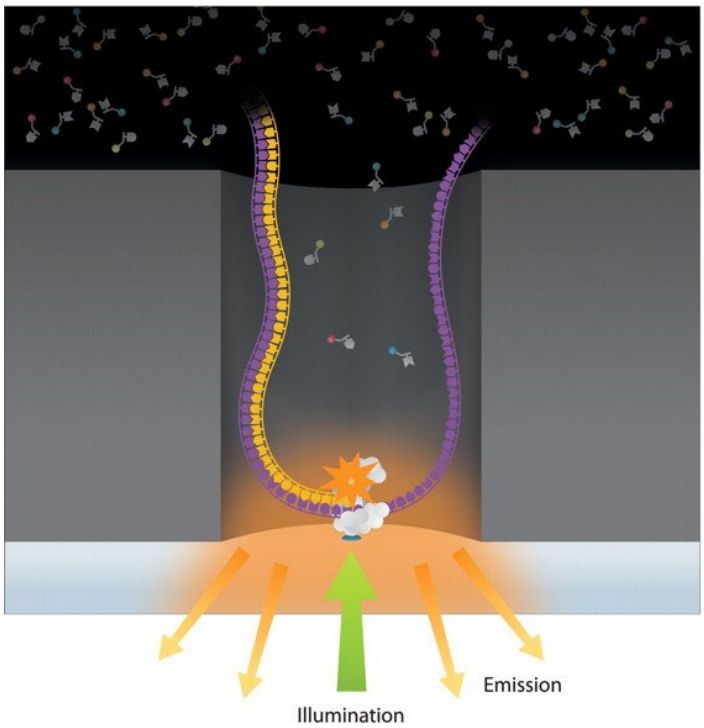



**Figure 12. Schematic representation of a ZMW, with the polymerase-primer-SMRT®bell complex bound to the bottom. The polymerase incorporates fluorescently-labelled nucleotides, emitting a fluorescent signal on illumination from below. The image was taken from Schdat et al. (106) with the permission of Oxford University Press.**

The SMRT technology has improved significantly over time, and new sequencers have been released: RS II, Sequel, and Sequel II. Notably, the throughput per run has increased, while the costs have decreased.

A further TGS technique, known as nanopore sequencing, has been developed by the firm Oxford Nanopore Technology (ONT) (103). Nanopore sequencing relies on both the transit through a hole of a DNA or RNA molecule or its component bases, and the detection of the bases via their effects on an electric current signal (106). Readouts are based on the size differences between all the nucleotides. So, characteristic current modulations are displayed for discrimination purposes for a given nucleotide (107). The ionic current is resumed after the trapped nucleotide is extracted in its entirety (107).

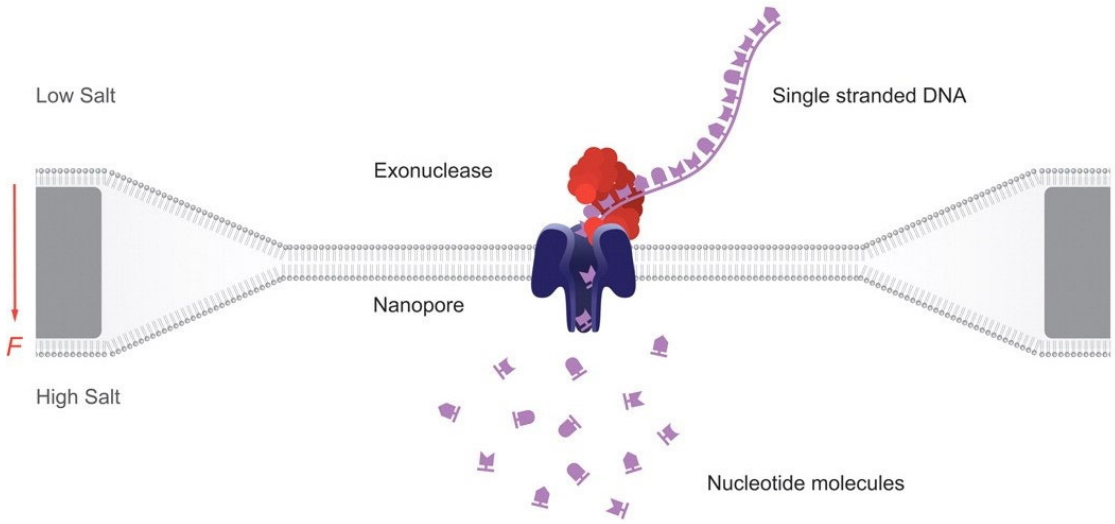



**Figure 13. Schematic representation of the Oxford Nanopore Technology. The image was taken from Schdat et al. (106) with the permission of Oxford University Press.**

In 2014, ONT released its first nanopore sequencer: the pocket-sized MinION (103). Thanks to the rapid evolution of its chemistries, a significant increase in throughput has been achieved. While the early tools produced ~184-450 million bases of sequence data per a 48-hour run, modern flow cells, in combination with the latest library-preparation kits, can produce up to 20 Gbps of sequence data (103). The translocation speed has also increased (103).

In order to achieve greater throughput, several ONT sequencers have been introduced to the market, such as PromethION and GridION X5 (103). The former can generate up to 125 Gbps (one flow cell) or even up to 15 Tbps, as it can contain 48 flow cells that can be run in parallel. Conversely, GridION X5 can produce up to 100 Gbps per run (five flow cells).

Long reads are the most attractive feature of SMS tools, overcoming the limitations of NGS and drastically improving the quality of the genome assembly (109). Moreover, the ONT tools have unique features like ultra-long reads (over 300 Kbs and some close to 1 million bps) and the capacity to directly sequence RNA molecules (109). These ultra-long reads have great potential when it comes to facilitating the assembly of large genomes (109). A further advantage of the ONT technology over the PacBio methods is its ability to differentiate between modified nucleotides at high speed (109).

TGS tools can be utilised for different purposes including: the genome assembly of complex organisms; resolving tandem repeats and complex structural rearrangements in human diseases; the phasing of haplotypes; and deciphering the MHC sequence or identifying the correct RNA isoforms (109). Nevertheless, these techniques also have limitations. The raw-read error rates in SMS technologies are generally ≥5%, although their extremely parallel nature can deliver high-fold coverage and a consensus-read accuracy >99% (106). The computational requirements for projects involving long reads are a further drawback (109). In conclusion, it is clear that these long-read technologies are revolutionising research into genomics since they enable scholars to explore genomes at unprecedented resolutions (103).

In table 2 there is a summary of the general characteristics, including advantages and drawbacks, of the main second- and third-generation sequencing technologies explained above.

**Table 2. Summary of the general characteristics of the main second- and third-generation sequencing technologies. The table was modified from Zaura et al. (98), an open-access article distributed under a Creative Commons Attribution 4.0 International (CC BY 4.0) license (https://creativecommons.org/licenses/by/4.0/).**

Bps= base pairs; d= days; EC/BBR= estimated cost/billion base read; EIC= estimated instrument cost; GC= guanine and cytosine; Gbps= gigabase pairs; h= hours; Max.= maximum; Mbps= megabase pairs; RL= read length; SMRT= single-molecule real-time; Tbps= terabase pairs.

| Company/ Technology | Platform | Max. RL (bps) | Error rate | Max. output | EIC | EC/ BBR | Max. run time | Advantages | Drawbacks |
|---|---|---|---|---|---|---|---|---|---|
| Roche/ pyro-sequencing | 454 FLX+ | 700 | 0.1% | 700 Mpbs | $100k | $10000 | 23 h | Long read length | Expensive runs; homopolymer errors; phased out |
| Illumina Solexa | MiSeq | 2x300 | 0.1% | 15 Gbps | $100k | $100 | 56 h | High accuracy | GC bias |
| | NextSeq 550 | 2x150 | 0.1% | 120 Gbps | $250k | $30 | 30 h | High accuracy; high throughput; low cost per base | |
| | HiSeq2500 | 2x125 | 0.1% | 1 Tbps | $750k | $30 | 6 d | | GC bias; short reads; expensive instrument |
| | NovaSeq6000 | 2x125 | 0.1% | 6 Tbps | $850k | $10 | 44 h | | |
| Pacific Biosciences/ SMRT sequencing | RS II | 15000 | 15% | 1 Gbps | $700k | $400 | 4 h | Very long reads; fast; no amplification bias | High error rate; Expensive instrument; high cost per base |
| | Sequel | 30000 | 15% | 10 Gbps | $350k | $85 | 20 h | Very long reads | High error rate |
| Oxford nanopore/ nanopore sequencing | MinION | >30000 | 10% | 20 Gbps | $1k | $75 | 48 h | Very long reads; cheapest instrument; no amplification bias | High error rate |
| | PromethION | >30000 | 10% | 15 Tbps | $160k | $10 | 48 h | Very long reads; lowest cost per base | |

### I.4.4. Fourth-generation sequencing

Recent years have seen the development of fourth-generation sequencing techniques, such as *in-situ* sequencing (ISS) (110) and fluorescent ISS (FISSEQ) (111). These methods utilise NGS chemistry to sequence single RNA molecules directly in fixed cells and tissues (110,111).

ISS is a targeted approach in which a specific or random primer is used to retrotranscribe mRNA. The synthesised complementary DNA (cDNA) is constructed as single-stranded, and a padlock probe is hybridised to complementary target sequences on the cDNA molecule, leaving a gap between the probe arms. This gap, which is the sequencing target, is filled by way of DNA polymerisation and then DNA ligation to form a complete DNA circle. This DNA circle can be replicated via target-primed, rolling-circle amplification, and the amplification product is subjected to sequencing by ligation chemistry (112).

In contrast, FISSEQ is a non-targeted method where the cDNA is synthesised by random primers containing a sequencing adaptor. The resulting cDNA is cross-linked to the cellular matrix with a cross-linking reagent to ensure its immobilisation. The spatial information is therefore conserved. The newly synthesised cDNA is self-circularised to form a DNA circle. This is followed by clonal amplification using the rolling-circle technique. Finally, sequencing by ligation is performed to sequence the target fragment (112).

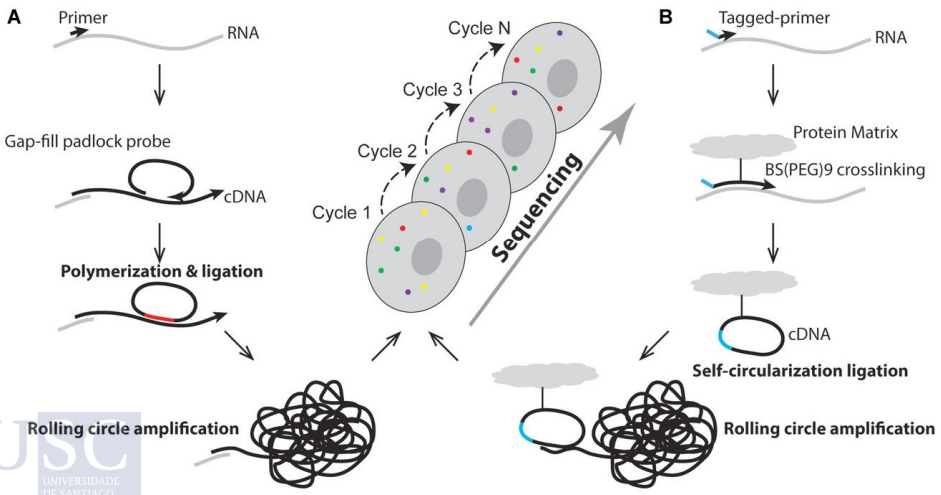



**Figure 14. Schematic representation of A) ISS, and B) FISSEQ. The image was taken from Ke et al. (112), an open-access article distributed under a Creative Commons Attribution 4.0 International (CC BY 4.0) license (https://creativecommons.org/licenses/by/4.0/).**

An alternative fourth-generation technique is a spatial transcriptomics. This enables the visualisation and quantitative analysis of the transcriptome, with spatial resolutions in individual tissue sections (113). Specifically, this method combines *in-situ* transcript mapping with *ex-situ* transcript identification by way of NGS (112).

These fourth-generation technologies, although still in their infancy, are now being used in projects aiming to unravel the functions of the brain as well as in cancer research (112).

### I.5. ANALYSIS OF SEQUENCING RESULTS

The NGS platforms generate a huge amount of genomic data that is unmanageable with an ordinary computer, meaning that informatics has become vital for processing and analysis purposes (114). In parallel with the use of high-throughput 16S rRNA gene sequencing, bioinformatics has emerged as a discipline that conceptualises biology in terms of macromolecules and then applies informatic techniques (applied maths, computer science, and statistics) to understand and organise the huge amount of data associated with these molecules (115). In other words, it is the tool used to make sense of sequencing results: signals are converted to data, data to interpretable information, and information into actionable knowledge (116).

The concept of a pipeline refers to a set of bioinformatic algorithms executed in a predefined sequence to process NGS data. Accordingly, the data flow is transformed into a process comprised of several sequential phases where the input of each is the output of the previous stage. Different local or web-based software packages have been developed to manage the amplicon sequence data from NGS, including the commonly used quantitative insights into microbial ecology (QIIME) (117), mothur (118), and the ribosomal database project (RDP) pipeline (119). QIIME 2 has recently emerged (120); the first version is therefore no longer supported, as efforts are now focused entirely on its successor (http://qiime.org/). This new pipeline has the potential to serve not only as a marker-gene analysis tool, but also as a multidimensional and powerful data-science platform that can be quickly adapted to analyse diverse microbiome features. It also makes use of many new interactive visualisation tools that facilitate exploratory analyses and the reporting of results. Furthermore, QIIME 2 provides a software-development kit that can be used both to integrate the technology with other systems and develop interfaces targeted towards users with different levels of computational knowledge and experience (120).

Regardless of the software or platform employed, the complete bioinformatics processing of 16S rRNA gene amplicon data usually encompasses three steps: 1) the pre-treatment or quality filtering of raw sequence data; 2) the construction of operational taxonomic units (OTUs) or single-nucleotide resolution table and 3) advanced data analysis and visualisation (98,121) (Figure 15).

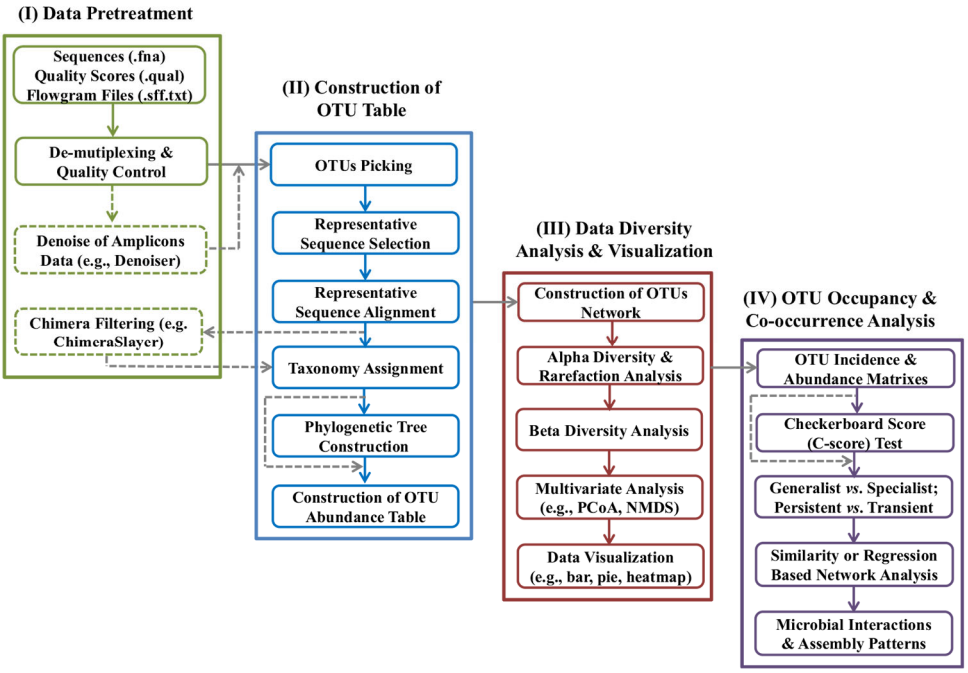



**Figure 15. Example of a flow chart of a 16S rRNA gene amplicon data analysis pipeline using OTUs. The image was taken from Ju and Zhang (121) with the permission of Springer Nature.**

### I.5.1. Pre-treatment of raw sequence data

The sequencing process generates a sff binary file. This provides general information on a run (number of flows, order of the nucleotides in the flows…) and follows it up with a description of each run sequence, indicating the position where it was generated, the pipeline used, and the total number of bases. It also provides information on the bases incorporated and the quality assigned to each of them. The pipeline transforms the sff file into readable fasta and qual versions by applying a particular command (121). This enables the first step to begin, i.e., processing the raw barcoded sequence data, which is a process known as demultiplexing. Once this phase has been completed, all the raw sequence data can be now demultiplexed, which is a procedure that sets it into individual subsets belonging to different samples based on specific barcodes (121).



Raw reads are then quality filtered by applying criteria such as the minimum average quality score allowed in a read, the maximum number of ambiguous bases, the minimum and maximum sequence lengths, and the maximum length of homopolymer and maximum mismatches in the primer or barcodes (121). Next, the sequence-alignment step determines

where each short DNA sequence aligns with the reference genome, i.e., the 16S rRNA gene (122). This makes it possible to discern the location of the sequenced fragment in the gene.

The PCR amplification and sequencing can introduce bias into the process, including PCR single-base errors, PCR chimeras, and sequencing errors, which must be checked and removed (121). If this is not done, the true diversity of the bacterial community would be overestimated. Particular attention should be paid to the chimeras or chimeric amplicons. These are artificial DNA sequences generated during PCR amplification and consist of a combination of two (or more) true underlying sequences (84). They appear when the extension step for an amplicon is brought to an end, with the short product obtained functioning as a primer in the next PCR cycle. This amplicon anneals to the incorrect DNA template and continues the extension, synthesising a single sequence sourced from two different templates. These chimeric amplicons can be overamplified in the steps that follow those described, thus creating unreal taxa and distorting the results.

AmpliconNoise (123) and Denoiser (implemented in QIIME) are two of the most widely used software applications to remove or correct the PCR and any sequencing errors; meanwhile, ChimeraSlayer (124) (QIIME default method) and UCHIME (125) (mothur default method) are some of the tools employed to filter the PCR chimeras (121).

### I.5.2. Sequence clustering

The second step in processing the 16S rRNA gene amplicon data begins by clustering the clean sequences into OTUs. An OTU is a cluster of organisms that are similar at the sequence level beyond a particular threshold, and which are intended to correspond to taxonomic clades (84,126). Sequence differences in the selected variability radius are assumed to be due to the variation within the taxonomic group or to random sequencer noise (127), which avoids the problem of differentiating biological from technical sequence variations but at the cost of taxonomic resolution (128).



Several identity cut-offs have been used for the different taxonomic ranks. Typically, sequences are clustered at the ≥97% similarity threshold, which has been conventionally regarded as the species-level correspondent (98,129). Conversely, the MEGAN pipeline recommends thresholds of ≥99% and ≥97% for the species and genus levels, respectively (121).

Nonetheless, the sequence-similarity levels used are imprecise measures of an imprecise concept of a "species", and the sequence identity of a given region of the 16S rRNA gene does not reflect the precise identity of the entire gene (105).

The assignment of sequences to OTUs is known as "binning" (84), and numerous OTU clustering algorithms have been integrated into the popular sequence-analysis pipelines, such as QIIME2 (120), mothur (118), and USEARCH (130). Overall, they use three different strategies (121):

- *De novo*: sequences are clustered without a reference database.
- Closed reference: sequences are matched against a reference database; those unmatched at the given identity cut-off are discarded.
- Open reference: sequences are first picked for closed-reference OTUs and the unmatched reads are subsequently clustered for *de novo* OTU versions.

Currently, there is mixed evidence on which strategy is best when attempting to define OTUs and reveal the observations closest to the true community (131). Although the *de novo* approach enables the exploration of uncharted territories in the microbiota (105) and has been shown to create higher quality OTU classifications (132), the reference-based method has several advantages. First, sequence data from different gene regions, or generated from distinct sequencing technologies, can be combined using reference databases (105). In these cases, *de novo* OTU-picking might wrongly assign the same organisms to different OTUs based solely on variations in the DNA region amplified or in the sequencing technique (105). Second, the reference-based approach is increasingly valuable as the scope of publicly available data is expands, enabling new research to be interpreted in the context of existing studies (105). Picking OTUs against a reference database can also diminish the impact of chimeras and noise data (105).

However, a single OTU can contain groups of sequences that could be individually assigned to different, related taxon (98) and the three OTU clustering approaches produce different results in terms of obtaining OTUs, even when using the same dataset (132,133). Moreover, the same method can yield distinct results after only a minor parameter change (134).

More recently, distinct error-correction or denoising approaches have become available, which are based on algorithms that use a single-nucleotide resolution (i.e., 100% sequence similarity) by generating amplicon sequence variants (ASVs), thus improving the taxonomic determination (128). These methods attempt to model the error of the sequencer and to cluster reads in a way that their distribution within clusters is consistent with such error (127).

Among the most widely-known ASV-based pipelines there are DADA2, Deblur, and UNOISE (135-137); and they differ in how the above-mentioned correction is done (128). For example, DADA2 generates a parametric error model that is trained on the entire sequencing run and then applies that model to correct and collapse the sequence errors into ASVs (135). For its part, Deblur aligns sequences together into "sub-OTUs" and, based on an upper error rate bound along with a constant probability of indels and the mean error rate, removes predicted error-derived reads from neighboring sequences (136). Also, the UNOISE3 pipeline uses a one-pass clustering strategy that depends on two parameters with pre-set values that were curated by its author to generate "zero-radius OTUs" (137). Lastly, other algorithms use the 100% sequence similarity to create oligotypes, or minimum entropy decomposition nodes (138). Despite the different nomenclatures indicated by the respective researchers to refer to the clusters, they all are commonly known as ASVs.

Different investigations have compared the two sequence clustering approaches to discern which performs better (127,128,139-142). In these studies, authors contrasted one (128,140-142), two (127) or three (139) OTU to one (140,142), two (141) or three (127,128,139) ASV clustering methods, using sequences derived from mock (127,128,139,141), human gut (128,139,141), shrimp gut (142) and soil (128) samples. Also, other researchers used 16S rRNA gene sequences from the rRNA operon copy number database (rrnDB) for its analysis (140,143). In general, the ASV pipelines had demonstrated superior sensitivity, specificity, and precision, and lower spurious sequence rates when compared to OTU algorithms (127,139). Moreover, they allow for easier inter-study integration of biological features as the ASVs have intrinsic meaning independent of the reference database used, contrary to the study-specific nature of OTUs (139,144). Still, ASV-level pipelines are not free of limitations and can fail to distinguish very closely related true biological sequences and clump them together into a single ASV (139). Also, when analysing 16S rRNA gene data, Schloss (140) has recently affirmed

that the risk of splitting a single genome into separate clusters when using ASVs is of higher importance than the risk of grouping together ASVs from distinct taxa into the same OTU.

Lastly, there is no consensus regarding the influence of the method chosen on the diversity results obtained. Meanwhile, some authors obtained minor differences between pipelines using the two clustering methods, with comparable alpha- and beta-diversity profiles (concepts that will be further explained) (141,142); others evidenced distinct results even among those from the same approach (128,139). In fact, Nearing et al. (128) found that, despite the similar general community structures, the alpha-diversity metrics varied considerably among all pipelines evaluated, even within the ASV-based DADA2 (135) and UNOISE (137). So, they concluded that the clustering pipeline choice will largely impact the alpha-diversity results among samples.

In table 3 there is a brief description of several tools available for sequence clustering into OTUs or single-nucleotide resolution. It should be noted that, as said above, the main pipelines QIIME (117), QIIME2 (120), mothur (118), and USEARCH (130) had their own approaches for OTU clustering.

**Table 3. Tools available for sequence clustering. The table was modified from Zaura et al. (98), an open-access article distributed under a Creative Commons Attribution 4.0 International (CC BY 4.0) license (https://creativecommons.org/licenses/by/4.0/).**

| Tool | Description |
|---|---|
| UPARSE (126) | Implemented in USEARCH (130). Algorithm for OTU clustering |
| QIIME BLAST (145) | An approach that matches the reads to the closest sequence in the database and groups the reads based on the BLAST label |
| CD-HIT-OTU-MiSeq (146) | Approach for clustering and annotation of MiSeq-based 16S sequence data |
| UNOISE (137) | Implemented in USEARCH (130). Creates high-resolution OTUs referred to as zOTUs |
| Minimum Entropy Decomposition (MED) (138) | Information theory-based clustering algorithm for sensitive partitioning of sequences Provides single-nucleotide resolution (oligotypes or MED nodes) |
| DADA2 (135) | Corrects Illumina-sequenced amplicon errors, providing single-nucleotide resolution as ASVs |
| Deblur (136) | Produces sOTU with single-nucleotide resolution (putative error-free sequences) |

Once the sequences are grouped, a single sequence is selected as a representative of each cluster. This sequence can be random, the longest, the most abundant, or the first in a cluster (121). The fact that each cluster is now represented by a single sequence also speeds up the posterior analysis.

### I.5.3. Taxonomy assignment

After clustering, a taxonomic identity has to be assigned to each of the representative sequences. Phylogenetic relationships among sequences can be inferred either *de novo* or by using a reference database with an associated phylogeny (105). In the latter case, the taxonomic assignment process can be performed via two different strategies, the best hit and the lowest common ancestor (LCA) (121). Both of these require an alignment tool to compare the sequences being analysed against a reference database containing related sequences whose taxonomy is known (121). The difference between the best hit and LCA is that the former assigns a sequence based on the alignment with the highest score, while in the latter this is achieved via multiple hits against a particular database (121).

The *de novo* approach can be carried out using tools like NAST (for sequence alignment) (147) and FastTree (for making phylogeny inferences from aligned sequences) (148). RDP (119), Greengenes (149), and SILVA (150) are among the most widely used databases at the taxonomic assignment stage and are used in combination with pairwise alignment tools like BLAST (145) or USEARCH/UCLUST (130).

Some databases, such as CORE (76) and the human oral microbiome database (HOMD), are specialised in oral microbiota (151). They emerged to provide a comprehensive and minimally redundant representation of the bacteria that usually reside in the human oral cavity, with computationally robust classifications at the genus and species levels. In fact, although larger public databases like GenBank (152) and RDP (119) return named matches for a slightly higher fraction of sequences identified in analyses of clinical samples, CORE and HOMD are much more likely to do so accurately (76). Nonetheless, the larger databases are still important supplements to the specialised versions when it comes to recognising rare species.

Performing a diversity analysis first requires the generation of a phylogenetic tree of OTUs or ASVs. Initially, the representative taxa sequences have to be aligned using tools like

MUSCLE (153) or PyNAST (154). The tree can then be constructed, which allows the relationships between the sequences to be visualised in terms of their evolutionary distance from a common ancestor (Figure 16). Many different packages have been developed for inferring phylogenies and building trees for multiple sequence alignments, with MEGA (155) being the most popular and versatile (121). At the very least, this type of software generates a table detailing the number of times that an OTU/ASV is observed and which taxa it represents (156).

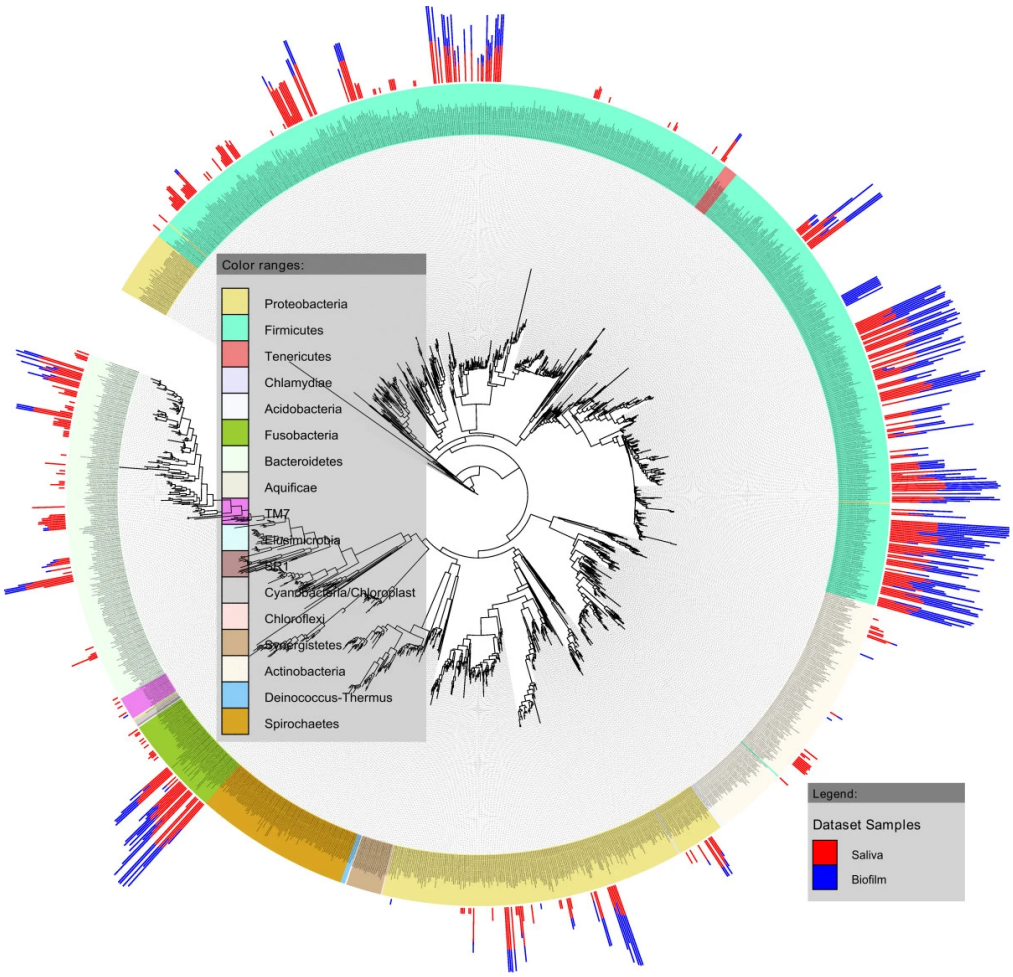



**Figure 16. Phylogenetic tree based on 1,642 HOMD (151) reference sequences. The image was taken from Edlund et al. (157), an open-access article distributed under a Creative Commons Attribution 2.0 Generic (CC BY 2.0) license (https://creativecommons.org/licenses/by/2.0/).**

### I.5.4. Advanced data analysis and visualisation

Understanding the compositional differences of microbial communities is essential in the field of microbial ecology (158). In this regard, an OTU or ASV table enables different taxonomic summaries to be obtained that show the bacteria present and their relative

abundances at all taxonomic levels (156). However, further analysis is required to understand the quality of the data, the diversity within and between samples, and, ultimately, which statistical comparisons are needed to determine whether the microbiota has experienced flux or dysbiosis (156).

A metadata table is also required to perform advanced exploratory and inferential analyses. The term “metadata” refers to the information associated with the sequences, including the environmental conditions and the time and location of the sample collection (105). Metadata is very important, as it is required if the aim is to replicate a particular investigation (159). As stated above, it is also essential for performing meaningful comparisons between samples or with specimens from other studies. Consequently, genomic sequence data that lack an environmental context have no value (159), meaning that aspects like the host’s health, sex, age, and diet, as well as the method of sampling, the size of the sample and its preparation, should all be recorded (159).

#### I.5.4.1. Analysis tools

There are several methods for comparing sequencing data, including QIIME (117), mothur (118), MEGAN (160), metagenomic-rapid annotation using subsystem technology (MG-RAST) (161), UniFrac (162), DOTUR (163) and Metastats (164). Most of these tools are used for conducting analyses of bacterial communities and can detect groups of related samples (165). However, they do not provide information on the phenotypes or environmental conditions associated with these communities, and do not usually identify the biological features responsible for bacterial group relationships (165). In fact, only Metastats (164) explicitly couples a statistical analysis (to assess if the metagenomes differ) with the identification of biomarkers (to detect features characterising the differences), based on repeated *t* statistics and Fisher's tests on random permutations (165).

Nevertheless, none of the aforementioned approaches produce explanations on biological classes with which to establish statistical significance and biological consistency, or estimate the size of the effects of predicted biomarkers. In 2011, Segata et al. (165) developed the linear discriminant analysis (LDA) effect size (LEfSe) method, which is a logarithm for detecting and explaining high-dimensional biomarkers. This couples standard tests for statistical significance

with tests encoding biological consistency and effect relevance, enabling the features (organisms, clades, OTUs, ASVs, genes, or functions) most likely to explain differences between two or more biological conditions (classes) to be identified. In particular, the effect size gives an estimation of the magnitude of an observed phenomenon that is due to each feature, and is thus a valuable tool for both ranking the relevance of different biological elements and designing further investigations and analyses.

Two years later, McMurdie and Holmes (166) developed the phyloseq software package for the R language (167). This is dedicated to the object-oriented representation and analysis of phylogenetic sequencing. One of its original aims was to leverage R-based resources for reproducible research and, by that means, improve the reproducibility and portability of the published microbial analysis. Phyloseq allows the importation of the most common output formats of the most common clustering applications, including QIIME (117), mothur (118), and the RDP pipeline (119). It also enables the OTU/ASV and taxonomy tables, phylogenetic tree, "representative sequence" fasta file, and metadata file to be incorporated in a single "phyloseq-class" R object (168). Thereafter, the researcher can utilise all the statistical and graphical tools available in R (167) to generate reproducible research reports with attractive graphics (168).

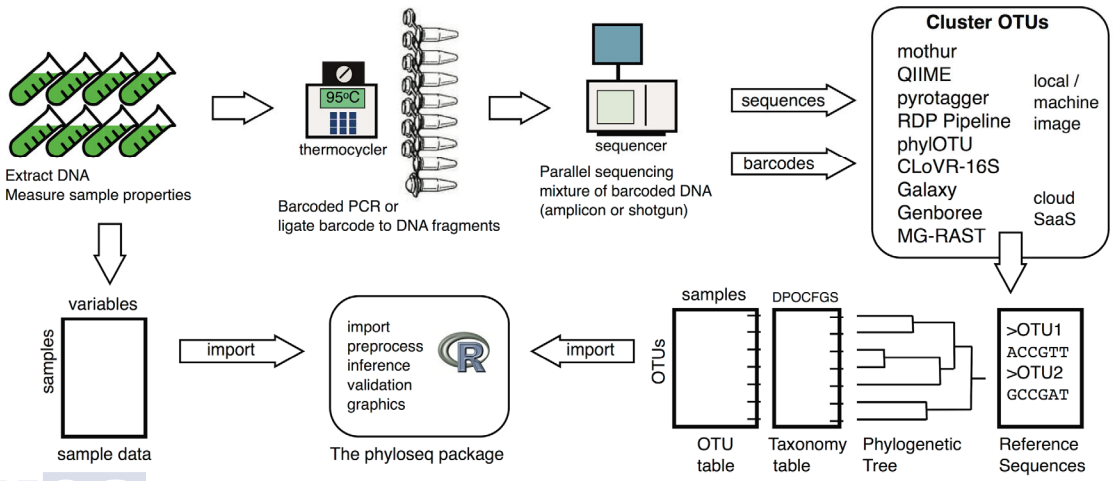



**Figure 17. Diagram of an experimental and analysis workflow for amplicon or shotgun phylogenetic sequencing, showing the intended role of phyloseq. The image was taken from McMurdie and Holmes (166), an open-access article distributed under a Creative Commons Attribution 4.0 International (CC BY 4.0) license (https://creativecommons.org/licenses/by/4.0/).**

Combining phyloseq with other important R packages like vegan (169), DADA2 (135), differential expression analysis for sequence count data version 2 (DESeq2) (170) and ggplot (171) makes it possible to perform powerful and specific analyses of amplicon-sequenced microbiota data (168). Figure 18 represents the workflow for the analysis of amplicon data within the R environment (167). This demonstrates how it takes the amplicon-sequencing reads and associated sample metadata as inputs, while its outputs are exploratory and inferential statistical analyses, as well as sharable analysis scripts and data files that fully reproduce those analyses (168).

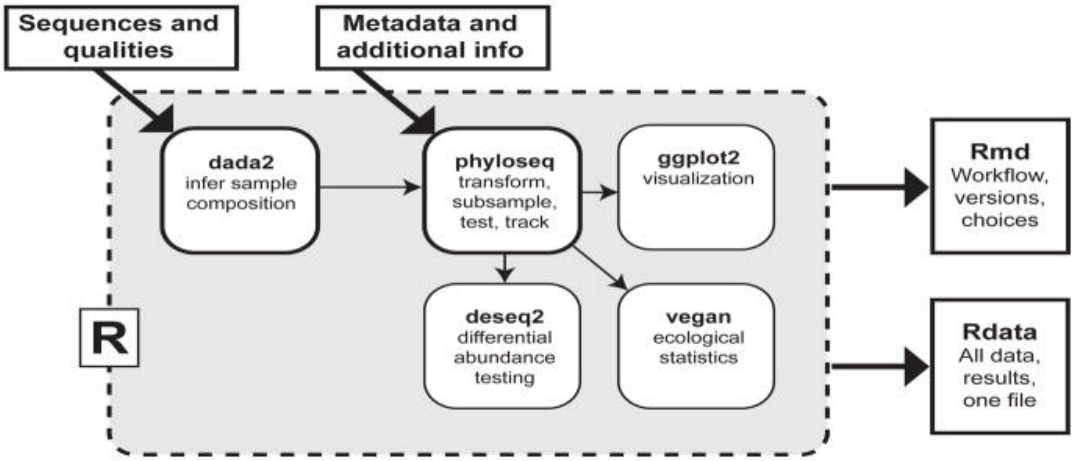



**Figure 18. Diagram of the R workflow for the analysis of amplicon data, including denoising, data integration, and statistical analysis. The image was taken from Callahan et al. (168), an open-access article distributed under a Creative Commons Attribution 4.0 International (CC BY 4.0) license (https://creativecommons.org/licenses/by/4.0/).**

In 2015, McMurdie and Holmes (172) presented Shiny-phyloseq, a shiny-based web application for dynamic interactions with data on the microbiota. For many researchers with classical training in biological science, learning a programming language like R (167) requires a huge investment of time and effort. However, many of the necessary computations are not only tractable by R (167) but are also fast enough for dynamic interactions via a graphical user interface (GUI). Consequently, Shiny-phyloseq is a web-browser GUI that leverages phyloseq and other R (167) resources for the analysis of data on the microbiota. Moreover, the application records the complete inputs and subsequent graphical results of a user's session, allowing them to be archived, shared, and reproduced without the need to write any new codes.

The microbiome R package was released later (167,173). This tool facilitates the manipulation, statistical analysis, and visualisation of taxonomic profiling data, in particular 16S taxonomic profiling. The package supports the independent phyloseq data format and expands the available toolkit to facilitate the standardisation of analyses and the development of best practices.

### I.5.4.2. Biodiversity of the bacterial community

The concept of diversity refers to the variability among organisms from ecological complexes, of which microbes are part (174). Biodiversity is one of the main indicators of microbiota health (114). In fact, several diseases are correlated with reduced microbiota diversity, presumably as one or a few microbes overgrow during immune-system or nutrient imbalances (84).

Numerous metrics have been developed to evaluate the diversity within (alpha-diversity) and between (beta-diversity) populations. This enables differences in diversity to be estimated qualitatively or quantitatively. In the former, only the presence/absence of taxa is considered, while the latter also takes into account the abundance of any observed microorganisms (114). Moreover, the diversity measurements can be further categorised as species-based or divergence-based (175). Species-based measures have been developed extensively and rely on the species as the fundamental unit of analysis (175). References to a cluster of 16S rRNA sequences often employ the terms OTU/ASV or phylotype instead of species. In contrast, divergence-based methods account for the fact that not all species or phylotypes within a sample are related to each other equally (175). The diversity metrics used most in the literature are explained below.

#### *I.5.4.2.1. Alpha-diversity*

Alpha-diversity is the measure of diversity within a single sample (156), and is a common first approach for assessing differences between environments (176). More specifically, the alpha-diversity metrics summarise the structure of an ecological community concerning its richness (number of taxonomic groups), evenness (distribution of abundances of the groups), or both (176). Accordingly, they act as a statistical summary of a single population (84). Figure 19 illustrates the richness and evenness measures, with each shape representing an organism and the colour and nature of the shape portraying an organism of a different type.

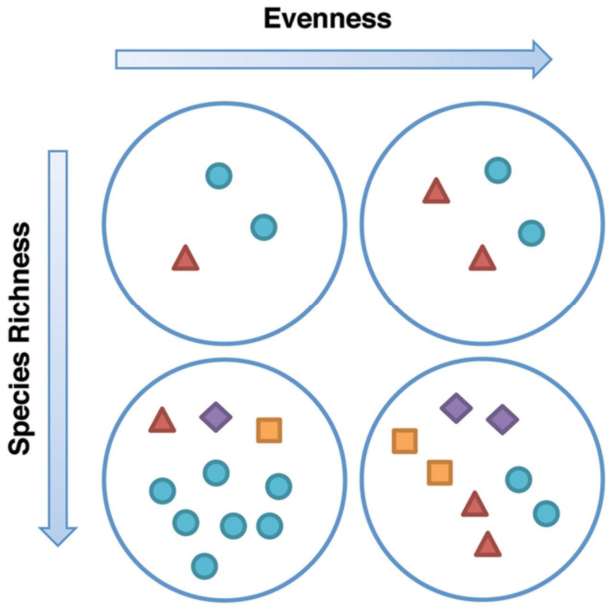



**Figure 19. Diagram demonstrating richness and evenness and how they describe the composition of a community. The image was taken from Cox et al. (177) with the permission of Oxford University Press.**

*A. Richness*

Sample richness is the most basic form of alpha-diversity (156) and answers the question: "*How many taxa are detected in a sample*?" The simplest way to measure this is using the "observed richness", which just involves counting how many different taxa are present in a sample (178). Consequently, the more taxa present, the richer the sample is. However, observed richness does not determine the number of individuals of each taxa, giving equal weight to those with very few individuals.

As more of a community of interest is sampled, the number of organisms observed increases (179) and its true diversity is closer to being revealed (176). The relationship between the number of species types observed and the sampling effort provides information about the total diversity of a sampled population. This pattern can be visualised by plotting an accumulation or a rank-abundance curve (179). The former illustrates the cumulative number of the types observed *vs.* the sampling effort (Figure 20.A). If all communities have a finite number of species and sampling is continued, the curve will eventually reach an asymptote at the point of actual richness (179). Accordingly, the curve provides information about how well communities have been sampled: the more concave-downwards it is, the better the sampling (179). The rank-abundance curve also shows how well ecosystems have been sampled. In this

representation, the species are ordered from most to least abundant on the x axis, and the abundance of each one is plotted on the y axis (Figure 20.B) (179).

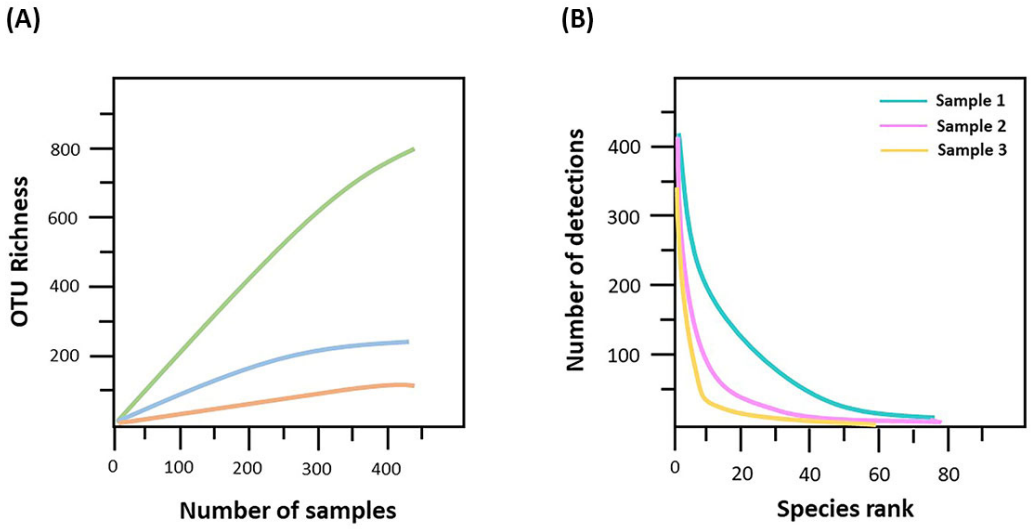



**Figure 20. Example of an accumulation curve (A) and a rank-abundance curve (B).**

Although these two curves provide us with information about the richness of a sample, other more robust measures should be employed. The rarefaction method proposed by Sanders in 1968 (180) compares the observed richness in sites, treatments, or habitats that have been sampled unequally (179). This involves selecting a minimum library size ($N_{L,min}$) and then discarding the libraries with fewer reads than this minimum number (181). The "library size" concept refers to the total reads per sample, and $N_{L,min}$ is often selected to be equal to the size of the smallest library that is considered not to be defective. The process of identifying defective samples thus carries a risk of subjectivity and bias. Finally, the remaining libraries are subsampled without replacement, which means that all of them are $N_{L,min}$ in size (181). These subsamples of equal size enable the calculation of diversity metrics in a way that contrasts ecosystems "fairly", irrespective of any initial differences in sample sizes (176).

Nonetheless, rarefaction is neither justifiable nor necessary (176). Imagine a comparison of two samples: one of 100 sequences and another of 1000. By rarefying, 100 sequences are taken from the second sample to ensure it has the same number as the first. This step cannot be reproduced because it is in itself random. Furthermore, McMurdie and Holmes (181) have

demonstrated statistically that rarefying is inappropriate since it affects variance and has a direct impact on statistical power (Figure 21).

| ORIGINAL ABUNDANCE | | |
|---|---|---|
| | Sample A | Sample B |
| OTU1 | 62 | 500 |
| OTU2 | 38 | 500 |
| **Total** | **100** | **1000** |

| RAREFIED ABUNDANCE | | |
|---|---|---|
| | Sample A | Sample B |
| OTU1 | 62 | 50 |
| OTU2 | 38 | 50 |
| **Total** | **100** | **100** |

| STANDARD TESTS FOR DIFFERENCE | | |
|---|---|---|
| P-value | Chi-2 | Fisher |
| Original | 0.0290 | 0.0272 |
| Rarefied | 0.1171 | 0.1169 |

**Figure 21. Example of the effect of rarefying on statistical power. The image was modified from McMurdie and Holmes, an open-access article distributed under a Creative Commons Attribution 4.0 International (CC BY 4.0) license (https://creativecommons.org/licenses/by/4.0/).**

Researchers cannot identify every microbe present in a community: samples are obtained and the microbial populations present are investigated (176). The findings gleaned from the samples enable inferences to be drawn about an environment of interest (176). As noted previously, more sampling takes researchers closer to achieving a true and complete understanding of the microbial community under study (176). However, it is almost impossible to identify every single taxon in a microbial sample, which requires the use of techniques that take into account the incompleteness of the inventory (178).

One way to calculate the true richness of a specimen is to consider the distribution of the tail of the taxa abundance; more specifically, the number of singletons (taxa observed once) and doubletons (taxa observed twice). In other words, the quantity of undetected taxa is estimated based on these numbers. This is achieved using the Chao1 estimator (182), the formula of which is as follows (178):



$$S_{est} = S_{obs} + \frac{f_1^2}{2f_2}$$

In the equation, $S_{est}$ is the estimated taxa richness, $S_{obs}$ is the observed taxa richness, $f_1$ is the number of singletons, and $f_2$ is the number of doubletons. This index is especially useful for datasets skewed towards low-abundance classes, as is likely to be the position in the case of microbes (179,182).

Related to Chao1 is the abundance-based coverage estimator (ACE) (183), which not only considers the ratio of singletons and doubletons but also of all the taxa observed up to an arbitrary count, usually set at 10 (178):

$$S_{ace} = f_{abund} + \frac{f_{rare}}{C_{ace}} + \gamma_{ace}^2 \frac{f_1}{C_{ace}}$$

$$C_{ace} = 1 - \frac{f_1}{n_{rare}}$$

$$\gamma_{ace}^2 = \max\left[0, \frac{f_{rare} \sum_{i=1}^{10} i\,(i-1)\,f_i}{C_{ace}\, n_{rare}\,(n_{rare}-1)} - 1\right]$$

In the ACE formula ($S_{ace}$), $f_{abund}$ is the number of taxa above the abundance threshold and $f_{rare}$ is the number below it (rare samples) (178). It should be noted that the sum of these two values equates to the total number of taxa observed (179). Additionally, $C_{ace}$ is a sample-coverage estimator and $y_{ace}$ is the estimated coefficient of variation for rare taxa (178). In their respective equations, $n_{rare}$ refers to the total number of individuals in rare taxa and $f_i$ to the number of taxa observed "i" times (178). Essentially, this index uses the number of rare taxa (≤10) and the number of singletons ($f_1$) to estimate how many more undiscovered taxa there might be. Nevertheless, both Chao1 and ACE underestimate true richness in small sample sizes (179).



Moreover, there are richness measures that consider the phylogenetic diversity (PD) of populations. The faith phylogenetic diversity index, first described in 1992, is a qualitative divergence-based measure that calculates the total branch length in a phylogenetic tree that

includes all the taxa in a sample (175,184). Although it fulfils the requirement of being a measure of taxon richness in a community, this index is highly sensitive to the sampling effort because it assumes that the total diversity of the population has been sampled (175). Moreover, the PD depends on the method employed to infer branch lengths on a tree, making it sensitive to errors during the tree's construction (175).

### *B. Evenness*

Along with richness, it is also important to measure the evenness of a sample's distribution (178). Take, for example, two specimens, A and B, with the following compositions:

**Table 4. Example explaining microbial evenness.**

| Bacterial phyla | Sample A (number) | Sample B (number) |
|---|---|---|
| Actinobacteria | 690 | 330 |
| Fusobacteria | 200 | 330 |
| Proteobacteria | 110 | 340 |
| *TOTAL* | 1000 | 1000 |

Both samples have the same richness (three types of bacteria) and the same total number of bacteria. However, B is more even than A, because the total number of bacteria is evenly distributed between the three phyla. Conversely, most of the bacteria present in sample A are *Actinobacteria*, with a few representatives of *Fusobacteria* and *Proteobacteria*. Consequently, specimen A is less diverse than B. As can, therefore, be seen, evenness is a measure of the relative abundance of the different taxa in a sample (174).

In general, when richness and evenness increase, so does diversity (174). Diversity can be viewed as a summary of a community's structure since membership, abundance, and evenness are taken into account (177). Traditionally, the Shannon-Weaver (185) and Simpson (186) indices have been used to estimate diversity (174) and, taken together, are a measure of species richness and evenness. However, neither of them is free of bias and, while the former attaches greater weight to species richness, the latter accounts more for species evenness (174).



The Shannon-Weaver (alternatively: Shannon entropy or, simply, Shannon diversity) index (185) was originally proposed as a measure of entropy within the text and quantifies the

uncertainty of predicting correctly what the next individual taken from a sample will be (177) (156). So, if a sample contains 1000 bacteria and 900 of them are *Fusobacteria*, the probability that the next one is also *Fusobacteria* is high and the Shannon index value would be low (close to 0). In contrast, if every 100 bacteria belong to 10 different species, the ability to estimate what the next one will be is low and the Shannon index score would be high. The value thus increases along with the number of species and as the distribution of individuals among the species becomes more even (174). The formula of the index is as follows, with *S* being the number of taxa and $p_i$ the proportion of the community represented by a taxa i:

$$H = -\sum_{i=1}^{S} (p_i \ln p_i)$$

This estimate enables the employment of a further measure: the Pielou evenness index (187), which divides the observed value of the Shannon index by the highest possible value, i.e., the value if all the species in a sample are equally abundant (178).

The Simpson Index (186), first described in 1949, estimates species dominance and reflects the probability that two individuals taken at random from a sample will belong to the same taxa (177,178). Its values range from 0 to 1, with 0 being “infinite diversity” and 1 “no diversity”. Consequently, the score produced by the index increases as diversity decreases (174). This is described mathematically as follows:

$$\lambda = \sum_{i=1}^{S} p_i^2$$

Here, λ represents the Simpson Index, *S* the total number of taxa in the community, and $p_i$ the proportional abundance of each taxa i (178). Since the value of the index increases as diversity decreases, it is usually represented as its inverse, which is known as the inverse Simpson index (1/λ). Accordingly, an increase in diversity is mirrored by an increased inverse Simpson value (177), which represents the probability that two individuals randomly selected from a specimen will belong to different taxa.



In contrast, Theta (Θ) is an example of an alpha-diversity measure that accounts for both evenness and the divergence between taxa (175). Simply put, it calculates the average

difference between two randomly chosen sequences or individuals in a population. Nonetheless, Theta has not been widely used to measure microbial diversity (175). Meanwhile, over recent years, Cadotte et al. (188) developed three indices of PD that also consider the relative abundance of each taxon in a community. Conversely, other authors have extended metrics like the Shannon and Simpson indices, transforming them into phylogenetically weighted equivalents. These have been shown to outperform the standard measures when it comes to distinguishing healthy from disease-associated human microbiota communities (178).

*I.5.4.2.2. Beta-diversity*

Beta-diversity is the measure of diversity between multiple samples (156), and describes how many taxa are shared between communities, including the absolute or relative overlap (84). Thus, a beta-diversity measure estimates the similarity between populations (84). Consequently, aspects of microbial ecology that are unapparent when examining the composition of individual specimens can be revealed by assessing the differences between samples (189).

There are many different approaches for evaluating the similarity between communities and some of those used the most are described below. Conceptually, these capture different aspects of diversity. Traditional measures like the Jaccard (190) or Bray-Curtis indices (191) focus on the taxa compositional overlap, which is quantified directly from the taxa-count data (192). Considered to be the earliest beta-diversity index, the Jaccard accounts for the relative taxa overlap between two samples, i.e., the ratio of shared taxa among all the organisms sampled (192). This is an incidence-based or unweighted (qualitative) index: it only considers the presence/absence of taxa. Over the years, different abundance-based or weighted variations of the original version have been proposed, including the Chao weighted Jaccard index (193) and the weighted Jaccard index (192). In addition, the widely employed Bray-Curtis similarity index describes the community overlap as the fractional minimum abundance of shared taxa between samples (191). The calculation is performed using the following formula:



$$BC\,(S_1,S_2) \;=\; \frac{\sum |S_{1i} - S_{2i}|}{\sum |S_{1i} + S_{2i}|}$$

$S_1$ and $S_2$ are two samples and $S_{1i}$ and $S_{2i}$ are the abundances of phylotype i in samples $S_1$ and $S_2$ (178). Although it is not very sensitive, this index is nonetheless appropriate for use with cero-inflated datasets (178).

Unlike the traditional measures, the recently developed phylogenetically-informed indices do not treat taxa independently. Instead, these metrics consider the phylogenetic relationships between taxa and quantify the shared evolutionary history between communities (192). Among these new measures is the widely known unique fraction metric (UniFrac), of which there are different versions. The unweighted form was the first to be released and only considers species presence/absence and counts the fraction of the branch length unique to either community (162). Conversely, the weighted UniFrac uses species-abundance data and weights the branch length with the abundance difference (194). In other words, it detects changes in the number of sequences from each lineage, as well as changes in the types of taxa that are present (175). The unweighted version is most efficient for detecting abundance changes in rare taxa, while its weighted counterpart is most sensitive for identifying differences in abundant organisms (158). Nevertheless, neither is particularly powerful when it comes to recognising changes in moderately abundant lineages (158). The third version released was the variance-adjusted weighted (VAW) UniFrac, which moderates the branch proportion difference by its variance, increasing the index's power over the weighted version for detecting the differences between two communities (195). The VAW-UniFrac was used by Chen et al. (158) to introduce generalised UniFrac distances that unify the weighted and unweighted UniFrac versions within a common framework. This combined metric adjusts the weight on the branches to cover a series of distances, ranging from weighted to unweighted, and is designed to apply to, and identify, a much wider range of biologically relevant changes in a microbiota's composition (158).

More recently, Schmidt et al. (192) proposed a novel family of beta-diversity indices that quantify community similarity in the context of taxa-interaction networks: the taxa interaction-adjusted (TINA) and the phylogenetic interaction-adjusted (PINA). The authors argued that because the indices take into account interactions between taxa, they are capable of quantifying new aspects of diversity and can expand possible biological interpretations of diversity patterns in new ways (192).

The distinct approaches to community dissimilarity described, i.e., count-based *vs.* phylogenetic, can highlight different aspects of a population and how it functions. Consequently, combining these different analyses to gain a deeper insight into the system under study may be a valuable next step (178).

### *A. Multivariate analysis*

Multivariate analyses are supplanting simple descriptive investigations of bacteria, and are widely used in microbial ecology, where complex, multidimensional datasets abound. However, the employment of OTUs or ASVs abundances makes it difficult to test the direct association between the composition of the microbiota and environmental factors, due to the high dimensionality, non-normality, and phylogenetic structure of the data (196). Consequently, multivariate analyses first require the researcher to select a methodology for measuring distance before conducting an analysis of estimated distances (196). A distance measure defined between any of two samples can be utilised. Among the numerous metrics that exist, the above-mentioned Bray-Curtis and UniFrac are two of those employed the most.

Many types of multivariate statistical analyses have been used for the assessment of high-throughput datasets, and novel approaches for analysing large-scale datasets are also being developed (197). These methodologies can be categorised based on criteria such as the technique's goal (e.g., interpret relationships, test statistical significance), the type of mathematical problem (regression, ordination, calibration, classification), or the variable response (e.g., linear, unimodal, mixture distribution) (197). These techniques can also be classified according to the primary research objectives, and three categories can be distinguished (197):

- Exploratory methods: these are used to explore the relationships among objects (e.g., samples or sites) based on the values of the variables measured in those objects. These techniques provide a valuable visualisation of object similarities since similar objects are usually positioned close together on the visualisation plot, while dissimilar objects are wide apart.
- Interpretive methods: these 'constrained' techniques use both the main set of measured variables and another of additional explanatory variables.

- Discriminatory methods: these are an extension of the former techniques and are usually known as discriminant analyses (DAs). The goal of DAs is to define discriminant functions (synthetic variables) or hyperspace planes that maximise the separation of objects among different classes (groups).

Of the exploratory approaches, the principal component analysis (PCA) is one of the most widely used and oldest (198). In the main, it is employed to calculate new synthetic variables (principal components), which are linear combinations of the original variables, and accounts for as much of the variance in the original data as possible (199). The first principal component (PC) represents the axis in the multidimensional data-space that would produce the largest dispersion of values. Other PCs are calculated as being orthogonal to their predecessors and are positioned along the largest remaining scatter plot of the values. Consequently, the PCA creates a rotation of the original system of coordinates, meaning that the PCs are orthogonal to one another and correspond to the directions of the greatest variance in the dataset (197). In this ordination plot, the first PC axis represents the largest variability gradient, PC2 is the second largest, and so on, until all the dataset's variability has been assessed (197). Each object can be given a new set of coordinates in the PC space, and its distribution in such a space will correspond to the similarity of the variables' scores for those objects (197).

A conceptual extension of the PCA is the principal coordinate analysis (PCoA) (200). Similarly, this is used to order objects along PC axes in an attempt to explain the variance in a dataset (197). While a PCA organises objects by analysing a correlation or covariance matrix, the PCoA can be applied to any type of distance metric (197). The use of this method has recently increased in ecology since it can employ measures of phylogenetic distance and community composition to calculate the similarity among populations (197). As a distance matrix is an input file, the PCoA cannot directly relate any of the measured variables to individual coordinate axes (199). Instead, an indirect correlation or regression analysis of object values *vs.* object scores can be used to estimate the contribution of a variable to object dispersion along a particular PC axis (197).

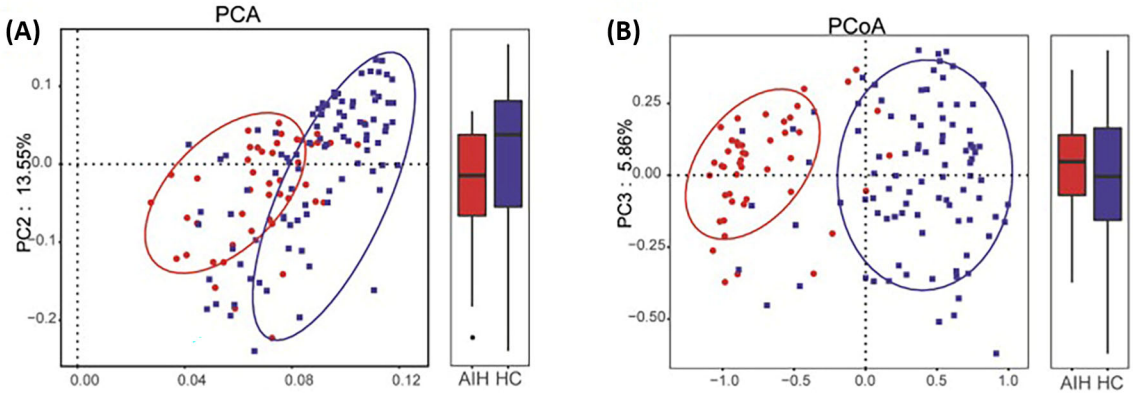



**Figure 22. Graphical representation of (A) a principal component analysis (PCA) and (B) a principal coordinate analysis (PCoA) plot. This image is adapted from Rao et al. (201), which is an open-access article distributed under a Creative Commons Attribution 4.0 International (CC BY 4.0) licence (https://creativecommons.org/licenses/by/4.0/).**

Finally, non-metric multidimensional scaling (NMDS) is another exploratory method in which a number of ordination axes are explicitly chosen in advance, after which data are fitted to those dimensions (202). As in the PCoA, a matrix of object dissimilarities is first calculated using a distance metric. The ranks of these distances for all the objects are calculated and then the algorithm identifies a configuration of objects in the N-dimensional ordination space that best matches the differences in ranks (197). In an NMDS ordination, the proximity between objects corresponds to their similarity, but the ordination distances do not correspond to the original distances between the objects (199).

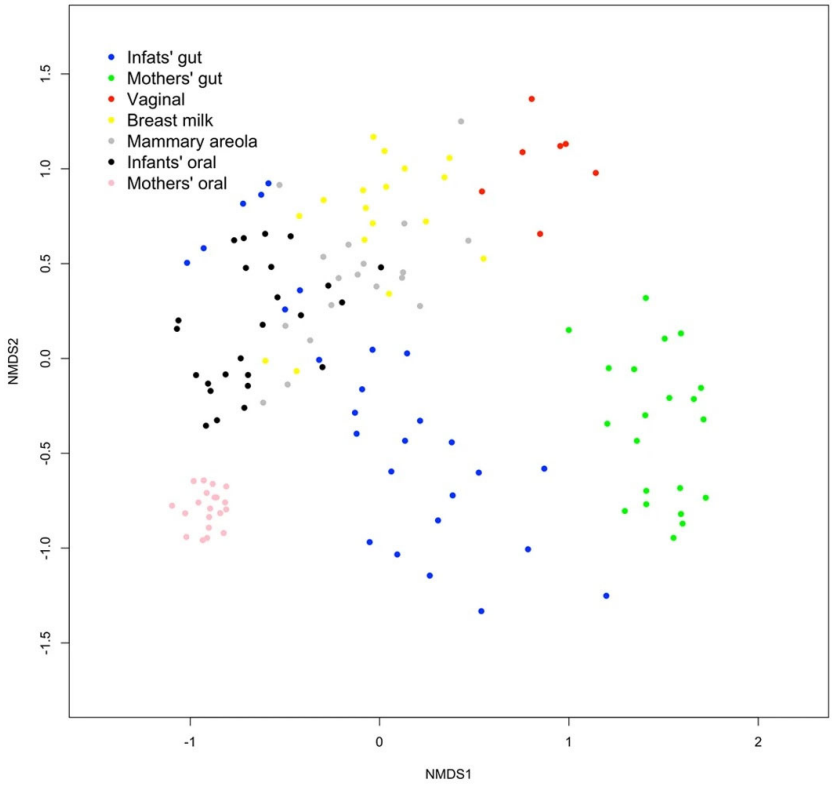



**Figure 23. Graphical representation of a non-metric multidimensional scaling (NMDS) plot. The image was taken from Drell et al. (203), an open-access article distributed under a Creative Commons Attribution 4.0 International (CC BY 4.0) license (https://creativecommons.org/licenses/by/4.0/).**

Interpretative methods for analysing large-scale datasets can be further subdivided into three types: symmetric, asymmetric, and statistical-significance testing. The first compares two datasets and does not distinguish between explanatory and response variables (197). Examples are the canonical correlation analysis (204) and the procrustes analysis (205). In contrast, the asymmetric approaches use two distinct sets of variables: one explanatory or independent and one response or dependent (197). The redundancy analysis (RDA) (206) and generalised linear models (207) are examples of asymmetric techniques. Specifically, the commonly used RDA is a type of constrained ordination that evaluates how much of the variation in one set of variables (response) can be explained by the variation in another set (explanatory) (206). This is a canonical version of the PCA and is based on similar principles, with the PCs constrained as linear combinations of the explanatory variables (197). The RDA provides a useful indication of, for example, how much the variation in species distribution is due to differences in the environmental factors between sites (197).

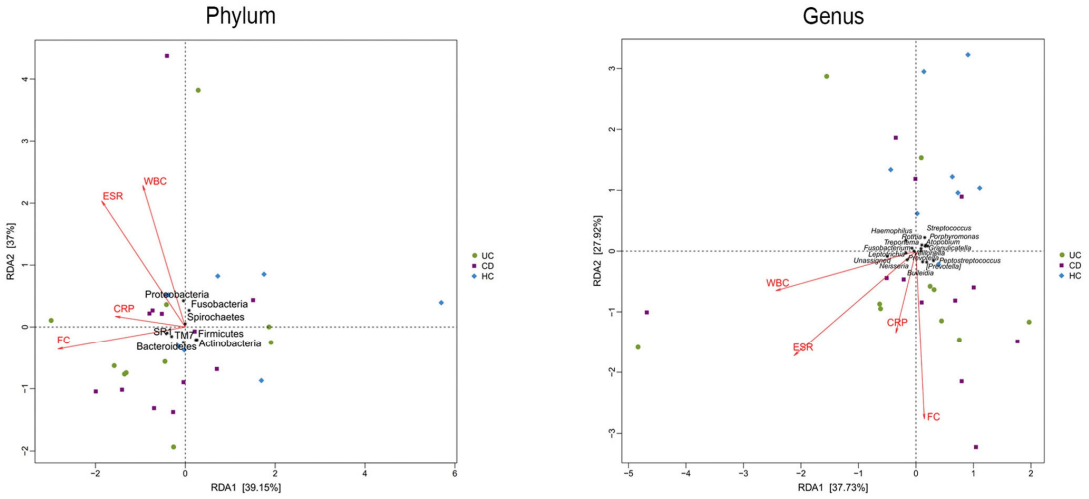


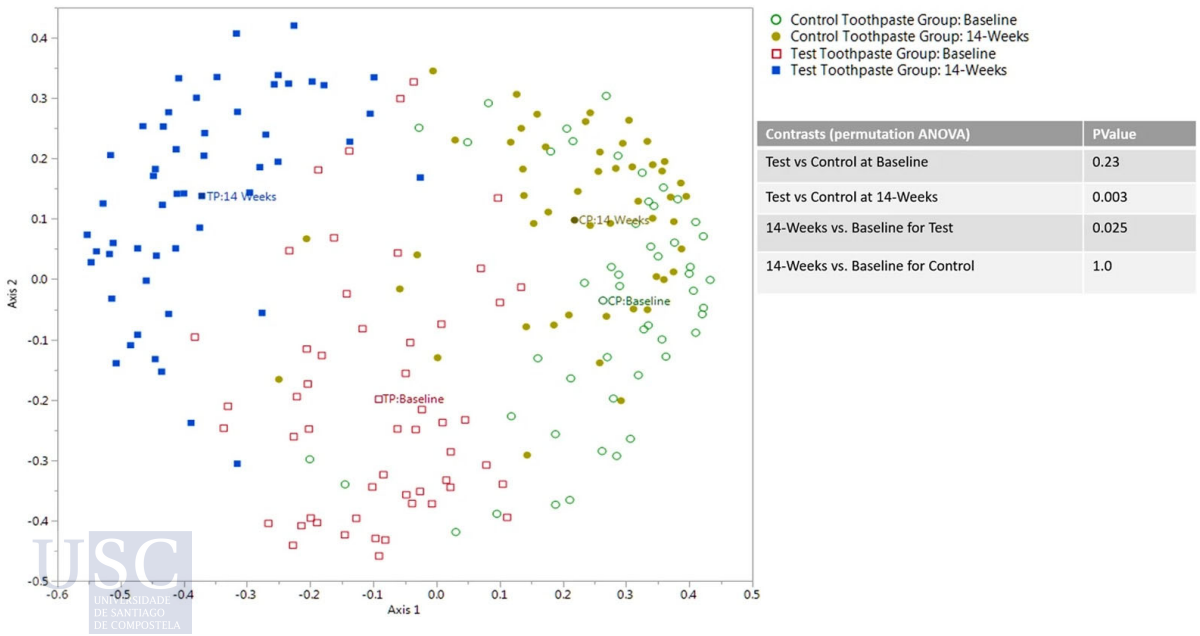

**Figure 24. Graphical representation of a redundancy analysis (RDA) plot. The image was taken from Qi et al.(208) with the permission of Elsevier.**

The third interpretative method involves the statistical-significance testing of multivariate datasets (197). Several approaches are available for analysing among-group differences in microbiota data, such as the permutational multivariate analysis of variance (PERMANOVA) (209), the analysis of similarities (ANOSIM) (210), the multi-response permutation procedure (MRPP) (211), and the Mantel test (212). Of these, the PERMANOVA and ANOSIM are the most widely used in microbiota studies and are generally employed with a distance measure (196). These tests make it possible to evaluate elements like microbial divergence or similarity in populations or the factors affecting such communities. The significance of the results can also be confirmed through visualisation methods.

Of the final examples of discriminatory methods, the discriminant function analysis (DFA) (213) and the random forest (214) should be highlighted. The DFA, better known as the LDA, is a method for evaluating how well a group of variables supports an *a priori* grouping of objects. Here, the measured variables are the predictor variables, while the variable defining the object classes is treated as the response variable (also called the grouping variable) (197). The LDA is closely related to other lineal methods like the PCA. However, unlike the PCA, it derives synthetic variables that specifically maximise the between-class group dispersion (197). As each discriminant function is a weighted linear combination of the measured predictor variables, the weights (called discriminant coefficients) can be used to define the contribution of each predictor variable to the observed discrimination between classes of objects (197). The

results of the LDA can be visualised in a scatter plot, where the axes are the discriminant functions (199).

Conversely, the random forest is an ensemble-learning approach based on the use of decision (classification) trees. Decision-tree learning seeks to construct a statistical model to predict the values of response variables based on the values given to predictor variables (197). The model is produced by iteratively partitioning the space of the predictor variables and establishing a value for the response variable within each partition (215). The results can be represented as a decision tree, which contains a set of "if-then" logical conditions (197). Many different classification trees are obtained for the same dataset. To classify a new object, the inputs (values for all the predictor variables) are assigned to each tree, which then generates an output classification or vote (197). Finally, the technique selects the classifications with the most votes among the trees in the forest (197). Although some individual trees may have a low classification accuracy, the voting step consolidates decisions across thousands of individual trees that, taken together, produce a very accurate overall classification score (197). The results of a random forest analysis can also be visualised using an MDS scatter plot of a matrix of proximities among subjects (197).




| Contrasts (permutation ANOVA) | PValue |
|---|---|
| Test vs Control at Baseline | 0.23 |
| Test vs Control at 14-Weeks | 0.003 |
| 14-Weeks vs. Baseline for Test | 0.025 |
| 14-Weeks vs. Baseline for Control | 1.0 |

**Figure 25. Graphical representation of the random forest (RF) analysis results in a scatter plot. The image was taken from Adams et al. (216), an open-access article distributed under a Creative Commons Attribution 4.0 International (CC BY 4.0) license (https://creativecommons.org/licenses/by/4.0/).**

### *B. Univariant: analysis of differential abundances*

Sometimes, it is not enough to simply determine how contextual data interact with microbiota at the community level. Indeed, it may also be important to identify which organisms contribute the most to community differences (178). The univariant analysis enables the calculation of the differential abundances of each taxa forming the populations of the distinct groups. The term "differential abundance" was coined as a direct analogy of the term "differential expression", which is used in RNA sequencing (164). Essentially, a taxon is considered to be differentially abundant if its mean proportion is significantly different in two or more sample classes in the experimental design (181).

There are three main problems from a mathematical perspective when attempting to identify differentially abundant taxa. First, the variance of each taxon is not independent of its measured value (heteroskedasticity). Second, most taxa are only present in numbers below the detection limit in the majority of samples (0-inflation or sparsity). Finally, if a normalisation procedure is conducted, the observed value for each taxon in a specimen depends on those of the other taxa in the sample (non-independence) (178). Moreover, distinct statistical tests perform quite differently in cases close to the detection limit (178).

In 2009, White et al. (164) released Metastats, a statistical method for comparing samples from two treatment populations that uses count data to detect differentially abundant features. This approach employs the false discovery rate (FDR) to improve specificity in high-complexity environments and, separately, uses the Fisher exact test to manage sparsely sampled features (164).

Other tools initially developed for RNA-seq analyses can also be utilised for microbiota investigations (178). An example is edgeR (217), which explicitly models the underlying distribution of each feature (e.g., gene or OTU/ASV) as a negative binomial distribution. This is achieved using an empirical Bayes procedure and by conditioning the variability in each OTU/ASV's abundance. Several other tools have been developed since edgeR's launch in 2010. These generally use similar procedures to model the distribution of each taxon, with DESeq2 (170) being one of the most popular (178). This method assumes that taxa with similar abundances will have similar dispersions, although such a supposition is over-ruled when the

observed variance is more than twice the mean variance (178). DESeq2 also considers noise levels to be higher when counts are low and is more aggressive in its approach to shrinkage variations in low-abundance taxa (178). A Wald test and multiple-testing correction are performed via the Benjamini-Hochberg methodology to evaluate the significance of differentially abundant taxa, but the FDR's accuracy is reduced by the previous removal of those that are only present in low numbers (178). DESeq2 also includes a tool to make the variability of each taxon independent of its mean (178).

In 2013, metagenomeSeq was released to assess differential abundances in the survey data on high-throughput microbial marker genes (218). This Bioconductor package has two novel aspects (178). First, it uses a percentile cut-off point instead of a process of normalising counts. This point is chosen automatically by selecting the highest percentile after which there is major instability between expected and observed values, which is suggestive of PCR biases. Second, it uses a zero-inflated Gaussian distribution to model data, although a subsequent study by McMurdie and Holmes (181) demonstrated that this produces a higher rate of false positives than negative binomial-based approaches. Consequently, edgeR (217) and DESeq2 (170) were recommended as best practices (181).

Conversely, the LEfSe method mentioned previously (165) uses a different approach to distinguish differentially abundant taxa. In particular, there is the first round of feature selection using the Kruskal-Wallis sum-rank test, which identifies taxa with differential abundances between conditions. Then, to remove spurious correlations, the pairwise Wilcoxon test is employed to discard the taxa present at inconsistent levels across sub-conditions (178). Finally, the LDA is employed to estimate the effect size of each taxa's differential abundance. This is an important step in the discovery of biomarkers since even a very significant marker is unlikely to be the driver of phenotypical changes if its effect size is too small (178).

#### *I.5.4.2.3. Core microbiota*

In our view, the analysis of core microbiota should also be highlighted when discussing the profiling of community diversity. As defined by Shade et al. (219), the "core" is typically described as the suite of members shared among microbial consortia from similar habitats. It is usually reported based on presence/absence datasets and visualised via a Venn diagram (219).

However, the definitions in the literature are heterogeneous and this step can also be performed based on the (219):

- Shared abundance.
- Shared composition: a combination of presence/absence and relative abundance.
- Incorporation of phylogenetic information: related taxa are counted towards a core as a single unit.
- Interaction: this only includes taxa that interact with the other community members (i.e., using network analysis).

The diverse meanings attributed to the concept of a "core" make associated findings difficult to compare. Nevertheless, it is important to describe the core microorganisms of a community to understand the stable, consistent components across complex microbial assemblages. This will enable researchers to predict the impact of global changes on biochemical cycling and make recommendations about how the human microbiota should be managed to improve human wellbeing (219).

Several metrics related to the concept of the core, including core_abundance, core_heatmap, core_matrix, and core_members, can be calculated using the microbiome R package (173).

#### *I.5.4.2.4. Co-occurrence network analysis*

The structure and functioning of complex microbial communities are heavily influenced by organism-organism and organism-environment interactions (220). However, despite the value of the diversity measures described above, they are unable to identify such interactions. Consequently, several analytical procedures have been developed to improve what is known about how microorganisms potentially cooperate in their environment (221). Specifically, microbial network analyses have been used to visualise the co-occurrence patterns among the members of communities (222). These networks permit the examination of more than the composition of microbial communities, also enabling the following: 1) the detection of "keystone species" (which will be explained later); 2) the identification of group dynamics; and 3) analyses of the effect of abiotic factors on the community (223).

A wide range of methods and algorithms are available to construct microbial networks. The simplest are the (dis)similarity- or distance-based techniques, but correlation-based methods, which detect significant pairwise associations between OTUs or ASVs using correlation coefficients, are the most popular (220). Nevertheless, the latter has limitations, as the detection of spurious correlations is possible with this methodology due to compositionality (220). This has led to the development of more robust techniques, including the sparse correlations for compositional data (SparCC) (224) and the sparse inverse covariance estimation for ecological association inference (SpiecEasi) (225). SparCC uses linear Pearson correlations between log-transformed components to infer associations in compositional datasets, and is particularly suitable for data that is compositionally diverse (224). In contrast, the SpiecEasi combines data transformations developed for compositional analyses with a graphical model inference framework that assumes the underlying association network is sparse (225). The SpiecEasi package (226) can be used to run either the *spiec.easi* or the *sparcc* function (220). In addition, there are two open-source and free network-analysis tools - igraph (227) and qgraph (228) - that can be employed in R to construct, simulate, analyse, and visualise networks (220).

Regardless of the method used for its construction, a co-occurrence network consists of: "nodes" or "vertexes", each of which represents an OTU or ASV; and "edges", which represent a relationship between the two connected OTUs or ASVs (223). This correlation can be either positive, indicating a direct or indirect relationship between taxa, or negative, suggesting a competitive interaction or that the taxa do not share a niche (223).

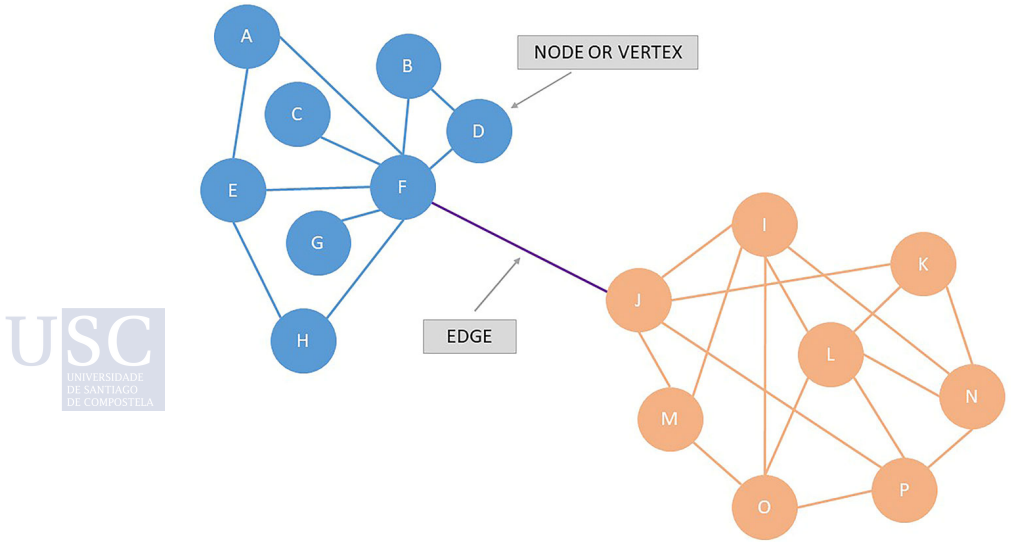



**Figure 26. Example of the principal components of a co-occurrence network.**

Different measures can be calculated to describe how nodes are connected to each other and to a network. It is also possible to characterise the structure of a network as a whole (next paragraph) (229). The node degree is perhaps the simplest measurement and represents the number of edges connected to a particular node (229). Another important metric is node centrality, which evaluates how central a vertex is in a network. There are many ways to calculate this: degree centrality (DC), closeness centrality (CC), betweenness centrality (BC), and eigenvector centrality (EC) (229). The DC of a node is equal to its degree. The CC is more interesting and is calculated as the average of the shortest path length from the node to every other node in the network. Consequently, this highlights how close a vertex is to all other network vertices and strongly corresponds to the visual centrality (229). The nodes visited more frequently have a high CC (223). Conversely, the BC calculates how important a node is to the shortest paths through the network (229). This is calculated as the total number of these paths between all the nodes passing through the one under consideration. Nodes with a high BC value are those that connect groups of nodes that support the network (223). Finally, the EC estimates the importance of a node by examining the value of its neighbours. The principle behind this measure states that links from significant vertices (determined by DC) are more important than links from unimportant ones (229). Since each measure defines nodes in different ways, more than one can be described to achieve a better perspective on the network (229).

The degree distribution can also be calculated to provide some idea of the degrees of all the nodes in a network. This measure reveals how many nodes have each possible degree and the results are usually presented in a bar graph (229). Researchers are also able to determine the minimum number of vertices that must be removed before the network becomes disconnected, which is achieved by estimating the connectivity or cohesion (229). One of the most common ways of describing a network, however, is to evaluate its density. This metric divides the number of existing edges portrayed in a graph by the maximum number of edges that might exist. Density values range between 0 (the lowest possible value) and 1 (the highest) (229).



Centralisation is another important tool and uses the distribution of a node-centrality measure (e.g., the BC) to understand the network as a whole. This is calculated as the sum of the differences in centrality between the most central node and all the others presented in the graph. The resulting value is then divided by the maximum possible difference in centrality in

the network (229). Centralisation is high when a vertex has high centrality values and those of the other vertices are low. Conversely, if the centrality is distributed more evenly, the network centralisation is low.

Additionally, groups of vertices may form a module or cluster, thus acting as a sub-network within the main network. A module is defined as a set of strongly related nodes that are, in turn, less related to those that do not belong to the group (223). A network is said to have high modularity if it presents dense connections within node clusters and sparse connections between different groups of vertices (223). All of the network measures described above can be calculated for the modules or clusters.

Finally, as referred to earlier, the capacity to identify hubs or keystone taxa, which are highly connected OTUs or ASVs in the microbiota, is one of the most useful features of a co-occurrence network analysis (230). Different measures have been adopted to define these hubs in microbial communities. Banerjee et al. (222), for example, aimed to provide a quantifiable threshold for the consistent identification and validation of keystone taxa. Their findings led to a recommendation that the high mean degree, high CC, and low BC scores should be combined to this end. Nevertheless, it should be noted that the identification of highly connected OTUs or ASVs in a microbial network does not necessarily reveal their role as keystone taxa (231), which are very closely linked species that exert a considerable influence on the structure and functioning of the microbiota, irrespective of its abundance (222). Although further experimental evidence is required before network hubs can be defined as keystone taxa (231,232), identifying them is nonetheless a valuable step, since this will help researchers to target key community members (231).

In conclusion, the findings from the co-occurrence network analyses described in the literature should be viewed with caution: they may be affected by methodological differences concerning, for example, the correlation values employed as cut-off points (233) or the use of different definitions of keystone taxa (222).

*I.5.4.2.5. Predictive models*

Machine learning (ML) is a computer science discipline in which computers are programmed to learn patterns from the data in a multi-dimensional dataset and produce classifications or predictions based on statistical associations (234). The field has two main approaches, supervised and unsupervised, and the goals of the research determine which is the most appropriate (Table 5). The latter is employed to identify the underlying structures or relationships between variables (samples) in a dataset and is well suited to the visualisation of high-dimensional input data (234,235). Indeed, the PCA (198) and PCoA (200) mentioned earlier are examples of unsupervised ML algorithms. In contrast, supervised learning involves the classification of an observation into one or more categories or outcomes (235). As a consequence, this requires training data, with each training sample having values for a number of independent variables or features, as well as an associated classification label (236).

Predictive modelling is a set of mathematical processes and computational techniques that enable the probability that an event will occur to be inferred from a set of previously obtained data. There is a close relationship between predictive analyses and ML, since predictive models typically include an ML algorithm. Unsupervised models do not require labelled data or the use of error-measurement metrics, with the algorithm instead searching for patterns among the input or the output variables (samples) during the modelling process. Conversely, a supervised model needs such data to generate a predictive model, as well as an error-measurement metric to improve it during the building process.

**Table 5. Summary of the supervised and unsupervised machine-learning approaches. The table was taken from Reel et al. (237) with the permission of Elsevier.**

| Learning approach | Goal | Description |
|---|---|---|
| Unsupervised | Identify clusters | Unsupervised learning employs input variables without a target/output variable to find the underlying patterns in unlabelled data. It can be used for clustering, anomaly detection and dimensionality reduction. |
| Supervised | Predict new data | Supervised learning involves fitting a model with labelled training data and then using it for predictive purposes. The problem it addresses can be classed in terms of either regression (the predicted variable is numeric) or classification (the predicted variable is categorical).<br><br>The three steps of supervised learning are: 1) fitting a model from the sample's input observations; 2) evaluating the model and then extensively tuning its hyper-parameters; and 3) setting up the model for the production stage and using it to make predictions.<br><br>When a model is accurate in relation to both the training and test data, it is said to have learned properly. However, a particular ML output might predict the training data with a high degree of accuracy, but nonetheless fails to produce precise predictions with the test data (overfitting); it may even be unable to predict the training data correctly (underfitting) (234). |

In comparison to the very low number of samples and clinical conditions typically evaluated in this kind of work, the oral microbiota high-throughput data in this study is characterised by a large quantity of independent and predictor variables (OTUs or ASVs). This often indicates a high degree of multicollinearity and, as a result, produces very poorly conditioned problems (238). In a supervised framework, one solution is to reduce the dimensionality of the data, either by feature selection or introducing artificial variables that summarise most of the relevant information (238). To this end, several tools for supervised predictive modelling have been proposed, including the aforementioned random forest plots (214), support vector machine (SVM) (239), and regression models like the sparse partial least-squares discriminant analysis (sPLS-DA) (238).

The SVM device is a method for identifying a decision boundary to enable the classification of data. An SVM training algorithm is applied to a training dataset with information about the class to which each piece of data belongs, establishing a hyperplane that separates two classes. Next, the SVM seeks to optimise the width of the gaps between classes, i.e., the maximum-margin hyperplane. The resulting model can be used to determine whether a new data element is, or is not, a member of a particular class (240). Among the advantages of

this method are its efficiency at learning complex classification functions and its employment of powerful regularisation principles to prevent overfitting (241). However, in the case of highly dimensional datasets, the results obtained are often difficult to interpret given the large number of variables (238). Furthermore, multiclass classification problems require either their decomposition into several binary problems or the definition of multiclass objective functions (238).

In contrast, the sPLS-DA (238), which is a natural extension of the PLS-DA, is based on the assumption that only a small number of features are responsible for driving a biological event or effect, enabling predictor variables to be selected and classified in a one-step procedure (238,242). The proper functioning of this method has been demonstrated previously (243), and multiple classes (e.g., clinical conditions) can be distinguished at the same time. Although there is some difficulty in construing the results obtained with this model compared to those that have only two classes, a graphical representation makes the interpretation process easier (238).

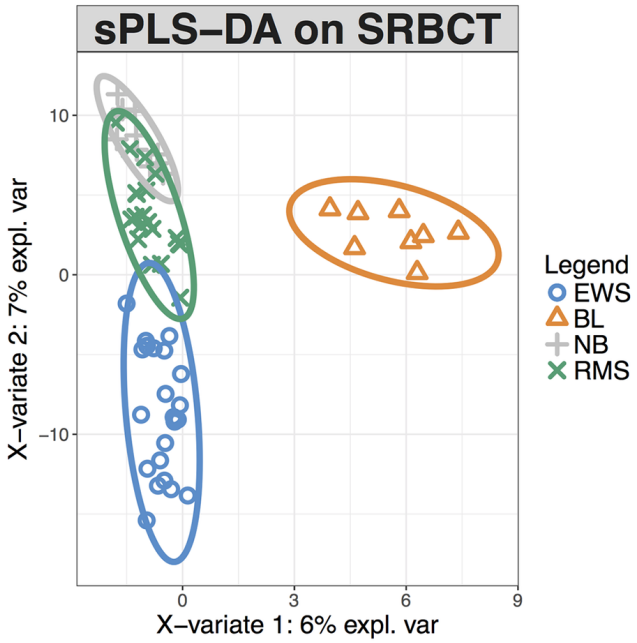



**Figure 27. Graphical representation of an sparse partial least-squares discriminant analysis (sPLS-DA). This image is adapted from Rohart et al. (244), which is an open-access article distributed under a Creative Commons Attribution 4.0 International (CC BY 4.0) licence (https://creativecommons.org/licenses/by/4.0/).**

The implementation of an sPLS-DA (238) in the mixOmics package (244) of R-Bioconductor (245) enables the following to be determined for each model (246):

- The number of components or latent variables. There are as many dimensions of the sPLS-DA model as required.
- A set of loading vectors, which are coefficients assigned to each independent variable (OTU or ASV) to define each component. In other words, they indicate the importance of each variable in the sPLS-DA and each loading vector is associated with a particular component. These vectors are obtained in such a way as to maximise the covariance between a linear combination of independent variables and the classes of interest.
- A list of designated variables associated with each component. The procedure implemented in the mixOmics package (244) uses a k-fold cross-validation technique to automatically calculate the appropriate number of dimensions for each model.
  The rule of thumb is that the number of dimensions will be K-1, where K indicates the number of classes (e.g., clinical conditions) included in each model. Accordingly, a variable (e.g., OTU or ASV) that is present in multiple models will be identified as important. It becomes critical or sees its value increase if its presence is always associated with the same class.
- Once the optimal parameters have been chosen (number of components and variables), the final model is run on the entire dataset. The model's classification error rate is then estimated (244), and an additional accuracy evaluation using the receiver operating characteristic (ROC) and AUC can be performed (244).

The use of predictive modelling to identify oral taxa that can distinguish between health conditions and are associated with specific disease states would be extremely valuable for determining the biomarkers of disease (236). However, to date, very few sequencing-based studies of the oral cavity have conducted predictivity analyses (247-250).

## I.6. AN OWN STUDY ON THE RELATIONSHIP BETWEEN DENTAL AND PERIODONTAL HEALTH STATUS AND THE SALIVARY MICROBIOTA

In a recently published 16S rRNA gene sequencing-based investigation of our investigation group (249), several of the above explained analysis tools were applied to examine the bacterial diversity and the co-occurrence network patterns of the salivary microbiota in patients who have been clinically classified by a self-designed and previously validated scale of overall oral health. Moreover, we evaluated the diagnostic potential of the salivary microbiota to discriminate between different clinical conditions.

The own scale of overall oral health consists of three dental- and three periodontal-associated parameters (Table 6). Taking this into account, the participants' dental and periodontal grades (DG and PG, respectively) corresponded to the grades assigned to at least two of the three variables analysed in each of these two categories. If there were differences between the grades allocated to each of the variables in a category, the parameters for "number of caries" and "number of periodontal pockets ≥4 mm" took precedence. If the same grade was allocated to two variables in a category, but the third variable's grade was two levels higher, the value assigned to it was one grade higher than that of the matching variables. Lastly, the oral health grade (oral grade, OG) was determined by the category (dental or periodontal) with the highest ranking, enabling patients to be classified based on the score for their dental and periodontal health and the combination of both conditions.

**Table 6. The scale of overall oral health, involving grades of dental and periodontal health. The table was taken from Relvas et al. (249), an open-access article distributed under a Creative Commons Attribution 4.0 International (CC BY 4.0) license (https://creativecommons.org/licenses/by/4.0/).**

Caries severity, 1= affecting the enamel, 2= affecting the enamel and dentine, and 3= affecting the enamel, dentine, and pulp.

| | Grade 0 | Grade 1 | Grade 2 | Grade 3 |
|---|---|---|---|---|
| **Grades of dental health** | | | | |
| Supragingival plaque (O'Leary index) (251) | 0 | 1–56 | 57–112 | >112 |
| Caries | 0 | 1–4 | 5–8 | ≥9 |
| Severity of the caries (median) | 0 | 1 | 2 | 3 |
| **Grades of periodontal health** | | | | |
| Gingival inflammation (Ainamo and Bay index) (252) | 0 | 1–56 | 57–112 | >112 |
| Periodontal pockets ≥4 mm | 0 | 1–56 | 57–112 | >112 |
| Severity of the pockets (mean) | <4mm | 4-4.9mm | 5–5.9mm | ≥6mm |

Unstimulated saliva samples were collected from each participant. Sequencing of the 3-4 region was performed in an Illumina MiSeq platform with 2 × 300 bps reads, while the raw reads were processed according to the mothur pipeline (118). The statistical analysis of the 16S rRNA sequencing data at the species level was conducted using the phyloseq (166), DESeq2 (170), Microbiome (173), SpiecEasi (226), igraph (227), and mixOmics (238) packages.

### I.6.1. Results and discussion

The overall oral health scale was used to produce a convenience sample of 81 patients that were given the following OGs: 0 for 17 of them; 1 for 25; 2 for 28; and 3 for 11. In relation to the subscales, 47 patients had a PG of 0 and different DGs (17 had a DG of 0, nine a DG of 1, 11 a DG of 2, and 10 a DG of 3), and 46 had a DG of 0 and different PGs (17 had a PG of 0, 14 a PG of 1; 14 a PG of 2; and one a PG of 3). The four patients excluded due to a low number of raw sequences obtained were: two of OG1 and two of OG2 (two of DG0, one of DG1, and one of DG2; two of PG0, one of PG1, and one of PG2).

#### I.6.1.1. Impact of the dental and periodontal subscales and the scale of overall oral health on the salivary microbiota: alpha diversity indicators and the structure of the bacterial community

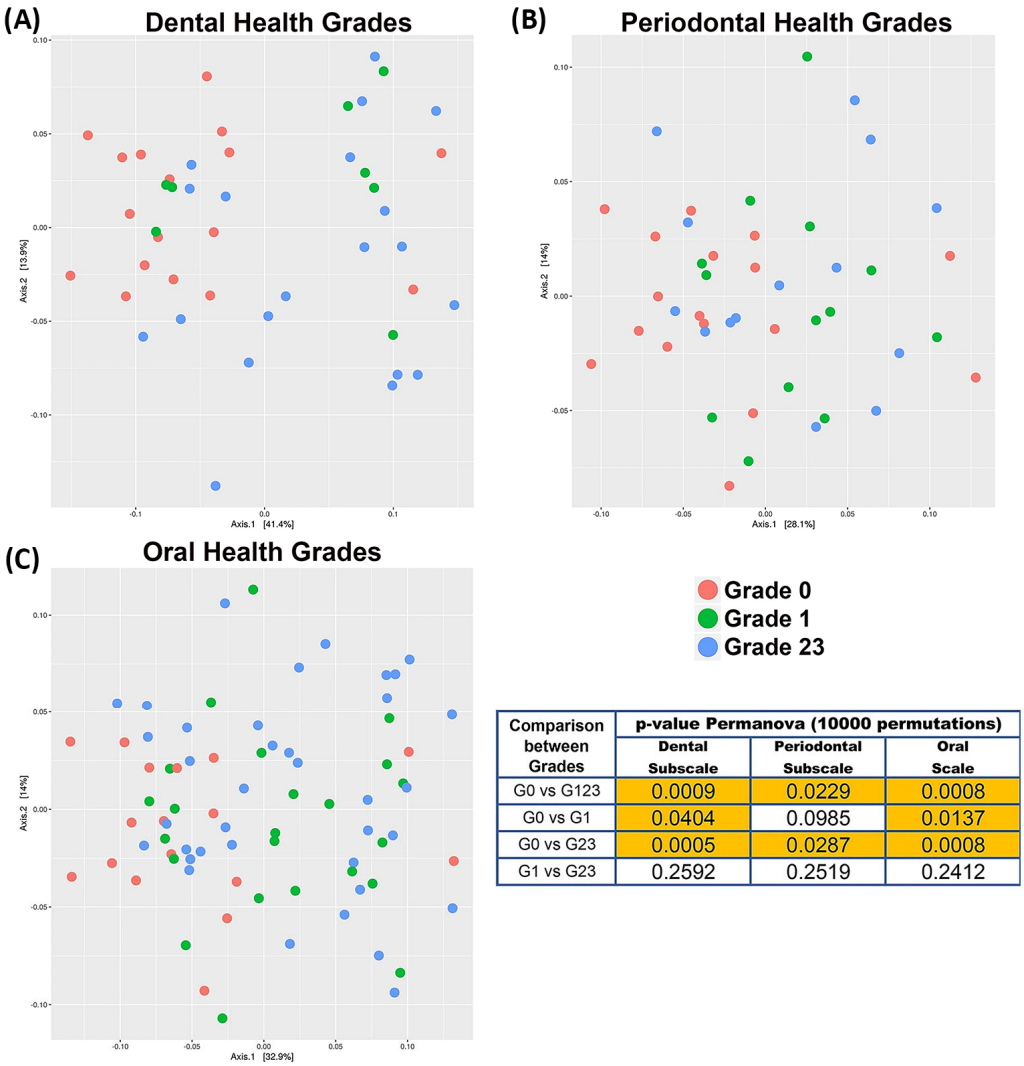
Worsening dental or periodontal health revealed a trend of increasing alpha diversity in both the dental and periodontal subscales; although significant differences in the number of OTUs were only observed in the former: DG0 *vs.* DG123 ($p = 0.009$), and DG0 *vs.* DG23 ($p = 0.006$). The DG23 also showed a trend towards increased diversity (Shannon Index) and evenness (Pielou Index).

On the contrary, other authors had observed that the saliva samples from the healthy and caries groups generally had similar levels of richness and diversity (253-256). This was true whether the diseased group was composed of subjects with active (254), inactive (255), or cavitated (256) caries. However, it should be noted that the dental health subscale used here incorporates variables that not only include the number of caries and their severity, but also the levels of supragingival plaque, all of which could affect the bacterial richness of the salivary community. In addition, there is some inconsistency between the alpha-diversity results of various studies in the literature and our findings when comparing periodontally healthy and periodontitis subjects. Several researchers described greater richness (248,257), diversity (248),

and evenness (257) in the saliva samples of patients with periodontitis than in those who were healthy; meanwhile, others, as observed here, only observed a trend of increased alpha diversity with worsening of periodontal health (258).

In the overall oral health scale comparisons, grade increments were linked to progressive increases in bacterial richness and the Shannon Index values, especially in OG0 *vs.* OG23 ($p= 0.013$ and $p= 0.026$, respectively). The potential impact of the simultaneous presence of dental and periodontal disease on the richness and diversity of the salivary microbiota is in line with the results of Takeshita et al. (259).

Like other studies in which the structure of the global salivary microbiota is similar in patients with good oral health (260-262), our PCoA revealed a grouping of the salivary samples taken from the participants with DGs, PGs, and OGs of 0. This contrasted with the picture for the other grades, whose compositional distributions were more diverse (Figure 28). The visual observation was confirmed by the PERMANOVA test, which produced significant results for the comparison of grades 0 and 123 (dental subscale, $p= 0.0009$; periodontal subscale, $p= 0.0229$; oral scale, $p= 0.0008$). These findings were mainly at the expense of the contrast between grades 0 and 23 (dental subscale, $p= 0.0005$; periodontal subscale, $p= 0.0287$; oral scale, $p= 0.0008$). Focusing on the group with the highest grade of oral pathology (OG23), PERMANOVA's test revealed that the structure of the salivary microbiota was different depending on the predominance of dental pathology (PG0_DG23) or periodontal pathology (DG0_PG23) ($p= 0.027$).




| Comparison between Grades | p-value Permanova (10000 permutations) | | |
|---|---|---|---|
| | Dental Subscale | Periodontal Subscale | Oral Scale |
| G0 vs G123 | 0.0009 | 0.0229 | 0.0008 |
| G0 vs G1 | 0.0404 | 0.0985 | 0.0137 |
| G0 vs G23 | 0.0005 | 0.0287 | 0.0008 |
| G1 vs G23 | 0.2592 | 0.2519 | 0.2412 |

**Figure 28. Principal Coordinate Analysis including PERMANOVA test values in the comparison between different grades of dental, periodontal, and oral health. The image was taken from Relvas et al. (249), an open-access article distributed under a Creative Commons Attribution 4.0 International (CC BY 4.0) license (https://creativecommons.org/licenses/by/4.0/).**

I.6.1.2. Impact of the dental and periodontal subscales and the scale of overall oral health on the salivary microbiota: composition of the core microbiota and testing differential abundance

There are numerous 16S rRNA-based microbiome studies on salivary microbiota in the literature that have only analysed the differential abundance of the taxa associated with various oral conditions (256,257,263,264). In our study, we considered that it was essential to evaluate salivary microbiota from a dual perspective: the prevalence of the taxa determining the core microbiota; and their differential abundance in relation to the different DGs, PGs, and OGs.

*I.6.1.2.1. Composition of the core microbiota*

The core microbiota associated with the participants' dental and periodontal health contained 57 species, representing 14.14% of the total number of OTUs and 63.06% of the total abundance. There were only nine taxa in DG0 and eight in PG0 (the specific core of grade 0), exemplifying abundances of 7.80% and 1.34%, respectively. Of these specific core species, five were common to both the dental and periodontal health conditions: *Neisseria macacae*, *Butyrivibrio* sp. HMT 455, *Campylobacter concisus*, *Porphyromonas catoniae,* and *Corynebacterium durum*.

There were 66 species in the core microbiota associated with the most severe dental disease (DG23) and 73 with the most serious periodontal disease (PG23), representing 16.37% and 18.11% of the total salivary microbiota, and 67.14% and 67.98% of the total abundance, respectively. There were only eight and 10 taxa present in DG23 and PG23 (the specific core of grade 23), exemplifying 2.54% and 2.46% of the abundance, respectively. Of these specific core species, only *P. endodontalis* was common to both pathological conditions.

There were 35 taxa common to both the dental and periodontal subscales, regardless of the grade (non-specific core), representing abundances of 50.83% and 52.75%, respectively. Of these non-specific core species, 25 were common to both subscales, with the most abundant (abundance > 1%) being: *Granulicatella adiacens*, *Haemophilus parainfluenzae*, *Leptotrichia* sp., *Porphyromonas pasteri*, *Prevotella* sp., *Prevotella melaninogenica*, *Prevotella salivae*, *Rothia mucilaginosa*, *Streptococcus* sp., *S. oralis* subsp. *dentisani* clade 058, *Streptococcus salivarius*, *Veillonella* sp., and *V. parvula*.

The core species in the literature on the salivary microbiome have various definitions and, as a consequence, any associated findings are difficult to compare (247,254,258,259). Despite this, our investigation demonstrated for the first time that the non-specific core of the salivary microbiota comprises a greater number of species in higher abundances than the specific-core associated with a particular dental or periodontal condition. Interestingly, more than half of the non-specific core species in the present series were the same as those previously identified by Takeshita in ≥ 75% of Japanese adults (259) as *G. adiacens*, *H. parainfluenzae*, or *R. mucilaginosa*, among others. These results confirm that several bacterial taxa in the salivary

microbiota could be present in ethnically diverse populations, regardless of their dental and periodontal health statuses.

*I.6.1.2.2. Testing differential abundance*

The results for the dental health subscale revealed differential abundances for the different grades in 102 species (25.31% of the total OTUs), 39 of which were core species and 63 non-core (9.67% and 15.63% of the total OTUs, respectively). In the periodontal health subscale, there were differential abundances for the different grades in 27 species (6.69% of the total OTUs), eight of which were core species and 19 non-core (1.98% and 4.71% of the total OTUs, respectively). Lastly, in the overall oral health scale, there were differential abundances for the different grades in 88 species (21.83% of the total OTUs), 22 of which were core species and 66 non-core (5.45% and 16.37% of the total OTUs, respectively).

As we can see, the number of taxa present in the salivary microbiota at significantly different abundances for the subscale or overall scale grades did not exceed 25%, and mainly included non-core species. Moreover, for the first time in the literature, we have reported that the such number of taxa was higher for the dental than the periodontal subscale. If the salivary microbiota comprises a mix of bacterial communities originating from various sites in the oral cavity (265), our observations provided evidence that the relative abundances of the most predominant bacteria in saliva are not strongly related to the grade of overall oral health.

Nevertheless, in the present series, specific bacteria were involved in different oral conditions. There were 36 species associated with dental health and 66 with some grade of dental pathology; 12 species were associated with periodontal health and 15 with some grade of periodontal pathology; and 22 species were associated with overall oral health and 66 with some grade of oral pathology.

In the comparison of DG0 *vs.* DG23, there were differential abundances in the main dental-health related OTUs (with >1% levels in DG0): *P. pasteri*, *F. periodonticum*, *V. parvula*, *Alloprevotella* sp. HMT 473, *Alloprevotella tannerae*, and *Neisseria subflava*. Comparing PG0 *vs*. PG23, there were differential abundances in the two main periodontal-health related OTUs (with >1% levels in PG0): *H. parainfluenzae* and *Capnocytophaga leadbetteri*. All these

bacteria continued to be associated with overall oral health (>1% levels in OG0 when compared to OG23), except for *F. periodonticum*, probably because it has been found to be more abundant in patients with periodontitis (266). These outcomes are in accordance with the findings of most 16S rRNA gene-based studies concerning the identification of the above-mentioned taxa as core species present in >70% of people (247,254,259,262). Other authors have, like us, also detected differential abundances of these species, which provide support for dental or periodontal health (248,254-256,266).

On the other hand, the taxa related to high grades of dental pathology (>0.5% levels in DG23) were: *Alloprevotella* sp. HMT 308, *Streptococcus parasanguinis* clade 411, *Atopobium* sp., *F. nucleatum* subsp. *vincentii*, *Megasphaera micronuciformis*, and *Alloprevotella rava*; some of which have been described as core species in other studies (259). The bacteria associated with high grades of periodontal pathology (with >0.1% levels in PG23) were: *T. forsythia*, *Mycoplasma faucium*, *Fretibacterium* sp., and *Bacteroidetes* [G-5] bacterium HMT 511. Of them, *T. forsythia* has been previously linked (60) and found to be more abundant in periodontitis patients (267). Finally, the taxa associated with high grades of oral pathology (>1% levels in OG23) were: *Alloprevotella* sp. HMT 308, *F. nucleatum* subsp. *vincentii*, and *P. gingivalis*. This latter bacteria was nine times more abundant for the highest oral pathology grades, revealing its role not only in periodontal but also in dental pathology, as reported in the literature (268,269).

I.6.1.3. Impact of the dental and periodontal subscales and the scale of overall oral health on the salivary microbiota: co-occurrence network patterns and discriminatory potential of the salivary microbiota.

*I.6.1.3.1. Co-occurrence network patterns*

Despite the importance of this type of analysis, we found that co-occurrence results were reported in only a few papers that evaluated the salivary microbiota of patients with and without caries, the periodontally healthy, and those with periodontitis, and adults with different oral health conditions (253,255). Similar to that observed in these investigations (253,255), the topological characteristics of our salivary microbiota networks differed between the statuses of oral health and the presence of high grades of oral or dental disease. The network coverage in OG0 was higher than in OG23 (65.75% and 43.42%, respectively), as was the number of edges

(1867 and 436, respectively). There was also a better balance between the number of positive and negative correlations in OG0 with respect to OG23 (56.5% and 43.5% and 83.1% and 16.9%, respectively). The network in OG0 had a higher density, an average number of neighbours, and higher centralization values than the network in OG23 (0.05, 15.09, and 0.16 *vs.* 0.02, 5.98, and 0.08); in contrast, the characteristic path lengths and modularity scores were lower (2.86 and 0.36 in OG0 *vs.* 4.27 and 0.62 in OG23). Also, OG0 network had fewer subnetworks and a higher number of modules (two and 28 *vs.* five and 19 in OG23). Given the disease-associated network was less dense and synergistic exchanges predominated, results suggest that the antagonism between the oral bacteria was not a major driving force in the formation of the disease-associated microbial community (257). All the above differences (except for the modularity values) were also observed between the OG0 and DG23 networks, albeit to a lesser extent. In contrast, all the other parameters had similar values in the comparison between OG0 and PG23.

Again, although one of the most useful features of a co-occurrence network analysis is that hubs or keystone OTUs can be identified; the salivary microbiota studies that included both co-occurrence network analyses and attempted to detect keystone OTUs were uncommon (257). In the OG0 network, the three main hubs or keystone OTUs, based on their combined scores in the three main modules, were *P. pasteri*, *P. endodontalis,* and *P. salivae*. All three species were part of the core microbiota, but only *P. pasteri* was present in copious numbers (relative abundance of 8.07%) and differentially abundant in relation to the oral health status. In the OG23 network, the main hubs were *F. periodonticum*, *Treponema socranskii*, and *Prevotella* sp. HTM 305. Only *F. periodonticum* was abundant (relative abundance of 2.95%) in the core microbiota, but none of the three species were differentially abundant. In the DG23 network, the main keystone taxa detected were *T. forsythia*, *F. nucleatum* subsp. *vincentii*, and *Prevotella oris*. Only *F. nucleatum* subsp. *vincentii* was part of the core microbiota and differentially abundant in relation to the dental disease status. In the PG23 network, the main hubs detected were *G. adiacens*, *P. endodontalis*, and *C. gracilis*. Of these three species, *G. adiacens* and *P. endodontalis* were abundant (relative abundances of 1.92% and 1.45%, respectively) and belonged to the core microbiota, but none of the three taxa were differentially abundant.

As we can see, seven of the 12 keystone OTUs identified (58.3%) were part of the salivary core microbiota. In this sense, it has been suggested that the contribution of keystone taxa will be greater if they are part of the core microbiota and consistently present, highlighting the importance of such taxa for microbiota functioning (270). However, eight and 10 of the 12 keystone OTUs identified (66.6% and 83.3%, respectively) were low-abundance species (<1%) with no associated differential abundance results; which confirms that keystone OTUs could have an impact on microbiota functioning, irrespective of the abundance parameters (222).

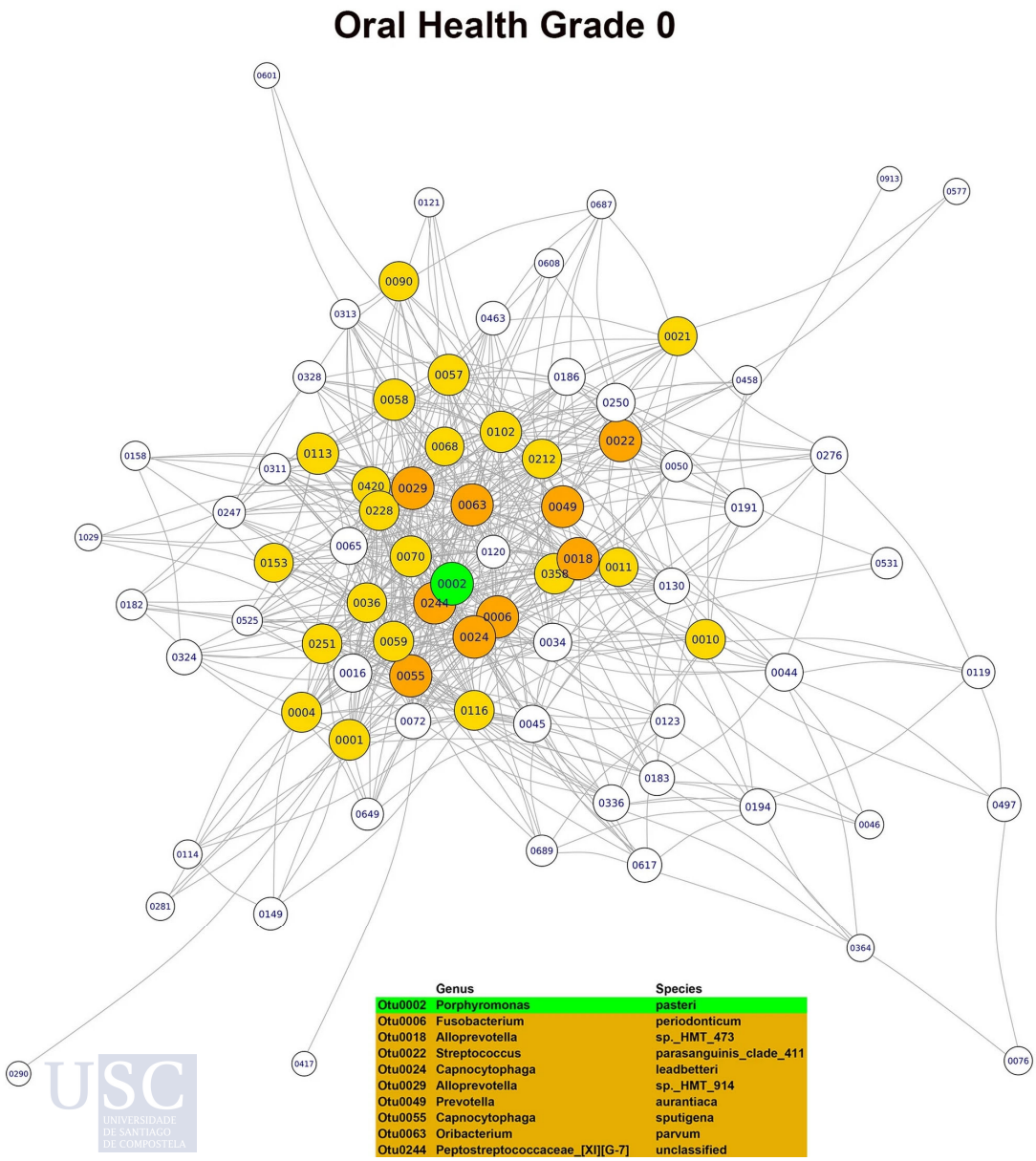



**Figure 29. Main module of the co-occurrence networks associated Grade 0 of the overall oral health scale (node = 79; degree = 1723). The image was taken from Relvas et al. (249), an open-access article distributed under a Creative Commons Attribution 4.0 International (CC BY 4.0) license (https://creativecommons.org/licenses/by/4.0/).**

*I.6.1.3.2. Discriminatory potential of the salivary microbiota*

As occurred with the co-occurrence network analysis, only a few studies attempted to evaluate the diagnostic accuracy of the salivary microbiome for diagnosing oral diseases like periodontitis (247,248). Moreover, to the best of our knowledge, our research was the first to assess the potential of the salivary microbiota for distinguishing different grades of dental and periodontal disease, or a combination thereof.

The model we obtained to distinguish the grades of the dental subscale was composed of 60 OTUs and the derived AUC values ranged from 0.93 (DG0 *vs.* others) to 0.99 (DG1 *vs.* others). In line with the differential abundance results, a higher number of predictive variables was required to discriminate the periodontal condition than the dental condition: the model to distinguish the grades of the periodontal subscale was composed of 140 OTUs and the derived AUC values ranged from 0.90 (PG0 *vs.* others) to 0.95 (PG1 *vs.* others). Interestingly, the best model of the salivary microbiota, i.e., the one with the lowest number of predictor variables, was the model for the scale of overall oral health, which was formed by 30 OTUs and had AUC values ranging from 0.88 (OG1 *vs.* others) to 0.95 (OG23 *vs.* others) (Figure 30). This evidences the impact of both dental and periodontal conditions on saliva. Focusing on the group with the highest grade of oral pathology (OG23), the predictive potential of PG0_DG23 subgroup and DG0_PG23 subgroup with respect to the others was similar, with AUC values of 0.96 and 0.97, respectively.

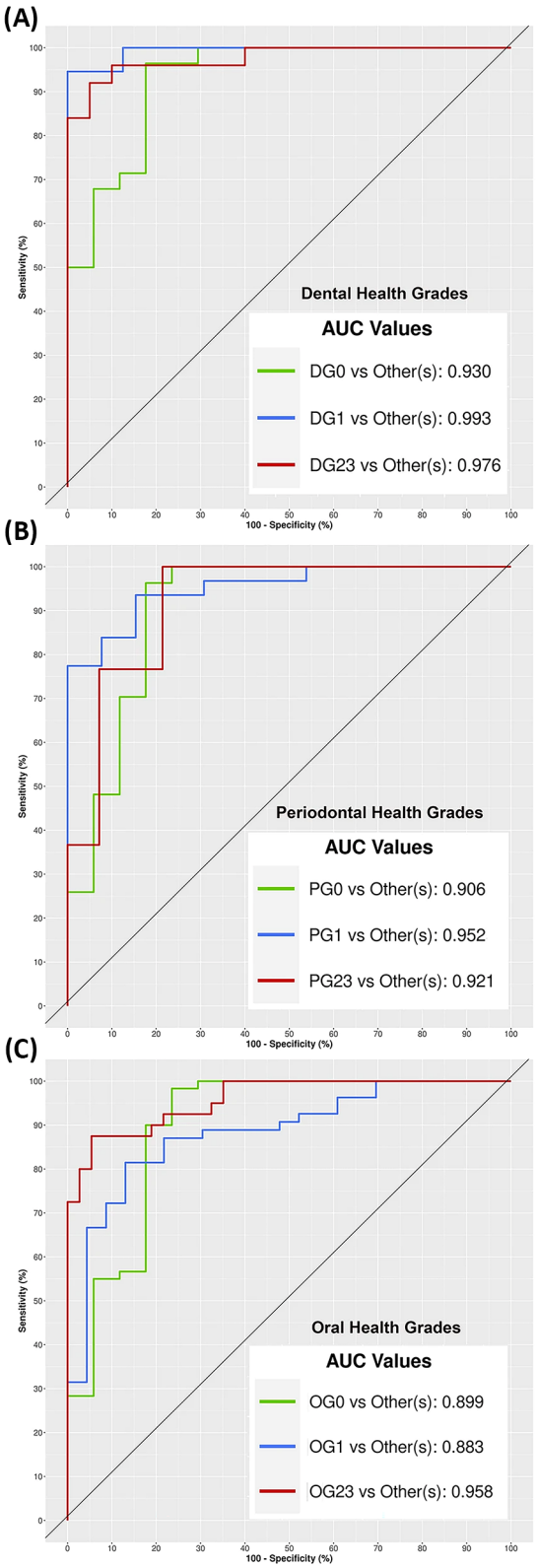





**Figure 30. Potential of the salivary microbiota to discriminate the different grades of dental and periodontal health and the combination of both: ROC curves and AUC values. The image was taken from Relvas et al. (249), an open-access article distributed under a Creative Commons Attribution 4.0 International (CC BY 4.0) license (https://creativecommons.org/licenses/by/4.0/).**

### I.6.2. Conclusions

This investigation concluded that the simultaneous presence of dental and periodontal pathology has a potentiating impact on the richness and diversity of the salivary microbiota. The structure of the bacterial community in oral health differs from that present in dental, periodontal, or oral disease, especially in high grades. The non-specific microbiota core comprises a greater number of species present in higher abundance than the specific core of a particular dental or periodontal condition (health or pathology). The number of taxa in the salivary microbiota with differential abundances between the DGs, PGs, or OGs represents, at most, a quarter of the bacterial community and are mainly non-core species. Supragingival dental parameters condition the abundance of the microbiota more than subgingival periodontal parameters, with the former contributing more to the impact that oral health has on the salivary microbiota.

The oral health-associated network has a bacterial community with more interconnections between its members, and a greater balance between its synergistic and competitive interactions, than the network associated with high grades of oral or dental disease; meanwhile, high grades of periodontal disease do not condition the characteristics of the salivary microbiota co-occurrence network. The possible keystone OTUs are different in relation to oral health and disease, and even these vary in different types of disease: half of them belong to the core microbiota and are independent of the abundance parameters.

The salivary microbiota, involving a considerable number of OTUs, shows an excellent discriminatory capability for distinguishing different grades of dental, periodontal, or oral disease; considering the number of predictive OTUs, the worst model is that which predicts the periodontal status, and the best model is that which predicts the combined dental and periodontal status.

## I.7. BIASES INTRODUCED IN THE 16S rRNA GENE SEQUENCING STUDIES

Although the recent advances in sequencing technologies have allowed the publication of hundreds of 16S rRNA gene-targeted investigations that studied the human oral microbiota profiles in different health states; the results of these research sometimes contradict other studies addressing the same disease or condition (98).

One possible reason for such disagreement is the multitude of systematic biases that can be introduced during each step of the 16S rRNA gene sequencing workflow (131,271). Some researchers have identified these weaknesses and reviewed the available evidence to provide the best quality practices to alleviate them and, thus, improve our understanding of the human microbiota (98,131,271,272). Here, we are going to provide a summary of the main findings of these investigations. Since the target-gene sequencing studies on the oral microbiome can be divided into three principal steps: clinical, laboratory, and bioinformatics -each with their corresponding sub-steps-; we elaborated three distinct summary tables.

Table 7 summarises the main points to consider in the clinical step, i.e., the selection of the study type before starting the investigation, the recording of all relevant patient data, the selection of the most suitable sample type and how it is going to be collected, and, lastly, the planning of the sample storage process according to the available resources.

**Table 7. Major clinical steps in the 16S rRNA gene sequencing workflow and main points to consider at each step in the oral studies. The table was constructed using the information from previous publications (120,125,216,217).**

| Clinical step | Considerations for oral studies |
|---|---|
| **Study design** | • The study type has to be selected according to the research question:<br>-Cross-sectional and case-control: it cannot assess temporality or causality<br>-Cohort: it assesses the role of the microbiota in aetiology of the disease and potential for disease risk prediction<br>-Interventional: it assesses the therapeutic or preventive effects of specific interventions |
| **Metadata collection** | • Several factors have been shown to affect the oral microbiota:<br>-Demographic factors: age, gender, socioeconomic status, education level, ethnicity<br>-General health factors: antibiotics, medication, systemic diseases, pregnancy<br>-Oral health factors: toothbrushing habits, tongue brushing, number of teeth, dentures, bleeding gums, tooth decay, oral hygiene level, diagnosis of periodontal disease, orthodontic appliances, intra-oral lesions, dry mouth, oral piercings<br>-Lifestyle factors: smoking, alcohol, sugar intake |
| **Sample type** | • A universal "oral microbiome" sample that would represent the entire ecosystem does not exist. The different intra-oral niches will result in different microbiota outcomes:<br>-Dental plaque (subgingival, supragingival), saliva, tongue dorsum, hard palate, keratinized gingiva, buccal mucosa… |
| **Sample collection** | • The sampling method should have a low risk of introducing contamination of the sample.<br>-Methods for collection: professional *vs.* self-sampling |
| **Sample storage** | • Storage temperature: ideally kept on ice and stored at -80ºC within 2 hours after collection<br>• Preservation methods if -80ºC storage is not possible: OMNI gene saliva kit, liquid dental transport medium (LDTM), RNAprotect solution... |

Next, table 8 illustrates the aspects to be regarded for the laboratory phase: the selection of the method to extract the genetic material from the samples, the elaboration of the PCR amplification protocol -mainly considering the hypervariable gene region to be amplified-, and the choice of the sequencing platform.

**Table 8. Major laboratory steps in the 16S rRNA gene sequencing workflow and main points to consider at each step in the oral studies. The table was constructed using the information from previous publications (120,125,216,217).**

| Laboratory step | Considerations for oral studies |
|---|---|
| **DNA extraction** | • Cell lysis method: chemical, mechanical, enzymatic, or a combination (preferable)<br>• DNA isolation kits: evidence is needed on the best option for different types of oral samples<br>• Eliminating contamination: positive and negative controls |
| **PCR amplification** | • Hypervariable region and primer pair choice affect the data obtained. Results from studies using distinct gene region and primer pairs should not be directly comparable<br>• Choice of polymerase can affect both the error rate of sequences and the abundance of chimeras |
| **Sequencing platform** | • Different platforms have distinct read lengths, throughput, and error rates/types. Its choice can impact the observed community. Results from studies using distinct platforms should not be directly comparable |

Lastly, table 9 lists the most relevant aspects to be considered during the bioinformatics analysis of the data, from the determination of the quality parameters to be applied to the sequences to the selection of the clustering method, the database to assign the taxonomy, and the analyses that will allow us to characterise the studied community.

**Table 9. Major bioinformatic steps in the 16S rRNA gene sequencing workflow and main points to consider at each step in the oral studies. The table was constructed using the information from previous publications (120,125,216,217).**

| Bioinformatic step | Considerations for oral studies |
|---|---|
| **Pre-processing of reads** | • An appropriate level of stringency when removing poor quality sequencing reads: the filtering depends on the sequencing platform, pipeline, and specific method used<br>• Removal of chimeras<br>• Sequence clustering of sequences into OTUs or by single-nucleotide resolution generates different final data tables |
| **Taxonomic classification** | • Choice of the reference database may alter the classifications: niche-specific had higher taxonomic resolution than the broad databases but might lose rare taxa<br>• Taxonomy assignment should be performed with broad-range databases in parallel to the oral-specific database |
| **Copy number correction** | • 16S rRNA gene experiments can be biased toward the detection of taxa with higher gene copy numbers<br>• There are phylogenetic methods for correcting this variation. While useful in well-studied environments, they generally introduce more noise than they correct |
| **Data analyses** | • Data normalization: the method used for normalization (as the commonly used rarefaction) may impact the study results<br>• The microbiota data should be treated as compositional; it cannot provide information on the absolute abundance of bacteria. The number of counts (reads) in the data set reflects the proportion of counts per feature (OTU, ASV, gene…) per sample, multiplied by the sequencing depth, thus the relative abundances. Ignoring this may lead to erroneous conclusions not based on true biological differences<br>• Downstream analysis tools. As we explained in the sections above, there is a great number of tools to analyse the data, each with its usefulness. The interpretation of differences in the ecological metrics should be done with care as it is still unclear how to translate these measures in clinical settings |

Although there are a number of promising approaches to help reduce the amount of systematic bias within the microbiome studies, the best way to help between-study comparisons is the creation of a comprehensive methodology section (131). Moreover, reports addressing the within-study biases that can lead to distorted observations of the true microbial communities compositions have to be performed (131).

## I.8. REFERENCES

(1) Kinane DF, Stathopoulou PG, Papapanou PN. Periodontal diseases. Nat Rev Dis Primers. 2017 Jun;3:17038. doi: 10.1038/nrdp.2017.38.

(2) World Health Organization. Global health estimates 2016: disease burden by cause, age, sex, by country, and by region, 2000-2016. Geneva, World Health Organization; 2018.

(3) Chapple ILC, Mealey BL, Van Dyke TE, Bartold PM, Dommisch H, Eickholz P, et al. Periodontal health and gingival diseases and conditions on an intact and a reduced periodontium: consensus report of workgroup 1 of the 2017 World Workshop on the Classification of Periodontal and Peri-Implant Diseases and Conditions. J Periodontol. 2018 Jun;89 Suppl 1:S74-84. doi: 10.1002/JPER.17-0719.

(4) Papapanou PN, Sanz M, Buduneli N, Dietrich T, Feres M, Fine DH, et al. Periodontitis: Consensus report of workgroup 2 of the 2017 World Workshop on the Classification of Periodontal and Peri-Implant Diseases and Conditions. J Periodontol. 2018 Jun;89 Suppl 1:S173-82. doi: 10.1002/JPER.17-0721.

(5) Peres MA, Macpherson LMD, Weyant RJ, Daly B, Venturelli R, Mathur MR, et al. Oral diseases: a global public health challenge. Lancet. 2019 Jul;394(10194):249-60.

(6) Kassebaum NJ, Smith AGC, Bernabé E, Fleming TD, Reynolds AE, Vos T, et al. Global, regional, and national prevalence, incidence, and disability-adjusted life years for oral conditions for 195 countries, 1990–2015: a systematic analysis for the global burden of diseases, injuries, and risk factors. J Dent Res. 2017 Apr;96(4):380-7.

(7) Eke PI, Borgnakke WS, Genco RJ. Recent epidemiologic trends in periodontitis in the USA. Periodontol 2000. 2020 Feb;82(1):257-67.

(8) Eke PI, Wei L, Borgnakke WS, Thornton-Evans G, Zhang X, Lu H, et al. Periodontitis prevalence in adults ≥ 65 years of age, in the USA. Periodontol 2000. 2016 Oct;72(1):76-95.

(9) Bravo-Pérez M, Almerich-Silla J, Ausina-Márquez V, Avilés-Gutiérrez P, Blanco-González J, Canorea-Díaz E, et al. Oral health survey in Spain 2015. RCOE. 2016 Jun;21(Suppl 1):8-48. Spanish.

(10) Llodra-Calvo JC. Oral health survey in Spain 2010. RCOE. 2012 Jan;17:13-41. Spanish.

(11) Ferreira MC, Dias-Pereira A, Branco-de-Almeida LS, Martins CC, Paiva SM. Impact of periodontal disease on quality of life: a systematic review. J Periodont Res. 2017 Aug;52(4):651-65.

(12) Chapple ILC, Bouchard P, Cagetti MG, Campus G, Carra M, Cocco F, et al. Interaction of lifestyle, behaviour or systemic diseases with dental caries and periodontal diseases: consensus report of group 2 of the joint EFP/ORCA workshop on the boundaries between caries and periodontal diseases. J Clin Periodontol. 2017 Mar;44 Suppl 18:S39-51. doi: 10.1111/jcpe.12685.

(13) Botelho J, Machado V, Proença L, Bellini DH, Chambrone L, Alcoforado G, et al. The impact of nonsurgical periodontal treatment on oral health-related quality of life: a systematic review and meta-analysis. Clin Oral Investig. 2020 Feb;24(2):585-96.

(14) AlJehani YA. Risk factors of periodontal disease: review of the literature. Int J Dent. 2021 Feb;2021:8735071. doi: 10.1155/2021/8735071.

(15) Nibali L, Bayliss-Chapman J, Almofareh SA, Zhou Y, Divaris K, Vieira AR. What is the heritability of periodontitis? a systematic review. J Dent Res. 2019 Jun;98(6):632-41.

(16) Nibali L, Di Iorio A, Tu Y, Vieira AR. Host genetics role in the pathogenesis of periodontal disease and caries. J Clin Periodontol. 2017 Mar;44 Suppl 18:S52-78. doi: 10.1111/jcpe.12639.



(17) Shiau HJ, Aichelmann-Reidy M, Reynolds MA. Influence of sex steroids on inflammation and bone metabolism. Periodontol 2000. 2014 Feb;64(1):81-94.

(18) Jepsen S, Blanco J, Buchalla W, Carvalho JC, Dietrich T, Dörfer C, et al. Prevention and control of dental caries and periodontal diseases at individual and population level: consensus report of group 3 of joint EFP/ORCA workshop on the boundaries between caries and periodontal diseases. J Clin Periodontol. 2017 Mar;44 Suppl 18:S85-93. doi: 10.1111/jcpe.12687.

(19) Zimmermann H, Zimmermann N, Hagenfeld D, Veile A, Kim T, Becher H. Is frequency of tooth brushing a risk factor for periodontitis? A systematic review and meta-analysis. Community Dent Oral Epidemiol. 2015 Apr;43(2):116-27.

(20) Trombelli L, Tatakis DN, Scapoli C, Bottega S, Orlandini E, Tosi M. Modulation of clinical expression of plaque-induced gingivitis. II. Identification of "high-responder" and "low-responder" subjects. J Clin Periodontol. 2004 Apr;31(4):239-52.

(21) Trombelli L, Scapoli C, Tatakis DN, Grassi L. Modulation of clinical expression of plaque-induced gingivitis: effects of personality traits, social support and stress. J Clin Periodontol. 2005 Nov;32(11):1143-50.

(22) Zhang Y, He J, He B, Huang R, Li M. Effect of tobacco on periodontal disease and oral cancer. Tob Induc Dis. 2019 May;17:40. doi: 10.18332/tid/106187.

(23) Leite FRM, Nascimento GG, Scheutz F, López R. Effect of smoking on periodontitis: a systematic review and meta-regression. Am J Prev Med. 2018 Jun;54(6):831-41.

(24) Leite F, Nascimento GG, Baake S, Pedersen LD, Scheutz F, López R. Impact of smoking cessation on periodontitis: a systematic review and meta-analysis of prospective longitudinal observational and interventional studies. Nicotine Tob Res. 2019 Nov;21(12):1600-8.

(25) Hanioka T, Morita M, Yamamoto T, Inagaki K, Wang P, Ito H, et al. Smoking and periodontal microorganisms. Jpn Dent Sci Rev. 2019 Nov;55(1):88-94.

(26) Wang J, Lv J, Wang W, Jiang X. Alcohol consumption and risk of periodontitis: a meta-analysis. J Clin Periodontol. 2016 Jul;43(7):572-83.

(27) Ryder MI, Couch ET, Chaffee BW. Personalized periodontal treatment for the tobacco- and alcohol-using patient. Periodontol 2000. 2018 Oct;78(1):30-46.

(28) Nyvad B, Takahashi N. Integrated hypothesis of dental caries and periodontal diseases. J Oral Microbiol. 2020 Jan;12(1):1710953. doi: 10.1080/20002297.2019.1710953.

(29) Woelber JP, Bremer K, Vach K, König D, Hellwig E, Ratka-Krüger P, et al. An oral health optimized diet can reduce gingival and periodontal inflammation in humans - a randomized controlled pilot study. BMC oral health. 2016 Jul;17(1):28. doi: 10.1186/s12903-016-0257-1.

(30) Woelber JP, Gärtner M, Breuninger L, Anderson A, König D, Hellwig E, et al. The influence of an anti-inflammatory diet on gingivitis. A randomized controlled trial. J Clin Periodontol. 2019 Apr;46(4):481-90.

(31) Decker A, Askar H, Tattan M, Taichman R, Wang H. The assessment of stress, depression, and inflammation as a collective risk factor for periodontal diseases: a systematic review. Clin Oral Investig. 2020 Jan;24(1):1-12.

(32) Albandar JM, Susin C, Hughes FJ. Manifestations of systemic diseases and conditions that affect the periodontal attachment apparatus: case definitions and diagnostic considerations. J Periodontol. 2018 Jun;89 Suppl 1:S183-203. doi: 10.1002/JPER.16-0480.

(33) Natto ZS, Abu Ahmad R,H., Alsharif LT, Alrowithi HF, Alsini DA, Salih HA, et al. Chronic periodontitis case definitions and confounders in periodontal research: a systematic assessment. Biomed Res Int. 2018 Nov;2018:4578782. doi: 10.1155/2018/4578782.



(34) Tonetti MS, Greenwell H, Kornman KS. Staging and grading of periodontitis: framework and proposal of a new classification and case definition. J Periodontol. 2018 Jun;89:S159-72. doi: 10.1002/JPER.18-0006.

(35) Arias-Bujanda N, Regueira-Iglesias A, Balsa-Castro C, Nibali L, Donos N, Tomás I. Accuracy of single molecular biomarkers in gingival crevicular fluid for the diagnosis of periodontitis: a systematic review and meta-analysis. J Clin Periodontol. 2019 Dec;46(12):1166-82.

(36) Arias-Bujanda N, Regueira-Iglesias A, Balsa-Castro C, Nibali L, Donos N, Tomás I. Accuracy of single molecular biomarkers in saliva for the diagnosis of periodontitis: a systematic review and meta-analysis. J Clin Periodontol. 2020 Jan;47(1):2-18.

(37) Williams RC, Offenbacher S. Periodontal medicine: the emergence of a new branch of periodontology. Periodontol 2000. 2000 Jun;23(1):9-12.

(38) Carrizales-Sepúlveda EF, Ordaz-Farías A, Vera-Pineda R, Flores-Ramírez R. Periodontal disease, systemic inflammation and the risk of cardiovascular disease. Heart Lung Circ. 2018 Nov;27(11):1327-34.

(39) Nascimento G, Leite F, Vestergaard P, Scheutz F, López R. Does diabetes increase the risk of periodontitis? A systematic review and meta-regression analysis of longitudinal prospective studies. Acta Diabetol. 2018 Jul;55(7):653-67.

(40) Gomes-Filho I, Cruz SSD, Trindade SC, Passos-Soares J, Carvalho-Filho P, Figueiredo ACMG, et al. Periodontitis and respiratory diseases: a systematic review with meta-analysis. Oral Dis. 2020 Mar;26(2):439-46.

(41) Fuggle NR, Smith TO, Kaul A, Sofat N. Hand to mouth: a systematic review and meta-analysis of the association between rheumatoid arthritis and periodontitis. Front Immunol. 2016 Mar;7:80. doi: 10.3389/fimmu.2016.00080.

(42) Leira Y, Domínguez C, Seoane J, Seoane-Romero J, Pías-Peleteiro JM, Takkouche B, et al. Is periodontal disease associated with alzheimer's disease? A systematic review with meta-analysis. Neuroepidemiology. 2017;48(1-2):21-31.

(43) Ide M, Papapanou PN. Epidemiology of association between maternal periodontal disease and adverse pregnancy outcomes – systematic review. J Clin Periodontol. 2013 Apr;40 Suppl 14:S181-94. doi: 10.1111/jcpe.12063.

(44) Nibali L, Tatarakis N, Needleman I, Tu Y, D'Aiuto F, Rizzo M, et al. Association between metabolic syndrome and periodontitis: a systematic review and meta-analysis. J Clin Endocrinol Metab. 2013 Mar;98(3):913-20.

(45) Martinez-Herrera M, Silvestre-Rangil J, Silvestre F. Association between obesity and periodontal disease. A systematic review of epidemiological studies and controlled clinical trials. Med Oral Patol Oral Cir Bucal. 2017 Nov;22(6):e708-15. doi: 10.4317/medoral.21786.

(46) Deschamps-Lenhardt S, Martin-Cabezas R, Hannedouche T, Huck O. Association between periodontitis and chronic kidney disease: systematic review and meta-analysis. Oral Dis. 2019 Mar;25(2):385-402.

(47) Hoare A, Soto C, Rojas-Celis V, Bravo D. Chronic inflammation as a link between periodontitis and carcinogenesis. Mediators Inflamm. 2019 Mar;2019:1029857. doi: 10.1155/2019/1029857.

(48) Van Dyke TE, van Winkelhoff AJ. Infection and inflammatory mechanisms. J Clin Periodontol. 2013 Apr;40 Suppl 14:S1-7. doi: 10.1111/jcpe.12088.

(49) Li X, Kolltveit KM, Tronstad L, Olsen I. Systemic diseases caused by oral infection. Clin Microbiol Rev. 2000 Oct;13(4):547-58.

(50) Petrosino JF, Highlander S, Luna RA, Gibbs RA, Versalovic J. Metagenomic pyrosequencing and microbial identification. Clin Chem. 2009 May;55(5):856-66.



(51) Sender R, Fuchs S, Milo R. Revised estimates for the number of human and bacteria cells in the body. PLoS Biol. 2016 Aug;14(8):e1002533. doi: 10.1371/journal.pbio.1002533.

(52) Verma D, Garg PK, Dubey AK. Insights into the human oral microbiome. Arch Microbiol. 2018 May;200(4):525-40.

(53) Sebastián-Domingo JJ. From the gut flora to the microbiome. Rev Esp Enferm Dig. 2018 Jan;110(1):51-6. Spanish.

(54) Teles R, Teles F, Frias-Lopez J, Paster B, Haffajee A. Lessons learned and unlearned in periodontal microbiology. Periodontol 2000. 2013 Jun;62(1):95-162.

(55) Hajishengallis G, Lamont RJ. Beyond the red complex and into more complexity: the polymicrobial synergy and dysbiosis (PSD) model of periodontal disease etiology. Mol Oral Microbiol. 2012 Dec;27(6):409-19.

(56) Blevins SM, Bronze MS. Robert Koch and the 'golden age' of bacteriology. Int J Infect Dis. 2010 Sep;14(9):e744-51. doi: 10.1016/j.ijid.2009.12.003.

(57) Krishnan K, Chen T, Paster BJ. A practical guide to the oral microbiome and its relation to health and disease. Oral Dis. 2017 Apr;23(3):276-86.

(58) Durán-Pinedo AE, Frias-Lopez J. Beyond microbial community composition: functional activities of the oral microbiome in health and disease. Microbes Infect. 2015 Jul;17(7):505-16.

(59) Aas JA, Paster BJ, Stokes LN, Olsen I, Dewhirst FE. Defining the normal bacterial flora of the oral cavity. J Clin Microbiol. 2005 Nov;43(11):5721-32.

(60) Socransky SS, Haffajee AD, Cugini MA, Smith C, Kent RL, Jr. Microbial complexes in subgingival plaque. J Clin Periodontol. 1998 Feb;25(2):134-44.

(61) Atieh MA. Accuracy of real-time polymerase chain reaction versus anaerobic culture in detection of Aggregatibacter actinomycetemcomitans and Porphyromonas gingivalis: a meta-analysis. J Periodontol. 2008 Sep;79(9):1620-9.

(62) Tomás I, Regueira-Iglesias A, López M, Arias-Bujanda N, Novoa L, Balsa-Castro C, et al. Quantification by qPCR of pathobionts in chronic periodontitis: development of predictive models of disease severity at site-specific level. Front Microbiol. 2017 Aug;8:1443. doi: 10.3389/fmicb.2017.01443.

(63) Belstrøm D, Fiehn N, Nielsen CH, Kirkby N, Twetman S, Klepac-Ceraj V, et al. Differences in bacterial saliva profile between periodontitis patients and a control cohort. J Clin Periodontol. 2014 Feb;41(2):104-12.

(64) Guzvic M. The history of DNA sequencing. J Med Biochem. 2012 Oct;32(4):301-12.

(65) Kumar PS, Griffen AL, Moeschberger ML, Leys EJ. Identification of candidate periodontal pathogens and beneficial species by quantitative 16S clonal analysis. J Clin Microbiol. 2005 Aug;43(8):3944-55.

(66) Human Microbiome Project Consortium. A framework for human microbiome research. Nature. 2012 Jun;486(7402):215-21.

(67) Human Microbiome Project Consortium. Structure, function and diversity of the healthy human microbiome. Nature. 2012 Jun;486(7402):207-14.

(68) Integrative HMP (iHMP) Research, Network Consortium. The Integrative Human Microbiome Project: dynamic analysis of microbiome-host omics profiles during periods of human health and disease. Cell Host Microbe. 2014 Sep;16(3):276-89.

(69) Fettweis JM, Serrano MG, Brooks JP, Edwards DJ, Girerd PH, Parikh HI, et al. The vaginal microbiome and preterm birth. Nat Med. 2019 Jun;25(6):1012-21.

(70) Lloyd-Price J, Arze C, Ananthakrishnan AN, Schirmer M, Avila-Pacheco J, Poon TW, et al. Multi-omics of the gut microbial ecosystem in inflammatory bowel diseases. Nature. 2019 May;569(7758):655-62.

(71) Zhou W, Sailani MR, Contrepois K, Zhou Y, Ahadi S, Leopold SR, et al. Longitudinal multi-omics of host–microbe dynamics in prediabetes. Nature. 2019 May;569(7758):663-71.

(72) The Integrative HMP (iHMP) Research Network Consortium. The Integrative Human Microbiome Project. Nature. 2019 May;569(7758):641-48.

(73) Mayanagi G, Sato T, Shimauchi H, Takahashi N. Detection frequency of periodontitis-associated bacteria by polymerase chain reaction in subgingival and supragingival plaque of periodontitis and healthy subjects. Oral Microbiol Immunol. 2004 Dec;19(6):379-85.

(74) Diaz PI, Chalmers NI, Rickard AH, Kong C, Milburn CL, Palmer RJ, Jr, et al. Molecular characterization of subject-specific oral microflora during initial colonization of enamel. Appl Environ Microbiol. 2006 Apr;72(4):2837-48.

(75) Dewhirst FE, Chen T, Izard J, Paster BJ, Tanner ACR, Yu W, et al. The human oral microbiome. J Bacteriol. 2010 Oct;192(19):5002-17

(76) Griffen A, Beall C, Firestone N, Gross E, Difranco J, Hardman J, et al. CORE: a phylogenetically-curated 16S rDNA database of the core oral microbiome. PLoS One. 2011 Apr;6(4):e19051. doi: 10.1371/journal.pone.0019051.

(77) Hajishengallis G. Periodontitis: from microbial immune subversion to systemic inflammation. Nature Rev Immunol. 2015 Jan;15(1):30-44.

(78) Van Dyke TE, Bartold PM, Reynolds EC. The nexus between periodontal inflammation and dysbiosis. Front Immunol. 2020 Mar;11:511. doi: 10.3389/fimmu.2020.00511.

(79) Lamont RJ, Koo H, Hajishengallis G. The oral microbiota: dynamic communities and host interactions. Nature Rev Microbiol. 2018 Dec;16(12):745-59.



(80) Silva N, Abusleme L, Bravo D, Dutzan N, Garcia-Sesnich J, Vernal R, et al. Host response mechanisms in periodontal diseases. J Appl Oral Sci. 2015 May-Jun;23(3):329-55.

(81) Cekici A, Kantarci A, Hasturk H, Van Dyke TE. Inflammatory and immune pathways in the pathogenesis of periodontal disease. Periodontol 2000. 2014 Feb;64(1):57-80.

(82) Pan W, Wang Q, Chen Q. The cytokine network involved in the host immune response to periodontitis. Int J Oral Sci. 2019 Nov;11(3):30. doi: 10.1038/s41368-019-0064-z.

(83) Garlet GP, Martins Jr W, Fonseca BAL, Ferreira BR, Silva JS. Matrix metalloproteinases, their physiological inhibitors and osteoclast factors are differentially regulated by the cytokine profile in human periodontal disease. J Clin Periodontol. 2004 Aug;31(8):671-9.

(84) Morgan XC, Huttenhower C. Chapter 12: human microbiome analysis. PLoS Comput Biol. 2012;8(12):e1002808. doi: 10.1371/journal.pcbi.1002808.

(85) Levy M, Thaiss CA, Elinav E. Metagenomic cross-talk: the regulatory interplay between immunogenomics and the microbiome. Genome Med. 2015 Nov;7:120. doi: 10.1186/s13073-015-0249-9.

(86) del Rosario-Rodicio M, del Carmen-Mendoza M. Bacterial identification by 16S rRNA sequencing: rationale, methodology and applications in clinical microbiology. Enferm Infecc Microbiol Clin. 2004 Apr;22(4):238-45. Spanish.

(87) Acinas SG, Marcelino LA, Klepac-Ceraj V, Polz MF. Divergence and redundancy of 16S rRNA sequences in genomes with multiple rrn operons. J Bacteriol. 2004 May;186(9):2629-35.

(88) Sun D, Jiang X, Wu QL, Zhou N. Intragenomic heterogeneity of 16S rRNA genes causes overestimation of prokaryotic diversity. Appl Environ Microbiol. 2013 Oct;79(19):5962-9.

(89) Bonk F, Popp D, Harms H, Centler F. PCR-based quantification of taxa-specific abundances in microbial communities: quantifying and avoiding common pitfalls. J Microbiol Methods. 2018 Oct;153:139-47.

(90) Stackebrandt E, Ebers J. Taxonomic parameters revisited: tarnished gold standards, Microbiol Today. 2006 Nov;33:152–55.

(91) Pei AY, Oberdorf WE, Nossa CW, Agarwal A, Chokshi P, Gerz EA, et al. Diversity of 16S rRNA genes within individual prokaryotic genomes. Appl Environ Microbiol. 2010 Jun;76(12):3886-97.

(92) Maxam AM, Gilbert W. A new method for sequencing DNA. Proc Natl Acad Sci U S A. 1977 Feb;74(2):560-4.

(93) Sanger F, Nicklen S, Coulson AR. DNA sequencing with chain-terminating inhibitors. Proc Natl Acad Sci U S A. 1977 Dec;74(12):5463-7.

(94) Siqueira JF,Jr, Fouad AF, Rocas IN. Pyrosequencing as a tool for better understanding of human microbiomes. J Oral Microbiol. 2012;4:10.3402/jom.v4i0.10743.

(95) Slatko BE, Gardner AF, Ausubel FM. Overview of next-generation sequencing technologies. Curr Protoc Mol Biol. 2018 Apr;122(1):e59. doi: 10.1002/cpmb.59.

(96) Paster BJ, Boches SK, Galvin JL, Ericson RE, Lau CN, Levanos VA, et al. Bacterial diversity in human subgingival plaque. J Bacteriol. 2001 Jun;183(12):3770-83.

(97) Mardis ER. Next-generation DNA sequencing methods. Annu Rev Genomics Hum Genet. 2008 Sep;9(1):387-402.

(98) Zaura E, Pappalardo VY, Buijs MJ, Volgenant CMC, Brandt BW. Optimizing the quality of clinical studies on oral microbiome: a practical guide for planning, performing, and reporting. Periodontol 2000. 2021 Feb;85(1):210-36.



(99) Kuczynski J, Costello EK, Nemergut DR, Zaneveld J, Lauber CL, Knights D, et al. Direct sequencing of the human microbiome readily reveals community differences. Genome Biol. 2010;11(5):210. doi: 10.1186/gb-2010-11-5-210.

(100) Lemos LN, Fulthorpe RR, Triplett EW, Roesch LF. Rethinking microbial diversity analysis in the high throughput sequencing era. J Microbiol Methods. 2011 Jul;86(1):42-51.

(101) Illumina, Inc. An introduction to next-generation sequencing technology. 2017; Available at: www.illumina.com.

(102) Chen Y, Liu T, Yu C, Chiang T, Hwang C. Effects of GC bias in next-generation-sequencing data on de novo genome assembly. PloS One. 2013 Apr;8(4):e62856. doi: 10.1371/journal.pone.0062856.

(103) van Dijk EL, Jaszczyszyn Y, Naquin D, Thermes C. The third revolution in sequencing technology. Trends Genet. 2018 Sep;34(9):666-81.

(104) Clarridge JE 3rd. Impact of 16S rRNA gene sequence analysis for identification of bacteria on clinical microbiology and infectious diseases. Clin Microbiol Rev. 2004 Oct;17(4):840-62.

(105) Kuczynski J, Lauber CL, Walters WA, Parfrey LW, Clemente JC, Gevers D, et al. Experimental and analytical tools for studying the human microbiome. Nat Rev Genet. 2011 Dec;13(1):47-58.

(106) Schadt EE, Turner S, Kasarskis A. A window into third-generation sequencing. Hum Mol Genet. 2010 Oct;19(R2):R227-40. doi: 10.1093/hmg/ddq416.

(107) Liu L, Li Y, Li S, Hu N, He Y, Pong R, et al. Comparison of next-generation sequencing systems. J Biomed Biotechnol. 2012;2012:251364. doi: 10.1155/2012/251364.

(108) Eid J, Fehr A, Gray J, Luong K, Lyle J, Otto G, et al. Real-time DNA sequencing from single polymerase molecules. Science. 2009 Jan;323(5910):133-8.

(109) Midha MK, Wu M, Chiu KP. Long-read sequencing in deciphering human genetics to a greater depth. Hum Genet. 2019 Dec;138(11-12):1201-15.

(110) Ke R, Mignardi M, Pacureanu A, Svedlund J, Botling J, Wählby C, et al. In situ sequencing for RNA analysis in preserved tissue and cells. Nat Methods. 2013 Sep;10(9):857-60.

(111) Lee JH, Daugharthy ER, Scheiman J, Kalhor R, Ferrante TC, Terry R, et al. Fluorescent in situ sequencing (FISSEQ) of RNA for gene expression profiling in intact cells and tissues. Nat Protoc. 2015 Mar;10(3):442-58.

(112) Ke R, Mignardi M, Hauling T, Nilsson M. Fourth generation of next-generation sequencing technologies: promise and consequences. Hum Mutat. 2016 Dec;37(12):1363-7.

(113) Ståhl PL, Salmén F, Vickovic S, Lundmark A, Navarro JF, Magnusson J, et al. Visualization and analysis of gene expression in tissue sections by spatial transcriptomics. Science. 2016 Jul;353(6294):78-82.

(114) Suárez-Moya A. Microbiome and next generation sequencing. Rev Esp Quimioter. 2017 Oct;30(5):305-11. Spanish.

(115) Luscombe N, Greenbaum D, Gerstein M. What is bioinformatics? A proposed definition and overview of the field. Methods Inf Med. 2001;40(4):346-58.

(116) Oliver GR, Hart SN, Klee EW. Bioinformatics for clinical next generation sequencing. Clin Chem. 2015 Jan;61(1):124-35.

(117) Caporaso JG, Kuczynski J, Stombaugh J, Bittinger K, Bushman FD, Costello EK, et al. QIIME allows analysis of high-throughput community sequencing data. Nat Methods. 2010 May;7(5):335-6.

(118) Schloss PD, Westcott SL, Ryabin T, Hall JR, Hartmann M, Hollister EB, et al. Introducing mothur: open-source, platform-independent, community-supported software for describing and comparing microbial communities. Appl Environ Microbiol. 2009 Dec;75(23):7537-41.

(119) Cole JR, Wang Q, Fish JA, Chai B, McGarrell DM, Sun Y, et al. Ribosomal Database Project: data and tools for high throughput rRNA analysis. Nucleic Acids Res. 2014 Jan;42(D1):D633-42. doi: 10.1093/nar/gkt1244.

(120) Bolyen E, Rideout JR, Dillon MR, Bokulich NA, Abnet CC, Al-Ghalith G, et al. Reproducible, interactive, scalable and extensible microbiome data science using QIIME 2. Nat Biotechnol. 2019 Aug;37(8):852-7.

(121) Ju F, Zhang T. 16S rRNA gene high-throughput sequencing data mining of microbial diversity and interactions. Appl Microbiol Biotechnol. 2015 May;99(10):4119-29.

(122) Roy S, Coldren C, Karunamurthy A, Kip NS, Klee EW, Lincoln SE, et al. Standards and guidelines for validating next-generation sequencing bioinformatics pipelines: a joint recommendation of the association for molecular pathology and the college of American pathologists. J Mol Diagn. 2018 Jan;20(1):4-27.

(123) Quince C, Lanzen A, Davenport RJ, Turnbaugh PJ. Removing noise from pyrosequenced amplicons. BMC Bioinformatics. 2011 Jan;12:38. doi: 10.1186/1471-2105-12-38.

(124) Haas BJ, Gevers D, Earl AM, Feldgarden M, Ward DV, Giannoukos G, et al. Chimeric 16S rRNA sequence formation and detection in Sanger and 454-pyrosequenced PCR amplicons. Genome Res. 2011 Mar;21(3):494-504.

(125) Edgar R, Haas BJ, Clemente JC, Quince C, Knight R. UCHIME improves sensitivity and speed of chimera detection. Bioinformatics. 2011 Aug;27(16):2194-200.

(126) Edgar R. UPARSE: highly accurate OTU sequences from microbial amplicon reads. Nat Methods. 2013 Oct;10(10):996-8.



(127) Caruso V, Song X, Asquith M, Karstens L. Performance of microbiome sequence inference methods in environments with varying biomass. mSystems. 2019 Feb;4(1):e00163-18. doi: 10.1128/mSystems.00163-18.

(128) Nearing JT, Douglas GM, Comeau AM, Langille MGI. Denoising the denoisers: an independent evaluation of microbiome sequence error-correction approaches. PeerJ. 2018 Aug;6:e5364. doi: 10.7717/peerj.5364.

(129) Stackebrandt E, Goebel, BM. Taxonomic note: a place for DNA-DNA reassociation and 16s rRNA sequence analysis in the present species definition in bacteriology. Int J Syst Evol Microbiol. 1994 Oct;44(4):846-9.

(130) Edgar R. Search and clustering orders of magnitude faster than BLAST. Bioinformatics. 2010 Oct;26(19):2460-1.

(131) Nearing JT, Comeau AM, Langille MGI. Identifying biases and their potential solutions in human microbiome studies. Microbiome. 2021 May;9(1):113. doi: 10.1186/s40168-021-01059-0.

(132) Westcott SL, Schloss PD. De novo clustering methods outperform reference-based methods for assigning 16S rRNA gene sequences to operational taxonomic units. PeerJ. 2015 Dec;3:e1487. doi: 10.7717/peerj.1487.

(133) He Y, Caporaso JG, Jiang XT, Sheng HF, Huse SM, Rideout JR, et al. Stability of operational taxonomic units: an important but neglected property for analyzing microbial diversity. Microbiome. 2015 May;3:20. doi: 10.1186/s40168-015-0081-x.

(134) Wei Z, Zhang X, Cao M, Liu F, Qian Y, Zhang S. Comparison of methods for picking the operational taxonomic units from amplicon sequences. Front Microbiol. 2021 Mar; 12:644012. doi: 10.3389/fmicb.2021.644012.

(135) 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.

(136) Amir A, McDonald D, Navas-Molina JA, Kopylova E, Morton JT, Zech Xu Z, et al. Deblur rapidly resolves single-nucleotide community sequence patterns. mSystems. 2017 Mar;2(2):e00191-16. doi: 10.1128/mSystems.00191-16.

(137) Edgar R. UNOISE2: improved error-correction for Illumina 16S and ITS amplicon sequencing. Preprint at bioRxiv. 2016. doi: 10.1101/081257.

(138) Eren AM, Morrison HG, Lescault PJ, Reveillaud J, Vineis JH, Sogin ML. Minimum entropy decomposition: unsupervised oligotyping for sensitive partitioning of high-throughput marker gene sequences. ISME J. 2015 Mar;9(4):968-79.

(139) Prodan A, Tremaroli V, Brolin H, Zwinderman AH, Nieuwdorp M, Levin E. Comparing bioinformatic pipelines for microbial 16S rRNA amplicon sequencing. PLoS One. 2020 Jan;15(1):e0227434. doi: 10.1371/journal.pone.0227434.

(140) Schloss PD. Amplicon sequence variants artificially split bacterial genomes into separate clusters. mSphere. 2021 Aug;6(4):e0019121. doi: 10.1128/mSphere.00191-21.

(141) Abellan-Schneyder I, Matchado MS, Reitmeier S, Sommer A, Sewald Z, Baumbach J, et al. Primer, pipelines, parameters: issues in 16S rRNA gene sequencing. mSphere. 2021 Feb;6(1):e01202-20. doi: 10.1128/mSphere.01202-20.

(142) García-López R, Cornejo-Granados F, Lopez-Zavala A, Cota-Huízar A, Sotelo-Mundo R, Gómez-Gil B, et al. OTUs and ASVs produce comparable taxonomic and diversity from shrimp microbiota 16S profiles using tailored abundance filters. Genes (Basel). 2021 Apr;12(4):564. doi: 10.3390/genes12040564.

(143) Stoddard SF, Smith BJ, Hein R, Roller BR, Schmidt TM. rrnDB: improved tools for interpreting rRNA gene abundance in bacteria and archaea and a new foundation for future development. Nucleic Acids Res. 2015 Jan;43(Database issue):D593-8. doi: 10.1093/nar/gku1201.

(144) Callahan BJ, McMurdie PJ, Holmes SP. Exact sequence variants should replace operational taxonomic units in marker-gene data analysis. ISME J. 2017 Dec;11(12):2639-43.

(145) Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. J Mol Biol. 1990 Oct;215(3):403-10.

(146) Li W, Chang Y. CD-HIT-OTU-MiSeq, an improved approach for clustering and analyzing paired end MiSeq 16S rRNA sequences. Preprint at bioRxiv. 2017. doi: 10.1101/153783.

(147) DeSantis TZ, Hugenholtz P, Keller K, Brodie EL, Larsen N, Piceno YM, et al. NAST: a multiple sequence alignment server for comparative analysis of 16S rRNA genes. Nucleic Acids Res. 2006 Jul;34(Web Server issue):W394-9. doi: 10.1093/nar/gkl244.

(148) Price MN, Dehal PS, Arkin AP. FastTree 2--approximately maximum-likelihood trees for large alignments. PLoS One. 2010 Mar;5(3):e9490. doi: 10.1371/journal.pone.0009490.

(149) DeSantis TZ, Hugenholtz P, Larsen N, Rojas M, Brodie EL, Keller K, et al. Greengenes, a chimera-checked 16S rRNA gene database and workbench compatible with ARB. Appl Environ Microbiol. 2006 Jul;72(7):5069-72.

(150) Quast C, Pruesse E, Yilmaz P, Gerken J, Schweer T, Yarza P, et al. 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.

(151) Chen T, Yu W, 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;2010:baq013. doi: 10.1093/database/baq013.



(152) Clark K, Karsch-Mizrachi I, Lipman DJ, Ostell J, Sayers EW. GenBank. Nucleic Acids Res. 2016 Jan;44:D67-72. doi: 10.1093/nar/gkv1276.

(153) Edgar R. MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res. 2004 Mar;32(5):1792-7.

(154) Caporaso JG, Bittinger K, Bushman FD, DeSantis TZ, Andersen GL, Knight R. PyNAST: a flexible tool for aligning sequences to a template alignment. Bioinformatics. 2010 Jan;26(2):266-7.

(155) Tamura K, Peterson D, Peterson N, Stecher G, Nei M, Kumar S. MEGA5: molecular evolutionary genetics analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods. Mol Biol Evol. 2011 Oct;28(10):2731-9.

(156) Ashton JJ, Beattie RM, Ennis S, Cleary DW. Analysis and interpretation of the human microbiome. Inflamm Bowel Dis. 2016 Jul;22(7):1713-22.

(157) Edlund A, Yang Y, Hall AP, Guo L, Lux R, He X, et al. An in vitro biofilm model system maintaining a highly reproducible species and metabolic diversity approaching that of the human oral microbiome. Microbiome. 2013 Oct;1(1):25. doi: 10.1186/2049-2618-1-25.

(158) Chen J, Bittinger K, Charlson ES, Hoffmann C, Lewis J, Wu GD, et al. Associating microbiome composition with environmental covariates using generalized UniFrac distances. Bioinformatics. 2012 Aug;28(16):2106-13.

(159) National Research Council (US) Committee on Metagenomics: Challenges and Functional Applications. The New Science of Metagenomics: Revealing the Secrets of Our Microbial Planet. Washington (DC): National Academies Press (US); 2007.

(160) Huson DH, Auch AF, Qi J, Schuster SC. MEGAN analysis of metagenomic data. Genome Res. 2007 Mar;17(3):377-86.



(161) Meyer F, Paarmann D, D'Souza M, Olson R, Glass EM, Kubal M, et al. The metagenomics RAST server - a public resource for the automatic phylogenetic and functional

analysis of metagenomes. BMC Bioinformatics. 2008 Sep;9:386. doi: 10.1186/1471-2105-9-386.

(162) Lozupone C, Knight R. UniFrac: a new phylogenetic method for comparing microbial communities. Appl Environ Microbiol. 2005 Dec;71(12):8228-35.

(163) Schloss PD, Handelsman J. Introducing DOTUR, a computer program for defining operational taxonomic units and estimating species richness. Appl Environ Microbiol. 2005 Mar;71(3):1501-6.

(164) White JR, Nagarajan N, Pop M. Statistical methods for detecting differentially abundant features in clinical metagenomic samples. PLoS Comput Biol. 2009 Apr; 5(4):e1000352. doi: 10.1371/journal.pcbi.1000352.

(165) Segata N, Izard J, Waldron L, Gevers D, Miropolsky L, Garrett WS, et al. Metagenomic biomarker discovery and explanation. Genome Biol 2011 Jun;12(6):R60. doi: 10.1186/gb-2011-12-6-r60.

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

(167) R Core Team. R: a language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing; 2021; Available at: https://www.R-project.org/.

(168) Callahan B, Proctor D, Relman D, Fukuyama J, Holmes S. Reproducible research workflow in R for the analysis of personalized human microbiome data. Pac Symp Biocomput. 2016;21:183-94.



(169) Oksanen J, Blanchet FG, Friendly M, Kindt R, Legendre P, McGlinn D, et al. vegan: community ecology package. 2020; Available at: https://cran.r-project.org, https://github.com/vegandevs/vegan.

(170) Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol. 2014 Dec;15(12)550. doi: 10.1186/s13059-014-0550-8.

(171) Wickham H. ggplot2: elegant graphics for data analysis. New York: Springer-Verlag; 2016; Available at: https://ggplot2.tidyverse.org.

(172) McMurdie PJ, Holmes S. Shiny-phyloseq: web application for interactive microbiome analysis with provenance tracking. Bioinformatics. 2015 Jan;31(2):282-3.

(173) Lahti L, Shetty S. Tools for microbiome analysis in R. Microbiome package. 2017; Available at: http://microbiome.github.com/microbiome.

(174) Kim BR, Shin J, Guevarra R, Lee JH, Kim DW, Seol KH, et al. Deciphering diversity indices for a better understanding of microbial communities. J Microbiol Biotechnol. 2017 Dec;27(12):2089-93.

(175) Lozupone CA, Knight R. Species divergence and the measurement of microbial diversity. FEMS Microbiol Rev. 2008 Jul;32(4):557-78.

(176) Willis AD. Rarefaction, alpha diversity, and statistics. Front Microbiol. 2019 Oct;10:2407. doi: 10.3389/fmicb.2019.02407.

(177) Cox MJ, Cookson WO, Moffatt MF. Sequencing the human microbiome in health and disease. Hum Mol Genet. 2013 Oct;22(R1):R88-94. doi: 10.1093/hmg/ddt398.

(178) Hugerth LW, Andersson AF. Analysing microbial community composition through amplicon sequencing: from sampling to hypothesis testing. Front Microbiol. 2017 Sep;8:1561. doi: 10.3389/fmicb.2017.01561.

(179) Hughes JB, Hellmann JJ, Ricketts TH, Bohannan BJM. Counting the uncountable: statistical approaches to estimating microbial diversity. Appl Environ Microbiol. 2001 Oct;67(10):4399-406.

(180) Sanders HL. Marine benthic diversity: a comparative study. Am Nat. 1968 May-Jun;102(925):243-82.

(181) McMurdie PJ, Holmes S. Waste not, want not: why rarefying microbiome data is inadmissible. PLOS Comput Biol. 2014 Apr;10(4):e1003531. doi: 10.1371/journal.pcbi.1003531.

(182) Chao A. Non-parametric estimation of the classes in a population. Scand J Stat. 1984 Jan;11(4):265-70.

(183) Chao A, Lee s. Estimating the number of classes via sample coverage. J Am Stat Assoc. 1992 Mar;87(417):210-7.

(184) Faith DP. Conservation evaluation and phylogenetic diversity. Biol Conserv. 1992;61(1):1-10.

(185) Shannon CE. A mathematical theory of communication. Bell Syst Tech J. 1948 Jul;27(3):379-423.

(186) Simpson EH. Measurement of Diversity. Nature. 1949 Apr;163(4148):688.

(187) Pielou EC. The measurement of diversity in different types of biological collections. J Theor Biol. 1966 Dec;13:131-44.

(188) Cadotte MW, Davies TJ, Regetz J, Kembel SW, Cleland E, Oakley TH. Phylogenetic diversity metrics for ecological communities: integrating species richness, abundance and evolutionary history. Ecol Lett. 2010 Jan;13(1):96-105.

(189) Goodrich J, Di Rienzi S, Poole A, Koren O, Walters W, Caporaso J, et al. Conducting a microbiome study. Cell. 2014 Jul;158(2):250-62.

(190) Jaccard P. Study of the floral distribution in a portion of the Alps and the Jura. Bull Soc Vaud Sci Nat. 1901 Jan;37(142):547-79. French.

(191) Bray JR, Curtis JT. An ordination of the upland forest communities of southern Wisconsin. Ecol Monogr. 1957 Oct;27(4):325-49.

(192) Schmidt TSB, Matias-Rodrigues JF, von Mering C. A family of interaction-adjusted indices of community similarity. ISME J. 2017 Mar;11(3):791-807.

(193) Chao A, Chazdon RL, Colwell RK, Shen T. A new statistical approach for assessing similarity of species composition with incidence and abundance data. Ecol Lett. 2005 Feb;8(2):148-59.

(194) Lozupone CA, Hamady M, Kelley ST, Knight R. Quantitative and qualitative beta diversity measures lead to different insights into factors that structure microbial communities. Appl Environ Microbiol. 2007 Mar;73(5):1576-85.

(195) Chang Q, Luan Y, Sun F. Variance adjusted weighted UniFrac: a powerful beta diversity measure for comparing communities based on phylogeny. BMC Bioinformatics. 2011 Apr;12:118. doi: 10.1186/1471-2105-12-118.

(196) Xia Y, Sun J. Hypothesis testing and statistical analysis of microbiome. Genes Dis. 2017 Sep;4(3):138-48.

(197) Paliy O, Shankar V. Application of multivariate statistical techniques in microbial ecology. Mol Ecol. 2016 Mar;25(5):1032-57.



(198) Pearson K. LIII. On lines and planes of closest fit to systems of points in space. London, Edinburgh Dublin Philos Mag J Sci. 1901 Nov;2(11):559-72.

(199) Ramette A. Multivariate analyses in microbial ecology. FEMS Microbiol Ecol. 2007 Nov;62(2):142-60.

(200) Gower JC. Some distance properties of latent root and vector methods used in multivariate analysis. Biometrika. 1966 Dec;53(3-4):325-38.

(201) Rao B, Lou J, Lu H, Liang H, Li J, Zhou H, et al. Oral microbiome characteristics in patients with autoimmune hepatitis. Front Cell Infect Microbiol. 2021 May;11:656674. doi: 10.3389/fcimb.2021.656674.

(202) Kruskal JB. Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis. Psychometrika. 1964 Mar;29(1):1-27.

(203) Drell T, Štšepetova J, Simm J, Rull K, Aleksejeva A, Antson A, et al. The influence of different maternal microbial communities on the development of infant gut and oral microbiota. Sci Rep. 2017 Aug;7(1):9940. doi: 10.1038/s41598-017-09278-y.

(204) Hotelling H. Relations between two sets of variates. Biometrika. 1936 Dec;28(3-4):321-77.

(205) Gower JC. Generalized procrustes analysis. Psychometrika. 1975 Mar;40(1):33-51.

(206) van den Wollenberg, Arnold L. Redundancy analysis an alternative for canonical correlation analysis. Psychometrika. 1977 Jun;42(2):207-19.

(207) Nelder JA, Wedderburn RWM. Generalized linear models. J R Stat Soc Series A Stat Methodol. 1972;135(3):370-84.

(208) Qi Y, Zang S, Wei J, Yu H, Yang Z, Wu H, et al. High-throughput sequencing provides insights into oral microbiota dysbiosis in association with inflammatory bowel disease. Genomics. 2021 Jan;113(1, Pt 2):664-76.

(209) Anderson M. A new method for non-parametric multivariate analysis of variance. Austral Ecol. 2001 Feb;26(1):32-46.

(210) Clarke KR. Non-parametric multivariate analyses of changes in community structure. Aust J Ecol 1993;18(1):117-43.

(211) Mielke PW, Berry KJ, Johnson ES. Multi-response permutation procedures for a priori classifications. Commun Stat - Theory Methods. 1976;5(14):1409-24.

(212) Mantel N. The detection of disease clustering and a generalized regression approach. Cancer Res. 1967 Feb;27(2):209-20.

(213) Fisher RA. The use of multiple measurements in taxonomic problems. Ann Eugen. 1936 Sep;7(2):179-88.

(214) Breiman L. Random forests. Mach Learn. 2001 Oct;45(1):5-32.

(215) Loh W. Classification and regression trees. WIREs Data Mining Knowl Discov. 2011 Jan-Feb;1(1):14-23.

(216) Adams SE, Arnold D, Murphy B, Carroll P, Green AK, Smith AM, et al. A randomised clinical study to determine the effect of a toothpaste containing enzymes and proteins on plaque oral microbiome ecology. Sci Rep. 2017 Feb;7:43344. doi: 10.1038/srep43344.

(217) Robinson MD, McCarthy DJ, Smyth GK. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics. 2010 Jan;26(1):139-40.

(218) Paulson JN, Stine OC, Bravo H, Pop M. Differential abundance analysis for microbial marker-gene surveys. Nat Methods. 2013 Dec;10(12):1200-2.



(219) Shade A, Handelsman J. Beyond the Venn diagram: the hunt for a core microbiome. Environ Microbiol. 2012 Jan;14(1):4-12.

(220) Layeghifard M, Hwang DM, Guttman DS. Constructing and analyzing microbiome networks in R. In: Beiko RG, Hsiao W, Parkinson J, editors. Microbiome analysis. Methods and protocols. New York: Springer Nature; 2018. p. 243-66.

(221) Jiang D, Armour CR, Hu C, Mei M, Tian C, Sharpton TJ, et al. Microbiome multi-omics network analysis: statistical considerations, limitations, and opportunities. Front Genet. 2019 Nov;10:995. doi: 10.3389/fgene.2019.00995.

(222) Banerjee S, Schlaeppi K, van der Heijden MGA. Keystone taxa as drivers of microbiome structure and functioning. Nat Rev Microbiol. 2018 Sep;16(9):567-76.

(223) Castro-Nallar E, Gutzwiller F, Mendez KN. Co-occurrence networks of micro-organisms. Center for Bioinformatics & Integrative Biology, Universidad Andrés Bello. 2019; Available at: http://www.castrolab.org/isme/microbial_networks/microbial_networks.html. Spanish.

(224) Friedman J, Alm EJ. Inferring correlation networks from genomic survey data. PLoS Comput Biol. 2012 Sep; 8(9):e1002687. doi: 10.1371/journal.pcbi.1002687.

(225) Kurtz ZD, Müller CL, Miraldi ER, Littman DR, Blaser MJ, Bonneau RA. Sparse and compositionally robust inference of microbial ecological networks. PLoS Comput Biol. 2015 May;11(5):e1004226. doi: 10.1371/journal.pcbi.1004226.

(226) Kurtz Z, Mueller C, Miraldi E, Bonneau R. SpiecEasi: sparse inverse covariance for ecological statistical inference. 2021; Available at: https://github.com/zdk123/SpiecEasi.

(227) Csardi G, Nepusz T. The Igraph software package for complex network research. InterJournal, Complex Systems. 2006; 1695; Available at: http://igraph.org.

(228) Epskamp S, Cramer AOJ, Waldorp LJ, Schmittmann VD, Borsboom D. qgraph: network visualizations of relationships in psychometric data. J Stat Softw. 2012 Apr;48(4):1–18.

(229) Golbeck J. Chapter 3 - Network structure and measures. In: Golbeck J, editor. Analyzing the social web. Boston: Morgan Kaufmann; 2013. p. 25-44.

(230) Manirajan BA, Maisinger C, Ratering S, Rusch V, Schwiertz A, Cardinale M, et al. Diversity, specificity, co-occurrence and hub taxa of the bacterial-fungal pollen microbiome. FEMS Microbiol Ecol. 2018 Aug;94(8):10.1093/femsec/fiy112.

(231) Banerjee S, Schlaeppi K, van der Heijden MGA. Reply to 'Can we predict microbial keystones?'. Nat Rev Microbiol. 2019 Mar;17(3):194. doi: 10.1038/s41579-018-0133-x.

(232) Röttjers L, Faust K. Can we predict keystones?. Nat Rev Microbiol. 2019 Mar;17(3):193. doi: 10.1038/s41579-018-0132-y.

(233) Lupatini M, Suleiman AKA, Jacques RJS, Antoniolli ZI, de Siquiera-Ferreira A, Kuramae EE, et al. Network topology reveals high connectance levels and few key microbial genera within soils. Front Environ Sci. 2014 May;2:10. doi: 10.3389/fenvs.2014.00010.

(234) Camacho DM, Collins KM, Powers RK, Costello JC, Collins JJ. Next-generation machine learning for biological networks. Cell. 2018 Jun;173(7):1581-92.

(235) Johnson KW, Torres-Soto J, Glicksberg BS, Shameer K, Miotto R, Ali M, et al. Artificial intelligence in cardiology. J Am Coll Cardiol. 2018 Jun;71(23):2668-79.

(236) Knights D, Costello EK, Knight R. Supervised classification of human microbiota. FEMS Microbiol Rev. 2011 Mar;35(2):343-59.

(237) Reel PS, Reel S, Pearson E, Trucco E, Jefferson E. Using machine learning approaches for multi-omics data analysis: a review. Biotechnol Adv. 2021 Jul-Aug;49:107739. doi: 10.1016/j.biotechadv.2021.107739.

(238) Lê Cao KA, Boitard S, Besse P. Sparse PLS discriminant analysis: biologically relevant feature selection and graphical displays for multiclass problems. BMC Bioinformatics. 2011 Jun;12:253. doi: 10.1186/1471-2105-12-253.

(239) Cortes C, Vapnik V. Support-vector networks. Mach Learn. 1995 Sep;20(3):273-97.

(240) Patel M, Gupta M. Chapter 7 - Caravan insurance customer profile modeling with R. In: Zhao Y, Cen Y, editors. Data mining applications with R. Boston: Academic Press; 2014. p. 181-227.

(241) Statnikov A, Aliferis CF, Tsamardinos I, Hardin D, Levy S. A comprehensive evaluation of multicategory classification methods for microarray gene expression cancer diagnosis. Bioinformatics. 2005 Mar;21(5):631-43.

(242) Chun H, Keleş S. Sparse partial least squares regression for simultaneous dimension reduction and variable selection. J R Stat Soc Series B Stat Methodol. 2010 Jan;72(1):3-25

(243) Chung D, Keles S. Sparse partial least squares classification for high dimensional data. Stat Appl Genet Mol Biol. 2010;9(1):Article17. doi: 10.2202/1544-6115.1492.

(244) Rohart F, Gautier B, Singh A, Lê Cao K. mixOmics: an R package for 'omics feature selection and multiple data integration. PLoS Comput Biol. 2017 Nov;13(11):e1005752. doi: 10.1371/journal.pcbi.1005752.

(245) Gentleman RC, Carey VJ, Bates DM, Bolstad B, Dettling M, Dudoit S, et al. Bioconductor: open software development for computational biology and bioinformatics. Genome Biol. 2004 Sep;5(10):R80. doi: 10.1186/gb-2004-5-10-r80.

(246) Lê Cao K, Dejean S, Abadi AJ. mixOmics vignette. 2019; Available at: https://mixomicsteam.github.io/Bookdown/index.html.

(247) Acharya A, Chen T, Chan Y, Watt RM, Jin L, Mattheos N. Species-level salivary microbial indicators of well-resolved periodontitis: a preliminary investigation. Front Cell Infect Microbiol. 2019 Oct;9:347. doi: 10.3389/fcimb.2019.00347.

(248) Chen H, Liu Y, Zhang M, Wang G, Qi Z, Bridgewater L, et al. A Filifactor alocis-centered co-occurrence group associates with periodontitis across different oral habitats. Sci Rep. 2015 Mar;5:9053. doi: 10.1038/srep09053.

(249) Relvas M, Regueira-Iglesias A, Balsa-Castro C, Salazar F, Pacheco JJ, Cabral C, et al. Relationship between dental and periodontal health status and the salivary microbiome: bacterial diversity, co-occurrence networks and predictive models. Sci Rep. 2021 Jan;11(1):929. doi: 10.1038/s41598-020-79875-x.

(250) Sisk-Hackworth L, Ortiz-Velez A, Reed MB, Kelley ST. Compositional data analysis of periodontal disease microbial communities. Front Microbiol. 2021 May;12:617949. doi: 10.3389/fmicb.2021.617949.

(251) O'Leary TJ, Drake RB, Naylor JE. The plaque control record. J Periodontol. 1972 Jan;43(1):38.

(252) Ainamo J, Bay I. Problems and proposals for recording gingivitis and plaque. Int Dent J. 1975 Dec;25(4):229-35.

(253) Zhou J, Jiang N, Wang S, Hu X, Jiao K, He X, et al. Exploration of human salivary microbiomes--insights into the novel characteristics of microbial community structure in caries and caries-free subjects. PLoS One. 2016 Jan;11(1):e0147039. doi: 10.1371/journal.pone.0147039.

(254) Yang F, Zeng X, Ning K, Liu KL, Lo CC, Wang W, et al. Saliva microbiomes distinguish caries-active from healthy human populations. ISME J. 2012 Jan;6(1):1-10.

(255) Zhou J, Jiang N, Wang Z, Li L, Zhang J, Ma R, et al. Influences of pH and iron concentration on the salivary microbiome in individual humans with and without caries. Appl Environ Microbiol. 2017 Feb;83(4):10.1128/AEM.02412-16.

(256) Rupf S, Laczny CC, Galata V, Backes C, Keller A, Umanskaya N, et al. Comparison of initial oral microbiomes of young adults with and without cavitated dentin caries lesions using an in situ biofilm model. Sci Rep. 2018 Sep;8(1):14010. doi: 10.1038/s41598-018-32361-x.

(257) Chen C, Hemme C, Beleno J, Shi ZJ, Ning D, Qin Y, et al. Oral microbiota of periodontal health and disease and their changes after nonsurgical periodontal therapy. ISME J. 2018 May;12(5):1210-24.

(258) Damgaard C, Danielsen AK, Enevold C, Massarenti L, Nielsen CH, Holmstrup P, et al. Porphyromonas gingivalis in saliva associates with chronic and aggressive periodontitis. J Oral Microbiol. 2019 Aug;11(1):1653123. doi: 10.1080/20002297.2019.1653123.

(259) Takeshita T, Kageyama S, Furuta M, Tsuboi H, Takeuchi K, Shibata Y, et al. Bacterial diversity in saliva and oral health-related conditions: the Hisayama study. Sci Rep. 2016 Feb;6:22164. doi: 10.1038/srep22164.

(260) Huse SM, Ye Y, Zhou Y, Fodor AA. A core human microbiome as viewed through 16S rRNA sequence clusters. PLoS One. 2012;7(6):e34242. doi: 10.1371/journal.pone.0034242.

(261) Ling Z, Liu X, Wang Y, Li L, Xiang C. Pyrosequencing analysis of the salivary microbiota of healthy Chinese children and adults. Microb Ecol. 2013 Feb;65(2):487-95.

(262) De Filippis F, Vannini L, La Storia A, Laghi L, Piombino P, Stellato G, et al. The same microbiota and a potentially discriminant metabolome in the saliva of omnivore, ovo-lacto-vegetarian and vegan individuals. PLoS One. 2014 Nov;9(11):e112373. doi: 10.1371/journal.pone.0112373.

(263) Chen T, Shi Y, Wang X, Wang X, Meng F, Yang S, et al. Highthroughput sequencing analyses of oral microbial diversity in healthy people and patients with dental caries and periodontal disease. Mol Med Rep. 2017 Jul;16(1):127-32.

(264) Ding T, Schloss PD. Dynamics and associations of microbial community types across the human body. Nature. 2014 May;509(7500):357-60.

(265) Gomar-Vercher S, Simon-Soro A, Montiel-Company JM, Almerich-Silla JM, Mira A. Stimulated and unstimulated saliva samples have significantly different bacterial profiles. PLoS One. 2018 Jun;13(6):e0198021. doi: 10.1371/journal.pone.0198021.

(266) Belstrom D, Paster BJ, Fiehn NE, Bardow A, Holmstrup P. Salivary bacterial fingerprints of established oral disease revealed by the human oral microbe identification using next generation sequencing (HOMINGS) technique. J Oral Microbiol. 2016 Jan;8:30170. doi: 10.3402/jom.v8.30170.

(267) Li Y, Feng X, Xu L, Zhang L, Lu R, Shi D, et al. Oral microbiome in Chinese patients with aggressive periodontitis and their family members. J Clin Periodontol. 2015 Nov;42(11):1015-23.

(268) Tezal M, Scannapieco FA, Wactawski-Wende J, Grossi SG, Genco RJ. Supragingival plaque may modify the effects of subgingival bacteria on attachment loss. J Periodontol. 2006 May;77(5):808-13.

(269) Martin FE, Nadkarni MA, Jacques NA, Hunter N. Quantitative microbiological study of human carious dentine by culture and real-time PCR: association of anaerobes with histopathological changes in chronic pulpitis. J Clin Microbiol. 2002 May;40(5):1698-1704.

(270) Shetty SA, Hugenholtz F, Lahti L, Smidt H, de Vos WM. Intestinal microbiome landscaping: insight in community assemblage and implications for microbial modulation strategies. FEMS Microbiol Rev. 2017 Mar;41(2):182-99.

(271) Robinson CK, Brotman RM, Ravel J. Intricacies of assessing the human microbiome in epidemiologic studies. Ann Epidemiol. 2016 May;26(5):311-21.

(272) de la Cuesta-Zuluaga J, Escobar JS. Considerations for optimizing microbiome analysis using a marker gene. Front Nutr. 2016 Aug;3:26. doi: 10.3389/fnut.2016.00026.

# JUSTIFICATION AND OBJECTIVES

# Justification and objectives

It is undeniable that the field of microbiology has seen enormous advances in the development and implementation of next-generation sequencing (NGS) technologies in recent decades (1). Specifically, sequencing of the 16S ribosomal RNA (rRNA) gene, which is regarded by many as the definitive phylogenetic marker due to its ubiquitous presence in bacteria and archaea and the intercalation of preserved and hypervariable zones (2); has enabled complex microbial communities like that in the oral environment to be studied in unprecedented detail (3).

Nevertheless, as discussed in the Introduction, an investigation's outcomes can be affected by multiple sources of possible bias during each step of the 16S rRNA gene sequencing workflow (4-6). The choice of primer is one such stage (4-6). The construction of primers is based on a consensus sequence, but they can be mismatched with some taxa, potentially leading to the over or underrepresentation of a particular microbial group (7). Consequently, using the wrong primer could lead to questionable biological conclusions about the niche under study (7). Nevertheless, there has been no exhaustive analysis of the coverage of the primers used to detect the prokaryotic microorganisms inhabiting the human mouth, i.e., the coverage percentage of matches for a particular group of sequences/taxonomic ranks.

In contrast, some types of bias are due to inherent limitations of the gene itself (8). Indeed, in a first example, various investigations have demonstrated the existence of multiple 16S rRNA gene copies in the prokaryotic genomes (9-14). This affects abundance estimates based on gene counts in such a manner that taxa with a low number of genes tend to be underestimated, while those with a high number are overestimated (9,12). Second, the nine hypervariable regions of the gene have different degrees of sequence heterogeneity (11,15). Furthermore, some distinct species may have matching amplicons (MAs), defined as those with 100% similarity and the same number of nucleotides. Other species, meanwhile, share highly similar sequences that even exceed the commonly used 97% threshold for constructing operational taxonomic units

(OTUs) (12,16), which means that they can be clustered erroneously in the same OTU. Such clustering affects the construction of OTU tables and, by extension, taxonomic assignments and findings on diversity. Nonetheless, despite the matters touched on above, there has been no exhaustive research on how to best evaluate the numbers of intragenomic 16S rRNA genes in the bacteria and archaea occupying the oral cavity, or on the impact that the primer chosen for the different regions has on the detection of either MAs or highly similar amplicons from distinct taxa. Furthermore, still requiring assessment is the issue of how many different oral taxa and which specific oral species can be grouped together erroneously in the same OTU, depending on the primer used.

The comparison of 16S rRNA gene sequencing-based studies on periodontal microbiome is controversial, due to significant methodological differences in the relevant steps within the typical workflow. Indeed, it is widely known that each sequencing technology performs differently in the trade-off between read length, sequence throughput, and the error rate (6), with Illumina generally preferred to Roche 454 or Ion Torrent (4). Moreover, as noted above, the different hypervariable regions and, as a consequence, the amplicons derived from them have discrepant levels of sequence heterogeneity (11,15). Accordingly, comparisons of oral-sample sequences obtained using distinct sequencing technologies and gene regions seems questionable. However, no research to date has evaluated the sequences obtained in studies on the periodontal microbiome present in different health conditions, in particular those generated using the Illumina high-throughput platform and distinguished by the most amplified 16S rRNA gene region.

Due to the lack of evidence on the issues highlighted above, this Thesis had the following objectives:

1) To analyse *in silico* the coverage of the primer pairs employed in sequencing-based studies of the oral microbiota; to be achieved using two oral-specific databases containing 16S rRNA gene sequences from bacterial and archaeal species.



2) To analyse *in silico* the number of 16S rRNA genes in the complete genomes of the bacterial and archaeal species inhabiting the human mouth. Moreover, to assess how

the use of different primer pairs targeting distinct regions affects the detection of MAs from different taxa, thereby identifying the oral species that have MAs.

3) To analyse *in silico* the performance of different primer pairs from distinct regions in order to identify distinct oral prokaryotic species with 16S rRNA gene amplicon similarity values ≥97%, thereby establishing the oral species that may be erroneously clustered in the same OTU.

4) To analyse the microbial community profiles in the supragingival and subgingival plaque of 2045 patients with different periodontal conditions (healthy periodontal, gingivitis, untreated and treated periodontitis) in relation to bacterial diversity, co-occurrence-network patterns, and predictive models; the sequences used are from the Illumina platform, have a focus on region 3-4, and are treated with the same bioinformatics protocol.

**REFERENCES**

(1) A genomic approach to microbiology. Nat Rev Genet. 2019 Jun;20(6):311. doi: 10.1038/s41576-019-0131-5.

(2) del Rosario-Rodicio M, del Carmen-Mendoza M. Bacterial identification by 16S rRNA sequencing: rationale, methodology and applications in clinical microbiology. Enferm Infecc Microbiol Clin. 2004 Apr;22(4):238-45. Spanish.

(3) Zaura E. Next-generation sequencing approaches to understanding the oral microbiome. Adv Dent Res. 2012 Sep;24(2):81-5.

(4) Nearing JT, Comeau AM, Langille MGI. Identifying biases and their potential solutions in human microbiome studies. Microbiome. 2021 May;9(1):113. doi: 10.1186/s40168-021-01059-0.

(5) Robinson CK, Brotman RM, Ravel J. Intricacies of assessing the human microbiome in epidemiologic studies. Ann Epidemiol. 2016 May;26(5):311-21.

(6) de la Cuesta-Zuluaga J, Escobar JS. Considerations for optimizing microbiome analysis using a marker gene. Front Nutr. 2016 Aug;3:26. doi: 10.3389/fnut.2016.00026.

(7) Hamady M, Knight R. Microbial community profiling for human microbiome projects: tools, techniques, and challenges. Genome Res. 2009 Jul;19(7):1141-52.

(8) Rajendhran J, Gunasekaran P. Microbial phylogeny and diversity: small subunit ribosomal RNA sequence analysis and beyond. Microbiol Res. 2011 Feb;166(2):99-110.

(9) Acinas SG, Marcelino LA, Klepac-Ceraj V, Polz MF. Divergence and redundancy of 16S rRNA sequences in genomes with multiple rrn operons. J Bacteriol. 2004 May;186(9):2629-35.

(10) Pei AY, Oberdorf WE, Nossa CW, Agarwal A, Chokshi P, Gerz EA, et al. Diversity of 16S rRNA genes within individual prokaryotic genomes. Appl Environ Microbiol. 2010 Jun;76(12):3886-97.

(11) Sun D, Jiang X, Wu QL, Zhou N. Intragenomic heterogeneity of 16S rRNA genes causes overestimation of prokaryotic diversity. Appl Environ Microbiol. 2013 Oct;79(19):5962-9.

(12) Větrovský T, Baldrian P. The variability of the 16S rRNA gene in bacterial genomes and its consequences for bacterial community analyses. PLoS One. 2013;8(2):e57923. doi: 10.1371/journal.pone.0057923.

(13) Case RJ, Boucher Y, Dahllöf I, Holmström C, Doolittle WF, Kjelleberg S. Use of 16S rRNA and rpoB genes as molecular markers for microbial ecology studies. Appl Environ Microbiol. 2007 Jan;73(1):278-88.

(14) Lee ZM, Bussema C 3rd, Schmidt TM. rrnDB: documenting the number of rRNA and tRNA genes in bacteria and archaea. Nucleic Acids Res. 2009 Jan;37(Database issue):D489-93. doi: 10.1093/nar/gkn689.

(15) Johnson JS, Spakowicz DJ, Hong BY, Petersen LM, Demkowicz P, Chen L, et al. Evaluation of 16S rRNA gene sequencing for species and strain-level microbiome analysis. Nat Commun. 2019 Nov;10(1):5029. doi: 10.1038/s41467-019-13036-1.

(16) Schloss PD. Amplicon sequence variants artificially split bacterial genomes into separate clusters. mSphere. 2021 Aug;6(4):e0019121. doi: 10.1128/mSphere.00191-21.

# OBJECTIVE 1

# Objective 1. *In silico* evaluation and selection of the best 16S rRNA gene primers for use in next-generation sequencing to detect oral bacteria and archaea

## 1.1 Abstract

**Aim:** To analyse *in silico* the coverage of the 16S rRNA gene primers used to study the composition of the oral microbiota, using two databases containing 16S rRNA sequences from oral bacteria and archaea; and to describe the best primer pairs for each domain.

**Material and methods:** Searches were conducted in PubMed to create a list of 1) 16S rRNA gene primers used in sequencing-based studies of the oral microbiome, and 2) oral-archaea species inhabiting the human mouth. The individual primers found were evaluated against a previously-reported database of 16S rRNA sequences from oral bacteria, which was modified by our group; and a self-created oral-archaea database, constructed based on the list of oral-archaeal species. Both databases contained the genomic variants detected for each included species. Primers were evaluated at the variant and species levels, and those with a species coverage (SC) ≥75.00% were selected for the pair analyses. All possible combinations of forward and reverse primers were identified and evaluated against the two databases.

**Results:** A total of 369 distinct individual primers were found in the literature. After applying the primer-pair formation criteria, 4638 primer pairs were identified. The best bacteria-specific pairs targeted the 3-4, 4-7, and 3-7 16S rRNA gene regions, with SC levels of 98.83% - 97.14%; meanwhile, the optimum archaea-specific primer pairs amplified regions 5-6, 3-6, and 3-6, with SC estimates of 95.88%. Finally, the best pairs for detecting both domains targeted regions 4-5, 3-5, and 5-9, and produced SC values of 95.71% - 94.54% and 99.48% - 96.91% for bacteria and archaea, respectively.

**Conclusions:** Given the three amplicon length categories (100-300, 301-600, and >600 base pairs), the primer pairs with the best coverage values for detecting oral bacteria were:

KP_F048-OP_R043 (region 3-4; primer pair position for *Escherichia coli* J01859.1: 342-529), KP_F051-OP_R030 (4-7; 514-1079), and KP_F048-OP_R030 (3-7; 342-1079). For detecting oral archaea, these were: OP_F066-KP_R013 (5-6; 784-undefined), KP_F020-KP_R013 (3-6; 518-undefined), and OP_F114-KP_R013 (3-6; 340-undefined). Lastly, for detecting both domains they were: KP_F020-KP_R032 (4-5; 518-801), OP_F114-KP_R031 (3-5; 340-801), and OP_F066-OP_R121 (5-9; 784-1405). The primer pairs with the best coverage identified herein are not among those described most widely in the oral microbiome literature.

### 1.1.1 Keywords

16S rRNA gene, primer, coverage, mouth, bacteria, archaea, database.

### 1.1.2 Declaration of conflict of interest

The doctoral candidate and the rest of the authors of the present study declare that they have no conflict of interest concerning the objectives proposed in this chapter.

### 1.1.3 Funding

This investigation was supported by the Instituto de Salud Carlos III (General Division of Evaluation and Research Promotion, Madrid, Spain) and co-financed by the European Regional Development Fund (ERDF) (“A way of making Europe”) under grant ISCIII/PI17/01722; Consellería de Cultura, Educación e Ordenación Universitaria (accreditation 2019-2022 ED431G-2019/04, group with growth potential ED431B 2020-2022 GPC2020/27; A. Regueira-Iglesias support ED481A-2017/233) and the ERDF, which acknowledges the CiTIUS-Research Center in Intelligent Technologies of the Universidade de Santiago de Compostela as a Research Center of the Galician University System.

The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

## 1.2 INTRODUCTION

The oral microbiota is the second largest and diverse in the human body, containing over 700 microbial species (1). It plays a critcal role in the onset and development of two of the most prevalent diseases in mankind: dental caries and periodontitis; both of which, if left untreated, can lead to tooth loss, edentulism, loss of masticatory function, poor nutrition status, loss of self-esteem, social difficulties and diminished quality of life (2,3). What is more, there is a body of evidence on the association between oral microorganisms and several systemic diseases (4).

The advent of high-throughput next-generation sequencing (NGS) technologies has enabled the characterisation of microbiota to unprecedented depths that are unachievable with previous methods (5). These revolutionary techniques enable large-scale projects to be completed in just a few days, or sometimes even hours (6). The NGS employed most at present - Illumina - can generate sequences with up to 2x300 base pairs (bps) (7). The NGS of the 16S ribosomal RNA (rRNA) marker-gene amplicons has been widely used to study the oral microbiota (8,9), allowing the detection of several bacterial and archaeal taxa in both the healthy human mouth and ones with various states of disease (10). Continuous improvements to the process have recently produced “third-generation sequencing” tools like those from Pacific Biosciences (PacBio) or Nanopore sequencing. These technologies have the objective of generating longer primary read lengths (600-1000 bps) or even the full-length sequence of the 16S rRNA gene (7).

Further advances in high-throughput sequencing have allowed the development of whole-genome shotgun (WGS) sequencing, which characterises genomes, genes, and genetic features in a sample. Although this technique has several advantages when compared to the 16S rRNA gene sequencing, the latter remains to be widely used in the oral microbiology field mainly due to the rapid processing, the simplicity in analysing the results, and the lower cost (11).

Nevertheless, marker-gene sequencing approaches are also not without shortcomings, with different challenges and pitfalls possible during each step of the gene-sequencing workflow (12). The primer chosen for the polymerase chain reaction (PCR) amplification step can greatly affect the diversity of an investigation’s findings (12,13). To amplify a 16S rRNA gene region of interest, “broad-range” (or universal) primers are designed to anneal with the conserved

regions flanking the hypervariable zone selected (13). Although these primers are based on a consensus sequence, some taxa can produce mismatches (12). Primer bias due to differential annealing can lead to the over- or under-representation of a particular microbial group and, occasionally, even the loss of some groups if there is a poor match with the consensus sequence (14). As a consequence, using an inadequate primer can lead to questionable biological conclusions (14).

If PCR results in microbial research are to be interpreted satisfactorily, conducting a comprehensive evaluation of a primer's coverage is essential (15). The concept of coverage has been defined heterogeneously as: the percentage of matches for certain taxonomic ranks (16,17); the number of sequences matched by at least one primer (18); or the proportion of species-level taxonomic entries for each phylum in a database where the prediction is that these will be amplified using a particular primer pair (19). The literature contains *in silico* research that analyses the coverage of 16S rRNA gene-targeting primers that are suitable for amplicon sequencing (15-25). These studies aim to identify the optimum primer pair(s) for sequencing the environmental (16,17,20,21), human (18,19,22-24), or combined environmental and human microbiota (15,25). For a few of them, the human mouth was an ecosystem of interest (19,22,23), and researchers employed non-oral-specific databases such as Silva (26) or a foregut dataset together with the ribosomal database project (RDP) database (27) for the coverage analysis. The application of phylogenetically diverse databases can, however, produce classification errors since they contain taxonomically misannotated 16S rRNA gene sequences (28). They also provide different levels of representation for each included environment, leading to substantial variations in the quality of the classifications (29).

Despite the above, to the best of our knowledge, there has been no exhaustive *in silico* evaluation of the coverage provided by the 16S rRNA gene primers employed in the massive sequencing of mouth specimens using oral-specific databases. Consequently, we aimed to investigate the coverage of primer pairs obtained from examinations of different oral niches in diverse health conditions and ecology studies. To this end, we used two databases containing 16S rRNA gene sequences taken from bacterial and archaeal species found in the human mouth.

## 1.3 MATERIAL AND METHODS

### 1.3.1 Computational search of scientific papers in PubMed and analysis of abstracts using text-mining techniques

We conducted systematic searches of articles in the PubMed database using the R statistical software (version 4.0.3) (30) and the RISmed package (version 2.1.7) (31). Two searches were conducted for two different purposes: 1) making a list of the 16S rRNA genes primers used to detect and amplify bacteria and archaea in oral samples before massive sequencing; and 2) making a list of the archaeal species reported to be inhabitants of the human mouth to create a database of oral-archaea 16S rRNA gene sequences. The groups of words employed in these searches can be found in appendices S1 and S2.

Text-mining techniques were applied to all the downloaded abstracts using the R package tm (version 0.7-7) (32). Specifically, the abstracts were tokenised, which involved the classification of the words and groups of two or three words contained within them. For purpose 1, publications on the study of bacterial microbiome received a score if their abstracts included terms associated with the oral cavity, and another if they contained terms related to the 16S rRNA gene and its different regions. A further score based on archaea-associated words was used for the articles about the oral archaeome. For purpose 2, the studies identified in the searches seeking to uncover oral-archaeal species were rated and assigned an oral and an archaeal word score. The terms used to calculate the scores were the same as those used in the searches. Repeated words were counted only once, meaning that articles with higher scores were purely those with a greater diversity of words. Ultimately, we were left with 129 bacterial and 16 archaeal studies that included the use of at least one different 16S rRNA gene primer, and 53 articles containing information on archaeal species (Figure 1). The references of all these papers are included in appendices S3 and S4.

| Purpose 1. To find 16S rRNA gene primers used to identify bacteria or archaea | | | |
|---|---|---|---|
| STEP | DESCRIPTION | BACTERIA | ARCHAEA |
| 1 | No. of computational searches in PubMed performed: | 2940 | 5796 |
| 2 | No. of abstracts and metadata of papers downloaded: | 3245 | 6405 |
| 3 | No. of papers processed by text mining techniques: | 2939 | 1687 |
| 4 | No. of papers with oral score ≥1 and gene score ≥3 (partial reading): | 576 | 44 |
| 5 | No. of papers reviewed for full text reading: | 323+15* | 22+12* |
| 6 | No. of papers with at least one different 16S rRNA gene primer: | 129 | 16 |

| Purpose 2. To create a list of oral archaea species | | |
|---|---|---|
| STEP | DESCRIPTION | ARCHAEA |
| 1 | No. of computational searches in PubMed performed: | 276 |
| 2 | No. of abstracts and metadata of papers downloaded: | 7548 |
| 3 | No. of papers processed by text mining techniques: | 6734 |
| 4 | No. of papers with oral score ≥1 and archaea score ≥3 (partial reading): | 200 |
| 5 | No. of papers reviewed for full text reading: | 60 |
| 6 | No. of papers with at least one oral archaea species: | 53 |

**Figure 1. Flowchart on the computational search of articles in PubMed and their analysis using text-mining techniques.**

Papers from "purpose 1" received one score for the oral cavity words included in their abstracts and another for the terms associated with the 16S rRNA gene and its different regions; papers from "purpose 2" received an oral- and an archaeal-word score. In each score, for each different related term included in the abstract, we gave one point with repeated words only counted once (i.e.: in a given abstract, the words "oral", "mouth" and "periodontitis" appear two, one and three times so the oral cavity score is equal to three). The terms used to give the punctuations were those used to conduct the searches. *Additional publications on the study of the oral microbiome using sequencing were considered for full-text reading; these were previously reviewed for other reasons (n= 15) or were found during the search for the oral-archaea species (n= 12).

### 1.3.2 Primer selection and creating a list of archaeal species found in the oral cavity

In total, we identified 444 16S rRNA gene primers: 204 forward (F), 230 reverse (R), and 12 unidentified (UI). Two hundred and seventy-eight of the primers were procured from the searches on PubMed, being 238 and 37 used for the detection of oral bacteria or archaea, respectively, and three to identify bacteria in the respiratory ecosystem. The remaining 166 were extracted from articles concerning different niches, mainly ecological, described in Klindworth et al. (16). Of them, 103 corresponded to the bacteria domain, 42 to the archaea domain, and 21 were universal. All 444 primers were assigned a unique identifier based on where they were sourced -"OP" for oral primer and "KP" for Klindworth primer (16)- and their direction (F, R, or UI), followed by a three-digit number (Appendix S5). The 5'-3' sequences of all 444 primers were then compared to identify repeats, with 75 identified as having the same sequences (Appendix S5). This left us with 369 16S rRNA gene primers with different sequences (with at least one nucleotide difference).



The publications in our final selection were read by two researchers to produce a list of archaeal species found in the human mouth. This gave us 177 different archaea names at the species level (Appendix S6).

### 1.3.3 16S rRNA gene-sequence databases of oral bacteria and archaea for the primer-coverage analysis

#### 1.3.3.1 Modification of an existing 16S rRNA gene-sequence database of oral bacteria

A total of 223,143 amplicon sequence variants (ASVs) of fasta-formatted 16S rRNA gene sequences were included in the Escapa et al. database (33). The file had been constructed using sequences from the expanded human oral microbiome database (eHOMD) (34) to then conduct a BLASTN search (35,36) of the National Centre for Biotechnology Information (NCBI) non-redundant nucleotide database (37). The header line of each sequence had an ASV identifier (from TS000001 to TS223143), followed by a RefSeq (38) or GenBank (39) identifier and an assignment to a seven-level taxonomic hierarchy. The format was as indicated on the DADA2 website (40). The sequences in the Escapa et al. database (33) were obtained mainly from GenBank (39), and we found that they contain annotation errors that make it impossible to calculate the correct position of the primers within each sequence in the case of a match.

We developed scripts in Python (version 3.9.0) (41) and Bash (version 5.1) (42) to improve the Escapa et al. database (33). First, we separated the 16S rRNA gene sequences from ASVs belonging to the same hierarchical level into 769 different fasta files. Second, a species identifier, from SP00001 to SP00769, was attached to all the sequences before the taxonomic hierarchy. Third, sequences from the same hierarchy were aligned simultaneously using Clustal Omega (43) against a set of 16S rRNA gene sequences of *Escherichia coli*: three from GenBank (39) and one from HOMD (44). We installed Clustal Omega (43) in the local mode (45) to enable its use with Biopython (46). The default characteristics were employed to carry out the alignments. Fourth, all the gaps created by Clustal Omega (43) were removed, save for those inserted from the start up to the first nucleotide of each sequence. Fifth, the aligned fasta files were combined in a single file to create a database of fully-aligned *E. coli* ASVs, with position one being the first nucleotide of *E. coli* J01859.1. Lastly, we trimmed the aligned sequences with bases in a lower position than the first nucleotide of J01859.1, as well as those with nucleotides above position 2000 (Figure 2). The resultant oral-bacteria database is available for consultation in appendix S7.

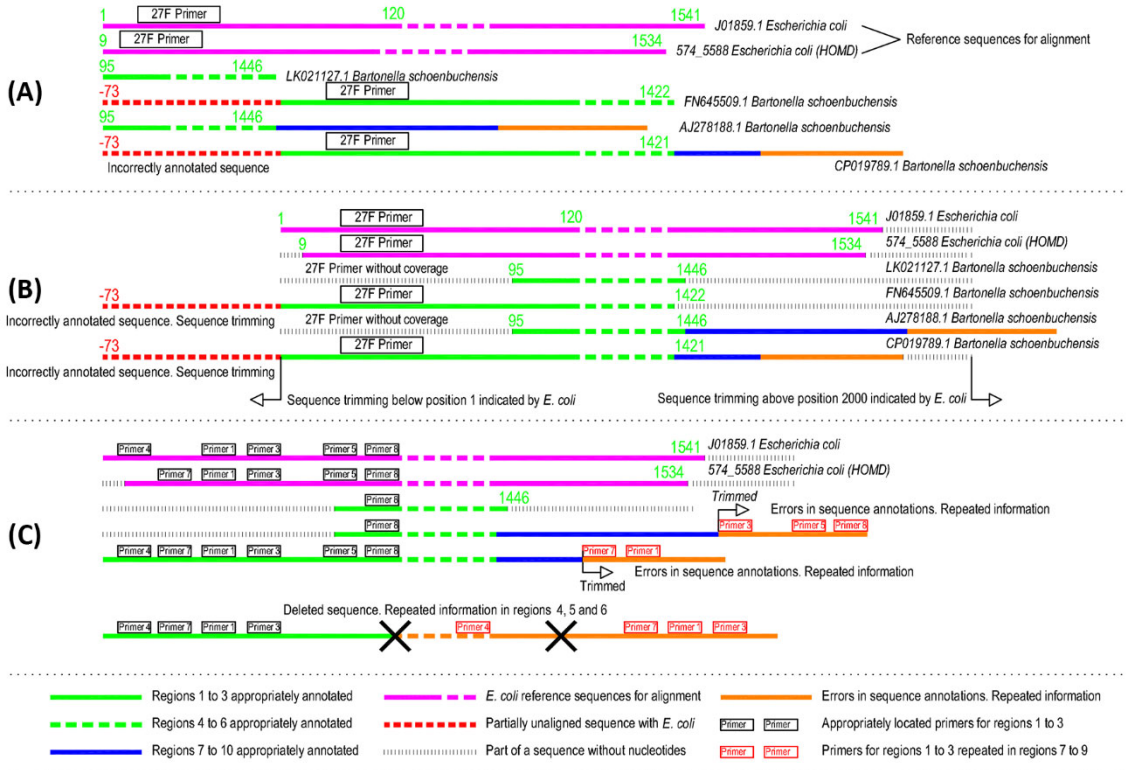



**Figure 2. Processing of errors in annotations of oral-bacterial and archaeal sequences.**

(A) Unaligned sequences with missing information at the first and last positions of the 16S rRNA gene, and the presence of redundant information. (B) Alignment of sequences with respect to *Escherichia coli* and trimming of sequences below position 1 and above position 2000 indicated by *Escherichia coli*. (C) Trimming of sequences with redundant information in high regions; and removal of a sequence with repeated information in regions 4, 5, and 6.

#### 1.3.3.2 Creation of a 16S rRNA gene-sequence database of archaea

We searched the NCBI nucleotide database (37) for the complete genomes of the archaeal species found in the human mouth. Along with a script developed in Python (41), these identifiers enabled us to download 193 genomes from RefSeq (38) and eight from GenBank (39).

The script was completed using a free downloadable module, search_16S.py (47), which is based on the algorithm created by Edgar (48). This allowed us to detect and extract the 16S rRNA gene sequences from the complete downloaded genomes, remove all the repeated sequences, and then store all the variants identified in a fasta file. Prior to use, the search_16S.py algorithm was trained with a RefSeq database containing 16S rRNA gene sequences of archaea stored in the NCBI database (49). The module and integrating the "Entrez Programming Utilities (E-utilities)" tool (50) into Biopython (46) meant we could easily and automatically

obtain and assign the complete taxonomic rank to the 16S rRNA genes. Biopython (46) also enabled us to access the information of interest requested from the different NCBI databases, such as Taxonomy (51), RefSeq (38), and Genbank (39).

Additionally, the 16S rRNA gene sequences of species without complete genome identifiers in RefSeq (38) or GenBank (39) were searched for in the aforementioned RefSeq archaeal database or, if not found, the Silva database (version 138) (26) or the Genome Taxonomy Database (GTDB) (52). Finally, all the 16S rRNA gene sequences of the oral-archaeal species were grouped into a single fasta file (Appendix S8).

These sequences were employed to BLASTN (35,36) against the NCBI non-redundant nucleotide database (37). Then, we downloaded the 16S rRNA gene sequences with a query coverage ≥98% and a percentage identity ≥99%. The regions aligned with complete genomes were also downloaded using these parameters. Both sequence types were treated as ASVs. We created the oral-archaea database using another script developed in Python (41). This contains 2842 sequences and all the ASVs presenting with a unique identifier with values between AS00001 and AS002842 (Appendix S9). The sequences in the database were aligned in relation to *E. coli* and were improved for the posterior-coverage analysis following the same steps used for the bacteria database (Figure 2). The definitive oral-archaea database is available for consultation in appendix S10.

### 1.3.4 Coverage ratios of the 16S rRNA gene primers

1.3.4.1 Concept and definition of the coverage ratios calculated for the 16S rRNA gene primers

A sequence was considered covered by a primer when all nucleotides of the primer showed a match with the sequence (mismatches were not allowed). Two types of coverage were defined for the *in silico* analysis. First, the coverage at the variant level (VC) equated to the percentage of matches of a particular primer in relation to the total sequences in the database. In order to minimise the effect on the VC of the absence of information at the ends of sequences, the concept of species-level coverage (SC) was defined as the percentage of species with matches in at least one of its sequence variants when a particular primer is used.

Matches between the analysed primers and sequences in the databases were evaluated by applying the regular expressions of Python's regex module (53). The results were then stored in the Excel format with xlsxwriter (54), which is a Python (41) package that allows the creation and formatting of xlsx files.

#### 1.3.4.2 Selection of primer pairs and analysis of their coverage

All the information related to the coverage analysis of individual primers is included in appendix S11.

The individual primers with an SC ≥75.00% were chosen in this stage of the research and all the possible combinations between F and R were identified. We then estimated the mean length between the two positions using the mean position of the first nucleotide of the F primer and that of the last nucleotide of the R primer. The primer pairs had to fulfil two conditions: 1) the mean position of the F primer's first nucleotide had to be lower than that of the R primer's last; and 2) the minimum distance between the two means had to be ≥100 nucleotides. The calculated average length was used to classify the primer pairs into one of three categories relating to the mean amplicon lengths: 1) Short (S): 100 to 300 bps; 2) Medium (M): 301 to 600 bps; and 3) Long (L): more than 600 bps.

Primer pairs obtained were evaluated against both the bacteria and archaea databases to calculate the coverage parameters defined above. This step enabled us to determine whether a primer pair was bacteria-specific, archaea-specific, or suitable for both domains. A primer pair was assigned the concept of "specific" for bacteria when it had an SC value of 0.00% for archaea, and it was "specific" for archaea when it had an SC value of 0.00% for bacteria. A primer pair was considered as "non-specific" if it showed SC values >0.00% in both domains.

Taxa covered and not covered by the different primer pairs evaluated were described, the latter being those with 16S rRNA sequences showing at least one mismatch with the tested primer pair.

## 1.4 RESULTS

Of the 369 individual primers, 178 (103 F, 75 R) and 50 (33 F, 17 R) showed some coverage value for only bacteria or only archaea, respectively. One hundred and twenty-four (30 F, 94 R) were able to detect both oral-bacterial and archaeal species, while 17 (9 F, 8 R) were not able to detect any such organisms.

The metrics obtained using the two databases for individual primers as well as all the possible combinations of primer pairs are included in appendices S12-S14. The bacterial and archaeal SC values were ≥75.00% in 148 (67 F, 81 R) and 65 (19 F, 46 R) individual primers, respectively. After applying the primer-pair formation criteria, 3993 bacterial and 645 archaeal combinations were possible. Of these, 156 were repeated primer pairs in both domains, and the rest (i.e., 3837 and 489) were obtained exclusively when searching for bacterial or archaeal primer pairs.

### 1.4.1 Evaluation of 16S rRNA gene primer pairs for the detection of oral bacteria, archaea, and both domains

Results obtained in the analysis of the individual primers are included in appendix S11.

#### 1.4.1.1 Bacteria-specific primer pairs

The pair's analysis revealed that 3218 of the 3837 bacteria-specific primer-pair candidates had an archaeal VC and SC of 0.00%. On the other hand, 619 had some coverage value for oral archaea (archaea VC range= 52.25% - 0.04%; archaea SC range= 70.62% - 0.52%). Relative to the mean lengths of the generated amplicons: 840 primer combinations had bps of 100 to 300; 1374 from 301 to 600; and 1623 more than 600.

In the short mean amplicon length category, 139 pairs had bacterial SC values ≥95.00% (bacterial SC range= 99.09% - 95.19%), while 33 also had an archaeal SC of 0.00%. The latter were used to amplify gene regions 3-4 or 5-7 and had bacterial SC values ranging from 97.92% to 95.58%, which meant that 16 to 34 oral-bacterial species were not covered. For most of these, the mean read length of their amplicons was around 186 (range= 189 - 182). However, the pair OP_F009-OP_R030 from region 5-7 stood out, with a mean read length of 297 and a bacterial

SC value of 96.88%, with only 24 bacterial species from the oral cavity were not covered by this pairing.

Sixty-eight primer pairs in the medium mean amplicon length category had bacterial SC values ≥95.00% (range= 98.83% - 95.06%). Of these, 45 did not amplify any archaeal species and could therefore be treated as bacteria-specific. Their bacterial SC values also ranged from 98.83% to 95.06%, meaning that between nine and 38 species were not covered. In addition, these pairs targeted gene regions 3-5, 3-6, or 4-7, and had maximum (max.) and minimum (min.) mean read lengths of 566 and 454, respectively. Of those with the longest mean amplicon lengths, the pairs providing the best coverage were, in order: KP_F051-OP_R030; OP_F021-OP_R030; KP_F048-OP_R073; KP_F051-KP_R053; OP_F021-KP_R053; and OP_F050-OP_R073 (bacterial SC range= 98.83% - 96.23%; mean read length range= 566 - 546). These pairs, which amplified regions 3-6 or 4-7, did not cover between nine and 29 oral-bacteria species.

Lastly, 20 primer pairs with mean amplicon lengths >600 bps had bacterial SC values ≥95.00% (range= 97.14% - 95.06%), while 17 also had an archaeal SC value of 0.00%. These pairs had the same bacterial SC range and left between 22 and 38 species uncovered. All of them targeted gene region 3-7 and had max. and min. mean read lengths of 772 and 732, respectively. The primers with the best balance between the mean read length and the coverage were: KP_F048-KP_R074 (bacterial SC = 97.01%; mean read length= 767); and OP_F050-KP_R074 (bacterial SC= 96.36%; mean read length= 766). There were, however, interesting options for the bacteria-specific pairs with mean amplicon lengths >1000 bps and bacterial SC values ≥90.00% (bacterial SC range= 93.37% - 90.64%; mean read length range= 1066 - 1059). In this sense, the pairs KP_F048-KP_R060, KP_F048-KP_R076, and KP_F048-OP_R121 from region 3-9 had mean read lengths of 1061, 1060, and 1060, respectively, and bacterial SC values of 93.37%; these pairings left a total of 51 oral-bacteria species uncovered.

For each amplicon-length category, we selected at least one primer pair suitable for detecting only bacteria (archaeal SC= 0.00%) and which targeted distinct 16S rRNA gene regions (Table 1). The pairs had to have a bacterial SC ≥90.00% and were chosen based on their

coverage and mean amplicon lengths. The VC results of these selected primers are detailed in appendix S15.

**Table 1. Selected primer pairs for detecting oral bacteria in different amplicon-length categories.**

| | | Bacteria | | | | | Archaea | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| ALC | Primer pair | Gene region | SC (%) | Covered | Not covered | Mean length | Gene region | SC (%) | Covered | Not covered | Mean length |
| S | KP_F048-OP_R043 | 3-4 | 97.92 | 753 | 16 | 183 | - | 0.00 | 0 | 194 | 0 |
| | OP_F098-OP_R119 | 4-5 | 94.54 | 727 | 42 | 289 | - | 0.00 | 0 | 194 | 0 |
| | OP_F066-KP_R040 | 5-6 | 90.25 | 694 | 75 | 142 | - | 0.00 | 0 | 194 | 0 |
| | OP_F009-OP_R030 | 5-7 | 96.88 | 745 | 24 | 297 | - | 0.00 | 0 | 194 | 0 |
| | OP_F101-OP_R030 | 6-7 | 93.63 | 720 | 49 | 164 | - | 0.00 | 0 | 194 | 0 |
| | KP_F061-KP_R074 | 6-7 | 91.94 | 707 | 62 | 206 | - | 0.00 | 0 | 194 | 0 |
| M | KP_F048-KP_R031 | 3-5 | 97.53 | 750 | 19 | 455 | - | 0.00 | 0 | 194 | 0 |
| | KP_F048-OP_R073 | 3-6 | 96.88 | 745 | 24 | 547 | - | 0.00 | 0 | 194 | 0 |
| | KP_F048-OP_R050 | 3-6 | 90.25 | 694 | 75 | 579 | - | 0.00 | 0 | 194 | 0 |
| | KP_F051-KP_R041 | 4-6 | 90.77 | 698 | 71 | 411 | - | 0.00 | 0 | 194 | 0 |
| | KP_F051-OP_R030 | 4-7 | 98.83 | 760 | 9 | 566 | - | 0.00 | 0 | 194 | 0 |
| | OP_F116-KP_R060 | 7-9 | 94.02 | 723 | 46 | 308 | - | 0.00 | 0 | 194 | 0 |
| L | KP_F048-OP_R030 | 3-7 | 97.14 | 747 | 22 | 733 | - | 0.00 | 0 | 194 | 0 |
| | KP_F048-KP_R074 | 3-7 | 97.01 | 746 | 23 | 767 | - | 0.00 | 0 | 194 | 0 |
| | KP_F048-KP_R060 | 3-9 | 93.37 | 718 | 51 | 1061 | - | 0.00 | 0 | 194 | 0 |
| | KP_F056-KP_R077 | 4-9 | 91.94 | 707 | 62 | 845 | - | 0.00 | 0 | 194 | 0 |

Species coverage was estimated as the number of species with at least one match in an ASV divided by the number of species included in the database. Our bacterial and archaeal databases contained 769 and 194 species, respectively, each of which had between one and 4000 ASVs. The location of the first and last nucleotides of each primer within each sequence with a match was calculated and the mode values for these positions were determined. If there was more than one mode for a position, we chose the one closest to the mean position value. As all the sequences in the two databases were aligned with the 16S rRNA *Escherichia coli* gene, the mode values obtained for each primer enabled us to allocate them to one of the gene regions defined for that organism by Baker et al. (55). The reference sequence utilised had 1542 base pairs distributed in 10 conserved (C1-C10) and nine hypervariable regions (V1-V9). The sequences of these selected primer pairs are described in appendix S16.
ALC= amplicon length category; F= forward; KP= Klindworth primer; L= long mean amplicon length category, >600 base pairs; M= medium mean amplicon length category, 301-600 base pairs; OP= oral primer; R= reverse; S= short mean amplicon length category,100-300 base pairs; SC= coverage at the species level.

Appendix S17 includes the oral species not detected by the primer pairs that were found to achieve a bacterial SC ≥95.00% and an archaeal SC= 0.00%, as well as those named previously or included in table 1 that produced a bacterial SC ≥90.00% and an archaeal SC= 0.00%.

1.4.1.2 Archaea-specific primer pairs

Of the 489 primer pairs that were specifically archaea-domain candidates, 359 simultaneously had a bacterial VC and an SC of 0.00%. Conversely, 130 had some coverage value for oral bacteria (bacterial VC range= 9.98% - 0.01%; bacterial SC range= 74.64% - 0.13%). Classification of all the pairs based on their mean amplicon lengths revealed that: 77 had 100 to 300 bps; 209 had 301 to 600; and 203 had more than 600.

Twelve primer pairs in the 100-300 bps category had archaeal SC values ≥95.00% (range= 98.45% - 95.36%). Of these, eight had bacterial SC values of 0.00% and should therefore be defined as archaea-specific: OP_F066-KP_R013; KP_F059-KP_R013; KP_F016-KP_R002; KP_F018-KP_R003; OP_F066-KP_R006; KP_F018-OP_R102; KP_F059-KP_R006; and KP_F018-KP_R002. Their archaeal SC ranged from 95.88% to 95.36%, their max. and min. read lengths from 275 to 144, and they were employed to amplify gene regions 3 or 5-6. The use of these pairs would leave between eight and nine oral-archaeal species uncovered.

Nineteen primer pairs in the medium mean amplicon length category had archaeal SC values ≥95.00% (archaeal SC range= 97.42% - 95.36%). Among these, nine also had a bacterial SC value of 0.00%: KP_F018-KP_R031; KP_F018-KP_R032; KP_F018-KP_R035; KP_F018-OP_R020; KP_F018-OP_R070; KP_F020-KP_R006; KP_F020-KP_R013; KP_F016-KP_R032; and OP_F114-KP_R006. These targeted gene regions 3-5 or 3-6 and had mean amplicon lengths of 551 to 414 bps. The pairs covered 95.88% to 95.36% of the oral-archaea species in our database, leaving between eight and nine uncovered.

Only one primer pair in the >600 bps category had an SC value ≥95.00% in the archaea database: OP_F114-KP_R013. Interestingly, it also had a bacterial SC value of 0.00%. This pair was used to amplify gene region 3-6, had a mean length of 679 bps, and left eight archaeal species uncovered. We obtained 27 pairs of primer combinations with an archaeal SC ≥90.00%, a bacterial SC of 0.00%, and a mean length >679 bps, 10 of which were longer than 1100 bps (max. mean length= 1131; min. mean length= 681). Of these, the best balance between coverage and the mean amplicon length was found in: KP_F016-KP_R066; KP_F016-KP_R063; KP_F018-KP_R066; and KP_F018-KP_R063. Their archaeal SC was 92.78% for the first two pairs and 93.81% for the second two, leaving 14 or 12 species, respectively, uncovered. All of

these pairs targeted gene region 3-9 and had, in order, mean amplicon lengths of 1129, 1128, 1119, and 1118.

At least one primer pair suitable for detecting only archaea (bacterial SC= 0.00%) in the different 16S rRNA gene regions was selected (Table 2). They had to present an archaeal SC ≥90.00% and were chosen based on both their coverage and mean amplicon lengths. The VC results of these selected primers are detailed in appendix S18.

**Table 2. Selected primer pairs for detecting oral archaea in different amplicon-length categories.**

| | | Bacteria | | | | | Archaea | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| ALC | Primer pair | Gene region | SC (%) | Covered | Not covered | Mean length | Gene region | SC (%) | Covered | Not covered | Mean length |
| S | KP_F018-KP_R002 | - | 0.00 | 0 | 769 | 0 | 3 | 95.88 | 186 | 8 | 144 |
| | KP_F016-KP_R003 | - | 0.00 | 0 | 769 | 0 | 3 | 94.85 | 184 | 10 | 158 |
| | OP_F066-KP_R013 | - | 0.00 | 0 | 769 | 0 | 5-6 | 95.88 | 186 | 8 | 275 |
| M | KP_F018-KP_R032 | - | 0.00 | 0 | 769 | 0 | 3-5 | 95.88 | 186 | 8 | 414 |
| | KP_F018-OP_R073 | - | 0.00 | 0 | 769 | 0 | 3-5 | 90.72 | 176 | 18 | 510 |
| | KP_F020-KP_R013 | - | 0.00 | 0 | 769 | 0 | 3-6 | 95.88 | 186 | 8 | 542 |
| | OP_F114-KP_R007 | - | 0.00 | 0 | 769 | 0 | 3-6 | 92.78 | 180 | 14 | 557 |
| | KP_F022-OP_R016 | - | 0.00 | 0 | 769 | 0 | 5-9 | 92.78 | 180 | 14 | 490 |
| | KP_F022-KP_R063 | - | 0.00 | 0 | 769 | 0 | 5-9 | 91.75 | 178 | 16 | 585 |
| L | OP_F114-KP_R013 | - | 0.00 | 0 | 769 | 0 | 3-6 | 95.88 | 186 | 8 | 679 |
| | KP_F018-KP_R063 | - | 0.00 | 0 | 769 | 0 | 3-9 | 93.81 | 182 | 12 | 1118 |
| | KP_F016-KP_R063 | - | 0.00 | 0 | 769 | 0 | 3-9 | 92.78 | 180 | 14 | 1128 |
| | OP_F066-OP_R016 | - | 0.00 | 0 | 769 | 0 | 5-9 | 92.78 | 180 | 14 | 624 |

Species coverage was estimated as the number of species with at least one match in an ASV divided by the number of species included in the database. Our bacterial and archaeal databases contained 769 and 194 species, respectively, each of which had between one and 4000 ASVs. The location of the first and last nucleotides of each primer within each sequence with a match was calculated and the mode values for these positions were determined. If there was more than one mode for a position, we chose the one closest to the mean position value. As all the sequences in the two databases were aligned with the 16S rRNA *Escherichia coli* gene, the mode values obtained for each primer enabled us to allocate them to one of the gene regions defined for that organism by Baker et al. (55). The reference sequence utilised had 1542 base pairs distributed in 10 conserved (C1-C10) and nine hypervariable regions (V1-V9). The sequences of these selected primer pairs are described in appendix S16.
ALC= amplicon length category; F= forward; KP= Klindworth primer; L= long mean amplicon length category, >600 base pairs; M= medium mean amplicon length category, 301-600 base pairs; OP= oral primer; R= reverse; S= short mean amplicon length category,100-300 base pairs; SC= coverage at the species level.



Appendix S19 contains the species not covered by the pairs that achieved an archaeal SC ≥95.00% and a bacterial SC= 0.00%, as well as those named above or included in table 2 with an archaeal SC ≥90.00% and a bacterial SC= 0.00%.

1.4.1.3 Bacterial and archaeal primer pairs

The 156 primer combinations that were candidates for detecting both bacteria and archaea had bacterial and archaeal SC values ranging from 98.57% to 75.42% and 99.48% to 73.71%, respectively. Our classification of the combinations based on their mean amplicon lengths revealed that: 40 pairs had between 100 and 300 bps; 42 from 301 to 600; and 74 more than 600.

Ten pairs in the short mean amplicon length category had bacterial and archaeal SC values ≥95.00%, with a range from 95.97% to 95.32% for the former and from 99.48% to 97.94% for the latter. The max. mean length was 288 bps and the min. 284. All the pairs targeted gene region 4-5 and had been assigned the following identifiers: KP_F020-KP_R031; KP_F020-OP_R070; KP_F020-KP_R032; KP_F020-KP_R035; KP_F020-OP_R020; KP_F020-KP_R038; KP_F020-OP_R010; KP_F020-OP_R014; KP_F020-OP_R036; and KP_F020-OP_R048. The number of bacterial species that were not covered by these pairs ranged from 31 to 36; for the oral-archaeal species, this range was one to four.

Two primer pairs in the 301-600 bps category had bacterial and archaeal SC estimates ≥95.00%: OP_F114-OP_R070 (bacterial SC= 95.58%; archaeal SC= 98.45%); and OP_F114-KP_R031 (bacterial SC= 95.71%; archaeal SC= 98.45%). Both were used to amplify gene region 3-5 and had mean lengths of 460 and 457, respectively. Thirty-three (OP_F114-KP_R031) or 34 (OP_F114-OP_R070) bacterial and three archaeal species from the oral cavity were not covered by these pairs. Lowering the cut-off level to SC ≥90.00% revealed six pairs with a longer mean sequence. For five of these, the difference was irrelevant (461 bps); but the pair OP_F114-OP_R073 had a mean length of 549. This combination targeted gene region 3-6 and had bacterial and archaeal SC values of 94.80% and 93.30%, respectively. The number of non-covered species increased to 40 for the bacteria and 13 for the archaea.

No primer pair from the >600 bps category had SC values ≥95.00% in either database. Conversely, 28 had bacterial and archaeal SC values ≥90.00% (bacterial SC range= 94.54% - 90.64%; archaeal SC range= 96.91% - 96.39%). These pairs amplified gene regions 3-9, 4-9, or 5-9, had mean lengths between 622 and 1063 bps, and did not cover from 42 to 72 bacterial and six to seven archaeal species. The combination of OP_F066 with KP_R060, KP_R076, and

OP_R121 yielded the highest coverage values, with all of them targeting region 5-9. Forty-two bacterial and six archaeal species were not covered by these primers. However, their mean sequence lengths were 623, 622 and 622 bps, respectively, which was close to the lower limit of this category. Primer pairs formed by OP_F114 with KP_R060, KP_R076, or OP_R121, which targeted region 3-9, had a better balance between the coverage results (bacteria SC= 91.42%, archaea SC= 96.91%) and the mean sequence lengths (1063, 1062, and 1062 bps). Sixty-six bacteria and six archaea were not covered by these pairs.

For each amplicon-length category, we selected at least one primer pair suitable for detecting bacteria and archaea in the distinct 16S rRNA gene regions (Table 3). The pairs had to have SC ≥90.00% in both domains and were chosen based on their coverage and mean amplicon lengths. The VC results of these selected primers are detailed in appendix S20.

**Table 3. Selected primer pairs for simultaneously detecting oral bacteria and archaea in different amplicon-length categories.**

| | | Bacteria | | | | | Archaea | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| ALC | Primer pair | Gene region | SC (%) | Covered | Not covered | Mean length | Gene region | SC (%) | Covered | Not covered | Mean length |
| S | OP_F114-KP_R002 | 3-4 | 92.46 | 711 | 58 | 188 | 3 | 98.97 | 192 | 2 | 152 |
| | KP_F020-KP_R032 | 4-5 | 95.58 | 735 | 34 | 284 | 3-5 | 99.48 | 193 | 1 | 285 |
| | OP_F066-OP_R073 | 5-6 | 98.31 | 756 | 13 | 110 | 5 | 92.78 | 180 | 14 | 114 |
| M | OP_F114-KP_R031 | 3-5 | 95.71 | 736 | 33 | 457 | 3-5 | 98.45 | 191 | 3 | 422 |
| | OP_F114-OP_R073 | 3-6 | 94.80 | 729 | 40 | 549 | 3-5 | 93.30 | 181 | 13 | 518 |
| | KP_F020-OP_R073 | 4-6 | 95.71 | 736 | 33 | 376 | 3-5 | 92.78 | 180 | 14 | 381 |
| L | OP_F114-OP_R121 | 3-9 | 91.42 | 703 | 66 | 1062 | 3-9 | 96.91 | 188 | 6 | 1035 |
| | KP_F020-OP_R121 | 4-9 | 91.55 | 704 | 65 | 888 | 3-9 | 96.91 | 188 | 6 | 897 |
| | OP_F066-OP_R121 | 5-9 | 94.54 | 727 | 42 | 622 | 5-9 | 96.91 | 188 | 6 | 630 |

Species coverage was estimated as the number of species with at least one match in an ASV divided by the number of species included in the database. Our bacterial and archaeal databases contained 769 and 194 species, respectively, each of which had between one and 4000 ASVs. The location of the first and last nucleotides of each primer within each sequence with a match was calculated and the mode values for these positions were determined. If there was more than one mode for a position, we chose the one closest to the mean position value. As all the sequences in the two databases were aligned with the 16S rRNA *Escherichia coli* gene, the mode values obtained for each primer enabled us to allocate them to one of the gene regions defined for that organism by Baker et al. (55). The reference sequence utilised had 1542 base pairs distributed in 10 conserved (C1-C10) and nine hypervariable regions (V1-V9). The sequences of these selected primer pairs are described in appendix S16.

ALC= amplicon length category; F= forward; KP= Klindworth primer; L= long mean amplicon length category, >600 base pairs; M= medium mean amplicon length category, 301-600 base pairs; OP= oral primer; R= reverse; S= short mean amplicon length category,100-300 base pairs; SC= coverage at the species level.

Appendix S21 is comprised of the species not covered by the pairs with a bacterial and an archaeal SC ≥95.00%, and by the combinations with a bacterial and archaeal SC ≥90.00% referred to in this section or included in table 3.

Finally, appendix 22 contains a list of the primer pairs used in the reviewed studies on 16S rRNA sequencing of the oral microbiome; and appendix 23 details the species not covered by the three most frequently employed primer combinations in the literature.

## 1.5 DISCUSSION

To the best of our knowledge, this is the first study that evaluates *in silico* the coverage of 16S rRNA gene primers for the detection of oral-bacterial and archaeal species. The primer sequences were obtained not only from sequencing-based studies of the microbiome inhabiting the human mouth, but also from an article containing primers used in ecosystems as dissimilar as the marine, geothermal, human gut, or cattle gut (16). Thus, numerous primers from diverse ecosystems were analysed to find those which performed better in the oral cavity. Moreover, to perform the analysis, we improved an earlier database of 16S rRNA gene sequences of oral bacteria (33) and created another from scratch that contained sequences from archaeal species found in the oral cavity.

We identified a series of individual primers that performed well in the detection of oral bacteria and/or archaea and combined them to create primer pairs. These were defined as "bacteria-specific", "archaea-specific", or "bacterial and archaeal" based on the results of their levels of coverage set out in the two databases. We also produced a series of primer pairs that may be the most suitable combinations for use when sequencing the oral ecosystem. These were classified according to the domain targeted, their mean amplicon length category, and the 16S rRNA gene region amplified.

### 1.5.1 Comparative analysis of our coverage results of 16S rRNA gene primers with the literature

The investigation by Klindworth et al. (16) is perhaps the most comprehensive to date on the coverage and phylum spectrum of 16S rRNA primers. These authors assessed 175 primers and 512 primer pairs *in silico* against the Silva non-redundant reference database (version 108) (26), producing a selection of those that performed best for bacteria and archaea. Like us, this group organised the most suitable primer combinations for the different sequencing technologies into three categories according to their amplicon length (100–400, 400–1000, >1000 bps). They then re-evaluated their analysis using the Global Ocean Sampling (GOS) dataset (56,57), which is limited to the marine habitat, and they examined experimentally the primer pair that performed best (16).

We identified two investigations involving the oral ecosystem that used the Silva database (26) -versions 111 (19) or 132 (22)- to analyse the efficiency of 16S rRNA gene primers for detecting the archaea diversity in oral samples (22), or for reconstructing the microbiota of ancient dental calculus specimens (19). Also, a third study evaluated the potential of seven primer pairs for detecting 219 species in a foregut dataset they created, which included oral, oesophageal, and gastric 16S rRNA gene sequences (23). The pair with the best results for classifying the foregut genes was also analysed against the RDP database (27).

The numbers of 16S rRNA gene primer pairs evaluated in these three oral-related studies are substantially lower than in the present investigation: 12 individual primers combined to form 12 primer pairs (22); 25 individual primers combined into 14 pairs (19); and 14 individual primers grouped into 14 pairs (23). In our study, we analysed 369 distinct individual primers and 4638 different primer-pair combinations. On the other hand, two investigations used the Silva database (19,22,26) which has broad phylogenetic diversity and contains information applicable to many environments but also includes 16S rRNA gene sequences that are misannotated taxonomically (33). Specifically, comprehensive databases such as Silva (26), RDP (27), or Greengenes (58) have been estimated to have annotation error rates ranging from 10-17% (28), and their accuracy may also be reduced because they contain numerous sequences derived from some environments and only a few from others (29). Furthermore, the evaluation of the primers' coverage using an ecosystem-specific database, as in our study, would allow researchers to identify the species covered and not covered by a particular primer pair. In this sense, only Nossa et al. (23) evaluated the primer pairs against a self-created database containing sequences from their three niches of interest: oesophageal, oral, and gastric (together described as the foregut). Nevertheless, although this database contained 9484 sequences, only 2373 were oral and, overall, they represented just 219 bacterial species. These numbers are much lower than those in the bacteria database used in our study, which is based on eHOMD (34), and to which we added sequences from our self-created archaeal dataset (bacteria: 223143 sequences, 769 species; archaea: 2842 sequences, 194 species).



In the present study, none of the individual primers yielded an SC= 100% when analysed against the oral bacteria or archaea databases. Due to this, it would not be possible to obtain a primer combination with such value as the coverage estimates of primer pair are always lower

than the values of the individual primers that form it. However, we do not know whether any of the primers included in this research could obtain such a value if mismatches were admitted.

#### 1.5.1.1 Bacteria-specific primer pairs

Table 4 summarises our results and those in other publications, with the primer pairs ordered by the mean amplicon length and the domain targeted. Concerning the bacteria-specific candidates, our coverage estimates for KP_F047-KP_R021 (S), KP_F049-KP_R033 (M), KP_F056-KP_R074 (M), KP_F033-KP_R060 (L) and KP_F047-KP_R053 (L) were similar to those of other studies, with differences no greater than 5.00% for both the bacteria and archaea domains (16). It should be noted that the latter, classified here as having a mean amplicon length >600 bps but put in the medium-length category by Klindworth (16), had a lower bacterial coverage when analysed against the GOS database (56,57). This was also the case for KP_F056-KP_R074 (16). Moreover, the coverage values of the pairs KP_F077-KP_R071 (S) and KP_F047-KP_R035 (M) in other studies were similar to those in our research, but the archaeal coverage was notably higher -~31.00% (19) and ~83.00% (21,25) more, respectively-. KP_F047-KP_R035, which had an archaeal SC= 0.00% in our analysis, has been described elsewhere as having universal coverage for both archaea and bacteria (25). We, therefore, believe that KP_F047-KP_R035 has value for detecting archaea in environmental (21,25) or human gut (25) specimens, but not in samples from the oral cavity.

**Table 4. Coverage findings are described in the literature for the gene primer pairs analysed in the present study.**

| Present study | | Other studies | Results present study | | Results other studies | | |
|---|---|---|---|---|---|---|---|
| Primer pair | ALC | Primer pair name | Bacterial SC (%) | Archaeal SC (%) | Bacterial coverage (%) | Archaeal coverage (%) | Ref. |
| KP_F044-KP_R023 | S | S-D-Bact-0337-a-S-20/ S-D-Bact-0518-a-A-17 | 87.52 | 0.00 | 80.90 | 0.00 | (12) |
| KP_F044-KP_R021 | S | S-D-Bact-0337-a-S-20/ S-D-Bact-0515-a-A-19 | 92.46 | 0.00 | 85.80 | 0.00 | (12) |
| KP_F046-KP_R023 | S | S-D-Bact-0341-a-S-17/ S-D-Bact-0518-a-A-17 | 87.52 | 0.00 | 81.30 | 0.00 | (12) |
| KP_F046-KP_R021 | S | S-D-Bact-0341-a-S-17/ S-D-Bact-0515-a-A-19 | 92.46 | 0.00 | 86.20 | 0.00 | (12) |
| KP_F046-OP_R045 | S | S-D-Bact-0341-a-S-17/ NA | 87.52 | 0.00 | 81.50 | 0.00 | (12) |
| KP_F047-KP_R021 | S | S-D-Bact-0341-b-S-17/ S-D-Bact-0515-a-A-19 | 92.59 | 0.00 | 91.20[a] | 0.00[a] | (11) |
| KP_F056-KP_R032 | S | S-D-Bact-0564-a-S-15/ S-D-Bact-0785-b-A-18 | 96.23 | 8.76 | 89.00[a]; 83.40[b] | 14.60[a]; 0.00[b] | (11) |
| | | S-D-Bact-0564-a-S-15/ S-D-Bact-0785-b-A-18 | | | 88.10 | 14.40 | (12) |
| KP_F058-KP_R053 | S | S-D-Bact-0784-a-S-19/ S-D-Bact-1061-a-A-17 | 84.40 | 0.00 | 78.60 | 0.00 | (12) |
| KP_F077-KP_R071 | S | U341F - 534R | 95.58 | 59.79 | 98.00 | ~ 91.00 | (14) |
| KP_F078-OP_R010 | S | 515F - 806 R (original) | 95.32 | 62.89 | 86.80 | 52.90 | (17) |
| | | S-*-Univ-0515-a-S-19/ NA | | | 86.10 | 52.00 | (12) |
| KP_F078-KP_R037 | S | S-*-Univ-0515-a-S-19/ S-D-Bact-0787-a-A-20 | 87.39 | 0.52 | 77.10 | 0.00 | (12) |
| KP_F018-KP_R002 | S | S-D-Arch-0349-a-S-17/ S-D-Arch-0519-a-A-16 | 0.00 | 95.88 | 0.00[a]; 0.00[b] | 76.80[a]; 74.50[b] | (11) |
| KP_F020-KP_R032 | S | S-D-Arch-0519-a-S-15/ S-D-Bact-0785-b-A-18 | 95.58 | 99.48 | 89.10[a]; 83.40[b] | 88.00[a]; 76.50[b] | (11) |
| | | 519F - 785R | | | 88.80 | 88.90 | (17) |
| KP_F020-KP_R035 | S | S-D-Arch-0519-a-S-15/ S-D-Bact-0785-a-A-21 | 95.45 | 98.97 | 87.10[a] | 86.50[a] | (11) |
| OP_F014-OP_R014 | S | 515F - 806 R (modified) | 95.32 | 88.14 | 87.70 | 85.70 | (17) |
| | | 515F - 806R | | | 96.20 | 96.39 | (20) |
| OP_F066-OP_R073 | S | NA/ NA | 98.31 | 92.78 | 88.90 | 75.30 | (12) |
| KP_F044-KP_R032 | M | S-D-Bact-0337-a-S-20/ S-D-Bact-0785-b-A-18 | 94.28 | 0.00 | 84.30 | 0.00 | (12) |
| KP_F046-OP_R010 | M | S-D-Bact-0341-a-S-17/ NA | 93.89 | 0.00 | 83.30 | 0.10 | (12) |
| KP_F047-KP_R035 | M | S-D-Bact-0341-b-S-17/ S-*-D-Bact-0785-a-A-21 | 94.15 | 0.00 | 86.20[a]; 43.10[b] | 0.50[a]; 0.00[b] | (11) |
| | | 341F -805R | | | 96.69 | 83.59 | (20) |
| | | 341F -785R | | | 96.51 | 82.96 | (16) |
| | | S-D-Bact-0341-b-S-17/ S-D-Bact-0785-a-A-21 | | | 86.00 | 0.50 | (12) |
| KP_F049-KP_R033 | M | S-D-Bact-0347-a-S-19/ S-D-Bact-0785-a-A-19 | 76.59 | 0.00 | 76.50[a] | 0.00[a] | (11) |
| KP_F056-KP_R074 | M | S-D-Bact-0564-a-S-15/ S-Univ-1100-a-A-15 | 97.27 | 7.73 | 92.70[a]; 76.20[b] | 8.00[a]; 0.00[b] | (11) |
| OP_F021-OP_R050 | M | NA / NA | 91.68 | 1.03 | 86.50 | 0.50 | (12) |
| KP_F020-KP_R013 | M | S-D-Arch-0519-a-S-15/ S-D-Arch-1041-a-A-18 | 0.00 | 95.88 | 0.00[a] | 76.60[a] | (11) |
| KP_F032-KP_R063 | L | S-D-Bact-0008-b-S-20/ S-D-Bact-1492-a-A-16 | 60.47 | 0.00 | 17.30 | 0.00 | (12) |
| KP_F033-KP_R060 | L | S-D-Bact-0008-c-S-20/ S-D-Bact-1391-a-A-17 | 74.90 | 0.00 | 78.00[a] | 0.10[a] | (11) |
| KP_F033-KP_R050 | L | S-D-Bact-0008-c-S-20/ S-D-Bact-1046-a-A-19 | 72.56 | 0.00 | 81.30[a] | 0.00[a] | (11) |
| KP_F047-KP_R053 | L | S-D-Bact-0341-b-S-17/ S-D-Bact-1061-a-A-17 | 93.11 | 0.00 | 91.90[a]; 58.90[b] | 0.00[a]; 0.00[b] | (11) |
| KP_F051-KP_R057 | L | S-D-Bact-0515-a-S-16/ S-D-Bact-1100-a-A-15 | 82.70 | 0.00 | 77.30 | 0.00 | (12) |
| KP_F018-KP_R078 | L | S-D-Arch-0349-a-S-17/ S-*-Univ-1392-a-A-15 | 0.00 | 93.30 | 0.00[a] | 65.80[a] | (11) |
| KP_F059-KP_R078 | L | S-D-Bact-0785-a-S-18/ S-*-Univ-1392-a-A-15 | 93.63 | 96.39 | 74.10[a] | 72.30[a] | (11) |

The coverage findings from the other investigations are those obtained when zero mismatches were accepted. a= Silva database; ALC= amplicon length category; b= GOS database; F= forward; KP= Klindworth primer; L= long mean amplicon length category, >600 base pairs; M= medium mean amplicon length category, 301-600 base pairs; NA= not available; OP= oral primer; R= reverse; Ref.= references; S= short mean amplicon length category,100-300 base pairs; SC= coverage at the species level.

The remaining candidates to be bacteria-specific primer pairs in all the amplicon-length categories herein were better at detecting bacteria than in other studies, with differences of 5.00% to 43.00%. Only Klindworth's (16) estimate for the bacterial coverage of KP_F033-KP_R050 was better than ours, with an approximate difference of ~9.00% between the studies (Table 4).

#### 1.5.1.2 Archaea-specific primer pairs

Two of the primer pairs selected herein to detect oral-archaea species – KP_F018-KP_R002 (S) and KP_F020-KP_R013 (M) – have previously been described in other studies as the best options for targeting this domain (16). Nonetheless, the archaeal SC values we obtained exceed Klindworth's (16) by ~ 20.00% (Table 4). Klindworth's group (16) also found that KP_F018-KP_R078 had the highest overall archaeal coverage for the amplicons with a long mean length. However, they did not recommend its use due to its low phylum spectrum. This combination produced bacterial SC values of 0.00% and an archaeal SC of 93.30% when analysed against our database (Table 4). Although this is a good result, we prefer OP_F114-KP_R013, which achieves better archaeal SC, or KP_F018-KP_R063, which produces both better archaeal SC and a greater mean amplicon length.

#### 1.5.1.3 Bacterial and archaeal primer pairs

KP_F020-KP_R032 and KP_F020-KP_R035 (S), with bacterial and archaeal coverage estimates >85.00% (mainly considering the Silva database (26)), have been proposed previously as suitable for the detection of both domains (16). As seen in table 4, the SC values obtained herein, ≥95.00%, are better than those in other studies (16,22). OP_F066-OP_F073 is also among the favoured primers in our study for the detection of bacteria and archaea when using short amplicon lengths (SC= 98.31% and 92.78%, respectively), achieving better coverage than in the research by Zhang et al. (17) (SC= 88.90% and 75.30%, respectively). Meanwhile, although other studies' *in silico* analyses of OP_F014-OP_R014 have described it as a good primer pair for detecting the two domains (22,25), it is not among our recommended primers, since others in the same length and gene-region categories achieved better archaeal coverage. KP_F059-KP_R078 has been proposed by Klindworth et al. (16) as suitable for use with both the bacteria and archaea domains when employing medium mean amplicon lengths (608 bps). However, its length was 622 bps in our study, meaning that it was included in the

>600 bps group. In any case, although our coverage values were higher than those obtained previously (SC= 93.63% and 96.39% *vs.* 74.10 and 72.30%, respectively) (Table 4), other primer pairs performed better in both categories as OP_F114-KP_R031 (M; bacterial SC= 95.71% and archaeal SC= 98.45%) and OP_F066-OP_R121 (L; bacterial SC= 94.54% and archaeal SC= 96.91%).

### 1.5.2 Non-covered species by the 16S rRNA gene primer pairs

The *in silico* analysis has enabled us to verify that, among the pairs achieving better coverage, the species not covered by the primers targeting a particular region tend to be covered by others relating to a different zone. In this sense, most of the species that were not covered by the bacteria-specific primer pairs from regions 3-4 (S), 3-5 (M), or 3-7 (L) were by those from 5-7 and 6-7 (S), 4-7 and 7-9 (M), or 4-9 (L), and vice versa. This was also seen in the archaea-specific primers, where taxa not detected by the pairs from regions 3 (S), 3-5 (M), or 3-6 (L) were covered by those from 5-6 (S), 3-6 and 5-9 (M), or 3-9 and 5-9 (L), and vice versa. Lastly, the pairs for the two domains combined also demonstrated that species not covered by primers from zones 4-5 (S) or 3-5 (M) were by those from 5-6 (S) or 4-6 (M), and vice versa. In the combinations with mean amplicon lengths >600 bps, half the taxa that were not covered using primers for amplifying region 3-9 were detected when targeting 5-9. However, in this case, the opposite was not true.

There were exceptions to this general rule, which demonstrated that even for two primer pairs targeting the same gene region, one would be able to cover most of the species that were not detected by the other. As an example, the bacteria-specific pair KP_F048-OP_R043 detected 18 of the 33 species non covered by 18 different primer pairs formed by combining KP_F044, 046, 047, OP_F048, 096, or 108 and KP_R071, OP_R040, or 146 (3-4; S); and, also, OP_F101-OP_R030 covered 33 of the 62 species not detected by KP_F061-KP_R074 (6-7; S). Furthermore, 54 of the 75 bacteria not covered by KP_F048-OP_R050 were detected using KP_F048-OP_R073 and OP_F050-OP_R073 (3-6; M); and 22 of the 28 species not detected by KP_F051-KP_R053 and OP_F021-KP_R053 were covered by all the other primers targeting region 4-7 (M). Also, in the long primer pairs, KP_F048-OP_R030 (3-7) covered 20 of the 37 non-detected taxa by the combinations of KP_F044, 046, 047, OP_F048, 096, or 108 with KP_R074. Meanwhile, the archaea-specific primer KP_F016-KP_R002 covered almost all

(6/8) of the taxa not detected using KP_F018-KP_R002, -KP_R003, and -OP_R102 (3; S). In addition, KP_F016-KP_R032 was the only primer from region 3-5 (M) able to identify six archaea: *Candidatus Korarchaeum cryptofilum*, *Ferroplasma acidarmanus*, *Fervidicoccus fontis*, *Metallosphaera cuprina*, *Methanocorpusculum labreanum*, and *Thermophilus pendens*; that were not detected by the rest of the primers from the same region. Nevertheless, these exceptions were not observed in the bacterial and archaeal primer pairs.

It is clear that most of the taxa not detected by these well-performing primer pairs must have been identified as present in the oral cavity at some point, or they would not have been included in the databases used for our *in silico* analysis. However, some of them are microbes associated with prevalent oral pathologies, such as periodontal disease or dental caries. We distinguished four recognised *Campylobacter* species among the bacteria not detected by some of the bacteria-specific primer pairs: *concisus*, *gracilis*, *rectus*, and *showae*. The first of these, as part of the Socransky green complex, has traditionally been associated with periodontal health; the remaining three are components of the orange complex, which is related to periodontitis (59). A further three bacteria commonly found in the healthy periodontium, *Leptotrichia bucalis* (60), *Leptotrichia hofstadii* (60) and *Rothia dentocariosa* (61,62), were also missed by some of the bacteria-specific and/or bacterial and archaeal primer pairs. Conversely, a few failed to cover bacterial taxa isolated from periodontally-diseased sites (in teeth or implants) or those regarded as novel periodontal pathogens, e.g., *Actinomyces dentalis* (63), *Actinomyces israelii* (63), *Desulfomicrobium orale* (64), *Mogibacterium timidum* (65), *Solobacterium moorei* (66,67), *Treponema lecithinolyticum* (61,65,67) and *Treponema maltophilum* (68). A further *Actinomyces* species, previously classified as *naeslundii* WVA 963 and now known as *johnsonii* (69), which has been encountered in both healthy and periodontitis sites (63), was not detected by some of the pairs that produced better coverage estimates. Moreover, different taxa from the phyla *Saccharibacteria* (TM7), which growing evidence links to periodontal disease (70), were also not covered. Meanwhile, the caries-associated bacterial species that were not detected by some of the primer pairs included *Bifidobacterium dentium* (71,72), *Lactobacillus reuteri* (73), *Leptotrichia buccalis* (74), *Parascardovia denticolens* (73,75), *R. dentocariosa* (76) and *Scardovia wiggsiae* (77).

The undetected archaeal species by some of the archaea-specific and/or bacterial and archaeal primer pairs included *Methanobrevibacter gottschalkii*, *Methanopyrus kandleri*, *Nitrosoarchaeum limnia*, and *Nitrososphaera evergladensis*; these species have been found, in order, in inflamed pulp tissue (78), periodontitis samples (79), endodontic infections (80) and ancient dental calculus (81,82). The rest of the non-detected archaea were extracted from the same publication (83) and, as far as we know, not reported by other authors so their role in the oral cavity has yet to be investigated.

Consequently, it would be preferable to choose a primer pair based on the health or disease condition being investigated. If it is known which oral species are not covered by each primer pair in the oral-specific database as we demonstrated for the first time in this study, and which taxa are most commonly associated with the target oral condition, it is possible to select the most optimal primer pair to use in the sequencing-based studies of the oral microbiota.

### 1.5.3 Primer pairs frequently used in the oral microbiome literature

Finally, our review of the literature found that 206 distinct primer pairs have been utilised to study the oral microbiome via massive sequencing techniques. The combinations employed most commonly were KP_F078-OP_R010 and KP_F047-KP_R035, which were repeated 33 and 21 times, respectively. These were followed by KP_F014-KP_R011, KP_F034-KP_R065, KP_F031-KP_R021, and OP_F009-OP_R029, which appeared in eight, eight, seven, and seven articles, respectively. Four, three, four, 10, and 21 distinct pairs were repeated six, five, four, three, and two times in the sequencing-based studies of the oral microbiome. Lastly, 158 were found only once.

Only 67 of these 206 pairs were evaluated in the present study. This means that at least one of the individual primers from the remaining 139 combinations had a bacterial and archaeal SC < 75.00%. The widely-employed primer KP_F078-OP_R010, which targets region 4 and is typically found as 515F-806R, was developed by Caporaso et al. (84) for use in the Illumina sequencing platform. The *in silico* analysis herein revealed bacterial and archaeal SC estimates of 95.32% and 62.89%, respectively (mean amplicon length= 292 bps), but failed to detect *M. kandleri*, *N. limnia*, and *N. evergladensis*, among other archaeal species. Numerous primer combinations in the same amplicon length category (S) and targeting the same gene region (4-

5) provided better SC for both domains, e.g., KP_F020 and KP_R031, KP_R032, or OP_R070 (bacterial SC range= 95.97% - 95.58%; archaeal SC range= 99.48% - 98.97%; mean amplicon length range= 287 - 284 bps). If only bacteria are to be detected, the primer pair from the same region, OP_F098-OP_R119, is preferrable; although it had a slightly lower bacterial SC (94.54%), its archaeal SC was 0.00%, meaning that no 16S rRNA gene sequence from an oral archaeon would limit the sequencing depth.

KP_F047-KP_R035, directed to amplify region 3-4, has been referred to as 341F-785R, 341F-805R, or 341F-806R and is the pair proposed in the Illumina protocol for the preparation of the sequencing library (85). In the *in silico* analysis, it achieved species coverages of 94.15% and 0.00% in the oral-bacteria and oral-archaea databases, respectively, (mean amplicon length= 460 bps). Surprisingly, this pair did not cover the previously mentioned bacterial species *A. dentalis*, *A. israelii*, *A. johnsonii*, *D. orale*, *L. reuteri*, *M. timidum*, *T. lecithinolyticum* and *T. maltophilum*. Although it has been used extensively in oral microbiome studies, in our investigation other primers in the same mean amplicon length category (M) and the same gene region (3-5) had better bacterial coverage values and a similar mean amplicon length as KP_F048-KP_R031 (SC= 97.53%). This latter pairing, as well as all those in the same length category and targeting the same region, also failed to detect *A. dentalis*, *A. israelii*, *A. johnsonii* and *D. orale.* However, unlike KP_F047-KP_R035, *L. reuteri*, *M. timidum*, *T. lecithinolyticum* and *T. maltophilum* were covered.

Another widely used primer is 785F-1175R, which has been employed to amplify gene region 5-7. The *in silico* evaluation of the pair named herein as OP_F009-OP_R029 yielded bacterial SC values of 88.30% and an archaeal SC of 0.00% (mean amplicon length= 410 bps). This, along with all the other bacteria-specific combinations within the same gene region (5-8) and amplicon length category (M), was not among the best primers in our investigation (bacterial SC range= 89.60% - 79.97%; archaeal SC= 0.00%; mean amplicon length range= 411 - 408 bps). In fact, OP_F009-OP_R029 not only failed to detect *A. dentalis*, *A. israelii*, *A. johnsonii, C. concisus, C. gracilis, C. rectus* and *C. showae*, but also the microbes that are widely known to be associated with periodontitis, *Porphyromonas endodontalis* (86-88), *Porphyromonas gingivalis* (59,86,88,89) and *Tannerella forsythia* (59,88,89). Consequently, it is preferable to amplify region 4-7 using the pairs KP_F051-OP_R030 or OP_F021-OP_R030,

which had better bacterial SC (98.83% and 98.70%, respectively) and mean amplicon lengths (566 bps), and also detected the bacterial species referred to above.

### 1.5.4 Factors to consider when selecting a 16S rRNA gene primer pair

Although we defined which primer pairs had the highest coverage results for detecting oral bacteria and archaea, this does not necessarily mean that they would be always the best option for any sequencing-based research on the oral microbiota. Other factors such as the amplicon length or gene region targeted, should also be taken into account when selecting the optimum primer pair as we did for constructing tables 1, 2, and 3. Although PCR efficiency decreases when the amplicon length increases (90), in general terms, the longer the fragment sequenced, the lower the taxonomic level that can be achieved (17). Indeed, sequencing full-lengths, as is possible with PacBio, is regarded as the solution to the limitations of taxonomic classification (17). Nevertheless, Soergel et al. (29) evaluated primer pairs in common use and found that longer gene amplicons did not necessarily confer better classifications, with the target region (depending on the sample's origin) impacting taxonomic assignment the most. Similarly, other authors have recently found that the different 16S rRNA gene regions contain varying amounts of information, which significantly affects the composition of the bacterial community (17). Consequently, we agree that the choice of the target region is also an important factor (17,29).

In this sense, our study provides the scientific community with information on these three aspects to consider in the selection of a primer pair for a total of 4638 primer pairs, adding the description of the covered and non-covered taxa for each primer pair. Our goal is to enable researchers to select the best primer pair to meet their research expectations.

In view of our results, the bacteria-specific primer pairs showed very similar average coverage values in the three mean amplicon length categories (S: 94.19%; M: 94.71%; L: 94.87%). However, the archaea-specific primer combinations with short mean amplicon lengths had a slightly higher overall coverage than those from the other two categories (S: 95.54% *vs.* M: 93.30% and L: 93.81%); and the bacterial and archaeal primer pairs with short and long mean amplicon lengths performed better than those from the medium category (S: 97.08% and L: 96.91% *vs.* M: 94.83%). Given the above, as the differences in overall coverage between the three amplicon length categories were not too large, the choice of amplicon length, and

consequently the sequencing platform, is left to the discretion of the researchers. On the other hand, the gene regions showing the greatest coverage values in the three mean amplicon length categories (in order S, M, and L) were: 3-4, 4-7, and 3-7 (bacteria-specific); 3 or 5-6, 3-5 or 3-6, and 3-6 (archaea-specific); and 4-5, 3-5, and 5-9 (bacterial and archaeal). Thus, for the different lengths of the bacteria-specific primers and the bacterial and archaeal primers, there was a specific region associated with the highest coverage, which was not observed in the archaea-specific primers where greater variability was detected. Moreover, except for the archaea-specific and the bacterial and archaeal primer pairs in the medium mean amplicon length category, in which region 3-5 showed the highest coverage estimates; there was no consensus between the three types of primer pairs on the most informative region for a given category.

### 1.5.5 Limitations of the present study

The main limitation of the present study arises from the lack of information on the first and last positions in the sequence annotations stored by the NCBI (37), which suggests that primers targeting these gene regions may have lower VC values. We, therefore, calculated the SC estimates, since a particular species would be regarded as covered by a particular primer if at least one of its variants is amplified. In addition, we were unable to identify the complete genome of some archaeal species in our database (*C. K. cryptofilum*, *M. gottschalkii strain HO*, *Methanobrevibacter oralis*, *Methanobrevibacter thaueri strain CW* and *N. limnia*). Given that the gene sequences from these taxa were not obtained in the same way as for the other species, we cannot be sure that there are no sequence variants in addition to those found in our investigation. It should, however, be noted that the oral-archaea database developed by our group is the first proposal and may, therefore, be subject to change. Moreover, additional scientific evidence on the archaea species associated with the oral cavity and its diseases is required to increase the amount of information contained in the 16S rRNA sequence databases. In consequence, the results of our *in silico* analysis have potential use in studies of the oral microbiome and need to be confirmed in other experimental studies using omics techniques.

## 1.6 CONCLUSIONS

Considering the three amplicon category lengths (100-300, 301-600, and >600), the primer pairs with the best estimated coverage for detecting oral bacteria targeted regions 3-4, 4-7, and 3-7, and these were: KP_F048-OP_R043 (primer pair position for *E. coli* J01859.1: 342-529), KP_F051-OP_R030 (514-1079), and KP_F048-OP_R030 (342-1079). For the detection of oral archaea, the pairs with the best coverage amplified regions 5-6, 3-6, and 3-6, and these were: OP_F066-KP_R013 (784-undefined), KP_F020-KP_R013 (518-undefined) and OP_F114-KP_R013 (340-undefined). The pairs with the best coverage of the bacteria and archaea domains jointly were found in regions 4-5, 3-5, and 5-9, and these were: KP_F020-KP_R032 (518-801), OP_F114-KP_R031 (340-801), and OP_F066-OP_R121 (784-1405). The primer pairs with the best coverage identified herein are not among those described most widely in the oral microbiome literature.

## 1.7 REFERENCES

(1) Deo PN, Deshmukh R. Oral microbiome: Unveiling the fundamentals. J Oral Maxillofac Pathol. 2019 Jan-Apr;23(1):122-8. doi: 10.4103/jomfp.JOMFP_304_18.

(2) Valm AM. the structure of dental plaque microbial communities in the transition from health to dental caries and periodontal disease. J Mol Biol. 2019 Jul;431(16):2957-69.

(3) Tonetti MS, Bottenberg P, Conrads G, Eickholz P, Heasman P, Huysmans M, et al. Dental caries and periodontal diseases in the ageing population: call to action to protect and enhance oral health and well-being as an essential component of healthy ageing – Consensus report of group 4 of the joint EFP/ORCA workshop on the boundaries between caries and periodontal diseases. J Clin Periodontol. 2017 Mar;44 Suppl 18:S135-44.

(4) Peng X, Cheng L, You Y, Tang C, Ren B, Li Y, et al. Oral microbiota in human systematic diseases. Int J Oral Sci. 2022 Mar;14(1):14. doi: 10.1038/s41368-022-00163-7.

(5) Durán-Pinedo AE, Frias-Lopez J. Beyond microbial community composition: functional activities of the oral microbiome in health and disease. Microbes Infect. 2015 Jul;17(7):505-16.

(6) Krishnan K, Chen T, Paster BJ. A practical guide to the oral microbiome and its relation to health and disease. Oral Dis. 2017 Apr;23(3):276-86.

(7) Slatko BE, Gardner AF, Ausubel FM. Overview of next-generation sequencing technologies. Curr Protoc Mol Biol. 2018 Apr;122(1):e59. doi: 10.1002/cpmb.59.

(8) Willis JR, Gabaldón T. The human oral microbiome in health and disease: from sequences to ecosystems. Microorganisms. 2020 Feb;8(2):308. doi: 10.3390/microorganisms8020308.

(9) Verma D, Garg PK, Dubey AK. Insights into the human oral microbiome. Arch Microbiol. 2018 May;200(4):525-40.

(10) Wade WG. The oral microbiome in health and disease. Pharmacol Res. 2013 Mar;69(1):137-43.

(11) Pérez-Cobas AE, Gomez-Valero L, Buchrieser C. Metagenomic approaches in microbial ecology: an update on whole-genome and marker gene sequencing analyses. Microb Genom. 2020 Aug;6(8):mgen000409. doi: 10.1099/mgen.0.000409.

(12) Robinson CK, Brotman RM, Ravel J. Intricacies of assessing the human microbiome in epidemiologic studies. Ann Epidemiol. 2016 May;26(5):311-21.

(13) de la Cuesta-Zuluaga J, Escobar JS. Considerations for optimizing microbiome analysis using a marker gene. Front Nutr. 2016 Aug;3:26. doi: 10.3389/fnut.2016.00026.

(14) Hamady M, Knight R. Microbial community profiling for human microbiome projects: tools, techniques, and challenges. Genome Res. 2009 Jul;19(7):1141-52.

(15) Mao DP, Zhou Q, Chen CY, Quan ZX. Coverage evaluation of universal bacterial primers using the metagenomic datasets. BMC Microbiol. 2012 May;12:66. doi: 10.1186/1471-2180-12-66.

(16) Klindworth A, Pruesse E, Schweer T, Peplies J, Quast C, Horn M, et al. Evaluation of general 16S ribosomal RNA gene PCR primers for classical and next-generation sequencing-based diversity studies. Nucleic Acids Res. 2013 Jan;41(1):e1. doi: 10.1093/nar/gks808.

(17) Zhang J, Ding X, Guan R, Zhu C, Xu C, Zhu B, et al. Evaluation of different 16S rRNA gene V regions for exploring bacterial diversity in a eutrophic freshwater lake. Sci Total Environ. 2018 Mar;618:1254-67.

(18) Sambo F, Finotello F, Lavezzo E, Baruzzo G, Masi G, Peta E, et al. Optimizing PCR primers targeting the bacterial 16S ribosomal RNA gene. BMC Bioinformatics. 2018 Sep;19(1):343. doi: 10.1186/s12859-018-2360-6.

(19) Ziesemer KA, Mann AE, Sankaranarayanan K, Schroeder H, Ozga AT, Brandt BW, et al. Intrinsic challenges in ancient microbiome reconstruction using 16S rRNA gene amplification. Sci Rep. 2015 Nov;5:16498. doi: 10.1038/srep16498.

(20) Bahram M, Anslan S, Hildebrand F, Bork P, Tedersoo L. Newly designed 16S rRNA metabarcoding primers amplify diverse and novel archaeal taxa from the environment. Environ Microbiol Rep. 2019 Aug;11(4):487-94.

(21) Thijs S, Op De Beeck M, Beckers B, Truyens S, Stevens V, Van Hamme JD, et al. Comparative evaluation of four bacteria-specific primer pairs for 16S rRNA gene surveys. Front Microbiol. 2017 Mar;8:494. doi: 10.3389/fmicb.2017.00494

(22) Pausan MR, Csorba C, Singer G, Till H, Schöpf V, Santigli E, et al. Exploring the archaeome: detection of archaeal signatures in the human body. Front Microbiol. 2019 Dec;10:2796. doi: 10.3389/fmicb.2019.02796.

(23) Nossa CW, Oberdorf WE, Yang L, Aas JA, Paster BJ, Desantis TZ, et al. Design of 16S rRNA gene primers for 454 pyrosequencing of the human foregut microbiome. World J Gastroenterol. 2010 Sep;16(33):4135-44.

(24) Ku HJ, Lee JH. Development of a novel long-range 16S rRNA universal primer set for metagenomic analysis of gastrointestinal microbiota in newborn infants. J Microbiol Biotechnol. 2014 Jun;24(6):812-22.

(25) Wasimuddin, Schlaeppi K, Ronchi F, Leib SL, Erb M, Ramette A. Evaluation of primer pairs for microbiome profiling from soils to humans within the One Health framework. Mol Ecol Resour. 2020 Nov;20(6):1558-71.

(26) Quast C, Pruesse E, Yilmaz P, Gerken J, Schweer T, Yarza P, et al. 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.

(27) Cole JR, Wang Q, Fish JA, Chai B, McGarrell DM, Sun Y, et al. Ribosomal Database Project: data and tools for high throughput rRNA analysis. Nucleic Acids Res. 2014 Jan;42(D1):D633-42. doi: 10.1093/nar/gkt1244.

(28) Edgar R. Taxonomy annotation and guide tree errors in 16S rRNA databases. PeerJ. 2018 Jun;6:e5030. doi: 10.7717/peerj.5030.

(29) Soergel DA, Dey N, Knight R, Brenner SE. Selection of primers for optimal taxonomic classification of environmental 16S rRNA gene sequences. ISME J. 2012 Jul;6(7):1440-4.

(30) R Core Team. R: a language and environment for statistical computing. R package version 4.0.3. Vienna, Austria: R Foundation for Statistical Computing; 2020; Available at: https://www.R-project.org/.

(31) Kovalchik S. RISmed: download content from NCBI databases. R package version 2.1.7. 2017; Available at: http://www.CRAN.R-project.org/.

(32) Feinerer, I., Hornik, K., Meyer, D. Text mining infrastructure in R. J Stat Softw. 2008 Mar;25(5):1-54. doi: 10.18637/jss.v025.i05.

(33) F Escapa I, Huang Y, Chen T, Lin M, Kokaras A, Dewhirst FE, et al. Construction of habitat-specific training sets to achieve species-level assignment in 16S rRNA gene datasets. Microbiome. 2020 May;8(1):65. doi: 10.1186/s40168-020-00841-w.

(34) F Escapa I, 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;3(6):e00187-18. doi: 10.1128/mSystems.00187-18.



(35) Morgulis A, Coulouris G, Raytselis Y, Madden TL, Agarwala R, Schäffer AA. Database indexing for production MegaBLAST searches. Bioinformatics. 2008 Aug;24(16):1757-64.

(36) Chen Y, Ye W, Zhang Y, Xu Y. High speed BLASTN: an accelerated MegaBLAST search tool. Nucleic Acids Res. 2015 Sep;43(16):7762-8.

(37) NCBI Resource Coordinators. Database resources of the National Center for Biotechnology Information. Nucleic Acids Res. 2016 Jan;44(D1):D7-19. doi: 10.1093/nar/gkv1290.

(38) O'Leary NA, Wright MW, Brister JR, Ciufo S, Haddad D, McVeigh R, et al. Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. Nucleic Acids Res. 2016 Jan;44(D1): D733-45. doi: 10.1093/nar/gkv1189.

(39) Clark K, Karsch-Mizrachi I, Lipman DJ, Ostell J, Sayers EW. GenBank. Nucleic Acids Res. 2016 Jan;44:D67-72. doi: 10.1093/nar/gkv1276.

(40) 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.

(41) Python Software Foundation. Python. Version 3.9.0. 2020; Available at: http://www.python.org/.

(42) GNU P. Free Software Foundation. Bash. Version 5.1. 2020; Available at: http://www.gnu.org/.

(43) Sievers F, Wilm A, Dineen D, Gibson TJ, Karplus K, Li W, et al. Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. Mol Syst Biol. 2011 Oct;7:539. doi: 10.1038/msb.2011.75.

(44) Chen T, Yu W, 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;2010:baq013. doi: 10.1093/database/baq013.

(45) Conway Institute UCD Dublin. Clustal Omega installation instructions. 2018. http://www.clustal.org/omega/INSTALL.

(46) Cock PJ, Antao T, Chang JT, Chapman BA, Cox CJ, Dalke A, et al. Biopython: freely available Python tools for computational molecular biology and bioinformatics. Bioinformatics. 2009 Jun;25(11):1422-23.

(47) Lyalina S. Search 16S py algorithm. 2019; Available at: https://github.com/slyalina/search_16S_py.

(48) Edgar R. SEARCH_16S: A new algorithm for identifying 16S ribosomal RNA genes in contigs and chromosomes. Preprint at bioRxiv 2017:124131. doi: 10.1101/124131.

(49) National Center for Biotechnology Information. NCBI RefSeq targeted loci project. Archaea FTP. 2008. ftp://ftp.ncbi.nlm.nih.gov/refseq/TargetedLoci/Archaea/.

(50) National Center for Biotechnology Information. Entrez programming utilities help. 2010; Available at: https://www.ncbi.nlm.nih.gov/books/NBK25501/.

(51) Schoch CL, Ciufo S, Domrachev M, Hotton CL, Kannan S, Khovanskaya R, et al. NCBI Taxonomy: a comprehensive update on curation, resources and tools. Database (Oxford). 2020 Jan;2020:baaa062. doi: 10.1093/database/baaa062.

(52) Parks DH, Chuvochina M, Chaumeil P, Rinke C, Mussig AJ, Hugenholtz P. A complete domain-to-species taxonomy for bacteria and archaea. Nat Biotechnol. 2020 Sep;38(9):1079-86.

(53) Barnett M. regex. 2020; Available at: https://pypi.org/.

(54) McNamara J. xlsxwriter. 2013; Available at: https://xlsxwriter.readthedocs.io/.

(55) Baker GC, Smith JJ, Cowan DA. Review and re-analysis of domain-specific 16S primers. J Microbiol Methods. 2003 Dec;55(3):541-55.

(56) Yooseph S, Sutton G, Rusch DB, Halpern AL, Williamson SJ, Remington K, et al. The Sorcerer II Global Ocean Sampling expedition: expanding the universe of protein families. PLoS Biol. 2007 Mar;5(3):e16. doi: 10.1371/journal.pbio.0050016.

(57) Rusch DB, Halpern AL, Sutton G, Heidelberg KB, Williamson S, Yooseph S, et al. The Sorcerer II Global Ocean Sampling expedition: northwest Atlantic through eastern tropical Pacific. PLoS Biol. 2007 Mar;5(3):e77. doi: 10.1371/journal.pbio.0050077

(58) DeSantis TZ, Hugenholtz P, Larsen N, Rojas M, Brodie EL, Keller K, et al. Greengenes, a chimera-checked 16S rRNA gene database and workbench compatible with ARB. Appl Environ Microbiol. 2006 Jul;72(7):5069-72.

(59) Socransky SS, Haffajee AD, Cugini MA, Smith C, Kent RL, Jr. Microbial complexes in subgingival plaque. J Clin Periodontol. 1998 Feb;25(2):134-44.

(60) Acharya A, Chen T, Chan Y, Watt RM, Jin L, Mattheos N. Species-level salivary microbial indicators of well-resolved periodontitis: a preliminary investigation. Front Cell Infect Microbiol. 2019 Oct;9:347. doi: 10.3389/fcimb.2019.00347.

(61) Velsko IM, Harrison P, Chalmers N, Barb J, Huang H, Aukhil I, et al. Grade C molar-incisor pattern periodontitis subgingival microbial profile before and after treatment. J Oral Microbiol. 2020 Sep;12(1):1814674. doi: 10.1080/20002297.2020.

(62) Papapanou PN, Park H, Cheng B, Kokaras A, Paster B, Burkett S, et al. Subgingival microbiome and clinical periodontal status in an elderly cohort: The WHICAP ancillary study of oral health. J Periodontol. 2020 Oct; 91 Suppl 1(Suppl 1):S56-7. doi: 10.1002/JPER.20-0194.



(63) Vielkind P, Jentsch H, Eschrich K, Rodloff AC, Stingu CS. Prevalence of Actinomyces spp. in patients with chronic periodontitis. Int J Med Microbiol. 2015 Oct;305(7):682-8.

(64) Maruyama N, Maruyama F, Takeuchi Y, Aikawa C, Izumi Y, Nakagawa I. Intraindividual variation in core microbiota in peri-implantitis and periodontitis. Sci Rep. 2014 Oct;4:6602. doi: 10.1038/srep06602.

(65) Marchesan JT, Morelli T, Moss K, Barros SP, Ward M, Jenkins W, et al. Association of Synergistetes and Cyclodipeptides with periodontitis. J Dent Res. 2015 Oct;94(10):1425-31.

(66) Komatsu K, Shiba T, Takeuchi Y, Watanabe T, Koyanagi T, Nemoto T, et al. Discriminating microbial community structure between peri-implantitis and periodontitis with integrated metagenomic, metatranscriptomic, and network analysis. Front Cell Infect Microbiol. 2020 Dec;10:596490. doi: 10.3389/fcimb.2020.596490.

(67) Hiranmayi KV, Sirisha K, Ramoji Rao MV, Sudhakar P. Novel pathogens in periodontal microbiology. J Pharm Bioallied Sci. 2017;9(3):155-63.

(68) Wyss C, Choi BK, Schüpbach P, Guggenheim B, Göbel UB. Treponema maltophilum sp. nov., a small oral spirochete isolated from human periodontal lesions. Int J Syst Bacteriol. 1996 Jul;46(3):745-52.

(69) Henssge U, Do T, Radford DR, Gilbert SC, Clark D, Beighton D. Emended description of Actinomyces naeslundii and descriptions of Actinomyces oris sp. nov. and Actinomyces johnsonii sp. nov., previously identified as Actinomyces naeslundii genospecies 1, 2 and WVA 963. Int J Syst Evol Microbiol. 2009 Mar;59(Pt 3):509-16.

(70) Bor B, Bedree JK, Shi W, McLean JS, He X. Saccharibacteria (TM7) in the human oral microbiome. J Dent Res. 2019 May;98(5):500-9.

(71) Mantzourani M, Fenlon M, Beighton D. Association between Bifidobacteriaceae and the clinical severity of root caries lesions. Oral Microbiol Immunol. 2009 Feb;24(1):32-7.

(72) Mantzourani M, Gilbert SC, Sulong HN, Sheehy EC, Tank S, Fenlon M, et al. The isolation of bifidobacteria from occlusal carious lesions in children and adults. Caries Res. 2009;43(4):308-13.

(73) Skelly E, Johnson NW, Kapellas K, Kroon J, Lalloo R, Weyrich L. Response of salivary microbiota to caries preventive treatment in aboriginal and torres strait islander children. J Oral Microbiol. 2020 Oct;12(1):1830623. doi: 10.1080/20002297.2020.1830623.

(74) Caneppele TMF, de Souza LG, Spinola MDS, de Oliveira FE, de Oliveira LD, Carvalho CAT, et al. Bacterial levels and amount of endotoxins in carious dentin within reversible pulpitis scenarios. Clin Oral Investig. 2020 Oct; doi: 10.1007/s00784-020-03624-7.

(75) Belstrøm D, Holmstrup P, Fiehn NE, Kirkby N, Kokaras A, Paster BJ, et al. Salivary microbiota in individuals with different levels of caries experience. J Oral Microbiol. 2017 Jan;9(1):1270614. doi: 10.1080/20002297.2016.1270614.

(76) Jiang S, Gao X, Jin L, Lo EC. Salivary microbiome diversity in caries-free and caries-affected children. Int J Mol Sci. 2016 Nov;17(12):1978. doi: 10.3390/ijms17121978.

(77) Tanner AC, Mathney JM, Kent RL, Chalmers NI, Hughes CV, Loo CY, et al. Cultivable anaerobic microbiota of severe early childhood caries. J Clin Microbiol. 2011 Apr;49(4):1464-74.

(78) Efenberger M, Agier J, Pawłowska E, Brzezińska-Błaszczyk E. Archaea prevalence in inflamed pulp tissues. Cent Eur J Immunol. 2015;40(2):194-200.

(79) Horz HP, Seyfarth I, Conrads G. McrA and 16S rRNA gene analysis suggests a novel lineage of archaea phylogenetically affiliated with Thermoplasmatales in human subgingival plaque. Anaerobe. 2012 Jun;18(3):373-7.

(80) Keskin C, Demiryürek EÖ, Onuk EE. Pyrosequencing analysis of cryogenically ground samples from primary and secondary/persistent endodontic infections. J Endod. 2017 Aug;43(8):1309-16.

(81) Huynh HT, Verneau J, Levasseur A, Drancourt M, Aboudharam G. Bacteria and archaea paleomicrobiology of the dental calculus: a review. Mol Oral Microbiol. 2016 Jun;31(3):234-42.

(82) Huynh HT, Nkamga VD, Signoli M, Tzortzis S, Pinguet R, Audoly G, et al. Restricted diversity of dental calculus methanogens over five centuries, France. Sci Rep. 2016 May;6:25775. doi: 10.1038/srep25775.

(83) Deng ZL, Szafrański SP, Jarek M, Bhuju S, Wagner-Döbler I. Dysbiosis in chronic periodontitis: key microbial players and interactions with the human host. Sci Rep. 2017 Jun;7(1):3703. doi: 10.1038/s41598-017-03804-8.

(84) Caporaso JG, Lauber CL, Walters WA, Berg-Lyons D, Lozupone CA, Turnbaugh PJ, et al. Global patterns of 16S rRNA diversity at a depth of millions of sequences per sample. Proc Natl Acad Sci U S A. 2011 Mar;108 Suppl 1(Suppl 1):4516-22.

(85) Illumina, Inc. 16S Metagenomic Sequencing Library Preparation. 2013; Available at: https://support.illumina.com/content/dam/illumina-support/documents/documentation/chemistry_documentation/16s/16s-metagenomic-library-prep-guide-15044223-b.pdf.

(86) Damgaard C, Danielsen AK, Enevold C, Massarenti L, Nielsen CH, Holmstrup P, et al. Porphyromonas gingivalis in saliva associates with chronic and aggressive periodontitis. J Oral Microbiol. 2019 Aug;11(1):1653123. doi: 10.1080/20002297.2019.

(87) Li Y, Feng X, Xu L, Zhang L, Lu R, Shi D, et al. Oral microbiome in chinese patients with aggressive periodontitis and their family members. J Clin Periodontol. 2015 Nov;42(11):1015-23.

(88) Chen H, Liu Y, Zhang M, Wang G, Qi Z, Bridgewater L, et al. A Filifactor alocis-centered co-occurrence group associates with periodontitis across different oral habitats. Sci Rep. 2015 Mar;5:9053. doi: 10.1038/srep09053.

(89) Belstrom D, Sembler-Moller ML, Grande MA, Kirkby N, Cotton SL, Paster BJ, et al. Microbial profile comparisons of saliva, pooled and site-specific subgingival samples in periodontitis patients. PLoS One. 2017 Aug;12(8):e0182992. doi: 10.1371/journal.pone.0182992.

(90) Dieffenbach CW, Lowe TM, Dveksler GS. General concepts for PCR primer design. PCR Methods Appl. 1993 Dec;3(3):S30-7. doi: 10.1101/gr.3.3.s30.

# OBJECTIVE 2

# Objective 2. Impact of 16S rRNA gene redundancy and primer pair selection on the quantification and classification of oral microbiota in next-generation sequencing


## 2.1. ABSTRACT

**Aims:** To evaluate the number of 16S rRNA genes in the complete genomes of all the bacterial and archaeal species ever detected in the human oral cavity; and to assess how the use of different primer pairs would affect the detection and classification of redundant amplicons and matching amplicons (MAs) from different taxa.

**Material and methods:** A total of 709 complete genomes (518 oral bacteria, 191 oral archaea) were downloaded from the NCBI database, and their complete 16S rRNA genes were extracted. The total number of genes and variants per genome were calculated. Next, 33 primer pairs selected from a previous reserach, and 6 commonly employed in the oral literature were used against all the genomes to obtain amplicons. For each primer pair, we calculated the number of 16S rRNA gene amplicons, variants, genomes, and species detected, as well as the percentage of coverage at the species level with no matching amplicons (SC-NMA).

**Results:** 94.1% of oral bacteria and 52.59% of oral archaea had more than one 16S rRNA gene in their respective genomes. Between 46.70% - 1.29% of the bacterial species and between 38.89% - 4.65% of the archaeal species detected by the evaluated primer pairs had MAs, affecting relevant genera present in the oral environment such as *Actinomyces, Fusobacterium, Lactobacillus, Methanosarcina, Staphylococcus,* and *Streptococcus.* The best primer pairs were (SC-NMA; region; primer pair position for *Escherichia coli* J01859.1): KP_F048-OP_R030 for bacteria (93.55%; 3-7; 342-1079), KP_F018-KP_R063 for archaea (89.63%; 3-9; undefined-1506), and OP_F114-OP_R121 for both bacteria and archaea (92.52%; 3-9; 340-1405).

**Conclusions**: In addition to the 16S rRNA gene redundancy, the considerable presence of MAs must be controlled to ensure the accurate interpretation of microbial diversity data. The

SC-NMA is a more useful parameter than the conventional coverage percentage for selecting the best primer pairs. The performance of the primer pairs to detect non-MA species increases as the average length of the amplicons increases; none of these being the most widely used primer pairs in the oral literature. The choice of primer pair significantly affects diversity estimates and taxonomic classification, conditioning the comparability of oral microbiome studies using different primer pairs.

### 2.1.1. Keywords

16S rRNA gene, gene variant, matching amplicon, oral microbiota, overestimation factor, primer, redundancy, sequence analysis.

### 2.1.2. Declaration of conflict of interest

The doctoral candidate and the rest of the authors of the present study declare that they have no conflict of interest concerning the objectives proposed in this chapter.

### 2.1.3. Funding

This investigation was supported by the Instituto de Salud Carlos III (General Division of Evaluation and Research Promotion, Madrid, Spain) and co-financed by the FEDER (European Regional Development Fund, ERDF) ("A way of making Europe") under grant ISCIII/PI17/01722; the Consellería de Cultura, Educación e Ordenación Universitaria de la Xunta de Galicia (accreditation 2019-2022 ED431G-2019/04, group with growth potential ED431B 2020-2022 GPC2020/27; A. Regueira-Iglesias support ED481A-2017/233) and the ERDF, which acknowledges the CiTIUS-Research Center in Intelligent Technologies of the Santiago de Compostela University as a Research Center of the Galician University System.

The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

## 2.2. INTRODUCTION

The 16S ribosomal RNA (rRNA) gene has been widely used to estimate bacterial diversity in different environments (1) ever since its promotion as an "evolutionary clock" some three decades ago (2). This gene, which has an average length of approximately 1,500 base pairs (bps), has several characteristics that have led to its identification as a reliable phylogenetic marker. These are: the ubiquitous presence of the 16S rRNA gene in bacteria and archaea; its relative stability in combining conserved and hypervariable regions; and the existence of complete and easily accessible databases (3).

However, the use of the 16S rRNA gene does not come without limitations, and various investigations have demonstrated the existence of up to 15 gene copies per genome in bacteria (4-8) and up to four in archaea (4,6,9). It is well known that this intragenomic gene redundancy affects estimates of microbial abundance that are based on gene counts (4,7). Overall, there is a tendency for the taxa with a low number of 16S rRNA genes to be underestimated, while those with high numbers are overestimated (7). In addition, as the different gene regions do not have the same levels of sequence heterogeneity (6,10), the primer pair employed in the amplification stage may influence both the detection of redundant amplicons as well as matching amplicons (MAs) from different taxa.

A recent study has reported the existence of a maximum of four different genes per genome (hereafter: genes/genome) in 32 species isolated from periodontal abscesses (11). However, this limited approach does not reflect the complexity of the oral microbiota where around 700 species have been identified (12,13). On the other hand, the identification, at least at the species level, is highly desirable in 16S rRNA sequencing-based studies of the oral microbiome (14). This is because it has been demonstrated how different species from the same genus are associated with different oral conditions (15-17). Our results revealed that *Porphyromonas catoniae* is a core species linked to dental and periodontal health, while *Porphyromonas endodontalis* is associated with dental and periodontal pathology. About the differential abundance data, while *Fusobacterium periodonticum* is present in significantly higher numbers in the dentally healthy, *Fusobacterium nucleatum* subsp. *vicentii* is present in significantly higher numbers in individuals with high degrees of dental pathology (15). However, the taxonomic resolution at the species level could be affected by the presence of MAs.

To the best of our knowledge, there has not yet been an exhaustive *in silico* evaluation of the number of 16S rRNA genes present in the complete genomes of the bacteria and archaea inhabiting the human mouth. Moreover, we have been unable to identify any study in the oral microbiology field that has examined the impact of which primer pair is selected for use to detect and classify redundant amplicons. Consequently, the aims of this investigation were to: 1) evaluate the number of 16S rRNA genes in the complete genomes of all the bacterial and archaeal species ever detected in the human oral cavity; and 2) assess how the use of different primer pairs would affect the detection of redundant amplicons and MAs from different taxa.

## 2.3. MATERIAL AND METHODS

### 2.3.1. Obtaining complete oral-bacteria and oral-archaea genomes

All the information available on the bacterial taxa present in the oral cavity was obtained from the expanded human oral microbiome database (eHOMD) website (18). All genomes with the complete sequencing status indicated by eHOMD were chosen. A total of 518 complete genomes were identified among 2074 on the eHOMD website.

The complete genomes indicated in the eHOMD, have one or more Genbank (19) identifiers, which were used to access the complete sequences stored in the National Center for Biotechnology Information (NCBI) database (20). In general, these complete genomes consisted of one or two identifiers corresponding to their circular chromosomes; in many cases, however, the genomes had plasmid identifiers as well, which were also investigated.

An initial list of 177 different oral archaea and their corresponding GenBank (19) identifiers, obtained as part of a objective 1 (21), enabled us to access their complete sequences in the NCBI database (20).

Integrating the "Entrez Programming Utilities (E-utilities)" tool (22) in the Python (version 3.9.0) (23) script allowed us to acquire the URLs needed to retrieve the information of interest from the various NCBI databases, including Taxonomy (24), RefSeq (25), and GenBank (19). The oral-bacteria and oral-archaea genomes were then downloaded, and finally, the taxonomy of each of them was obtained.

### 2.3.2. Detection and extraction of 16S rRNA genes

There were a number of International Union of Pure and Applied Chemistry (IUPAC) non-specific nucleotides distributed along some of the genomes. Consequently, we developed a Python (23) script to detect and then randomly replace them with one of the specific equivalent nucleotides. Other genomes were excluded because they had an excess of IUPAC nucleotides, mainly of "N" bases.



Our Python (23) script was completed with a free downloadable module known as search_16S.py (26), which is based on Edgar's algorithm (27). This algorithm looks for the 16S

rRNA genes in the genomes, identifying sections with a high frequency of 13-mers in known 16S rRNA genes and then, searches within each segment for conserved motifs close to the beginning and end of the gene. The obtention of a pair of motifs within the expected length range confirms the presence of the gene and provides consistent and homologous endpoints (25). Applying this algorithm, the 16S rRNA gene sequences from the complete downloaded genomes were detected and extracted, while the variants were stored in a fasta file. All the 16S rRNA gene variants identified were designated taxonomically at the strain or the species level if no designated strain name existed. This left us with the following for inclusion in subsequent analyses: 518 oral-bacteria genomes, corresponding to 186 species; and 191 oral-archaea genomes, corresponding to 135 species. Their taxonomy and NCBI identifiers are included in appendices S1 and S2, respectively.

For each genome evaluated, we calculated: its size; the sizes of the 16S rRNA genes detected; the total number of 16S rRNA genes; the number of different variants; and the number of 16S rRNA genes in each strand. The averages of the data obtained were subsequently determined using Python's NumPy (28) and pandas modules (29) for hierarchical levels above the strain level.

### 2.3.3. Evaluation of a selection of primer pairs for the detection of 16S rRNA genes

We selected the primer pairs with the best *in silico* coverage values, as identified in objective 1, as well as those used most in the oral microbiome literature (21). This left us with 33 and 6 primer pairs, respectively, for this stage of the study, which were classified according to the average length of the amplicons into: short (S; 100-300 bps), medium (M; 301-600 bps), and long (L; >600 bps) primer pairs (21) (Table 1).

**Table 1. Selected primer pairs with high *in silico* coverage percentages targeting oral bacteria and/or archaea and those most used in the sequencing-based studies of the oral microbiome.**

| Bacterial-specific primer pair | ALC | F identifier | F Sequence 5-3 | F First post | F Last Post | R identifier | R Sequence 5-3 | R First post | R Last Post | Length (bps) | Region |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | S | KP_F048 | TACGGRAGGCAGCAG | 342 | 356 | OP_R043 | CCGCGRCTGCTGGCAC | 514 | 529 | 187 | 3-4 |
| | | OP_F098 | CCAGCAGCYGCGGTAAN | 517 | 533 | OP_R119 | GGACTACCRGGGTATCTAA | 787 | 805 | 288 | 4-5 |
| | | OP_F066 | GGMTTAGATACCC | 784 | 796 | KP_R040 | CCGTCAATTCMTTTGAGTTT | 906 | 925 | 141 | 5-6 |
| | | OP_F009 | GGATTAGATACCCBRGTAGTC | 784 | 804 | OP_R030 | TCACRRCACGAGCTGWCGAC | 1060 | 1079 | 295 | 5-7 |
| | | KP_F061 | ACTCAAAKGAATWGACGG | 908 | 925 | KP_R074 | GGGTYKCGCTCGTTR | 1099 | 1113 | 205 | 6-7 |
| | | OP_F101 | GAATTGRCGGGGRCC | 916 | 930 | OP_R030 | TCACRRCACGAGCTGWCGAC | 1060 | 1079 | 163 | 6-7 |
| | M | OP_F053 | GRGTTYGATYMTGGCTCAG | 9 | 27 | KP_R020 | CTGCTGCCTYCCGTA | 342 | 356 | 347 | 1-3 |
| | | KP_F048 | TACGGRAGGCAGCAG | 342 | 356 | KP_R031 | TACHVGGGTATCTAAKCC | 784 | 801 | 459 | 3-5 |
| | | KP_F048 | TACGGRAGGCAGCAG | 342 | 356 | OP_R073 | CRTACTHCHCAGGYG | 879 | 893 | 551 | 3-6 |
| | | KP_F051 | GTGCCAGCMGCNGCGG | 514 | 529 | KP_R041 | CGTCAATTCMTTTGAGTT | 907 | 924 | 410 | 4-6 |
| | | KP_F051 | GTGCCAGCMGCNGCGG | 514 | 529 | OP_R030 | TCACRRCACGAGCTGWCGAC | 1060 | 1079 | 565 | 4-7 |
| | | OP_F116 | YAACGAGCGCAACCC | 1099 | 1113 | KP_R060 | GACGGGCGGTGWGTRCA | 1390 | 1406 | 307 | 7-9 |
| | L | KP_F048 | TACGGRAGGCAGCAG | 342 | 356 | OP_R030 | TCACRRCACGAGCTGWCGAC | 1060 | 1079 | 737 | 3-7 |
| | | KP_F048 | TACGGRAGGCAGCAG | 342 | 356 | KP_R060 | GACGGGCGGTGWGTRCA | 1390 | 1406 | 1064 | 3-9 |
| | | KP_F056 | AYTGGGYDTAAAGNG | 572 | 576 | KP_R077 | GACGGGCGGTGTGTACAA | 1389 | 1406 | 834 | 4-9 |
| **Archaeal-specific primer pair** | **ALC** | **F identifier** | **F Sequence 5-3** | **F First post** | **F Last Post** | **R identifier** | **R Sequence 5-3** | **R First post** | **R Last Post** | **Length (bps)** | **Region** |
| | S | KP_F018 | GYGCASCAGKCGMGAAW | U | U | KP_R002 | TTACCGCGGCKGCTG | 518 | 532 | - | - 4* |
| | | OP_F066 | GGMTTAGATACCC | 784 | 796 | KP_R013 | GGCCATGCACCWCCTCTC | U | U | - | 5-6 |
| | M | KP_F018 | GYGCASCAGKCGMGAAW | U | U | KP_R032 | TACNVGGGTATCTAATCC | 784 | 801 | - | 3-5 |
| | | KP_F018 | GYGCASCAGKCGMGAAW | U | U | OP_R073 | CRTACTHCHCAGGYG | 879 | 893 | - | 3-5 |
| | | KP_F020 | CAGCMGCCGCGGTAA | 518 | 532 | KP_R013 | GGCCATGCACCWCCTCTC | U | U | - | 3-6 |
| | | KP_F022 | AGGAATTGGCGGGGGAGCA | U | U | KP_R063 | TACCTTGTTACGACTT | 1491 | 1506 | - | 5-9 |
| | L | OP_F114 | CCTAYGGGRBGCASCAG | 340 | 356 | KP_R013 | GGCCATGCACCWCCTCTC | U | U | - | 3-6 |
| | | KP_F018 | GYGCASCAGKCGMGAAW | U | U | KP_R063 | TACCTTGTTACGACTT | 1491 | 1506 | - | 3-9 |
| | | OP_F066 | GGMTTAGATACCC | 784 | 796 | OP_R016 | CGGTGTGTGCAAGGAG | U | U | - | 5-9 |
| **Bacterial and archaeal primer pair** | **ALC** | **F identifier** | **F Sequence 5-3** | **F First post** | **F Last Post** | **R identifier** | **R Sequence 5-3** | **R First post** | **R Last Post** | **Length (bps)** | **Region** |
| | S | OP_F114 | CCTAYGGGRBGCASCAG | 340 | 356 | KP_R002 | TTACCGCGGCKGCTG | 518 | 532 | 192 | 3-4 |
| | | KP_F020 | CAGCMGCCGCGGTAA | 518 | 532 | KP_R032 | TACNVGGGTATCTAATCC | 784 | 801 | 283 | 4-5 |
| | | OP_F066 | GGMTTAGATACCC | 784 | 796 | OP_R073 | CRTACTHCHCAGGYG | 879 | 893 | 109 | 5-6 |
| | M | OP_F114 | CCTAYGGGRBGCASCAG | 340 | 356 | KP_R031 | TACHVGGGTATCTAAKCC | 784 | 801 | 461 | 3-5 |
| | | OP_F114 | CCTAYGGGRBGCASCAG | 340 | 356 | OP_R073 | CRTACTHCHCAGGYG | 879 | 893 | 553 | 3-6 |
| | | KP_F020 | CAGCMGCCGCGGTAA | 518 | 532 | OP_R073 | CRTACTHCHCAGGYG | 879 | 893 | 375 | 4-6 |
| | L | OP_F114 | CCTAYGGGRBGCASCAG | 340 | 356 | OP_R121 | ACGGGCGGTGWGTRC | 1391 | 1405 | 1065 | 3-9 |
| | | KP_F020 | CAGCMGCCGCGGTAA | 518 | 532 | OP_R121 | ACGGGCGGTGWGTRC | 1391 | 1405 | 887 | 4-9 |
| | | OP_F066 | GGMTTAGATACCC | 784 | 796 | OP_R121 | ACGGGCGGTGWGTRC | 1391 | 1405 | 621 | 5-9 |
| **Most used primer pair** | **ALC** | **F identifier** | **F Sequence 5-3** | **F First post** | **F Last Post** | **R identifier** | **R Sequence 5-3** | **R First post** | **R Last Post** | **Length (bps)** | **Region** |
| | S | KP_F078 | GTGCCAGCMGCCGCGGTAA | 514 | 532 | OP_R010 | GGACTACHVGGGTWTCTAAT | 786 | 805 | 291 | 4-5 |
| | M | KP_F031 | AGAGTTTGATCCTGGCTCAG | 8 | 27 | KP_R021 | TTACCGCGGCTGCTGGCAC | 515 | 532 | 524 | 1-4 |
| | | KP_F047 | CCTACGGGNGGCWGCAG | 340 | 356 | KP_R035 | GACTACHVGGGTATCTAATCC | 784 | 804 | 464 | 3-5 |
| | | OP_F009 | GGATTAGATACCCBRGTAGTC | 784 | 868 | OP_R029 | ACGTCRTCCCCDCCTTCCTC | 1174 | 1193 | 409 | 5-8 |
| | L | KP_F014 | TCCAGGCCCTACGGG | U | U | KP_R011 | YCCGGCGTTGAMTCCAATT | U | U | - | 3-6 |
| | | KP_F034 | AGAGTTTGATCMTGGCTCAG | 8 | 27 | KP_R065 | TACGGYTACCTTGTTACGACTT | 1491 | 1512 | 1504 | 1-9** |

Primer pairs were selected based on the species coverage values (number of species detected /total species evaluated) in objective 1 (21). They were individually evaluated through regular expressions against *Escherichia coli* J01859 to define their positions. The U values represent a mismatch on the assessment and, therefore, the position cannot be confirmed with a guarantee. Gene regions were delimited as described by Baker et al. (30). *The primer pair gene region was 3 and **1-10 when considering the mode positions obtained in the analysis against the oral-bacteria and oral-archaea databases in objective 1 (21).

ALC= amplicon length category; bps= base pairs; F= forward; KP= Klindworth primer; L= long mean amplicon length category, >600 base pairs; M= medium mean amplicon length category, 301-600 base pairs; OP= oral primer; Post= position; R= reverse; S= short mean amplicon length category, 100-300 base pairs; U= unaligned with Escherichia coli; V= variable.

The direct and reverse sequences of each primer pair selected were used in combination with Python's regex module (31) to obtain, *in silico*, the amplicons of the 16S rRNA genes identified in all of the chosen genomes. For each primer pair, we determined: the mean size and number of the 16S rRNA gene amplicons; the number of gene variants; the number of genomes and species detected; and the percentage of coverage at the species level with no MAs (SC-NMA). This coverage value was calculated as:

SC-NMA (%)= [(Number of species detected - Number of species with MAs)/Total number of species evaluated] x 100

The overestimation of abundance at the species level (the overestimation factor -OF-) was also calculated. This represented for each species the combination of the number of copies of the 16S rRNA gene amplicons and the number of MAs. To remove the overestimation derived from the intragenomic gene redundancy, the OF of each species was divided by the number of gene copies, resulting in OF caused by the presence of MAs (OF-MA). Species with values equal to 1.00 did not have amplicons that matched other species for the corresponding primer pair, while those with estimates greater than 1.00 did. For each primer pair, both parameters were expressed cumulatively and as an average. The best primer pairs selected first were those with the highest SC-NMA value and of these, those with the lowest OF-MA value. The worst primer pairs were those with the lowest SC-NMA and the highest OF-MA.

## 2.4. RESULTS

### 2.4.1. Number of intragenomic 16S rRNA genes in oral-bacteria and oral-archaea genomes

Table 2 details the mean number of intragenomic 16S rRNA genes in the bacterial and archaeal phyla through seven taxonomic ranks. The 518 oral-bacteria genomes examined had a mean size of 2,933,660.68 bps and an average number of 4.55 intragenomic 16S rRNA genes, which in turn had a mean size of 1,501.32 bps and an average of 2.60 variants. Eleven of the 186 bacterial species (5.91%) had one gene/genome, 159 species (85.49%) showed a mean between two and six genes, and 16 species (8.60%), mean values of seven or more genes. The maximum mean number of intragenomic 16S rRNA genes observed was 10.83 in *Bacillus anthracis*, with five strains of this species having a total of 11 genes/genome. Concerning the average number of intragenomic gene variants, 63 bacterial species (33.87%) presented one variant/genome, 118 species (63.44%), between two and six, and five species (2.69%), seven or more.

The 191 oral-archaea genomes had a mean size of 2,545,441.40 bps and an average of 1.95 intragenomic 16S rRNA genes, which in turn had a mean size of 1,471.25 bps and an average of 1.44 variants. Sixty-four out of the 135 archaeal species (47.41%) had a mean of one gene/genome, 67 species (49.63%) showed an average between two and three genes, and 4 species (2.96%) mean values above three (*Methanobacterium formicicum*, *Methanococcus vannielii*, *Methanosphaera stadtmanae,* and *Methanospirillum hungatei)*. At the strain level, the maximum total number of genes/genome increased to five in *Methanococcus maripaludis* (unknown strain) and *Sulfolobus acidocaldarius* (unknown strain). Concerning the average number of intragenomic gene variants, 93 species (68.89%) had an average number of one variant/genome and 42 (31.11%) had between two and three.

Appendices S3 and S4 contain the sizes of the bacterial and archaeal genomes and genes, the number of genes/genome, and the number of gene variants/genome across eight taxonomic ranks.

**Table 2. Intragenomic 16S rRNA genes in the bacterial and archaeal phyla through seven taxonomy ranks.**

| | Mean number of intragenomic 16S rRNA genes | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | Taxonomy level | | | | | | | |
| Phylum | Phylum | Class | Order | Family | Genera | Species | Strain | No. genomes |
| Actinobacteria | 3.12 | 3.19 - 2.00 | 3.41 - 1.33 | 4.55 - 1.10 | 4.55 - 1.00 | 5.00 - 1.00 | 5 - 1 | 91 |
| Bacteroidetes | 3.68 | 4.75 - 3.44 | 4.75 - 3.44 | 4.75 - 2.00 | 4.75 - 2.00 | 7.00 - 2.00 | 7 - 2 | 24 |
| C. Saccharibacteria | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1 | 1 |
| Chlamydiae | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1 - 1 | 5 |
| Chlorobi | 2.00 | 2.00 | 2.00 | 2.00 | 2.00 | 2.00 | 2 | 1 |
| Chloroflexi | 2.00 | 2.00 - 2.00 | 2.00 - 2.00 | 2.00 - 2.00 | 2.00 - 2.00 | 2.00 - 2.00 | 2 - 2 | 2 |
| Firmicutes | 5.43 | 5.52 - 3.25 | 6.61 - 3.25 | 9.85 - 2.00 | 10.18 - 2.00 | 10.83 - 2.00 | 11 - 2 | 177 |
| Fusobacteria | 4.35 | 4.35 | 4.35 | 4.40 - 4.33 | 4.75 - 3.00 | 5.00 - 3.00 | 5 - 2 | 21 |
| Ignavibacteriae | 1.00 | 1.00 | 1.00 | 1.00 - 1.00 | 1.00 - 1.00 | 1.00 - 1.00 | 1 - 1 | 2 |
| Proteobacteria | 5.21 | 6.13 - 2.17 | 6.98 - 2.17 | 7.14 - 2.00 | 8.00 - 2.00 | 8.00 - 2.00 | 8 - 2 | 170 |
| Spirochaetes | 2.00 | 2.00 | 2.00 | 2.00 | 2.00 | 2.00 - 2.00 | 2 - 2 | 11 |
| Tenericutes | 1.23 | 1.23 | 1.23 | 1.23 | 1.67 - 1.10 | 2.00 - 1.00 | 2 - 1 | 13 |
| C. Thermoplasmatota | 1.00 | 1.00 | 1.00 - 1.00 | 1.00 - 1.00 | 1.00 - 1.00 | 1.00 - 1.00 | 1 - 1 | 7 |
| Crenarchaeota | 1.10 | 1.10 | 1.25 - 1.00 | 1.25 - 1.00 | 1.29 - 1.00 | 2.00 - 1.00 | 5 - 1 | 43 |
| Euryarchaeota | 2.29 | 2.67 - 1.00 | 2.89 - 1.00 | 4.00 - 1.00 | 4.00 - 1.00 | 4.00 - 1.00 | 5 - 1 | 138 |
| Thaumarchaeota | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 - 1.00 | 1 - 1 | 3 |
| | Mean number of intragenomic 16S rRNA gene variants | | | | | | | |
| | Taxonomy level | | | | | | | |
| Phylum | Phylum | Class | Order | Family | Genera | Species | Strain | No. genomes |
| Actinobacteria | 1.54 | 1.56 - 1.20 | 2.00 - 1.00 | 3.00 - 1.00 | 4.00 - 1.00 | 4.00 - 1.00 | 4 - 1 | 91 |
| Bacteroidetes | 1.77 | 2.50 - 1.61 | 2.50 - 1.61 | 2.50 - 1.00 | 2.50 - 1.00 | 5.00 - 1.00 | 5 - 1 | 24 |
| C. Saccharibacteria | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1 | 1 |
| Chlamydiae | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1 - 1 | 5 |
| Chlorobi | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1 | 1 |
| Chloroflexi | 1.50 | 2.00 - 1.00 | 2.00 - 1.00 | 2.00 - 1.00 | 2.00 - 1.00 | 2.00 - 1.00 | 2 - 1 | 2 |
| Firmicutes | 3.18 | 3.50 - 1.75 | 4.85 - 1.75 | 8.00 - 1.00 | 9.00 - 1.00 | 9.00 - 1.00 | 10 - 1 | 177 |
| Fusobacteria | 3.55 | 3.55 | 3.55 | 3.73 - 3.00 | 3.73 - 3.00 | 5.00 - 1.00 | 5 - 1 | 21 |
| Ignavibacteriae | 1.00 | 1.00 | 1.00 | 1.00 - 1.00 | 1.00 - 1.00 | 1.00 - 1.00 | 1 - 1 | 2 |
| Proteobacteria | 2.87 | 3.55 - 1.00 | 4.93 - 1.00 | 5.45 - 1.00 | 6.17 - 1.00 | 8.00 - 1.00 | 8 - 1 | 170 |
| Spirochaetes | 1.36 | 1.36 | 1.36 | 1.36 | 1.36 | 2.00 - 1.22 | 2 - 1 | 11 |
| Tenericutes | 1.15 | 1.15 | 1.15 | 1.15 | 1.67 - 1.00 | 1.67 - 1.00 | 2 - 1 | 13 |
| C. Thermoplasmatota | 1.00 | 1.00 | 1.00 - 1.00 | 1.00 - 1.00 | 1.00 - 1.00 | 1.00 - 1.00 | 1 - 1 | 7 |
| Crenarchaeota | 1.00 | 1.00 | 1.00 - 1.00 | 1.00 - 1.00 | 1.00 - 1.00 | 1.00 - 1.00 | 1 - 1 | 43 |
| Euryarchaeota | 1.61 | 1.86 - 1.00 | 2.00 - 1.00 | 3.00 - 1.00 | 3.00 - 1.00 | 3.00 - 1.00 | 3 - 1 | 138 |
| Thaumarchaeota | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 - 1.00 | 1 - 1 | 3 |

Ranges at the strain level are not mean values, they correspond to the maximum and minimum numbers of intragenomic genes in all strains from a given phylum.
C. Saccharibacteria= Candidatus Saccharibacteria; C. Thermoplasmatota= Candidatus Thermoplasmatota; No.= number.

### 2.4.2. Evaluation of the primer pairs taken from our previous research and those used most in oral microbiome studies

Tables 3 and 4 detail the size and number of 16S rRNA gene amplicons detected by the primer pairs in the oral-bacteria and oral-archaea genomes. The mean number of 16S rRNA gene amplicons varied from 4.84 to 4.39 for bacteria (mean amplicon variants/genome= 2.69 to 1.09) and 2.43 to 1.58 for archaea (mean amplicon variants/genome= 1.34 to 1.08). All the primer combinations identified the maximum mean numbers of intragenomic genes for the bacterial and archaeal species examined (10.83 and 4.00, respectively). However, although most of the primer pairs were able to detect the highest mean value of the gene variants/genome for the archaeal species (i.e., 3.00), only one primer pair detected this maximum value for bacterial species (i.e., 9.00).

**Table 3. Size and number of 16S rRNA gene amplicons detected by the primer pairs in the oral-bacteria genomes.**

| | | | Superkingdom level | | | Species level | | |
|---|---|---|---|---|---|---|---|---|
| ALC | Bacteria-specific primer pairs | Gene region | Amplicon length (mean, bps) | g/G (mean) | gv/G (mean) | Amplicon length (mean, range, bps) | g/G (mean, range) | gv/G (mean, range) |
| S | KP_F048-OP_R043 | 3 - 4 | 182.99 | 4.62 | 1.25 | 190.00 – 162.00 | 10.83 – 1.00 | 4.00 – 1.00 |
| | OP_F098-OP_R119 | 4 - 5 | 288.86 | 4.68 | 1.22 | 290.00 - 287.98 | 10.83 – 1.00 | 3.00 – 1.00 |
| | OP_F066-KP_R040 | 5 - 6 | 142.07 | 4.84 | 1.11 | 152.00 – 135.00 | 10.83 – 1.00 | 2.00 – 1.00 |
| | OP_F009-OP_R030 | 5 - 7 | 296.13 | 4.60 | 1.38 | 307.00 – 283.00 | 10.83 – 1.00 | 5.00 – 1.00 |
| | KP_F061-KP_R074 | 6 - 7 | 206.30 | 4.76 | 1.31 | 212.00 – 202.00 | 10.83 – 1.00 | 4.00 – 1.00 |
| | OP_F101-OP_R030 | 6 - 7 | 164.06 | 4.77 | 1.31 | 170.00 – 160.00 | 10.83 – 1.00 | 3.00 – 1.00 |
| M | OP_F053-KP_R020 | 1 - 3 | 351.74 | 4.61 | 1.97 | 547.00 – 315.00 | 10.83 – 1.00 | 8.00 – 1.00 |
| | KP_F048-KP_R031 | 3 – 5 | 454.84 | 4.61 | 1.42 | 462.00 – 433.00 | 10.83 – 1.00 | 5.00 – 1.00 |
| | KP_F048-OP_R073 | 3 - 6 | 546.81 | 4.67 | 1.49 | 554.20 – 520.00 | 10.83 – 1.00 | 5.00 – 1.00 |
| | KP_F051-KP_R041 | 4 - 6 | 410.90 | 4.84 | 1.32 | 421.00 – 404.00 | 10.83 – 1.00 | 3.00 – 1.00 |
| | KP_F051-OP_R030 | 4 - 7 | 566.17 | 4.62 | 1.55 | 577.00 – 552.00 | 10.83 – 1.00 | 5.00 – 1.00 |
| | OP_F116_KP_R060 | 7 - 9 | 308.84 | 4.54 | 1.35 | 1012.50 – 285.00 | 10.83 – 1.00 | 4.00 – 1.00 |
| L | KP_F048-OP_R030 | 3 - 7 | 733.00 | 4.61 | 1.71 | 742.56 – 707.00 | 10.83 – 1.00 | 5.00 – 1.00 |
| | KP_F048-KP_R060 | 3 – 9 | 1059.48 | 4.59 | 1.93 | 1070.00 – 1016.00 | 10.83 – 1.00 | 6.00 – 1.00 |
| | KP_F056-KP_R077 | 4 – 9 | 846.21 | 4.67 | 1.81 | 1551.50 – 821.00 | 10.83 – 1.00 | 6.00 – 1.00 |
| | | | Superkingdom level | | | Species level | | |
| ALC | Bacterial and archaeal primer pairs | Gene region | Amplicon length (mean, bps) | g/G (mean) | gv/G (mean) | Amplicon length (mean, range, bps) | g/G (mean, range) | gv/G (mean, range) |
| S | OP_F114-KP_R002 | 3 - 4 | 188.27 | 4.74 | 1.27 | 194.50 – 167.00 | 10.83 – 1.00 | 4.00 – 1.00 |
| | KP_F020-KP_R032 | 4 - 5 | 283.86 | 4.71 | 1.22 | 285.00 – 282.98 | 10.83 – 1.00 | 3.00 – 1.00 |
| | OP_F066-OP_R073 | 5 - 6 | 110.11 | 4.65 | 1.09 | 120.00 – 101.00 | 10.83 – 1.00 | 2.00 – 1.00 |
| M | OP_F114-KP_R031 | 3 - 5 | 456.84 | 4.60 | 1.42 | 464.00 – 435.00 | 10.83 – 1.00 | 5.00 – 1.00 |
| | OP_F114-OP_R073 | 3 - 6 | 548.82 | 4.67 | 1.49 | 556.20 – 522.00 | 10.83 – 1.00 | 5.00 – 1.00 |
| | KP_F020-OP_R073 | 4 - 6 | 375.93 | 4.72 | 1.29 | 386.00 – 366.00 | 10.83 – 1.00 | 3.00 – 1.00 |
| L | OP_F114-OP_R121 | 3 - 9 | 1060.47 | 4.59 | 1.93 | 1071.00 – 1017.00 | 10.83 – 1.00 | 6.00 – 1.00 |
| | KP_F020-OP_R121 | 4 - 9 | 889.30 | 4.70 | 1.82 | 1594.50 – 864.00 | 10.83 – 1.00 | 6.00 – 1.00 |
| | OP_F066-OP_R121 | 5 - 9 | 623.05 | 4.55 | 1.65 | 1328.50 – 598.00 | 10.83 – 1.00 | 6.00 – 1.00 |
| | | | Superkingdom level | | | Species level | | |
| ALC | Most used primer pairs | Gene region | Amplicon length (mean, bps) | g/G (mean) | gv/G (mean) | Amplicon length (mean, range, bps) | g/G (mean, range) | gv/G (mean, range) |
| S | KP_F078-OP_R010$^{B+A}$ | 4 – 5 | 291.86 | 4.71 | 1.22 | 293.00 – 290.98 | 10.83 – 1.00 | 3.00 – 1.00 |
| M | KP_F031-KP_R021$^{B}$ | 1 - 4 | 525.59 | 4.39 | 2.03 | 700.00 – 467.00 | 10.83 – 1.00 | 8.00 – 1.00 |
| | KP_F047-KP_R035$^{B}$ | 3 - 5 | 460.25 | 4.61 | 1.42 | 467.00 – 438.00 | 10.83 – 1.00 | 5.00 – 1.00 |
| | OP_F009-OP_R029$^{B}$ | 5 - 8 | 409.87 | 4.76 | 1.51 | 417.00 – 395.00 | 10.83 – 1.00 | 5.00 – 1.00 |
| L | KP_F034-KP_R065$^{B}$ | 1 – 9* | 1505.85 | 4.81 | 2.69 | 1677.00 – 1429.00 | 10.83 – 1.00 | 9.00 – 1.00 |

The amplicon length category and gene regions were determined in objective 1 according to the mean size of the amplicons generated by a given primer and to the mode first position of the forward primer and the mode last of the reverse primer, respectively. The most commonly used primer pairs in the literature were detected in objective 1 (21). *The primer pair gene region was 1-10 when considering the mode positions obtained in the analysis against the oral-bacteria and oral-archaea databases in objective 1 (21).
A= archaea; ALC= amplicon length category; B= bacteria; bps= base pairs; F= forward; g/G= number of 16S rRNA gene amplicons per genome; gv/G= number of 16S rRNA gene variant amplicons per genome; KP= Klindworth primer; L= long mean amplicon length category, >600 base pairs; M= medium mean amplicon length category, 301-600 base pairs; OP= oral primer; R= reverse; S= short mean amplicon length category,100-300 base pairs.

**Table 4. Size and number of 16S rRNA gene amplicons detected by the primer pairs in the oral-archaea genomes.**

| | | | Superkingdom level | | | Species level | | |
|---|---|---|---|---|---|---|---|---|
| ALC | Archaea-specific primer pairs | Gene region | Amplicon length (mean, bps) | g/G (mean) | gv/G (mean) | Amplicon length (mean, range, bps) | g/G (mean, range) | gv/G (mean, range) |
| S | KP_F018-KP_R002 | 4* | 154.68 | 1.96 | 1.09 | 860.00 - 138.00 | 4.00 - 1.00 | 3.00 - 1.00 |
| | OP_F066-KP_R013 | 5 - 6 | 274.32 | 1.99 | 1.11 | 277.00 - 274.00 | 4.00 - 1.00 | 2.00 - 1.00 |
| M | KP_F018-KP_R032 | 3 - 5 | 429.16 | 1.95 | 1.18 | 1914.00 - 407.00 | 4.00 - 1.00 | 3.00 - 1.00 |
| | KP_F018-OP_R073 | 3 - 5 | 526.11 | 1.98 | 1.22 | 2010.00 - 503.00 | 4.00 - 1.00 | 3.00 - 1.00 |
| | KP_F020-KP_R013 | 3 - 6 | 545.97 | 1.99 | 1.21 | 1325.00 - 539.00 | 4.00 - 1.00 | 3.00 - 1.00 |
| | KP_F022-KP_R063 | 5 - 9 | 586.51 | 2.00 | 1.22 | 670.00 - 530.00 | 4.00 - 1.00 | 3.00 - 1.00 |
| L | OP_F114-KP_R013 | 3 - 6 | 693.70 | 1.99 | 1.26 | 2178.00 - 671.00 | 4.00 - 1.00 | 3.00 - 1.00 |
| | KP_F018-KP_R063 | 3 - 9 | 1131.87 | 1.96 | 1.34 | 1828.00 - 1056.00 | 4.00 - 1.00 | 3.00 - 1.00 |
| | OP_F066-OP_R016 | 5 - 9 | 623.27 | 2.01 | 1.22 | 626.33 - 620.00 | 4.00 - 1.00 | 3.00 - 1.00 |
| | | | Superkingdom level | | | Species level | | |
| ALC | Bacterial and archaeal primer pairs | Gene region | Amplicon length (mean, bps) | g/G (mean) | gv/G (mean) | Amplicon length (mean, range, bps) | g/G (mean, range) | gv/G (mean, range) |
| S | OP_F114-KP_R002 | 3 - 4 | 162.29 | 1.95 | 1.10 | 868.00 - 146.00 | 4.00 - 1.00 | 3.00 - 1.00 |
| | KP_F020-KP_R032 | 4 - 5 | 289.43 | 1.95 | 1.14 | 1069.00 - 283.00 | 4.00 - 1.00 | 3.00 - 1.00 |
| | OP_F066-OP_R073 | 5 - 6 | 114.03 | 1.98 | 1.08 | 115.00 - 114.00 | 4.00 - 1.00 | 2.00 - 1.00 |
| M | OP_F114-KP_R031 | 3 - 5 | 436.72 | 1.95 | 1.19 | 1922.00 - 415.00 | 4.00 - 1.00 | 3.00 - 1.00 |
| | OP_F114-OP_R073 | 3 - 6 | 533.62 | 1.98 | 1.23 | 2018.00 - 511.00 | 4.00 - 1.00 | 3.00 - 1.00 |
| | KP_F020-OP_R073 | 4 - 6 | 385.76 | 1.99 | 1.18 | 1165.00 - 379.00 | 4.00 - 1.00 | 3.00 - 1.00 |
| L | OP_F114-OP_R121 | 3 - 9 | 1042.81 | 1.98 | 1.33 | 1741.00 - 1023.00 | 4.00 - 1.00 | 3.00 - 1.00 |
| | KP_F020-OP_R121 | 4 - 9 | 907.41 | 1.98 | 1.29 | 2279.00 - 891.00 | 4.00 - 1.00 | 3.00 - 1.00 |
| | OP_F066-OP_R121 | 5 - 9 | 635.78 | 1.98 | 1.22 | 1323.00 - 625.00 | 4.00 - 1.00 | 3.00 - 1.00 |
| | | | Superkingdom level | | | Species level | | |
| ALC | Most used primer pairs | Gene region | Amplicon length (mean, bps) | g/G (mean) | gv/G (mean) | Amplicon length (mean, range, bps) | g/G (mean, range) | gv/G (mean, range) |
| S | KP_F078-OP_R010$^{B+A}$ | 4 - 5 | 292.79 | 2.43 | 1.21 | 294.00 - 291.50 | 4.00 - 1.00 | 3.00 - 1.00 |
| L | KP_F014-KP_R011$^{A}$ | 3 - 6 | 606.18 | 1.58 | 1.14 | 603.00 - 608.00 | 4.00 - 1.00 | 3.00 - 1.00 |

The amplicon length category and gene regions were determined in objective 1 according to the mean size of the amplicons generated by a given primer and to the mode first position of the forward primer and the mode last of the reverse primer, respectively. The most commonly used primer pairs in the literature were detected in objective 1 (21). *The primer pair gene region was 3 when considering the mode positions obtained in the analysis against the oral-bacteria and oral-archaea databases in objective 1 (21).
A= archaea; ALC= amplicon length category; B= bacteria; bps= base pairs; F= forward; g/G= number of 16S rRNA gene amplicons per genome; gv/G= number of 16S rRNA gene variant amplicons per genome; KP= Klindworth primer; L= long mean amplicon length category, >600 base pairs; M= medium mean amplicon length category, 301-600 base pairs; OP= oral primer; R= reverse; S= short mean amplicon length category,100-300 base pairs.

Tables 5 and 6 show the percentages of detected taxa with and without MAs and overabundance estimators obtained by the primer pairs tested on the oral-bacteria and oral-archaea genomes. Our selected primer pairs detected 16S rRNA gene amplicons in a range from 99.46% to 88.71% for the bacterial species and 99.26% to 90.37% for the archaeal species;

these percentages were lower for the primer pairs used most in the oral microbiome literature (95.16% - 74.19% for the bacteria, and 63.70% and 30.37% for the archaea).

Overall, excluding the most commonly used primer pairs in the literature, unlike the coverage values, the SC-NMA values increased as the mean length of the amplicons obtained by the primer pair increased. If we contrast the percentages of species detected with their respective SC-NMA, all the short primer pairs analysed showed the largest differences between both parameters (average difference= 21.34% for bacteria and 23.70% for archaea), followed by those of medium length (7.30% and 13.75%, respectively). The long primer pairs presented the smallest differences between the coverage and SC-NMA values (4.30% and 5.82%, respectively).

According to the SC-NMA values, the best three bacteria-specific primer pairs were: KP_F048-OP_R030 and KP_F048-KP_R060 (L, SC-NMA= 93.55%, six MAs, OF-MA= 1.06 for both) and OP_F053-KP_R020 (M, 93.01%, six, 1.06). In contrast, the worst primer pair was OP_F066-KP_R040 (S, 47.31%, 77, 2.78). The most commonly used primer pairs in the literature did not stand out for their SC-NMA values among those in their category.

Considering the three categories of amplicon lengths, the SC-NMA values for the archaea-specific primer pairs ranged from 89.63% for the KP_F018-KP_R063 (L, six MAs, OF-MA= 1.11), 85.93% for KP_F022-KP_R063 (M, eight, 1.14) to 69.63% for the OP_F066-KP_R013 (S, 35, 1.99). Interestingly, the long primer pair KP_F014-KP_R011, which is the one used most in the literature to detect oral archaea, was only able to identify 30.37% of the species tested in this study, resulting in the lowest SC-NMA value (26.67%, five MAs, OF-MA= 1.14).

In relation to the bacterial and archaeal primer pairs, the overall SC-NMA values ranged from 92.52% for OP_F114_OP_R121 (L, 12 MAs, OF-MA= 1.08), 88.79% for OP_F114-KP_R031 (M, 29, 1.26) to 54.21% for OP_F066-OP_R073 (S, 134, 3.45). In terms of overall SC-NMA, the second worst was KP_F078-OP_R010 (S, 66.67%, 48 MAs, OF-MA= 1.68), mainly due to its low capacity to detect archaea (63.70%), which directly affected the SC-NMA value for archaea (48.89%) (Table 6). However, this primer pair is the most widely used in the literature to detect bacteria and archaea.

**Table 5. Detected taxa with and without matching amplicons and overabundance estimators obtained by the primer pairs tested on the oral-bacteria genomes.**

| ALC | Bacteria-specific primer pairs | Detected genomes (%) | Detected species (%) | Detected species with MAs (%) | SC-NMA (%) | OF | OF-MA |
|---|---|---|---|---|---|---|---|
| S | KP_F048-OP_R043 | 508 (98.07) | 180 (96.77) | 22 (12.22) | 158 (84.95) | 5.82 | 1.30 |
| | OP_F098-OP_R119 | 493 (95.17) | 177 (95.16) | 28 (15.82) | 149 (80.11) | 8.28 | 1.73 |
| | OP_F066-KP_R040 | 455 (87.84) | 165 (88.71) | 77 (46.67) | 88 (47.31) | 13.16 | 2.78 |
| | OP_F009-OP_R030 | 504 (97.30) | 181 (97.31) | 29 (16.02) | 152 (81.72) | 6.01 | 1.32 |
| | KP_F061-KP_R074 | 468 (90.35) | 169 (90.86) | 39 (23.08) | 130 (69.89) | 7.48 | 1.61 |
| | OP_F101-OP_R030 | 460 (88.80) | 167 (89.78) | 39 (23.35) | 128 (68.82) | 7.44 | 1.61 |
| M | OP_R053-KP_R020 | 506 (97.68) | 179 (96.24) | 6 (3.35) | 173 (93.01) | 4.80 | 1.06 |
| | KP_F048-KP_R031 | 507 (97.88) | 180 (96.77) | 9 (5.00) | 171 (91.94) | 5.04 | 1.12 |
| | KP_F048-OP_R073 | 498 (96.14) | 178 (95.70) | 6 (3.37) | 172 (92.47) | 4.72 | 1.06 |
| | KP_F051-KP_R041 | 456 (88.03) | 166 (89.25) | 20 (12.05) | 146 (78.50) | 7.45 | 1.58 |
| | KP_F051-OP_R030 | 508 (98.07) | 184 (98.92) | 19 (10.33) | 165 (88.71) | 5.53 | 1.22 |
| | OP_F116_KP_R060 | 516 (99.61) | 185 (99.46) | 31 (16.76) | 154 (82.80) | 6.13 | 1.37 |
| L | KP_F048-OP_R030 | 507 (97.88) | 180 (96.77) | 6 (3.33) | 174 (93.55) | 4.72 | 1.06 |
| | KP_F048-KP_R060 | 507 (97.88) | 180 (96.77) | 6 (3.33) | 174 (93.55) | 4.72 | 1.06 |
| | KP_F056-KP_R077 | 495 (95.56) | 180 (96.77) | 10 (5.56) | 170 (91.40) | 4.89 | 1.10 |
| **ALC** | **Bacterial and archaeal primer pairs** | **Detected genomes (%)** | **Detected species (%)** | **Detected species with MAs (%)** | **SC-NMA (%)** | **OF** | **OF-MA** |
| S | OP_F114-KP_R002 | 485 (93.63) | 172 (92.47) | 22 (12.79) | 150 (80.65) | 5.82 | 1.30 |
| | KP_F020-KP_R032 | 488 (94.21) | 176 (94.62) | 28 (15.91) | 148 (79.57) | 8.28 | 1.73 |
| | OP_F066-OP_R073 | 502 (96.91) | 182 (97.85) | 85 (46.70) | 97 (52.15) | 15.90 | 3.31 |
| M | OP_F114-KP_R031 | 507 (97.88) | 180 (96.77) | 9 (5.00) | 171 (91.94) | 5.04 | 1.12 |
| | OP_F114-OP_R073 | 498 (96.14) | 178 (95.70) | 6 (3.77) | 172 (92.47) | 4.72 | 1.06 |
| | KP_F020-OP_R073 | 488 (94.21) | 176 (94.62) | 22 (12.50) | 154 (82.80) | 7.52 | 1.61 |
| L | OP_F114-OP_R121 | 507 (97.88) | 180 (96.77) | 6 (3.33) | 174 (93.55) | 4.72 | 1.06 |
| | KP_F020-OP_R121 | 489 (94.40) | 177 (95.16) | 10 (5.65) | 167 (89.79) | 4.89 | 1.10 |
| | OP_F066-OP_R121 | 516 (99.61) | 185 (99.46) | 16 (8.65) | 169 (90.86) | 5.20 | 1.16 |
| **ALC** | **Most used primer pairs** | **Detected genomes (%)** | **Detected species (%)** | **Detected species with MAs (%)** | **SC-NMA (%)** | **OF** | **OF-MA** |
| S | KP_F078-OP_R010$^{B+A}$ | 488 (94.21) | 176 (94.62) | 28 (15.91) | 148 (79.57) | 8.28 | 1.73 |
| M | KP_F031-KP_R021$^{B}$ | 347 (66.99) | 138 (74.19) | 2 (1.45) | 136 (73.12) | 4.50 | 1.02 |
| | KP_F047-KP_R035$^{B}$ | 500 (96.53) | 177 (95.16) | 9 (5.09) | 168 (90.32) | 5.04 | 1.12 |
| | OP_F009-OP_R029$^{B}$ | 469 (90.54) | 164 (88.17) | 24 (14.63) | 140 (75.27) | 5.62 | 1.24 |
| L | KP_F034-KP_R065$^{B}$ | 440 (84.94) | 155 (83.33) | 2 (1.29) | 153 (82.26) | 4.50 | 1.02 |

A= archaea; ALC= amplicon length category; B= bacteria; F= forward; KP= Klindworth primer; L= long mean amplicon length category, >600 base pairs; M= medium mean amplicon length category, 301-600 base pairs; MAs= matching amplicons; OF= overestimation factor; OF-MA= overestimation factor associated with matching amplicons; OP= oral primer; R= reverse; S= short mean amplicon length category,100-300 base pairs; SC-NMA= species coverage with no matching amplicons.

**Table 6. Detected taxa with and without matching amplicons and overabundance estimators obtained by the primer pairs tested on the oral-archaea genomes.**

| ALC | Archaea-specific primer pairs | Detected genomes (%) | Detected species (%) | Detected species with MAs (%) | SC-NMA (%) | OF | OF-MA |
|---|---|---|---|---|---|---|---|
| S | KP_F018-KP_R002 | 185 (96.86) | 129 (95.56) | 29 (22.48) | 100 (74.07) | 3.30 | 1.76 |
| | OP_F066-KP_R013 | 184 (96.34) | 129 (95.56) | 35 (27.13) | 94 (69.63) | 4.02 | 1.99 |
| M | KP_F018-KP_R032 | 186 (97.38) | 130 (96.30) | 20 (15.39) | 110 (81.48) | 2.68 | 1.49 |
| | KP_F018-OP_R073 | 177 (92.67) | 122 (90.37) | 18 (14.75) | 104 (77.04) | 2.65 | 1.46 |
| | KP_F020-KP_R013 | 183 (95.81) | 128 (94.81) | 20 (15.63) | 108 (80.00) | 2.61 | 1.35 |
| | KP_F022-KP_R063 | 180 (94.24) | 124 (91.85) | 8 (6.45) | 116 (85.93) | 2.26 | 1.14 |
| L | OP_F114-KP_R013 | 184 (96.34) | 129 (95.56) | 16 (12.40) | 113 (83.70) | 2.47 | 1.28 |
| | KP_F018-KP_R063 | 183 (95.81) | 127 (94.07) | 6 (4.72) | 121 (89.63) | 2.16 | 1.11 |
| | OP_F066-OP_R016 | 180 (94.24) | 124 (91.85) | 8 (6.45) | 116 (85.93) | 2.26 | 1.14 |
| **ALC** | **Bacterial and archaeal primer pairs** | **Detected genomes (%)** | **Detected species (%)** | **Detected species with MAs (%)** | **SC-NMA (%)** | **OF** | **OF-MA** |
| S | OP_F114-KP_R002 | 190 (99.48) | 134 (99.26) | 29 (21.64) | 105 (77.78) | 3.30 | 1.76 |
| | KP_F020-KP_R032 | 190 (99.48) | 134 (99.26) | 30 (22.39) | 104 (77.04) | 3.88 | 1.92 |
| | OP_F066-OP_R073 | 181 (94.76) | 126 (93.33) | 49 (38.89) | 77 (57.04) | 8.37 | 3.71 |
| M | OP_F114-KP_R031 | 190 (99.48) | 134 (99.26) | 20 (14.93) | 114 (84.44) | 2.68 | 1.49 |
| | OP_F114-OP_R073 | 181 (94.76) | 126 (93.33) | 18 (14.29) | 108 (80.00) | 2.65 | 1.46 |
| | KP_F020-OP_R073 | 180 (94.24) | 125 (92.59) | 26 (20.80) | 99 (73.33) | 3.74 | 1.85 |
| L | OP_F114-OP_R121 | 185 (96.86) | 129 (95.56) | 6 (4.65) | 123 (91.11) | 2.16 | 1.11 |
| | KP_F020-OP_R121 | 185 (96.86) | 129 (95.56) | 6 (4.65) | 123 (91.11) | 2.16 | 1.11 |
| | OP_F066-OP_R121 | 186 (97.38) | 130 (96.30) | 8 (6.15) | 122 (90.37) | 2.26 | 1.14 |
| **ALC** | **Most used primer pairs** | **Detected genomes (%)** | **Detected species (%)** | **Detected species with MAs (%)** | **SC-NMA (%)** | **OF** | **OF-MA** |
| S | KP_F078-OP_R010$^{B+A}$ | 123 (66.40) | 86 (63.70) | 20 (23.26) | 66 (48.89) | 3.56 | 1.60 |
| L | KP_F014-KP_R011$^{A}$ | 44 (23.04) | 41 (30.37) | 5 (12.20) | 36 (26.67) | 2.00 | 1.14 |

A= archaea; ALC= amplicon length category; B= bacteria; F= forward; KP= Klindworth primer; L= long mean amplicon length category, >600 base pairs; M= medium mean amplicon length category, 301-600 base pairs; MAs= matching amplicons; OF= overestimation factor; OF-MA= overestimation factor associated with matching amplicons; OP= oral primer; R= reverse; S= short mean amplicon length category,100-300 base pairs; SC-NMA= species coverage with no matching amplicons.

Appendices S5-S11 contain more detailed information on the results of MA- and MA-free species coverage and the overabundance parameters (OF and OF-MA values) obtained by the primer pairs tested against the oral-bacteria and oral-archaea genomes. Appendices S5, S9, and S11 also include the results obtained by the bacterial and archaeal primer pairs for both domains.

Depending on the primer pair tested, between 46.70% - 1.29% of the bacterial species and between 38.89% - 4.65%. of the archaeal species had MAs (Tables 5 and 6). Figure 1 shows the bacterial and archaeal species with MAs obtained with at least 10 primer pairs. To the bacteria, these species belonged to the following genera: *Actinomyces*, *Cronobacter*, *Fusobacterium*, *Klebsiella*, *Lacticaseibacillus*, *Lactobacillus*, *Staphylococcus,* and *Streptococcus*. In the archaea, these genera were: *Haloarcula*, *Halomicrobium*, *Methanosarcina*, *Nitrososphaera*, *Pyrococcus,* and *Thermococcus*. Appendices S12-S14 define in detail which species from the same or different genera shared MAs, depending on the primer pair evaluated.

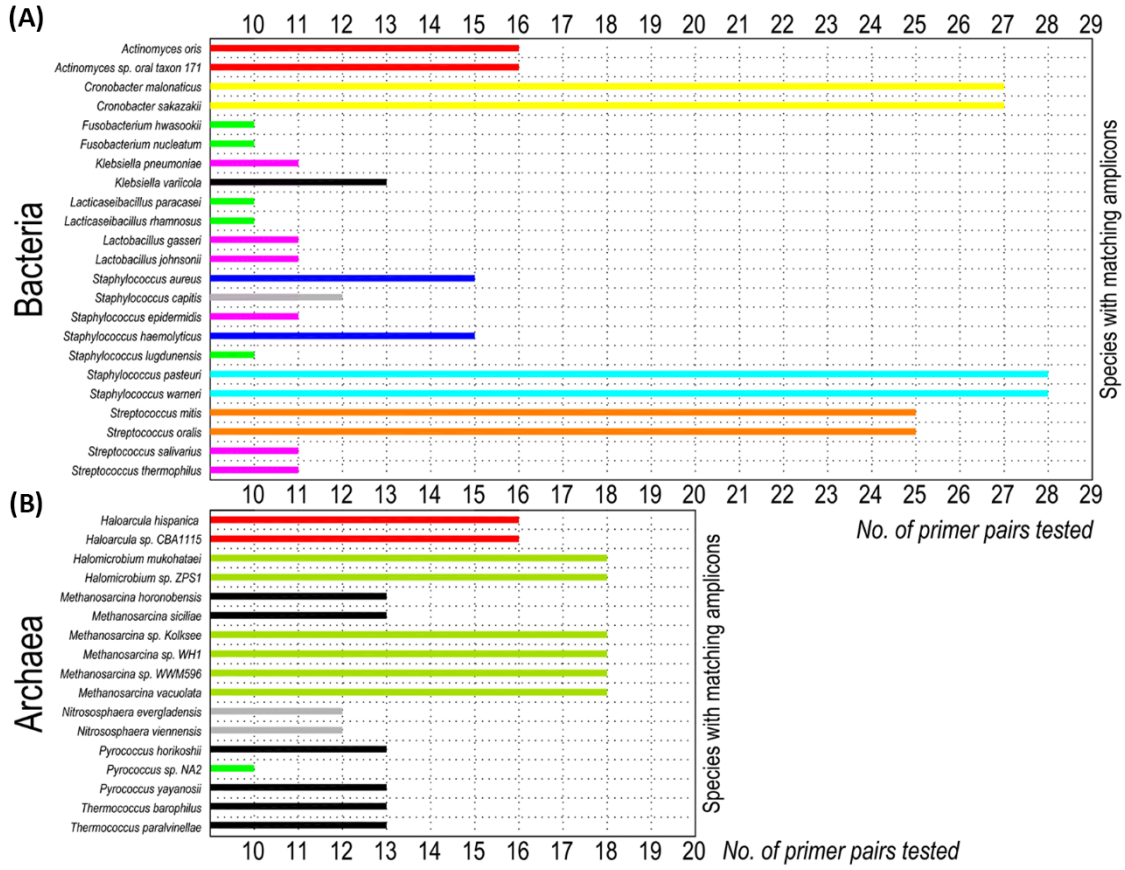



**Figure 1. Taxa with matching amplicons using at least 10 primer pairs: bacterial species (A) and archaeal species (B).**

To detect bacteria and archaea, 29 and 20 primer pairs, respectively, were tested, including those specific, bacterial and archaeal, and the most widely used in the literature.

## 2.5. DISCUSSION

### 2.5.1. Number of intragenomic 16S rRNA genes in oral-bacteria and oral-archaea genomes

The intragenomic redundancy of the 16S rRNA gene has been evaluated previously using genomes from diverse sources such as GenBank (5,7,19,32), the NCBI's microbial genome database (4,6,8,20), or the ribosomal RNA operon copy number database (rrnDB) (9,33,34). These *in silico* investigations extracted the gene sequences from the complete genomes through tools such as Kodon 2.0 (32,35) or RNAmmer (6,36), or by using a primer pair targeting the regions 4-6 (7). However, none of these studies focused on the genomes of microorganisms living in a specific environment. As gene redundancy has been proven to affect abundance estimates based on gene counts (4,7), variations in the number of genes/genome of the microbes inhabiting the ecosystem of interest must be examined to ensure proper descriptions of the microbial community. To the best of our knowledge, this study is the first to investigate the number of intragenomic 16S rRNA genes in the microbiota of the oral environment.

Through chromatograms derived from direct sequencing or cloning, recent research identified a maximum of four different 16S rRNA genes/genome in 138 clinical isolates taken from periodontal abscesses (11). However, the low number of species evaluated (n= 32) and the focus on a specific niche and health condition within the mouth limit the applicability of the findings to the oral microbiota more generally. In contrast, the present study evaluated all of the complete bacterial genomes described in an oral-specific database (18) and a series of genomes taken from archaeal species previously identified in the human mouth (21); all these bacterial and archaeal genomes were downloaded from the NCBI website (20). Moreover, for the first time in the oral microbiome literature, we extracted the 16S rRNA genes using a special and easily accessible tool based on Edgar's algorithm, which has an estimated sensitivity >99% for identifying known genes (27). In our opinion, this algorithm represents a significant improvement in the detection of the 16S rRNA genes over previously used methods (7,32) since it constitutes a specialised tool for this purpose.



Our study identified that 94.09% of the oral-bacteria species had more than one 16S rRNA gene/genome and an 8.60% had seven or more; which are values similar to previously reported in non-oral specific investigations (95.53% and ~9.50%, respectively) (5,7). Conversely, other

authors found greater percentages of bacteria with one intragenomic gene (15.00% *vs.* 5.91% in the present study)(4,7) or with seven or more (17.80% *vs.* 8.60% in the present study) (4). Also, we detected that 47.41% of the oral-archaea species had one gene/genome, which is considerably lower as reported before in non-oral studies (65.20% and 57.00%) (4,9). Consequently, we found that several species traditionally associated with different oral-health conditions had more than one intragenomic 16S rRNA gene, meaning they may have been overcounted in previous sequencing-based investigations. Included in these species were bacteria that are widely known to be associated with periodontitis, such as: *Aggregatibacter actinomycetemcomitans* (mean genes/genome=5.75), *F. nucleatum* (=4.25), *Filifactor alocis* (=4.00), *Porphyromonas gingivalis* (=4.00), *Tannerella forsythia* (=2.00), and *Treponema denticola* (=2.00) (37); the caries-associated bacteria *Streptococcus mutans* (=5.00) (38) and *Rothia dentocariosa* (=3.00) (39); and the commensal bacteria *Streptococcus mitis* (=4.00) (40) and *Streptococcus oralis* (=4.00) (41). Some archaeal species that can be found in healthy subjects or those with periodontitis (42) also had more than one gene/genome. These included: *M. stadtmanae* (=4.00), *Methanosarcina mazei* (=3.00), *M. maripaludis* (=2.88), *Methanobrevibacter smithii* (=2.00), and *S. acidocaldarius* (=2.00).

### 2.5.2. Evaluation of the primer pairs taken from our previous research and those used most in oral microbiome studies

There is a lack of literature on how the 16S rRNA gene primer pair influences the detection of redundant amplicons and MAs from different taxa. Recognising the importance of conducting 16S rRNA gene-based research using habitat-specific databases (14), the present study constitutes the first to evaluate the above-mentioned topic focusing on the oral microbiota.

Through this analysis, we discovered that all the primer combinations amplified the maximum mean number of genes/genome in both the bacterial and archaeal species (10.83 and 4.00, respectively). However, the great majority of them could not detect the maximum mean number of variants/genome in bacteria (i.e., 9.00), which was not the case for the archaea.



The presence of amplicons with matching 16S rRNA gene sequences in different species is a challenge for researchers, as they could be inappropriately misclassified, thus artificially increasing the number of counts in operational taxonomic units (OTUs) or amplicon sequence

variants (ASVs) (7,43), depending on the bioinformatics pipeline used. As amplicons derived from distinct regions have different degrees of heterogeneity (6,10), the primer pair employed may affect both estimates of diversity and taxonomic identifications. Despite the lack of literature on the subject, it is important from a clinical applicability point of view to conduct studies that take into account the quality of the primer pairs, in our case those specific to the oral microbiota.

As previously described in objective 1 (21), the primer pairs that identified >90% of the species in a dataset were evenly distributed across the different amplicon length categories. However, these findings do not reflect the influence of MAs. To ensure that this factor was taken into account in the present study, we considered, for the first time in this type of research, the values of the percentage of coverage at the species level with no matching amplicons (SC-NMA), the overestimation factor, which combines the copy number of the 16S rRNA gene amplicons and the number of MAs, and the OF caused by the presence of MAs (OF-MA). The lack of studies employing these parameters makes it impossible to conduct a relevant comparative analysis.

However, the estimation tools that we newly describe have allowed us to demonstrate in general terms that the long primer pairs with >600 bps, followed by those of medium length of 301-600 bps, showed the greatest ability to detect bacterial and archaeal species with no MAs in contrast to the short primer pairs of 100-300 bps (the mean differences between coverage and SC-NMA percentages were 22.52%, 10.52%, and 5.06%, respectively). These discrepancies between the two coverage parameters were confirmed by considering the coverage results of objective 1 in which we analysed a larger number of oral taxa based on 16S rRNA gene sequences instead of complete genomes (21). Assessing the impact that MAs could have on species abundance, long primer pairs had OF-MA values as low as 1.08 (e.g. with OP_F114_OP_R121), meaning that the small number of MAs detected did not influence abundance. By contrast, short primer pairs had a maximum value of 3.45 (e.g. with OP_F066-OP_R073), meaning that abundance was tripled by the presence of MAs.



These findings reveal that the SC-NMA parameter is more useful than the conventional coverage percentage in selecting the best primer pairs because this last value does not

discriminate the presence of MAs for different taxa. If there are several primer pairs with similar SC-NMA values, the OF-MA values would be the appropriate parameter to use to choose between them.

Specifically, the best primer pairs presented a mean amplicon lengths >600 bps and were: KP_F048-OP_R030 (for bacteria (B), SC-NMA= 93.55%, OF-MA= 1.06), KP_F018-KP_R063 (for archaea (A), 89.63%, 1.11); and OP_F114_OP_R121 (for bacteria and archaea jointly (B+A), 92.52%, 1.08). In consequence, we were thus able to demonstrate that sequencing longer fragments enable the identification of lower taxonomy levels (44), reducing the probability of overestimation and classification bias related to MAs.

None of the pairs used most in the oral microbiome sequencing-based studies reported in the literature were able to detect the highest possible numbers of total genomes and species. We might assume *a priori* that the lower the number of taxa detected, the fewer the opportunities to misclassify them but, in fact, the best SC-NMA estimates were also not obtained with these primer pairs. The pair employed most in the literature is KP_F078-OP_R010 (B+A, region 4-5), as described by Caporaso et al. (45). This showed a SC-NMA score of 66.67% and an OF-MA of 1.68 and was the second worst primer pair at detecting both the bacterial and archaeal superkingdoms. It was even outperformed by other primers from the same region, such as: OP_F098-OP_R119 (B, SC-NMA= 80.11%, OF-MA= 1.73) and KP_F020-KP_R032 (B+A, 78.51%, 1.79). The next most-used primer pair was KP_F047-KP_R035 (B, 3-5, 90.32%, 1.12). This was recommended by Illumina (46), but produced poorer estimates than KP_F048-KP_R031 (B, 3-5, 91.94%, 1.12). Another primer pair used in the literature, albeit to a lesser extent, is KP_F014-KP_R011 (A, 3-6, OF-MA= 1.14), which has produced a SC-NMA value of 26.67%. This is considerably lower than the 80.00% achieved by KP_F020-KP_R013 (A, 3-6, OF-MA= 1.35). Consequently, the data derived from the primer combinations employed most in the literature could be improved upon, in some cases significantly, by using the alternative primer pairs presented in this study. Moreover, these results highlight that primer pairs targeting the same gene region do not distinguish equally between taxa with MAs.



Therefore, comparisons of data from studies assessing the same region may be biased and abundance may be inaccurate but close. In the case of comparing amplicons from different

regions, the results could be vastly different and may even lead to opposite biological conclusions. Consequently, the comparison of oral microbiome studies using the same primer pairs would be the most recommendable methodological approach.

Our research proves that the detection of MAs is not a one-off issue, with 46.70% - 1.29% of the species detected by the primer pairs having them. Indeed, this number may be an underestimate, given that we were only able to examine a third of the genomes contained in the eHOMD (18). Despite this, relevant genera present in the oral environment were identified, including: *Actinomyces, Fusobacterium, Lactobacillus, Methanosarcina, Staphylococcus,* and *Streptococcus* (13,47) that had MAs in at least 10 primer pairs. A 3.00%, 2.00%, 4.00%, 19.30%, 9.00% and 15.00% of all different bacterial or archaeal species with MAs detected by all the primer pairs belonged, respectively, to such genera. Among them, there were health-associated species as *S. mitis* (40) and *S. oralis* (41), disease-associated taxa as *F. nucleatum* (37) and *S. mutans* (38), or abundant in both states as *Methanosarcina vacuolata* (42); which, as shown above, had problems related to the presence of more than one 16S rRNA gene/genome. Other relevant species from distinct genera as *Capnocytophaga ochracea*, *T. forsythia,* and *T. denticola* (37) also presented both intragenomic gene redundancy and MAs.

### 2.5.3. Limitations of the present study

The main limitation of the present research is that only 25% of the oral microorganism genomes listed on the eHOMD website were evaluated, as the remaining ones were not fully sequenced. This lack of complete genomes reduced the number of species evaluated to 35% of those listed on eHOMD. Although the analysis could have been performed with a fasta file containing 16S rRNA gene sequences from oral microbes, we preferred to use complete genomes, thereby ensuring the high quality of the gene sequences reviewed.

In our opinion, these results are only the tip of the iceberg, and the problematic issue of MAs is likely to affect more taxa as the number of genomes examined increases. Moreover, it is important to note that our study has only taken into account the MAs from different species. Our next objective will be to assess the impact of OTU clustering, which will undoubtedly increase the complexity of the issues involved.

## 2.6. CONCLUSIONS

In conclusion, nearly all oral bacteria and about half of the oral archaea have more than one 16S rRNA gene in their respective genomes. Depending on the primer pair used, up to almost half of the species present MAs, affecting relevant genera present in the oral environment such as *Actinomyces, Fusobacterium, Lactobacillus, Methanosarcina, Staphylococcus,* and *Streptococcus.* The performance of the primer pairs to detect non-MA species increases as the average length of the amplicons increases; none of these being the most widely used primer pairs in the oral microbiome literature. The best primer pairs were: KP_F048-OP_R030 (for bacteria; region 3-7; primer pair position for *Escherichia coli* J01859.1: 342-1079), KP_F018-KP_R063 (for archaea; 3-9; undefined-1506), and OP_F114_OP_R121 (for both bacteria and archaea; 3-9; 340-1405). In addition to the 16S rRNA gene redundancy, the considerable presence of MAs must be controlled to ensure the accurate interpretation of microbial diversity data. The SC-NMA is a more useful parameter than the conventional coverage percentage for selecting the best primer pairs. The choice of primer pair affects significantly diversity estimates and taxonomic classification, conditioning the comparability of oral microbiome studies using different primer pairs.

## 2.7. REFERENCES

(1) Rajendhran J, Gunasekaran P. Microbial phylogeny and diversity: small subunit ribosomal RNA sequence analysis and beyond. Microbiol Res. 2011 Feb;166(2):99-110.

(2) Woese CR. Bacterial evolution. Microbiol Rev. 1987 Jun;51(2):221-71.

(3) del Rosario-Rodicio M, del Carmen-Mendoza M. Bacterial identification by 16S rRNA sequencing: rationale, methodology and applications in clinical microbiology. Enferm Infecc Microbiol Clin. 2004 Apr;22(4):238-45. Spanish.

(4) Acinas SG, Marcelino LA, Klepac-Ceraj V, Polz MF. Divergence and redundancy of 16S rRNA sequences in genomes with multiple rrn operons. J Bacteriol. 2004 May;186(9):2629-35.

(5) Pei AY, Oberdorf WE, Nossa CW, Agarwal A, Chokshi P, Gerz EA, et al. Diversity of 16S rRNA genes within individual prokaryotic genomes. Appl Environ Microbiol. 2010 Jun;76(12):3886-97.

(6) Sun D, Jiang X, Wu QL, Zhou N. Intragenomic heterogeneity of 16S rRNA genes causes overestimation of prokaryotic diversity. Appl Environ Microbiol. 2013 Oct;79(19):5962-9.

(7) Větrovský T, Baldrian P. The variability of the 16S rRNA gene in bacterial genomes and its consequences for bacterial community analyses. PLoS One. 2013;8(2):e57923. doi: 10.1371/journal.pone.0057923.

(8) Case RJ, Boucher Y, Dahllöf I, Holmström C, Doolittle WF, Kjelleberg S. Use of 16S rRNA and rpoB genes as molecular markers for microbial ecology studies. Appl Environ Microbiol. 2007 Jan;73(1):278-88.



(9) Lee ZM, Bussema C 3rd, Schmidt TM. rrnDB: documenting the number of rRNA and tRNA genes in bacteria and archaea. Nucleic Acids Res. 2009 Jan;37(Database issue):D489-93. doi: 10.1093/nar/gkn689.

(10) Johnson JS, Spakowicz DJ, Hong BY, Petersen LM, Demkowicz P, Chen L, et al. Evaluation of 16S rRNA gene sequencing for species and strain-level microbiome analysis. Nat Commun. 2019 Nov;10(1):5029. doi: 10.1038/s41467-019-13036-1.

(11) Chen J, Miao X, Xu M, He J, Xie Y, Wu X, et al. Intra-genomic heterogeneity in 16S rRNA genes in strictly anaerobic clinical isolates from periodontal abscesses. PLoS One. 2015 Jun;10(6):e0130265. doi: 10.1371/journal.pone.0130265.

(12) Durán-Pinedo AE, Frias-Lopez J. Beyond microbial community composition: functional activities of the oral microbiome in health and disease. Microbes Infect. 2015 Jul;17(7):505-16.

(13) Verma D, Garg PK, Dubey AK. Insights into the human oral microbiome. Arch Microbiol. 2018 May;200(4):525-40.

(14) F Escapa I, Huang Y, Chen T, Lin M, Kokaras A, Dewhirst FE, et al. Construction of habitat-specific training sets to achieve species-level assignment in 16S rRNA gene datasets. Microbiome. 2020 May;8(1):65. doi: 10.1186/s40168-020-00841-w.

(15) Relvas M, Regueira-Iglesias A, Balsa-Castro C, Salazar F, Pacheco JJ, Cabral C, et al. Relationship between dental and periodontal health status and the salivary microbiome: bacterial diversity, co-occurrence networks and predictive models. Sci Rep. 2021 Jan;11(1):929. doi: 10.1038/s41598-020-79875-x.

(16) Camelo-Castillo A, Novoa L, Balsa-Castro C, Blanco J, Mira A, Tomas I. Relationship between periodontitis-associated subgingival microbiota and clinical inflammation by 16S pyrosequencing. J Clin Periodontol. 2015 Dec;42(12):1074-82.

(17) Camelo-Castillo AJ, Mira A, Pico A, Nibali L, Henderson B, Donos N, et al. Subgingival microbiota in health compared to periodontitis and the influence of smoking. Front Microbiol. 2015 Feb;6:119. doi: 10.3389/fmicb.2015.00119.

(18) F Escapa I, 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;3(6):e00187-18. doi: 10.1128/mSystems.00187-18.

(19) Clark K, Karsch-Mizrachi I, Lipman DJ, Ostell J, Sayers EW. GenBank. Nucleic Acids Res. 2016 Jan;44:D67-72. doi: 10.1093/nar/gkv1276.

(20) NCBI Resource Coordinators. Database resources of the National Center for Biotechnology Information. Nucleic Acids Res. 2016 Jan;44(D1):D7-19. doi: 10.1093/nar/gkv1290.

(21) Regueira-Iglesias A, Vázquez-González L, Balsa-Castro C, Vila-Blanco N, Blanco-Pintos T, Tamames J, et al. In silico evaluation and selection of the best 16S rRNA gene primers for use in next-generation sequencing to detect oral bacteria and archaea. Accepted for publication in Microbiome. Preprint at Research Square. 2021. doi: 10.21203/rs.3.rs-516961/v1.

(22) National Center for Biotechnology Information. Entrez programming utilities help. 2010; Available at: https://www.ncbi.nlm.nih.gov/books/NBK25501/.

(23) Python Software Foundation. Python. Version 3.9.0. 2020. http://www.python.org/.

(24) Schoch CL, Ciufo S, Domrachev M, Hotton CL, Kannan S, Khovanskaya R, et al. NCBI Taxonomy: a comprehensive update on curation, resources and tools. Database (Oxford). 2020 Jan;2020:baaa062. doi: 10.1093/database/baaa062.

(25) O'Leary NA, Wright MW, Brister JR, Ciufo S, Haddad D, McVeigh R, et al. Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. Nucleic Acids Res. 2016 Jan;44(D1):D733-45. doi: 10.1093/nar/gkv1189.



(26) Lyalina S. Search 16S py algorithm. 2019; Available at: https://github.com/slyalina/search_16S_py.

(27) Edgar R. SEARCH_16S: A new algorithm for identifying 16S ribosomal RNA genes in contigs and chromosomes. Preprint at bioRxiv 2017:124131. doi: 10.1101/124131.

(28) Harris CR, Millman KJ, van der Walt, Stéfan J, Gommers R, Virtanen P, Cournapeau D, et al. Array programming with NumPy. Nature. 2020 Sep;585(7825):357-62.

(29) McKinney W. Data structures for statistical computing in Python. In: van der Walt S, Millman J, editors. Proceedings of the 9th Python in Science Conference; 2010; Austin. Texas: SciPy; 2010. doi: 10.25080/Majora-92bf1922-00a.

(30) Baker GC, Smith JJ, Cowan DA. Review and re-analysis of domain-specific 16S primers. J Microbiol Methods. 2003 Dec;55(3):541-55.

(31) Barnett M. regex. 2020; Available at: https://pypi.org/.

(32) Coenye T, Vandamme P. Intragenomic heterogeneity between multiple 16S ribosomal RNA operons in sequenced bacterial genomes. FEMS Microbiol Lett. 2003 Nov;228(1):45-9.

(33) Klappenbach JA, Saxman PR, Cole JR, Schmidt TM. rrnDB: the ribosomal RNA operon copy number database. Nucleic Acids Res. 2001 Jan;29(1):181-4.

(34) Stoddard SF, Smith BJ, Hein R, Roller BR, Schmidt TM. rrnDB: improved tools for interpreting rRNA gene abundance in bacteria and archaea and a new foundation for future development. Nucleic Acids Res. 2015 Jan;43(Database issue):D593-8. doi: 10.1093/nar/gku1201.

(35) Applied Maths NV. Kodon 2.0; Available at: https://www.applied-maths.com.

(36) Lagesen K, Hallin P, Rødland EA, Stærfeldt H, Rognes T, Ussery DW. RNAmmer: consistent and rapid annotation of ribosomal RNA genes. Nucleic Acids Res. 2007;35(9):3100-8.

(37) Teles R, Teles F, Frias-Lopez J, Paster B, Haffajee A. Lessons learned and unlearned in periodontal microbiology. Periodontol 2000. 2013 Jun;62(1):95-162.

(38) Abranches J, Zeng L, Kajfasz JK, Palmer SR, Chakraborty B, Wen ZT, et al. Biology of oral streptococci. Microbiol Spectr. 2018 Oct;6(5):10.1128/microbiolspec.GPP3-0042-2018.

(39) Jiang S, Gao X, Jin L, Lo EC. Salivary microbiome diversity in caries-free and caries-affected children. Int J Mol Sci. 2016 Nov;17(12):1978. doi: 10.3390/ijms17121978.

(40) Mitchell J. Streptococcus mitis: walking the line between commensalism and pathogenesis. Mol Oral Microbiol. 2011 Apr;26(2):89-98.

(41) Thurnheer T, Belibasakis GN. Streptococcus oralis maintains homeostasis in oral biofilms by antagonizing the cariogenic pathogen Streptococcus mutans. Mol Oral Microbiol. 2018 Jun;33(3):234-9.

(42) Deng ZL, Szafrański SP, Jarek M, Bhuju S, Wagner-Döbler I. Dysbiosis in chronic periodontitis: key microbial players and interactions with the human host. Sci Rep. 2017 Jun;7(1):3703. doi: 10.1038/s41598-017-03804-8.

(43) Schloss PD. Amplicon sequence variants artificially split bacterial genomes into separate clusters. mSphere. 2021 Aug;6(4):e0019121. doi: 10.1128/mSphere.00191-21.

(44) Zhang J, Ding X, Guan R, Zhu C, Xu C, Zhu B, et al. Evaluation of different 16S rRNA gene V regions for exploring bacterial diversity in a eutrophic freshwater lake. Sci Total Environ. 2018 Mar;618:1254-67.

(45) Caporaso JG, Lauber CL, Walters WA, Berg-Lyons D, Lozupone CA, Turnbaugh PJ, et al. Global patterns of 16S rRNA diversity at a depth of millions of sequences per sample. Proc Natl Acad Sci U S A. 2011 Mar;108 Suppl 1(Suppl 1):4516-22.

(46) Illumina, Inc. 16S Metagenomic Sequencing Library Preparation. 2013; Available at: https://support.illumina.com/content/dam/illumina-support/documents/documentation/chemistry_documentation/16s/16s-metagenomic-library-prep-guide-15044223-b.pdf.

(47) Zhang Y, Wang X, Li H, Ni C, Du Z, Yan F. Human oral microbiota and its modulation for oral health. Biomed Pharmacother. 2018 Mar;99:883-93.

# OBJECTIVE 3

# Objective 3. *In silico* detection of oral prokaryotic species with highly similar 16S rRNA sequence segments using different primer pairs


## 3.1. ABSTRACT

**Aims:** To evaluate *in silico* the coverage of a set of previously selected primer pairs to detect oral species having 16S rRNA sequence segments with ≥97% similarity; and to describe oral species with highly similar sequence segments and determine whether they belong to distinct genera or other higher taxonomic ranks.

**Material and methods:** Thirty-nine primer pairs were employed to obtain the *in silico* amplicons from the complete genomes of 186 bacterial and 135 archaeal species. Each fasta file for the same primer pair was inserted as subject and query in BLASTN for obtaining the similarity percentage between amplicons belonging to different oral species. Amplicons with 100% alignment coverage of the query sequences and with a similarity value ≥97% (ASI97) were selected. For each primer, the species coverage with no ASI97 (SC-NASI97) was calculated.

**Results:** Based on the SC-NASI97 parameter, the best primer pairs were OP_F053-KP_R020 for bacteria (region 1-3; primer pair position for *Escherichia coli* J01859.1: 9-356); KP_F018-KP_R002 for archaea (4; undefined-532); and OP_F114-KP_R031 for both (3-5; 340-801). Around 80% of the oral-bacteria and oral-archaea species analysed had an ASI97 with at least one other species. These very similar species play different roles in the oral microbiota and belong to bacterial genera such as *Campylobacter*, *Rothia*, *Streptococcus*, and *Tannerella*, and archaeal genera such as *Halovivax*, *Methanosalsum*, and *Methanosarcina*. Moreover, ~20% and ~30% of these two-by-two similarity relationships were established between species from different bacterial and archaeal genera, respectively. Even taxa from distinct families, orders, and classes could be grouped in the same possible OTU.

**Conclusions:** Regardless of the primer pair used, sequence clustering with a 97% similarity provides an inaccurate description of oral-bacterial and oral-archaeal species, which can greatly affect microbial diversity parameters. As a result, OTU clustering conditions the credibility of associations between some oral species and certain health and disease conditions. This significantly limits the comparability of the microbial diversity findings reported in oral microbiome literature.

### 3.1.1. Keywords

Computational biology; DNA primers; genes; high-throughput nucleotide sequencing; mouth; microbiota; 16S rRNA.

### 3.1.2. Declaration of conflict of interest

The doctoral candidate and the rest of the authors of the present study declare that they have no conflict of interest concerning the objectives proposed in this chapter.

### 3.1.3. Funding

This investigation was supported by the Instituto de Salud Carlos III (General Division of Evaluation and Research Promotion, Madrid, Spain) and co-financed by the FEDER (European Regional Development Fund, ERDF) ("A way of making Europe") under grant ISCIII/PI17/01722; the Consellería de Cultura, Educación e Ordenación Universitaria de la Xunta de Galicia (accreditation 2019-2022 ED431G-2019/04, group with growth potential ED431B 2020-2022 GPC2020/27; A. Regueira-Iglesias support ED481A-2017/233) and the ERDF, which acknowledges the CiTIUS-Research Center in Intelligent Technologies of the Santiago de Compostela University as a Research Center of the Galician University System.

The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

## 3.2. INTRODUCTION

Since the introduction of the Sanger method in 1977, the sequencing technologies have undergone substantial improvements as the automatisation and parallelisation, which have allowed the characterisation of the microbiota to unprecedent depths rapidly and cost-effectively (1). At present, the targeted amplicon sequencing of the phylogenetic marker 16S ribosomal RNA (rRNA) gene is, by far, one of the most commonly used techniques to determine the structure and composition of the prokaryote communities (2).

Studies published during the last decade have assessed the mouth's microbiome using high-throughput 16S rRNA gene sequencing (3). To facilitate the analysis of complex microbial communities like the oral environment, amplicons derived from this technology are typically clustered into operational taxonomic units (OTUs) that are intended to correspond to taxonomic clades or monophyletic groups (4). Specifically, sequences are clustered based on a given similarity threshold, usually set at 97%, which has been conventionally regarded as the species-level correspondent (3,5).

Numerous OTU clustering algorithms have been integrated into the popular sequence-analysis pipelines, such as QIIME 2 (6), mothur (7), and USEARCH (8). Overall, existing methods for grouping 16S rRNA gene amplicons into OTUs can be categorised in three ways: *de novo*, closed-reference, and open-reference (9). While the first approach groups amplicons based on pairwise sequence distances, the second approach groups sequences that match a reference sequence from a database in the same OTU. The open-reference represents a combination of the two others, where sequences that do not adequately match the reference are grouped using a *de novo* method (9). However, none of these approaches produce the same results in terms of obtaining OTUs, even when using the same dataset (10,11). Moreover, even the same method can yield distinct results after only a minor parameter change (9).

In addition, it has been reported that different species can have very highly similar 16S rRNA gene sequences (12,13), which may lead to the grouping of distinct taxa in the same OTU. In fact, around 25% of OTUs constructed using the widely adopted ≥97% similarity threshold have been found to contain gene sequences from multiple species (12,13). These estimates were slightly different depending on the gene region studied but reached up to 35%

for variable regions 4-5 (13). Consequently, the construction of an OTU table can be affected, as can, by extension, taxonomic assignments, and microbial diversity results.

Due to the limitations of OTU clustering, other analyses based on establishing 100% sequence identity or single-nucleotide resolution have been proposed, such as zero-radius operational taxonomic units (14), oligotypes, or minimum entropy decomposition nodes (15), amplicon sequence variants (16) or suboperational taxonomic units (17). The most widely known pipelines based on the single-nucleotide resolution are DADA2 (16), Deblur (17), and UNOISE (14). These attempt to model the error of the sequencer and to cluster reads in a way that their distribution within clusters is consistent with such error (18); however, they differ in how this correction is done (19).

Several investigations have compared the two clustering approaches (OTUs *versus* single-nucleotide resolution) to discern which performs better (13,18-22). In general, the pipelines based on the single-nucleotide resolution have demonstrated superior sensitivity, specificity, and precision, and lower spurious sequence rates when compared to OTU algorithms (18,20). They allow for easier inter-study integration of biological features as amplicon sequence variants have intrinsic meaning independent of the reference database used, contrary to the study-specific nature of OTUs (20,23). However, single-nucleotide resolution algorithms, when analysing 16S rRNA gene data, can split a single genome into separate clusters (13). Furthermore, there is no consensus regarding the influence of the method chosen on the diversity results obtained. Meanwhile, some authors obtained minor differences between pipelines using the two clustering methods, with comparable alpha- and beta-diversity profiles (21,22); others evidenced distinct results even among those from the same approach (20).

Currently, more than 80% of recent studies on the oral microbiome performed their analyses based on OTU clustering. However, to our knowledge, no research has specifically evaluated the extent of the limitations of these sequence clustering-based methods by a similarity threshold in the oral microbiota. Consequently, the objectives of the present *in silico* investigation were to: 1) evaluate the coverage of a set of previously selected primer pairs to detect oral species having 16S rRNA sequence segments with ≥97% similarity; 2) describe oral

species with highly similar sequence segments and determine whether they belong to distinct genera or other higher taxonomic ranks.

### 3.3. MATERIALS AND METHODS

The complete analysis protocol applied in the present study is detailed in Figure 1.

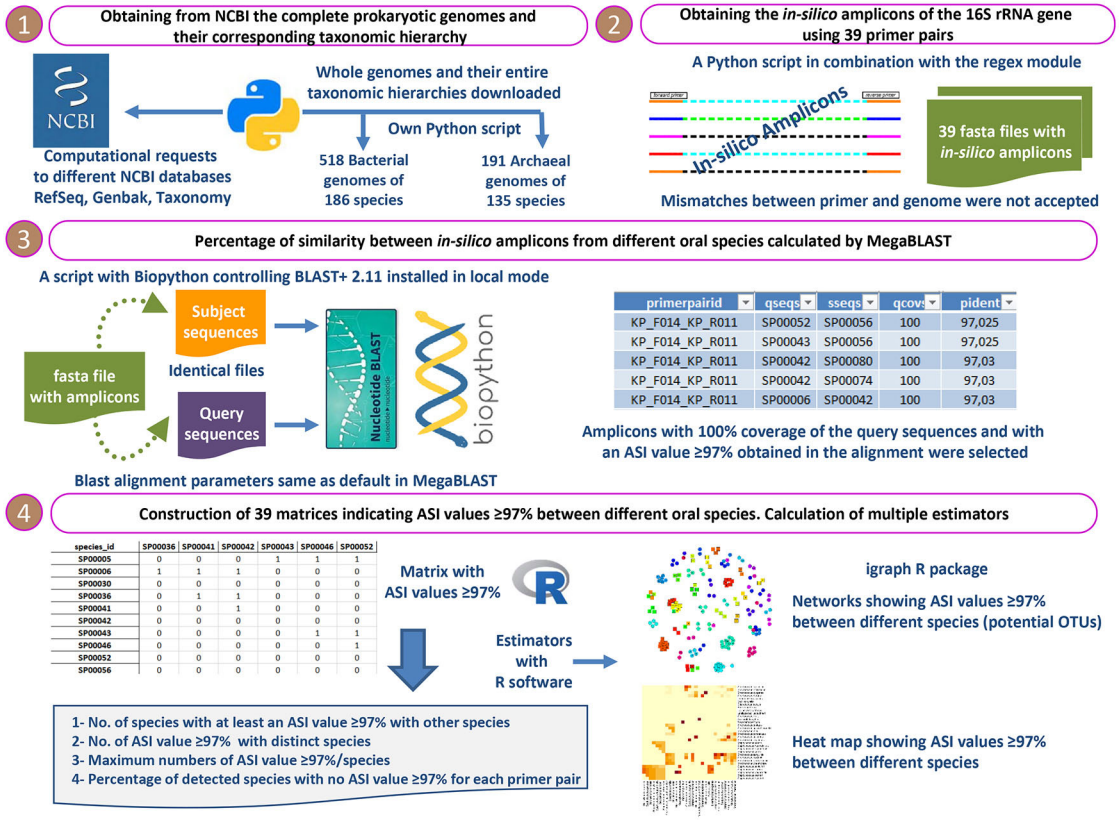



**Figure 1. The complete analysis protocol applied in the present *in silico* study.**

#### 3.3.1. Obtaining complete genomes of oral bacteria and oral archaea

The information on the bacterial taxa present in the oral cavity was obtained from the expanded human oral microbiome database (eHOMD) website (24). Of the 2074 genomes available on the site, we only selected the 518 that had a complete sequencing status for use in the research. These complete genomes have one or more GenBank identifiers (25), which were employed to access the complete sequences stored in the National Center for Biotechnology Information (NCBI) database (26). Additionally, an initial list of 177 different oral archaea and their corresponding GenBank identifiers (25), obtained as part of objective 1 (27), enabled us to access their complete sequences and annotations in the NCBI database (26). Integrating the "Entrez Programming Utilities (E-utilities)" tool (28) in our Python (version 3.9.0) (29) script allowed us to acquire the URLs needed to retrieve the information of interest from various NCBI databases, including Taxonomy (30), RefSeq (31), and GenBank (25).

Therefore, a total of 709 complete prokaryotic genomes from a total of 321 oral species were downloaded (more than one complete genome was analysed for several species). Finally, the complete taxonomic hierarchy (from superkingdom to strain) of all downloaded complete genomes was designated by the taxonomic identifier included in the annotated information in the NCBI database (26), all computationally performed with our script above. The taxonomy and NCBI identifiers of the oral-bacteria and archaea genomes are included in appendices S1 and S2, respectively.

### 3.3.2. Selecting the primer pairs and obtaining the *in silico* amplicons of the 16S rRNA gene

Thirty-three primer pairs with the best *in silico* coverage, as identified in objective 1, were selected, along with the six primer pairs used the most in the oral microbiome literature (27). These primer pairs were classified according to the mean length of their amplicons into: short primer pairs (S, 100-300 base pairs), medium primer pairs (M, 301-600 bps), and long primer pairs (L, >600 bps); and the domain targeted (bacteria, archaea, or both) (Table 1).

**Table 1. Selected primer pairs with high *in silico* coverage percentages targeting oral bacteria and/or archaea and those most used in the sequencing-based studies of the oral microbiome.**

| Bacterial-specific primer pair | ALC | F identifier | F Sequence 5-3 | F First post | F Last Post | R identifier | R Sequence 5-3 | R First post | R Last Post | Length (bps) | Region |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | S | KP_F048 | TACGGRAGGCAGCAG | 342 | 356 | OP_R043 | CCGCGRCTGCTGGCAC | 514 | 529 | 187 | 3-4 |
| | | OP_F098 | CCAGCAGCYGCGGTAAN | 517 | 533 | OP_R119 | GGACTACCRGGGTATCTAA | 787 | 805 | 288 | 4-5 |
| | | OP_F066 | GGMTTAGATACCC | 784 | 796 | KP_R040 | CCGTCAATTCMTTTGAGTTT | 906 | 925 | 141 | 5-6 |
| | | OP_F009 | GGATTAGATACCCBRGTAGTC | 784 | 804 | OP_R030 | TCACRRCACGAGCTGWCGAC | 1060 | 1079 | 295 | 5-7 |
| | | KP_F061 | ACTCAAAKGAATWGACGG | 908 | 925 | KP_R074 | GGGTYKCGCTCGTTR | 1099 | 1113 | 205 | 6-7 |
| | | OP_F101 | GAATTGRCGGGGRCC | 916 | 930 | OP_R030 | TCACRRCACGAGCTGWCGAC | 1060 | 1079 | 163 | 6-7 |
| | M | OP_F053 | GRGTTYGATYMTGGCTCAG | 9 | 27 | KP_R020 | CTGCTGCCTYCCGTA | 342 | 356 | 347 | 1-3 |
| | | KP_F048 | TACGGRAGGCAGCAG | 342 | 356 | KP_R031 | TACHVGGGTATCTAAKCC | 784 | 801 | 459 | 3-5 |
| | | KP_F048 | TACGGRAGGCAGCAG | 342 | 356 | OP_R073 | CRTACTHCHCAGGYG | 879 | 893 | 551 | 3-6 |
| | | KP_F051 | GTGCCAGCMGCNGCGG | 514 | 529 | KP_R041 | CGTCAATTCMTTTGAGTT | 907 | 924 | 410 | 4-6 |
| | | KP_F051 | GTGCCAGCMGCNGCGG | 514 | 529 | OP_R030 | TCACRRCACGAGCTGWCGAC | 1060 | 1079 | 565 | 4-7 |
| | | OP_F116 | YAACGAGCGCAACCC | 1099 | 1113 | KP_R060 | GACGGGCGGTGWGTRCA | 1390 | 1406 | 307 | 7-9 |
| | L | KP_F048 | TACGGRAGGCAGCAG | 342 | 356 | OP_R030 | TCACRRCACGAGCTGWCGAC | 1060 | 1079 | 737 | 3-7 |
| | | KP_F048 | TACGGRAGGCAGCAG | 342 | 356 | KP_R060 | GACGGGCGGTGWGTRCA | 1390 | 1406 | 1064 | 3-9 |
| | | KP_F056 | AYTGGGYDTAAAGNG | 572 | 576 | KP_R077 | GACGGGCGGTGTGTACAA | 1389 | 1406 | 834 | 4-9 |
| **Archaeal-specific primer pair** | **ALC** | **F identifier** | **F Sequence 5-3** | **F First post** | **F Last Post** | **R identifier** | **R Sequence 5-3** | **R First post** | **R Last Post** | **Length (bps)** | **Region** |
| | S | KP_F018 | GYGCASCAGKCGMGAAW | U | U | KP_R002 | TTACCGCGGCKGCTG | 518 | 532 | - | - 4* |
| | | OP_F066 | GGMTTAGATACCC | 784 | 796 | KP_R013 | GGCCATGCACCWCCTCTC | U | U | - | 5-6 |
| | M | KP_F018 | GYGCASCAGKCGMGAAW | U | U | KP_R032 | TACNVGGGTATCTAATCC | 784 | 801 | - | 3-5 |
| | | KP_F018 | GYGCASCAGKCGMGAAW | U | U | OP_R073 | CRTACTHCHCAGGYG | 879 | 893 | - | 3-5 |
| | | KP_F020 | CAGCMGCCGCGGTAA | 518 | 532 | KP_R013 | GGCCATGCACCWCCTCTC | U | U | - | 3-6 |
| | | KP_F022 | AGGAATTGGCGGGGGAGCA | U | U | KP_R063 | TACCTTGTTACGACTT | 1491 | 1506 | - | 5-9 |
| | L | OP_F114 | CCTAYGGGRBGCASCAG | 340 | 356 | KP_R013 | GGCCATGCACCWCCTCTC | U | U | - | 3-6 |
| | | KP_F018 | GYGCASCAGKCGMGAAW | U | U | KP_R063 | TACCTTGTTACGACTT | 1491 | 1506 | - | 3-9 |
| | | OP_F066 | GGMTTAGATACCC | 784 | 796 | OP_R016 | CGGTGTGTGCAAGGAG | U | U | - | 5-9 |
| **Bacterial and archaeal primer pair** | **ALC** | **F identifier** | **F Sequence 5-3** | **F First post** | **F Last Post** | **R identifier** | **R Sequence 5-3** | **R First post** | **R Last Post** | **Length (bps)** | **Region** |
| | S | OP_F114 | CCTAYGGGRBGCASCAG | 340 | 356 | KP_R002 | TTACCGCGGCKGCTG | 518 | 532 | 192 | 3-4 |
| | | KP_F020 | CAGCMGCCGCGGTAA | 518 | 532 | KP_R032 | TACNVGGGTATCTAATCC | 784 | 801 | 283 | 4-5 |
| | | OP_F066 | GGMTTAGATACCC | 784 | 796 | OP_R073 | CRTACTHCHCAGGYG | 879 | 893 | 109 | 5-6 |
| | M | OP_F114 | CCTAYGGGRBGCASCAG | 340 | 356 | KP_R031 | TACHVGGGTATCTAAKCC | 784 | 801 | 461 | 3-5 |
| | | OP_F114 | CCTAYGGGRBGCASCAG | 340 | 356 | OP_R073 | CRTACTHCHCAGGYG | 879 | 893 | 553 | 3-6 |
| | | KP_F020 | CAGCMGCCGCGGTAA | 518 | 532 | OP_R073 | CRTACTHCHCAGGYG | 879 | 893 | 375 | 4-6 |
| | L | OP_F114 | CCTAYGGGRBGCASCAG | 340 | 356 | OP_R121 | ACGGGCGGTGWGTRC | 1391 | 1405 | 1065 | 3-9 |
| | | KP_F020 | CAGCMGCCGCGGTAA | 518 | 532 | OP_R121 | ACGGGCGGTGWGTRC | 1391 | 1405 | 887 | 4-9 |
| | | OP_F066 | GGMTTAGATACCC | 784 | 796 | OP_R121 | ACGGGCGGTGWGTRC | 1391 | 1405 | 621 | 5-9 |
| **Most used primer pair** | **ALC** | **F identifier** | **F Sequence 5-3** | **F First post** | **F Last Post** | **R identifier** | **R Sequence 5-3** | **R First post** | **R Last Post** | **Length (bps)** | **Region** |
| | S | KP_F078 | GTGCCAGCMGCCGCGGTAA | 514 | 532 | OP_R010 | GGACTACHVGGGTWTCTAAT | 786 | 805 | 291 | 4-5 |
| | M | KP_F031 | AGAGTTTGATCCTGGCTCAG | 8 | 27 | KP_R021 | TTACCGCGGCTGCTGGCAC | 515 | 532 | 524 | 1-4 |
| | | KP_F047 | CCTACGGGNGGCWGCAG | 340 | 356 | KP_R035 | GACTACHVGGGTATCTAATCC | 784 | 804 | 464 | 3-5 |
| | | OP_F009 | GGATTAGATACCCBRGTAGTC | 784 | 868 | OP_R029 | ACGTCRTCCCCDCCTTCCTC | 1174 | 1193 | 409 | 5-8 |
| | L | KP_F014 | TCCAGGCCCTACGGG | U | U | KP_R011 | YCCGGCGTTGAMTCCAATT | U | U | - | 3-6 |
| | | KP_F034 | AGAGTTTGATCMTGGCTCAG | 8 | 27 | KP_R065 | TACGGYTACCTTGTTACGACTT | 1491 | 1512 | 1504 | 1-9** |

Primer pairs were selected based on the species coverage values (number of species detected /total species evaluated) in objective 1 (27). They were individually evaluated through regular expressions against *Escherichia coli* J01859 to define their positions. The U values represent a mismatch on the assessment and, therefore, the position cannot be confirmed with a guarantee. Gene regions were delimited as described by Baker et al. (32). *The primer pair gene region was 3 and **1-10 when considering the mode positions obtained in the analysis against the oral-bacteria and oral-archaea databases in objective 1 (27).

ALC= amplicon length category; bps= base pairs; F= forward; KP= Klindworth primer; L= long mean amplicon length category, >600 base pairs; M= medium mean amplicon length category, 301-600 base pairs; OP= oral primer; Post= position; R= reverse; S= short mean amplicon length category, 100-300 base pairs; U= unaligned with Escherichia coli; V= variable.

Applying our script in combination with Python's regex module (33), the direct and reverse sequences of each primer pair were used to obtain *in silico* sequence segments of the whole genomes analysed (hereafter referred to as *in silico* amplicons). An *in silico* amplicon was considered for subsequent analysis when the following conditions were present: 1) there is a zero mismatch of both primers (forward and reverse) of each pair; 2) the distance between the starting position of the forward primer and the ending position of the reverse primer is less than 2300 and higher than 100 nucleotides; 3) the *in silico* amplicon does not repeat within the same species.

All *in silico* amplicons from the same species, even if from different strains, were considered for analysis. For each primer pair, a fasta file was created where all *in silico* amplicons found were stored. The stored sequences were identified with the same taxonomic hierarchy as the genomes from which they were detected. As many *in silico*-amplicons were detected within the same species (differing by at least 1 nucleotide), the sequence variants were identified with correlative numbering at a new hierarchical level below the species name.

In the fasta files, all sequences included a species identifier (SPn) and a variant identifier (Vn) in their header, and then the header of each sequence also included the taxonomic hierarchy up to the variant level within each species. Finally, the *in silico* amplicons were obtained from 186 oral-bacterial and 135 oral-archaeal species.

### 3.3.3. Determination of the percentage of similarity between *in silico* amplicons of different oral species by MegaBLAST

A script with the NcbiblastnCommandline wrapper from Biopython (34) was developed to manage BLAST+ 2.11 (35) in the local mode from Biopython. This enabled the data obtained in the alignments to be easily transferred for later analysis on Python (29). The alignment parameters were configured to be the same as the default settings in MegaBLAST (36) since these settings were appropriate for the alignment between sequences with a similarity ≥95%.



All sequences belonging to the same fasta file for the same primer pair were aligned against themselves; in order to do this, each fasta file was inserted as subject and query in BLASTN

(37) for obtaining the percentage of similarity between *in silico* amplicons belonging to different oral species.

From the results obtained, *in silico* amplicons with 100% alignment coverage of the query sequences and with a similarity value ≥97% were selected. That is, alignments with the following BLAST+ estimates (35): qcovs= 100%, qcovhsp= 100%, qcovus= 100%, and pident ≥97% were selected. Of the above alignments obtained, the following were discarded as they were not of interest: 1) *in silico* amplicons with the same unique identifier (SPn + Vn); 2) *in silico* amplicons with the same species identifier; and 3) duplicate alignments.

If two different species had more than one *in silico* amplicon similarity value ≥97% (ASI97) among them, one of the alignments was chosen at random. The results of the highly similar species pairs, including taxonomic hierarchy data for both species, were then stored using the pandas (38) and xlsxwriter (39) Python modules.

**3.3.4. Construction of a matrix with oral species showing *in silico* amplicon similarity values ≥97% and calculation of descriptive statistical estimators**

A similarity matrix was created for each primer pair, where rows and columns had the species identifiers, and cells indicated with a number 1 the presence of an ASI97 between two different species. We then developed a script in R (version 4.0.3) (40) through which we calculated the following estimates for each analysed primer pair: 1) the number of species with at least one ASI97 with other species; 2) the total number of ASI97 between different species; 3) the mean and maximum numbers of ASI97 per species. In addition, we estimated the percentage of detected species (species coverage, SC) and the percentage of detected species without ASI97 for each primer pair (species coverage no ASI97, SC-NASI97). This last parameter was then used as a criterion for selecting the primers associated with a smaller number of oral species that may be erroneously clustered. The SC-NASI97 parameter will be influenced not only by the number of species with ASI, but also by the coverage percentages of each primer pair.



Finally, the bacterial and archaeal species pairs that showed an ASI97 were described and assessed whether they belonged to different genera or higher taxonomic ranks.

## 3.4. RESULTS

### 3.4.1. Evaluation of the primer pairs for detecting oral species with *in silico* amplicon similarity values ≥97%

The primer pairs that targeted bacteria had a mean of 91.88 (49.40%) bacterial species with an ASI97 and an average of 153.46 ASI97 containing distinct species. For those targeting archaea, these numbers were 65.60 (48.59%) and 162.26, respectively. If the primers used most in the oral microbiome literature were excluded, those with short amplicon lengths (unlike the SC percentages) had the lowest SC-NASI97 values for both bacteria (S= 39.54%) and archaea (S= 40.44%) compared to the medium length and long primers (M= 45.82% and 46.35%, respectively; L= 48.39% and 44.32%, respectively).

Figures 2 and 3 show the number of species with ASI97 and the number of ASI97 with each primer pair evaluated against bacteria and archaea, respectively; while figures 4 and 5 detail the percentages of coverage and coverage considering the presence or absence of ASI97 for both domains. Concerning the bacteria-specific primer pairs, the number of bacterial species with an ASI97 and the total number of ASI97 ranged from 37 and 32 with the most widely used primer KP_F031-KP_R021 (M; SC-NASI97= 54.30%), to 120 and 277 with OP_F066-KP_R040 (S; SC-NASI97= 24.19%), respectively. This latter primer also had the lowest SC-NASI97 value, while OP_F053-KP_R020 detected the highest number of species with no ASI97 (M; SC-NASI97= 65.05%). In addition, except for OP_F053-KP_R020, all the bacteria-specific primers had a maximum number of ASI97/species above five (range= 4 - 15 ASI97/species).

Concerning the archaea-specific primer pairs, the number of archaeal species with an ASI97 and the total number of ASI97 ranged from 24 and 96 with the widely used KP_F014-KP_R011 (L; SC-NASI97= 12.59%) to 89 and 240 with OP_F066-KP_R013 (S; SC-NASI97= 29.63%), respectively. The former primer detected the lowest number of species without an ASI97, and KP_F018-KP_R002 the highest (S; SC-NASI97= 51.11%). Moreover, all the archaea-specific primers had a maximum number of ASI97/species ≥10 (range= 10 - 13 ASI97/species).

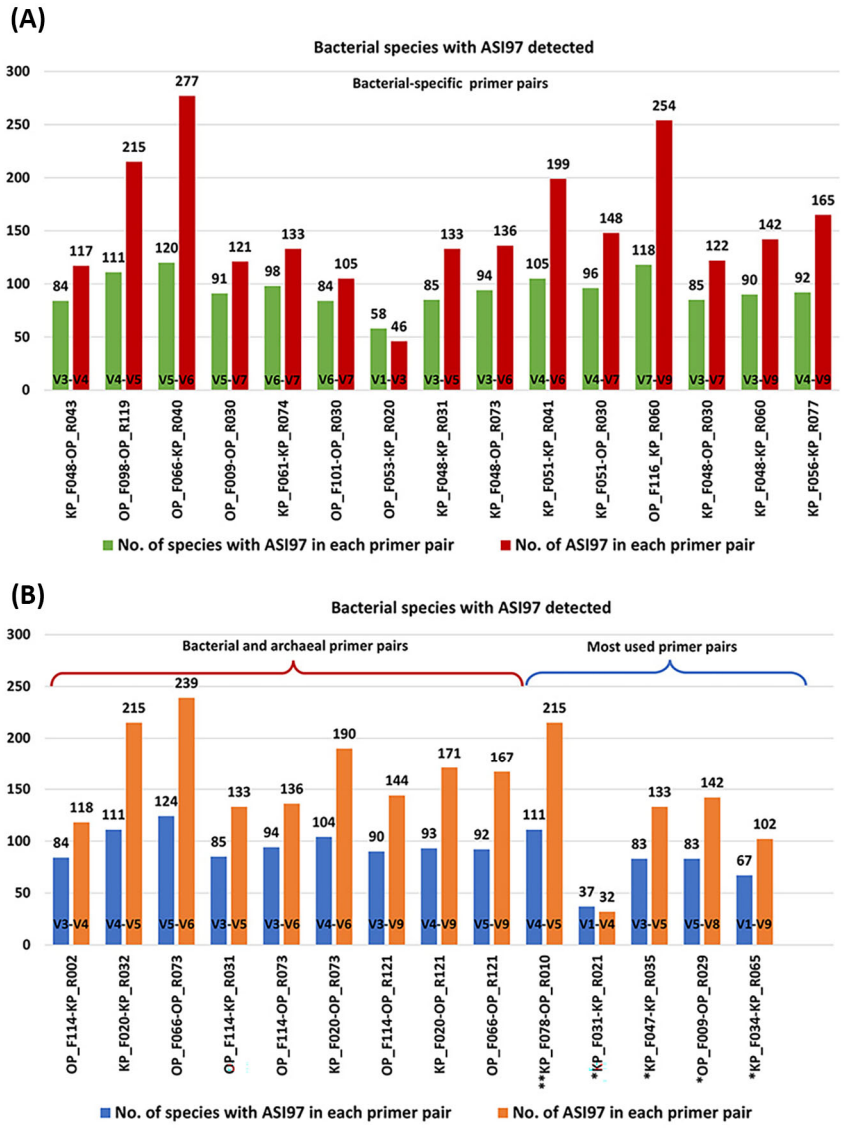



**Figure 2. Number of bacterial species with *in silico* amplicon similarity values ≥97% and number of *in silico* amplicon similarity values ≥97% with the primer pairs evaluated against the oral-bacteria genomes. (A) Estimates were obtained by the selected bacteria-specific primer pairs. (B) Estimates were obtained by the selected bacterial and archaeal primer pairs and the primer pairs used the most in the oral microbiome literature.**

Among the most commonly used primer pairs in the literature, those marked with an * are bacteria-specific and those with ** target both bacterial and archaea. ASI97= *in silico* amplicon similarity values ≥97%; F= forward; KP= Klindworth primer; No.= number; OP= oral primer; R= reverse.

Finally, using both the bacterial and archaeal primer pairs, the number of bacterial and archaeal species with an ASI97 and the total number of ASI97 ranged from 84 and 60 and 118 and 126, respectively, with OP_F114-KP_R002 (S; SC-NASI≥97= 47.31% for bacteria and 54.81% for archaea) to 124 and 95 and 239 and 286, respectively, with OP_F066-OP_R073 (S; SC-NASI≥97= 31.18% for bacteria and 22.96% for archaea). The latter primer also detected

the lowest number of species without an ASI97 and OP_F114-KP_R031 the highest (M; SC-NASI97= 51.08% for bacteria and 53.33% for archaea) (Figures 2b-5b). Most bacterial and archaeal primer combinations had maximum numbers of ASI97/species ≥10 (range= 9 - 14 ASI/species and 11 - 14 ASI/ species for both domains, respectively).

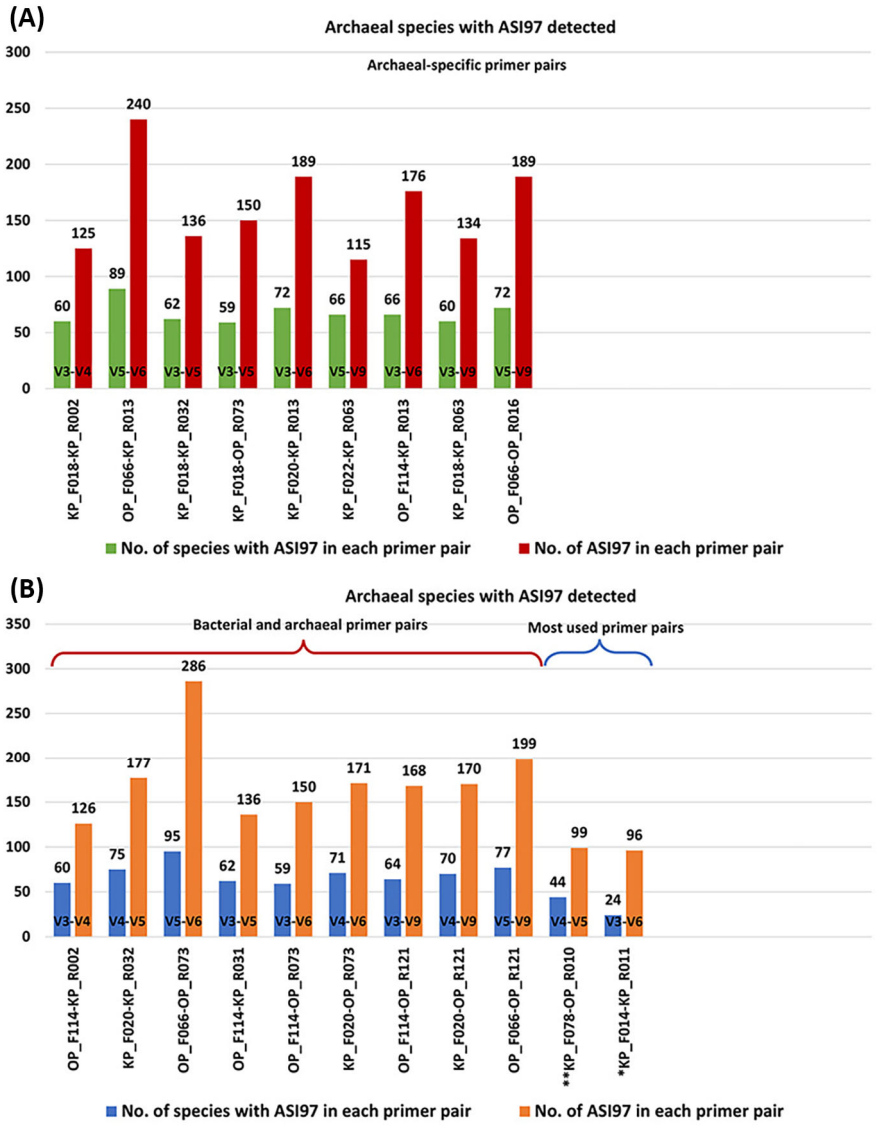



**Figure 3. Number of archaeal species with *in silico* amplicon similarity values ≥97% and number of *in silico* amplicon similarity values ≥97% with the primer pairs evaluated against the oral-archaea genomes. (A) Estimates were obtained by the selected archaea-specific primer pairs. (B) Estimates were obtained by the selected bacterial and archaeal primer pairs and the primer pairs used the most in the oral microbiome literature.**

Among the most commonly used primer pairs in the literature, those marked with an * are archaea-specific and those with ** target both bacterial and archaea. ASI97= *in silico* amplicon similarity values ≥97%; F= forward; KP= Klindworth primer; No.= number; OP= oral primer; R= reverse.

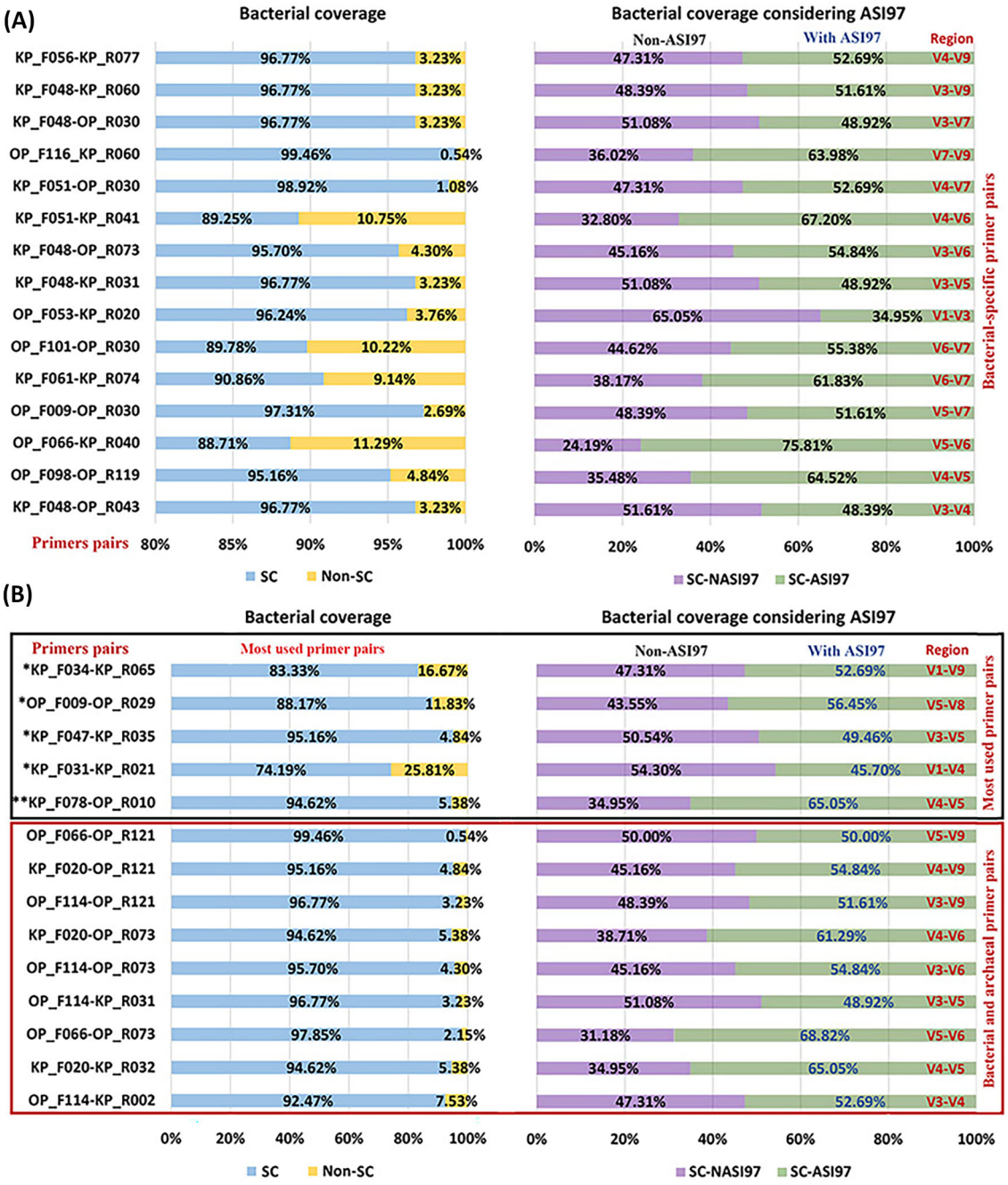



**Figure 4. Percentages of coverage and coverage considering the species with *in silico* amplicon similarity values ≥97% of the primer pairs evaluated against the oral-bacteria genomes. (A) Percentages were obtained by the selected bacteria-specific primer pairs. (B) Percentages were obtained by the selected bacterial and archaeal primer pairs and the primer pairs used the most in the oral microbiome literature.**

Among the most commonly used primer pairs in the literature, those marked with an * are bacteria-specific and those with ** target both bacterial and archaea. ASI97= *in silico* amplicon similarity values ≥97%; F= forward; KP= Klindworth primer; Non-SC= non-coverage of species; OP= oral primer; R= reverse; SC= species coverage; SC-ASI97= species coverage with *in silico* amplicon similarity values ≥97%; SC-NASI97= species coverage with no *in silico* amplicon similarity values ≥97%.

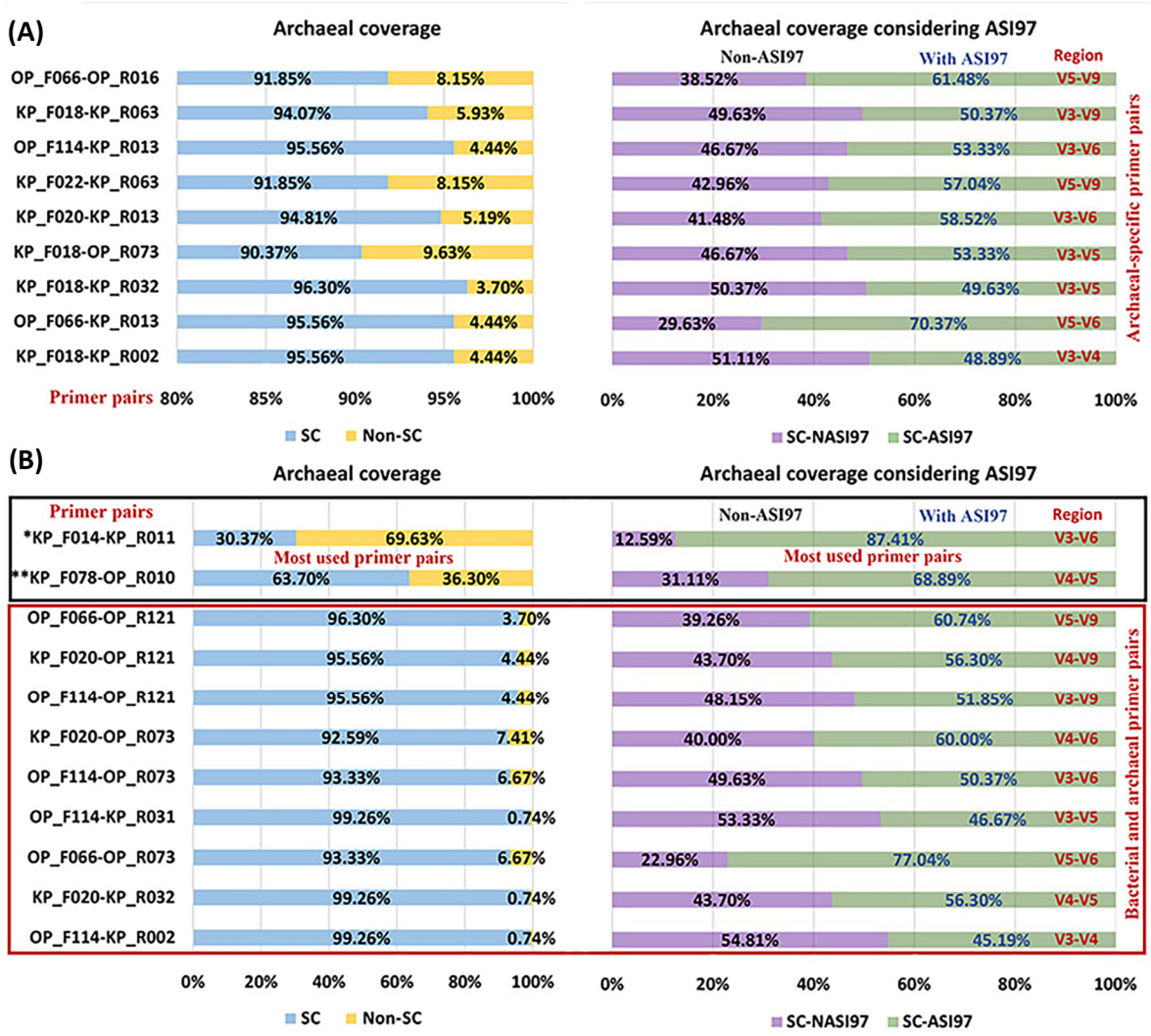



**Figure 5. Percentages of coverage and coverage considering the species with *in silico* amplicon similarity values ≥97% of the primer pairs evaluated against the oral-archaea genomes. (A) Percentages obtained by the selected archaea-specific primer pairs. (B) Percentages obtained by the selected bacterial and archaeal primer pairs and the primer pairs used the most in the oral microbiome literature.**

Among the most commonly used primer pairs in the literature, those marked with an * are archaea-specific and those with ** target both bacterial and archaea. ASI97= *in silico* amplicon similarity values ≥97%; F= forward; KP= Klindworth primer; Non-SC= non-coverage of species; OP= oral primer; R= reverse; SC= species coverage; SC-ASI97= species coverage with *in silico* amplicon similarity values ≥97%; SC-NASI97= species coverage with no *in silico* amplicon similarity values ≥97%.

Figures 6 and 7 are networks showing the potential clusters (hereinafter referred to as potential OTUs) with a ≥97% similarity threshold obtained with the primer pairs that presented the lowest SC-NASI97 values (F066-KP_R040 for bacteria, OP_F066-KP_R013 for archaea, and OP_F066-OP_R073 for bacteria and archaea), as well as one of the most used primer pairs in the oral microbiome literature, KP_F078-OP_R010. Thus, for example, for the primer pair F066-KP_R040, focusing on the one indicated by a dashed dotted line, 24 bacteria formed a

potential OTU, in which 10 genera, five families, and two orders were involved. As can be seen, there were species such as *Ligilactobacillus salivarius* (spp. 162) that presented high similarity only with two others, *Lacticaseibacillus paracasei* (spp. 196) and *Lacticaseibacillus rhamnosus* (spp. 243); while *Staphylococcus cohnii* (spp. 297) presented high similarity with 11 species (among which, *Enterococcus faecalis*, spp. 155; *Staphylococcus aureus*, spp. 163; *Levilactobacillus brevis*, spp. 181; *Lentilactobacillus buchneri*, spp. 237) belonging to four genera, three families and two orders. For the primer pair OP_F066-KP_R013, the potential OTU indicated was formed by nine archaea, involving seven genera and two families. Thus, *Desulfurococcus amylolyticus* (spp. 37) showed high similarity only with *Desulfurococcus mucosus* (spp. 68), while *Thermogladius caldera* had high similarity with five species (*Hyperthermus butylicus*, spp. 24; *Staphylothermus marinus*, spp. 26; *Staphylothermus hellenicus*, spp. 57; *Desulfurococcus mucosus*, spp. 68; *Pyrolobus fumarii*; sp. 82) belonging to four genera and two families.

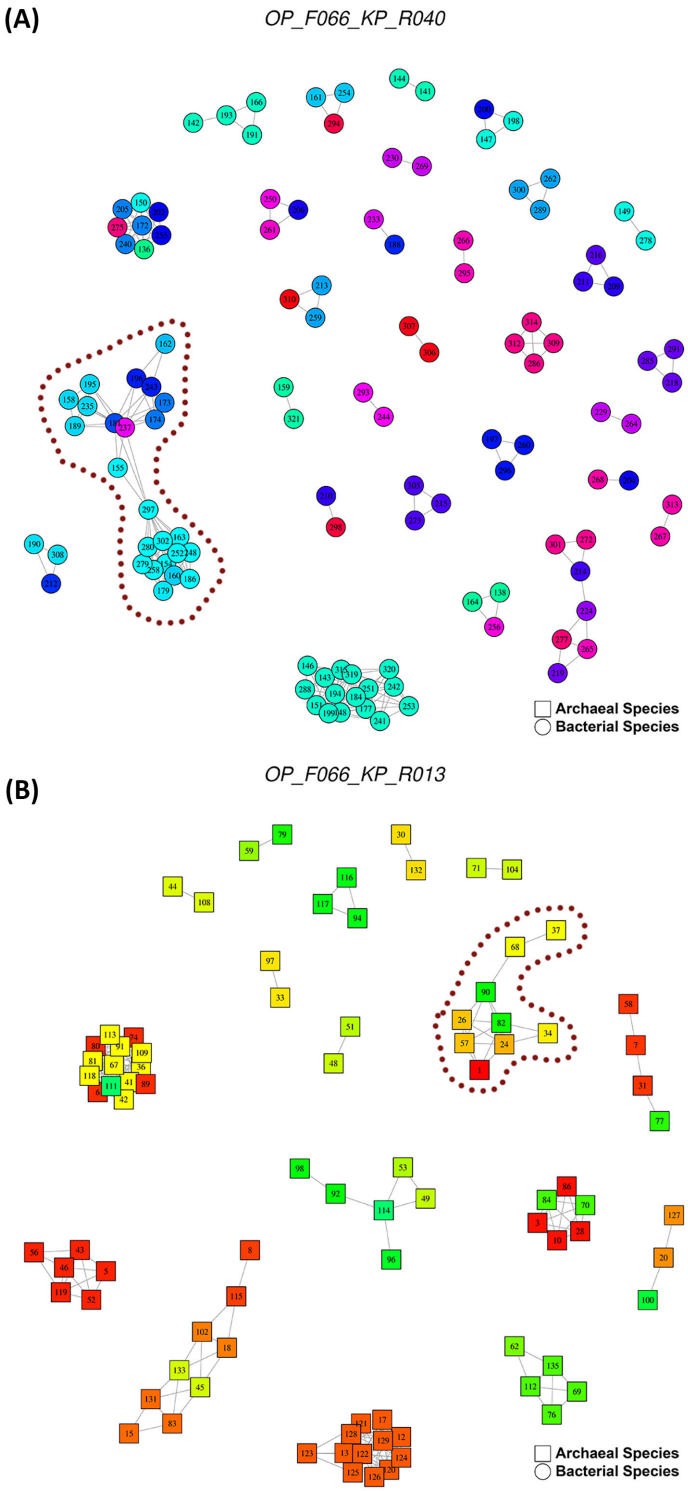



**Figure 6. Networks showing the potential OTUs with a ≥97% similarity threshold obtained with the specific primer pairs. (A) OP_F066-KP_R040 for bacteria (120 species with ASI97, 277 ASI97). (B) OP_F066-KP_R013 for archaea (89 species with ASI97, 240 ASI97).**

In the graphs, each node represents an oral species, the color indicates the genus and the number refers to the species identifier, whose assigned species are detailed in appendices S3 and S4. Each edge represents the presence of a ≥97% similarity between different species. resulting in clusters of possible OTUs. The graphs were made using the igraph package (version 1.2.6) (41).

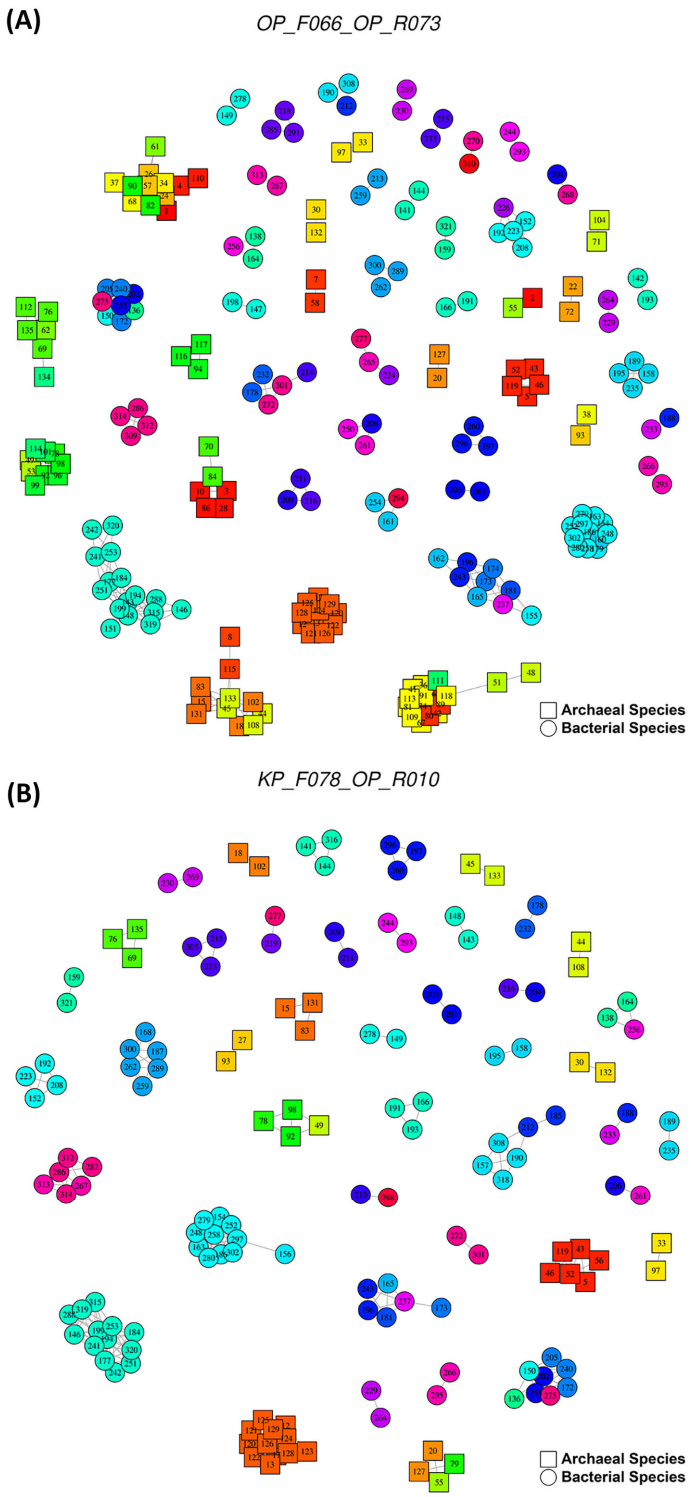



**Figure 7. Networks showing the potential OTUs with a ≥97% similarity threshold obtained with the primer pairs for bacteria and archaea. (A) OP_F066-OP_R073 (219 species with ASI97, 525 ASI97). (B) KP_F078-OP_R010 (primer pair widely used in the oral microbiome literature; 155 species with ASI97, 314 ASI97).**

In the graphs, each node represents an oral species, the color indicates the genus and the number refers to the species identifier, whose assigned species are detailed in appendices S3 and S4. Each edge represents the presence of a ≥97% similarity between different species. resulting in clusters of possible OTUs. The graphs were made using the igraph package (version 1.2.6) (41).

### 3.4.2. Description of the distinct pairs of oral-bacteria species and oral-archaea species with *in silico* amplicon similarity values ≥97%

One-hundred and forty-nine (80.11%) of the oral-bacteria species and 108 (80.00%) of the oral-archaea species analysed had an ASI97 with at least one distinct species (Figures 8 and 9; Appendices S3 and S4). Among them, it is worth mentioning because of their importance in both oral health and disease, *Aggregatibacter actinomycetemcomitans, Campylobacter concisus, Campylobacter curvus, Fusobacterium nucleatum, Rothia dentocariosa, Streptococcus mitis, Streptococcus mutans, Streptococcus oralis, Tannerella forsythia,* and *Treponema denticola*; regarding archaea, *Candidatus Nitrososphaera evergladensis, Halovivax ruber, Methanobrevibacter smithii, Methanococcus maripaludis, Methanosalsum zhiliniae, Methanosarcina barkeri, Methanosarcina mazei, Methanosarcina vacuolata, Methanosphaera stadtmanae, Natronococcus occultus.* There were 30 distinct bacterial and 27 distinct archaeal species that could be clustered with a maximum of ≥10 different species when all the analysed primer pairs were used. Most of these bacterial species belonged to genera *Streptococcus* and *Staphylococcus*; as for archaeal species, to genera *Methanosarcina*, *Thermococococcus,* and *Pyrococcus* (Appendices S3 and S4).

Conversely, 37 (19.89%) bacterial species and 27 (20.00%) archaeal species, including *Filifactor alocis*, *Porphyromonas gingivalis*, *Prevotella intermedia*, *Treponema pallidum*, *Veillonella parvula*, and *Sulfolobus acidocaldarius* did not have ASI≥97% with other taxa (Appendices S5 and S6).

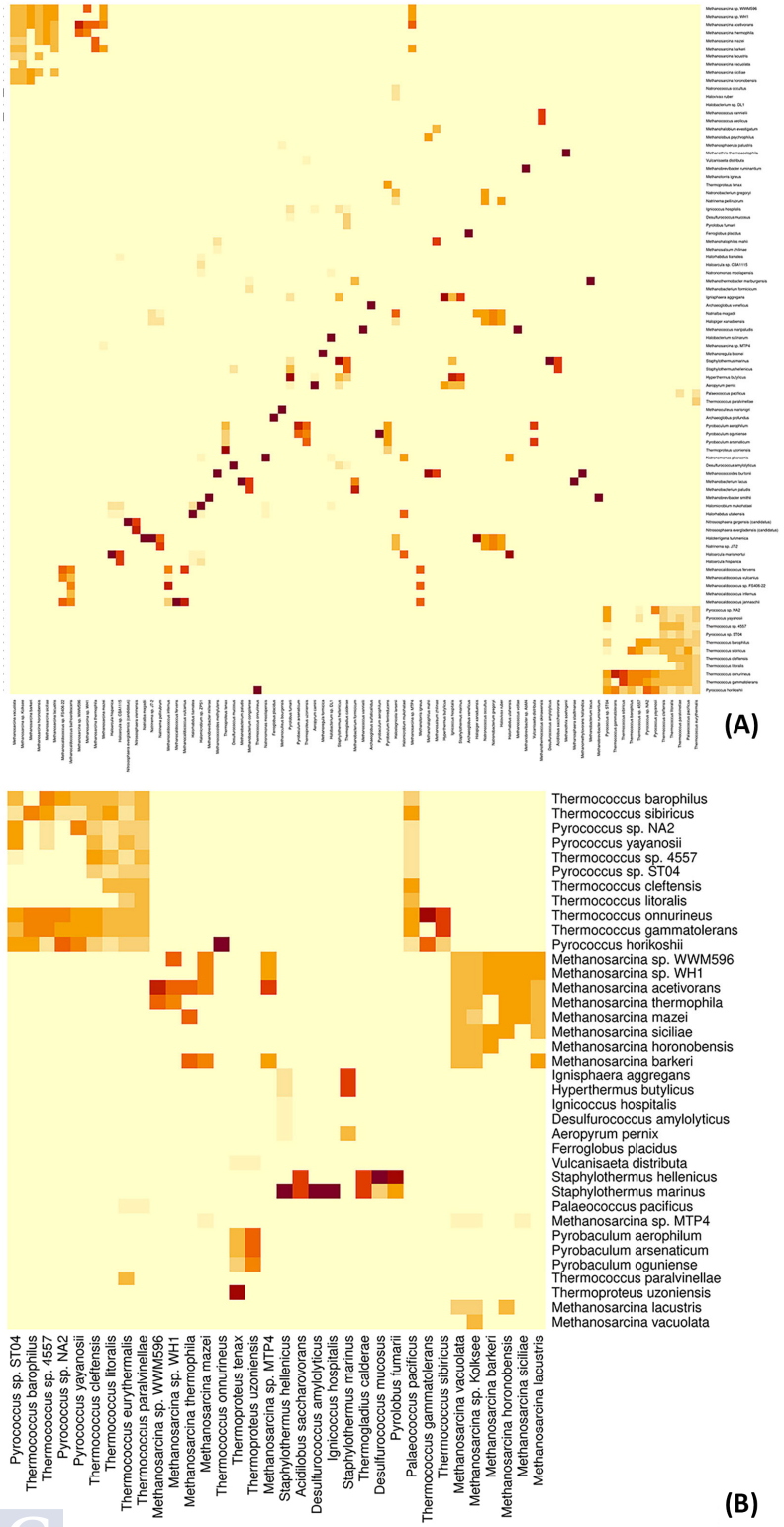



**Figure 8. Heat map showing the presence of *in silico* amplicon similarity values ≥97% between pairs of different bacterial species. (A) Global perspective. (B) Partial perspective, involving species belonging to genera *Staphylococcus*, *Streptococcus*, *Tannerella*, and *Treponema*.**

The intensity of the colour indicates the frequency, i.e. the number of times a species pair had an *in silico* amplicon similarity value ≥97% in the different primer pairs evaluated.

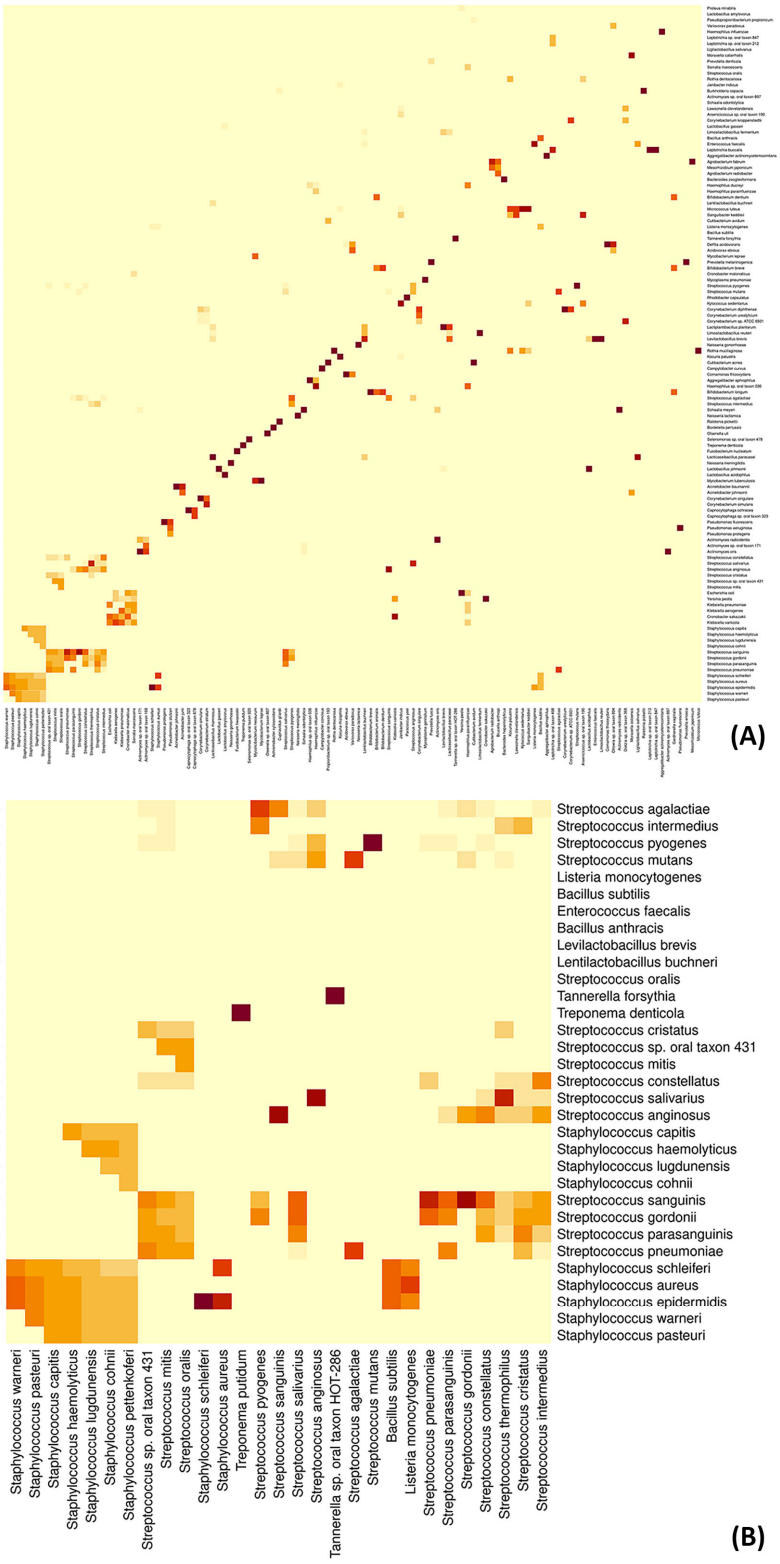



**Figure 9. Heat map showing the presence of *in silico* amplicon similarity values ≥97% between pairs of different archaeal species. (A) Global perspective. (B) Partial perspective, involving species belonging to genera *Methanosarcina*, *Pyrococcus*, *Staphylothermus*, *Thermococcus*, and *Thermoproteus*.**

The intensity of the colour indicates the frequency, i.e. the number of times a species pair had an *in silico* amplicon similarity value ≥97% in the different primer pairs evaluated.

All the primers targeting bacteria enabled us to detect 4450 two-on-two relationships between 408 distinct pairs of oral-bacteria species with an ASI97. Eighteen of these different taxa pairs were obtained with the 29 primer pairs analysed (frequency= 29, parameter defined as the number of times that a pair of species had an ASI97 in the different primer pairs evaluated), which belonged to the genera *Actinomyces*, *Lactobacillus*, *Neisseria*, *Staphylococcus*, and *Streptococcus*. Conversely, 50 species pairs with an ASI97 were detected once by only one primer pair (frequency= 1) (Appendix S7). Although the two-on-two relationships mostly involved species from the same genera (3641; 81.82%), 809 relationships (18.18%) were constituted by taxa from different genera. Thus, the combination of species from *Klebsiella* with others from *Cronobacter* occurred most frequently (frequency= 99) followed by *Klebsiella-Serratia*, *Escherichia-Klebsiella*, *Cronobacter-Escherichia*, and *Aggregatibacter-Haemophilus* (frequencies= 67 - 28). For higher taxonomic ranks, 293 (6.58%) two-on-two relationships were between species pairs with an ASI97 belonged to distinct families, with *Enterobacteriaceae* and *Yersiniaceae* being the most frequently detected (frequency= 153); even, there were 26 (0.58%) relationships between species pairs with an ASI97 from different orders, like *Bacillales* and *Lactobacillales* or *Enterobacterales* and *Pasteurellales* (frequencies= 10 and 10, respectively) (Appendix S8).

The primers targeting archaea enabled us to detect 3232 two-on-two relationships between 340 different pairs of archaeal species with an ASI97. All primer pairs analysed identified seven pairs of species (frequency= 20), which belonged to the genera *Methanobrevibacter* and *Methanocaldococcus*. There were 66 species pairs detected only once by only one primer pair (frequency= 1) (Appendix S9). Again, most of the two-on-two relationships were between archaeal species from the same genera (2359, 72.99%), but 873 (27.01%) relationships involved taxa pairs with an ASI97 from distinct genera. The combination of species from *Pyrococcus* and *Thermococcus* occurred the most, by far, (frequency= 428), followed by *Palaeococcus* and *Thermococcus* (frequency= 109). For higher taxonomic ranks, 35 (1.08%) relationships were species pairs with an ASI97 from distinct families, such as *Desulfurococcaceae* and *Pyrodictiaceae* (frequency= 27), also belonged to distinct orders (3 relationships; 0.09%), or even classes (1; 0.03%) (Appendix S10).

## 3.5. DISCUSSION

The high degree of similarity between full-length 16S rRNA sequences from distinct species, or even genera, has been reported in the literature (12,13), leading to questions about the reliability of diversity estimates derived from sequence clustering methods based on a given similarity threshold. Using full-length genes and a ≥97% similarity threshold, some authors have detected that around a quarter of constructed OTUs contain sequences from multiple species (12,13) and about a tenth from distinct genera (12). These estimates were obviously higher when gene regions were assessed instead of full sequences. Schloss et al. (13) found that, with a ≥97% similarity threshold and applying the OptiClust algorithm (42), 31.7%, 34.3% and 34.8% of the OTUs assessed had 16S rRNA amplicons from distinct species in the variable regions 3-4, 4, and 4-5, respectively (13). However, these investigations did not focus on taxa inhabiting a specific environment, despite the importance of conducting 16S rRNA gene-based research using habitat-specific databases (43). Consequently, we used primer pairs targeting several variable regions of the 16S rRNA gene (27) to determine the number of different oral-bacterial and oral-archaeal species with *in silico* amplicon similarity values ≥97% (ASI97), as well as the potential OTUs that might contain distinct species. Moreover, for the first time in this kind of analysis, we described the specific taxa of the oral ecosystem with highly similar sequence segments, specifying if they belong to different genera or other higher taxonomic ranks.

### 3.5.1. Evaluation of the primer pairs for detecting oral species with *in silico* amplicon similarity values ≥97%

In the present study, the primer pairs that targeted bacteria had a mean of 91.88 (49.40%) bacterial species with an ASI97 and an average of 153.46 potential OTUs containing distinct species. For those targeting archaea, these numbers were 65.60 (48.59%) and 162.26, respectively. Using the percentage species coverage with no *in silico* amplicons similarity ≥97% (SC-NASI97) as a selection criterion, the optimum primer pair for detecting oral bacteria was OP_F053-KP_R020. Although the primer used most in the oral microbiome studies, KP_F031-KP_R021 identified slightly fewer species with an ASI97 (37 *vs.* 58) and number of ASI97 (32 *vs.* 46); its SC-NASI97 was also lower than that of OP_F053-KP_R020 (54.30% *vs.* 65.05%). The primer pair producing the best estimates for detecting oral archaea was KP_F018-KP_R002. Again, the widely used primer KP_F014-KP_R011, although it only detected a few

species with an ASI97 (24 *vs.* 60) and number of ASI97 (96 *vs.* 125), however, also had a considerably lower SC-NASI97 than that of KP_F018-KP_R002 (12.59% *vs.* 51.11%). Lastly, we recommend the primer OP_F114-KP_R031 for detecting oral bacteria and archaea simultaneously. OP_F114-KP_R002, meanwhile, identified slightly fewer taxa with an ASI≥97% (for bacteria= 84 and for archaea= 60 *vs.* 85 and 62) and number of ASI97 (118 and 126 *vs.* 133 and 136) but had a lower SC-NASI≥97% (47.31% and 54.81% *vs.* 51.08% and 53.33%). In addition, as previously observed in objectives 1 and 2 (27,44), none of the primer combinations that are most commonly employed in sequencing-based studies of the oral microbiome were among the best. Specifically, the species coverage of KP_F078-OP_R010, a primer described by Caporaso (45), fell from 94.62% for bacteria and 63.70% for archaea as described in objective 2 (44) to 34.95% and 31.11%, respectively; when considering the species with an ASI97, possibly generating as many as 215 and 99 potential bacterial and archaeal OTUs, respectively; that contain different species.

### 3.5.2. Description of the distinct pairs of oral-bacteria species and oral-archaea species with *in silico* amplicon similarity values ≥97%

Around 80% of the oral-bacteria and oral-archaea species analysed had an ASI97 with at least another species. The widely-known bacterial periodontopathogens *F. nucleatum* and *T. denticola* (46-48) had similar *in silico* amplicons to *Fusobacterium hwasookii* and *Treponema putidum*, respectively, which have also been detected in periodontal lesions (49,50). Interestingly, other bacteria with high *in silico* amplicon similarities had antagonistic roles in oral health and disease. Examples are: the health-associated *C. concisus* and the initially periodontitis-associated *C. curvus* (51); the health-related *Rothia mucilaginosa* (52) and the decay-abundant *R. dentocariosa* (53,54); the commensal *S. mitis, oralis,* and *salivarius*; the caries-associated *S. mutans* (46,55,56); and the periodontal health-related *Tannerella* sp. oral taxon HOT-286 (57,58) and the periodontitis-related *T. forsythia* (46-48). Furthermore, relevant oral-disease associated species, such as *A. actinomycetemcomitans* (46,59) and *R. dentocariosa* (53,54), were among those that had an ASI97 with taxa from distinct genera. Regarding the archaea, we found that four Methanosarcina species found in healthy and periodontitis pockets, namely *barkeri*, *lacustris*, *mazeii,* and *vacuolata* (60), were highly similar. Moreover, *H. ruber*, *Methanotorris igneus*, *M. zhilinae,* and *N. occultus*, which are reported to be among the 10 most

abundant species in both healthy and periodontitis subjects (60), had an ASI97 with several taxa from distinct genera.

Schloss (13) has recently stated that the risks of artificially splitting a genome into multiple amplicon sequence variants (ASVs) are greater than those of clustering ASVs from different species into the same OTU when using broad distance thresholds. However, considering the results obtained in the present study, our opinion is that the latter approach should be avoided in the analysis of the oral microbiota if the aim is to associate species with specific clinical conditions. *In silico* amplicons from species traditionally associated with contrary health conditions, like those described above, can be grouped with a ≥97% similarity threshold. This would result in both an overabundance of the single species representing the OTU and an underestimation of the diversity of the community, with other species within the OTU overlooked. Consequently, it would be better to use the lowest possible level of resolution, i.e., the variant level (23), and databases specifically designed for taxonomic identifications of taxa at this level (43).

It has been demonstrated that distinct OTU clustering approaches, or even the same method, can yield uneven results for the same dataset (9-11). Therefore, we decided to analyse the 97% similarity relationships between oral species, without considering the influence of any clustering algorithm. Consequently, the results presented here are an approximation of the different oral species that could be grouped in potential OTUs.

### 3.5.3. Limitations of the present study

The main limitation of our study is that we have only considered one, randomly selected, of all possible *in silico* amplicons with ASI97 between two different species to establish the existence of a close relationship between the two. Another consideration is that we were only able to evaluate 25% of the oral microorganism genomes listed on the eHOMD website, as the remainder were not fully sequenced. This absence of complete genomes reduced the number of species investigated to 35% of those set out on the site. Although the analysis could have been performed on annotations of the 16S rRNA gene sequences from oral microbes, we preferred to use complete genomes, thereby ensuring the high quality of the sequences reviewed. The reasons why we adopted this approach were: 1) Edgar (61) estimated that the taxonomy

annotation error rate of the ribosomal database project (RDP) database (62) is ~10%; on the other hand, he found 249,490 identical sequences with conflicting annotations in SILVA v128 (63) and Greengenes v13.5 (64) at ranks up to phylum (7,804 conflicts), indicating that the annotation error rate in these databases is ~17%; 2) we have verified in objective 1 that a very high percentage of 16S rRNA gene annotations present a loss of information of up to 60 - 70 nucleotides in regions 1 and 9 of the sequences, which invalidates their use (27); 3) most of the complete genomes evaluated here are isolates that were sequenced with Sanger technology or with second-generation technology (shorter sequences than Sanger). In both cases, contig scaffolding algorithms were used to construct the complete genomes from the sequences with a minimum coverage of 8x for Sanger sequences and 30x in the case of second-generation technologies (65). In these types of assemblies, positions within the genome that did not have high coverage included non-specific nucleotides. In the present study, we discarded genomes that included more than 20 consecutive unspecific positions; 4) in addition, many genomes were downloaded from the NCBI RefSef database (31), where the annotations of the complete genomes were manually curated or re-annotated concerning the information provided by the original author, including their taxonomic hierarchy. Thus, our results highlight only part of a much more extensive problem.

### 3.6. CONCLUSIONS

In conclusion, the tested primer pairs targeting bacteria and/or archaea detected an average of more than 150 potential OTUs that might contain different species, when ≥97% similarity threshold was used. According to the SC-NASI97 parameter, the best primer pairs were: OP_F053-KP_R020 for bacteria (region 1-3; primer pair position for *Escherichia coli* J01859.1: 9-356); KP_F018-KP_R002 for archaea (4 undefined-532); and OP_F114-KP_R031 for both (3-5; 340-801). Around 80% of the oral-bacteria and oral-archaea species analysed had an ASI97 with at least one other species. These very similar species play different roles in the oral microbiota and belong to bacterial genera such as *Campylobacter*, *Rothia*, *Streptococcus*, and *Tannerella*, and archaeal genera such as *Halovivax*, *Methanosalsum*,and *Methanosarcina*. Moreover, ~20% and ~30% of these two-by-two similarity relationships were established between species from different bacterial and archaeal genera, respectively. Even taxa from distinct families, orders, and classes could be grouped in the same potential OTU. Consequently, regardless of the primer pair used, sequence-clustering with ≥97% similarity provides an inaccurate description of oral-bacterial and oral-archaeal species, which can greatly affect microbial diversity parameters. As a result, OTU clustering conditions the credibility of associations between some oral species and certain health and disease conditions. This significantly limits the comparability of the microbial diversity findings reported in oral microbiome literature.

### 3.7. REFERENCES

(1) Midha MK, Wu M, Chiu KP. Long-read sequencing in deciphering human genetics to a greater depth. Hum Genet. 2019 Dec;138(11-12):1201-15.

(2) Davidson RM, Epperson LE. Microbiome sequencing methods for studying human diseases. Methods Mol Biol. 2018;1706:77-90.

(3) Zaura E, Pappalardo VY, Buijs MJ, Volgenant CMC, Brandt BW. Optimizing the quality of clinical studies on oral microbiome: a practical guide for planning, performing, and reporting. Periodontol 2000. 2021 Feb;85(1):210-36.

(4) Edgar R. UPARSE: highly accurate OTU sequences from microbial amplicon reads. Nat Methods. 2013 Oct;10(10):996-8.

(5) Stackebrandt E, Goebel, BM. Taxonomic note: a place for DNA-DNA reassociation and 16s rRNA sequence analysis in the present species definition in bacteriology. Int J Syst Evol Microbiol. 1994 Oct;44(4):846-9.

(6) Bolyen E, Rideout JR, Dillon MR, Bokulich NA, Abnet CC, Al-Ghalith G, et al. Reproducible, interactive, scalable and extensible microbiome data science using QIIME 2. Nat Biotechnol. 2019 Aug;37(8):852-7.

(7) Schloss PD, Westcott SL, Ryabin T, Hall JR, Hartmann M, Hollister EB, et al. Introducing mothur: open-source, platform-independent, community-supported software for describing and comparing microbial communities. Appl Environ Microbiol. 2009 Dec;75(23):7537-41.

(8) Edgar R. Search and clustering orders of magnitude faster than BLAST. Bioinformatics. 2010 Oct;26(19):2460-1.



(9) Wei Z, Zhang X, Cao M, Liu F, Qian Y, Zhang S. Comparison of methods for picking the operational taxonomic units from amplicon sequences. Front Microbiol. 2021 Mar; 12:644012. doi: 10.3389/fmicb.2021.644012.

(10) He Y, Caporaso JG, Jiang XT, Sheng HF, Huse SM, Rideout JR, et al. Stability of operational taxonomic units: an important but neglected property for analyzing microbial diversity. Microbiome. 2015 May;3:20. doi: 10.1186/s40168-015-0081-x.

(11) Westcott SL, Schloss PD. De novo clustering methods outperform reference-based methods for assigning 16S rRNA gene sequences to operational taxonomic units. PeerJ. 2015 Dec;3:e1487. doi: 10.7717/peerj.1487.

(12) Větrovský T, Baldrian P. The variability of the 16S rRNA gene in bacterial genomes and its consequences for bacterial community analyses. PLoS One. 2013;8(2):e57923. doi: 10.1371/journal.pone.0057923.

(13) Schloss PD. Amplicon sequence variants artificially split bacterial genomes into separate clusters. mSphere. 2021 Aug;6(4):e0019121. doi: 10.1128/mSphere.00191-21.

(14) Edgar R. UNOISE2: improved error-correction for Illumina 16S and ITS amplicon sequencing. Preprint at bioRxiv. 2016. doi: 10.1101/081257.

(15) Eren AM, Morrison HG, Lescault PJ, Reveillaud J, Vineis JH, Sogin ML. Minimum entropy decomposition: unsupervised oligotyping for sensitive partitioning of high-throughput marker gene sequences. ISME J. 2015 Mar;9(4):968-79.

(16) 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.

(17) Amir A, McDonald D, Navas-Molina JA, Kopylova E, Morton JT, Zech Xu Z, et al. Deblur rapidly resolves single-nucleotide community sequence patterns. mSystems. 2017 Mar;2(2):e00191-16. doi: 10.1128/mSystems.00191-16.



(18) Caruso V, Song X, Asquith M, Karstens L. Performance of microbiome sequence inference methods in environments with varying biomass. mSystems. 2019 Feb;4(1):e00163-18. doi: 10.1128/mSystems.00163-18.

(19) Nearing JT, Douglas GM, Comeau AM, Langille MGI. Denoising the denoisers: an independent evaluation of microbiome sequence error-correction approaches. PeerJ. 2018 Aug;6:e5364. doi: 10.7717/peerj.5364.

(20) Prodan A, Tremaroli V, Brolin H, Zwinderman AH, Nieuwdorp M, Levin E. Comparing bioinformatic pipelines for microbial 16S rRNA amplicon sequencing. PLoS One. 2020 Jan;15(1):e0227434. doi: 10.1371/journal.pone.0227434.

(21) Abellan-Schneyder I, Matchado MS, Reitmeier S, Sommer A, Sewald Z, Jan Baumbach J, et al. Primer, pipelines, parameters: issues in 16S rRNA gene sequencing. mSphere. 2021 Feb;6(1):e01202-20. doi: 10.1128/mSphere.01202-20.

(22) García-López R, Cornejo-Granados F, Lopez-Zavala A, Cota-Huízar A, Sotelo-Mundo R, Gómez-Gil B, et al. OTUs and ASVs produce comparable taxonomic and diversity from shrimp microbiota 16S profiles using tailored abundance filters. Genes (Basel). 2021 Apr;12(4):564. doi: 10.3390/genes12040564.

(23) Callahan BJ, McMurdie PJ, Holmes SP. Exact sequence variants should replace operational taxonomic units in marker-gene data analysis. ISME J. 2017 Dec;11(12):2639-43.

(24) F Escapa I, 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;3(6):e00187-18. doi: 10.1128/mSystems.00187-18.

(25) Clark K, Karsch-Mizrachi I, Lipman DJ, Ostell J, Sayers EW. GenBank. Nucleic Acids Res. 2016 Jan;44:D67-72. doi: 10.1093/nar/gkv1276.

(26) NCBI Resource Coordinators. Database resources of the National Center for Biotechnology Information. Nucleic Acids Res. 2016 Jan;44(D1):D7-19. doi: 10.1093/nar/gkv1290.

(27) Regueira-Iglesias A, Vázquez-González L, Balsa-Castro C, Vila-Blanco N, Blanco-Pintos T, Tamames J, et al. In silico evaluation and selection of the best 16S rRNA gene primers for use in next-generation sequencing to detect oral bacteria and archaea. Accepted for publication in Microbiome. Preprint at Research Square. 2021. doi: 10.21203/rs.3.rs-516961/v1.

(28) National Center for Biotechnology Information. Entrez programming utilities help. 2010; Available at: https://www.ncbi.nlm.nih.gov/books/NBK25501/.

(29) Python Software Foundation. Python. Version 3.9.0. 2020; Available at: http://www.python.org/.

(30) Schoch CL, Ciufo S, Domrachev M, Hotton CL, Kannan S, Khovanskaya R, et al. NCBI Taxonomy: a comprehensive update on curation, resources and tools. Database (Oxford). 2020 Jan;2020:baaa062. doi: 10.1093/database/baaa062.

(31) O'Leary NA, Wright MW, Brister JR, Ciufo S, Haddad D, McVeigh R, et al. Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. Nucleic Acids Res. 2016 Jan;44(D1): D733-45. doi: 10.1093/nar/gkv1189.

(32) Baker GC, Smith JJ, Cowan DA. Review and re-analysis of domain-specific 16S primers. J Microbiol Methods. 2003 Dec;55(3):541-55.

(33) Barnett M. regex. 2020; Available at: https://pypi.org/.

(34) Cock PJ, Antao T, Chang JT, Chapman BA, Cox CJ, Dalke A, et al. Biopython: freely available Python tools for computational molecular biology and bioinformatics. Bioinformatics. 2009 Jun;25(11):1422-23.

(35) Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, et al. BLAST+: architecture and applications. BMC Bioinformatics. 2009 Dec;10:421. doi: 10.1186/1471-2105-10-421.

(36) Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. J Mol Biol. 1990 Oct;215(3):403-10.

(37) Chen Y, Ye W, Zhang Y, Xu Y. High speed BLASTN: an accelerated MegaBLAST search tool. Nucleic Acids Res. 2015 Sep;43(16):7762-8.

(38) McKinney W. Data structures for statistical computing in Python. In: van der Walt S, Millman J, editors. Proceedings of the 9th Python in Science Conference; 2010; Austin. Texas: SciPy; 2010. doi: 10.25080/Majora-92bf1922-00a.

(39) McNamara J. xlsxwriter. 2013; Available at: https://xlsxwriter.readthedocs.io/.

(40) R Core Team. R: a language and environment for statistical computing. R package version 4.0.3. Vienna, Austria: R Foundation for Statistical Computing; 2020; Available at: https://www.R-project.org/.

(41) Csardi G, Nepusz T. The Igraph software package for complex network research. InterJournal, Complex Systems. 2006; 1695; Available at: http://igraph.org.

(42) Westcott SL, Schloss PD. OptiClust, an improved method for assigning amplicon-based sequence data to operational taxonomic units. mSphere. 2017 Mar;2(2):e00073-17. doi: 10.1128/mSphereDirect.00073-17.

(43) F Escapa I, Huang Y, Chen T, Lin M, Kokaras A, Dewhirst FE, et al. Construction of habitat-specific training sets to achieve species-level assignment in 16S rRNA gene datasets. Microbiome. 2020 May;8(1):65. doi: 10.1186/s40168-020-00841-w.

(44) Regueira-Iglesias A, Vázquez-González L, Balsa-Castro C, Blanco-Pintos T, Vila-Blanco N, Carreira MJ, et al. Impact of 16S rRNA gene redundancy and primer pair selection on the quantification and classification of oral microbiota in next-generation sequencing. Preprint at Research Square. 2021. doi: 10.21203/rs.3.rs-662236/v1.

(45) Caporaso JG, Lauber CL, Walters WA, Berg-Lyons D, Lozupone CA, Turnbaugh PJ, et al. Global patterns of 16S rRNA diversity at a depth of millions of sequences per sample. Proc Natl Acad Sci U S A. 2011 Mar;108 Suppl 1(Suppl 1):4516-22.

(46) Teles R, Teles F, Frias-Lopez J, Paster B, Haffajee A. Lessons learned and unlearned in periodontal microbiology. Periodontol 2000. 2013 Jun;62(1):95-162.

(47) Socransky SS, Haffajee AD, Cugini MA, Smith C, Kent RL, Jr. Microbial complexes in subgingival plaque. J Clin Periodontol. 1998 Feb;25(2):134-44.

(48) Na HS, Kim SY, Han H, Kim HJ, Lee JY, Lee JH, et al. Identification of potential oral microbial biomarkers for the diagnosis of periodontitis. J Clin Med. 2020 May;9(5):1549. doi: 10.3390/jcm9051549.

(49) Cho E, Park SN, Lim YK, Shin Y, Paek J, Hwang CH, et al. Fusobacterium hwasookii sp. nov., isolated from a human periodontitis lesion. Curr Microbiol. 2015 Feb;70(2):169-75.

(50) Wyss C, Moter A, Choi BK, Dewhirst FE, Xue Y, Schüpbach P, et al. Treponema putidum sp. nov., a medium-sized proteolytic spirochaete isolated from lesions of human periodontitis and acute necrotizing ulcerative gingivitis. Int J Syst Evol Microbiol. 2004 Jul;54(Pt 4):1117-22.

(51) Henne K, Fuchs F, Kruth S, Horz HP, Conrads G. Shifts in Campylobacter species abundance may reflect general microbial community shifts in periodontitis progression. J Oral Microbiol. 2014 Nov;6:25874. doi: 10.3402/jom.v6.25874.

(52) Zhang Y, Wang X, Li H, Ni C, Du Z, Yan F. Human oral microbiota and its modulation for oral health. Biomed Pharmacother. 2018 Mar;99:883-93.

(53) Jiang S, Gao X, Jin L, Lo EC. Salivary microbiome diversity in caries-free and caries-affected children. Int J Mol Sci. 2016 Nov;17(12):1978. doi: 10.3390/ijms17121978.

(54) Inquimbert C, Bourgeois D, Bravo M, Viennot S, Tramini P, Llodra JC, et al. The oral bacterial microbiome of interdental surfaces in adolescents according to carious risk. Microorganisms. 2019 Sep;7(9):319. doi: 10.3390/microorganisms7090319.

(55) Abranches J, Zeng L, Kajfasz JK, Palmer SR, Chakraborty B, Wen ZT, et al. Biology of oral streptococci. Microbiol Spectr. 2018 Oct;6(5):10.1128/microbiolspec.GPP3-0042-2018.

(56) Lemos JA, Palmer SR, Zeng L, Wen ZT, Kajfasz JK, Freires IA, et al. The biology of Streptococcus mutans. Microbiol Spectr. 2019 Jan;7(1):10.1128/microbiolspec.GPP3-0051-2018.

(57) Vartoukian SR, Moazzez RV, Paster BJ, Dewhirst FE, Wade WG. First cultivation of health-associated Tannerella sp. HOT-286 (BU063). J Dent Res. 2016 Oct;95(11):1308-13.

(58) Lenartova M, Tesinska B, Janatova T, Hrebicek O, Mysak J, Janata J, et al. The oral microbiome in periodontal health. Front Cell Infect Microbiol. 2021 Mar; 11:629723. doi: 10.3389/fcimb.2021.629723.

(59) Åberg CH, Kelk P, Johansson A. Aggregatibacter actinomycetemcomitans: virulence of its leukotoxin and association with aggressive periodontitis. Virulence. 2015;6(3):188-95.

(60) Deng ZL, Szafrański SP, Jarek M, Bhuju S, Wagner-Döbler I. Dysbiosis in chronic periodontitis: key microbial players and interactions with the human host. Sci Rep. 2017 Jun;7(1):3703. doi: 10.1038/s41598-017-03804-8.

(61) Edgar R. Taxonomy annotation and guide tree errors in 16S rRNA databases. PeerJ. 2018 Jun;6:e5030. doi: 10.7717/peerj.5030.

(62) Cole JR, Wang Q, Fish JA, Chai B, McGarrell DM, Sun Y, et al. Ribosomal Database Project: data and tools for high throughput rRNA analysis. Nucleic Acids Res. 2014 Jan;42(D1):D633-42. doi: 10.1093/nar/gkt1244.

(63) Quast C, Pruesse E, Yilmaz P, Gerken J, Schweer T, Yarza P, et al. 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.

(64) DeSantis TZ, Hugenholtz P, Larsen N, Rojas M, Brodie EL, Keller K, et al. Greengenes, a chimera-checked 16S rRNA gene database and workbench compatible with ARB. Appl Environ Microbiol. 2006 Jul;72(7):5069-72.

(65) Schatz MC, Delcher AL, Salzberg SL. Assembly of large genomes using second-generation sequencing. Genome Res. 2010 Sep;20(9):1165-73.

# OBJECTIVE 4

# Objective 4. A large-scale meta-omics analysis of plaque microbiota in periodontal diseases

## 4.1. ABSTRACT

**Aims:** To analyse the supragingival and subgingival plaque microbiota at ASV level of different periodontal conditions (periodontal health, gingivitis, and untreated and treated periodontitis) in terms of bacterial diversity, co-occurrence networks, and predictive models.

**Material and methods**: A total of 120 patients (55 controls, 65 periodontitis) were selected for subgingival plaque collection. Sequencing of the 3-4 16S rRNA gene region was performed in Illumina MiSeq. The obtained sequences and metadata were uploaded to the sequence read archive (SRA). Searches were performed in PubMed, Scopus, Embase, and the SRA to identify previously published Illumina 3-4 sequencing studies on the supragingival and subgingival plaque microbiome in distinct periodontal conditions. Research that met the criteria for sequences and metadata were included in the meta-omics analysis, comprising a total of 2045 samples. Sequences were processed under the same bioinformatics protocol, which included the ASV-level classification and the use of an oral-specific database for taxonomic classification. The statistical analysis was conducted using the phyloseq, DESeq2, microbiome, mixOmics, vegan, SpiecEasi, and igraph packages.

**Results and conclusions:** Bacterial richness associated with periodontitis was higher than in health in supragingival plaque and lower in subgingival, but evenness was higher in disease in both niches. The supragingival microbiota was richer and more diverse than the subgingival for the same periodontal condition. The structure of the bacterial community differed among conditions in the supra- and subgingival plaque, as well as for the same health status between the two niches. In addition, the core microbiota of dental plaque did not allow the characterisation of periodontal health and disease; and the proportion of the bacterial community organised in co-occurrence networks at the ASV level was very small. However, a small proportion of supra- and subgingival taxa had outstanding ability to distinguish between

periodontal conditions, and a relevant percentage of them were core members. Supragingival plaque was a better bacterial biomarker than subgingival for discriminating periodontal health from untreated and treated periodontitis. The main health-predictor ASVs in supragingival and subgingival plaque were: *R. dentocariosa* ASV2, *H. parainfluenzae* ASV3, ASV78, ASV45, and ASV46, *K. oralis* ASV66, *S. vestibularis* ASV27, and *A.* HMT170 ASV119. The main predictor ASVs of periodontitis in dental plaque were: *T. forsythia* ASV15, *F. alocis* ASV19, *T. denticola* ASV38 and ASV150, *F. fastidiosum* ASV97, *P.* HMT369 ASV124, *S. anginosus* ASV142, and *P. nodatum* ASV189.

**4.1.1. Keywords**

Meta-omics analysis; next-generation sequencing; 16S rRNA gene; dental plaque; supragingival; subgingival; microbiota; periodontal diseases.

**4.1.2. Declaration of conflict of interest**

The doctoral candidate and the rest of the authors of the present study declare that they have no conflict of interest concerning the objectives proposed in this chapter.

**4.1.3. Funding**

This investigation was supported by the Instituto de Salud Carlos III (General Division of Evaluation and Research Promotion, Madrid, Spain) and co-financed by the FEDER (European Regional Development Fund, ERDF) ("A way of making Europe") under grant ISCIII/PI21/00588; the Consellería de Cultura, Educación e Ordenación Universitaria de la Xunta de Galicia (group with growth potential ED431B 2020-2022 GPC2020/27; A. Regueira-Iglesias support ED481A-2017/233) and the ERDF, which acknowledges the CiTIUS-Research Center in Intelligent Technologies of the Santiago de Compostela University as a Research Center of the Galician University System.

The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

## 4.2. INTRODUCTION

Hundreds of articles have been published in the last two decades on the use of next-generation sequencing (NGS) of the 16S ribosomal RNA (rRNA) gene as a way to study the oral microbiome. These have generally analysed different intra-oral niches like dental plaque (1), tongue coatings (2), the soft tissues (3), and saliva (4) to determine the microbial diversity associated with distinct conditions, including: periodontal health (5); periodontal and peri-implant diseases (6); dental caries (7); and oral cancer (8). The large amount of such scientific production and the variation in the results obtained have caused researchers to conduct numerous narrative reviews in an attempt to achieve a consensus when defining the microbial profiles for distinct periodontal health statuses (9-12).

However, these published studies on the periodontal microbiome vary in terms of the relevant steps undertaken within a typical 16S rRNA gene sequencing workflow. This has had significant effects on the diversity of the results obtained, making comparisons very difficult (13-15). It is well known that each sequencing technology performs differently in the trade-off between read length, sequence throughput, and error rate (13); being Illumina that with preferable performance over Roche 454 or Ion Torrent (15). On the other hand, we have recently demonstrated through the *in silico* analysis performed in objective 1 (16) that, even among primer pairs with coverage values ≥90%, the oral species detected by primers targeting a particular region tended to be not covered by others amplifying a different zone and vice versa. Consequently, it can be said that it is rather questionable to compare sequences and consequently microbial diversity data derived from distinct sequencing technologies and gene regions.

On the other hand, more than 80% of recently published studies of the periodontal microbiome used the clustering of operational taxonomic units (OTUs) to perform their analyses. However, the 97% similarity threshold that is typically employed means that community descriptions based on this approach are wildly inaccurate, since 80% of oral-bacterial and archaeal species have an amplicon sequence similarity ≥97% to at least one other oral species as described in objective 3 (17). It is therefore necessary to conduct periodontal microbiota analyses using techniques that are currently considered to be more reliable, for example by examining levels of amplicon sequence variants (ASVs) (18-20). Furthermore,

high-quality, oral-specific databases are required if accurate classifications of these ASVs are to be achieved (21).

In an attempt to produce the strongest evidence to date on the periodontal microbiota, we conducted the large-scale meta-omics research described herein. This had the following objective: 1) to analyse the supragingival and subgingival plaque microbiota at ASV level of different periodontal conditions (healthy periodontal, gingivitis, periodontitis, and treated periodontitis) in terms of bacterial diversity, co-occurrence networks, and predictive models. To achieve our objective, we re-analysed sequences stored in public repositories from previously published Illumina 3-4 sequencing studies on the periodontal microbiome in supragingival and subgingival plaque. Our sample also included a bioproject with in-house sequences of the same region, which were taken from the subgingival plaque of periodontally healthy and periodontitis patients from our setting. The meta-omics analysis employed a unique bioinformatics protocol for high-quality filtering and sequence analysis.

## 4.3. MATERIAL AND METHODS

The complete analysis protocol applied in the present study is detailed in figure 1.

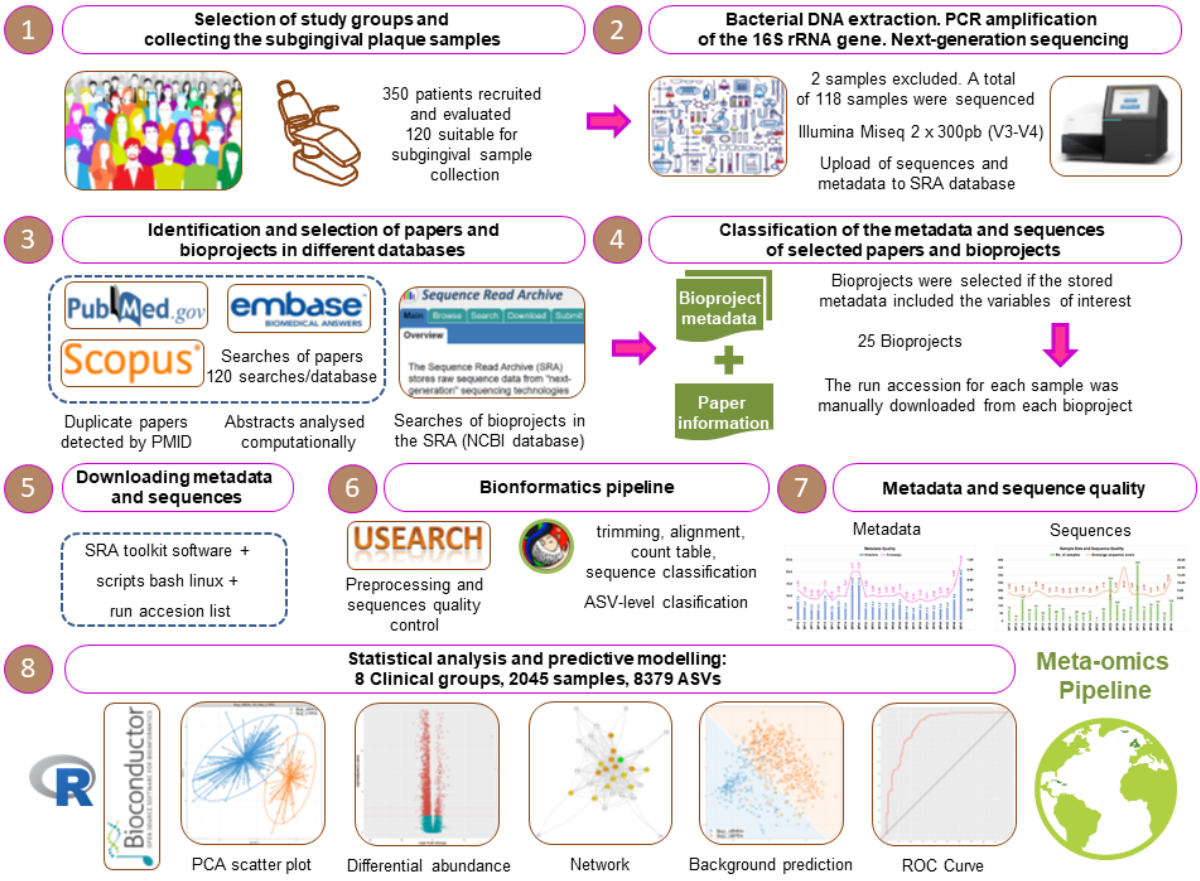



**Figure 1. The complete analysis protocol applied in the present meta-omics study.**

### 4.3.1. Selection of study groups and collecting the subgingival plaque samples

A convenience sample of 120 eligible participants, comprising 55 periodontally healthy controls (control group) and 65 subjects affected by untreated periodontitis (periodontitis group), were recruited from 350 consecutive patients in the general population who were referred to the School of Medicine and Dentistry (Universidade de Santiago de Compostela, Spain) and the Instituto Superior de Ciências da Saúde Norte, Cooperativa de Ensino Superior, Politécnico e Universitário (CESPU, Gandra, Paredes, Portugal) between 2018 and 2019 for an assessment of their oral health status.



Patients were recruited if they fulfilled the following inclusion criteria: 1) age 24 to 75; 2) the presence of at least 15 natural teeth; 3) no previous periodontal treatment; 4) no medical

history of diabetes mellitus, hepatic or renal disease, or other severe medical conditions or transmittable diseases; 5) no intake of systemic antimicrobials during the previous six months; 6) no intake of anti-inflammatory medication in the previous four months; 7) no routine use of oral antiseptics; 8) no history of alcohol or drug abuse; 9) no pregnancy or breastfeeding; 10) no presence of implants or orthodontic appliances; 11) have smoked for at least one year; and 12) have never smoked or stopped more than three years ago.

Two experienced dentists performed all the periodontal diagnoses. The bleeding on probing (BOP) and the bacterial plaque level (BPL) were recorded for the full mouth on a binary scale (presence/absence) at six sites per tooth. We also documented the probing pocket depth (PPD) and clinical attachment level (CAL) throughout the mouth, again at six sites per tooth, using a PCP-UNC 15 probe. Standardised radiographs of all the teeth were obtained to assess the alveolar bone status. The diagnosis of periodontitis was based on the clinical and radiographic information obtained. The control group included periodontally healthy patients who had: BOP ≤20%, no location with a PPD ≥4 mm, and no radiographic evidence of alveolar bone loss. The presence of periodontal health or moderate to severe generalised chronic periodontitis was established according to the clinical/radiographic information, applying previously published criteria (22,23).

The "smoking habit" of the participants was evaluated using a questionnaire, with information collected on its extent, i.e., non-smoker, former smoker, current smoker, time spent as a former or current smoker, and the number of cigarettes consumed per day.

The research was conducted following the principles of the Declaration of Helsinki (revised in 2000) on studies involving human experimentation (24), and its protocol was approved by the Galician Clinical Research Ethics Committee (registration number 2018/295) and the Instituto Superior de Ciências da Saúde-Norte, CESPU (registration number 35/CE-IUCS/2019) (Appendix S1). All the participants provided their written informed consent to their involvement in the study.



The plaque collection took place one or two weeks after the initial examination. Subgingival plaque samples from the controls and periodontal patients were collected and

pooled from eight non-adjacent proximal sites using two paper strips inserted into the gingival sulcus or periodontal pocket for 30 seconds. In the first case, samples were taken from subgingival healthy sites in quadrants one and three, and in the second case from sites with the most in-depth PPD in each quadrant. The strips used with each recruit were inserted into labelled tubes with 300 ml of 0.01M phosphate-buffered saline (PBS) (pH=7.2) and frozen at -80ºC until further genomic analysis.

### 4.3.2. 16S rRNA gene amplicon sequencing of subgingival samples

Total DNA was extracted from the subgingival plaque samples using a commercial kit (MasterPure Complete DNA and RNA Purification Kit; Epicentre,Wisconsin, USA) according to the manufacturer's instructions, albeit with minor modifications, including a mechanical disruption of bacteria (Pathogen Lysis Tube S; Qiagen, Hilden, Germany), and the addition of a lysozyme treatment (20 mg/ml at 37 ºC for 30 minutes). The isolated DNA was eluted in 50 µl of distilled and apyrogenic water, and its quality and concentration were assessed using a Nanodrop spectrophotometer (ND-2000 Spectrophotometer, Wilmington, USA). DNA samples with spectrophotometer ratios (Abs 260/280) between 1.5 and 2.0 were considered to be acceptable for inclusion in the study. Two subgingival samples from the control group were excluded due to non-compliance with this requirement.

A polymerase chain reaction (PCR) amplification of the 16S rRNA gene was performed with the KAPA HiFi HotStart ReadyMixPCR Kit (Cat. No. KK2602, 7958935001; Kapa Biosystems, F. Hoffmann-La Roche Ltd, Basel, Switzerland). The 3-4 hypervariable region was amplified as previously described (25) using the following primers in a limited-cycle PCR:

3-4-Forward (5′ -CCT ACG GGNGGC WGC AG-3).

3-4-Reverse (5′ -GAC TAC HVGGG TAT CTA ATC C-3).

A set of modified primers, 3-4-F and 3-4-R, were also used. This set contained a 1-3 base pair (bp) "heterogeneity spacer" that we designed to mitigate the issues caused by low-sequence diversity amplicons.

Each PCR amplification was carried out on a total volume of 10 µl, which comprised 4 µl of DNA, 0.2-Mm from each forward and reverse primer, and a Kapa ready mix (Kapa Biosystems). The PCR conditions were modified by conducting: 1) an initial denaturation at 95°C for 3 minutes; 2) 25 three-step cycles at 95°C for 30 seconds, 55°C for 30 seconds and 72°C for 30 seconds; and 3) a final 5-minute extension at 72ºC. Water, up to a total volume of 50 µl, was added after the first PCR step. The reactions were purified using AMPure XP beads (Beckman Coulter, Brea, CA, USA) with a 0.9X (3-4 amplicon) ratio, according to the manufacturer's instructions.

The PCR products were eluted from the magnetic beads with 32 µl of Buffer EB (Qiagen N.V, Hilden; Germany), with 30 µl of the eluate transferred to a fresh 96-well plate. The primers described above contain overhangs that enable the addition of full-length Nextera adapters. Barcodes are available for multiplex sequencing in a second PCR step, which produces sequencing-ready libraries. To this end, 5 µl of the first amplification was used as a template for the second PCR, with Nextera XT v2 adaptor primers added up to a final volume of 50 µl. The PCR mix and thermal profile employed for the first PCR were also used for the second, but only for eight cycles. After the second PCR, 25 µl of the final product was purified and normalised with the SequalPrep normalisation kit (Invitrogen, Carlsbad, CA, USA), according to the manufacturer's protocol. Libraries were eluted in a 20 µl volume and pooled for sequencing.

Final pools were quantified with a quantitative PCR (qPCR) using the Kapa library quantification kit for Illumina Platforms (Kapa Biosystems) on an ABI 7900HT real-time cycler (Applied Biosystems, Foster City, CA, USA). Sequencing using v3 chemistry with a loading concentration of 18 pM was performed in Illumina MiSeq (Illumina Inc., San Diego, CA, USA) with 2x300 bps reads. In all cases, 10% of the PhIX control libraries were spiked to increase the diversity of the sequenced samples.

In parallel, negative control tests of the sample-collection buffer, DNA-extraction and PCR-amplification steps were conducted routinely under the same conditions and using reagents. One such non-template control was subjected to the library preparation and then

sequenced. As expected, this yielded very few reads (1611 per sample). This was in contrast to an average of 249,747 reads/library in the sample-derived collections.

The bacterial mock community as a positive control for the downstream procedures were taken from the ZymoBIOMICS Microbial Community DNA Standard (Catalog Number D6306, Zymo Research, Irvine, CA, USA), which is a mix of genomic DNA isolated from pure cultures of eight bacterial and two fungal strains. Mock DNAs were amplified and sequenced in the same way as all the other samples used in the experiment.

The sequences obtained were deposited in the sequence read archive (SRA) database (26) under accession number PRJNA773202.

### 4.3.3. Characteristics of the studies for the meta-omics analysis: inclusion and exclusion criteria

Studies (cross-sectional, longitudinal, or interventional) on the microbial diversity in both supragingival and subgingival plaque in adult individuals with different periodontal conditions were included in our research (periodontal health, gingivitis, periodontitis, treated periodontitis, periimplantitis, and treated periimplantitis). We incorporated all the studies in which the diversity of the periodontal microbiome was assessed using primers from the 3-4 region and the Illumina-sequencing technology. An associated bioproject number indicates the repository in which the sequences are stored.

Studies were included in our analysis if the reference standard for diagnosing a periodontal condition was based on only clinical (PPD or CAL) or clinical and radiographic parameters (bone loss -BL-), irrespective of the diagnostic benchmarks applied. Consequently, in the absence of homogeneous criteria, any definition based on the author's reported standards was accepted. Studies without a reference for diagnosing the periodontal condition were ineligible for inclusion, as were those that failed to assess the periodontal status of patients using at least one clinical parameter (either the PPD or CAL).

### 4.3.4. Characteristics of the metadata table and the stored sample sequences: inclusion and exclusion criteria

After applying the criteria described above to the studies in the literature, we further selected those where the metadata of interest per sample was properly assigned in the repository.

In relation to the characteristics of the stored sequences, the inclusion and exclusion criteria were as follows: 1) direct and reverse sequences were accepted, whether or not the primer-pair sequence was included; 2) contigs with or without primer pairs, whose minimum average length had to be ≥350 bps; 3) the primer sequences must have been aligned with the complete *Escherichia coli* J01859.1 16S rRNA gene sequence using BLAST (27) to determine their initial and end positions; if they corresponded to the region of interest, the bioproject was included in the analysis; 4) study samples without primers were accepted for the analysis if multiple sequences were selected and aligned with the full 16S rRNA gene of *E. coli* J01859.1 to confirm that they belonged to the region of interest; 5) bioprojects in which most of the samples had a very low number of stored sequences (≤7000 sequences) were rejected; and 6) bioprojects were excluded if the samples were multiplexed or had different barcodes in each file.

### 4.3.5. Search methods for the identification and selection of studies and bioprojects

#### *4.3.5.1. Information sources and search strategy*

The searches were conducted in July 2021 using the electronic databases PubMed, Scopus, and Embase. The search strategy to identify Illumina sequencing-based studies of the periodontal microbiome encompassed two sets of terms relating to: 1) periodontal health conditions, oral niches and microbiota; and 2) the 16S rRNA gene sequencing technology (Appendix S2). All the searches in the three databases were filtered according to the publication year - 2000 to 2021 (inclusive). The searches of Scopus and Embase were also filtered by the type of: document/publication (Embase: article, article in the press; or review); source (journal); and language (English).



Additional searches of the SRA database (26) were performed using the terms “periodontitis”, “periodontal health”, “periodontal disease”, “peri-implantitis”, “gingivitis”,

and “gingival health” to ensure we had examined all of the potential bioprojects of interest, including those published as pre-prints.

*4.3.5.2. Search process and data mining*

The manipulation of the data identified in the searches described above was performed using the R software, version 4.1.2 (28). A total of 120 searches based on combinations of previously defined terms were performed in each database and the results were stored individually in a txt. (PubMed) or csv. (Scopus and Embase) file. Duplicates were detected via their PubMed unique identifiers (PMIDs) and removed.

The abstracts were analysed computationally using seven sets of positive terms (Appendix S3). Each word in the abstract that belonged to the “oral health”, “gene”, and/or “microbiome” categories was assigned 100 points; terms from the remaining groups were given one point each. Publications with both “oral health” and “microbiome” scores ≥200 and ≥300, respectively, were selected for subsequent manual evaluations of their abstracts and full-text. The automated data-mining process had previously been validated in a limited series of papers that met the inclusion criteria. The analysis of the positive words was carried out using the packages tm (version 0.7-8) and natural language processing (NLP, version 0.2-1) (29,30).

The identifiers of the chosen bioprojects were used to access the SRA database (26) and the SRA run selector (https://www.ncbi.nlm.nih.gov/Traces/study/). They were then downloaded and the metadata tables deposited there were assessed. The authors were contacted, if required, to obtain metadata or for clarification purposes. At this point, our own bioproject, PRJNA773202, was added to the total.

The manual selection of published studies and bioprojects was conducted by two independent reviewers (in the first case, ARI and IT; in the second, ARI and TBP). The reasons why studies and bioprojects were excluded were also recorded.



**4.3.6. Classification of the metadata and sequences of selected studies and bioprojects**

The information in the SRA database (26) was used to manually construct a metadata table for each of the bioprojects in the initial selection. This table contained information on the

following variables: 1) a bioproject's assigned code; 2) the run; 3) the bioproject; 4) the biosample; 5) the SRA study; 6) the PMID or digital object identifier (doi); 7) the 16S rRNA gene region targeted; 8) the periodontal diagnosis; 9) the health condition of the sampling site; 10) the periodontitis type and severity (if applicable); 11) the type of plaque sampled (supragingival or subgingival); 12) the therapy (yes/no); 13) the sampling moment; and 14) the patient's age, sex, ethnicity, systemic condition, and smoking habit. If the metadata did not contain sufficient information, it was supplemented with data from the studies and the authors who were available for correspondence.

Concerning the sequences stored for each bioproject, the identifiers list (run accession list) corresponding to the study samples of interest was downloaded in the txt. format. To download and store the sequences using the aforementioned accession lists, the free SRA Toolkit software (31) was installed in the local mode. A script was then developed in Bash (version 5.0.17) (32) in combination with the prefetch and fastq-dump commands from the SRA Toolkit (31). The plaque samples from each bioproject were stored in individual fastq files for subsequent manipulation.

Next, the bioprojects were classified according to the type of sequences stored:

1) Type 1: contig sequences not containing the primer pairs.
2) Type 2: contig sequences containing the primer pairs.
3) Type 3: paired-end 2x250 bps sequences containing the primer pairs.
4) Type 4: paired-end 2x300 bps sequences containing the primer pairs.
5) Type 5: paired-end 2x250 bps sequences not containing the primer pairs.
6) Type 6: paired-end 2x300 bps sequences not containing the primer pairs.

#### 4.3.7. Preprocessing and quality control of the sequences

The preprocessing and quality assessments of the sequences from each fastq file were performed using USEARCH (33). The sequences were aligned and assembled, with a maximum of five mismatches and a minimum similarity percentage of 90% deemed acceptable for the 2x250 bps, and 10 bps and 80% for the 2x300 bps.

A maximum of two mismatches in the sequence of each individual primer and a total of four in a pair were allowed. Finally, we discarded all the sequences with a maximum expected error >1 or with a minimum length <300 bps. The quality values were removed after preprocessing and quality control.

### 4.3.8. Identification of the processed sequences

We developed a script in R-Bioconductor (34) using Biostrings package (version 2.60.2) (35), with all the sequences including an identifying header with three clearly defined parts, as shown below:

(1) BPnn + (2) run_accession_id: (3) 15 digits, for example:
BP12ERR1735576:000037345541741.

The first part was the same for all the sequences within a bioproject and corresponded to the bioproject identifier, where "nn" are numbers between 0 and 9. The second part was the same for all the sequences within a fastq file (i.e., the sample) and corresponded to the run-accession identifier. Finally, each sequence within a file was given a unique sequence identifier consisting of 15 digits with values between 0 and 9.

### 4.3.9. Obtention of fasta, group, and name files for the analysis in mothur

All the fasta files from a particular bioproject were merged using Bash (32). These, and a script developed in R (28), enabled us to create a group file for each bioproject. Every such file included the unique sequence identifiers in the first column and the identifier of the sample to which the sequence belonged in the second; for example:

| Column 1 | Column 2 |
|---|---|
| BP12ERR1735576_000037345541741 | BP12ERR1735576 |

This allowed us to produce aperfect mapping between each sequence requiring analysis and its corresponding sample, and there was no case where a sequence identifier was repeated in its respective sample identifier.

The commands screen.seqs and unique.seqs from mothur (36) were executed to obtain the names file of each bioproject. These archives displayed all the sequence identifiers with the same sequences in the same row, and had as many rows as unique sequences identified in all the samples taken from a particular bioproject. The first sequence identifier of each row represents the next identifier in the analyses that followed; for example:

Seq1 Seq2 Seq3 Seq4 Seq5
Seq6 Seq7 Seq8
Seq9 Seq10
Seq11

Seq1 represents seqs 1 to 5, Seq6, seqs 6 to 8, and Seq11 only represents itself.

All the fasta files of all the bioprojects were then merged into a single file, along with the "groups" and "names" files to form the so-called meta-omics files.

#### 4.3.10. Low abundance filtering and mothur pipeline on meta-omics files

First, sequences in the overall samples with abundance values <500 counts were removed. This rare-abundance filtering produced almost 82 million sequences and 10577 ASVs in a total of 2186 samples.

The mothur pipeline (36) for ASVs was applied with slight modifications, including the application of the oral-specific database for the taxonomic classification of ASVs described by Escapa et al. (21). Sequences with a length >400 were permitted; those with more than eight homopolymers, those considered to be chimeras after implementing the mothur (36) VSEARCH algorithm (37), and those classified as unknown taxa at the highest hierarchical level (i.e., bacteria) were all removed. The sequences were not clustered to any level, as was also the case for the singletons and doubletons, since our objective was to identify and classify the highest possible number of sequences at the ASV level.

Once the mothur pipeline was completed, the following meta-omics files were exported to R-bioconductor (34) for further analysis: the count table, the taxonomic hierarchy at the ASV level, the phylogenetic tree, and the metadata table.

### 4.3.11. Assessment of the methodological quality of selected studies and bioprojects

#### *4.3.11.1. Quality of the metadata stored in repositories*

Using a checklist we designed for this purpose, the authors (ARI and TBP) independently assessed the quality of the metadata obtained from the bioprojects included in our research. This checklist covered 19 variables relating to the data available on the sampled subjects: 1) and 2) overall and sampling-site periodontal health; 3) periodontitis type and severity (if applicable), 4) sample type; 5) therapy; 6) age; 7) sex; 8) ethnicity; 9) systemic health condition; 10) smoking habit; and 11) the following periodontal parameters - number of teeth, BPL, total and site BOP, PPD, and CAL.

Each variable within a bioproject was given a value from 1 to 0: 1= the information was clearly specified in the metadata table downloaded from the repository; 0.6= the data was obtained from the published article (meaning of the code used in the sequence repository or general statements in the manuscript applicable to all participants); 0.3= the information was retrieved after contacting the authors; and 0= the data was not available.

Then, all the values from each bioproject were added together (maximum= 19), and this number was divided by the total number of variables applicable to it (i.e., “periodontitis type” did not apply to bioprojects containing only healthy samples). The final number obtained represented the quality of the bioproject’s metadata and was categorised as follows: low-quality score= 0.00 - 0.33; medium-quality score= 0.34 - 0.66; and high-quality score= 0.67 - 1.00.

#### *4.3.11.2. Sample size and number of sequences stored in repositories*

After applying the quality control of the bioinformatics protocol, the number of samples per bioproject and the average number of high-quality sequences per sample in each project were evaluated as quality parameters. The average number of sequences was divided by 10,000 (the number of sequences required to obtain proportional abundance to the niche being analysed); this parameter was given the name average sequence score (ASS). The ASS values

were interpreted as representing: 1) <0.25, very low-quantity sequences; 2) 0.25 - 0.75, - low-quantity sequences; 3) 0.75 - 1.0 - acceptable-quantity sequences; 4) 1.0 - 2.0 - high-quantity sequences; and 5) >2.0 - very high-quantity sequences.

#### 4.3.12. Statistical analysis with R-Bioconductor

The statistical analysis of the 16S rRNA sequencing data at the ASV level was performed according to the protocol proposed by McMurdie and Holmes (38), using implementations in R (28) that included the phyloseq, DESeq2 and microbiome packages (versions 1.36.0, 1.32.0 and 1.14.0, respectively) (39-41).

To eliminate samples with a low number of sequences, those with less than 2500 were excluded (n= 62), meaning that 2124 samples remained. Groups were created according to the type of dental plaque (supragingival, subgingival, or submucosal) and the periodontal health condition of the participants. A total of 11 groups were obtained (ordered alphabetically):

1) Supragingival plaque of periodontally healthy subjects - healthy sites (Sup_x0HHx; n= 210).
2) Supragingival plaque of the gingivitis subjects - diseased sites (Sup_x0GDx; n= 79).
3) Supragingival plaque of the periodontitis subjects - diseased sites (Sup_x0PDx; n= 493).
4) Supragingival plaque of the periodontitis subjects - diseased sites after periodontal therapy (Sup_x1PDx; n= 81).
5) Subgingival plaque of the periodontally healthy subjects - healthy sites (Sub_x0HHx; n= 155).
6) Subgingival plaque of the gingivitis subjects - diseased sites (Sub_x0GDx; n= 20).
7) Subgingival plaque of the periodontitis subjects - healthy sites (Sub_x0PHx; n= 62).
8) Subgingival plaque of the periodontitis subjects - diseased sites (Sub_x0PDx; n= 768).
9) Subgingival plaque of the periodontitis subjects - diseased sites after periodontal therapy (Sub_x1PDx; n= 197).
10) Submucosal plaque of the peri-implantitis subjects - healthy sites (Imp_x0IHx; n= 18).
11) Submucosal plaque of the peri-implantitis subjects - diseased sites (Imp_x0IDx; n= 41).

The groups Sub_x0GDx, Imp_x0IHx, and Imp_x0IDx were removed due to their low sample sizes (n= <50), leaving a total of 2045 samples to be analysed.

An independent filter had previously excluded from the statistical analysis the ASVs with an abundance of ≤10 counts and a presence in ≤2 samples (42), leaving a final total of 8379 ASVs.

The relationship between the different periodontal health conditions and the plaque microbiota was investigated from several perspectives: 1) the alpha diversity indicators and the structure of the bacterial community; 2) the composition of the core microbiota and the testing of differential abundance; 3) the co-occurrence network patterns; and 4) the predictive capacity of the plaque microbiota for discriminating the periodontal health condition. In general, where applicable, the comparative analyses were first performed between different clinical conditions within the same niche (supragingival plaque or subgingival plaque), and then in the same clinical condition between the two different plaques.

*4.3.12.1. Alpha diversity indicators and the structure of the bacterial community*

The phyloseq and microbiome packages were used to obtain the alpha diversity data (39,41). As indicators of taxa richness, we calculated the absolute count data ("observed") and the coverage index, which defines how many of the more abundant ASVs are required to achieve a particular proportion of the occupied ecosystem (95%). The Shannon and Pielou indices were determined as indicators of diversity and the evenness of the ASVs present in the samples (43,44). The Mann-Whitney U test (two-tailed) was used to conduct different comparative analyses.

A principal component analysis (PCA) was employed to visualise the clustering of the plaque samples in relation to their respective periodontal health condition. The mixOmics package (version 6.16.3) (45) was used to obtain the scatter plots of the first two principal components based on the relative abundance of the ASVs, showing the centroids of each clinical group and the ellipses representing the 95% confidence interval. A non-parametric permutational multivariate analysis of variance (PERMANOVA) (46) was used to measure the

multivariate community-level differences between the groups. These analyses were performed using the vegan package (version 2.5-7) (47).

*4.3.12.2. Composition of the core plaque microbiota and testing the differential abundance*

The microbiome package (41) was used to identify the core ASVs present at a prevalence rate of ≥75% in each plaque type and each periodontal clinical condition.

The DESeq2 package (40) was used to identify the ASVs with the most significant changes in differential abundance for the different periodontal conditions. Improvements to the stability and dispersion of the counts (variance) were required before it was possible to calculate the differential abundances. To this end, we used the estimate SizeFactors function in DESeq2 (40) to transform the stabilisation of the variance. The differential abundances were measured with the log2foldchange (log2FC) value and the different conditions were compared using the Wald test with the Benjamini–Hochberg correction (Q parameter= 0.1, false discovery rate (FDR) <10%). The differential-abundance measurements were statistically significant if the adjusted p-value was < 0.01 (−log10 adjusted p-value= 2).

*4.3.12.3. Co-occurrence networks in the plaque microbiota*

Co-occurrence network analyses were performed with the clinical groups with more than 100 samples, filtering out ASVs with an abundance of 0.01%. The SparCC method was used to generate the networks (48), as this allows researchers to detect with a high degree of accuracy the linear relationships in both a set of samples and a compositional dataset (49).

The default parameters and the SpiecEasi package (version 1.1.1) (50) were used to run SparCC, and the correlation matrix obtained was filtered using an absolute correlation score greater than or equal to 0.5. The networks were then visualised with the igraph package (version 1.2.6) (51), where each node represents an ASV and each edge represents the correlations between the abundances of the ASVs.



A set of measures was calculated to describe the topology of the resulting networks: 1) the network coverage, defined as the percentage of ASVs present in the co-occurrence network

concerning the total number of ASVs detected in the corresponding group; 2) the number of nodes and edges; 3) the number of sub-networks; and 4) the number of modules (52).

We calculated the betweenness centrality (BC) (53) to measure the relative importance of each ASV within the network (how influential a taxon is within a network). This determines the fraction of the shortest paths through one particular bacterial taxon to another. The BC of a taxon in a network reflects the importance of the control exerted by the taxon over the interactions of other taxa in the same network (53). In line with Banerjee et al. (54), a combined score based on a high degree value and a high BC value was used as a threshold to define the hub or keystone ASVs in the microbial communities.

*4.3.12.4. Predictive capacity of the plaque microbiota for discriminating the clinical condition.*

We conducted a supervised classification in the form of a sparse partial least-squares discriminant analysis (sPLS-DA) (55) to facilitate the categorisation of the different clinical groups and identify the ASVs that best distinguished two groups within each plaque niche (supragingival and subgingival plaques); consequently, the discriminant models were calculated two-to-two.

The sPLS-DA was performed using the mixOmics package (45), which is dedicated to the integrative examination of "omics" data. The ASVs with a relative abundance of less than 0.1% in the total samples were previously excluded from the development of predictive models. The number of components in each model was determined by applying the rule of thumb K-1, where K is the number of classes (in our case, two clinical groups). Consequently, all the predictive models were of one component.

Receiver operating characteristic (ROC) curves were constructed with the true positivity rate (sensitivity) as a function of the false positivity rate (1-specificity), while area under the curve (AUC) values were used to distinguish between each clinical group in the supragingival and subgingival plaque. It should be noted that simulations with an AUC value equal to or higher than 0.70 are generally considered to be acceptable predictive models (56).

## 4.4. RESULTS

### 4.4.1. Clinical characteristics of the study groups from our setting

Subjects in our own-recruited periodontitis group had a higher mean age and a higher number of smokers than those in the healthy group. Smokers with periodontitis consumed more cigarettes per day and had been smoking for more months. Regarding the clinical parameters associated with the periodontal status, patients in the periodontitis group had significantly higher BPL in the full-mouth, and BOP, PPD, and CAL values both in the full mouth and sampled sites records than subjects in the healthy group (p<0.001; Table 1).

**Table 1. Age, sex, smoking habit, and clinical characteristics associated with periodontal status in our own-recruited healthy and periodontitis groups.**

| Clinical parameters | Study groups | | |
|---|---|---|---|
| | Control (n= 53)* | Periodontitis (n= 65) | p value |
| **Age (years)** | 46.13 (12.11) | 52.31 (9.78) | 0.002 |
| **Sex** | | | |
| Female | 27 | 38 | NS |
| Male | 26 | 27 | |
| **Smoking habit** | | | |
| Non-smokers | 44 | 30 | <0.001 |
| Smokers | 9 | 35 | |
| Cigarettes/day (no.) | 1.02 (2.78) | 8.62 (9.89) | <0.001 |
| Months of smoking (no.) | 39.55 (105.66) | 175.75 (185.12) | <0.001 |
| **No. of teeth** | 26.98 (2.38) | 25.22 (4.01) | 0.012 |
| **Full mouth** | | | |
| BOP (%) | 10.91 (6.27) | 50.29 (20.43) | <0.001 |
| BPL (%) | 22.62 (17.48) | 55.02 (27.21) | <0.001 |
| PPD (mm) | 2.03 (0.26) | 3.61 (0.72) | <0.001 |
| CAL (mm) | 2.20 (0.40) | 4.40 (1.15) | <0.001 |
| **Sampled sites** | | | |
| BOP (%) | 5.92 (7.23) | 66.38 (24.88) | <0.001 |
| PPD (mm) | 2.20 (0.25) | 5.58 (0.76) | <0.001 |
| CAL (mm) | 2.29 (0.32) | 6.11 (1.05) | <0.001 |

*Of the 55 initial control subjects, two were excluded due to non-compliance with the requirements for the amount of DNA extracted. Values indicate means (standard deviations) and the number of subjects. After applying the Shapiro-Wilks test and verifying the non-normal distribution of almost all the clinical variables, the Mann-Whitney U test (two-tailed) was used to compare the quantitative clinical variables between the control and periodontitis groups. The Fisher's exact test (two-tailed) was used to assess the association of the qualitative variables between the two study groups. A significance level of $p<0.05$ was established.
BOP= bleeding on probing; BPL= bacterial plaque level; CAL= clinical attachment level; mm= milimetres; n= sample size; No.= number; NS= No significant; PPD= probing pocket depth.



### 4.4.2. Studies and bioprojects obtained in the search process

Figure 2 shows the flowchart of the search process, including the number of results obtained from each step.

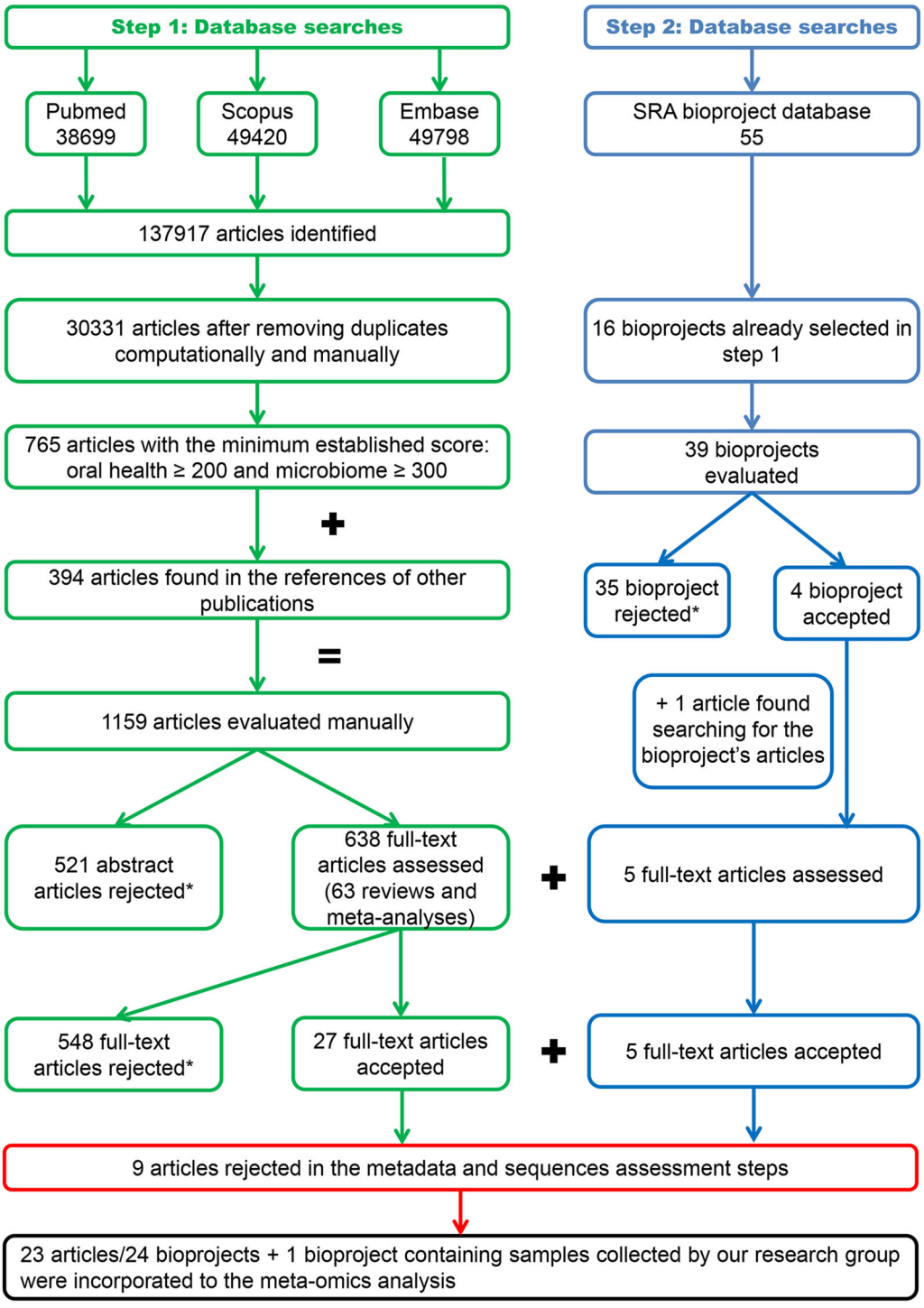



**Figure 2. Flowchart of the search process.**

*The exclusion reasons of the rejected articles and bioprojects are indicated in appendices S4, S5, and S6.

The abstracts of 30331 articles derived from the searches of the electronic databases were analysed computationally, using seven sets of positive terms to select the candidates for evaluation. A total of 1159 articles from these databases and 39 bioprojects from the SRA were

evaluated. These are listed in appendices S4 and S5, respectively; and the exclusion reason is indicated if applicable. Ultimately, 32 articles in which the sequence data had been deposited in 32 different bioprojects met the inclusion criteria (Appendix S6). The bioproject containing our sequences was added at this point.

Ten authors were contacted to obtain or clarify data and three provided the information required. Five articles were excluded in the metadata assessment step, and a further four after the samples and sequences were evaluated. We were ultimately left with 23 articles (6,57-78) and 25 bioprojects for inclusion in the meta-analysis, involving a total of 2045 samples distributed in eight periodontal clinical groups (four groups in supragingival plaque and the remaining four in subgingival plaque).

### 4.4.3. Quality assessment of metadata and sequences

#### *4.4.3.1. Quality of the metadata stored in repositories*

Three of the 25 included bioprojects had high-quality metadata (range= 0.95 - 0.77), four were medium quality (range= 0.45 - 0.34), and 18 low quality (range= 0.30 - 0.15). Overall, the those with medium and low-quality metadata did not include information about the periodontitis type and severity, ethnicity, or clinical periodontal parameters (total and sampling site). Moreover, most of the bioprojects with low-quality metadata did not provide information on the age or sex of the participants.

#### *4.4.3.2. Sample size and number of sequences stored in repositories*

From a sample size point of view, 10 bioprojects had <50 samples (40%), 9 between 50 and 100 (36%), and six >100 (24%). No bioprojects had an ASS of <0.25, as these had been eliminated in previous steps because of their very low-quantity sequences. Four bioprojects (BP20, BP21, BP22, and BP25) involving a total of 167 samples, representing 8.17% of all of those analysed, had ASS values from 0.75 - 1.0. These were therefore of an acceptable quantity, with more than 7500 sequences per sample. Eight bioprojects (BP12, BP14, BP18, BP19, BP23, BP27, BP44, and BP45) and 422 samples, representing 20.63% of all of those processed, had ASS values from 1.0 - 2.0. These were thus deemed to be high quantity, with 10,000 - 20,000 sequences per sample. Finally, 13 bioprojects, with an overall total of 1456 samples,

representing 71.20% of all of those processed, had an ASS >2.0, making them very high quantity, with more than 20,000 sequences per sample.

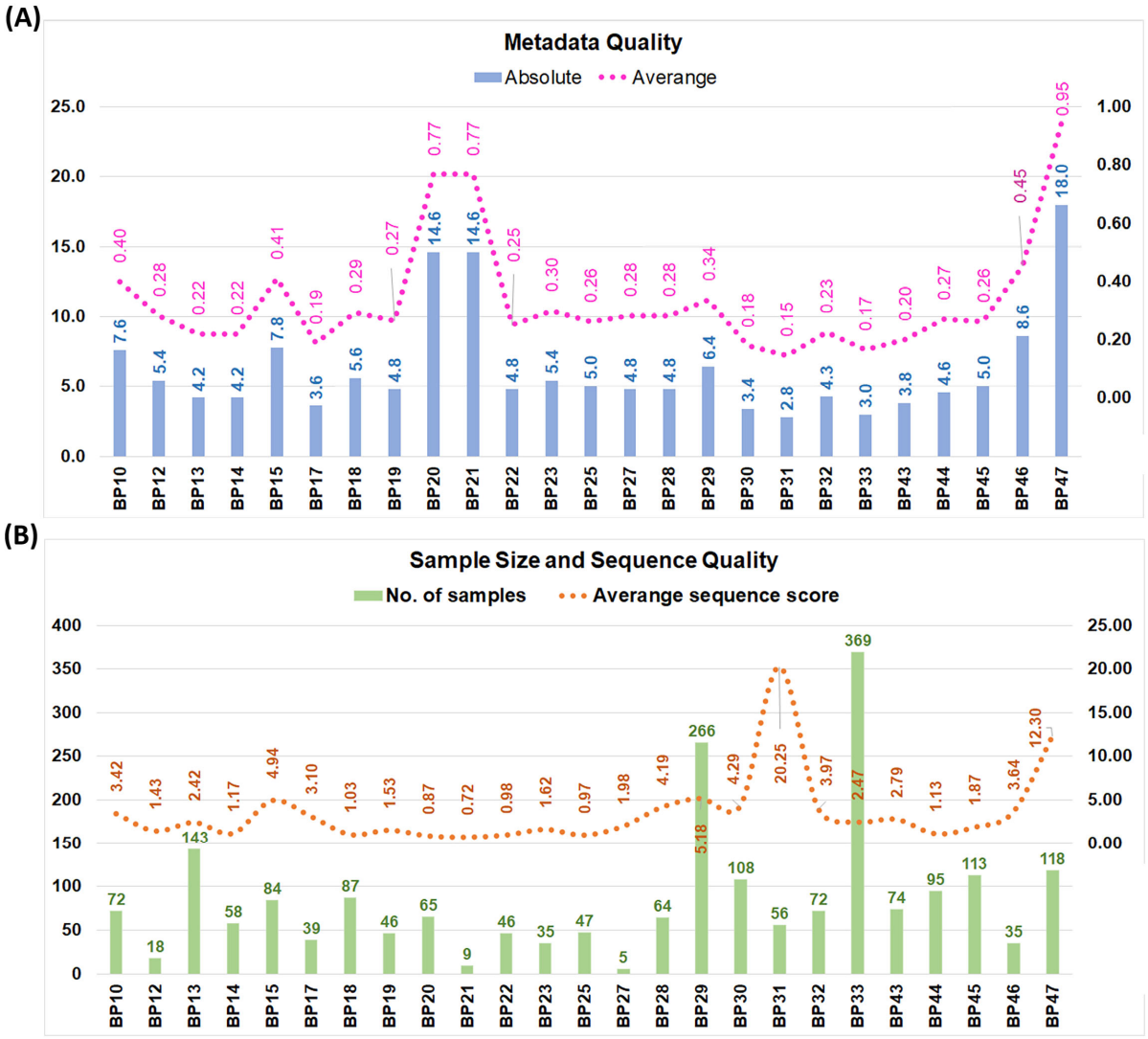



**Figure 3. The methodological quality of selected studies and bioprojects: (A) metadata; and (B) sample size and sequence quantity.**

Twenty-three articles and 24 bioprojects were included since one article referred to two distinct bioprojects (+ 1 bioproject associated with our samples).

### 4.4.4. Characteristics of the selected studies and bioprojects

Appendix S7 contains a quantitative summary of the main descriptive characteristics of the sequencing-based studies of the periodontal microbiome that formed part of our meta-omics research. Less than a third (7/23 articles + 1 own unpublished bioproject; 29.17%) were able to establish the periodontal diagnosis with the new Classification of Periodontal and Peri-implant

Diseases and Conditions (79), with most using earlier classifications or the authors' own criteria (17/24; 70.83%). In 13/24 investigations (54.17%), there was a comparison of the microbial profiles in relation to states of periodontal health and disease, while 10/24 (41.66%) only evaluated periodontitis. There was also one article (4.17%) where only five healthy samples were selected, as these were the only ones that could be assigned to a specific health condition. Subgingival plaque was used the most to study the periodontal microbiota (16/24; 66.66%), followed by supragingival plaque (4/24; 16.67%) or both types (4/24; 16.67%). Moreover, 4/24 studies (16.67%) assessed the changes produced in the microbiota after non-surgical periodontal therapy, including (in some cases) the adjuvant effect of antibiotics or toothpastes.

### 4.4.5. Alpha-diversity in supragingival and subgingival plaque microbiota

#### *4.4.5.1. Supragingival plaque microbiota*

As shown in table 2, supragingival plaque richness decreased significantly from periodontal health to gingivitis and then increased strongly in the periodontitis condition (median number of ASVs observed= 610.50, 474.00, and 892.00, respectively; 95% coverage index= 220.00, 130.00, and 288.00, respectively). There was a significant decrease in both the number of ASVs and the 95% coverage index in the post-periodontal therapy samples compared to those collected before treatment (Sup_x1PDx *vs.* Sup_x0PDx: 781.00 and 263.00 *vs.* 892.00 and 288.00). However, these post-treatment estimates of richness did not reach the levels of the healthy group, with significant differences remaining between the two clinical conditions (Sup_x1PDx *vs.* Sup_x0HHx, 781.00 and 263.00 *vs.* 610.50 and 220.00).

Conversely, the diversity and evenness indexes showed a continuous upwards trend from health to disease and even continued to improve after treatment (Shannon index range between 4.75 and 4.07; Pielou index range between 0.70 and 0.62). All the two-by-two group comparisons were significantly different, excepting Sup_x0HHx and Sup_x0GDx (Shannon index) and Sup_x0GDx and Sup_x0PDx (Pielou index).

**Table 2. Alpha diversity indicators in the different periodontal health conditions and dental plaque types.**

| Groups (n) | No. Observed ASVs | Coverage index (95%) | Shannon index | Pielou index |
|---|---|---|---|---|
| | Median values and IQR | | | |
| Sup_x0HHx (210) | 610.50 (458.00) | 220.00 (145.00) | 4.07 (0.75) | 0.62 (0.13) |
| Sup_x0GDx (79) | 474.00 (162.50) | 130.00 (89.50) | 4.19 (0.65) | 0.68 (0.11) |
| Sup_x0PDx (493) | 892.00 (912.00) | 288.00 (165.00) | 4.51 (0.65) | 0.67 (0.12) |
| Sup_x1PDx (81) | 781.00 (333.00) | 263.00 (114.00) | 4.75 (0.57) | 0.70 (0.08) |
| Sub_x0HHx (155) | 478.00 (1142.50) | 171.00 (169.00) | 4.15 (1.18) | 0.65 (0.12) |
| Sub_x0PHx (62) | 474.00 (320.50) | 120.00 (93.75) | 4.05 (0.94) | 0.65 (0.13) |
| Sub_x0PDx (768) | 417.50 (455.25) | 142.00 (130.00) | 4.17 (0.98) | 0.68 (0.10) |
| Sub_x1PDx (197) | 507.00 (461.00) | 129.00 (105.00) | 4.20 (0.83) | 0.69 (0.10) |
| **Comparison of distinct periodontal health conditions in the supragingival plaque (p-value)** | | | | |
| Sup_x0HHx *vs.* Sup_x0GDx | $1.0227E^{-06}$ | $8.0438E^{-12}$ | NS | 0.0002 |
| Sup_x0HHx *vs.* Sup_x0PDx | $7.0505E^{-14}$ | $7.7671E^{-13}$ | $8.5206E^{-24}$ | $2.7619E^{-08}$ |
| Sup_x0HHx *vs.* Sup_x1PDx | 0.0010 | 0.0008 | $9.9932E^{-20}$ | $2.9953E^{-16}$ |
| Sup_x0GDx *vs.* Sup_x0PDx | $9.6812E^{-21}$ | $1.2208E^{-29}$ | $4.3632E^{-08}$ | NS |
| Sup_x0GDx *vs.* Sup_x1PDx | $5.4703E^{-15}$ | $1.7601E^{-17}$ | $4.5964E^{-11}$ | 0.0003 |
| Sup_x0PDx *vs.* Sup_x1PDx | 0.0030 | 0.0466 | 0.0003 | $1.2601E^{-07}$ |
| **Comparison of distinct periodontal health conditions in the subgingival plaque (p-value)** | | | | |
| Sub_x0HHx *vs.* Sub_x0PHx | NS | 0.0129 | NS | NS |
| Sub_x0HHx *vs.* Sub_x0PDx | 0.0005 | 0.0158 | NS | 0.0001 |
| Sub_x0HHx *vs.* Sub_x1PDx | 0.0139 | 0.0384 | NS | $2.4036E^{-05}$ |
| Sub_x0PHx *vs.* Sub_x0PDx | NS | NS | NS | 0.0017 |
| Sub_x0PHx *vs.* Sub_x1PDx | NS | NS | 0.0475 | 0.0003 |
| Sub_x0PDx *vs.* Sub_x1PDx | NS | NS | NS | NS |
| **Comparison of the same periodontal health condition between the supragingival and subgingival plaques (p-value)** | | | | |
| Sup_x0HHx *vs.* Sub_x0HHx | 0.0363 | $9.2241E^{-06}$ | NS | 0.0050 |
| Sup_x0PDx *vs.* Sub_x0PDx | $8.0165E^{-72}$ | $5.6815E^{-97}$ | $2.2485E^{-23}$ | $8.4811E^{-05}$ |
| Sup_x1PDx *vs.* Sub_x1PDx | $6.2408E^{-13}$ | $3.8383E^{-20}$ | $1.8787E^{-12}$ | 0.0194 |

A significance level of $p<0.05$ was established.
ASVs= amplicon sequence variants; IQR= interquartile range; n= sample size; No.= number; NS= No significant; Sub_x0HHx= subgingival plaque of periodontally healthy subjects, healthy sites; Sub_x0PHx= subgingival plaque of periodontitis subjects, healthy sites; Sub_x0PDx= subgingival plaque of periodontitis subjects, diseased sites; Sub_x1PDx= subgingival plaque of periodontitis subjects, diseased sites after therapy; Sup_x0GDx= supragingival plaque of gingivitis subjects, diseased sites; Sup_x0HHx= supragingival plaque of periodontally healthy subjects, healthy sites; Sup_x0PDx= supragingival plaque of periodontitis subjects, diseased sites; Sup_x1PDx= supragingival plaque of periodontitis subjects, diseased sites after therapy.

#### *4.4.5.2. Subgingival plaque microbiota*

Regarding the subgingival microbiota, significantly lower values were detected in the number of ASVs and the 95% coverage index in diseased sites from periodontitis patients with respect to healthy patients (417.50 and 142.00 *vs.* 478.00 and 171.00, respectively). After periodontal treatment, there was a significant increase in bacterial richness, surpassing even the healthy levels (507.00 compared to 478.00), although the 95% coverage index remained lower than the healthy levels (129.00 compared to 171.00). The diversity and evenness indices tended to increase in the groups of diseased tooth sites, although only the Pielou index comparisons

were significant (0.65 in Sub_x0HHx *vs.* 0.68 and 0.69 in Sub_x0PDx and Sub_x1PDx, respectively).

The number of observed ASVs and the diversity and evenness indices of the subgingival microbiota did not vary significantly between Sub_x0HH and Sub_x0PHx. Similarly, the alpha diversity indicators did not show significant variations between the different periodontal groups, except for the Pielou index in the comparisons Sub_x0PHx *vs.* Sub_x0PDx (0.65 *vs.* 0.68) and Sub_x0PHx *vs.* Sub_x1PDx (0.65 *vs.* 0.69), and the Shannon in Sub_x0PHx *vs.* Sub_x1PDx (4.05 *vs.* 4.20) (Table 2).

*4.4.5.3. Supragingival and subgingival plaque*

When contrasting the "supra" and "sub" plaques of the subjects with the same periodontal health status, we observed that the number of ASVs, the 95% coverage index values, and the Shannon diversity scores were significantly higher in the supragingival niche than in the subgingival niche ("supra" *vs.* "sub": ASV number range= 892.00 - 610.50 *vs.* 507.00 - 417.50; 95% coverage index range= 288.00 - 220.00 *vs.* 171.00 - 129.00; Shannon diversity range= 4.75 - 4.07 *vs.* 4.20 - 4.15); the exception was represented by the Shannon index of the healthy groups for both plaques. Conversely, the evenness values were significantly higher in the subgingival environment, except for the case of Sup_x1PDx (Table 2).

**4.4.6. Structure of the bacterial community in supragingival and subgingival plaque microbiota**

The PCAs revealed a grouping of the supragingival and subgingival samples according to the periodontal health condition of the subject and the sampled site (the latter in the case of the Sub_x0PHx group) (Figures 4 and 5). The visual observations were confirmed by the PERMANOVA, which produced significant results for all the two-by-two group comparisons (Table 3).

In the comparison between the different niches of the same periodontal health condition, the PCA revealed a clustering of the samples according to the type of plaque collected from the subject for the same periodontal health status (Figure 6). The visual observations were

confirmed by the PERMANOVA, which yielded significant results for all the two-by-two group comparisons (Table 3).

**Table 3. PERMANOVA test values in the comparison between the different periodontal health conditions and dental plaque types.**

| Groups | p-value PERMANOVA |
|---|---|
| **Comparison of distinct periodontal health conditions in the supragingival plaque** | |
| Sup_x0HHx *vs.* Sup_x0GDx | 0.0001 |
| Sup_x0HHx *vs.* Sup_x0PDx | 0.0001 |
| Sup_x0HHx *vs.* Sup_x1PDx | 0.0001 |
| Sup_x0GDx *vs.* Sup_x0PDx | 0.0001 |
| Sup_x0GDx *vs.* Sup_x1PDx | 0.0001 |
| Sup_x0PDx *vs.* Sup_x1PDx | 0.0001 |
| **Comparison of distinct periodontal health conditions in the subgingival plaque** | |
| Sub_x0HHx *vs.* Sub_x0PHx | 0.0001 |
| Sub_x0HHx *vs.* Sub_x0PDx | 0.0001 |
| Sub_x0HHx *vs.* Sub_x1PDx | 0.0001 |
| Sub_x0PHx *vs.* Sub_x0PDx | 0.0001 |
| Sub_x0PHx *vs.* Sub_x1PDx | 0.0001 |
| Sub_x0PDx *vs.* Sub_x1PDx | 0.0001 |
| **Comparison of the same periodontal health condition between the supragingival and subgingival plaques** | |
| Sup_x0HHx *vs.* Sub_x0HHx | 0.0001 |
| Sup_x0PDx *vs.* Sub_x0PDx | 0.0001 |
| Sup_x1PDx *vs.* Sub_x1PDx | 0.0001 |

A significance level of $p<0.05$ was established.
PERMANOVA= permutational multivariate analysis of variance; Sub_x0HHx= subgingival plaque of periodontally healthy subjects, healthy sites; Sub_x0PHx= subgingival plaque of periodontitis subjects, healthy sites; Sub_x0PDx= subgingival plaque of periodontitis subjects, diseased sites; Sub_x1PDx= subgingival plaque of periodontitis subjects, diseased sites after therapy; Sup_x0GDx= supragingival plaque of gingivitis subjects, diseased sites; Sup_x0HHx= supragingival plaque of periodontally healthy subjects, healthy sites; Sup_x0PDx= supragingival plaque of periodontitis subjects, diseased sites; Sup_x1PDx= supragingival plaque of periodontitis subjects, diseased sites after therapy.

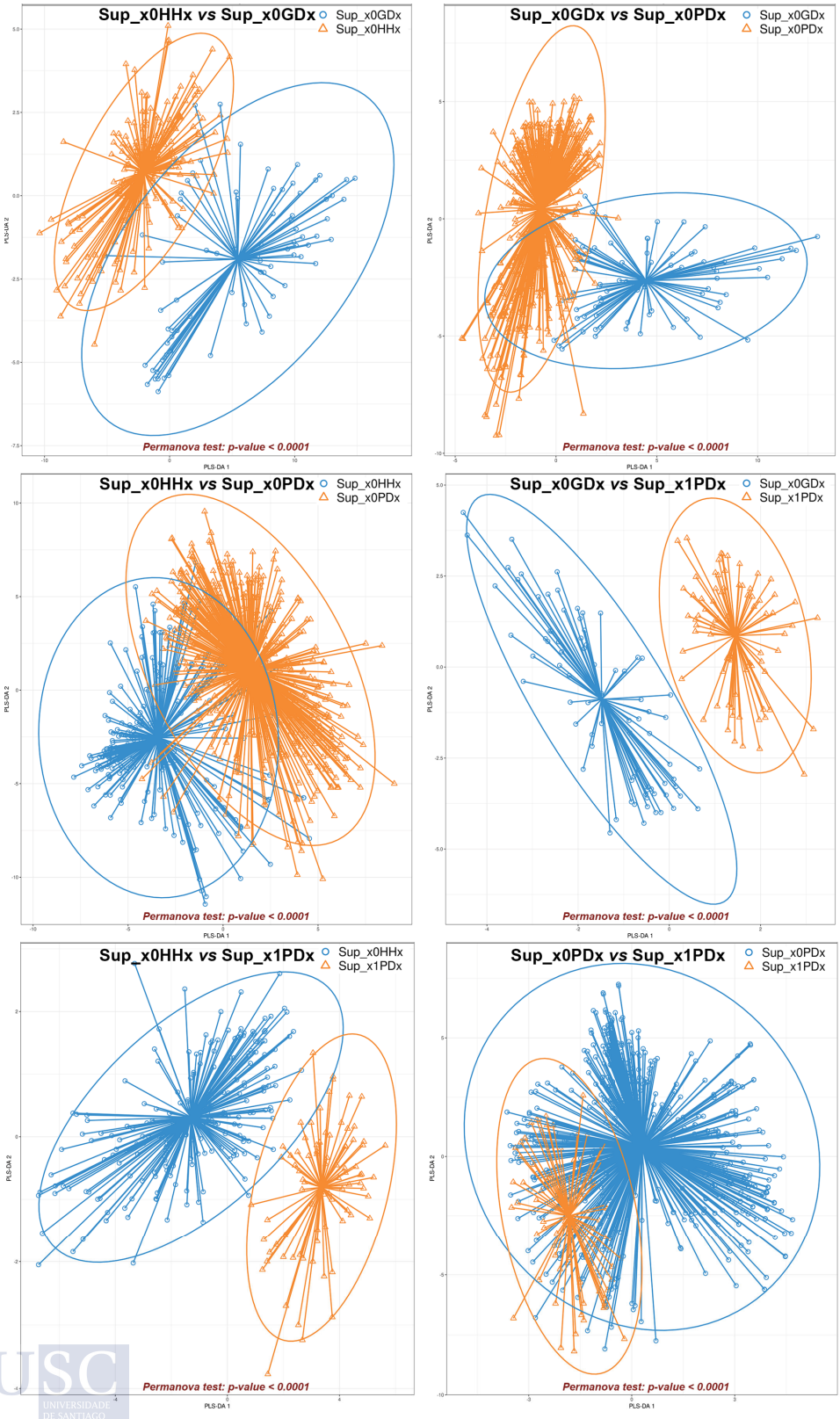





**Figure 4. Principal component analysis (PCA) of the different periodontal health conditions in supragingival plaque.**

Sup_x0GDx= supragingival plaque of gingivitis subjects, diseased sites; Sup_x0HHx= supragingival plaque of periodontally healthy subjects, healthy sites; Sup_x0PDx= supragingival plaque of periodontitis subjects, diseased sites; Sup_x1PDx= supragingival plaque of periodontitis subjects, diseased sites after therapy.

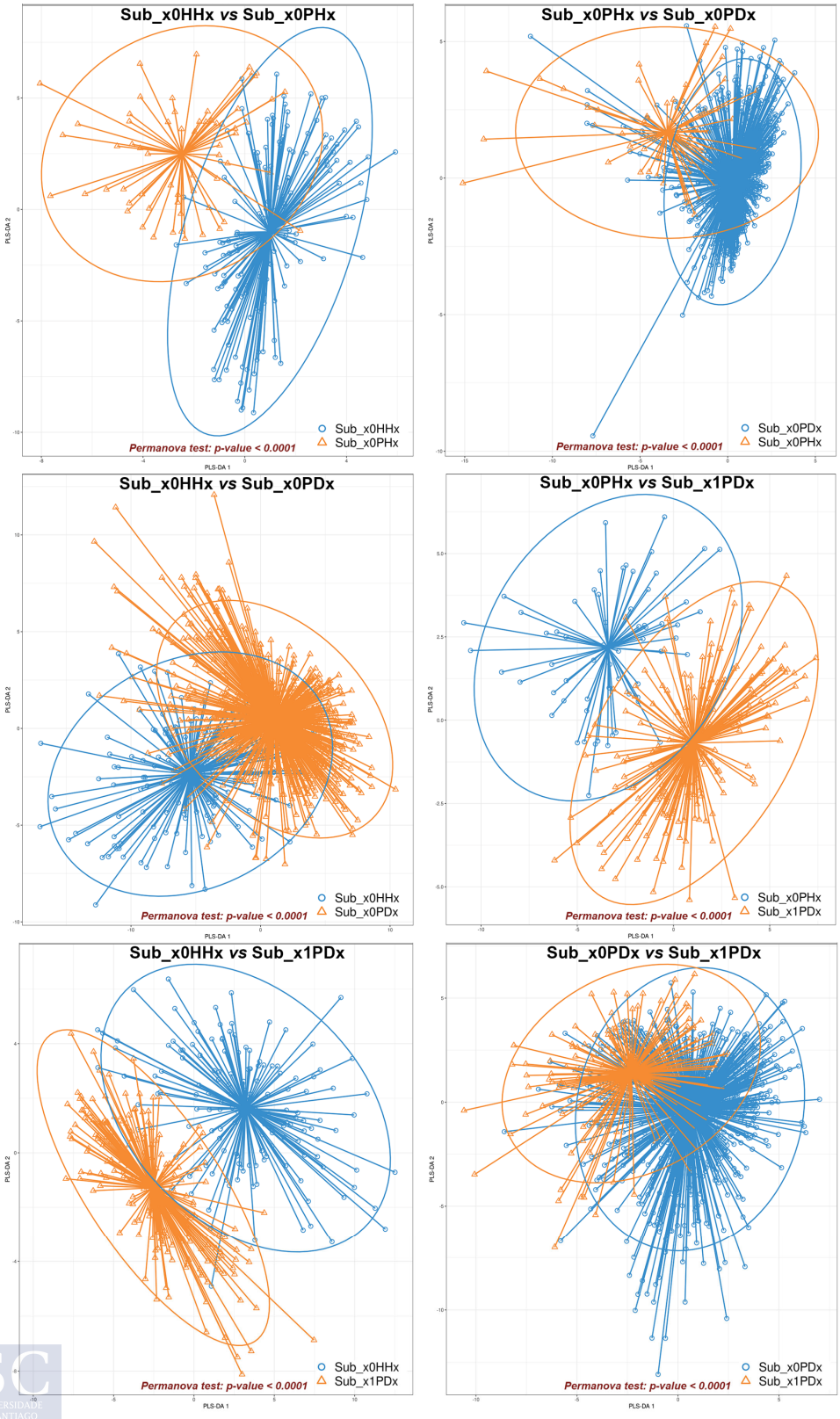



**Figure 5. Principal component analysis (PCA) of the different periodontal health conditions in subagingival plaque.**

Sub_x0HHx= subgingival plaque of periodontally healthy subjects, healthy sites; Sub_x0PDx= subgingival plaque of periodontitis subjects, diseased sites; Sub_x0PHx= subgingival plaque of periodontitis subjects, healthy sites; Sub_x1PDx= subgingival plaque of periodontitis subjects, diseased sites after therapy.

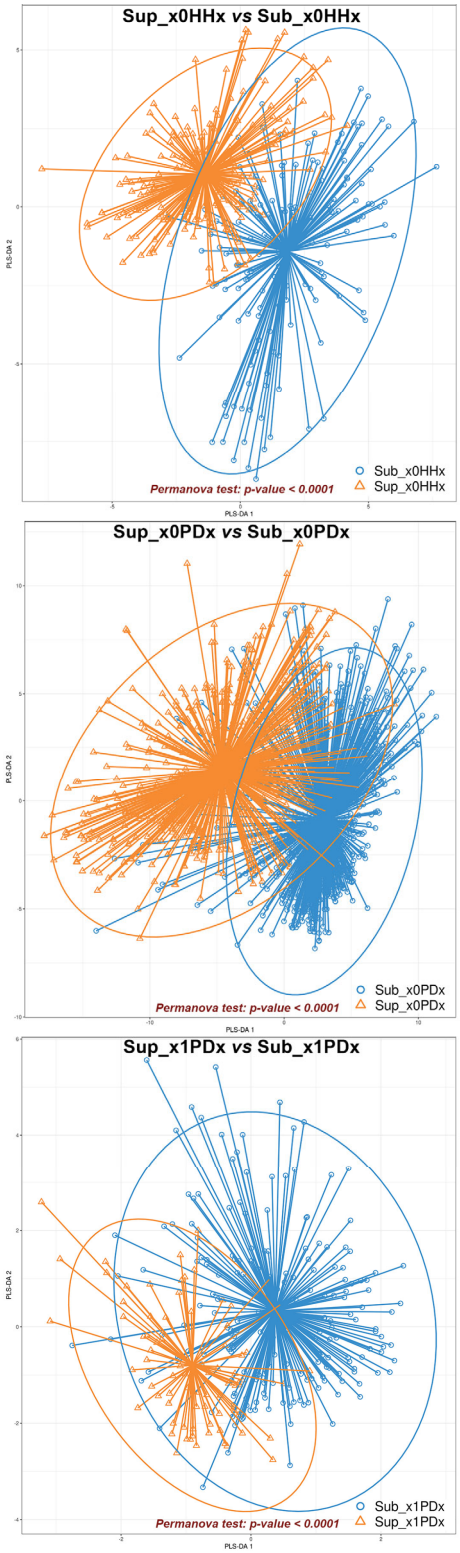



**Figure 6. Principal component analysis (PCA) of the same periodontal health conditions in supragingival and subgingival plaque.**

Sub_x0HHx= subgingival plaque of periodontally healthy subjects, healthy sites; Sup_x0HHx= supragingival plaque of periodontally healthy subjects, healthy sites; Sub_x0PDx= subgingival plaque of periodontitis subjects, diseased sites; Sup_x0PDx= supragingival plaque of periodontitis subjects, diseased sites; Sub_x1PDx= subgingival plaque of periodontitis subjects, diseased sites after therapy; Sup_x1PDx= supragingival plaque of periodontitis subjects, diseased sites after therapy.

### 4.4.7. Core microbiota in supragingival and subgingival plaque microbiota

Table 4 portrays the number and percentage of core ASVs and core species constituting the supragingival, subgingival, and both plaque types in the different periodontal health conditions, as well as the relative abundances of the ASVs. Appendices S8 and S9 contain the list of ASVs in the dental plaque core microbiota and their relative abundances in the different periodontal health groups.

**Table 4. Number and percentage of taxa of the core microbiota in the different periodontal health conditions and dental plaque types, and the relative abundance values they represented.**

| Group (n) | No. Core ASVs (% detected) | No. Core species (% detected) | Abundance of core ASVs total (range) |
|---|---|---|---|
| **Core microbiota in the supragingival plaque** | | | |
| Sup_x0HHx (210) | 46 (0.77%) | 28 (5.76%) | 48.08% (11.28 - 0.03%) |
| Sup_x0GDx (79) | 41 (1.57%) | 27 (6.78%) | 28.34% (3.55 - 0.02%) |
| Sup_x0PDx (493) | 51 (0.69%) | 33 (6.29%) | 39.58% (9.39 - 0.01%) |
| Sup_x1PDx (81) | 76 (1.85%) | 46 (10.02%) | 38.14% (4.57 - 0.02 %) |
| Common all groups | 37 (0.48%) | 28 (5.31%) | 36.44% (9.02 - 0.02%) |
| **Core microbiota in the subgingival plaque** | | | |
| Sub_x0HHx (155) | 44 (0.74%) | 29 (5.57%) | 30.84% (7.28 - 0.02%) |
| Sub_x0PHx (62) | 31 (0.85%) | 18 (3.78%) | 32.30% (7.87 - 0.02%) |
| Sub_x0PDx (768) | 29 (0.34%) | 22 (3.94%) | 23.50% (3.63 - 0.14%) |
| Sub_x1PDx (197) | 27 (0.53%) | 22 (4.46%) | 25.49% (4.05 - 0.11%) |
| Common all groups | 24 (0.28%) | 18 (3.27%) | 23.70% (4.57 - 0.15%) |
| **Core microbiota in the supragingival and subgingival plaques** | | | |
| Common all groups | 26 (0.29%) | 20 (3.62%) | 27.50% (6.58 – 0.12%) |

The percentages of detected core ASVs and species are calculated with respect to the total number of ASVs and species detected by the correspondent group(s). The taxa that could not be classified at the species level ("unclassified") were counted once so the number of species detected is the minimum that could be obtained.
ASVs= amplicon sequence variants; n= sample size; No.= number; Sub_x0HHx= subgingival plaque of periodontally healthy subjects, healthy sites; Sub_x0PHx= subgingival plaque of periodontitis subjects, healthy sites; Sub_x0PDx= subgingival plaque of periodontitis subjects, diseased sites; Sub_x1PDx= subgingival plaque of periodontitis subjects, diseased sites after therapy; Sup_x0GDx= supragingival plaque of gingivitis subjects, diseased sites; Sup_x0HHx= supragingival plaque of periodontally healthy subjects, healthy sites; Sup_x0PDx= supragingival plaque of periodontitis subjects, diseased sites; Sup_x1PDx= supragingival plaque of periodontitis subjects, diseased sites after therapy.

#### *4.4.7.1. Supragingival plaque microbiota*

A total of 37 ASVs from 28 different species (0.48% and 5.31% of the detected taxa, respectively) constituted the supragingival core microbiota, regardless of periodontal condition, and represented 36.44% of the total abundance.



Considering the different clinical groups in the supragingival plaque, the number of core ASVs ranged between 51 - 41 (1.57% - 0.69% of the detected taxa) and the number of core species, between 33 - 27 (6.78% - 5.76%), with the Sup_x1PDx group showing the highest

values of core ASVs and core species (76 and 46, respectively; 1.85% and 10.02%, respectively); with respect to the relative abundance of core ASVs, this varied between 28.34% - 39.58%, with the Sup_x0HHx group exhibiting the highest abundance percentage (48.08%) (Table 4).

*4.4.7.2. Subgingival plaque microbiota*

The subgingival core microbiota groups consisted of 24 ASVs from 18 different species (0.28% and 3.27% of the detected taxa, respectively) and accounted for 23.70% of the total abundance.

Considering the different clinical groups, the number of ASVs and core species ranged from 44 - 27 (0.85% - 0.34%) to 29 - 18 (5.57% - 3.78%), and their respective relative abundance, between 32.30% - 23.50% (Table 4).

*4.4.7.3. Supragingival and subgingival plaque microbiota*

When both supragingival and subgingival plaque were considered together and regardless of periodontal health status, the core microbiota consisted of 26 ASVs of 20 different species (0.29% and 3.62% of the detected taxa, respectively), representing 27.50% of the total abundance.

**4.4.8. Differential abundance in supragingival and subgingival plaque microbiota**

Table 5 shows the results for the differential abundances (p-value <0.01) for the distinct dental plaque and periodontal health groups, while appendix S10 illustrates the main taxa (p-value <0.001 and ≥0.5% relative abundance) associated with each group in the two-by-two comparisons. Appendix S11 contains the relative abundance and log2FC values of the main taxa associated with each group.

*4.4.8.1. Supragingival plaque microbiota*

The results for the supragingival plaque revealed that the highest numbers of ASVs and species with differential abundances were obtained when two-by-two comparisons were conducted for the periodontal health, gingivitis, and periodontitis groups. Accordingly, Sup_x0GDx *vs.* Sup_x0PDx produced a total of 1290 ASVs and 243 species with differential

abundances (16.96% and 45.42% of the total detected by the two groups, respectively), while Sup_x0HHx *vs.* Sup_x0GDx, 945 ASVs and 198 species (15.09% and 39.52%, respectively). The comparison of Sup_x0HHx and Sup_x1PDx revealed 926 ASVs and 210 species both differentially abundant (13.12% and 41.02% of the detected taxa, respectively), and, again, Sup_x0HHx *vs.* Sup_x0PDx, a total of 918 (12.17%) and 272 (51.52%) ASVs and species, respectively. In contrast, the lowest relative numbers of ASVs and species with differential abundances were observed in the analysis of Sup_x0PDx *vs.* Sup_x1PDx (total= 660 ASVs, 8.95%; 145 species, 27.62%).

The percentages of core ASVs and core species showing differential abundance ranged from 6.31% - 2.87% and 14.65% - 7.35%, respectively (Table 5).

*4.4.8.2. Subgingival plaque microbiota*

The results for the subgingival plaque demonstrated that the highest relative numbers of ASVs and species with differential abundances were obtained when comparing the periodontal health group to both the non-treated and treated periodontitis groups. Accordingly, the comparison of Sub_x0HHx *vs.* Sub_x0PDx revealed a total of 1074 ASVs (12.64% of the total detected by the two groups) from 273 species with differential abundances (48.75%), while Sub_x0HHx *vs.* Sub_x1PDx had 1015 ASVs (14.45%) from 225 species (41.67%). Conversely, the lowest relative numbers of ASVs and species with differential abundances were observed in the analysis of Sub_x0PHx *vs.* both Sub_x0PDx (total= 364 ASVs; 4.28%; 156 species; 27.81%) and Sub_x1PDx (total= 339 ASVs; 5.64%; 117 species; 22.20%).

The percentages of core ASVs and core species showing differential abundance ranged from 6.59% - 2.06% and 10.26% - 5.13%, respectively (Table 5).

**Table 5. Number of total and core taxa that presented differential abundances in the different periodontal health conditions and dental plaque types, and the relative abundance values they represented.**

| | No. ASVs (% detected) | No. Species (% detected) | No. Core ASVs (% detected)* | No. Core species (% detected)* |
|---|---|---|---|---|
| **Differential abundances of distinct periodontal health conditions in the supragingival plaque** | | | | |
| Sup_x0HHx *vs.* Sup_x0GDx | 945 (15.09%) | 198 (39.52%) | 41 (4.34%) | 29 (14.65%) |
| Sup_x0HHx *vs.* Sup_x0PDx | 918 (12.17%) | 272 (51.52%) | 33 (3.59%) | 20 (7.35%) |
| Sup_x0HHx *vs.* Sup_x1PDx | 926 (13.12%) | 210 (41.02%) | 33 (3.56%) | 21 (10.00%) |
| Sup_x0GDx *vs.* Sup_x0PDx | 1290 (16.96%) | 243 (45.42%) | 37 (2.87%) | 25 (10.29%) |
| Sup_x0GDx *vs.* Sup_x1PDx | 507 (10.08%) | 163 (33.61%) | 32 (6.31%) | 16 (9.82%) |
| Sup_x0PDx *vs.* Sup_x1PDx | 660 (8.95%) | 145 (27.62%) | 19 (2.88%) | 14 (9.66%) |
| **Differential abundances of distinct periodontal health conditions in the subgingival plaque** | | | | |
| Sub_x0HHx *vs.* Sub_x0PHx | 425 (6.62%) | 160 (29.52%) | 13 (3.06%) | 12 (7.50%) |
| Sub_x0HHx *vs.* Sub_x0PDx | 1074 (12.64%) | 273 (48.75%) | 36 (3.35%) | 25 (9.16%) |
| Sub_x0HHx *vs.* Sub_x1PDx | 1015 (14.45%) | 225 (41.67%) | 27 (2.66%) | 18 (8.00%) |
| Sub_x0PHx *vs.* Sub_x0PDx | 364 (4.28%) | 156 (27.81%) | 24 (6.59%) | 16 (10.26%) |
| Sub_x0PHx *vs.* Sub_x1PDx | 339 (5.64%) | 117 (22.20%) | 7 (2.06%) | 6 (5.13%) |
| Sub_x0PDx *vs.* Sub_x1PDx | 604 (7.14%) | 189 (33.87%) | 14 (2.32%) | 12 (6.35%) |
| **Differential abundances of the same periodontal health condition between supragingival and subgingival plaques** | | | | |
| Sup_x0HHx *vs.* Sub_x0HHx | 802 (10.57%) | 255 (46.88%) | 36 (4.49%) | 25 (9.80%) |
| Sup_x0PDx *vs.* Sub_x0PDx | 2367 (27.34%) | 349 (62.21%) | 48 (2.03%) | 33 (9.46%) |
| Sup_x1PDx *vs.* Sub_x1PDx | 198 (3.85%) | 72 (14.55%) | 14 (7.07%) | 4 (5.56%) |

The percentages of detected ASVs and species are calculated concerning the total number of different ASVs and species detected by at least one of the groups to be compared. *The percentages of core ASVs and species are calculated concerning the total number of different ASVs and species that showed differential abundances in the two groups compared. The taxa that could not be classified at the species level ("unclassified") were counted once so the number of species detected is the minimum that could be obtained.
ASVs= amplicon sequence variants; No.= number; Sub_x0HHx= subgingival plaque of periodontally healthy subjects, healthy sites; Sub_x0PHx= subgingival plaque of periodontitis subjects, healthy sites; Sub_x0PDx= subgingival plaque of periodontitis subjects, diseased sites; Sub_x1PDx= subgingival plaque of periodontitis subjects, diseased sites after therapy; Sup_x0GDx= supragingival plaque of gingivitis subjects, diseased sites; Sup_x0HHx= supragingival plaque of periodontally healthy subjects, healthy sites; Sup_x0PDx= supragingival plaque of periodontitis subjects, diseased sites; Sup_x1PDx= supragingival plaque of periodontitis subjects, diseased sites after therapy.

*4.4.8.3. Supragingival and subgingival plaque microbiota*

As shown in table 5, the comparison of Sup_x0PDx *vs.* Sub_x0PDx revealed the highest relative numbers of ASVs and species with differential abundances (total= 2367 ASVs, 27.34% of the total detected by the two groups; 349 species, 62.21%). Conversely, the lowest relative estimates were observed in the analysis of Sup_x1PDx *vs.* Sub_x1PDx (total= 198 ASVs, 3.85%; 72 species, 14.55%).

The percentages of core ASVs and core species showing differential abundance ranged from 7.07% - 2.03% and 9.80% - 5.56%, respectively (Table 5).

### 4.4.9. Co-occurrence networks in supragingival and subgingival plaque microbiota

Table 6 shows the topological parameters of the co-occurrence networks in the two supragingival and subgingival plaque groups that met the inclusion criteria for this analysis.

**Table 6. Topological parameters of the co-occurrence networks in the different periodontal health conditions and dental plaque types.**

| | Supragingival plaque | | Subgingival plaque | | |
|---|---|---|---|---|---|
| | Sup_x0HHx | Sup_x0PDx | Sub_x0HHx | Sub_x0PDx | Sub_x1PDx |
| Network coverage* | 2.26% | 2.54% | 2.75% | 0.63% | 1.54% |
| Number of nodes | 136 | 187 | 163 | 53 | 78 |
| Number of edges | 290 | 959 | 387 | 80 | 111 |
| Number of positive correlations (%) | 290 (100.0%) | 958 (99.9%) | 387 (100.0%) | 80 (100.0%) | 111 (100.0%) |
| Number of negative correlations (%) | 0 (0.0%) | 1 (0.1%) | 0 (0.0%) | 0 (0.0%) | 0 (0.0%) |
| Ratio of positive correlations and nodes | 2.13% | 5.12% | 2.37% | 1.51% | 1.42% |
| Number of subnetworks | 18 | 12 | 12 | 10 | 12 |
| Number of modules | 25 | 55 | 25 | 11 | 13 |
| Number of modules with more than 3 nodes | 12 | 12 | 11 | 4 | 6 |

*Percentage of ASVs present in the co-occurrence network with respect to the total number of ASVs detected in the correspondent group.
Sub_x0HHx= subgingival plaque of periodontally healthy subjects, healthy sites; Sub_x0PDx= subgingival plaque of periodontitis subjects, diseased sites; Sub_x1PDx= subgingival plaque of periodontitis subjects, diseased sites after therapy; Sup_x0HHx= supragingival plaque of periodontally healthy subjects, healthy sites; Sup_x0PDx= supragingival plaque of periodontitis subjects, diseased sites.

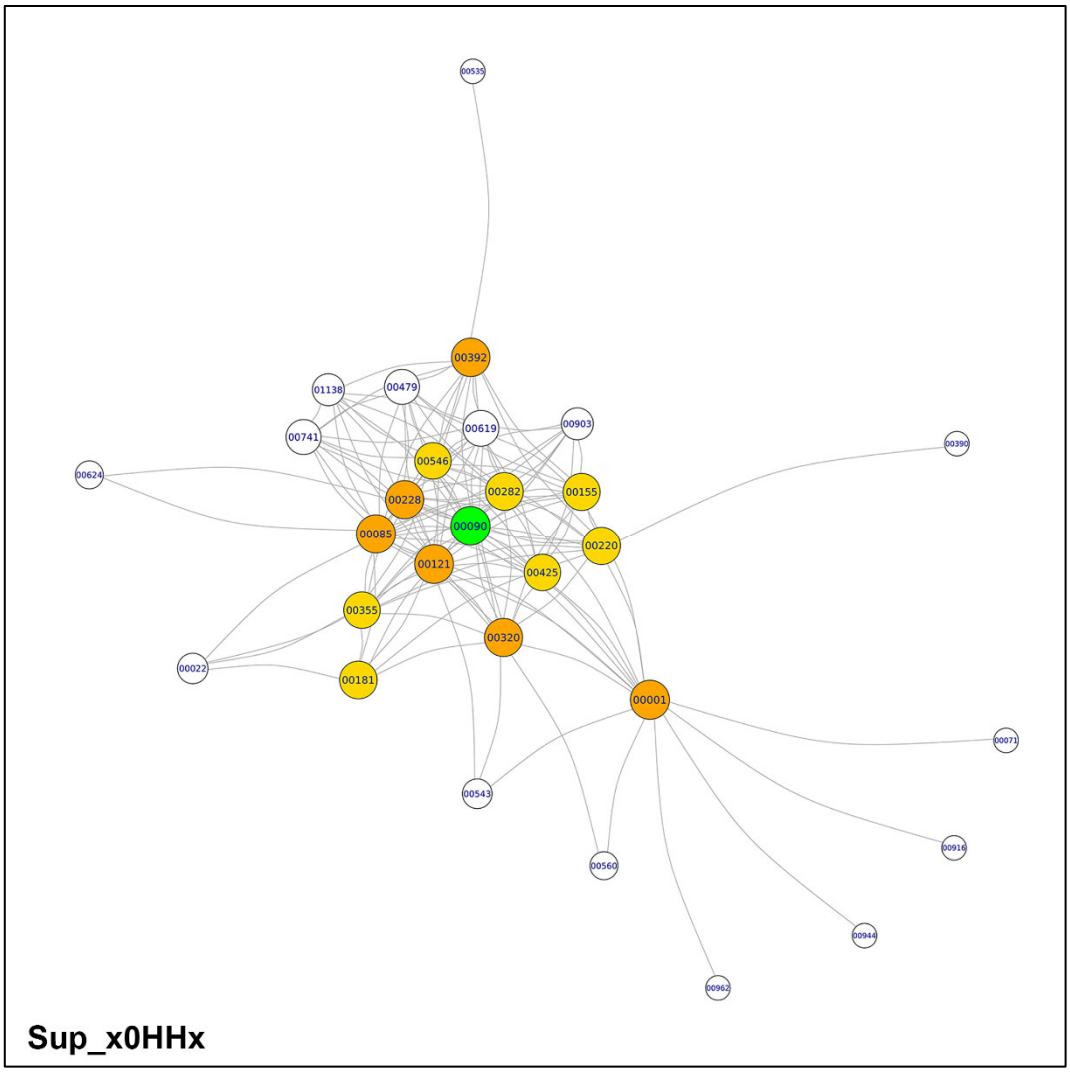
#### *4.4.9.1. Supragingival plaque microbiota*

The network coverage and the number of nodes in Sup_x0PDx were slightly higher than in Sup_x0HHx (2.54% and 187 *vs.* 2.26% and 136, respectively). Moreover, the number of edges was more than three times greater in the diseased than in the healthy group (959 *vs.* 290, respectively). Practically all these correlations were positive in both groups, except for that of *Dialister invisus* ASV68 and *Streptococcus unclassified* ASV4 in Sup_x0PDx (correlation value= -0.51). The diseased network had fewer subnetworks and a higher number of modules (12 and 55 *vs.* 18 and 25 in Sup_x0HHx), but both groups had the same number of modules with more than three nodes (Table 6).

In the Sup_x0HHx network, the three main hubs or keystone ASVs were: *Streptococcus unclassified* ASV90, *Rothia dentocariosa* ASV2, and *Streptococcus oralis* subsp. *dentisani* clade 058 ASV1. All were part of the Sup_x0HHx core microbiota but, despite having an abundance ≥0.5%, none of the three taxa were differentially abundant when compared to Sup_x0PDx.

The principal keystone ASVs in the Sup_x0PDx network were: *Streptococcus* unclassified ASV85, *Streptococcus sanguinis* ASV228, and *Streptococcus* unclassified ASV121. Only the former had an abundance ≥0.5% in Sup_x0PDx, but none were core members or had differential abundance if compared to Sup_x0HHx.

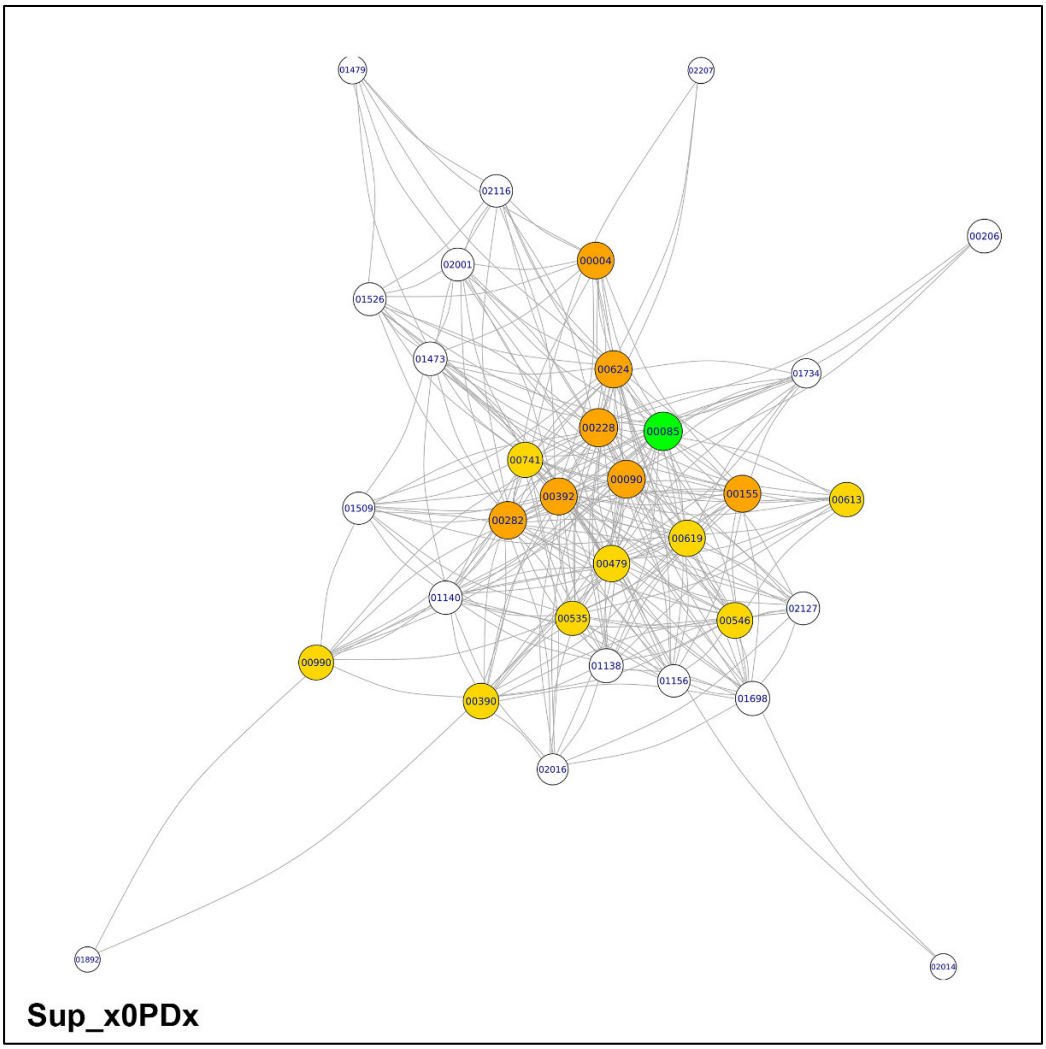
Figures 7 and 8 represent the main modules of the co-occurrence networks associated with Sup_x0HHx and Sup_x0PDx, respectively.




| Sup_x0HHx Samples= 210 | | | | | |
|---|---|---|---|---|---|
| ASVid | Genus | Species | ASV | Core | Relative abundance |
| AV00001 | Streptococcus | oralis_subsp.dentisani _clade_058 | BTASV016027 | Y | 11.2800 |
| AV00085 | Streptococcus | Unclassified | unclassified | Y | 0.6454 |
| AV00090 | Streptococcus | Unclassified | unclassified | Y | 0.5700 |
| AV00121 | Streptococcus | Unclassified | unclassified | Y | 0.3356 |
| AV00155 | Streptococcus | Unclassified | unclassified | Y | 0.2059 |
| AV00181 | Streptococcus | Unclassified | unclassified | N | 0.1094 |
| AV00220 | Streptococcus | Unclassified | unclassified | N | 0.1252 |
| AV00228 | Streptococcus | Sanguinis | unclassified | Y | 0.2314 |
| AV00282 | Streptococcus | Unclassified | unclassified | Y | 0.1702 |
| AV00320 | Streptococcus | oralis_subsp.dentisani _clade_058 | unclassified | Y | 0.1281 |
| AV00355 | Streptococcus | Unclassified | unclassified | N | 0.0609 |
| AV00392 | Streptococcus | Sanguinis | unclassified | Y | 0.1019 |
| AV00425 | Streptococcus | oralis_subsp.dentisani _clade_058 | unclassified | N | 0.0682 |
| AV00546 | Streptococcus | Unclassified | unclassified | N | 0.0861 |

**Figure 7. Main module of the co-occurrence network associated with the supragingival plaque of periodontally healthy subjects.**

In the graph, the most important taxa are highlighted in green, orange, and yellow according to the score obtained in the analysis group.The highest value is shown in green, the values belonging to the first quartile in orange, and those belonging to the second quartile, i.e. up to the median, in yellow.
ASV= amplicon sequence variant; ASVid= amplicon sequence variant identifier; N= no; subsp.= subspecies; Sup_x0HHx= supragingival plaque of periodontally healthy subjects, healthy sites; Y= yes.




| Sup_x0PDx Samples= 493 | | | | | |
|---|---|---|---|---|---|
| ASVid | Genus | Species | ASV | Core | Relative abundance |
| AV00004 | Streptococcus | unclassified | unclassified | Y | 4.0960 |
| AV00085 | Streptococcus | unclassified | unclassified | N | 0.5505 |
| AV00090 | Streptococcus | unclassified | unclassified | N | 0.5082 |
| AV00155 | Streptococcus | unclassified | unclassified | N | 0.1948 |
| AV00228 | Streptococcus | sanguinis | unclassified | N | 0.1995 |
| AV00390 | Streptococcus | unclassified | unclassified | N | 0.0686 |
| AV00392 | Streptococcus | sanguinis | unclassified | N | 0.1129 |
| AV00479 | Streptococcus | unclassified | unclassified | N | 0.0901 |
| AV00535 | Streptococcus | unclassified | unclassified | N | 0.0837 |
| AV00546 | Streptococcus | unclassified | unclassified | N | 0.0796 |
| AV00613 | Streptococcus | unclassified | unclassified | N | 0.0663 |
| AV00619 | Streptococcus | sanguinis | unclassified | N | 0.0534 |
| AV00624 | Streptococcus | unclassified | unclassified | N | 0.0492 |
| AV00741 | Streptococcus | unclassified | unclassified | N | 0.0558 |
| AV00990 | Streptococcus | unclassified | unclassified | N | 0.0325 |

**Figure 8. Main module of the co-occurrence network associated with the supragingival plaque of periodontitis subjects in the diseased sites.**

In the graph, the most important taxa are highlighted in green, orange, and yellow according to the score obtained in the analysis group.The highest value is shown in green, the values belonging to the first quartile in orange, and those belonging to the second quartile, i.e. up to the median, in yellow.
ASV= amplicon sequence variant; ASVid= amplicon sequence variant identifier; N= no; subsp.= subspecies; Sup_x0PDx= supragingival plaque of periodontitis subjects, diseased sites; Y= yes.

*4.4.9.2. Subgingival plaque microbiota*

On the contrary to the supragingival plaque, the subgingival network's coverage and numbers of nodes and edges decreased with worsening health. In this sense, the Sub_x0PDx network showed the lowest coverage and number of nodes and edges followed by Sub_x1PDx network with respect to Sub_x0HHx network (network coverage= 0.63% and 1.54% *vs.* 2.75%; number of nodes= 53 and 78 *vs.* 163; number of edges= 80 and 111 *vs.* 387). All the correlations in this niche were positive. Although similar numbers of subnetworks were observed in the three clinical groups (12 and 10), the Sub_x0PDx and Sub_x1PDx networks presented lower numbers of modules and modules with more than three nodes than the periodontal health group (11 and 13 *vs.* 25, respectively; 4 and 6 *vs.* 11, respectively).

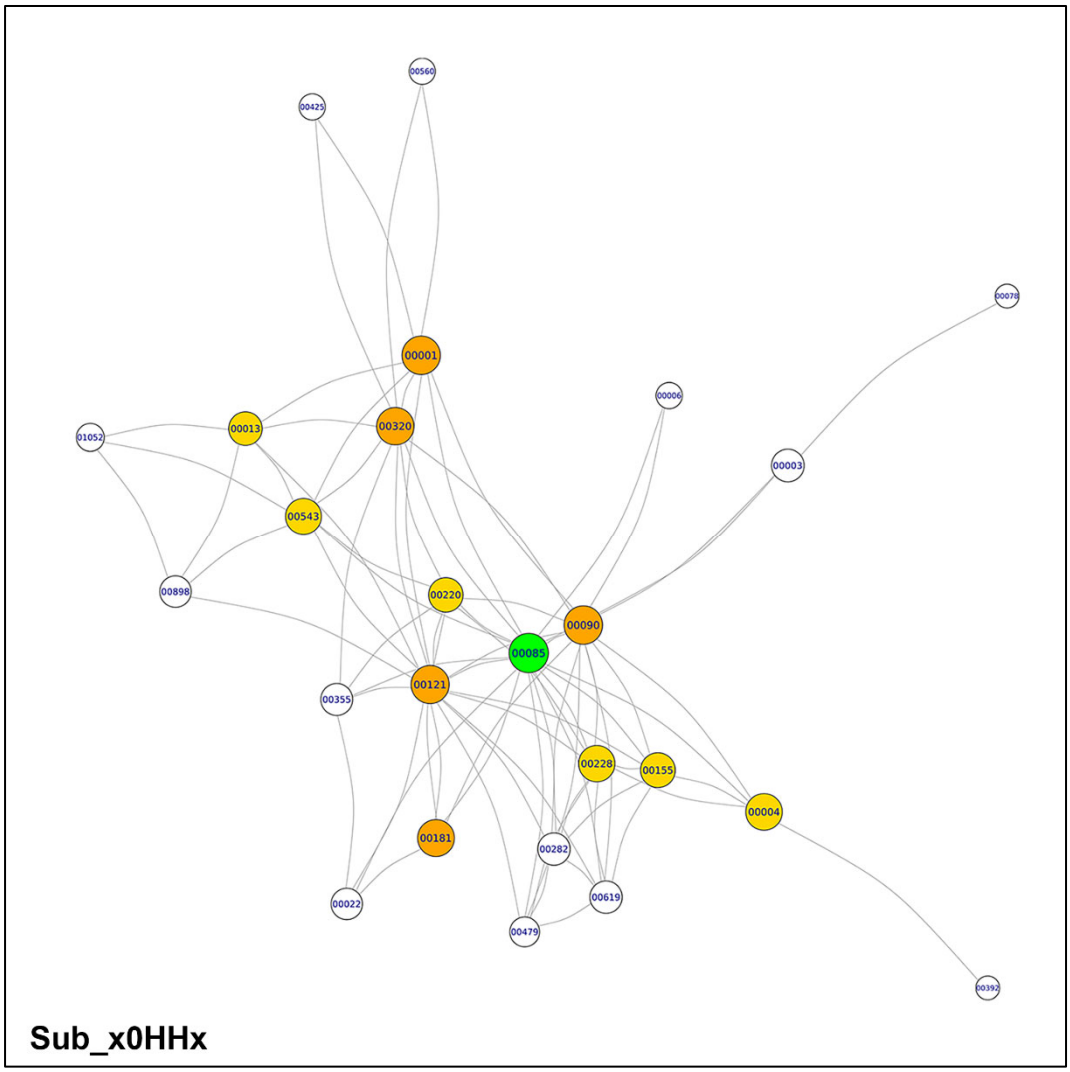
The main hubs or keystone taxa in the Sub_x0HHx network were: *Streptococcus unclassified* ASV85, *Fusobacterium unclassified* ASV14, and *Streptococcus unclassified* ASV90. All were part of the Sub_x0HHx core microbiota and were present in higher relative abundances in this group than in Sub_x0PDx and Sub_x1PDx, although only the *Fusobacterium* was present in an abundance ≥0.5%.

In the Sub_x0PDx network, the main keystone ASVs were: *Streptococcus unclassified* ASV121, *Tannerella forsythia* ASV15, and *Streptococcus unclassified* ASV85. All of them had an abundance ≥0.5% in Sub_x0PDx and were differentially abundant when compared to Sub_x1PDx; but only *T. forsythia* was part of the core microbiota and differentially abundant when compared to Sub_x0HHx.

Lastly, the main hubs or keystone taxa in the Sub_x1PDx network were: *T. forsythia* ASV15, *Fusobacterium nucleatum* subsp. *vincentii* ASV10, and *S. oralis* subsp. *dentisani* clade 058 ASV1. The two later taxa belonged to the Sub_x1PDx core but, despite all of them had an abundance ≥0.5% in Sub_x1PDx, none had significantly greater abundance in this group *vs.* Sub_x0HHx or Sub_x0PDx.



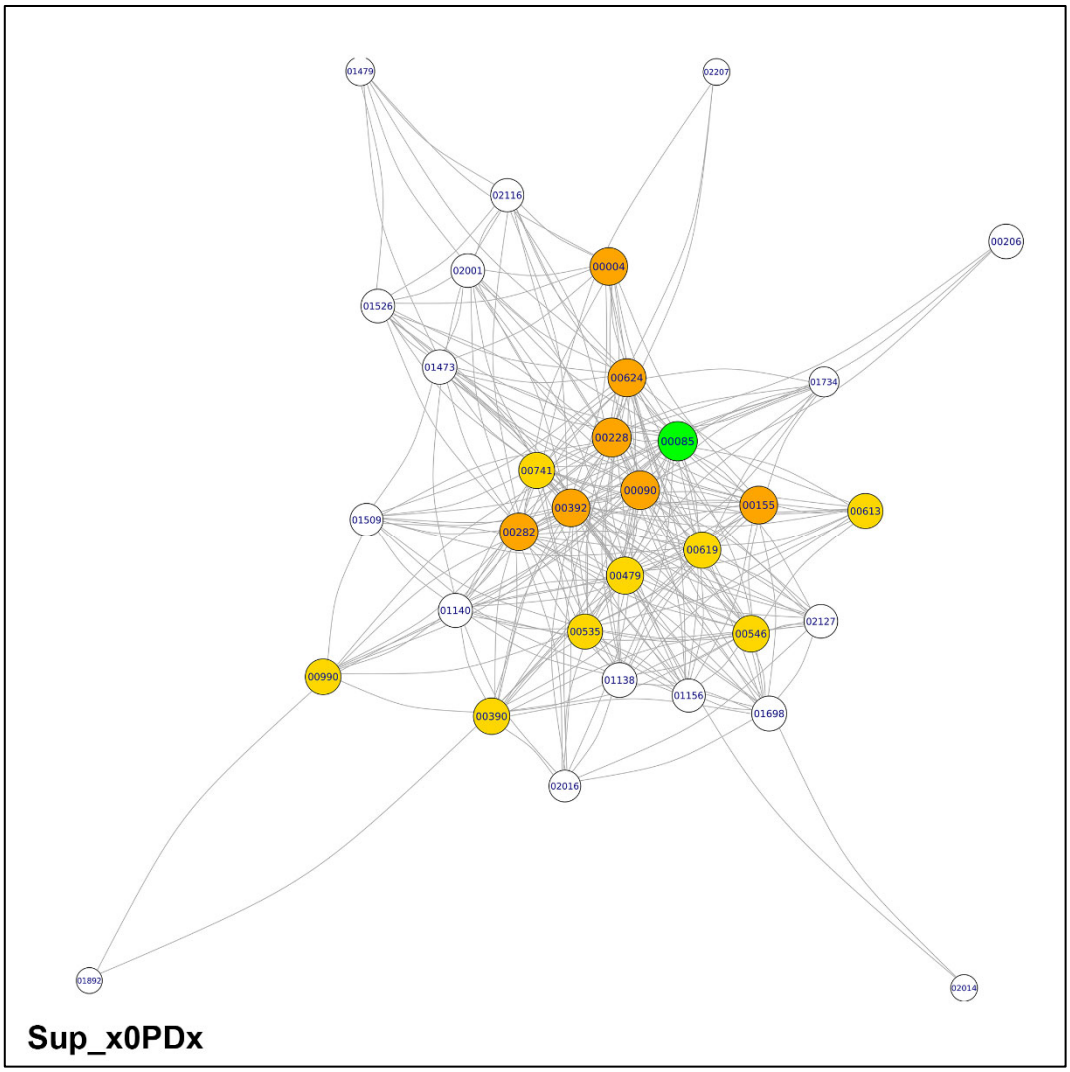
Figures 9, 10, and 11 represent the main modules of the co-occurrence networks associated with Sub_x0HHx, Sub_x0PDx, and Sub_x1PDx, respectively.




| Sub_x0HHx Samples= 155 | | | | | |
|---|---|---|---|---|---|
| ASVid | Genus | Species | ASV | Core | Relative abundance |
| AV00001 | Streptococcus | oralis_subsp.dentisani clade_058 | BTASV016027 | Y | 7.2773 |
| AV00004 | Streptococcus | unclassified | unclassified | Y | 1.5538 |
| AV00013 | Granulicatella | adiacens | unclassified | Y | 1.1744 |
| AV00085 | Streptococcus | unclassified | unclassified | Y | 0.1300 |
| AV00090 | Streptococcus | unclassified | unclassified | Y | 0.1240 |
| AV00121 | Streptococcus | unclassified | unclassified | Y | 0.0817 |
| AV00155 | Streptococcus | unclassified | unclassified | N | 0.1341 |
| AV00181 | Streptococcus | unclassified | unclassified | N | 0.0635 |
| AV00220 | Streptococcus | unclassified | unclassified | N | 0.0343 |
| AV00228 | Streptococcus | sanguinis | unclassified | N | 0.0344 |
| AV00320 | Streptococcus | oralis_subsp.dentisani clade_058 | unclassified | Y | 0.0271 |
| AV00543 | Streptococcus | unclassified | unclassified | N | 0.0227 |

**Figure 9. Main module of the co-occurrence network associated with the subgingival plaque of periodontally healthy subjects.**

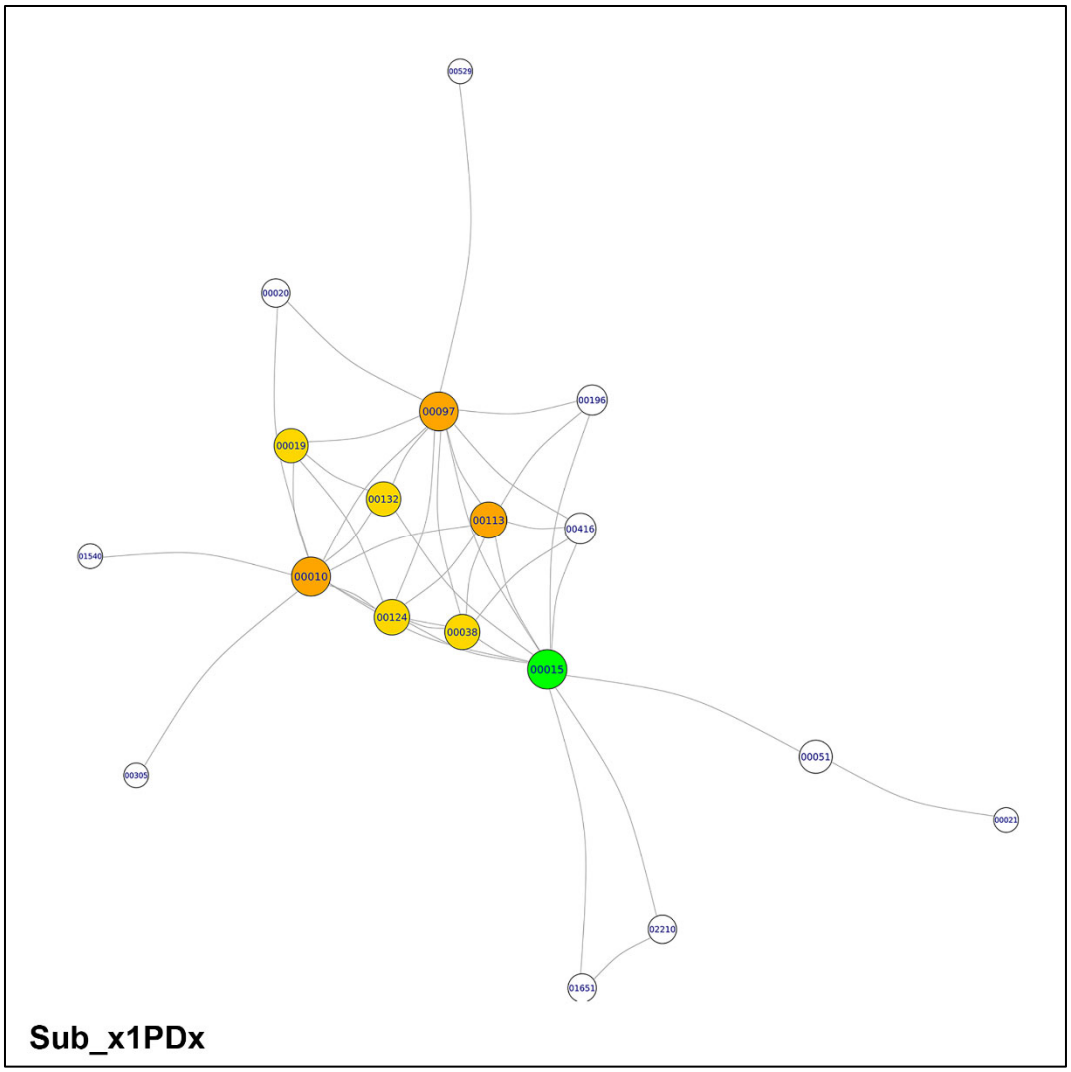
In the graph, the most important taxa are highlighted in green, orange, and yellow according to the score obtained in the analysis group.The highest value is shown in green, the values belonging to the first quartile in orange, and those belonging to the second quartile, i.e. up to the median, in yellow.
ASV= amplicon sequence variant; ASVid= amplicon sequence variant identifier; N= no; subsp.= subspecies; Sub_x0HHx= subgingival plaque of periodontally healthy subjects, healthy sites; Y= yes.




| Sub_x0PDx Samples= 768 | | | | | |
|---|---|---|---|---|---|
| **ASVid** | **Genus** | **Species** | **ASV** | **Core** | **Relative abundance** |
| AV00010 | Fusobacterium | nucleatum_subsp. vincentii | unclassified | Y | 1.8295 |
| AV00015 | Tannerella | forsythia | BTASV153103 | Y | 1.8566 |
| AV00051 | Peptostreptococcaceae [XI][G-9] | brachy | BTASV129419 | Y | 0.6348 |
| AV00097 | Fretibacterium | fastidiosum | unclassified | Y | 0.4676 |
| AV00124 | Peptostreptococcaceae [XI][G-4] | bacterium_HMT369 | BTASV096563 | Y | 0.3447 |

**Figure 10. Main module of the co-occurrence network associated with the subgingival plaque of periodontitis subjects in the diseased sites.**

In the graph, the most important taxa are highlighted in green, orange, and yellow according to the score obtained in the analysis group.The highest value is shown in green, the values belonging to the first quartile in orange, and those belonging to the second quartile, i.e. up to the median, in yellow.
ASV= amplicon sequence variant; ASVid= amplicon sequence variant identifier; subsp.= subspecies; Sub_x0PDx= subagingival plaque of periodontitis subjects, diseased sites; Y= yes.




| Sub_x1PDx Samples= 197 | | | | | |
|---|---|---|---|---|---|
| ASVid | Genus | Species | ASV | Core | Relative abundance |
| AV00010 | Fusobacterium | nucleatum_subsp .vincentii | unclassified | Y | 1.7567 |
| AV00015 | Tannerella | forsythia | BTASV153103 | N | 1.0663 |
| AV00019 | Filifactor | alocis | BTASV124203 | N | 0.7452 |
| AV00038 | Treponema | denticola | BTASV138814 | N | 0.5607 |
| AV00097 | Fretibacterium | fastidiosum | unclassified | N | 0.3035 |
| AV00113 | Fretibacterium | unclassified | unclassified | N | 0.2714 |
| AV00124 | Peptostreptococcaceae [XI][G-4] | bacterium_HMT369 | BTASV096563 | N | 0.2289 |
| AV00132 | Treponema | lecithinolyticum | BTASV162382 | N | 0.2025 |

**Figure 11. Main module of the co-occurrence network associated with the subgingival plaque of periodontitis subjects in the diseased sites after therapy.**

In the graph, the most important taxa are highlighted in green, orange, and yellow according to the score obtained in the analysis group.The highest value is shown in green, the values belonging to the first quartile in orange, and those belonging to the second quartile, i.e. up to the median, in yellow.
ASV= amplicon sequence variant; ASVid= amplicon sequence variant identifier; N= no; subsp.= subspecies; Sub_x1PDx= subagingival plaque of periodontitis subjects, diseased sites after therapy; Y= yes.

### 4.4.10. Predictive models in supragingival and subgingival plaque microbiota

#### *4.4.10.1. Supragingival plaque microbiota*

The predictive models for the supragingival microbiota had AUC values ranging from 0.818 to 0.996, representing, at most, 2.39% and 17.37% of the ASVs and species, respectively. The models used to distinguish periodontal health from periodontitis after therapy and gingivitis from both untreated and treated periodontitis had AUC values >0.970, indicating the presence of very low numbers of ASVs (range= 30 - 20; 0.40% - 0.28%). In contrast, the models for differentiating periodontal health from gingivitis and periodontitis had more (150 and 70, respectively), with AUC values of around 0.895 (Table 7). Appendix S12 contains a list of all taxa that were part of these models and the group they predicted.

**Table 7. Number of taxa that composed the predictive models to distinguish the periodontal health conditions in the supragingival plaque, and the derived AUC values.**

| | No. ASVs (% detected) | No. Species (% detected) | No. Core ASVs (% detected)* | No. Core species (% detected)* | AUC |
|---|---|---|---|---|---|
| **Predictive models to distinguish periodontal health conditions in the supragingival plaque** | | | | | |
| Sup_x0HHx *vs.* Sup_x0GDx | 150 (2.39%) | 87 (17.37%) | 48 (32.00%) | 33 (37.93%) | 0.8948 |
| Sup_x0HHx *vs.* Sup_x0PDx | 70 (0.93%) | 50 (9.47%) | 19 (27.14%) | 16 (32.00%) | 0.8986 |
| Sup_x0HHx *vs.* Sup_x1PDx | 20 (0.28%) | 12 (2.34%) | 17 (85.00%) | 11 (91.67%) | 0.9884 |
| Sup_x0GDx *vs.* Sup_x0PDx | 30 (0.39%) | 23 (4.30%) | 8 (26.67%) | 5 (21.74%) | 0.9724 |
| Sup_x0GDx *vs.* Sup_x1PDx | 20 (0.40%) | 17 (3.51%) | 8 (40.00%) | 7 (41.18%) | 0.9962 |
| Sup_x0PDx *vs.* Sup_x1PDx | 8 (0.11%) | 4 (0.76%) | 1 (12.50%) | 1 (25.00%) | 0.8185 |

The percentages of detected ASVs and species are calculated with respect to the total number of different ASVs and species detected by at least one of the groups to be compared. *The percentages of core ASVs and species are calculated with respect to the total number of different ASVs and species that showed predictivity in the two groups compared. The taxa that could not be classified at the species level ("unclassified") were counted once so the number of species detected is the minimum that could be obtained.
ASVs= amplicon sequence variants; AUC= area under the curve; No.= number; Sup_x0GDx= supragingival plaque of gingivitis subjects, diseased sites; Sup_x0HHx= supragingival plaque of periodontally healthy subjects, healthy sites; Sup_x0PDx= supragingival plaque of periodontitis subjects, diseased sites; Sup_x1PDx= supragingival plaque of periodontitis subjects, diseased sites after therapy.

The ROC curves of the predictive models in supragingival plaque and their derived AUC values are represented in figures 12 and 13.

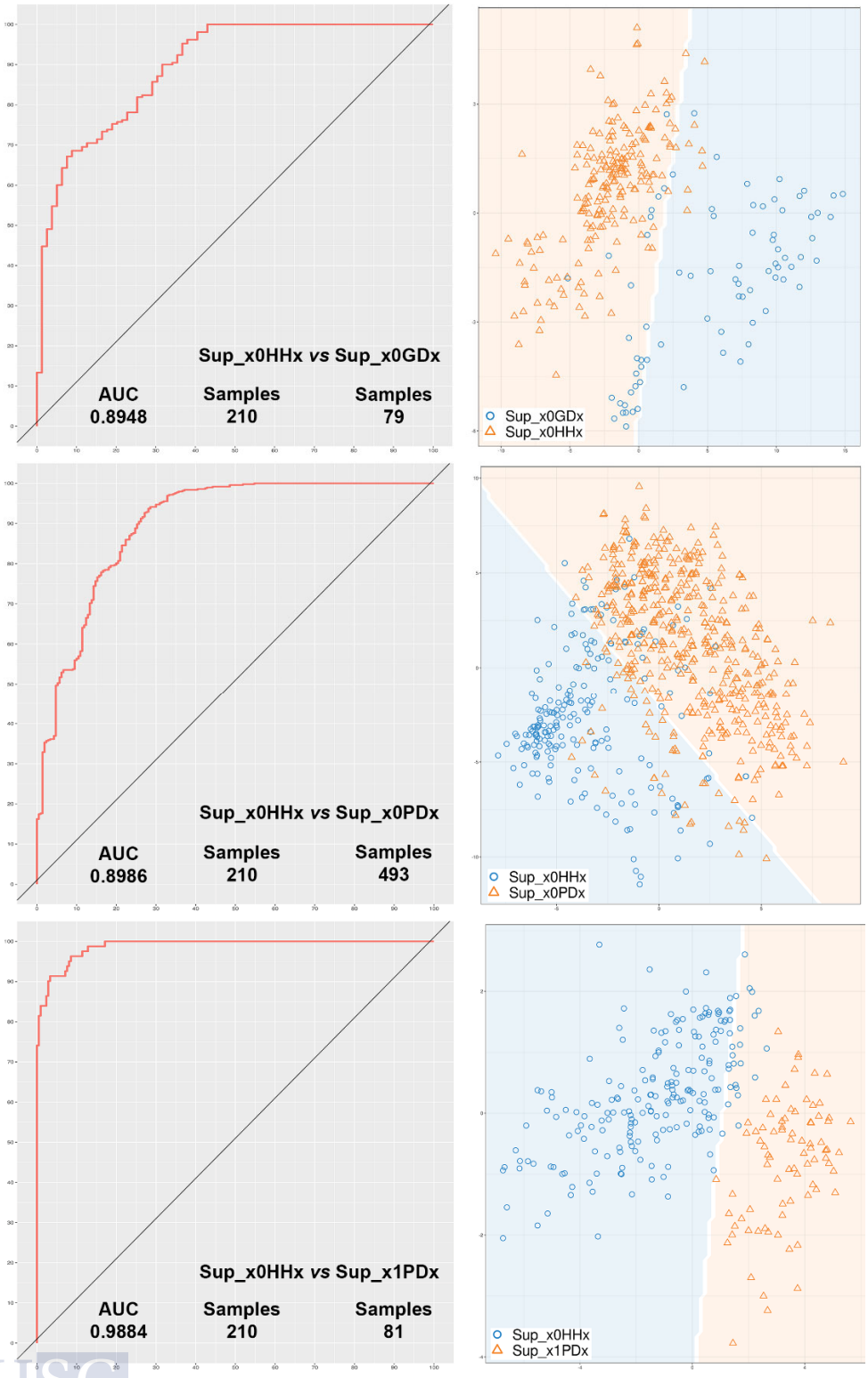



**Figure 12. Potential of the supragingival plaque microbiota to discriminate periodontal health from gingivitis and non-treated and treated periodontitis: ROC curves and AUC values.**

AUC= area under the curve; Sup_x0GDx= supragingival plaque of gingivitis subjects, diseased sites; Sup_x0HHx= supragingival plaque of periodontally healthy subjects, healthy sites; Sup_x0PDx= supragingival plaque of periodontitis subjects, diseased sites; Sup_x1PDx= supragingival plaque of periodontitis subjects, diseased sites after therapy.

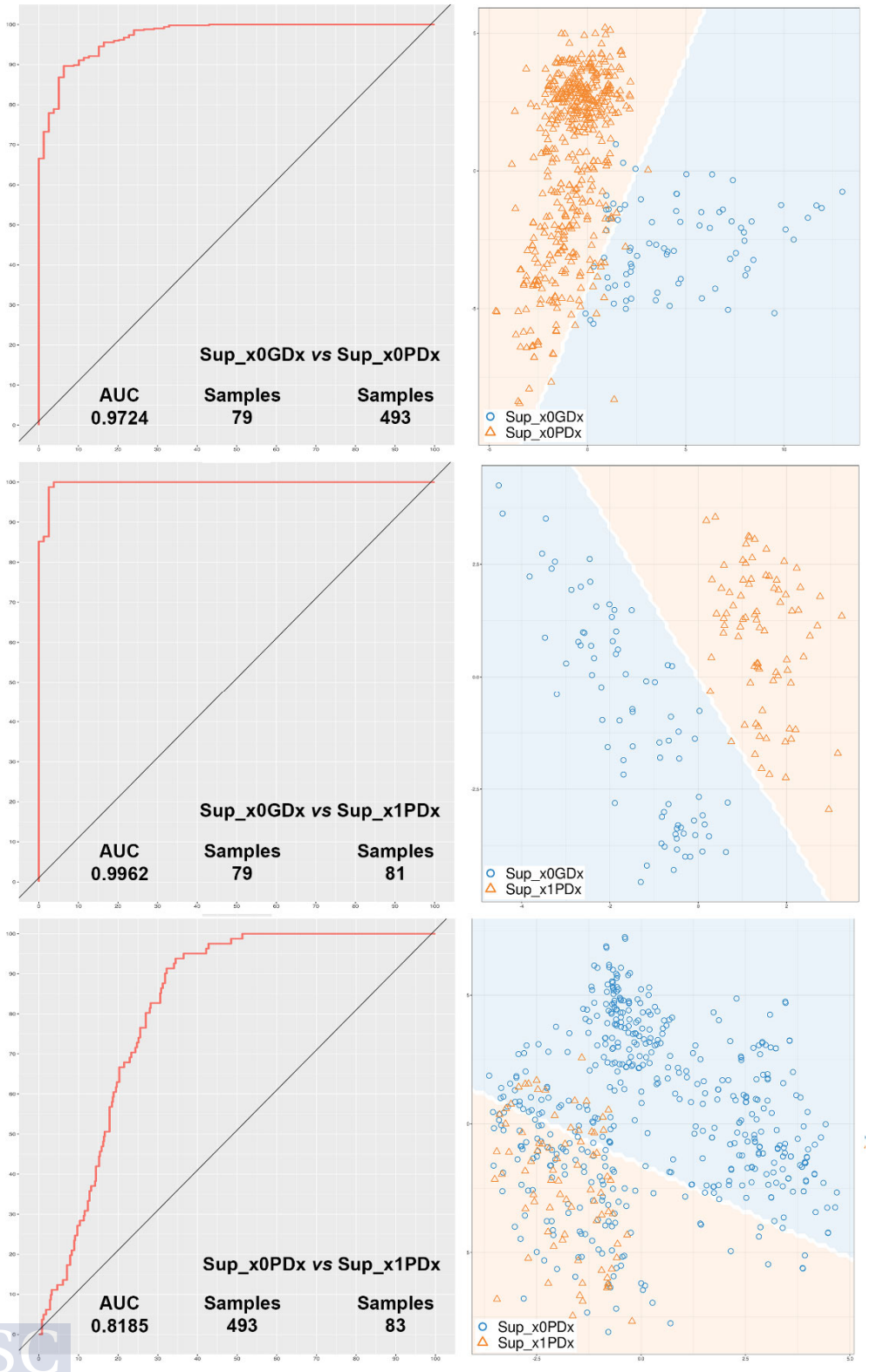



**Figure 13. Potential of the supragingival plaque microbiota to discriminate the different periodontal disease groups: ROC curves and AUC values.**

AUC= area under the curve; Sup_x0GDx= supragingival plaque of gingivitis subjects, diseased sites; Sup_x0PDx= supragingival plaque of periodontitis subjects, diseased sites; Sup_x1PDx= supragingival plaque of periodontitis subjects, diseased sites after therapy.

The supragingival plaque ASVs of the genera *Bacteroides*, *Capnoctophaga*, *Haemophilus*, *Kingella*, *Neisseria*, *Leptotrichia*, *Sacchari*, and *Serratia* acted as predictors of periodontal health. Of these, the most important (in order of relative abundance) were: 1) *Leptotrichia hongkongensis* ASV24 (1.19%); 2) *Neisseria macacae* ASV9 (0.75%); 3) *Neisseria bacilliformis* ASV281 (0.57%); 4) *Capnocytophaga sputigena* ASV56 (0.29%); and 5) *Capnocytophaga gingivalis* ASV93 (0.28%). All of these taxa, apart from *N. bacilliformis*, were core ASVs (Table 8).

**Table 8. Main taxa predictive of periodontal health in supragingival plaque.**

| ASVid | Genus | Species | ASV | Sup_x0HHx_vs_Sup_x0GDx | Sup_x0HHx_vs_Sup_x0PDx | Sup_x0HHx_vs_Sup_x1PDx | Sup_x0GDx_vs_Sup_x0PDx | Sup_x0GDx_vs_Sup_x1PDx | Sup_x0PDx_vs_Sup_x1PDx | Sub_x0HHx_vs_Sub_x0PHx | Sub_x0HHx_vs_Sub_x0PDx | Sub_x0HHx_vs_Sub_x1PDx | Sub_x0PHx_vs_Sub_x0PDx | Sub_x0PHx_vs_Sub_x1PDx | Sub_x0PDx_vs_Sub_x1PDx |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| ASV1908 | Bacteroides | heparinolyticus | unclassified | | | | | | | | | | | | |
| ASV0093 | Capnocytophaga | gingivalis | unclassified | C | | | | | | | | | | | |
| ASV0239 | Capnocytophaga | gingivalis | unclassified | | | | | | | | | | | | |
| ASV0120 | Capnocytophaga | granulosa | BTASV089348 | | | | | | | | | | | | |
| ASV0212 | Capnocytophaga | granulosa | BTASV089343 | | | | | | | | | | | | |
| ASV0056 | Capnocytophaga | sputigena | unclassified | C | | | | | | | | | | | |
| ASV0494 | Capnocytophaga | sputigena | unclassified | | | | | | | | | | | | |
| ASV0564 | Haemophilus | sputorum | unclassified | | | | | | | | | | | | |
| ASV0430 | Kingella | denitrificans | BTASV138706 | | | | | | | | | | | | |
| ASV0024 | Leptotrichia | hongkongensis | unclassified | C | | | | | | | | | | | |
| ASV0480 | Leptotrichia | hongkongensis | unclassified | | | | | | | | | | | | |
| ASV0281 | Neisseria | bacilliformis | BTASV002174 | | | | | | | | | | | | |
| ASV0586 | Neisseria | bacilliformis | unclassified | | | | | | | | | | | | |
| ASV0780 | Neisseria | bacilliformis | BTASV002176 | | | | | | | | | | | | |
| ASV0928 | Neisseria | bacilliformis | BTASV002174 | | | | | | | | | | | | |
| ASV0009 | Neisseria | macacae | unclassified | C | | | | | | | | | | | |
| ASV0302 | Neisseria | macacae | BTASV035730 | | | | | | | | | | | | |
| ASV0751 | Neisseria | sp.HMT499 | unclassified | | | | | | | | | | | | |
| ASV0345 | Sacchari | BTASV095519 | unclassified | | | | | | | | | | | | |
| ASV0938 | Serratia | marcescens | unclassified | | | | | | | | | | | | |
| ASV0970 | Serratia | marcescens | unclassified | | | | | | | | | | | | |

Cells are coloured according to the periodontal health condition predicted by the taxon in question. The green colour was associated with periodontally healthy subjects, healthy sites; pink with periodontitis, healthy sites; and orange with periodontitis, diseased sites after therapy.
ASV= amplicon sequence variant; ASVid= amplicon sequence variant identifier; C= core member; Sub_x0HHx= subgingival plaque of periodontally healthy subjects, healthy sites; Sub_x0PHx= subgingival plaque of periodontitis subjects, healthy sites; Sub_x0PDx= subgingival plaque of periodontitis subjects, diseased sites; Sub_x1PDx= subgingival plaque of periodontitis subjects, diseased sites after therapy; Sup_x0GDx= supragingival plaque of gingivitis subjects, diseased sites; Sup_x0HHx= supragingival plaque of periodontally healthy subjects, healthy sites; Sup_x0PDx= supragingival plaque of periodontitis subjects, diseased sites; Sup_x1PDx= supragingival plaque of periodontitis subjects, diseased sites after therapy.



The supragingival plaque ASVs of the genera *Campylobacter*, *Leptotrichia*, *Parvimonas*, *Prevotella*, and *Selenomonas* acted as predictors of periodontal disease, mainly gingivitis; three also predicted periodontitis. A focus on the relative abundance values observed in Sup_x0GDx

revealed that the most relevant ASVs were *Campylobacter rectus* ASV20 and *Parvimonas* HMT110 ASV21, both of which not only had the ability to distinguish gingivitis (core ASV20 - abundances: 0.49% and 0.40%, respectively), but also periodontitis (both ASVs were core species for this condition: 0.35% and 0.51%, respectively). *P.* HMT110 ASV21 also acted as a predictor of treated periodontitis. Other abundant ASVs associated with gingivitis were *Prevotella nigrescens* ASV25 and ASV86 (core ASV25 - 0.42% and 0.36%, respectively), which also appeared to distinguish treated periodontitis, and *Selenomonas sputigena* ASV243 (0.55%) (Table 9).

**Table 9. Main taxa predictive of periodontal diseases in supragingival plaque.**

| ASVid | Genus | Species | ASV | Sup_x0HHx_vs_Sup_x0GDx | Sup_x0HHx_vs_Sup_x0PDx | Sup_x0HHx_vs_Sup_x1PDx | Sup_x0GDx_vs_Sup_x0PDx | Sup_x0GDx_vs_Sup_x1PDx | Sup_x0PDx_vs_Sup_x1PDx | Sub_x0HHx_vs_Sub_x0PHx | Sub_x0HHx_vs_Sub_x0PDx | Sub_x0HHx_vs_Sub_x1PDx | Sub_x0PHx_vs_Sub_x0PDx | Sub_x0PHx_vs_Sub_x1PDx | Sub_x0PDx_vs_Sub_x1PDx |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| ASV0020 | Campylobacter | rectus | BTASV188427 | C | C | | | | | | | C | | C | |
| ASV0055 | Leptotrichia | wadei | BTASV075529 | | | | | | | C | | | | | |
| ASV0146 | Leptotrichia | wadei | BTASV075583 | | | | | | | | | | | | |
| ASV0021 | Parvimonas | sp.HMT110 | unclassified | | C | C | | | | | | | | C | |
| ASV0100 | Prevotella | intermedia | unclassified | | | | | | | | | | | | |
| ASV0347 | Prevotella | intermedia | unclassified | | | | | | | | | | | | |
| ASV0406 | Prevotella | intermedia | unclassified | | | | | | | | | | | | |
| ASV0437 | Prevotella | intermedia | unclassified | | | | | | | | | | | | |
| ASV1461 | Prevotella | intermedia | unclassified | | | | | | | | | | | | |
| ASV0025 | Prevotella | nigrescens | unclassified | C | | C | | | C | | | | | | |
| ASV0086 | Prevotella | nigrescens | BTASV174229 | | | | | | | | | | | | |
| ASV0428 | Prevotella | nigrescens | unclassified | | | | | | | | | | | | |
| ASV1550 | Prevotella | nigrescens | unclassified | | | | | | | | | | | | |
| ASV0144 | Prevotella | oris | BTASV085458 | | | C | | C | | | | | | | |
| ASV0205 | Prevotella | oris | unclassified | | | | | | | | | | | | |
| ASV0376 | Prevotella | oris | unclassified | | | | | | | | | | | | |
| ASV1007 | Prevotella | oris | BTASV085459 | | | | | | | | | | | | |
| ASV1619 | Prevotella | oris | unclassified | | | | | | | | | | | | |
| ASV0156 | Selenomonas | noxia | unclassified | | | | | | | | | | | | |
| ASV0261 | Selenomonas | noxia | BTASV049415 | | | | | | | | | | | | |
| ASV1336 | Selenomonas | noxia | unclassified | | | | | | | | | | | | |
| ASV0214 | Selenomonas | sputigena | BTASV050374 | | | | | | | | | | | | |
| ASV0243 | Selenomonas | sputigena | unclassified | | | | | | | | | | | | |
| ASV0340 | Selenomonas | sputigena | unclassified | | | | | | | | | | | | |
| ASV0837 | Selenomonas | sputigena | unclassified | | | | | | | | | | | | |

Cells are coloured according to the periodontal health condition predicted by the taxon in question. The yellow colour was associated with gingivitis, diseased sites; pink with periodontitis, healthy sites; red with periodontitis, diseased sites; and orange with periodontitis, diseased sites after therapy.
ASV= amplicon sequence variant; ASVid= amplicon sequence variant identifier; C= core member; Sub_x0HHx= subgingival plaque of periodontally healthy subjects, healthy sites; Sub_x0PHx= subgingival plaque of periodontitis subjects, healthy sites; Sub_x0PDx= subgingival plaque of periodontitis subjects, diseased sites; Sub_x1PDx= subgingival plaque of periodontitis subjects, diseased sites after therapy; Sup_x0GDx= supragingival plaque of gingivitis subjects, diseased sites; Sup_x0HHx= supragingival plaque of periodontally healthy subjects, healthy sites; Sup_x0PDx= supragingival plaque of periodontitis subjects, diseased sites; Sup_x1PDx= supragingival plaque of periodontitis subjects, diseased sites after therapy.

The supragingival plaque ASVs of the genera *Corynebacterium*, *Eikenella*, *Streptococcus*, and *Veillonella* emerged as predictors of both periodontal health and gingivitis or periodontitis. For example, *S. oralis* subsp. *dentisani* clade 058 had four core ASVs which were predictors of health (highly abundant core ASV1: 11.27%); and three others, two of them core ASVs, were predictors of periodontitis (ASV425, ASV877, and ASV1534: abundances of 0.07%, 0.01%, and 0.01%, respectively). Similarly, while the highly abundant *Veillonella dispar* ASV5 (core, 3.55%) was associated with gingivitis, two other ASVs of this species (ASV16: 0.23%; and ASV89: 0.17%) distinguished health. We even detected a single core ASV of *S. sanguinis* that had the capacity to predict the contrasting clinical conditions of health and periodontitis (abundance >0.20% in both conditions) (Table 10).

**Table 10. Main taxa predictive of periodontal health and diseases in supragingival plaque.**

| ASVid | Genus | Species | ASV | Sup_x0HHx_vs_Sup_x0GDx | Sup_x0HHx_vs_Sup_x0PDx | Sup_x0HHx_vs_Sup_x1PDx | Sup_x0GDx_vs_Sup_x0PDx | Sup_x0GDx_vs_Sup_x1PDx | Sup_x0PDx_vs_Sup_x1PDx | Sub_x0HHx_vs_Sub_x0PHx | Sub_x0HHx_vs_Sub_x0PDx | Sub_x0HHx_vs_Sub_x1PDx | Sub_x0PHx_vs_Sub_x0PDx | Sub_x0PHx_vs_Sub_x1PDx | Sub_x0PDx_vs_Sub_x1PDx |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| ASV0123 | Corynebacterium | matruchotii | BTASV165891 | | | | | | | | | | | | |
| ASV0177 | Corynebacterium | matruchotii | Unclassified | | | | | | | | | | | | |
| ASV0262 | Corynebacterium | matruchotii | Unclassified | | | | | | | | | | | | |
| ASV0287 | Corynebacterium | matruchotii | Unclassified | | | | | | | | | | | | |
| ASV0381 | Corynebacterium | matruchotii | Unclassified | | | | | | | | | | | | |
| ASV0803 | Corynebacterium | matruchotii | Unclassified | | | | | | | | | | | | |
| ASV0851 | Corynebacterium | matruchotii | Unclassified | | | | | | | | | | | | |
| ASV0870 | Corynebacterium | matruchotii | Unclassified | | | | | | | | | | | | |
| ASV1347 | Corynebacterium | matruchotii | Unclassified | | | | | | | | | | | | |
| ASV2602 | Corynebacterium | matruchotii | Unclassified | | | | | | | | | | | | |
| ASV0361 | Eikenella | corrodens | Unclassified | C | | | C | | | | | | | | |
| ASV0472 | Eikenella | corrodens | BTASV138315 | | | | | | | | | | | | |
| ASV0548 | Eikenella | corrodens | BTASV138321 | | | | | | | | | | | | |
| ASV0001 | Streptococcus | oralis_subsp.dentisani_clade_058 | BTASV016027 | C | | | | | | | C | C | | C | |
| ASV0114 | Streptococcus | oralis_subsp.dentisani_clade_058 | Unclassified | | | | | | | | | | | | |
| ASV0320 | Streptococcus | oralis_subsp.dentisani_clade_058 | Unclassified | C | | C | | | | | C | | | | |
| ASV0356 | Streptococcus | oralis_subsp.dentisani_clade_058 | Unclassified | | | | | | | | | | | | |
| ASV0425 | Streptococcus | oralis_subsp.dentisani_clade_058 | Unclassified | | | | | | | | | | | | |
| ASV0560 | Streptococcus | oralis_subsp.dentisani_clade_058 | Unclassified | | C | | | | | | C | | | | |
| ASV0674 | Streptococcus | oralis_subsp.dentisani_clade_058 | Unclassified | | | | | | | | | | | | |
| ASV0877 | Streptococcus | oralis_subsp.dentisani_clade_058 | Unclassified | | | | C | | | | | | | | |
| ASV0962 | Streptococcus | oralis_subsp.dentisani_clade_058 | Unclassified | | C | | | | | | | | | | |
| ASV1534 | Streptococcus | oralis_subsp.dentisani_clade_058 | Unclassified | | C | | | | | | | | | | |
| ASV0228 | Streptococcus | sanguinis | Unclassified | C | C | C | | | | | | | | | |
| ASV0392 | Streptococcus | sanguinis | Unclassified | C | | C | | | | | | | | | |
| ASV0619 | Streptococcus | sanguinis | Unclassified | | | | | | | | | | | | |
| ASV0005 | Veillonella | dispar | BTASV053366 | C | | | | | | C | | | C | C | C |
| ASV0016 | Veillonella | dispar | BTASV053367 | | | | | C | | | | | | | |
| ASV0089 | Veillonella | dispar | unclassified | | | | | | | C | | | C | C | |
| ASV0131 | Veillonella | dispar | unclassified | | | | | | | C | | | | C | |

Cells are coloured according to the periodontal health condition predicted by the taxon in question. The green colour was associated with periodontally healthy subjects, healthy sites; yellow with gingivitis, diseased sites; pink with periodontitis, healthy sites; red with periodontitis, diseased sites; and orange with periodontitis, diseased sites after therapy.

ASV= amplicon sequence variant; ASVid= amplicon sequence variant identifier; C= core member; Sub_x0HHx= subgingival plaque of periodontally healthy subjects, healthy sites; Sub_x0PHx= subgingival plaque of periodontitis subjects, healthy sites; Sub_x0PDx= subgingival plaque of periodontitis subjects, diseased sites; Sub_x1PDx= subgingival plaque of periodontitis subjects, diseased sites after therapy; Sup_x0GDx= supragingival plaque of gingivitis subjects, diseased sites; Sup_x0HHx= supragingival plaque of periodontally healthy subjects, healthy sites; Sup_x0PDx= supragingival plaque of periodontitis subjects, diseased sites; Sup_x1PDx= supragingival plaque of periodontitis subjects, diseased sites after therapy.

*4.4.10.2. Subgingival plaque microbiota*

The predictive models of the subgingival microbiota had AUC values ranging from 0.796 to 0.902, representing, at most, 2.85% and 21.67% ASVs and species, respectively. In relation to their predictive potential, the models requiring lower numbers of ASVs were Sub_x0PHx *vs.* Sub_x0PDx (20 ASVs) and Sub_x0HHx *vs.* Sub_x0PHx (50 ASVs), with AUC values of 0.885 and 0.902, respectively. The remaining models identified higher numbers of predictor ASVs, with the range being from 80 for distinguishing between Sub_x0PDx *vs.* Sub_x1PDx (AUC= 0.796) to 200 for Sub_x0HHx *vs.* Sub_x1PDx (AUC= 0.888) (Table 11). Appendix S12 contains a list of all taxa that were part of these models and the group they predicted.

**Table 11. Number of taxa that composed the predictive models to distinguish the periodontal health conditions in the subgingival plaque, and the derived AUC values.**

| | No. ASVs (% detected) | No. Species (% detected) | No. Core ASVs (% detected)* | No. Core species (% detected)* | AUC |
|---|---|---|---|---|---|
| **Predictive models to distinguish periodontal health conditions in the subgingival plaque** | | | | | |
| Sub_x0HHx *vs.* Sub_x0PHx | 50 (0.78%) | 42 (7.75%) | 16 (32.00%) | 12 (28.57%) | 0.9024 |
| Sub_x0HHx *vs.* Sub_x0PDx | 140 (1.65%) | 87 (15.54%) | 28 (20.00%) | 21 (24.14%) | 0.8698 |
| Sub_x0HHx *vs.* Sub_x1PDx | 200 (2.85%) | 117 (21.67%) | 21 (10.50%) | 14 (11.97%) | 0.8883 |
| Sub_x0PHx *vs.* Sub_x0PDx | 20 (0.24%) | 15 (2.67%) | 10 (50.00%) | 7 (46.67%) | 0.8850 |
| Sub_x0PHx *vs.* Sub_x1PDx | 90 (1.50%) | 62 (11.76%) | 18 (20.00%) | 11 (17.74%) | 0.8803 |
| Sub_x0PDx *vs.* Sub_x1PDx | 80 (0.95%) | 52 (9.32%) | 7 (8.75%) | 6 (11.54%) | 0.7966 |

The percentages of detected ASVs and species are calculated with respect to the total number of different ASVs and species detected by at least one of the groups to be compared. *The percentages of core ASVs and species are calculated with respect to the total number of different ASVs and species that showed predictivity in the two groups compared. The taxa that could not be classified at the species level ("unclassified") were counted once so the number of species detected is the minimum that could be obtained.
ASVs= amplicon sequence variants; AUC= area under the curve; No.= number; Sub_x0HHx= subgingival plaque of periodontally healthy subjects, healthy sites; Sub_x0PHx= subgingival plaque of periodontitis subjects, healthy sites; Sub_x0PDx= subgingival plaque of periodontitis subjects, diseased sites; Sub_x1PDx= subgingival plaque of periodontitis subjects, diseased sites after therapy.

The ROC curves of the predictive models in subgingival plaque and their derived AUC values are represented in figures 14 and 15.


**Figure 14. Potential of the subgingival plaque microbiota to discriminate periodontal health from the different periodontitis groups: ROC curves and AUC values.**

AUC= area under the curve; Sub_x0HHx= subgingival plaque of periodontally healthy subjects, healthy sites; Sub_x0PHx= subgingival plaque of periodontitis subjects, healthy sites; Sub_x0PDx= subgingival plaque of periodontitis subjects, diseased sites; Sub_x1PDx= subgingival plaque of periodontitis subjects, diseased sites after therapy.


**Figure 15. Potential of the subgingival plaque microbiota to discriminate the different periodontitis groups: ROC curves and AUC values.**

AUC= area under the curve; Sub_x0PHx= subgingival plaque of periodontitis subjects, healthy sites; Sub_x0PDx= subgingival plaque of periodontitis subjects, diseased sites; Sub_x1PDx= subgingival plaque of periodontitis subjects, diseased sites after therapy.

The subgingival plaque ASVs of the genera *Aggregatibacter*, *Capnocytophaga*, *Cupriavidus*, *Gemella*, *Granulicatella*, *Lachnospiraceae* [G-3], *Leptotrichia*, *Oribacterium*, *Porphyromonas*, *Prevotella*, *Pseudomonas*, *Pseudopropionibacterium*, *Sphingomonas*, and *Veillonella* acted as predictors of periodontal health. Of these, the most important in order of relative abundance were *Granulicatella adiacens* ASV13 (core, 0.52%), *Gemella haemolysans* ASV26 (core, 0.36%), *Capnocytophaga leadbetteri* ASV126 (0.30%), *Aggregatibacter* HMT458 ASV145 (0.18%), and *Prevotella melaninogenica* ASV7 (core, 0.16%). Both *G. adiacens* ASV13 and *C. leadbetteri* ASV126 also had a predictive capacity in the Sub_x0PHx and Sub_x1PDx groups, respectively (Table 12).

**Table 12. Main taxa predictive of periodontal health in subgingival plaque.**

| ASVid | Genus | Species | ASV | Sup_x0HHx_vs_Sup_x0GDx | Sup_x0HHx_vs_Sup_x0PDx | Sup_x0HHx_vs_Sup_x1PDx | Sup_x0GDx_vs_Sup_x0PDx | Sup_x0GDx_vs_Sup_x1PDx | Sup_x0PDx_vs_Sup_x1PDx | Sub_x0HHx_vs_Sub_x0PHx | Sub_x0HHx_vs_Sub_x0PDx | Sub_x0HHx_vs_Sub_x1PDx | Sub_x0PHx_vs_Sub_x0PDx | Sub_x0PHx_vs_Sub_x1PDx | Sub_x0PDx_vs_Sub_x1PDx |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| ASV0145 | Aggregatibacter | sp.HMT458 | unclassified | | | | | | | | | | | | |
| ASV0331 | Aggregatibacter | sp.HMT458 | unclassified | | | | | | | | | | | | |
| ASV0126 | Capnocytophaga | leadbetteri | unclassified | | | | | | | | | | | | |
| ASV0140 | Capnocytophaga | leadbetteri | unclassified | | | | | | | | | | | | |
| ASV0133 | Cupriavidus | gilardii | unclassified | | | | | | | | | | | | |
| ASV1234 | Cupriavidus | gilardii | unclassified | | | | | | | | | | | | |
| ASV2179 | Cupriavidus | gilardii | unclassified | | | | | | | | | | | | |
| ASV0026 | Gemella | haemolysans | unclassified | | | | | | | C | C | C | | | |
| ASV0724 | Gemella | haemolysans | unclassified | | | | | | | | | | | | |
| ASV0013 | Granulicatella | adiacens | unclassified | C | | | | | | | C | C | | C | |
| ASV0898 | Granulicatella | adiacens | unclassified | | | | | | | | | | | | |
| ASV0476 | Lachnospiraceae [G-3] | bacterium_HMT100 | unclassified | | | | | | | | | | | | |
| ASV0512 | Lachnospiraceae [G-3] | bacterium_HMT100 | unclassified | | | | | | | | | | | | |
| ASV0389 | Leptotrichia | goodfellowii | BTASV213085 | | | | | | | | | | | | |
| ASV0533 | Leptotrichia | sp.HMT212 | unclassified | | | | | | | | | | | | |
| ASV0526 | Leptotrichia | sp.HMT392 | unclassified | | | | | | | | | | | | |
| ASV0117 | Oribacterium | sinus | BTASV107685 | | | | | | | | | | | | |
| ASV0218 | Porphyromonas | sp.HMT275 | BTASV079830 | | | | | | | | | | | | |
| ASV0007 | Prevotella | melaninogenica | BTASV111236 | | | | | | | | C | | | | |
| ASV0514 | Prevotella | melaninogenica | BTASV111262 | | | | | | | | | | | | |
| ASV0588 | Prevotella | sp.HMT472 | unclassified | | | | | | | | | | | | |
| ASV0717 | Prevotella | sp.HMT472 | unclassified | | | | | | | | | | | | |
| ASV0829 | Prevotella | sp.HMT472 | unclassified | | | | | | | | | | | | |
| ASV0108 | Pseudomonas | fluorescens | unclassified | | | | | | | | | | | | |
| ASV0610 | Pseudomonas | fluorescens | unclassified | | | | | | | | | | | | |
| ASV0429 | Pseudopropionibacterium | propionicum | unclassified | | | | | | | | | | | | |
| ASV0961 | Pseudopropionibacterium | propionicum | unclassified | | | | | | | | | | | | |
| ASV1055 | Sphingomonas | echinoides | unclassified | | | | | | | | | | | | |
| ASV0072 | Veillonella | rogosae | unclassified | | | | | | | | | | | | |
| ASV0084 | Veillonella | rogosae | unclassified | | | | | | | | | | | | |
| ASV0485 | Veillonella | sp.HMT780 | unclassified | | | | | | | | | | | | |

Cells are coloured according to the periodontal health condition predicted by the taxon in question. The green colour was associated with periodontally healthy subjects, healthy sites; pink with periodontitis, healthy sites; and orange with periodontitis, diseased sites after therapy.
ASV= amplicon sequence variant; ASVid= amplicon sequence variant identifier; C= core member; Sub_x0HHx= subgingival plaque of periodontally healthy subjects, healthy sites; Sub_x0PHx= subgingival plaque of periodontitis subjects, healthy sites; Sub_x0PDx= subgingival plaque of periodontitis subjects, diseased sites; Sub_x1PDx= subgingival plaque of periodontitis subjects, diseased sites after therapy; Sup_x0GDx= supragingival plaque of gingivitis subjects, diseased sites; Sup_x0HHx= supragingival plaque of periodontally healthy subjects, healthy sites; Sup_x0PDx= supragingival plaque of periodontitis subjects, diseased sites; Sup_x1PDx= supragingival plaque of periodontitis subjects, diseased sites after therapy.

The subgingival plaque ASVs of the genera *Actinomyces*, *Anaerolineae* [G-1], *Bacteroidaceae* [G-1], *Bacteroidetes* [G-3], *Catonella*, *Desulfobulbus*, *Dialister*, *Fretibacterium*, *Mogibacterium*, *Mycoplasma*, *Olsenella*, *Peptostreptococcaceae* [XI][G-2], *Peptostreptococcaceae* [XI][G-5], *Peptostreptococcaceae* [XI][G-6], *Porphyromonas*, *Prevotella*, *Pseudoramibacter*, *Stomatobaculum*, *Treponema*, and *Veillonellaceae* [G-1] acted as predictors of periodontitis. A focus on the relative abundance values observed in Sub_x0PDx

revealed the most relevant to be *Peptostreptococcaceae* [XI][G-5] *saphenum* ASV129 (0.11%), *Dialister pneumosintes* ASV194 (0.05%), *Desulfobulbus* HMT041 ASV149 (0.03%), and *Mogibacterium timidum* ASV640 (0.03%). A comparison of both models to periodontal health highlighted that the first ASV similarly had a role in identifying treated and healthy sites of periodontitis, while the second and third ASVs were also predictors of treated periodontitis (Table 13).

**Table 13. Main taxa predictive of periodontitis in subgingival plaque.**

| ASVid | Genus | Species | ASV | Sup_x0HHx_vs_Sup_x0GDx | Sup_x0HHx_vs_Sup_x0PDx | Sup_x0HHx_vs_Sup_x1PDx | Sup_x0GDx_vs_Sup_x0PDx | Sup_x0GDx_vs_Sup_x1PDx | Sup_x0PDx_vs_Sup_x1PDx | Sub_x0HHx_vs_Sub_x0PHx | Sub_x0HHx_vs_Sub_x0PDx | Sub_x0HHx_vs_Sub_x1PDx | Sub_x0PHx_vs_Sub_x0PDx | Sub_x0PHx_vs_Sub_x1PDx | Sub_x0PDx_vs_Sub_x1PDx |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| AV02271 | Actinomyces | israelii | BTASV162329 | | | | | | | | | | | | |
| AV00291 | Anaerolineae [G-1] | bacterium_HMT439 | BTASV107527 | | | | | | | | | | | | |
| AV00242 | Bacteroidaceae [G-1] | bacterium_HMT272 | BTASV079645 | | | | | | | | | | | | |
| AV00452 | Bacteroidetes [G-3] | bacterium_HMT280 | BTASV080885 | | | | | | | | | | | | |
| AV00563 | Catonella | sp.HMT451 | BTASV107675 | | | | | | | | | | | | |
| AV00149 | Desulfobulbus | sp.HMT041 | unclassified | | | | | | | | | | | | |
| AV00194 | Dialister | pneumosintes | unclassified | | | | | | | | | | | | |
| AV00328 | Fretibacterium | sp.HMT358 | BTASV096156 | | | | | | | | | | | | |
| AV00300 | Fretibacterium | sp.HMT360 | unclassified | | | | | | | | | | | | |
| AV00640 | Mogibacterium | timidum | BTASV011251 | | | | | | | | | | | | |
| AV00213 | Mycoplasma | faucium | unclassified | | | | | | | | | | | | |
| AV00670 | Olsenella | uli | BTASV011234 | | | | | | | | | | | | |
| AV01263 | Peptostreptococcaceae [XI][G-2] | bacterium_HMT091 | BTASV033613 | | | | | | | | | | | | |
| AV00129 | Peptostreptococcaceae [XI][G-5] | saphenum | unclassified | | | | | | | | | | | | |
| AV00793 | Peptostreptococcaceae [XI][G-6] | minutum | BTASV167025 | | | | | | | | | | | | |
| AV00069 | Porphyromonas | endodontalis | unclassified | | | | | | | | | | | | |
| AV00395 | Porphyromonas | endodontalis | unclassified | | | | | | | | | | | | |
| AV00692 | Porphyromonas | endodontalis | unclassified | | | | | | | | | | | | |
| AV00945 | Prevotella | baroniae | BTASV128220 | | | | | | | | | | | | |
| AV00217 | Prevotella | sp.HMT304 | unclassified | | | | | | | | | | | | |
| AV00550 | Pseudoramibacter | alactolyticus | BTASV124199 | | | | | | | | | | | | |
| AV00766 | Stomatobaculum | sp.HMT373 | unclassified | | | | | | | | | | | | |
| AV00329 | Treponema | maltophilum | unclassified | | | | | | | | | | | | |
| AV00427 | Treponema | maltophilum | unclassified | | | | | | | | | | | | |
| AV00583 | Treponema | maltophilum | unclassified | | | | | | | | | | | | |
| AV00861 | Treponema | parvum | BTASV182612 | | | | | | | | | | | | |
| AV00416 | Treponema | socranskii | unclassified | | | | | | | | | | | | |
| AV00422 | Treponema | socranskii | unclassified | | | | | | | | | | | | |
| AV01208 | Treponema | sp.HMT256 | BTASV077288 | | | | | | | | | | | | |
| AV00520 | Treponema | sp.HMT258 | BTASV077301 | | | | | | | | | | | | |
| AV00974 | Veillonellaceae [G-1] | bacterium_HMT145 | unclassified | | | | | | | | | | | | |

Cells are coloured according to the periodontal health condition predicted by the taxon in question. The pink colour was associated with periodontitis, healthy sites; red with periodontitis, diseased sites; and orange with periodontitis, diseased sites after therapy.
ASV= amplicon sequence variant; ASVid= amplicon sequence variant identifier; Sub_x0HHx= subgingival plaque of periodontally healthy subjects, healthy sites; Sub_x0PHx= subgingival plaque of periodontitis subjects, healthy sites; Sub_x0PDx= subgingival plaque of periodontitis subjects, diseased sites; Sub_x1PDx= subgingival plaque of periodontitis subjects, diseased sites after therapy; Sup_x0GDx= supragingival plaque of gingivitis subjects, diseased sites; Sup_x0HHx= supragingival plaque of periodontally healthy subjects, healthy sites; Sup_x0PDx= supragingival plaque of periodontitis subjects, diseased sites; Sup_x1PDx= supragingival plaque of periodontitis subjects, diseased sites after therapy.

*F. nucleatum* subsp. *vincentii* was the main species with a capacity to distinguish contrasting conditions. The highly abundant core ASV10 (0.86%) was a strong predictor of periodontitis, as was also the case for several other models. However, further, less abundant, ASVs (>0.04%), including ASV77, ASV200, and ASV204, simultaneously predicted both periodontal health and periodontitis (Table 14).

**Table 14. Main taxa predictive of periodontal health and periodontitis in subgingival plaque.**

| ASVid | Genus | Species | ASV | Sup_x0HHx_vs_Sup_x0GDx | Sup_x0HHx_vs_Sup_x0PDx | Sup_x0HHx_vs_Sup_x1PDx | Sup_x0GDx_vs_Sup_x0PDx | Sup_x0GDx_vs_Sup_x1PDx | Sup_x0PDx_vs_Sup_x1PDx | Sub_x0HHx_vs_Sub_x0PHx | Sub_x0HHx_vs_Sub_x0PDx | Sub_x0HHx_vs_Sub_x1PDx | Sub_x0PHx_vs_Sub_x0PDx | Sub_x0PHx_vs_Sub_x1PDx | Sub_x0PDx_vs_Sub_x1PDx |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| AV00335 | Actinomyces | gerencseriae | unclassified | | | | | | | | | | | | |
| AV00010 | Fusobacterium | nucleatum_subsp.vincentii | unclassified | | C | C | | | | C | C | C | | C | C |
| AV00012 | Fusobacterium | nucleatum_subsp.vincentii | unclassified | C | | | | | | | | | | | |
| AV00047 | Fusobacterium | nucleatum_subsp.vincentii | unclassified | C | | | | | | | | | | | |
| AV00077 | Fusobacterium | nucleatum_subsp.vincentii | unclassified | | | | | | | | | | | | |
| AV00098 | Fusobacterium | nucleatum_subsp.vincentii | unclassified | | | | | | | C | | C | | | |
| AV00130 | Fusobacterium | nucleatum_subsp.vincentii | unclassified | C | | | | | | | | | | | |
| AV00192 | Fusobacterium | nucleatum_subsp.vincentii | unclassified | | | | | | | | | | | | |
| AV00200 | Fusobacterium | nucleatum_subsp.vincentii | unclassified | C | | | | | | | C | | | | C |
| AV00204 | Fusobacterium | nucleatum_subsp.vincentii | unclassified | | | | | | | | | | | | |
| AV00415 | Fusobacterium | nucleatum_subsp.vincentii | unclassified | | | | | | | | | | | | |
| AV00547 | Fusobacterium | nucleatum_subsp.vincentii | unclassified | | | | | | | | | | | | |
| AV00573 | Fusobacterium | nucleatum_subsp.vincentii | unclassified | | | | | | | | | | | | |
| AV01540 | Fusobacterium | nucleatum_subsp.vincentii | unclassified | | | | | | | | | | | | |

Cells are coloured according to the periodontal health condition predicted by the taxon in question. The green colour was associated with periodontally healthy subjects, healthy sites; yellow with gingivitis, diseased sites; pink with periodontitis, healthy sites; red with periodontitis, diseased sites; and orange with periodontitis, diseased sites after therapy.
ASV= amplicon sequence variant; ASVid= amplicon sequence variant identifier; C= core member; Sub_x0HHx= subgingival plaque of periodontally healthy subjects, healthy sites; Sub_x0PHx= subgingival plaque of periodontitis subjects, healthy sites; Sub_x0PDx= subgingival plaque of periodontitis subjects, diseased sites; Sub_x1PDx= subgingival plaque of periodontitis subjects, diseased sites after therapy; Sup_x0GDx= supragingival plaque of gingivitis subjects, diseased sites; Sup_x0HHx= supragingival plaque of periodontally healthy subjects, healthy sites; Sup_x0PDx= supragingival plaque of periodontitis subjects, diseased sites; Sup_x1PDx= supragingival plaque of periodontitis subjects, diseased sites after therapy.

*4.4.10.3. Supragingival and subgingival microbiota*

The ASVs belonging to the genera *Actinomyces*, *Cardiobacterium*, *Corynebacterium*, *Escherichia*, *Haemophilus*, *Kingella*, *Leptotrichia*, *Neisseria*, *Prevotella*, *Rothia*, and *Streptococcus* were strong predictors of periodontal health in both plaque types. Of these, there were high relative abundances in the supragingival or subgingival plaque, or both, in the following species (note: abbreviated to "supra" and "sub", in that order):1) *R. dentocariosa* ASV2 (core in both plaque types: "supra" 5.56%, "sub" 2.04%); 2) several ASVs of *Haemophilus parainfluenzae* (ASV3 - core in the "supra" and "sub" plaque: 4.51% and 1.84%; ASV78: 0.80% and 0.23%; ASV45: 0.71% and 0.29%; and ASV46: 0.25% and 0.24%); 3) *Kingella oralis* ASV66 (core in both the "supra" and "sub" types: 0.89% and 0.38%); 4) *Streptococcus vestibularis* ASV27 (0.80% and 0.11%); and 5) *Actinomyces* HMT170 ASV119 (0.61% and 0.05%). Some of these ASVs, specifically *R. dentocariosa* ASV2, *S. vestibularis* ASV27, and *A.* HMT170 ASV119, also behaved as strong predictors of healthy sites of periodontitis (Table 15).

**Table 15. Main taxa predictive of periodontal health in supragingival and subgingival plaques.**

| ASVid | Genus | Species | ASV | Sup_x0HHx_vs_Sup_x0GDx | Sup_x0HHx_vs_Sup_x0PDx | Sup_x0HHx_vs_Sup_x1PDx | Sup_x0GDx_vs_Sup_x0PDx | Sup_x0GDx_vs_Sup_x1PDx | Sup_x0PDx_vs_Sup_x1PDx | Sub_x0HHx_vs_Sub_x0PHx | Sub_x0HHx_vs_Sub_x0PDx | Sub_x0HHx_vs_Sub_x1PDx | Sub_x0PHx_vs_Sub_x0PDx | Sub_x0PHx_vs_Sub_x1PDx | Sub_x0PDx_vs_Sub_x1PDx |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| AV00206 | Actinomyces | massiliensis | BTASV213424 | C | C | C | | | | | | | | | |
| AV00119 | Actinomyces | sp.HMT170 | BTASV057334 | | | | | | | | | | | | |
| AV00103 | Cardiobacterium | hominis | BTASV158475 | C | | | | | | | | | | | |
| AV00595 | Cardiobacterium | hominis | BTASV158478 | | | | | | | | | | | | |
| AV01577 | Cardiobacterium | hominis | unclassified | | | | | | | | | | | | |
| AV00102 | Corynebacterium | durum | BTASV140751 | C | | | | | | | | | | | |
| AV00199 | Corynebacterium | durum | unclassified | C | C | C | | | | | | | | | |
| AV00202 | Corynebacterium | durum | unclassified | | | | | | | | | | | | |
| AV00116 | Escherichia | coli | BTASV133996 | | | | | | | | | | | | |
| AV00003 | Haemophilus | parainfluenzae | unclassified | C | | | | | | C | C | C | | | C |
| AV00045 | Haemophilus | parainfluenzae | unclassified | | | | | | | | | | | | |
| AV00046 | Haemophilus | parainfluenzae | unclassified | | | | | | | | | | | | |
| AV00078 | Haemophilus | parainfluenzae | unclassified | | | | | | | | | | | | |
| AV00094 | Haemophilus | parainfluenzae | unclassified | | | | | | | | | | | | |
| AV00109 | Haemophilus | parainfluenzae | unclassified | | | | | | | | | | | | |
| AV00226 | Haemophilus | parainfluenzae | unclassified | | | | | | | | | | | | |
| AV00552 | Haemophilus | parainfluenzae | unclassified | | | | | | | | | | | | |
| AV00233 | Haemophilus | sp.HMT036 | BTASV010962 | | | | | | | | | | | | |
| AV00268 | Haemophilus | sp.HMT036 | BTASV010999 | | | | | | | | | | | | |
| AV00066 | Kingella | oralis | BTASV175208 | C | C | | | | | | C | | | | C |
| AV01319 | Kingella | oralis | unclassified | | | | | | | | | | | | |
| AV00139 | Leptotrichia | hofstadii | unclassified | | | | | | | | | | | | |
| AV01562 | Leptotrichia | hofstadii | BTASV075636 | | | | | | | | | | | | |
| AV00173 | Neisseria | oralis | BTASV002194 | | | | | | | | | | | | |
| AV00018 | Neisseria | perflava | unclassified | | | | | | | | | | | | |
| AV00035 | Neisseria | perflava | BTASV036108 | | | | | | | | | | | | |
| AV02105 | Neisseria | perflava | unclassified | | | | | | | | | | | | |
| AV00267 | Prevotella | oralis | BTASV175162 | | | | | | | | | | | | |
| AV00002 | Rothia | dentocariosa | BTASV138915 | C | C | | | | | | | C | C | C | |
| AV00886 | Rothia | dentocariosa | unclassified | | | | | | | | | | | | |
| AV01258 | Rothia | dentocariosa | unclassified | | | | | | | | | | | | |
| AV00027 | Streptococcus | vestibularis | BTASV004875 | | | | | | | C | | | C | C | |

Cells are coloured according to the periodontal health condition predicted by the taxon in question. The green colour was associated with periodontally healthy subjects, healthy sites; pink with periodontitis, healthy sites; and orange with periodontitis, diseased sites after therapy.
ASV= amplicon sequence variant; ASVid= amplicon sequence variant identifier; C= core member; Sub_x0HHx= subgingival plaque of periodontally healthy subjects, healthy sites; Sub_x0PHx= subgingival plaque of periodontitis subjects, healthy sites; Sub_x0PDx= subgingival plaque of periodontitis subjects, diseased sites; Sub_x1PDx= subgingival plaque of periodontitis subjects, diseased sites after therapy; Sup_x0GDx= supragingival plaque of gingivitis subjects, diseased sites; Sup_x0HHx= supragingival plaque of periodontally healthy subjects, healthy sites; Sup_x0PDx= supragingival plaque of periodontitis subjects, diseased sites; Sup_x1PDx= supragingival plaque of periodontitis subjects, diseased sites after therapy.

The ASVs belonging to the genera *Filifactor*, *Fretibacterium*, *Lachnospiraceae* [G-8], *Peptostreptococcaceae* [XI][G-4], *Peptostreptococcaceae* [XI][G-6], *Prevotella*, *Streptococcus*, *Tannerella*, and *Treponema* were strong predictors of periodontitis in both the supragingival and subgingival plaque. The abundance values of the following taxa deserve special attention: 1) *T. forsythia* ASV15 (core in all the plaque models where it acted as a predictor of periodontitis: 0.35% and 1.85%); 2) *Filifactor alocis* ASV19 (core in some "sub"

models: 0.38% and 1.33%); 3) two ASVs of *Treponema denticola* (ASV38: 0.21% and 0.84%; ASV150: 0.07% and 0.31%); 4) *Fretibacterium fastidiosum* ASV97 (core in the models where it acted as a periodontitis predictor: 0.08% and 0.46%): 5) *Peptostreptococcaceae* [XI][G-4] HMT369 ASV124 (core in the models where it acted as a periodontitis predictor: 0.09% and 0.34%); 6) *Streptococcus anginosus* ASV142 (0.21% and 0.20%); and 7) *Peptostreptococcaceae* [XI][G-6] *nodatum* ASV189 (0.11% and 0.18%). *F. fastidiosum* ASV97 and *S. anginosus* ASV142 were also predictors of gingivitis in the supragingival plaque samples. All of these taxa similarly had a capacity to identify treated periodontitis in the subgingival plaque, with *T. forsythia* ASV15 also being able to do so in the supragingival samples (Table 16).

**Table 16. Main taxa predictive of periodontitis in supragingival and subgingival plaques.**

| ASVid | Genus | Species | ASV | Sup_x0HHx_vs_Sup_x0GDx | Sup_x0HHx_vs_Sup_x0PDx | Sup_x0HHx_vs_Sup_x1PDx | Sup_x0GDx_vs_Sup_x0PDx | Sup_x0GDx_vs_Sup_x1PDx | Sup_x0PDx_vs_Sup_x1PDx | Sub_x0HHx_vs_Sub_x0PHx | Sub_x0HHx_vs_Sub_x0PDx | Sub_x0HHx_vs_Sub_x1PDx | Sub_x0PHx_vs_Sub_x0PDx | Sub_x0PHx_vs_Sub_x1PDx | Sub_x0PDx_vs_Sub_x1PDx |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| AV00019 | Filifactor | alocis | BTASV124203 | | | | | | | | | | | | C |
| AV00097 | Fretibacterium | fastidiosum | Unclassified | | C | | | | | | C | | | | |
| AV00278 | Lachnospiraceae [G-8] | bacterium_HMT500 | Unclassified | | | | | | | | | | | | |
| AV00503 | Lachnospiraceae [G-8] | bacterium_HMT500 | BTASV114386 | | | | | | | | | | | | |
| AV00124 | Peptostreptococcaceae [XI][G-4] | bacterium_HMT369 | BTASV096563 | | C | | | | | | | | | | C |
| AV00189 | Peptostreptococcaceae [XI][G-6] | nodatum | BTASV174812 | | | | | | | | | | | | |
| AV00308 | Prevotella | denticola | Unclassified | | | | | | | | | | | | |
| AV00142 | Streptococcus | anginosus | Unclassified | | | | | | | | | | | | |
| AV00015 | Tannerella | forsythia | BTASV153103 | | C | C | C | C | | | C | | C | | C |
| AV01651 | Tannerella | forsythia | BTASV153103 | | | | | | | | | | | | |
| AV00038 | Treponema | denticola | BTASV138814 | | | | | | | | | | | | |
| AV00150 | Treponema | denticola | Unclassified | | | | | | | | | | | | |
| AV00184 | Treponema | denticola | Unclassified | | | | | | | | | | | | |
| AV00269 | Treponema | denticola | Unclassified | | | | | | | | | | | | |
| AV00344 | Treponema | denticola | BTASV138816 | | | | | | | | | | | | |

Cells are coloured according to the periodontal health condition predicted by the taxon in question. The yellow colour was associated with gingivitis, diseased sites; pink with periodontitis, healthy sites; red with periodontitis, diseased sites; and orange with periodontitis, diseased sites after therapy.
ASV= amplicon sequence variant; ASVid= amplicon sequence variant identifier; C= core member; Sub_x0HHx= subgingival plaque of periodontally healthy subjects, healthy sites; Sub_x0PHx= subgingival plaque of periodontitis subjects, healthy sites; Sub_x0PDx= subgingival plaque of periodontitis subjects, diseased sites; Sub_x1PDx= subgingival plaque of periodontitis subjects, diseased sites after therapy; Sup_x0GDx= supragingival plaque of gingivitis subjects, diseased sites; Sup_x0HHx= supragingival plaque of periodontally healthy subjects, healthy sites; Sup_x0PDx= supragingival plaque of periodontitis subjects, diseased sites; Sup_x1PDx= supragingival plaque of periodontitis subjects, diseased sites after therapy.

The ASVs of the genera *Alloprevotella*, *Fusobacterium*, *Gemella*, *Granulicatella*, *Lachnoanaerobaculum*, and *Ruminococcaceae* [G-1] had properties that enabled them to distinguish different clinical conditions in both plaque types: "supra" - gingivitis; “sub” - periodontal health. Examples are: *Fusobacterium periodonticum* ASV11 (core in all the models: "supra" 0.3%; "sub"0.24%); *Lachnoanaerobaculum umeaense* ASV152 (core in all the models: 0.33% and 0.17%); and *Granulicatella elegans* ASV207 (0.17% and 0.20%) (Table 17).

**Table 17. Main taxa predictive of periodontal health and periodontitis in supragingival and subgingival plaques.**

| ASVid | Genus | Species | ASV | Sup_x0HHx_vs_Sup_x0GDx | Sup_x0HHx_vs_Sup_x0PDx | Sup_x0HHx_vs_Sup_x1PDx | Sup_x0GDx_vs_Sup_x0PDx | Sup_x0GDx_vs_Sup_x1PDx | Sup_x0PDx_vs_Sup_x1PDx | Sub_x0HHx_vs_Sub_x0PHx | Sub_x0HHx_vs_Sub_x0PDx | Sub_x0HHx_vs_Sub_x1PDx | Sub_x0PHx_vs_Sub_x0PDx | Sub_x0PHx_vs_Sub_x1PDx | Sub_x0PDx_vs_Sub_x1PDx |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| AV00327 | Alloprevotella | sp.HMT473 | unclassified | | | | | | | | | | | | |
| AV00978 | Alloprevotella | sp.HMT473 | unclassified | | | | | | | | | | | | |
| AV00011 | Fusobacterium | periodonticum | BTASV066878 | | | | | C | | C | | C | | | |
| AV00188 | Fusobacterium | periodonticum | unclassified | | | | | | | | | | | | |
| AV00499 | Fusobacterium | periodonticum | unclassified | | | | | | | | | | | | |
| AV00118 | Gemella | morbillorum | BTASV011336 | | | | | | | C | | | | | |
| AV00136 | Gemella | morbillorum | unclassified | | | | | | | | | | | | |
| AV00207 | Granulicatella | elegans | unclassified | | | | | | | | | | | | |
| AV00152 | Lachnoanaerobaculum | umeaense | unclassified | | | | | C | | C | | C | | | |
| AV00081 | Ruminococcaceae [G-1] | bacterium_HMT075 | unclassified | | | | | | | | | | | | |

Cells are coloured according to the periodontal health condition predicted by the taxon in question. The green colour was associated with periodontally healthy subjects, healthy sites; yellow with gingivitis, diseased sites; and orange with periodontitis, diseased sites after therapy.
ASV= amplicon sequence variant; ASVid= amplicon sequence variant identifier; C= core member; Sub_x0HHx= subgingival plaque of periodontally healthy subjects, healthy sites; Sub_x0PHx= subgingival plaque of periodontitis subjects, healthy sites; Sub_x0PDx= subgingival plaque of periodontitis subjects, diseased sites; Sub_x1PDx= subgingival plaque of periodontitis subjects, diseased sites after therapy; Sup_x0GDx= supragingival plaque of gingivitis subjects, diseased sites; Sup_x0HHx= supragingival plaque of periodontally healthy subjects, healthy sites; Sup_x0PDx= supragingival plaque of periodontitis subjects, diseased sites; Sup_x1PDx= supragingival plaque of periodontitis subjects, diseased sites after therapy.

There were numerous biological inconsistencies in relation to the correspondence between the results obtained with the main predictive ASV models and their respective differential abundances. In supragingival plaque, 17 of the 21 ASVs that predicted health had differential abundances, with 88.23% (15/17) of them showing significantly elevated abundances in both health and periodontitis. In the 25 disease-predictor ASVs, 22 were differentially abundant and, of these, four (18.18%) had a tendency to high levels in the two opposing conditions.

In the subgingival plaque, 29 of the 31 health-predictor ASVs were differentially abundant, with 31.03% (9/29) showing significantly elevated abundances in both health and periodontitis. In the 31 disease-predictor ASVs, 30 exhibited differential abundances and, of these, three (10.00%) presented significantly elevated levels in the two opposing conditions.

In both plaque types, 28 of the 32 health-predictor ASVs showed differential abundance and, of these, 71.42% (20/28) were differentially abundant for both health and periodontitis. The 15 disease-predictor ASVs in both plaques were also differentially abundant and, of these, three (20.0%) had significantly elevated levels in both clinical conditions.

### 4.5. DISCUSSION

Methodological differences mean it is controversial to compare the results of 16S rRNA gene sequencing-based studies of the periodontal microbiome (13-15). Moreover, the small sample sizes typically analysed make it difficult to understand and distinguish any health or disease conditions, since individual variations can be confused with real biological differences, introducing bias into the conclusions presented (80,81).

Aware of these limitations, and in an attempt to minimise methodological-associated preconceptions, some authors have reanalysed sequences taken from distinct investigations using the same bioinformatics protocol and statistical approach (80-83). In general, these works have mainly examined subgingival microbiota profiles in states of health and/or periodontal disease (80-83).

The samples analysed in such research comprised between 273 (82) and 943 specimens (81), with sequences from distinct sequencing platforms (81-83) and/or hypervariable regions (80,81,83) evaluated together, combining up to six gene zones (80,81). If specified in these studies, the amplicon clustering step involved the employment of OTUs with a sequence similarity of 97% (81) or 100% (83) and taxonomy was assigned using the human oral microbiome database (HOMD) (82,84), Greengenes (81,83,85) or the ribosomal database project (RDP) (80,86). Surprisingly, the diversity, co-occurrence, and predictivity results of these re-analyses are mostly expressed at the genus level (80,81,83), with only Abusleme et al. (82) providing species-level descriptions.

Nevertheless, the joint assessment of data derived from distinct platforms and gene zones is problematic, with discrepancies in both the performances of the sequencing technologies (13) and the detection of oral species using primers that target different regions as observed in objective 1 (16). The OTU algorithms adopted have been shown to be less sensitive, specific, and precise, with highly spurious sequence rates, than ASV-based pipelines (18,19). Furthermore, as demonstrated in the objective 3 (17), OTU clustering at the commonly used 97% similarity threshold is very inaccurate, given the high percentage of different oral species that nonetheless share amplicon similarities above that level.

Additionally, phylogenetically diverse databases like Greengenes (85) or the RDP (86) contain 16S rRNA gene sequences that are taxonomically misannotated (87), as well as different levels of representation of each included environment (88). Their use in taxonomic assignments can thus produce classification errors (87) and lead to substantial variations in the quality of any categorisations (88). Furthermore, descriptions of sequencing-derived results at the genus level may not be particularly valuable, since different species from the same genus are associated with contradictory oral conditions (89-91). Identification at, at least, the species level is therefore desirable in sequencing-based studies of the oral microbiota (21).

Our comprehensive review of the literature, with >1000 articles evaluated, ensured the obtention of all the available evidence of interest. Datasets from 25 Illumina 3-4 gene region bioprojects were merged and we analysed >2000 samples. These were then assigned to four supra- and four subgingival plaque groups that represented distinct clinical conditions. The evaluation of a broad spectrum of patients with different conditions and degrees of disease severity is desirable in predictive analyses since studies that include healthy controls and severe cases tend to overestimate diagnostic performance (92,93).

Consequently, in view of the issues raised above, the meta-omics approach of the present study has produced the strongest evidence to date on the diversity, co-occurrence patterns, and predictive capacity of the supra- and subgingival plaque microbiota present in different periodontal statuses. All the sequences were processed under a strict and unified bioinformatics protocol that involved the use of ASVs and an oral-specific database that was employed for the purpose of taxonomic assignment (21). Our results, which are described at both the species and ASV levels, are only compared to those derived from other meta-omics analyses. However, this is the first meta-omics study reporting supragingival plaque microbial profiles for different periodontal conditions, so our findings in this niche could not be compared with those of other studies.



### 4.5.1. Quality assessment of metadata and sequences

Although Cai et al. (81) assessed the risk of bias of the articles included in their investigation using the Downs-Black checklist (94) for randomized and nonrandomized comparative studies; to the best of our knowledge, our investigation is the first to propose a

specific checklist with which to assess the quality of the metadata obtained from sequencing-based examinations of the periodontal microbiota. Such a list can also serve as a guide to the variables that should be included in metadata tables uploaded to a sequence's repositories. In fact, the authors of 10 of the 32 articles (~31%) that met our inclusion criteria had to be contacted to obtain the metadata or for clarification purposes. Half of these 10 studies were then ultimately excluded because either there was no metadata table at all or the one available contained incompatible or inadequate information.

The metadata assessment revealed that 88% of the bioprojects included in the meta-omics analysis were of a low or medium quality. In particular, they lacked relevant information such as the periodontitis type and severity or the total and sampling site clinical parameters. In addition, we often had to review an entire paper to identify either the meaning of the codes used in the table uploaded to the database or general phrases that could be applied to all of the study's subjects. Based on our experience in this study, we consider that a higher quality of metadata tables uploaded to repositories must be required to facilitate reproducible research and meta-omics approaches; without quality descriptive information, stored sequences are of little use.

The strict quality filter we applied to the 16S rRNA gene sequences described in the selected bioprojects resulted in the elimination of approximately 50% of them, with very high-quality sequences remaining. On the other hand, 90% of the bioprojects had an average number of ≥10,000 sequences (range of sequence average per bioproject= 202,516 - 7,233). This high quality of the sequences and their high number per sample guaranteed the robustness of the results obtained.

#### 4.5.2. Alpha-diversity in supragingival and subgingival plaque microbiota

In supragingival plaque groups, the richness estimators decreased form health to gingivitis and then, increased in periodontitis. Although they diminished again after therapy, the values did not reach those in health. The diversity (Shannon) and evenness (Pielou) estimators demonstrated an upward trend from health to disease that continued even after treatment.



There is controversy in the literature regarding the alpha-diversity results of the subgingival plaque for the different periodontal conditions. Like the findings in this study, Meuric et al. (80)

observed less richness in their untreated periodontitis samples than in those that were healthy. However, other authors have described the opposite position (81,82). Indeed, we observed an increase in richness after periodontal therapy to levels that even exceeded those seen in the healthy samples. This occurred at the expense of non-abundant species, as the number of ASVs increased but the 95% coverage index did not rise to the same extent. On the other hand, we agree with the findings in studies by other authors which state that there are no differences in diversity between healthy and periodontitis samples (81,82). However, unlike such research, we did note an increase in evenness in diseased and treated sites compared to healthy ones. In contrast to the supragingival plaque samples, the richness and diversity of the subgingival plaque did not vary in the periodontitis-related groups. Consequently, in terms of alpha-diversity, only the subgingival periodontitis niche is stable in the healthy, diseased, and treated sites.

Our meta-omics analysis describes for the first time how supragingival plaque is richer and more diverse than subgingival plaque in relation to the same periodontal conditions. In addition, we found that the microbiota associated with periodontitis *vs.* health had contrasting levels of richness in the niches: 1) with non-treated periodontitis, these values increased in the supragingival plaque and decreased in the subgingival samples; and 2) for treated periodontitis, they decreased in the "supra" and increased in the "sub" samples. These differences can be explained by the different biological characteristics of the niches and the different repercussions of periodontal therapy on each of them.

### 4.5.3. Structure of the bacterial community in supragingival and subgingival plaque microbiota

The PERMANOVA analyses revealed differences in the structure of the bacterial community for all the pairwise comparisons of the distinct periodontal conditions seen in a particular plaque type, as well as for the same health statuses in the different plaques.



In general, other studies on the subgingival niche have also revealed differences in the bacterial community structure for periodontal health and non-treated periodontitis (80-82). As we observed herein, Sisk-Hackworth et al. (83), who used PERMANOVA to evaluate periodontitis samples that had undergone distinct periodontal therapies, also identified

differences between the treatment groups (before *vs.* after). However, their non-parametric multidimensional scaling (NMDS) analysis found no clustering by treatment, overall response, or pocket depth.

#### 4.5.4. Core microbiota in supragingival and subgingival plaque microbiota

Various definitions of the core microbiome can be found in the literature and, as a consequence, the associated findings are difficult to compare (82,95,96).

In this study, the prevalence of the supra- and subgingival plaque core microbiota in the different periodontal health statuses represented at most 2% of the detected ASVs; 10% of the species; and 50% of the total abundance. Accordingly, the greatest proportion of the plaque microbiota in the groups assessed was not part of the core and represented at least 50% of the total abundance. Moreover, in relation to the presence and total abundance, no clinical condition could be characterised by either a broad or a very specific core microbiota in the supra- or subgingival plaque.

These outcomes indicate that the presence of taxa is heterogeneous and subject to inter-individual variations, complicating any comparisons of different periodontal conditions. None of these discoveries have previously been reported in a meta-omics analysis and, as a consequence, constitute novel findings.

#### 4.5.5. Differential abundance in supragingival and subgingival plaque microbiota

This analysis describes the magnitude of the differences in the microbial composition of the clinical conditions or plaques examined in our investigation. In the supragingival plaque, the taxa with the greatest differential abundances were found among health *vs.* disease-related groups, and gingivitis *vs.* non-treated periodontitis. This represented at most, 17% of the detected ASVs (of which no more than 7% were core microbiota) and 50% of the species (15% core). In contrast, the lowest numbers were found for non- *vs.* treated periodontitis, involving 9% of the ASVs (3% core) and 28% of the species (10% core). As the latter groups were the most comparable in terms of differential abundance, this might indicate either the ineffectiveness of the treatment or that the two conditions remain very similar to each other with respect to other conditions, even after therapy.

The highest number of taxa with differential abundances in the subgingival niche were found for the health *vs.* non- and treated periodontitis groups, involving, at most, 14% of the ASVs (3% core) and 42% of the species (8% core). Conversely, the lowest numbers were identified in the healthy periodontitis sites *vs.* the non- and treated periodontitis sites, comprising, at most, 6% of the ASVs (2% core) and 22% of the species (5% core). It is thus contended that, as for microbial composition, the healthy periodontitis sites are more similar to the non- and treated sites than to health.

Lastly, the greatest differences between the two plaque types were observed for non-treated periodontitis, involving 27% of the ASVs (2% core) and 62% of the species (9% core); and the fewest were seen in the treated group, and comprised 4% of the ASVs (7% core) and 15% of the species (6% core). The supra- and subgingival niches in the non-treated periodontitis group differed more in terms of microbial composition than the healthy or periodontitis sites after therapy.

Overall, the two-by-two comparisons revealed that the core microbiota members did not include most of the taxa with differential abundances. Additionally, the results for the supragingival plaque were in line with those for the subgingival plaque, even though the former is more strongly affected by external factors like saliva or anti-plaque agents in toothpastes (97) which might influence its composition.

### 4.5.6. Co-occurrence networks in supragingival and subgingival plaque microbiota

Network analyses of microbe-microbe interactions have improved our knowledge of how they potentially intermingle in their ecosystem (98). The correlations identified are able to describe the tendencies of different species to co-occur in a variety of circumstances. In this sense, two species with a significantly positive correlation might indicate a shared preference for a particular environmental condition or a true ecological interaction. Conversely, negative correlations may signify competition for nutrients or differences in their physiological requirements that lead to them never occupying the same niche (99). Moreover, the identification of hubs or keystones, which are highly associated taxa in a microbiota, is one of the most useful functions of the co-occurrence network analysis (100).

Given the importance of the issue, other meta-omics studies have evaluated bacterial co-occurrence patterns in the subgingival microbiota of patients with distinct periodontal health statuses (81-83). Nevertheless, comparisons of their findings should be interpreted with caution, as they may be affected by methodological differences, including the correlation values used as cut-off points (101) or how the keystone taxa are defined (54). Besides, there may be an inversely proportional affect caused by the sample size, meaning that smaller samples would produce more correlations.

In the supragingival niche, the bacterial community that was organised in a positive-correlation network at the ASV level was present in a (low) similar proportion (~2.4%) in the health and non-treated periodontitis groups. However, in disease, there was a greater degree of interconnexion among the community members at the cost of any bi- or tri-directional relationships. Such a deduction can be made because the latter network had twice as many total modules as the healthy condition, but both had the same number of modules with >3 nodes.

The position was similar in the subgingival plaque, where a low proportion of the bacterial community at the ASV level was likewise organised into a positive-correlation network (~1.6%). However, unlike the supragingival networks, the topological characteristics of the subgingival versions were mainly affected in the non-treated periodontitis and (albeit to a lesser extent) treated groups. In contrast to what was observed in the healthy samples, these networks were characterised by lower numbers of both the connexions between their members and the interconnected bacterial clusters. This finding is in line with that of Sisk-Hackworth et al. (83), who reported that fewer internodal connections and less connectedness in the subgingival networks from the deeper pockets were, respectively, indicative of a more random or a less stable biofilm and an absence of taxa interdependence in the later stages of disease.

A comparison of the results obtained for the two niches, revealed that the non-treated periodontitis network in the subgingival plaque was present to a lesser extent and involved fewer nodes, interconnexions, and interconnected bacterial clusters (including those with >3 nodes) than in the supragingival plaque.

Furthermore, the description of the hub taxa in the meta-omics studies of the periodontal ecosystem that we examined was uncommon (82). Authors have found that their gingivitis and periodontitis-associated species *S. sputigena* and *Selenomonas noxia* acted as the main hubs in the healthy co-occurrence network, which could be an indicator of risk for dysbiosis (82).

In our network associated with periodontal health in supragingival plaque the main keystone ASVs were: *R. dentocariosa* ASV2, and *S. oralis* subsp. *dentisani* clade 058 ASV1.

In contrast, the principal keystone ASVs in the untreated periodontitis network in supragingival plaque was *S. sanguinis* ASV228; and in subgingival plaque it was *T. forsythia* ASV15. Moreover, the network associated with treated periodontitis in subgingival plaque obtained *T. forsythia* ASV15, *F. nucleatum* subsp. *vincentii* ASV10, and *S. oralis* subsp. *dentisani* clade 058 ASV1; as major keystones.

### 4.5.7. Predictive models in supragingival and subgingival plaque microbiota

The clinical metagenomic NGS is an emerging discipline consisting of the comprehensive analysis of microbial or host genetic material present within a clinical sample, with the purpose of recovering clinically relevant information that can drive the accurate diagnosis of infectious diseases (102,103). This approach has already been applied to a wide range of conditions such as respiratory, gastrointestinal, and bloodstream infections (102); and it has the potential to detect and contain disease outbreaks at an earlier stage than conventional diagnostic techniques (103). In the case of periodontitis, this could result in an improvement in the quality of life of patients suffering from this disease, with more teeth saved and cost savings. Ultimately, a precision diagnosis can advance precision medicine to personalise patient care (103).

In this regard, different tools have been developed that, based on data derived from omics techniques, allow the creation of predictive models to classify health conditions on the basis of the microbiota composition. Among them is the mixOmics package (45) used in this study, which constructs ROC curves that allow easy assessment of the degree of discrimination (i.e., AUC) of each health condition with respect to the others.

Our review of the literature identified only one meta-omics investigation that evaluated the predictive capacity of the subgingival microbiota for detecting the genera that distinguish between, on the one hand, the timing of and response to periodontal treatment and, on the other, the disease classes (A, maximum pocket depth= <6 mm *vs.* C= >8 mm) (83). In fact, to the best of our knowledge, our research is the first to assess the potential of supragingival and subgingival plaque for identifying species and ASVs that can be employed to distinguish distinct periodontal conditions. We used both the AUC value and the number of ASVs required to discriminate between two conditions to identify the best and worst predictive models.

A small proportion of the supragingival taxa, involving a maximum of 2% of the detected ASVs (32% core) and 17% of the species (38% core), had an outstanding (ROC≥0.9) or excellent (0.8≤ROC<0.9) ability (56) to distinguish between the clinical conditions evaluated. The best models were those that compared either health, gingivitis, or non- *vs.* treated periodontitis. These results suggest that periodontal therapy might induce a significant change in the supragingival plaque's microbiota, making it very different from the others, meaning that only a few predictor ASVs are required to distinguish between this and the rest of groups. It could also indicate the presence of ASVs inherent to the physiological condition that, despite the human intervention through treatment, are very defining of health, gingivitis, or untreated periodontitis. Conversely, the worst models were those that contrasted health *vs.* gingivitis and non-treated periodontitis. The fact that these two groups of diseased patients were made up of samples representing different condition severities may introduce greater variability and heterogeneity into the microbiota (104). This means that more predictor taxa would be required to differentiate between these diseased and healthy patients, even though they involve contrasting conditions.

Likewise, a small proportion of subgingival taxa, at most 3% of the detected ASVs (11% core) and 22% of species (12% core), had outstanding (ROC≥0.9) or excellent (0.8≤ROC<0.9) ability (56) to discriminate among periodontal conditions. The best model was that comparing healthy *vs.* diseased sites in periodontitis. Despite ~700 diseased samples did not have their same-mouth healthy "equivalent", most of the specimens from healthy sites were obtained from patients who also contributed diseased samples to our analysis. This may lead to greater control of the inter-subject variability (105), which could result in less predictors needed. In contrast,

the worst model was that contrasting health *vs.* non- and treated periodontitis. As occurred in the supragingival environment, the different degrees of severity within the non-treated group may have introduced variability and heterogeneity into the microbiota, possibly leading to a need for more predictor taxa. In the case of the treated group, the difficulties of access for plaque removal in the subgingival environment and the subsequent different responses to therapy may condition the microbiota diversity in different ways (106), meaning that more predictors would be needed to distinguish it from the healthy group.

A comparison of the results obtained for the two niches revealed that the supragingival plaque models performed better in terms of both AUC values and number of predictor ASVs in differentiating the periodontal health of untreated and treated periodontitis, as well as in distinguishing between the latter two groups. This may be due to the established disease-associated dysbiotic microbiome in subgingival plaque (107), with a higher number of potential predictor ASVs contributing to explain the biological condition than in supragingival plaque. Surprisingly, according to our results, for the development of clinical metagenomics in the near future, supragingival plaque samples would act as a better bacterial biomarker to discriminate different periodontal conditions than subgingival plaque samples. This would benefit practitioners in the sense of facilitating the sample collection process, which is much easier in the supragingival than in the subgingival niche.

Finally, in contrast to our observations in the differential abundance analyses, the core microbiota members generally represented a relevant percentage of the predictive ASVs and species, reaching values of up to 85% and 92%, respectively. More importantly, multivariate predictive modelling allowed in the present series a better understanding of the health and periodontitis associated taxa on each and both plaques than the univariate analysis of differential abundances. Despite the filters applied in the bioinformatics protocol, this univariate analysis is exposed to a higher number of incongruences, i.e. taxa that simultaneously had results in favour of health and periodontitis, which is difficult for the biological interpretation of the results. This was most strikingly observed in the ASVs that showed predictivity for health regardless of the niche, with ~64% of them having inconsistent results in favour of opposite conditions compared to ~16% of disease-predictor ASVs that showed results in favour of opposite conditions. Furthermore, the association of certain predictor taxa to health

or periodontitis allowed us to understand the role they played when predicting the healthy and treated sites in periodontitis.

### 4.5.8. Description of health and disease associated taxa in supragingival and subgingival plaque microbiota

Due to the greater robustness of the results obtained, the description of the oral species and ASVs associated with the different clinical conditions will be based on predictive modeling.

Our predictive modelling confirmed the previously reported associations in differential abundance analyses of *Capnocytophaga* (80-82), *Kingella* (80,81), *Neisseria* (80,81), and *Rothia* (80-82) to periodontal health; and of *Campylobacter* (80), *Catonella* (81), *Desulfobulbus* (81), *Tannerella* (80,81), and *Treponema* (80,81) to periodontitis. The analyses of Sisk-Hackworth et al. (83) found *Haemophilus* to be a predictor of post-treatment sites. Our investigation similarly identified this genus as a predictor of health, suggesting that its presence may be a positive indicator of the effectiveness of treatment.

However, our predictive modelling analyses could not confirm the associations identified between health and *Actinomyces* (80,82), *Corynebacterium* (80-82), *Gemella* (80,82), and *Veillonella* (80,81), and between periodontitis and *Aggregatibacter*, *Bacteroides*, and *Eikenella* (81); because different species from these genera were found to predict distinct health conditions. Moreover, although the analyses of Sisk-Hackworth et al. (83) revealed that *Prevotella* could identify pre-treatment samples and *Fusobacterium* and *Porphyromonas* worsening periodontitis after therapy and/or deeper pocket depths, we found that different species from these genera were predictors of distinct health and disease states. These observations thus highlight the importance of reporting the results of sequencing-based studies at, at least, the species level.

In our research, a series of ASVs from certain species were defined as predictors of periodontal health. In the supragingival samples, we highlighted due to their abundance: *L. hongkonensis* ASV24, *N. macacae* ASV9, *N. baciliformis* ASV281, *C. sputigena* ASV56, and *C. gingivalis* ASV93. In subgingival plaque, there were: *G. adiacens* ASV13, *G. haemosylans* ASV26, *C. leadbetteri* ASV126, *A.* HMT458 ASV145, and *P. melaninogenica* ASV7.

Moreover, *R. dentocariosa* ASV2, several ASVs from *H. parainfluenzae* (ASV3, ASV78, ASV45, and ASV46), *K. oralis* ASV66, *S. vestibularis* ASV27, and *A.* HMT170 ASV119; were predictors of periodontal health in the two dental plaque types.

Some of these health-predictive taxa were also found to be predictors of the healthy sites of periodontitis in subgingival plaque (alphabetically ordered): *A.* HMT170 ASV119, *G. adiacens* ASV13, *R. dentocariosa* ASV2, and *S. vestibularis* ASV27. Furthermore, *C. leadbetteri* ASV126 predicted the treated sites of periodontitis in the subgingival niche. The presence of health-predictive ASVs in the healthy and treated sites of periodontitis suggests they may have a positive or protective effect for health.

Of the taxa referred to above as health-related, *K. oralis* was found in a previous meta-omics analysis (82) to be associated with health (i.e., it was more abundant). In addition, via *in-vitro* investigations, *G. heamolysans* has been shown to inhibit the growth of the periodontopathogen *P. gingivalis* (108), while *H. parainfluenzae* has been associated with beneficial immunomodulatory effects, helping to maintain and regulate the healthy immune state of the host (109).

In contrast, a series of ASVs were defined as predictors of periodontal diseases. In supragingival plaque, *C. rectus* ASV20, *P.* HMT110 ASV21, *P. nigrescens* ASV25 and ASV86, and *S. sputigena* ASV243; were the most relevant considering their abundance in the predicted group, i.e, gingivitis. Of these, the two first ASVs also predicted periodontitis. In the subgingival plaque samples, the main periodontitis predictors were: *P. saphenum* ASV129, *D. pneumosintes* ASV194, *D.* HMT041 ASV149, and *M. timidum* ASV640. Besides, we found taxa which predicted periodontitis both in supragingival and subgingival plaque as: *T. forsythia* ASV15, *F. alocis* ASV19, *T. denticola* ASV38 and ASV150, *F. fastidiosum* ASV97, *P.* HMT369 ASV124, *S. anginosus* ASV142, and *P. nodatum* ASV189. Of these, *F. fastidiosum* ASV97 and *S. anginosus* ASV142 also acted as predictors of gingivitis in the supragingival niche.



Some of these disease-predictive taxa were also found to be predictors of the treated sites of periodontitis in supragingival (alphabetically ordered): *P.* HMT ASV21 and *P. nigrescens* ASV25; or in subgingival plaque: *D.* HMT041 ASV149, *D. pneumosintes* ASV194, *F. alocis*

ASV19, *F. fastidiosum* ASV97, *P.* HMT369 ASV124, *P. nodatum* ASV189, *P. saphenum* ASV129, *S. anginosus* ASV142, and *T. denticola* ASV38 and ASV150. The only taxa which predicted the treated sites of periodontitis in the two dental plaque types was *T. forsythia* ASV15. Moreover, *P. saphenum* ASV129 also acted as predictor of the healthy sites of periodontitis in subgingival plaque. Oppositely as observed in health, the presence of disease-predictive ASVs in the healthy and treated sites of periodontitis may suggest a negative effect or a risk of disease progression.

Of the above described disease-associated taxa, *C. rectus* and *S. putigena* have been reported to have, respectively, mechanisms such as a surface layer that can evade the human immune system (110), or components like the lipopolysaccharide, which is a virulence factor (111). Furthermore, *F. alocis*, *T. denticola*, and *T. forsythia* have been found as periodontitis-related in previous meta-omics analysis (82); and constitute widely known periodontopathogens which present with evasion strategies to disarm the neutrophil effector functions to maintain homeostasis in the oral cavity (112).

Interestingly, there were some species for which different ASVs predicted distinct periodontal health conditions. In supragingival plaque, four ASVs of *S. oralis* subsp. *dentisani* clade 058 and two of *V. dispar* predicted periodontal health; meanwhile other three and one ASVs, respectively, were disease-predictors. Moreover, even the same ASV could predict opposite clinical conditions. In the subgingival plaque samples, *F. nucleatum* subsp. *vincentii* ASV10 was a strong predictor of periodontitis; but other three ASVs from this species acted as predictors of periodontal health and periodontitis. Similarly, *F. periodonticum* ASV11, *L. umeanense* ASV152 and *G. elegans* ASV207 predicted gingivitis in supragingival plaque but were predictors of periodontal health in the subgingival niche.

Different reports have supported an association between *S. oralis* subsp. *dentisani* clade 058 and oral health, referring not only to its inhibitory activity over oral pathogens like *Aggregatibacter actinomycetemcomitans*, *F. nucleatum, Prevotella intermedia*, *Streptococcus mutans*, and *Streptococcus sobrinus* (113,114), but also to its ability to alkalinise the extracellular environment via the arginine deiminase system (113). In our analysis, the highly-abundant-core *S. oralis* subsp. *dentisani* clade 058 ASV1 together with ASV320, ASV560, and

ASV960 predicted periodontal health in supragingival plaque; but three other less abundant ASVs (ASV425, ASV877, and ASV1534; <0.07%) predicted periodontitis in the same niche. Moreover, former meta-omics reports have found *F. periodonticum* as gingivitis-associated and all subspecies from *F. nucleatum* as periodontitis-related (i.e. more abundant) (82); but we observed how these species predicted distinct conditions as health and gingivitis, or health and periodontitis, respectively. The prediction of opposite periodontal statuses might corroborate the true biological existence of the bioinformatics concept of ASVs.

### 4.5.9. Limitations of the present study

One of the main limitations of the present study was the unequal sample sizes for the different groups being studied. Specifically, the supragingival gingivitis, treated periodontitis, and subgingival healthy periodontitis sites were represented by between 60 and 90 samples; meanwhile, the rest of the groups had >150 and even the two periodontitis groups ("supra" and "sub") had around 500 and 770 samples, respectively.

Moreover, we could not assess the subgingival microbiota associated with gingivitis and healthy and diseased peri-implantitis, since these groups were represented by <45 samples. We were also unable to either identify any investigations of the supragingival microbiota associated with peri-implantitis or conduct analyses of the impact of different degrees of periodontitis severity. Taken together, these issues highlight the need for more 16S rRNA gene research on dental plaque obtained from sites of gingivitis, healthy areas of periodontitis, treated periodontitis, and peri-implantitis, potentially leading to the development of an improved meta-omics approach in the future.

In relation to the above, it should be noted that ~17% of the full text excluded articles used Illumina technology but had not stored their sequences in databases. Also, as stated in previous sections, other publications were excluded for inadequate metadata reporting and many others had low metadata quality. Thus, the potential number of studies and, in consequence, groups and samples with different periodontal conditions assessed could have been higher. Therefore, we believe that the storage of metadata and sequences in public repositories should be mandatory, as well as the fulfillment of quality requirements to facilitate the reproducibility of future research and the conduct of large-scale meta-omics research.

Lastly, keeping in mind the principles of clinical metagenomics and given saliva has been considered a potential diagnostic tool for systemic and oral diseases (115,116); it would be interesting to perform a meta-omics analysis on the diagnostic potential of the salivary microbiota at ASV level to distinguish different periodontal conditions. Indeed, first preliminary results from our research group have shown that the salivary microbiota may have an outstanding (ROC≥0.9) ability (56) to discriminate between periodontal health, gingivitis and non-treated periodontitis.

## 4.6. CONCLUSIONS

The bacterial richness associated with periodontitis is higher than in periodontal health in supragingival plaque and lower in subgingival plaque; evenness is higher in disease than in health in both niches. The supragingival microbiota is richer and more diverse than its subgingival counterpart for the same periodontal health status. The bacterial community's structure is different for distinct periodontal conditions in supragingival and subgingival plaque, as well as for the same health status between the two niches.

The core microbiota of supragingival and subgingival plaque does not allow the characterisation of periodontal health and disease, revealing the high heterogeneity of the oral microbiota. The percentage of the dental-plaque bacterial community that is organised into co-occurrence networks at the ASV level is very small; the untreated periodontitis network of supragingival plaque is more extensive, containing more nodes, interconnections, and interconnected bacterial clusters than its subgingival counterpart. The main keystone ASVs in the periodontal health networks of the supragingival plaque are *R. dentocariosa* ASV2 and *S. oralis* subsp. *dentisani* clade 058 ASV1. The principal hub in the untreated periodontitis networks of the supragingival plaque is *S. sanguinis* ASV228; in the subgingival plaque, it is *T. forsythia* ASV15. The main keystone taxa in the treated periodontitis network of the subgingival niche are *T. forsythia* ASV15, *F. nucleatum* subsp. *vincentii* ASV10, and *S. oralis* subsp. *dentisani* clade 058 ASV1.

A small proportion of the supra- and subgingival taxa have an outstanding ability to distinguish between periodontal conditions, and a relevant percentage are members of the core microbiota. From a clinical metagenomics point of view, supragingival plaque is a better bacterial biomarker than its subgingival counterpart for differentiating periodontal health from untreated and treated periodontitis.

The main periodontal-health-predictor ASVs in supragingival and subgingival plaque are: *R. dentocariosa* ASV2; *H. parainfluenzae* ASV3, ASV78, ASV45, and ASV46; *K. oralis* ASV66; *S. vestibularis* ASV27; and *A.* HMT170 ASV119. In contrast, the main periodontitis-predictor ASVs in both plaque types are: *T. forsythia* ASV15; *F. alocis* ASV19; *T. denticola* ASV38 and ASV150; *F. fastidiosum* ASV97; *P.* HMT369 ASV124; *S. anginosus* ASV142; and

*P. nodatum* ASV189. Of these, *F. fastidiosum* ASV97 and *S. anginosus* ASV142 also acted as predictors of gingivitis in the supragingival niche.

## 4.7. REFERENCES

(1) Campisciano G, Toschetti A, Comar M, Taranto RD, Berton F, Stacchi C. Shifts of subgingival bacterial population after nonsurgical and pharmacological therapy of localized aggressive periodontitis, followed for 1 year by Ion Torrent PGM platform. Eur J Dent. 2017 Jan-Mar;11(1):126-9.

(2) Iwauchi M, Horigome A, Ishikawa K, Mikuni A, Nakano M, Xiao JZ, et al. Relationship between oral and gut microbiota in elderly people. Immun Inflamm Dis. 2019 Sep;7(3):229-36.

(3) Huse SM, Ye Y, Zhou Y, Fodor AA. A core human microbiome as viewed through 16S rRNA sequence clusters. PLoS One. 2012 Jun;7(6):e34242. doi: 10.1371/journal.pone.0034242.

(4) Lee WH, Chen HM, Yang SF, Liang C, Peng CY, Lin FM, et al. Bacterial alterations in salivary microbiota and their association in oral cancer. Sci Rep. 2017 Nov;7(1):16540. doi: 10.1038/s41598-017-16418-x.

(5) Ye W, Zhang Y, He M, Zhu C, Feng XP. Relationship of tongue coating microbiome on volatile sulfur compounds in healthy and halitosis adults. J Breath Res. 2019 Nov;14(1):016005. doi: 10.1088/1752-7163/ab47b4.

(6) Yu XL, Chan Y, Zhuang L, Lai HC, Lang NP, Keung Leung W, et al. Intra-oral single-site comparisons of periodontal and peri-implant microbiota in health and disease. Clin Oral Implants Res. 2019 Aug;30(8):760-76.

(7) Jiang W, Ling Z, Lin X, Chen Y, Zhang J, Yu J, et al. Pyrosequencing analysis of oral microbiota shifting in various caries states in childhood. Microb Ecol. 2014 May;67(4):962-9.

(8) Kageyama S, Nagao Y, Ma J, Asakawa M, Yoshida R, Takeshita T, et al. Compositional shift of oral microbiota following surgical resection of tongue cancer. Front Cell Infect Microbiol. 2020 Nov;10:600884. doi: 10.3389/fcimb.2020.600884.

(9) Ng E, Tay JRH, Balan P, Ong MMA, Bostanci N, Belibasakis GN, et al. Metagenomic sequencing provides new insights into the subgingival bacteriome and aetiopathology of periodontitis. J Periodont Res. 2021 Apr;56(2):205-18.

(10) de Melo F, Milanesi FC, Angst PDM, Oppermann RV. A systematic review of the microbiota composition in various peri-implant conditions: data from 16S rRNA gene sequencing. Arch Oral Biol. 2020 Sep;117:104776. doi: 10.1016/j.archoralbio.2020.104776.

(11) Bhaumik D, Manikandan D, Foxman B. Cariogenic and oral health taxa in the oral cavity among children and adults: a scoping review. Arch Oral Biol. 2021 Sep;129:105204. doi: 10.1016/j.archoralbio.2021.105204.

(12) Ramos RT, Sodré CS, de Sousa-Rodrigues PMGR, da Silva AMP, Fuly MS, dos Santos HF, et al. High-throughput nucleotide sequencing for bacteriome studies in oral squamous cell carcinoma: a systematic review. Oral Maxillofac Surg. 2020 Dec;24(4):387-401.

(13) de la Cuesta-Zuluaga J, Escobar JS. Considerations for optimizing microbiome analysis using a marker gene. Front Nutr. 2016 Aug;3:26. doi: 10.3389/fnut.2016.00026.

(14) Robinson CK, Brotman RM, Ravel J. Intricacies of assessing the human microbiome in epidemiologic studies. Ann Epidemiol. 2016 May;26(5):311-21.

(15) Nearing JT, Comeau AM, Langille MGI. Identifying biases and their potential solutions in human microbiome studies. Microbiome. 2021 May;9(1):113. doi: 10.1186/s40168-021-01059-0.

(16) Regueira-Iglesias A, Vázquez-González L, Balsa-Castro C, Vila-Blanco N, Blanco-Pintos T, Tamames J, et al. In silico evaluation and selection of the best 16S rRNA gene primers for use in next-generation sequencing to detect oral bacteria and archaea. Accepted for publication in Microbiome. Preprint at Research Square. 2021. doi: 10.21203/rs.3.rs-516961/v1.

(17) Regueira-Iglesias A, Vázquez-González L, Balsa-Castro C, Blanco-Pintos T, Martín-Biedma B, Arce VM, et al. In silico detection of oral prokaryotic species with highly similar 16S rRNA sequence segments using different primer pairs. Front Cell Infect Microbiol. 2022 Feb;11:770668. doi: 10.3389/fcimb.2021.770668.

(18) Caruso V, Song X, Asquith M, Karstens L. Performance of microbiome sequence inference methods in environments with varying biomass. mSystems. 2019 Feb;4(1):e00163-18. doi: 10.1128/mSystems.00163-18.

(19) Prodan A, Tremaroli V, Brolin H, Zwinderman AH, Nieuwdorp M, Levin E. Comparing bioinformatic pipelines for microbial 16S rRNA amplicon sequencing. PLoS One. 2020 Jan;15(1):e0227434. doi: 10.1371/journal.pone.0227434.

(20) Callahan BJ, McMurdie PJ, Holmes SP. Exact sequence variants should replace operational taxonomic units in marker-gene data analysis. ISME J. 2017 Dec;11(12):2639-43.

(21) F Escapa I, Huang Y, Chen T, Lin M, Kokaras A, Dewhirst FE, et al. Construction of habitat-specific training sets to achieve species-level assignment in 16S rRNA gene datasets. Microbiome. 2020 May;8(1):65. doi: 10.1186/s40168-020-00841-w.

(22) Armitage GC. Development of a classification system for periodontal diseases and conditions. Ann Periodontol. 1999 Dec;4(1):1-6.

(23) Page RC, Eke PI. Case definitions for use in population-based surveillance of periodontitis. J Periodontol. 2007 Jul;78(7 Suppl):1387-99.

(24) World Medical Association. World Medical Association Declaration of Helsinki: ethical principles for medical research involving human subjects. JAMA. 2013 Nov;310(20):2191-4.



(25) Willis JR, González-Torres P, Pittis AA, Bejarano LA, Cozzuto L, Andreu-Somavilla N, et al. Citizen science charts two major “stomatotypes” in the oral microbiome of adolescents

and reveals links with habits and drinking water composition. Microbiome. 2018 Dec;6(1):218. doi: 10.1186/s40168-018-0592-3.

(26) Leinonen R, Sugawara H, Shumway M, International Nucleotide Sequence, Database Collaboration. The sequence read archive. Nucleic Acids Res. 2011 Jan;39(Database issue):D19-21. doi: 10.1093/nar/gkq1019.

(27) Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. J Mol Biol. 1990 Oct;215(3):403-10.

(28) R Core Team. R: a language and environment for statistical computing. R package version 4.1.2. Vienna, Austria: R Foundation for Statistical Computing; 2021; Available at: https://www.R-project.org/.

(29) Feinerer, I., Hornik, K., Meyer, D. Text mining infrastructure in R. J Stat Softw. 2008 Mar;25(5):1-54. doi: 10.18637/jss.v025.i05.

(30) Hornik, K. NLP: natural language processing infrastructure. R package version 0.2-1. 2020; Available at: https://CRAN.R-project.org/package=NLP.

(31) SRA Toolkit Development Team. Sequence read archive toolkit. Available at: http://www.ncbi.nlm.nih.gov/Traces/sra/sra.cgi?cmd=show&f=software&m=software&s=software.

(32) GNU P. Free Software Foundation. Bash. Version 5.0.17. 2019; Available at: http://www.gnu.org/.

(33) Edgar R. Search and clustering orders of magnitude faster than BLAST. Bioinformatics. 2010 Oct;26(19):2460-1.

(34) Gentleman RC, Carey VJ, Bates DM, Bolstad B, Dettling M, Dudoit S, et al. Bioconductor: open software development for computational biology and bioinformatics. Genome Biol. 2004 Sep;5(10):R80. doi: 10.1186/gb-2004-5-10-r80.

(35) Pagès H, Aboyoun P, Gentleman R, DebRoy S. Biostrings: efficient manipulation of biological strings. R package version 2.60.2. 2021; Available at: https://bioconductor.org/packages/Biostrings.

(36) Schloss PD, Westcott SL, Ryabin T, Hall JR, Hartmann M, Hollister EB, et al. Introducing mothur: open-source, platform-independent, community-supported software for describing and comparing microbial communities. Appl Environ Microbiol. 2009 Dec;75(23):7537-41.

(37) Rognes T, Flouri T, Nichols B, Quince C, Mahe F. VSEARCH: a versatile open source tool for metagenomics. PeerJ. 2016 Oct;4:e2584. doi: 10.7717/peerj.2584.

(38) McMurdie PJ, Holmes S. Waste not, want not: why rarefying microbiome data is inadmissible. PLoS Comput Biol. 2014 Apr;10(4):e1003531. doi: 10.1371/journal.pcbi.1003531.

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

(40) Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol. 2014 Dec;15(12):550. doi: 10.1186/s13059-014-0550-8.

(41) Lahti L, Shetty S. Tools for microbiome analysis in R. Microbiome package. R package version 1.14.0. 2017; Available at: http://microbiome.github.com/microbiome.



(42) Bourgon R, Gentleman R, Huber W. Independent filtering increases detection power for high-throughput experiments. Proc Natl Acad Sci U S A. 2010 May;107(21):9546-51.

(43) Spellerberg IF, Fedor PJ. A tribute to Claude Shannon (1916–2001) and a plea for more rigorous use of species richness, species diversity and the 'Shannon–Wiener' index. Glob Ecol Biogeogr. 2003 Apr;12(3):177-9.

(44) Pielou EC. The measurement of diversity in different types of biological collections. J Theor Biol. 1966 Dec;13:131-44.

(45) Rohart F, Gautier B, Singh A, Lê Cao K. mixOmics: an R package for 'omics feature selection and multiple data integration. PLoS Comput Biol. 2017 Nov;13(11):e1005752. doi: 10.1371/journal.pcbi.1005752.

(46) Anderson M. A new method for non-parametric multivariate analysis of variance. Austral Ecol. 2001 Feb;26(1):32-46.

(47) Oksanen J, Blanchet FG, Friendly M, Kindt R, Legendre P, McGlinn D, et al. vegan: community ecology package. R package version 2.5-7. 2020; Available at: https://cran.r-project.org, https://github.com/vegandevs/vegan.

(48) Friedman J, Alm EJ. Inferring correlation networks from genomic survey data. PLoS Comput Biol. 2012 Sep; 8(9):e1002687. doi: 10.1371/journal.pcbi.1002687.

(49) Weiss S, Van Treuren W, Lozupone C, Faust K, Friedman J, Deng Y, et al. Correlation detection strategies in microbial data sets vary widely in sensitivity and precision. ISME J. 2016 Jul;10(7):1669-81.

(50) Kurtz Z, Mueller C, Miraldi E, Bonneau R. SpiecEasi: sparse inverse covariance for ecological statistical inference. R package version 1.1.1. 2021; Available at: https://github.com/zdk123/SpiecEasi.



(51) Csardi G, Nepusz T. The Igraph software package for complex network research. InterJournal, Complex Systems. 2006; 1695; Available at: http://igraph.org.

(52) Golbeck J. Chapter 3 - Network structure and measures. In: Golbeck J, editor. Analyzing the social web. Boston: Morgan Kaufmann; 2013. p. 25-44.

(53) Martín-González AM, Dalsgaard B, Olesen JM. Centrality measures and the importance of generalist species in pollination networks. Ecol Complex. 2010 Mar;7(1):36-43.

(54) Banerjee S, Schlaeppi K, van der Heijden MGA. Keystone taxa as drivers of microbiome structure and functioning. Nat Rev Microbiol. 2018 Sep;16(9):567-76.

(55) Lê Cao KA, Boitard S, Besse P. Sparse PLS discriminant analysis: biologically relevant feature selection and graphical displays for multiclass problems. BMC Bioinformatics. 2011 Jun;12:253. doi: 10.1186/1471-2105-12-253.

(56) Hosmer DJ, Lemeshow S, Sturdivant R. Applied logistic regression. 3rd ed. Hoboken (NJ): John Wiley & Sons, Inc.; 2013. 528 p.

(57) Coretti L, Cuomo M, Florio E, Palumbo D, Keller S, Pero R, et al. Subgingival dysbiosis in smoker and nonsmoker patients with chronic periodontitis. Mol Med Rep. 2017 Apr;15(4):2007-14.

(58) Hagenfeld D, Koch R, Junemann S, Prior K, Harks I, Eickholz P, et al. Do we treat our patients or rather periodontal microbes with adjunctive antibiotics in periodontal therapy? A 16S rDNA microbial community analysis. PLoS One. 2018 Apr;13(4):e0195534. doi: 10.1371/journal.pone.0195534.

(59) Pei J, Li F, Xie Y, Liu J, Yu T, Feng X. Microbial and metabolomic analysis of gingival crevicular fluid in general chronic periodontitis patients: lessons for a predictive, preventive, and personalized medical approach. EPMA J. 2020 Apr;11(2):197-215.



(60) Johnston W, Rosier BT, Artacho A, Paterson M, Piela K, Delaney C, et al. Mechanical biofilm disruption causes microbial and immunological shifts in periodontitis patients. Sci Rep. 2021 May;11(1):9796. doi: 10.1038/s41598-021-89002-z.

(61) Liu H, Chen H, Liao Y, Li H, Shi L, Deng Y, et al. Comparative analyses of the subgingival microbiome in chronic periodontitis patients with and without gingival erosive oral lichen planus based on 16S rRNA gene sequencing. Biomed Res Int. 2021 Jun;2021:9995225. doi: 10.1155/2021/9995225.

(62) Liu G, Luan Q, Chen F, Chen Z, Zhang Q, Yu X. Shift in the subgingival microbiome following scaling and root planing in generalized aggressive periodontitis. J Clin Periodontol. 2018 Apr;45(4):440-52.

(63) Komatsu K, Shiba T, Takeuchi Y, Watanabe T, Koyanagi T, Nemoto T, et al. Discriminating microbial community structure between peri-implantitis and periodontitis with integrated metagenomic, metatranscriptomic, and network analysis. Front Cell Infect Microbiol. 2020 Dec;10:596490. doi: 10.3389/fcimb.2020.596490.

(64) Boyer E, Martin B, Le Gall-David S, Fong SB, Deugnier Y, Bonnaure-Mallet M, et al. Periodontal pathogens and clinical parameters in chronic periodontitis. Mol Oral Microbiol. 2020 Jan;35(1):19-28.

(65) Balmasova IP, Olekhnovich EI, Klimina KM, Korenkova AA, Vakhitova MT, Babaev EA, et al. Drift of the subgingival periodontal microbiome during chronic periodontitis in type 2 diabetes mellitus patients. Pathogens. 2021 Apr;10(5):504. doi: 10.3390/pathogens10050504.

(66) Kharitonova M, Vankov P, Abdrakhmanov A, Mamaeva E, Yakovleva G, Ilinskaya O. The composition of microbial communities in inflammatory periodontal diseases in young adults Tatars. AIMS Microbiol. 2021 Jan;7(1):59-74.

(67) Liu G, Chen F, Cai Y, Chen Z, Luan Q, Yu X. Measuring the subgingival microbiota in periodontitis patients: comparison of the surface layer and the underlying layers. Microbiol Immunol. 2020 Feb;64(2):99-112.

(68) Pyysalo MJ, Mishra PP, Sundström K, Lehtimäki T, Karhunen PJ, Pessi T. Increased tooth brushing frequency is associated with reduced gingival pocket bacterial diversity in patients with intracranial aneurysms. PeerJ. 2019 Jan;e6316. doi: 10.7717/peerj.6316.

(69) Wirth R, Maróti G, Lipták L, Mester M, Al Ayoubi A, Pap B, et al. Microbiomes in supragingival biofilms and saliva of adolescents with gingivitis and gingival health. Oral Dis. 2021 Apr. doi: 10.1111/odi.13883.

(70) Na HS, Kim SY, Han H, Kim HJ, Lee JY, Lee JH, et al. Identification of potential oral microbial biomarkers for the diagnosis of periodontitis. J Clin Med. 2020 May;9(5):1549. doi: 10.3390/jcm9051549.

(71) Na HS, Kim S, Kim S, Yu Y, Kim SY, Kim HJ, et al. Molecular subgroup of periodontitis revealed by integrated analysis of the microbiome and metabolome in a cross-sectional observational study. J Oral Microbiol. 2021 Mar;13(1):1902707. doi: 10.1080/20002297.2021.1902707.

(72) Na HS, Jung N, Choi S, Kim Sy, Kim H, Lee JY, et al. Analysis of oral microbiome in chronic periodontitis with Alzheimer's disease: pilot study. Preprint at Research Square. 2020. doi: 10.21203/rs.3.rs-24938/v1.

(73) Stephen AS, Dhadwal N, Nagala V, Gonzales-Marin C, Gillam DG, Bradshaw DJ, et al. Interdental and subgingival microbiota may affect the tongue microbial ecology and oral malodour in health, gingivitis and periodontitis. J Periodontal Res. 2021 Dec;56(6):1174-84.

(74) Hagenfeld D, Prior K, Harks I, Jockel-Schneider Y, May TW, Harmsen D, et al. No differences in microbiome changes between anti-adhesive and antibacterial ingredients in toothpastes during periodontal therapy. J Periodontal Res. 2019 Aug;54(4):435-43.



(75) Annavajhala MK, Khan SD, Sullivan SB, Shah J, Pass L, Kister K, et al. Oral and gut microbial diversity and immune regulation in patients with HIV on antiretroviral therapy. mSphere. 2020 Feb;5(1):e00798-19. doi: 10.1128/mSphere.00798-19.

(76) Hall MW, Singh N, Ng KF, Lam DK, Goldberg MB, Tenenbaum HC, et al. Inter-personal diversity and temporal dynamics of dental, tongue, and salivary microbiota in the healthy oral cavity. NPJ Biofilms Microbiomes. 2017 Jan;3:2. doi: 10.1038/s41522-016-0011-0

(77) Wei Y, Shi M, Zhen M, Wang C, Hu W, Nie Y, et al. Comparison of subgingival and buccal mucosa microbiome in chronic and aggressive periodontitis: a pilot study. Front Cell Infect Microbiol. 2019 Mar;9:53. doi: 10.3389/fcimb.2019.00053.

(78) Amado PPP, Kawamoto D, Albuquerque-Souza E, Franco DC, Saraiva L, Casarin RCV, et al. Oral and fecal microbiome in molar-incisor pattern periodontitis. Front Cell Infect Microbiol. 2020 Oct;10:583761. doi: 10.3389/fcimb.2020.583761.

(79) Tonetti MS, Greenwell H, Kornman KS. Staging and grading of periodontitis: framework and proposal of a new classification and case definition. J Periodontol. 2018 Jun;89 Suppl 1:S159-S172. doi: 10.1002/JPER.18-0006.

(80) Meuric V, Le Gall-David S, Boyer E, Acuña-Amador L, Martin B, Fong SB, et al. Signature of microbial dysbiosis in periodontitis. Appl Environ Microbiol 2017 Jun 30;83(14):e00462-17. Print 2017 Jul 15.

(81) Cai Z, Lin S, Hu S, Zhao L. Structure and function of oral microbial community in periodontitis based on integrated data. Front Cell Infect Microbiol. 2021 Jun;11:663756. doi: 10.3389/fcimb.2021.663756.

(82) Abusleme L, Hoare A, Hong BY, Diaz PI. Microbial signatures of health, gingivitis, and periodontitis. Periodontol 2000. 2021 Jun;86(1):57-78.

(83) Sisk-Hackworth L, Ortiz-Velez A, Reed MB, Kelley ST. Compositional data analysis of periodontal disease microbial communities. Front Microbiol. 2021 May;12:617949. doi: 10.3389/fmicb.2021.617949.

(84) 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;2010:baq013. doi: 10.1093/database/baq013.

(85) DeSantis TZ, Hugenholtz P, Larsen N, Rojas M, Brodie EL, Keller K, et al. Greengenes, a chimera-checked 16S rRNA gene database and workbench compatible with ARB. Appl Environ Microbiol. 2006 Jul;72(7):5069-72.

(86) Cole JR, Wang Q, Fish JA, Chai B, McGarrell DM, Sun Y, et al. Ribosomal Database Project: data and tools for high throughput rRNA analysis. Nucleic Acids Res. 2014 Jan;42(D1):D633-42. doi: 10.1093/nar/gkt1244.

(87) Edgar R. Taxonomy annotation and guide tree errors in 16S rRNA databases. PeerJ. 2018 Jun;6:e5030. doi: 10.7717/peerj.5030.

(88) Soergel DA, Dey N, Knight R, Brenner SE. Selection of primers for optimal taxonomic classification of environmental 16S rRNA gene sequences. ISME J. 2012 Jul;6(7):1440-4.

(89) Relvas M, Regueira-Iglesias A, Balsa-Castro C, Salazar F, Pacheco JJ, Cabral C, et al. Relationship between dental and periodontal health status and the salivary microbiome: bacterial diversity, co-occurrence networks and predictive models. Sci Rep. 2021 Jan;11(1):929. doi: 10.1038/s41598-020-79875-x.

(90) Camelo-Castillo A, Novoa L, Balsa-Castro C, Blanco J, Mira A, Tomas I. Relationship between periodontitis-associated subgingival microbiota and clinical inflammation by 16S pyrosequencing. J Clin Periodontol. 2015 Dec;42(12):1074-82.

(91) Camelo-Castillo AJ, Mira A, Pico A, Nibali L, Henderson B, Donos N, et al. Subgingival microbiota in health compared to periodontitis and the influence of smoking. Front Microbiol. 2015 Feb;6:119. doi: 10.3389/fmicb.2015.00119.

(92) Dinnes J, Deeks J, Kirby J, Roderick P. A methodological review of how heterogeneity has been examined in systematic reviews of diagnostic test accuracy. Health Technol Assess. 2005 Mar;9(12):1-113, iii. doi: 10.3310/hta9120.

(93) Reitsma JB, Rutjes AWS, Whiting P, Vlassov VV, Leeflang M, Deeks JJ. Chapter 9: Assessing methodological quality. In: Deeks JJ, Bossuyt PM, Gatsonis C, editors. Cochrane handbook for systematic reviews of diagnostic test accuracy. 1st ed. Cham: The Cochrane Collaboration; 2009. p. 1–29. Available: http://srdta.cochrane.org/.

(94) Downs SH, Black N. The feasibility of creating a checklist for the assessment of the methodological quality both of randomised and non-randomised studies of health care interventions. J Epidemiol Community Health. 1998 Jun;52(6):377-84.

(95) Acharya A, Chen T, Chan Y, Watt RM, Jin L, Mattheos N. Species-level salivary microbial indicators of well-resolved periodontitis: a preliminary investigation. Front Cell Infect Microbiol. 2019 Oct;9:347. doi: 10.3389/fcimb.2019.00347.

(96) Damgaard C, Danielsen AK, Enevold C, Massarenti L, Nielsen CH, Holmstrup P, et al. Porphyromonas gingivalis in saliva associates with chronic and aggressive periodontitis. J Oral Microbiol. 2019 Aug;11(1):1653123. doi: 10.1080/20002297.2019.1653123.

(97) Jin Y, Yip H. Supragingival calculus: formation and control. Crit Rev Oral Biol Med. 2002 Sep;13(5):426-41.

(98) Jiang D, Armour CR, Hu C, Mei M, Tian C, Sharpton TJ, et al. Microbiome multi-omics network analysis: statistical considerations, limitations, and opportunities. Front Genet. 2019 Nov;10:995. doi: 10.3389/fgene.2019.00995.

(99) Das P, Ji B, Kovatcheva-Datchary P, Bäckhed F, Nielsen J. In vitro co-cultures of human gut bacterial species as predicted from co-occurrence network analysis. PLoS One. 2018 Mar;13(3):e0195161. doi: 10.1371/journal.pone.0195161.

(100) Manirajan BA, Maisinger C, Ratering S, Rusch V, Schwiertz A, Cardinale M, et al. Diversity, specificity, co-occurrence and hub taxa of the bacterial-fungal pollen microbiome. FEMS Microbiol Ecol. 2018 Aug;94(8):10.1093/femsec/fiy112.

(101) Lupatini M, Suleiman AKA, Jacques RJS, Antoniolli ZI, de Siquiera-Ferreira A, Kuramae EE, et al. Network topology reveals high connectance levels and few key microbial genera within soils. Front Environ Sci. 2014 May;2:10. doi: 10.3389/fenvs.2014.00010.

(102) Forbes JD, Knox NC, Peterson CL, Reimer AR. Highlighting clinical metagenomics for enhanced diagnostic decision-making: a step towards wider implementation. Comput Struct Biotechnol J. 2018 Feb;16:108-20.

(103) Chiu CY, Miller SA. Clinical metagenomics. Nat Rev Genet. 2019 Jun;20(6):341-55.

(104) Genco RJ, LaMonte MJ, McSkimming DI, Buck MJ, Li L, Hovey KM, et al. The subgingival microbiome relationship to periodontal disease in older women. J Dent Res 2019 Aug;98(9):975-84.

(105) Diaz PI. Microbial diversity and interactions in subgingival biofilm communities. Front Oral Biol. 2012 Jan;15:17-40.

(106) Drisko CH. Nonsurgical periodontal therapy. Periodontol 2000. 2001 Feb;25(1):77-88.

(107) Van Dyke TE, Bartold PM, Reynolds EC. The nexus between periodontal inflammation and dysbiosis. Front Immunol. 2020 Mar;11:511. doi: 10.3389/fimmu.2020.00511.

(108) Miyoshi T, Oge S, Nakata S, Ueno Y, Ukita H, Kousaka R, et al. Gemella haemolysans inhibits the growth of the periodontal pathogen Porphyromonas gingivalis. Sci Rep. 2021 Jun;11(1):11742. doi: 10.1038/s41598-021-91267-3.



(109) Tseng YC, Yang HY, Lin WT, Chang CB, Chien HC, Wang HP, et al. Salivary dysbiosis in Sjögren's syndrome and a commensal-mediated immunomodulatory effect of salivary gland

epithelial cells. NPJ Biofilms Microbiomes. 2021 Mar;7(1):21. doi: 10.1038/s41522-021-00192-w.

(110) Thompson SA. Campylobacter surface-layers (S-layers) and immune evasion. Ann Periodontol. 2002 Dec;7(1):43-53.

(111) Kumada H, Watanabe K, Nakamu A, Haishima Y, Kondo S, Hisatsune K, et al. Chemical and biological properties of lipopolysaccharide from Selenomonas sputigena ATCC 33150. Oral Microbiol Immunol. 1997 Jun;12(3):162-7.

(112) Miralda I, Uriarte SM. Periodontal pathogens' strategies disarm neutrophils to promote dysregulated inflammation. Mol Oral Microbiol. 2021 Apr;36(2):103-20.

(113) López-López A, Camelo-Castillo A, Ferrer MD, Simon-Soro Á, Mira A. Health-associated niche inhabitants as oral probiotics: the case of Streptococcus dentisani. Front Microbiol. 2017 Mar;8:379. doi: 10.3389/fmicb.2017.00379.

(114) Jansen PM, Abdelbary MMH, Conrads G. A concerted probiotic activity to inhibit periodontitis-associated bacteria. PLoS One. 2021 Mar;16(3):e0248308. doi: 10.1371/journal.pone.0248308.

(115) Kaczor-Urbanowicz KE, Martin Carreras-Presas C, Aro K, Tu M, Garcia-Godoy F, Wong DT. Saliva diagnostics - Current views and directions. Exp Biol Med (Maywood). 2017 Mar;242(5):459-72. doi: 10.1177/1535370216681550.

(116) Javaid MA, Ahmed AS, Durand R, Tran SD. Saliva as a diagnostic tool for oral and systemic diseases. J Oral Biol Craniofac Res. 2016 Jan-Apr;6(1):66-75. doi: 10.1016/j.jobcr.2015.08.006.

# CONCLUSIONS

# Conclusions

The conclusions derived from the methodological studies are:

1) For the three mean amplicon length categories (100-300, 301-600, and >600 bps), the primer pairs producing the best coverage estimates for detecting oral-bacteria targeted regions 3-4, 4-7, and 3-7. These species were: KP_F048-OP_R043 (primer pair position for *E. coli* J01859.1: 342-529); KP_F051-OP_R030 (514-1079); and KP_F048-OP_R030 (342-1079). For the detection of oral archaea, the pairs with the best coverage amplified regions 5-6 and 3-6, and were: OP_F066-KP_R013 (784-undefined); KP_F020-KP_R013 (518-undefined); and OP_F114-KP_R013 (340-undefined). The pairs that provided the optimum coverage of the bacteria and archaea domains jointly were found in regions 4-5, 3-5, and 5-9, and were: KP_F020-KP_R032 (518-801); OP_F114-KP_R031 (340-801); and OP_F066-OP_R121 (784-1405).

2) Nearly all the oral bacteria and about half the oral archaea have more than one 16S rRNA gene in their respective genomes. Depending on the primer pair used, up to almost half of the species have MAs that affect the relevant genera present in the oral environment, including *Actinomyces*, *Fusobacterium*, *Lactobacillus*, *Methanosarcina*, *Staphylococcus*, and *Streptococcus.* In terms of the SC-NMA, the best primer pairs were: KP_F048-OP_R030 for bacteria (region 3-7; primer pair position for *Escherichia coli* J01859.1: 342-1079); KP_F018-KP_R063 for archaea (region 3-9; undefined-1506); and OP_F114_OP_R121 for bacteria and archaea jointly (region 3-9; 340-1405). Consequently, in addition to the 16S rRNA gene redundancy, the considerable presence of MAs must be controlled to ensure the accurate interpretation of the microbial diversity data. The choice of primer pair significantly affects diversity estimates and taxonomic classification, conditioning the comparability of oral microbiota studies that employ different primer pairs.

3) The tested primer pairs targeting oral bacteria and/or archaea detected an average of more than 150 potential OTUs that may contain different species when the ≥97% similarity threshold is used. For SC-NASI97, the best primer pairings were OP_F053-KP_R020 for bacteria (region 1-3; primer pair position for *Escherichia coli* J01859.1: 9-356); KP_F018-KP_R002 for archaea (region 4; undefined-532); and OP_F114-KP_R031 for both (region 3-5; 340-801). Around 80% of the oral bacteria and oral archaea analysed had an ASI97 with at least one other species. These very similar species play different roles in the oral microbiota and belong to bacterial genera such as *Campylobacter*, *Rothia*, *Streptococcus*, and *Tannerella*, and archaeal genera like *Halovivax*, *Methanosarcina*, and *Methanosalsum*. Moreover, ~20% and ~30% of these two-by-two similarity relationships were established between species from different bacterial and archaeal genera, respectively. Even taxa from distinct families, orders, and classes could be grouped in the same OTU. Consequently, regardless of the primer pair used, sequence clustering with ≥97% similarity produces an inaccurate account of oral-bacterial and oral-archaeal species, which can greatly affect microbial diversity parameters. As a result, OTU clustering conditions the credibility of the associations between some oral species and certain health and disease conditions. This significantly limits the comparability of the microbial diversity findings reported in the oral microbiota literature.

4) In order to determine the most appropriate primer pairs for their research, scholars must decide which aspect to prioritise: coverage; identifying the lowest possible number of MAs; or, in the case of clustering sequences into OTUs, detecting the lowest possible number of species with ASI97. As evidenced in this thesis, some primers are more suitable than others, depending on the parameter chosen. Surprisingly, the pairings used the most in the relevant literature were never among the best performers.

The conclusions of the meta-omics study are:

1) The bacterial richness associated with periodontitis is higher than in periodontal health in supragingival plaque and lower in subgingival plaque; evenness is higher in disease than in health in both niches. The supragingival microbiota is richer and more diverse than its subgingival counterpart for the same periodontal health status. The bacterial

community's structure is different for distinct periodontal conditions in supragingival and subgingival plaque, as well as for the same health status between the two niches.

2) The core microbiota of supragingival and subgingival plaque does not allow the characterisation of periodontal health and disease, revealing the high heterogeneity of the oral microbiota. The percentage of the dental-plaque bacterial community that is organised into co-occurrence networks at the ASV level is very small; the untreated periodontitis network of supragingival plaque is more extensive, containing more nodes, interconnections, and interconnected bacterial clusters than its subgingival counterpart. The main keystone ASVs in the periodontal health networks of the supragingival plaque are *R. dentocariosa* ASV2 and *S. oralis* subsp. *dentisani* clade 058 ASV1. The principal hub in the untreated periodontitis networks of the supragingival plaque is *S. sanguinis* ASV228; in the subgingival plaque, it is *T. forsythia* ASV15. The main keystone taxa in the treated periodontitis network of the subgingival niche are *T. forsythia* ASV15, *F. nucleatum* subsp. *vincentii* ASV10, and *S. oralis* subsp. *dentisani* clade 058 ASV1.

3) A small proportion of the supra- and subgingival taxa have an outstanding ability to distinguish between periodontal conditions, and a relevant percentage are members of the core microbiota. From a clinical metagenomics point of view, supragingival plaque is a better bacterial biomarker than its subgingival counterpart for differentiating periodontal health from untreated and treated periodontitis.

4) The main periodontal-health-predictor ASVs in supragingival and subgingival plaque are: *R. dentocariosa* ASV2; *H. parainfluenzae* ASV3, ASV78, ASV45, and ASV46; *K. oralis* ASV66; *S. vestibularis* ASV27; and *A.* HMT170 ASV119. In contrast, the main periodontitis-predictor ASVs in both plaque types are: *T. forsythia* ASV15; *F. alocis* ASV19; *T. denticola* ASV38 and ASV150; *F. fastidiosum* ASV97; *P.* HMT369 ASV124; *S. anginosus* ASV142; and *P. nodatum* ASV189. Of these, *F. fastidiosum* ASV97 and *S. anginosus* ASV142 also acted as predictors of gingivitis in the supragingival niche.

# APPENDICES

# Appendices Introduction

Appendix S1. Permissions, which were given by the publishers for the use of non-original figures and tables in the Introduction of the present Thesis.

# Appendix S1

SPRINGER NATURE LICENSE
TERMS AND CONDITIONS

Jan 18, 2022

---

This Agreement between Universidade de Santiago de Compostela -- Alba Regueira Iglesias ("You") and Springer Nature ("Springer Nature") consists of your license details and the terms and conditions provided by Springer Nature and Copyright Clearance Center.

| | |
|---|---|
| License Number | 5232021362024 |
| License date | Jan 18, 2022 |
| Licensed Content Publisher | Springer Nature |
| Licensed Content Publication | Nature Reviews Disease Primers |
| Licensed Content Title | Periodontal diseases |
| Licensed Content Author | Denis F. Kinane et al |
| Licensed Content Date | Jun 22, 2017 |
| Type of Use | Thesis/Dissertation |
| Requestor type | academic/university or research institute |
| Format | print and electronic |
| Portion | figures/tables/illustrations |
| Number of figures/tables/illustrations | 1 |
| High-res required | no |

SPRINGER NATURE LICENSE
TERMS AND CONDITIONS

Jan 18, 2022

---

This Agreement between Universidade de Santiago de Compostela -- Alba Regueira Iglesias ("You") and Springer Nature ("Springer Nature") consists of your license details and the terms and conditions provided by Springer Nature and Copyright Clearance Center.

| | |
|---|---|
| License Number | 5232041159721 |
| License date | Jan 18, 2022 |
| Licensed Content Publisher | Springer Nature |
| Licensed Content Publication | Nature Reviews Immunology |
| Licensed Content Title | Periodontitis: from microbial immune subversion to systemic inflammation |
| Licensed Content Author | George Hajishengallis |
| Licensed Content Date | Dec 23, 2014 |
| Type of Use | Thesis/Dissertation |
| Requestor type | academic/university or research institute |
| Format | print and electronic |
| Portion | figures/tables/illustrations |
| Number of figures/tables/illustrations | 1 |
| High-res required | no |

CCC Marketplace™

# Order Confirmation

Thank you, your order has been placed. An email confirmation has been sent to you. Your order license details and printable licenses will be available within 24 hours. Please access Manage Account for final order details.

**This is not an invoice. Please go to manage account to access your order history and invoices.**

## CUSTOMER INFORMATION

Payment by invoice: You can cancel your order until the invoice is generated by contacting customer service.

### Billing Address

**Ms. Alba Regueira Iglesias**
Universidade de Santiago de Compostela
Calle de Entrerríos, 1
Santiago De Compostela, A Coruña 15705
Spain

+34 622783071
albaregueira.iglesias@usc.es

### Customer Location

**Ms. Alba Regueira Iglesias**
Universidade de Santiago de Compostela
Calle de Entrerríos, 1
Santiago De Compostela, A Coruña 15705
Spain

### PO Number (optional)

N/A

### Payment options

Invoice

## PENDING ORDER CONFIRMATION

**Confirmation Number:** Pending

**Order Date:** 01-Sep-2021

### 1. Annual review of genomics and human genetics

0,00 EUR

| | | | |
|---|---|---|---|
| **Order License ID** | Pending | **Publisher** | ANNUAL REVIEWS |
| **ISSN** | 1545-293X | | |
| **Type of Use** | Republish in a thesis/dissertation | **Portion** | Image/photo/illustration |

### LICENSED CONTENT

| | | | |
|---|---|---|---|
| **Publication Title** | Annual review of genomics and human genetics | **Rightsholder** | Annual Reviews, Inc. |
| | | **Publication Type** | e-Journal |
| **Date** | 01/01/2000 | **URL** | http://arjournals.annualreviews.org/loi/genom/ |
| **Language** | English | | |

## Illumina Copyright Permission Agreement

| Parties | |
|---|---|
| Illumina | Illumina, Inc.<br>5200 Illumina Way, San Diego, CA 92122, US |
| Requestor | Alba Regueira Iglesias<br>University of Santiago de Compostela (Spain) |

| Start Date & Term | |
|---|---|
| Start Date | August 23, 2021 |
| Term | N/A |

**Permitted Content**

Figure 4 from "An Introduction to Next-Generation Sequencing Technology" (https://www.illumina.com/content/dam/illumina-marketing/documents/products/illumina_sequencing_introduction.pdf)

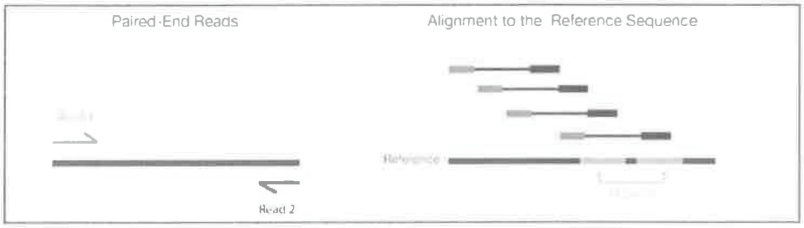



**Figure 4: Paired-End Sequencing and Alignment** – Paired-end sequencing enables both ends of the DNA fragment to be sequenced. Because the distance between each paired read is known, alignment algorithms can use this information to map the reads over repetitive regions more precisely. This results in better alignment of reads especially across difficult-to-sequence, repetitive regions of the genome.

**Permitted Territory/ Use**

Requester would like to use the image in a doctoral thesis about the periodontal microbiomes in states of health and disease through 16S rRNA gene sequencing

**Rights Granted to the Requestor**

Illumina grants you a limited, revocable, non-exclusive license to use Permitted Content subject to the terms of this Agreement.

The right and license granted herein is personal to you and is non-sublicensable, non-transferable and non-assignable by you to any other person or entity

**Attribution and Disclaimers**

Requestor must not remove the name of the author/creator when using Permitted Content in any Publication. Requestor shall acknowledge the source of the Permitted Content by including the following wording immediately below each publication of any Permitted Content in a Requestor Publication:
"Source: [Author/Creator] [Year] [Title] Used under license from Illumina, Inc. All Rights Reserved."; or such other form of words as may from time to time be reasonably specified by Illumina, Inc.

| Relationship Managers | |
|---|---|
| Illumina | none |
| Requestor | Alba Regueira Iglesias, albaregueira.iglesias@usc.es |

The parties each agree to be bound by this Agreement which shall come into effect on the date of the last signature below

| | |
|---|---|
| Illumina, Inc.<br>Signed: R. Schwillinski<br>Name: Roland Schwillinski<br>Title: VP Global IP & Litigation<br>Date: August 25, 2021 | Alba Regueira Iglesias<br>Signed:<br>Name: Alba Regueira Iglesias<br>Title: PhD Student. Predoctoral Researcher<br>Date: 25/08/2021 |

REVIEWED BY ILLUMINA LEGAL
(when digital signature is present)
Nichola Jamotillo
Digitally signed by Nichola Jamotillo
DN: cn=Nichola Jamotillo, o=Illumina, ou=Legal, emailAddress=njamotillo@illumina.com, c=US
Date: 2021.08.23 14:42:18 -07'00'

OXFORD UNIVERSITY PRESS LICENSE
TERMS AND CONDITIONS

Jan 18, 2022

---

This Agreement between Universidade de Santiago de Compostela -- Alba Regueira Iglesias ("You") and Oxford University Press ("Oxford University Press") consists of your license details and the terms and conditions provided by Oxford University Press and Copyright Clearance Center.

| | |
|---|---|
| License Number | 5232060421903 |
| License date | Jan 18, 2022 |
| Licensed content publisher | Oxford University Press |
| Licensed content publication | Human Molecular Genetics |
| Licensed content title | A window into third-generation sequencing |
| Licensed content author | Schadt, Eric E.; Turner, Steve |
| Licensed content date | Sep 21, 2010 |
| Type of Use | Thesis/Dissertation |
| Institution name | |
| Title of your work | Limitations of 16S rRNA gene as phylogenetic marker: a large-scale meta-omics analysis of plaque microbiota in periodontal diseases |
| Publisher of | Universidade de Santiago de Compostela |

SPRINGER NATURE LICENSE
TERMS AND CONDITIONS

Jan 18, 2022

---

This Agreement between Universidade de Santiago de Compostela -- Alba Regueira Iglesias ("You") and Springer Nature ("Springer Nature") consists of your license details and the terms and conditions provided by Springer Nature and Copyright Clearance Center.

| | |
|---|---|
| License Number | 5232070024167 |
| License date | Jan 18, 2022 |
| Licensed Content Publisher | Springer Nature |
| Licensed Content Publication | Applied Microbiology and Biotechnology |
| Licensed Content Title | 16S rRNA gene high-throughput sequencing data mining of microbial diversity and interactions |
| Licensed Content Author | Feng Ju et al |
| Licensed Content Date | Mar 27, 2015 |
| Type of Use | Thesis/Dissertation |
| Requestor type | academic/university or research institute |
| Format | print and electronic |
| Portion | figures/tables/illustrations |
| Number of figures/tables/illustrations | 1 |
| Will you be translating? | no |

OXFORD UNIVERSITY PRESS LICENSE
TERMS AND CONDITIONS

Jan 19, 2022

---

This Agreement between Universidade de Santiago de Compostela -- Alba Regueira Iglesias ("You") and Oxford University Press ("Oxford University Press") consists of your license details and the terms and conditions provided by Oxford University Press and Copyright Clearance Center.

| | |
|---|---|
| License Number | 5232420793324 |
| License date | Jan 19, 2022 |
| Licensed content publisher | Oxford University Press |
| Licensed content publication | Human Molecular Genetics |
| Licensed content title | Sequencing the human microbiome in health and disease |
| Licensed content author | Cox, Michael J.; Cookson, William O.C.M. |
| Licensed content date | Aug 13, 2013 |
| Type of Use | Thesis/Dissertation |
| Institution name | |
| Title of your work | Limitations of 16S rRNA gene as phylogenetic marker: a large-scale meta-omics analysis of plaque microbiota in periodontal diseases |
| Publisher of | Universidade de Santiago de Compostela |

ELSEVIER LICENSE
TERMS AND CONDITIONS

Jan 19, 2022

---

This Agreement between Universidade de Santiago de Compostela -- Alba Regueira Iglesias ("You") and Elsevier ("Elsevier") consists of your license details and the terms and conditions provided by Elsevier and Copyright Clearance Center.

| | |
|---|---|
| License Number | 5232451365955 |
| License date | Jan 19, 2022 |
| Licensed Content Publisher | Elsevier |
| Licensed Content Publication | Genomics |
| Licensed Content Title | High-throughput sequencing provides insights into oral microbiota dysbiosis in association with inflammatory bowel disease |
| Licensed Content Author | Ying Qi,Sheng-qi Zang,Juan Wei,Hong-chuan Yu,Zhao Yang,Hui-min Wu,Ying Kang,Hui Tao,Miao-fang Yang,Lei Jin,Ke Zen,Fang-yu Wang |
| Licensed Content Date | Jan 1, 2021 |
| Licensed Content Volume | 113 |
| Licensed Content Issue | 1 |
| Licensed Content Pages | 13 |
| Start Page | 664 |
| End Page | 676 |

ELSEVIER LICENSE
TERMS AND CONDITIONS

Jan 04, 2022

---

This Agreement between Universidade de Santiago de Compostela -- Alba Regueira Iglesias ("You") and Elsevier ("Elsevier") consists of your license details and the terms and conditions provided by Elsevier and Copyright Clearance Center.

| | |
|---|---|
| License Number | 5221890295571 |
| License date | Jan 04, 2022 |
| Licensed Content Publisher | Elsevier |
| Licensed Content Publication | Biotechnology Advances |
| Licensed Content Title | Using machine learning approaches for multi-omics data analysis: A review |
| Licensed Content Author | Parminder S. Reel,Smarti Reel,Ewan Pearson,Emanuele Trucco,Emily Jefferson |
| Licensed Content Date | July–August 2021 |
| Licensed Content Volume | 49 |
| Licensed Content Issue | n/a |
| Licensed Content Pages | 1 |
| Start Page | 107739 |
| End Page | 0 |
| Type of Use | reuse in a thesis/dissertation |

# Appendices Objective 1

Appendix S1. Words employed to find 16S rRNA gene primers used for detecting oral bacteria before sequencing.

Appendix S2. Words employed to find 16S rRNA gene primers used for detecting oral archaea before sequencing, and to elaborate a list of oral-archaea species.

Appendix S3. List of references from which we obtained at least one different 16S rRNA gene primer. doc. file stored in USB pendrive.

Appendix S4. List of references from which we obtained the archaeal species inhabiting different human mouth niches. doc. file stored in USB pendrive.

Appendix S5. Forward and reverse 16S rRNA gene primers evaluated in the study and the sequence comparison used to detect repeats. xlsx. file stored in USB pendrive.

Appendix S6. List of archaeal species present in the human mouth and the PMID of the investigations from which they were obtained. xlsx. file stored in USB pendrive.

Appendix S7. Oral-bacteria database of the 16S rRNA gene sequences used in the present study for the coverage analysis. fasta file stored in USB pendrive.

Appendix S8. 16S rRNA gene sequences from the oral archaea employed for the BLASTN search against the NCBI non-redundant nucleotide database. fasta file stored in USB pendrive.



Appendix S9. Oral-archaea database of the 16S rRNA gene sequences constructed by our group before alignment. fasta file stored in USB pendrive.

Appendix S10. Oral-archaea database of the 16S rRNA gene sequences used in the present study for the coverage analysis. fasta file stored in USB pendrive.

Appendix S11. Information related to the coverage analysis of the 16S rRNA gene individual primers.

Appendix S12. Evaluation of individual primers against the oral-bacteria database. xlsx. file stored in USB pendrive.

Appendix S13. Evaluation of individual primers against the oral-archaea database. xlsx. file stored in USB pendrive.

Appendix S14. Evaluation of primer pairs against the oral-bacteria and the oral-archaea databases. xlsx. file stored in USB pendrive.

Appendix S15. Coverage at the variant level of the selected primer pairs for detecting oral bacteria in different amplicon-length categories. xlsx. file stored in USB pendrive.

Appendix S16. Sequences of the selected primer pairs for detecting oral bacteria or/and archaea. xlsx. file stored in USB pendrive.

Appendix S17. Bacterial species non covered by the selected bacteria-specific primer pairs. xlsx. file stored in USB pendrive.

Appendix S18. Coverage at the variant level of the selected primer pairs for detecting oral archaea in different amplicon-length categories. xlsx. file stored in USB pendrive.

Appendix S19. Archaeal species non covered by the selected archaea-specific primer pairs. xlsx. file stored in USB pendrive.



Appendix S20. Coverage at the variant level of the selected primer pairs for detecting oral bacteria and archaea in different amplicon-length categories. xlsx. file stored in USB pendrive.

Appendix S21. Species non covered by the selected primer pairs for detecting both bacteria and archaea. xlsx. file stored in USB pendrive.

Appendix S22. Primer pairs utilised in the reviewed oral microbiome studies through 16S rRNA gene sequencing. xlsx. file stored in USB pendrive.

Appendix S23. Oral species non covered by KP_F078-OP_R010, KP_F047-KP_R035, and OP_F009-OP_R029. xlsx. file stored in USB pendrive.

**Appendix S1. Words employed to find 16S rRNA gene primers used for detecting oral bacteria before sequencing.**

| field1 | field2 | field3 | field4 |
|---|---|---|---|
| oral | microbiome | sequencing | region |
| mouth | microbiota | "sanger sequencing" | "hypervariable region" |
| caries | bacteria | "454 sequencing" | 16S |
| "periodontal disease*" | microbe | "454 pyrosequencing" | |
| periodontitis | "bacterial communit*" | Illumina | |
| gingivitis | "microbial communit*" | "ion torrent" | |
| oral cancer | "bacterial structure" | PacBio | |
| supragingival | "microbial structure" | | |
| subgingival | "bacterial ecology" | | |
| saliva | "microbial ecology" | | |
| tongue | | | |
| cheek | | | |
| "buccal mucosa" | | | |
| "oral mucosa" | | | |

Each search included one word from each of the four fields. Field one contains 14 words related to the oral cavity and its diseases, field two is constituted by 10 terms associated with the microbiota and the bacterial diversity, field three has seven words related to several types of sequencing platforms, and, finally, field four contains three terms associated with the 16S rRNA gene. All possible combinations between fields were performed, making a total of 2940 automated searches.

**Appendix S2. Words employed to find 16S rRNA gene primers used for detecting oral archaea before sequencing, and to elaborate a list of oral-archaea species.**

| field1 | field2 | field3 | field4 |
|---|---|---|---|
| oral | microbiome | sequencing | region |
| mouth | microbiota | "sanger sequencing" | "hypervariable region" |
| caries | microbe | "454 sequencing" | 16S |
| "periodontal disease*" | "microbial communit*" | "454 pyrosequencing" | |
| periodontitis | "microbial structure" | Illumina | |
| gingivitis | "microbial ecology" | "ion torrent" | |
| oral cancer | archaeome | PacBio | |
| supragingival | "archaeal communit*" | | |
| subgingival | "archaeal structure" | | |
| saliva* | "archaeal ecology" | | |
| tongue | archaea | | |
| cheek* | archaeal | | |
| "buccal mucosa" | | | |
| "oral mucosa" | | | |
| dental | | | |
| pulp | | | |
| peri-implantitis | | | |
| endodontic* | | | |
| periodontal | | | |
| "root canal" | | | |
| tooth | | | |
| teeth | | | |
| gingiva | | | |

Each search included one word from each of the four fields. Field one contains 23 words related to the oral cavity and its diseases, field two is constituted by 12 terms associated with the microbiota and the archaeal diversity, field three has seven words related to several types of sequencing platforms, and, lastly, field four contains three terms associated with the 16S rRNA gene. All possible combinations between the four fields were performed, meaning a total of 5796 automated searches.
To elaborate the list of archaeal species inhabiting the human mouth we used the fields one and two. The combination of these terms implied a total of 276 automated searches.

Appendix S11.

## MATERIAL AND METHODS

### Allocation of individual primers to the corresponding 16S rRNA gene regions

The location of the first and last nucleotides of each primer within each sequence with a match was calculated and the mode values for these positions were determined. If there was more than one mode for a position, we chose the one closest to the mean position value. As all the sequences in the two databases were aligned with the 16S rRNA *E. coli* gene, the mode values obtained for each primer enabled us to allocate them to one of the gene regions defined for that organism by Baker et al. (1). The reference sequence utilised had 1542 base pairs distributed in 10 conserved and nine hypervariable regions.

### Analysis of the coverage of individual primers

The script used for the VC analysis stored one xlsx file for each evaluated primer. In addition, the analysis produced a summary file in the same format as that which synthesised the results obtained for all the primers. Each line contained the information for each primer. The first columns included the primer identifier, direction, sequence, and length. These were followed by the number of allowed mismatches, the VC (%), the number of analysed sequences, and the number of sequences with and without matches. The file also included the mean and the mode positions of the first and last primer nucleotides in all the corresponding sequences, as well as the 16S rRNA gene region assigned to the initial and the end mode positions. Although we did not include any mismatches, the developed script allowed us to indicate the maximum number permitted.

The SC analysis used two Excel files for each evaluated primer: one with all the species for which at least one genomic variant matched the primer, and another with the non-matched or non-covered species. Again, a file in Excel format summarised the results obtained for all the primers assessed. The first five lines of the file were the same as those obtained from the VC analysis but only included: the SC (%), the number of analysed species, and the number of species with and without matches.

## RESULTS

### Evaluation of 16S rRNA gene individual primers for the detection of oral bacteria, archaea and both domains

Bacteria-specific individual primers

A total of 302 different individual primers (133 F, 169 R) had some coverage value for detecting oral bacteria (Appendix S11.1), while 67 (42 F and 25 R) provided no VC or SC. Fifty-nine primers (27 F, 32 R) localised in the gene regions 3, 4, 5, 6, 7 and 9 had bacterial SC values ≥95.00%. Thirteen of them (9 F, 4 R) had archaeal SC values of 0.00%. These bacteria-specific primers belonged to regions 3 and 7 and corresponded to: KP_F044, 046, 047, 048; OP_F048, 050, 096, 108, 116; KP_R018, 020; and OP_R054 and 116.

**Appendix S11.1. Individual primers within a particular bacterial coverage range in each 16S rRNA gene region.**

| Coverage (%) | 16S rRNA gene region | | | | | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | 1 | | 2 | | 3 | | 4 | | 5 | | 6 | | 7 | | 8 | | 9 | | 10 | |
| | Vr | Sp | Vr | Sp | Vr | Sp | Vr | Sp | Vr | Sp | Vr | Sp | Vr | Sp | Vr | Sp | Vr | Sp | Vr | Sp |
| ≥95 | 0 | 0 | 0 | 0 | 20 | 15 | 11 | 19 | 8 | 15 | 1 | 1 | 5 | 6 | 1 | 0 | 0 | 3 | 0 | 0 |
| ≥90 - <95 | 0 | 0 | 0 | 0 | 1 | 4 | 20 | 13 | 23 | 1 | 10 | 10 | 6 | 3 | 0 | 0 | 0 | 5 | 0 | 0 |
| ≥85 - <90 | 0 | 0 | 0 | 0 | 0 | 2 | 4 | 1 | 1 | 15 | 1 | 3 | 4 | 1 | 0 | 1 | 2 | 0 | 0 | 0 |
| ≥80 - <85 | 0 | 6 | 0 | 0 | 1 | 0 | 0 | 1 | 0 | 0 | 3 | 3 | 1 | 3 | 1 | 0 | 0 | 0 | 0 | 0 |
| ≥75 - <80 | 0 | 11 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 12 | 0 | 0 | 2 | 0 | 0 | 0 | 2 | 0 | 2 |
| <75% | 44 | 27 | 8 | 8 | 8 | 9 | 9 | 10 | 17 | 18 | 19 | 29 | 13 | 14 | 2 | 3 | 13 | 5 | 33 | 31 |
| Total | 44 | 44 | 8 | 8 | 30 | 30 | 44 | 44 | 49 | 49 | 46 | 46 | 29 | 29 | 4 | 4 | 15 | 15 | 33 | 33 |

Sp= species; Vr= variant.

Archaea-specific individual primers

One hundred and seventy-four individual primers (63 F, 111 R) had some coverage value for detecting oral archaea (Appendix S11.2). Conversely, 195 (112 F, 83 R) had an archaeal VC and SC of 0.00%. Thirty-three primers (7 F, 26 R) covered at least 95.00% of the oral-associated archaeal species in our database and were localised in gene regions 3, 5, 6, and 9 (Appendix S11.2). Of these, only KP_F016, KP_F018, KP_R006, and KP_R013 were specific to the archaea domain, with bacterial SC values of 0.00%. The two F primers belonged to gene region 3 and the two R to region 6.

**Appendix S11.2. Individual primers within a particular archaeal coverage range in each 16S rRNA gene region.**

| Coverage (%) | 16S rRNA gene region | | | | | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | 1 | | 2 | | 3 | | 4 | | 5 | | 6 | | 7 | | 8 | | 9 | | 10 | |
| | Vr | Sp | Vr | Sp | Vr | Sp | Vr | Sp | Vr | Sp | Vr | Sp | Vr | Sp | Vr | Sp | Vr | Sp | Vr | Sp |
| **≥95** | 0 | 0 | 0 | 0 | 0 | 7 | 0 | 0 | 0 | 14 | 0 | 3 | 0 | 0 | 0 | 0 | 0 | 9 | 0 | 0 |
| **≥90 - <95** | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 3 | 0 | 4 | 0 | 0 | 0 | 0 | 0 | 9 | 0 | 0 |
| **≥85 - <90** | 0 | 0 | 0 | 0 | 6 | 2 | 0 | 0 | 1 | 3 | 0 | 1 | 0 | 2 | 0 | 0 | 0 | 1 | 0 | 0 |
| **≥80 - <85** | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 19 | 0 | 7 | 3 | 0 | 0 | 0 | 0 | 9 | 0 | 0 | 0 |
| **≥75 - <80** | 0 | 0 | 0 | 0 | 3 | 2 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| **<75%** | 14 | 14 | 1 | 1 | 39 | 36 | 1 | 1 | 20 | 20 | 25 | 21 | 2 | 0 | 0 | 0 | 20 | 10 | 7 | 7 |
| **Total** | 14 | 14 | 1 | 1 | 48 | 48 | 1 | 1 | 40 | 40 | 32 | 32 | 2 | 2 | 0 | 0 | 29 | 29 | 7 | 7 |

Sp= species; Vr= variant.

Bacterial and archaeal individual primers

Nineteen different primers (4 F, 15 R), from gene regions 3, 5, and 9, had SC values ≥95.00% simultaneously in the two databases. The individual F and R primers with the best values in both domains were OP_F066 (bacterial SC= 99.48%, archaeal SC= 99.48%) and KP_R031 (bacterial SC= 99.35%, archaeal SC= 98.97%).

## DISCUSSION

### Comparative analysis of our coverage results of 16S rRNA gene primers with the literature

Bacteria-specific individual primers

Appendix S11.3 compares the results on individual primers analysed in both our research and the studies mentioned above. It is clear that our estimates of bacterial SC are similar to those of the other research, with differences no greater than 3.29% for: KP_F031; KP_F047; KP_R034; OP_R054; KP_F056; and KP_R053. The latter two primers achieved the highest overall coverage and specificity for bacteria in the study by Klindworth et al. (2); in our research, the best-performing primers for the bacteria domain were OP_F116, KP_F048, or KP_R020 (bacterial SC in order= 98.70%, 98,05%, 98,05%; archaeal SC= 0.00%). Moreover, while KP_R053 was also analysed by Ku et al. (3), the coverage values they obtained were poorer than those of both Klindworth (2) and in our study (Appendix S11.3); on the other hand, our bacterial coverage estimates for KP_F032, KP_F049, KP_R040, and KP_R075 were worse (Appendix S11.3). Nonetheless, the archaeal coverage of the latter primer in our database suggests it would be a good option for detecting both archaea and bacteria.

Archaea-specific individual primers

Concerning the archaeal coverage of individual primers, our coverage values were higher than those in the literature for KP_R003, KP_R005, and KP_F083, and similar for KP_F017 and KP_F082 (Appendix S11.3). Klindworth et al. (2) found that KP_F082 and KP_F083, which were previously regarded as targeting both bacteria and archaea, actually only targeted the latter. Although we agree with this about KP_F083, KP_F082 had a bacterial SC herein of 26.40%, which is more than four times higher than the coverage achieved by Klindworth. In any case, because of its poor bacterial coverage, we would not recommend these individual primers as a suitable option for only evaluating oral-archaeal species, or for studying the bacterial or archaeal domains together.



Finally, we confirmed the results obtained in other studies for KP_F020, OP_R014, and KP_R035, which are regarded as suitable for detecting both bacteria and archaea (Appendix S11.3). Although we found that OP_F014 and KP_F078 achieved reasonable archaeal SC, this was lower than the coverage described previously (Appendix S11.3). Conversely, our bacterial

estimates for KP_F078, and the bacterial and archaeal coverage of KP_R038, are better than those reported by Klindworth (2).

**Appendix S11.3. Coverage values obtained from the literature for the individual primers analysed in the present study.**

| Present study | Other studies | Results of the present study | | Results of the other studies | | |
|---|---|---|---|---|---|---|
| Primer identifier | Primer name | Bacterial SC (%) | Archaeal SC (%) | Bacterial coverage (%) | Archaeal coverage (%) | Ref. |
| KP_F031/OP_F023, 024, 034 | 8F | 66.19 | 0.00 | 62.90 | - | (3) |
| KP_F032/OP_F040 | 27F | 75.81 | 0.00 | 12.90 [e] | - | (4) |
| KP_F047/OP_F035 | 341F | 96.10 | 0.00 | 98.35 | 84.56 | (5) |
| KP_F049/OP_F038 | 347F | 88.43 | 0.00 | 93.60[a] | - | (6) |
| | | | | 91.10[bc]; 90.40[bd] | - | |
| KP_F056/OP_F083 | S-D-Bact-0564-a-S-15 | 98.31 | 8.76 | 96.00 | 16.30 | (2) |
| KP_R034/OP_R039 | 803R | 94.80 | 5.67 | 95.40[a] | - | (6) |
| | | | | 91.80[bc];84.90[bd] | - | |
| KP_R040 | 907R/926R | 90.9 | 0.00 | ~1.50 [e] | - | (4) |
| KP_R053/OP_R062 | 1061R | 96.62 | 3.61 | 75.80 | - | (3) |
| | S-D-Bact-1061-a-A-17 | | | 96.40 | 2.90 | (2) |
| OP_R054 | 338R | 96.10 | 0.00 | ~3.50 [e] | - | (4) |
| KP_F017/OP_F001 | 344F | 0.00 | 74.23 | - | 73.20 | (7) |
| KP_F082 | S-*-Univ-0789-a-S-18 | 26.40 | 88.66 | 6.80 | 86.10 | (2) |
| KP_F083 | S-*-Univ-0906-a-S-17 | 3.38 | 96.39 | 0.30 | 83.70 | (2) |
| KP_R003 | S-D-Arch-0519-a-A-19 | 10.27 | 98.97 | 0.10 | 91.30 | (2) |
| KP_R005/OP_R005, 061 | S-D-Arch-0786-a-A-20 | 26.14 | 97.94 | 7.80 | 87.40 | (2) |
| KP_F020/OP_F007 | 519F | 96.49 | 99.48 | 98.00 | 98.20 | (7) |
| KP_F078/OP_F005, 022 | S-*-Univ-0515-a-S-19 | 96.49 | 63.92 | 54.50 | 95.40 | (2) |
| KP_F081 | S-*-Univ-0779-a-S-20 | 0.00 | 0.00 | 0.00 | 0.00 | (2) |
| OP_F014/OP_F047 | 515F | 96.49 | 89.18 | 98.48 | 97.79 | (5) |
| KP_R035/OP_R035 | 805R | 97.79 | 98.97 | 98.17 | 98.40 | (5) |
| KP_R038 | S-D-Bact-0787-b-A-20 | 97.53 | 97.94 | 89.90 | 90.60 | (2) |
| KP_R075 | 1390R/1406R | 94.41 | 96.91 | ~1.00 [e] | - | (4) |
| OP_R014/OP_R120 | 806R | 97.53 | 98.45 | 97.50 | 98.43 | (5) |

The coverage values from the other investigations are those obtained when no mismatches were accepted.
a= foregut database; b= Ribosomal Database Project (RDP); c= species coverage; d= sequence coverage; e= non-coverage percentage; F= forward; KP= Klindworth primer; OP= oral primer; R= reverse; Ref.= references; SC= coverage at the species level.

**REFERENCES**

(1) Baker GC, Smith JJ, Cowan DA. Review and re-analysis of domain-specific 16S primers. J Microbiol Methods. 2003 Dec;55(3):541-55.

(2) Klindworth A, Pruesse E, Schweer T, Peplies J, Quast C, Horn M, et al. Evaluation of general 16S ribosomal RNA gene PCR primers for classical and next-generation sequencing-based diversity studies. Nucleic Acids Res. 2013 Jan;41(1):e1. doi: 10.1093/nar/gks808.

(3) Ku HJ, Lee JH. Development of a novel long-range 16S rRNA universal primer set for metagenomic analysis of gastrointestinal microbiota in newborn infants. J Microbiol Biotechnol. 2014 Jun;24(6):812-22.

(4) Mao DP, Zhou Q, Chen CY, Quan ZX. Coverage evaluation of universal bacterial primers using the metagenomic datasets. BMC Microbiol. 2012 May;12:66. doi: 10.1186/1471-2180-12-66.

(5) Wasimuddin, Schlaeppi K, Ronchi F, Leib SL, Erb M, Ramette A. Evaluation of primer pairs for microbiome profiling from soils to humans within the One Health framework. Mol Ecol Resour. 2020 Nov;20(6):1558-71.

(6) Nossa CW, Oberdorf WE, Yang L, Aas JA, Paster BJ, Desantis TZ, et al. Design of 16S rRNA gene primers for 454 pyrosequencing of the human foregut microbiome. World J Gastroenterol. 2010 Sep;16(33):4135-44.

(7) Pausan MR, Csorba C, Singer G, Till H, Schöpf V, Santigli E, et al. Exploring the archaeome: detection of archaeal signatures in the human body. Front Microbiol. 2019 Dec;10:2796. doi: 10.3389/fmicb.2019.02796.

## Appendices Objective 2

Appendix S1. NCBI taxonomy and identifiers of the oral-bacteria genomes. xlsx. file stored in USB pendrive.

Appendix S2. NCBI taxonomy and identifiers of the oral-archaea genomes. xlsx. file stored in USB pendrive.

Appendix S3. Sizes of the bacterial genomes and genes, the number of genes per genome, and the number of gene variants per genome across eight taxonomic ranks. xlsx. file stored in USB pendrive.

Appendix S4. Sizes of the archaeal genomes and genes, the number of genes per genome, and the number of gene variants per genome across eight taxonomic ranks. xlsx. file stored in USB pendrive.

Appendix S5. Species with and without matching amplicons and coverage estimators for all primer pairs tested. xlsx. file stored in USB pendrive.

Appendix S6. Overestimation factor of bacterial species using the bacteria-specific primer pairs. xlsx. file stored in USB pendrive.

Appendix S7. Overestimation factor of archaeal species using the archaea-specific primer pairs. xlsx. file stored in USB pendrive.



Appendix S8. Overestimation factor of the bacterial and archaeal species using the bacterial and archaeal primer pairs. xlsx. file stored in USB pendrive.

Appendix S9. Overestimation factor based on matching amplicons of bacterial species using the bacteria-specific primer pairs. xlsx. file stored in USB pendrive.

Appendix S10. Overestimation factor based on matching amplicons of archaeal species using the archaea-specific primer pairs. xlsx. file stored in USB pendrive.

Appendix S11. Overestimation factor based on matching amplicons of bacterial and archaeal species using the bacterial and archaeal primer pairs. xlsx. file stored in USB pendrive.

Appendix S12. Bacterial species with matching amplicons using the bacteria-specific primer pairs. xlsx. file stored in USB pendrive.

Appendix S13. Archaeal species with matching amplicons using the archaea-specific primer pairs. xlsx. file stored in USB pendrive.

Appendix S14. Bacterial and archaeal species with matching amplicons using the bacterial and archaeal primer pairs. xlsx. file stored in USB pendrive.

## Appendices Objective 3

Appendix S1. NCBI taxonomy and identifiers of the oral-bacteria genomes. xlsx. file stored in USB pendrive.

Appendix S2. NCBI taxonomy and identifiers of the oral-archaea genomes. xlsx. file stored in USB pendrive.

Appendix S3. Taxonomy of oral-bacteria species with *in silico* amplicon similarity values ≥97% with at least one different taxon, their mean number of intragenomic 16S rRNA genes, and taxa with a maximum number of *in silico* amplicon similarity values ≥97%/species ≥10. xlsx. file stored in USB pendrive.

Appendix S4. Taxonomy of oral-archaea species with *in silico* amplicon similarity values ≥97% with at least one different taxon, their mean number of intragenomic 16S rRNA genes, and taxa with a maximum number of *in silico* amplicon similarity values ≥97%/species ≥10. xlsx. file stored in USB pendrive.

Appendix S5. Taxonomy of oral-bacteria species with no *in silico* amplicon similarity values ≥97% with different taxa and their mean number of intragenomic 16S rRNA genes. xlsx. file stored in USB pendrive.

Appendix S6. Taxonomy of oral-archaea species with no *in silico* amplicon similarity values ≥97% with different taxa and their mean number of intragenomic 16S rRNA genes. xlsx. file stored in USB pendrive.



Appendix S7. Pairs of bacterial species with *in silico* amplicon similarity values ≥97% using the analysed primer pairs. xlsx. file stored in USB pendrive.

Appendix S8. Pairs of different bacterial genera, families, and orders with *in silico* amplicon similarity values ≥97%. xlsx. file stored in USB pendrive.

Appendix S9. Pairs of archaeal species with *in silico* amplicon similarity values ≥97% using the analysed primer pairs. xlsx. file stored in USB pendrive.

Appendix S10. Pairs of different archaeal genera, families, orders, and classes with *in silico* amplicon similarity values ≥97%. xlsx. file stored in USB pendrive.

# Appendices Objective 4

Appendix S1. Approval of the ethics committees.

Appendix S2. Search terms to identify Illumina sequencing-based studies on the periodontal microbiome. doc. file stored in USB pendrive.

Appendix S3. List of positive words in seven different sets used for punctuation of the candidate articles. xlsx. file stored in USB pendrive.

Appendix S4. Investigations obtained through the electronic databases searches and their exclusion reason if applicable. xlsx. file stored in USB pendrive.

Appendix S5. Investigations obtained through the bioproject database searches and their exclusion reason if applicable. xlsx. file stored in USB pendrive.

Appendix S6. List of the 32 articles that met the inclusion criteria and their exclusion reason if applicable. xlsx. file stored in USB pendrive.

Appendix S7. Characteristics of the sequencing-based studies on the periodontal microbiome included in the present meta-omics analysis. doc. file stored in USB pendrive.

Appendix S8. Taxa of the core microbiota in the different periodontal health conditions and dental plaque types.



Appendix S9. Relative abundance of the taxa of the core microbiota in the different periodontal health conditions and dental plaque types. xlsx. file stored in USB pendrive.

Appendix S10. Main taxa associated with each periodontal health condition and dental plaque type in the differential abundance analyses.

Appendix S11. Relative abundance and log2foldchange values of the main taxa associated with each periodontal health condition and dental plaque type in the differential abundance analyses. xlsx. file stored in USB pendrive.

Appendix S12. Taxa that formed part of the predictive models and the periodontal health condition they predicted in supragingival and subgingival plaque. xlsx. file stored in USB pendrive.

Appendix S1.

XUNTA DE GALICIA
CONSELLERÍA DE SANIDADE
Secretaría Xeral Técnica

Secretaria Técnica
Comité Autonómico de Ética da Investigación de Galicia
Secretaria Xeral. Consellería de Sanidade
Edificio Administrativo San Lázaro
15703 SANTIAGO DE COMPOSTELA
Tel: 881546425. Correo-e: ceic@sergas.es

galicia

**DICTAMEN DEL COMITÉ DE ÉTICA DE LA INVESTIGACIÓN DE SANTIAGO-LUGO**

Juan Manuel Vázquez Lago, Presidente del Comité de Ética de la Investigación de Santiago-Lugo,

**CERTIFICA:**

Que este Comité evaluó en su reunión del día 24/05/18 el estudio:

**Título:** Nuevos biomarcadores bacterianos, actividades funcionales, cambios dinámicos e interacciones bacterianas en la enfermedad periodontal crónica
**Versión:**
**Promotor/a**:Inmaculada Tomás Carmona
**Investigador/a:** Inmaculada Tomás Carmona
**Código de Registro**:2018/295

Y que este Comité, tomando en consideración la pertinencia del estudio, el conocimiento disponible, los requisitos legales aplicables y los Procedimientos Normalizados de Trabajo del Comité, emite un dictamen **FAVORABLE** para la realización del citado estudio.

XUNTA DE GALICIA
CONSELLERÍA DE SANIDADE
Secretaría Xeral Técnica

galicia

**Y HACE CONSTAR QUE:**

1.- El Comité Territorial de Ética de la Investigación de Santiago-Lugo cumple tanto en su composición como en sus PNTs los requisitos legales vigentes.

2.- La composición actual del Comité Territorial de Ética de la Investigación de Santiago-Lugo es:

- **Juan Manuel Vázquez Lago (Presidente).** Médico especialista en Medicina Preventiva y Salud Pública. Área de Gestión Integrada de Santiago.
- **Pilar Rodríguez Ledo (Vicepresidenta).** Médico especialista en Medicina Familiar y Comunitaria. Área de Gestión Integrada de Lugo.
- **Cristina Márquez Riveras (Secretaria).** Enfermera. Dirección Xeral de Saúde Pública.
- **Lorenzo Armenteros del Olmo (Secretario Suplente).** Médico especialista en Medicina Familiar y Comunitaria. Área de Gestión Integrada de Lugo.
- **Francisco Campos Pérez.** Biólogo. Fundación Instituto de Investigación Sanitaria de Santiago de Compostela.
- **Rosana Castelo Domínguez.** Farmacéutica de Atención Primaria. Área de Gestión Integrada de Santiago.
- **Ricardo García Martínez.** Licenciado en Derecho. Área de Gestión Integrada de Lugo.
- **Jaime Gulín Dávila**. Farmacéutico especialista en Farmacia Hospitalaria. Área de Gestión Integrada de Lugo.
- **Victor Herrán Carreira.** Paciente. ADIL-Asociación de Diabéticos Lucense.
- **María Jesús Lamas Díaz.** Farmacéutica especialista en Farmacia Hospitalaria. Área de Gestión Integrada de Santiago.
- **Guillermo José Prada Ramallal** Médico especialista en Farmacología Clínica. Área de Gestión Integrada de Santiago. Fundación Instituto de Investigación Sanitaria de Santiago de Compostela.
- **Carlos Rodríguez Moreno.** Médico especialista en Farmacología Clínica. Área de Gestión Integrada de Santiago.
- **Sandra Vidal Martínez.** Enfermera. Área de Gestión Integrada de Santiago
- **Rafael Carlos Vidal Pérez.** Médico especialista en Cardiología. Área de Gestión Integrada de Lugo.

Para que conste donde proceda, y a petición de quien proceda, en Santiago de Compostela,

El Presidente del Comité Territorial de Ética de la Investigación de Santiago Lugo,

Juan Manuel Vázquez Lago

VAZQUEZ LAGO JUAN MANUEL - 44829259M

XUNTA DE GALICIA
CONSELLERÍA DE SANIDADE
Secretaría Xeral Técnica

galicia

Se emite un informe FAVORABLE CONDICIONADO a que se tenga en cuenta lo siguiente:

1. Debe solucionarse la discrepancia en el tamaño muestral de los casos y controles: 60 controles y 30 casos según la página 7, y 30 + 30 en la página 10.

2. En la HIP debe especificarse qué pruebas son práctica clínica habitual (PCH) y cuáles no. Consecuentemente, en la HIP deben ajustarse los inconvenientes a las pruebas que no son PCH.

CESPU
INSTITUTO UNIVERSITÁRIO
DE CIÊNCIAS DA SAÚDE

Comissão de Ética
Instituto Universitário de Ciências da Saúde
Contacto: 224 157 136
E-mail: carla.ribeiro@cespu.pt

CARTA RESPOSTA

**Título do projeto:** Validação microbiológica de uma escala de saúde oral de potencialidade infecciosa-inflamatória mediante técnicas de metagenómica do gene ADNr 16S

**Investigador responsável:** Prof. Doutora Marta Mendonça Moutinho Relvas

**Nº Registo:** 35/CE-IUCS/2019

Parecer:

Exmo(a). Senhor(a),

Em resposta ao pedido efetuado por V. Exa. a esta Comissão de Ética, para emissão de parecer sobre o projeto de investigação supra identificado, somos a informar que, e de acordo com o regulamento, o mesmo recebeu parecer favorável por parte desta Comissão.

Gandra, 10 de de[illegible] de 2019

CESPU
INSTITUTO UNIVERSITÁRIO
DE CIÊNCIAS DA SAÚDE

Prof. Douto[illegible]
Presidente da C[illegible]

Rua Central de Gandra, 1317
4585-116 GANDRA PRD • Portugal
T. +351 224157100 • F. +351 224157101
www.cespu.pt

**Appendix S8. Taxa of the core microbiota in the different periodontal health conditions and dental plaque types.**

| ASVId | Genus | Species | ASV | Sup_x0HHx | Sup_x0GDx | Sup_x0PDx | Sup_x1PDx | Supragingival | Sub_x0HHx | Sub_x0PHx | Sub_x0PDx | Sub_x1PDx | Subgingival | Supra_Sub |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| ASV0206 | Actinomyces | massiliensis | BTASV213424 | | | | | | | | | | | |
| ASV0034 | Actinomyces | sp.HMT169 | BTASV057073 | | | | | | | | | | | |
| ASV0135 | Actinomyces | sp.HMT169 | unclassified | | | | | | | | | | | |
| ASV0360 | Actinomyces | sp.HMT169 | unclassified | | | | | | | | | | | |
| ASV0175 | Actinomyces | unclassified | unclassified | | | | | | | | | | | |
| ASV0107 | Bergeyella | sp.HMT322 | BTASV089026 | | | | | | | | | | | |
| ASV0080 | Campylobacter | concisus | BTASV137880 | | | | | | | | | | | |
| ASV0041 | Campylobacter | gracilis | BTASV155368 | | | | | | | | | | | |
| ASV0073 | Campylobacter | gracilis | unclassified | | | | | | | | | | | |
| ASV0020 | Campylobacter | rectus | BTASV188427 | | | | | | | | | | | |
| ASV0093 | Capnocytophaga | gingivalis | unclassified | | | | | | | | | | | |
| ASV0140 | Capnocytophaga | leadbetteri | unclassified | | | | | | | | | | | |
| ASV0056 | Capnocytophaga | sputigena | unclassified | | | | | | | | | | | |
| ASV0103 | Cardiobacterium | hominis | BTASV158475 | | | | | | | | | | | |
| ASV0179 | Catonella | sp.HMT164 | BTASV056976 | | | | | | | | | | | |
| ASV0102 | Corynebacterium | durum | BTASV140751 | | | | | | | | | | | |
| ASV0199 | Corynebacterium | durum | unclassified | | | | | | | | | | | |
| ASV0068 | Dialister | invisus | BTASV044355 | | | | | | | | | | | |
| ASV0361 | Eikenella | corrodens | unclassified | | | | | | | | | | | |
| ASV0019 | Filifactor | alocis | BTASV124203 | | | | | | | | | | | |
| ASV0097 | Fretibacterium | fastidiosum | unclassified | | | | | | | | | | | |
| ASV0014 | Fusobacterium | unclassified | unclassified | | | | | | | | | | | |
| ASV0044 | Fusobacterium | nucleatum_subsp.polymorphum | unclassified | | | | | | | | | | | |
| ASV0010 | Fusobacterium | nucleatum_subsp.vincentii | unclassified | | | | | | | | | | | |
| ASV0012 | Fusobacterium | nucleatum_subsp.vincentii | unclassified | | | | | | | | | | | |
| ASV0047 | Fusobacterium | nucleatum_subsp.vincentii | unclassified | | | | | | | | | | | |
| ASV0077 | Fusobacterium | nucleatum_subsp.vincentii | unclassified | | | | | | | | | | | |
| ASV0098 | Fusobacterium | nucleatum_subsp.vincentii | unclassified | | | | | | | | | | | |
| ASV0130 | Fusobacterium | nucleatum_subsp.vincentii | unclassified | | | | | | | | | | | |
| ASV0200 | Fusobacterium | nucleatum_subsp.vincentii | unclassified | | | | | | | | | | | |
| ASV0011 | Fusobacterium | periodonticum | BTASV066878 | | | | | | | | | | | |
| ASV0028 | Fusobacterium | unclassified | unclassified | | | | | | | | | | | |
| ASV0030 | Fusobacterium | unclassified | unclassified | | | | | | | | | | | |
| ASV0060 | Fusobacterium | unclassified | unclassified | | | | | | | | | | | |
| ASV0169 | Fusobacterium | unclassified | unclassified | | | | | | | | | | | |
| ASV0170 | Fusobacterium | unclassified | unclassified | | | | | | | | | | | |
| ASV0697 | Fusobacterium | unclassified | unclassified | | | | | | | | | | | |
| ASV0026 | Gemella | haemolysans | unclassified | | | | | | | | | | | |
| ASV0118 | Gemella | morbillorum | BTASV011336 | | | | | | | | | | | |
| ASV0136 | Gemella | morbillorum | unclassified | | | | | | | | | | | |
| ASV0013 | Granulicatella | adiacens | unclassified | | | | | | | | | | | |
| ASV0003 | Haemophilus | parainfluenzae | unclassified | | | | | | | | | | | |
| ASV0725 | Haemophilus | parainfluenzae | unclassified | | | | | | | | | | | |
| ASV1758 | Haemophilus | parainfluenzae | unclassified | | | | | | | | | | | |
| ASV0066 | Kingella | oralis | BTASV175208 | | | | | | | | | | | |
| ASV0152 | Lachnoanaerobaculum | umeaense | unclassified | | | | | | | | | | | |
| ASV0270 | Lachnoanaerobaculum | unclassified | unclassified | | | | | | | | | | | |
| ASV0017 | Lautropia | mirabilis | BTASV005234 | | | | | | | | | | | |

| | | | | |
|---|---|---|---|---|
| ASV0024 | Leptotrichia | hongkongensis | unclassified |  |
| ASV0033 | Leptotrichia | sp.HMT417 | BTASV102426 | |
| ASV0055 | Leptotrichia | wadei | BTASV075529 | |
| ASV0063 | Neisseria | elongata | unclassified | |
| ASV0009 | Neisseria | macacae | unclassified | |
| ASV0018 | Neisseria | perflava | unclassified | |
| ASV0035 | Neisseria | perflava | BTASV036108 | |
| ASV0165 | Oribacterium | sp.HMT078 | BTASV031505 | |
| ASV0021 | Parvimonas | sp.HMT110 | unclassified | |
| ASV0065 | Peptidiphaga | sp.HMT183 | unclassified | |
| ASV0124 | Peptostreptococcaceae [XI][G-4] | bacterium_HMT369 | BTASV096563 | |
| ASV0051 | Peptostreptococcaceae [XI][G-9] | brachy | BTASV129419 | |
| ASV0032 | Peptostreptococcus | stomatis | unclassified | |
| ASV0023 | Porphyromonas | endodontalis | unclassified | |
| ASV0008 | Porphyromonas | gingivalis | BTASV154369 | |
| ASV0031 | Porphyromonas | pasteri | BTASV079927 | |
| ASV0007 | Prevotella | melaninogenica | BTASV111236 | |
| ASV0025 | Prevotella | nigrescens | unclassified | |
| ASV0144 | Prevotella | oris | BTASV085458 | |
| ASV0006 | Rothia | aeria | BTASV063887 | |
| ASV0002 | Rothia | dentocariosa | BTASV138915 | |
| ASV0081 | Ruminococcaceae [G-1] | bacterium_HMT075 | unclassified | |
| ASV0050 | Sacchari | BTASV095406 | unclassified | |
| ASV0036 | Sacchari | BTASV095869 | unclassified | |
| ASV0037 | Sacchari | BTASV096126 | unclassified | |
| ASV0223 | Selenomonas | sp.HMT136 | BTASV049657 | |
| ASV0243 | Selenomonas | sputigena | unclassified | |
| ASV0197 | Solobacterium | moorei | BTASV171373 | |
| ASV0062 | Streptococcus | intermedius | BTASV162089 | |
| ASV0001 | Streptococcus | oralis_subsp.dentisani_clade_058 | BTASV016027 | |
| ASV0320 | Streptococcus | oralis_subsp.dentisani_clade_058 | unclassified | |
| ASV0560 | Streptococcus | oralis_subsp.dentisani_clade_058 | unclassified | |
| ASV0877 | Streptococcus | oralis_subsp.dentisani_clade_058 | unclassified | |
| ASV0962 | Streptococcus | oralis_subsp.dentisani_clade_058 | unclassified | |
| ASV1534 | Streptococcus | oralis_subsp.dentisani_clade_058 | unclassified | |
| ASV1622 | Streptococcus | oralis_subsp.dentisani_clade_058 | unclassified | |
| ASV0228 | Streptococcus | sanguinis | unclassified | |
| ASV0392 | Streptococcus | sanguinis | unclassified | |
| ASV0004 | Streptococcus | unclassified | unclassified | |
| ASV0022 | Streptococcus | unclassified | unclassified | |
| ASV0029 | Streptococcus | unclassified | unclassified | |
| ASV0042 | Streptococcus | unclassified | unclassified | |
| ASV0064 | Streptococcus | unclassified | unclassified | |
| ASV0071 | Streptococcus | unclassified | unclassified | |
| ASV0085 | Streptococcus | unclassified | unclassified | |
| ASV0090 | Streptococcus | unclassified | unclassified | |
| ASV0121 | Streptococcus | unclassified | unclassified | |
| ASV0155 | Streptococcus | unclassified | unclassified | |
| ASV0282 | Streptococcus | unclassified | unclassified | |
| ASV0541 | Streptococcus | unclassified | unclassified | |
| ASV0712 | Streptococcus | unclassified | unclassified | |
| ASV0862 | Streptococcus | unclassified | unclassified | |
| ASV0916 | Streptococcus | unclassified | unclassified | |
| ASV0944 | Streptococcus | unclassified | unclassified | |
| ASV0950 | Streptococcus | unclassified | unclassified | |
| ASV1000 | Streptococcus | unclassified | unclassified | |

| | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| ASV1012 | Streptococcus | unclassified | unclassified | | | | | | | | | | | |
| ASV1462 | Streptococcus | unclassified | unclassified | | | | | | | | | | | |
| ASV1597 | Streptococcus | unclassified | unclassified | | | | | | | | | | | |
| ASV2228 | Streptococcus | unclassified | unclassified | | | | | | | | | | | |
| ASV0027 | Streptococcus | vestibularis | BTASV004875 | | | | | | | | | | | |
| ASV0015 | Tannerella | forsythia | BTASV153103 | | | | | | | | | | | |
| ASV0141 | Tannerella | sp.HMT286 | unclassified | | | | | | | | | | | |
| ASV0038 | Treponema | denticola | BTASV138814 | | | | | | | | | | | |
| ASV0195 | Treponema | unclassified | unclassified | | | | | | | | | | | |
| ASV0005 | Veillonella | dispar | BTASV053366 | | | | | | | | | | | |
| ASV0016 | Veillonella | dispar | BTASV053367 | | | | | | | | | | | |
| ASV0089 | Veillonella | dispar | unclassified | | | | | | | | | | | |
| ASV0131 | Veillonella | dispar | unclassified | | | | | | | | | | | |
| ASV1765 | Veillonella | dispar | BTASV053366 | | | | | | | | | | | |
| ASV2353 | Veillonella | dispar | BTASV053366 | | | | | | | | | | | |
| ASV0057 | Veillonella | unclassified | unclassified | | | | | | | | | | | |
| ASV0075 | Veillonella | unclassified | unclassified | | | | | | | | | | | |
| ASV1320 | Veillonella | unclassified | unclassified | | | | | | | | | | | |

Cells are coloured if the taxon was part of the core microbiota of the group in question.
ASV= amplicon sequence variant; ASVid= amplicon sequence variant identifier; Sub_x0HHx= subgingival plaque of periodontally healthy subjects, healthy sites; Sub_x0PHx= subgingival plaque of periodontitis subjects, healthy sites; Sub_x0PDx= subgingival plaque of periodontitis subjects, diseased sites; Sub_x1PDx= subgingival plaque of periodontitis subjects, diseased sites after therapy; Sup_x0GDx= supragingival plaque of gingivitis subjects, diseased sites; Sup_x0HHx= supragingival plaque of periodontally healthy subjects, healthy sites; Sup_x0PDx= supragingival plaque of periodontitis subjects, diseased sites; Sup_x1PDx= supragingival plaque of periodontitis subjects, diseased sites after therapy; Sub= subgingival; Sup= Supragingival.

Appendix S10.

In the appendix S10.1 and appendix S10.2 tables, cells are coloured according to the periodontal health condition in which the taxon was most abundant. The green colour was associated with periodontally healthy subjects, healthy sites; yellow with gingivitis, diseased sites; pink with periodontitis, healthy sites; red with periodontitis, diseased sites; and orange with periodontitis, diseased sites after therapy. In the appendix S10.3 table, cells are coloured according to the dental plaque in which the taxon was most abundant. Blue was associated with supragingival and green with subgingival plaque.

**Appendix S10.1. Main taxa associated with each periodontal health condition in supragingival plaque.**

| ASVId | Genus | Species | ASV | x0HHx vs. x0GDx | x0HHx vs. x0PDx | x0HHx vs. x1PDx | x0GDx vs. x0PDx | x0GDx vs. x1PDx | x0PDx vs. x1PDx |
|---|---|---|---|---|---|---|---|---|---|
| ASV0034 | Actinomyces | sp.HMT169 | BTASV057073 | C | | C | C | | C |
| ASV0119 | Actinomyces | sp.HMT170 | BTASV057334 | | | | | | |
| ASV0110 | Aggregatibacter | unclassified | unclassified | | | | | | |
| ASV0128 | Alloprevotella | tannerae | unclassified | | | | | | |
| ASV0020 | Campylobacter | rectus | BTASV188427 | | | C | | C | |
| ASV0056 | Capnocytophaga | sputigena | unclassified | | C | | C | | C |
| ASV0068 | Dialister | invisus | BTASV044355 | | | C | | | |
| ASV0019 | Filifactor | alocis | BTASV124203 | | | | | | |
| ASV0010 | Fusobacterium | nucleatum_subsp.vincentii | unclassified | | C | C | | | |
| ASV0047 | Fusobacterium | nucleatum_subsp.vincentii | unclassified | C | | C | | | |
| ASV0499 | Fusobacterium | periodonticum | unclassified | | | | | | |
| ASV0060 | Fusobacterium | unclassified | unclassified | C | | | | | |
| ASV0099 | Fusobacterium | unclassified | unclassified | | | | | | |
| ASV0013 | Granulicatella | adiacens | unclassified | | | | C | | C |
| ASV0003 | Haemophilus | parainfluenzae | unclassified | C | C | C | | | |
| ASV0045 | Haemophilus | parainfluenzae | unclassified | | | | | | |
| ASV0078 | Haemophilus | parainfluenzae | unclassified | | | | | | |
| ASV0066 | Kingella | oralis | BTASV175208 | | | C | | | |
| ASV0017 | Lautropia | mirabilis | BTASV005234 | C | | C | C | | C |
| ASV0067 | Lautropia | mirabilis | unclassified | | | | | | |
| ASV0024 | Leptotrichia | hongkongensis | unclassified | C | | | C | C | |
| ASV0033 | Leptotrichia | sp.HMT417 | BTASV102426 | | | | | | |
| ASV0063 | Neisseria | elongata | unclassified | C | | | | | |
| ASV0009 | Neisseria | macacae | unclassified | C | | | C | C | |
| ASV0092 | Neisseria | unclassified | unclassified | | | | | | |
| ASV0021 | Parvimonas | sp.HMT110 | unclassified | | C | | | | |
| ASV0065 | Peptidiphaga | sp.HMT183 | unclassified | | | C | | | C |
| ASV0032 | Peptostreptococcus | stomatis | unclassified | | C | | C | | C |
| ASV0069 | Porphyromonas | endodontalis | unclassified | | | | | | |
| ASV0008 | Porphyromonas | gingivalis | BTASV154369 | | | | | | |
| ASV0025 | Prevotella | nigrescens | unclassified | | | C | | C | |
| ASV0086 | Prevotella | nigrescens | BTASV174229 | | | | | | |
| ASV0868 | Prevotella | pallens | unclassified | | | | | | |
| ASV0006 | Rothia | aeria | BTASV063887 | C | | C | C | | C |
| ASV0002 | Rothia | dentocariosa | BTASV138915 | C | | C | C | | |
| ASV0050 | Sacchari | | BTASV095406 | C | | | | | |
| ASV0243 | Selenomonas | sputigena | unclassified | | | | | | |
| ASV0001 | Streptococcus | oralis_subsp.dentisani_clade_058 | BTASV016027 | C | | C | C | | C |
| ASV0114 | Streptococcus | oralis_subsp.dentisani_clade_058 | unclassified | | | | | | |
| ASV0004 | Streptococcus | unclassified | unclassified | | | C | | C | C |
| ASV0022 | Streptococcus | unclassified | unclassified | | C | | C | | C |
| ASV0085 | Streptococcus | unclassified | unclassified | C | | C | | | |
| ASV0090 | Streptococcus | unclassified | unclassified | C | | C | | | |
| ASV0027 | Streptococcus | vestibularis | BTASV004875 | | | | | | |
| ASV0015 | Tannerella | forsythia | BTASV153103 | | | C | | | |
| ASV0038 | Treponema | denticola | BTASV138814 | | | C | | | |
| ASV0057 | Veillonella | unclassified | unclassified | C | | C | C | | |
| ASV0075 | Veillonella | unclassified | unclassified | C | | | | | |

ASV= amplicon sequence variant; ASVid= amplicon sequence variant identifier; C= core member; x0GDx= gingivitis subjects, diseased sites; x0HHx= periodontally healthy subjects, healthy sites; x0PDx= periodontitis subjects, diseased sites; x1PDx= periodontitis subjects, diseased sites after therapy.

**Appendix S10.2. Main taxa associated with each periodontal health condition in subgingival plaque.**

| ASVId | Genus | Species | ASV | x0HHx vs. x0PHx | x0HHx vs. x0PDx | x0HHx vs. x1PDx | x0PHx vs. x0PDx | x0PHx vs. x1PDx | x0PDx vs. x1PDx |
|---|---|---|---|---|---|---|---|---|---|
| ASV0034 | Actinomyces | sp.HMT169 | BTASV057073 | | | | | | |
| ASV0049 | Bacteroidales [G-2] | bacterium_HMT274 | unclassified | | | | | | |
| ASV0061 | Bacteroidales [G-2] | bacterium_HMT274 | unclassified | | | | | | |
| ASV0020 | Campylobacter | rectus | BTASV188427 | C | | | C | C | |
| ASV0133 | Cupriavidus | gilardii | unclassified | | | | | | |
| ASV0116 | Escherichia | coli | BTASV133996 | | | | | | |
| ASV0019 | Filifactor | alocis | BTASV124203 | | C | | C | | C |
| ASV0044 | Fusobacterium | nucleatum_subsp.polymorphum | unclassified | | C | C | | | |
| ASV0010 | Fusobacterium | nucleatum_subsp.vincentii | unclassified | | C | | C | | |
| ASV0012 | Fusobacterium | nucleatum_subsp.vincentii | unclassified | | | C | | | C |
| ASV0059 | Fusobacterium | nucleatum_subsp.vincentii | BTASV066150 | | | | | | |
| ASV0014 | Fusobacterium | unclassified | unclassified | | C | C | | | |
| ASV0028 | Fusobacterium | unclassified | unclassified | | | | | | |
| ASV0030 | Fusobacterium | unclassified | unclassified | | | | | | |
| ASV0074 | Fusobacterium | unclassified | unclassified | | | | | | |
| ASV0440 | unclassified | unclassified | unclassified | | | | | | |
| ASV0026 | Gemella | haemolysans | unclassified | | C | C | C | | |
| ASV0013 | Granulicatella | adiacens | unclassified | | C | C | | | |
| ASV0003 | Haemophilus | parainfluenzae | unclassified | C | C | C | | | |
| ASV0066 | Kingella | oralis | BTASV175208 | | | | | C | |
| ASV0017 | Lautropia | mirabilis | BTASV005234 | | | | | | |
| ASV0067 | Lautropia | mirabilis | unclassified | | | | | | |
| ASV0009 | Neisseria | macacae | unclassified | | C | C | | | |
| ASV0051 | Peptostreptococcaceae [XI][G-9] | brachy | BTASV129419 | | | | | | C |
| ASV0008 | Porphyromonas | gingivalis | BTASV154369 | | | | | | |
| ASV0104 | Porphyromonas | gingivalis | unclassified | | | | | | |
| ASV0025 | Prevotella | nigrescens | unclassified | | | | | | |
| ASV0108 | Pseudomonas | fluorescens | unclassified | | | | | | |
| ASV0002 | Rothia | dentocariosa | BTASV138915 | | C | | C | C | |
| ASV0037 | Sacchari | BTASV096126 | unclassified | | | | | | |
| ASV0062 | Streptococcus | intermedius | BTASV162089 | | C | | | | |
| ASV0105 | Streptococcus | mutans | BTASV173594 | | | | | | |
| ASV0001 | Streptococcus | oralis_subsp.dentisani_clade_058 | BTASV016027 | | C | C | C | C | |
| ASV0004 | Streptococcus | unclassified | unclassified | | C | | C | | |
| ASV0022 | Streptococcus | unclassified | unclassified | | | | C | | |
| ASV0042 | Streptococcus | unclassified | unclassified | | | | C | | |
| ASV0071 | Streptococcus | unclassified | unclassified | | | | | | |
| ASV0027 | Streptococcus | vestibularis | BTASV004875 | | | | C | C | |
| ASV0015 | Tannerella | forsythia | BTASV153103 | | C | | C | | C |
| ASV0038 | Treponema | denticola | BTASV138814 | | | | | | |
| ASV0005 | Veillonella | dispar | BTASV053366 | | | | C | | |
| ASV0089 | Veillonella | dispar | unclassified | | | | | C | |
| ASV0131 | Veillonella | dispar | unclassified | | | | C | C | |

ASV= amplicon sequence variant; ASVid= amplicon sequence variant identifier; C= core member; x0HHx= periodontally healthy subjects, healthy sites; x0PDx= periodontitis subjects, diseased sites; x0PHx= periodontitis subjects, healthy sites; x1PDx= periodontitis subjects, diseased sites after therapy.

**Appendix S10.3. Main taxa associated with each dental plaque in the same periodontal health condition.**

| | | | Supragingival | | |
|---|---|---|---|---|---|
| | | | Subgingival | | |
| ASVId | Genus | Species | ASV | x0HHx | x0PDx | x1PDx |
| ASV0034 | Actinomyces | sp.HMT169 | BTASV057073 | | C | |
| ASV0135 | Actinomyces | sp.HMT169 | unclassified | | | |
| ASV0119 | Actinomyces | sp.HMT170 | BTASV057334 | | | |
| ASV0049 | Bacteroidales [G-2] | bacterium_HMT274 | unclassified | | | |
| ASV0061 | Bacteroidales [G-2] | bacterium_HMT274 | unclassified | | | |
| ASV0041 | Campylobacter | gracilis | BTASV155368 | C | | |
| ASV0020 | Campylobacter | rectus | BTASV188427 | C | | |
| ASV0056 | Capnocytophaga | sputigena | unclassified | | C | |
| ASV0133 | Cupriavidus | gilardii | unclassified | | | |
| ASV0019 | Filifactor | alocis | BTASV124203 | | C | |
| ASV0012 | Fusobacterium | nucleatum_subsp.vincentii | unclassified | C | | |
| ASV0059 | Fusobacterium | nucleatum_subsp.vincentii | BTASV066150 | | | |
| ASV0014 | Fusobacterium | unclassified | unclassified | | C | |
| ASV0028 | Fusobacterium | unclassified | unclassified | | | |
| ASV0074 | Fusobacterium | unclassified | unclassified | | | |
| ASV0096 | Fusobacterium | unclassified | unclassified | | | |
| ASV0026 | Gemella | haemolysans | unclassified | C | | |
| ASV0013 | Granulicatella | adiacens | unclassified | C | | |
| ASV0003 | Haemophilus | parainfluenzae | unclassified | | C | |
| ASV0045 | Haemophilus | parainfluenzae | unclassified | | | |
| ASV0017 | Lautropia | mirabilis | BTASV005234 | | C | |
| ASV0053 | Lautropia | mirabilis | BTASV005235 | | | |
| ASV0067 | Lautropia | mirabilis | unclassified | | | |
| ASV0024 | Leptotrichia | hongkongensis | unclassified | C | C | |
| ASV0009 | Neisseria | macacae | unclassified | C | | |
| ASV0021 | Parvimonas | sp.HMT110 | unclassified | C | | |
| ASV0065 | Peptidiphaga | sp.HMT183 | unclassified | | C | |
| ASV0008 | Porphyromonas | gingivalis | BTASV154369 | | | |
| ASV0104 | Porphyromonas | gingivalis | unclassified | | | |
| ASV0025 | Prevotella | nigrescens | unclassified | | | |
| ASV0108 | Pseudomonas | fluorescens | unclassified | | | |
| ASV0006 | Rothia | aeria | BTASV063887 | C | C | |
| ASV0002 | Rothia | dentocariosa | BTASV138915 | | C | |
| ASV0062 | Streptococcus | intermedius | BTASV162089 | C | | |
| ASV0001 | Streptococcus | oralis_subsp.dentisani_clade_058 | BTASV016027 | | C | |
| ASV0004 | Streptococcus | unclassified | unclassified | | C | |
| ASV0022 | Streptococcus | unclassified | unclassified | | C | C |
| ASV0085 | Streptococcus | unclassified | unclassified | C | | |
| ASV0090 | Streptococcus | unclassified | unclassified | C | | |
| ASV0027 | Streptococcus | vestibularis | BTASV004875 | | | |
| ASV0015 | Tannerella | forsythia | BTASV153103 | | C | |
| ASV0038 | Treponema | denticola | BTASV138814 | | | |

ASV= amplicon sequence variant; ASVid= amplicon sequence variant identifier; C= core member; x0HHx= periodontally healthy subjects, healthy sites; x0PDx= periodontitis subjects, diseased sites; x1PDx= periodontitis subjects, diseased sites after therapy.

# Appendix: Publications derived from this Thesis

1. Relvas M[1], Regueira-Iglesias A[2], Balsa-Castro C[2], Salazar F[1], Pacheco JJ[1], Cabral C[1], Henriques C[1], Tomás I[2]. Relationship between dental and periodontal health status and the salivary microbiome: bacterial diversity, co-occurrence networks and predictive models. Sci Rep. 2021 Jan;11(1):929. doi: 10.1038/s41598-020-79875-x. ISSN: 2045-2322.

   1-Institute of Research and Advanced Training in Health Sciences and Technologies (IINFACTS), IUCS-Cespu-Instituto Universitário de Ciencias da Saúde; Gandra, Paredes, Portugal.
   2-Oral Sciences Research Group, Special Needs Unit, Department of Surgery and Medical-Surgical Specialties, School of Medicine and Dentistry, Universidade de Santiago de Compostela, Health Research Institute Foundation of Santiago (FIDIS); Santiago de Compostela, Spain.

The introduction (I.6) of the present Thesis partially reproduces the content of the publication: "Relationship between dental and periodontal health status and the salivary microbiome: bacterial diversity, co-occurrence networks and predictive models". Contribution of the PhD student: data acquisition, analysis, and interpretation, and drafted the manuscript.

Quality indexes: Impact factor: 4.380 (2020); Ranking: Q1, 17/72 (Category: Multidisciplinary Sciences).

2. Regueira-Iglesias A[1], Vázquez-González L[2], Balsa-Castro C[1], Vila-Blanco N[2], Blanco-Pintos T[1], Tamames J[3], Carreira MJ[2], Tomás I[1]. In silico evaluation and selection of the best 16S rRNA gene primers for use in next-generation sequencing to detect oral bacteria and archaea. Accepted for publication in Microbiome. Preprint at Research Square. 2021. doi: 10.21203/rs.3.rs-516961/v1. ISSN: 2049-2618.

1-Oral Sciences Research Group, Special Needs Unit, Department of Surgery and Medical-Surgical Specialties, School of Medicine and Dentistry, Universidade de Santiago de Compostela, Health Research Institute Foundation of Santiago (FIDIS); Santiago de Compostela, Spain.
2-Centro Singular de Investigación en Tecnoloxías Intelixentes and Departamento de Electrónica e Computación, Universidade de Santiago de Compostela; Health Research Institute Foundation of Santiago (FIDIS); Santiago de Compostela, Spain.
3-Microbiome Analysis Laboratory, Systems Biology Department, Centro Nacional de Biotecnología (CNB)-Consejo Superior de Investigaciones Científicas (CSIC); Madrid, Spain.

The objective 1 of the present Thesis reproduces the content of the publication: "*In silico* evaluation and selection of the best 16S rRNA gene primers for use in next-generation sequencing to detect oral bacteria and archaea". Contribution of the PhD student: data acquisition, analysis, and interpretation, and drafted the manuscript.

Quality indexes: Impact factor: 14.652 (2020); Ranking: D1, 8/136 (Category: Microbiology).

3. Regueira-Iglesias A[1], Vázquez-González L[2], Balsa-Castro C[1], Blanco-Pintos T[1], Martín-Biedma B[1], Arce VM[3], Carreira MJ[2], Tomás I[1]. In silico detection of oral prokaryotic species with highly similar 16S rRNA sequence segments using different primer pairs. Front Cell Infect Microbiol. 2022 Feb;11:770668. doi: 10.3389/fcimb.2021.770668. ISSN: 2235-2988.

1-Oral Sciences Research Group, Special Needs Unit, Department of Surgery and Medical-Surgical Specialties, School of Medicine and Dentistry, Universidade de Santiago de Compostela, Health Research Institute Foundation of Santiago (FIDIS); Santiago de Compostela, Spain.
2-Centro Singular de Investigación en Tecnoloxías Intelixentes and Departamento de Electrónica e Computación, Universidade de Santiago de Compostela; Health Research Institute Foundation of Santiago (FIDIS); Santiago de Compostela, Spain.
3-Department of Physiology and Center for Disease in Molecular Medicine and Chronic Diseases, Universidade de Santiago de Compostela; Santiago de Compostela, Spain.

The objective 3 of the present Thesis reproduces the content of the publication: "*In silico* detection of oral prokaryotic species with highly similar 16S rRNA sequence segments using different primer pairs". Contribution of the PhD student: data acquisition, analysis, and interpretation, and drafted the manuscript.

Impact factor: 5.293 (2020)
Ranking: Q1, 33/136 (Category: Microbiology)

# ABBREVIATIONS

# Glossary of Abbreviations

All abbreviations that have been used in this Thesis are listed here, excluding those used only in figures/tables of the main text and appendices, which are defined in the corresponding legend. Some of them were written in plural, for which the corresponding abbreviation was incorporated an "s" in lower case.

A: archaea
ACE: abundance-based coverage estimator
ANOSIM: analysis of similarities
APS: adenosine 5’ phosphosulfate
ASI97: amplicons with similarity value ≥97%
ASS: average sequence score
ASV(s): amplicon sequence variant(s)
ATP: adenosine triphosphate
AUC: area under the curve
B: bacteria
BC: betweenness centrality
BOP: bleeding on probing
Bp(s): base pair(s)
BPL: bacterial plaque level
CAL: clinical attachment level
CC: closeness centrality
cDNA: complementary deoxyribonucleic acid
CESPU: Cooperativa de Ensino Superior, Politécnico e Universitário
Chip: chromatin immunoprecipitation
CPD: classification of periodontal diseases
CRP: C-reactive protein
DA(s): discriminant analysis/es

DC: degree centrality

ddNTP(s): dideoxynucleotide-triphosphate(s)

DESeq2: differential expression analysis for sequence count data version 2

DFA: discriminant function analysis

DG: dental grade

DNA: deoxyribonucleic acid

dNTP(s): deoxyribonucleotide-triphosphate(s)

doi: digital object identifier

EC: eigenvector centrality

eHOMD: expanded human oral microbiome database

F: forward

Fc-cRIIAFc: gamma receptor IIA

FDR: false discovery rate

FISSEQ: fluorescent in-situ sequencing

Gbps: giga base pairs

GC: guanine-cytosine

GCF: gingival crevicular fluid

GOS: global ocean sampling

GS: genome sequencer

GTDB: genome taxonomy database

GUI: graphical user interface

HMP: human microbiome project

HOMD: human oral microbiome database

HOMIM: human oral microbe identification microarray

IFN: gammainterferon gamma

IL: interleukin

Imp_x0IDx: submucosal plaque of the peri-implantitis subjects - diseased sites

Imp_x0IHx: submucosal plaque of the peri-implantitis subjects - healthy sites

IMPEDE: inflammation-mediated polymicrobial-emergence and dysbiotic-exacerbation

ISS: in-situ sequencing

KP: Klindworth primer

L: long mean amplicon length category, >600 base pairs

LCA: lowest common ancestor

LDA: linear discriminant analysis

LDTM: liquid dental transport medium

LEfSe: effect size

log2FC: log2foldchange

M: medium mean amplicon length category, 301-600 base pairs

MA(s): matching amplicon(s)

Max.: maximum

Mbp(s): mega base pair(s)

MG-RAST: metagenomic rapid annotation using subsystem technology

MHC: major histocompatibility complex

Min.: minimum

ML: machine learning

MMP: matrix metalloproteinase

mRNA: messenger ribonucleic acid

MRPP: multi-response permutation procedure

NCBI: National Centre for Biotechnology Information

NGS: next-generation sequencing

NHI: National Institutes of Health

NLP: natural language processing

NMDS: non-metric multidimensional scaling

OBL(s): osteoblast(s)

OCL(s): osteoclast(s)

OF: overestimation factor

OF-MA: overestimation factor associated with matching amplicons

OG: oral grade

ONT: Oxford Nanopore Technology

OP: oral primer

OTU(s): operational taxonomic unit(s)

PacBio: Pacific Biosciences

PBS: phosphate-buffered saline

PC: principal component

PCA: principal component analysis

PCoA: principal coordinate analysis

PCR: polymerase chain reaction

PD: phylogenetic diversity

PERMANOVA: permutational multivariate analysis of variance

PG: periodontal grade

PINA: phylogenetic interaction-adjusted

PMID(s): PubMed unique identifier(s)

PPD: probing pocket depth

Ppi: pyrophosphate

PSD: polymicrobial synergy and dysbiosis

QIIME: quantitative insights into microbial ecology

qPCR: quantitative polymerase chain reaction

R: reverse

RANK: receptor activator of the nuclear factor-k B

RANKL: receptor activator of the nuclear factor-k B ligand

RDA: redundancy analysis

RDP: ribosomal database project

RNA: ribonucleic acid

ROC: receiver operating characteristic

rRNA: ribosomal ribonucleic acid

rrnDB: ribosomal ribonucleic acid operon copy number database

RT-PCR: real-time polymerase chain reaction

S: short mean amplicon length category, 100-300 base pairs

SC: coverage at the species level, i.e., species coverage

SC-NASI97: species coverage with no amplicons with similarity value ≥97%

SC-NMA: species coverage with no matching amplicons

SMRT: single-molecule real-time

SMS: single-molecule sequencing

SNP(s): single-nucleotide polymorphism(s)

SNV(s): single-nucleotide variation(s)

SparCC: sparse correlations for compositional data

SpiecEasi: sparse inverse covariance estimation for ecological association inference

sPLS-DA: sparse partial least-squares discriminant analysis

SPn: species identifier

SRA: sequence read archive database

Sub_x0GDx: subgingival plaque of the gingivitis subjects - diseased sites

Sub_x0HHx: subgingival plaque of the periodontally healthy subjects - healthy sites

Sub_x0PDx: subgingival plaque of the periodontitis subjects - diseased sites

Sub_x0PHx: subgingival plaque of the periodontitis subjects - healthy sites

Sub_x1PDx: subgingival plaque of the periodontitis subjects - diseased sites after periodontal therapy

Sup_x0GDx: supragingival plaque of the gingivitis subjects - diseased sites

Sup_x0HHx: supragingival plaque of periodontally healthy subjects - healthy sites

Sup_x0PDx: supragingival plaque of the periodontitis subjects - diseased sites

Sup_x1PDx: supragingival plaque of the periodontitis subjects - diseased sites after periodontal therapy

SVM: support vector machine

SV(s): structural variation(s)

Tbp(s): tera base pair(s)

TGS: third-generation sequencing

Th: T-helper

TIMP(s): tissue inhibitors of matrix metalloproteinase(s)

TINA: taxa interaction-adjusted

TNF alpha: tumour necrosis factor alpha

UI: unidentified

UniFrac: unique fraction metric

US: United States

VAW: variance-adjusted weighted

VC: coverage at the variant level

Vn: variant identifier

WGS: whole-genome shotgun

WHO: World Health Organisation

YLDs: years of healthy life lost due to disability

ZMWs: zero-mode waveguides



In the present Thesis, we describe the primer pairs with the best coverage values for detecting oral bacteria and/or archaea. Nearly all oral bacteria and about half archaea have more than one intragenomic 16S rRNA gene and, depending on the primer used, up to almost half of the species have matching amplicons. The tested primers targeting oral taxa detected an average of >150 potential OTUs that may contain different species when the ≥97% similarity threshold is used, and around 80% of the oral bacteria and archaea had an amplicon similarity ≥97% with at least one other species. Lastly, in a meta-omics analysis comprising 2045 dental plaque samples, we discovered that supragingival plaque is a better bacterial biomarker than subgingival for differentiating periodontal health from untreated and treated periodontitis.