In-Silico Evaluation and Selection of the Best 16S rRNA Gene Primers for Use in Next-Generation Sequencing to Detect Oral Bacteria and Archaea.

Regueira-Iglesias A
Universidade de Santiago de Compostela
Vázquez-González L
Universidade de Santiago de Compostela
Balsa-Castro C
Universidade de Santiago de Compostela
Vila-Blanco N
Universidade de Santiago de Compostela
Blanco-Pintos T
Universidade de Santiago de Compostela
Tamames J
Centro Nacional de Biotecnología (CNB)-CSIC
Carreira MJ
Universidade de Santiago de Compostela
Tomás I (inmaculada.tomas@usc.es)
Universidade de Santiago de Compostela https://orcid.org/0000-0002-3317-0853

Research

Keywords: 16S rRNA gene, primer, coverage, mouth, bacteria, archaea, database.

Posted Date: May 24th, 2021

DOI: https://doi.org/10.21203/rs.3.rs-516961/v1

TITLE: *In-silico* evaluation and selection of the best 16S rRNA gene primers for use in next-generation sequencing to detect oral bacteria and archaea.

AUTHORS: 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].

1-Oral Sciences Research Group, 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)-CSIC; Madrid, Spain.

Regueira-Iglesias A: albaregueira.iglesias@usc.es

Vázquez-González L: laramvazquez@usc.es

Balsa-Castro C: cbalsa@coitt.es

Vila-Blanco N: nicolas.vila@usc.es

Blanco-Pintos T: triana.blanco.pintos@usc.es

Tamames J: jtamames@cnb.csic.es

Carreira-Nouche MJ: mariajose.carreira@usc.es

Tomás I: inmaculada.tomas@usc.es

CORRESPONDENCE

Inmaculada Tomás

School of Medicine and Dentistry. Universidade de Santiago de Compostela

C/ Entrerrios s/n. 15872 Santiago de Compostela, Spain

Tel: +34 981 563100 ext: 12377

Email: inmaculada.tomas@usc.es

Maria José Carreira

Centro Singular de Investigación en Tecnoloxías Intelixentes. Universidade de Santiago de Compostela

Rúa de Jenaro de la Fuente, s/n, 15705 Santiago de Compostela, Spain

Tel: +34 981 563100 16431

Email: mariajose.carreira@usc.es

ABSTRACT

Background: Sequencing has been widely used to study the composition of the oral microbiome present in various health conditions. The extent of the coverage of the 16S rRNA gene primers employed for this purpose has not, however, been evaluated *in silico* using oral-specific databases. This paper analyses these primers using two databases containing 16S rRNA sequences from bacteria and archaea found in the human mouth and describes some of the best primers for each domain.

Results: A total of 369 distinct individual primers were identified from sequencing studies of the oral microbiome and other ecosystems. These were evaluated against a database reported in the literature of 16S rRNA sequences obtained from oral bacteria, which was modified by our group, and a self-created oral-archaea database. 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 the forward and reverse primers were identified, with the resulting 4638 primer pairs also evaluated using the two databases. The best bacteria-specific pairs targeted the 3-4, 4-7 and 3-7 16S rRNA gene regions, with SC levels of 97.14-98.83%; meanwhile, the optimum archaea-specific primer pairs amplified regions 5-6, 3-5 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 94.54-95.71% and 96.91-99.48% for bacteria and archaea, respectively.

Conclusions: Given the three amplicon length categories (100-300, 301-600 and >600 bps), 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 jointly 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.

Keywords: 16S rRNA gene, primer, coverage, mouth, bacteria, archaea, database.

INTRODUCTION

The advent of high-throughput next-generation sequencing (NGS) technologies has enabled the characterisation of microbiomes to unprecedented depths that are unachievable with previous methods (1). These revolutionary techniques enable large-scale projects to be completed in just a few days, or sometimes even hours (2). The NGS employed most at present- Illumina and Ion Torrent - are able to generate sequences with up to 2x300 or 200 base pairs (bps), respectively (3). In the oral-microbiology field, NGS of the 16S ribosomal RNA (rRNA) gene amplicons has been widely used to study the mouth microbiota (4,5), allowing the detection of several bacterial and archaeal taxa in both the healthy human mouth and ones with various states of disease (6). 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 bp) or even the full-length sequence of 16S rRNA gene (V1-V9 regions) through a high throughput process (3).

Nevertheless, sequencing technologies are not without shortcomings, with different challenges and pitfalls possible during each step of the gene-sequencing workflow (7). The primer chosen for the polymerase chain reaction (PCR) amplification step can greatly affect the diversity of an investigation's findings (7,8). 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 (8). Although these primers are based on a consensus sequence, some taxa can produce mismatches (7). 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 (9). As a consequence, using an inadequate primer can lead to questionable biological conclusions (9).

If PCR results in microbial research are to be interpreted satisfactorily, conducting a comprehensive evaluation of a primer's coverage is essential (10). The concept of coverage has been defined heterogeneously as: the percentage of matches for certain taxonomic ranks (11,12); the number of sequences matched by at least one primer (13); 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 (14). The literature contains *in-silico* research that analyses the coverage of 16S rRNA gene-targeting primers that are suitable for amplicon sequencing (10-20). These studies aim to identify the optimum primer pair(s) for sequencing the environmental (11,12,15,16), human (13,14,17-19), or combined environmental and human microbiomes (10,20). For a few of them, the human mouth was an ecosystem of interest (14,17,18), and researchers employed non-oral-specific databases as Silva (21) or a foregut dataset together with the Ribosomal Database Project (RDP) database (22) for the coverage analysis. The application of phylogenetically diverse databases can, however, produce classification errors since they contain taxonomically misannotated 16S rRNA gene sequences (23). They also provide different levels of representation for each included environment, leading to substantial variations in the quality of the classifications (24).

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.

MATERIALS AND METHODS

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) (25) and the RISmed package (version 2.1.7) (26). 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 additional file 1.

Text-mining techniques were applied to all the downloaded abstracts using the R package "tm" (version 0.7-7) (27). 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 microbiota received a score if their abstracts included terms associated with the oral cavity, and another if they contained terms related to a 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 involved

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 Additional files 2 and 3.

Figure 1. Flowchart on the computational search of articles in PubMed and their analysis using text-mining techniques.

Fig.1

Purpose 1. To find 16s rRNA gene primers used to identify bacteria or archaea

a

STEP	DESCRIPTION	BACTERIA	ARCHAEA
1	No. of computational searches in PubMed performed:	2940	5796
2	No. of abstract 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

b

STEP	DESCRIPTION	ARCHAEA
1	No. of computational searches in PubMed performed:	276
2	No. of abstract 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

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 microbiota 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).

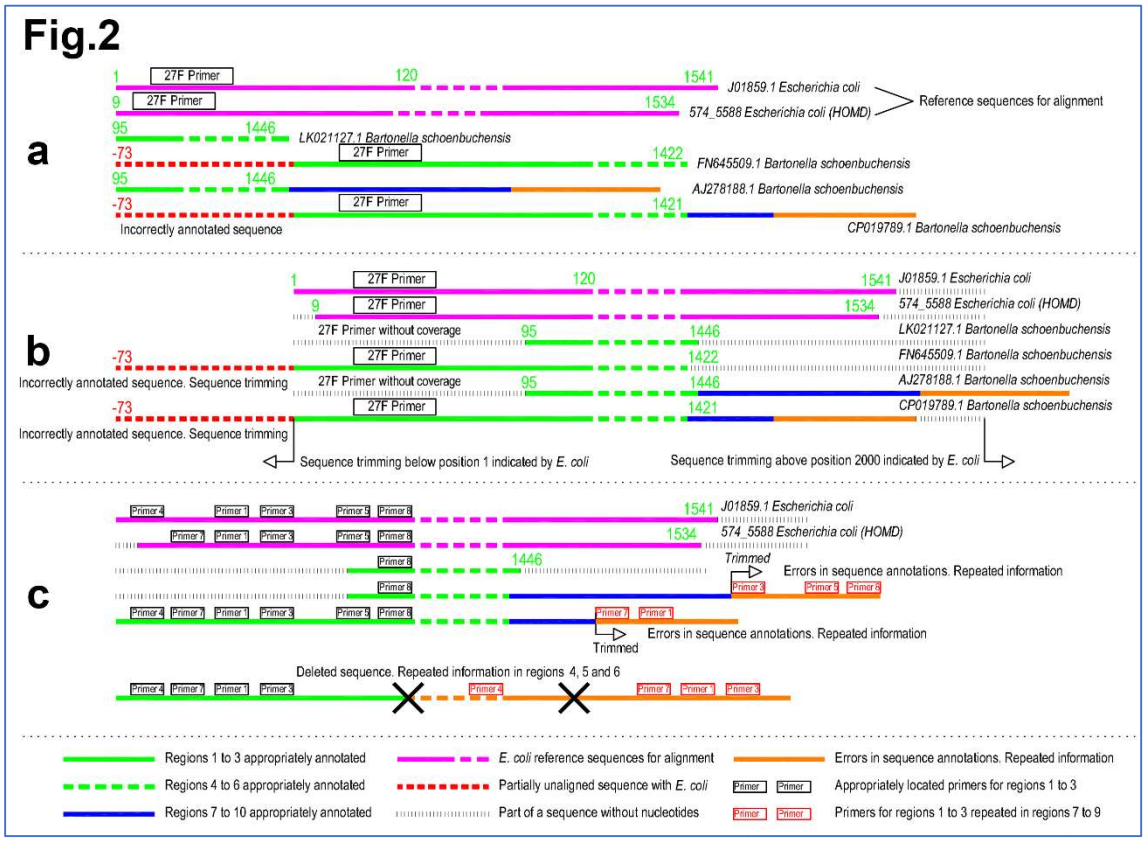
Primer selection and creating a list of archaeal species found in the oral cavity

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

196 remaining 166 were extracted from articles concerning different niches, mainly
197 ecological, described in Klindworth et al. (11). Of them, 103 corresponded to the bacteria
198 domain, 42 to the archaea domain, and 21 were universal. All 444 primers were assigned
199 a unique identifier based on where they were sourced -"OP" for oral primer and "KP" for
200 Klindworth primer (11)- and their direction (F, R or UI), followed by a three-digit number
201 (Additional file 4). The 5'-3' sequences of all 444 primers were then compared to identify
202 repeats, with 75 identified as having the same sequences (Additional file 4). This left us
203 with 369 16S rRNA gene primers with different sequences.
204 The publications in our final selection were read by two researchers to produce a list of
205 archaeal species found in the human mouth. This gave us 177 different archaea names at
206 the species level (Additional file 5).
207 **16S rRNA gene-sequence databases of oral bacteria and archaea for the primer-**
208 **coverage analysis**
209 ***Modification of an existing 16S rRNA gene-sequence database of oral bacteria***
210 A total of 223,143 amplicon sequence variants (ASVs) of fasta-formatted 16S rRNA gene
211 sequences were included in the Escapa et al. database (28). The file had been constructed
212 using sequences from the extended Human Oral Microbiome Database (eHOMD) (29) to
213 then conduct a BlastN search (30,31) of the National Centre for Biotechnology
214 Information (NCBI) non-redundant nucleotide database (32). The header line of each
215 sequence had an ASV identifier (from TS000001 to TS223143), followed by a RefSeq
216 (33) or GenBank (34) identifier and an assignment to a seven-level taxonomic hierarchy.
217 The format was as indicated on the DADA2 website (35). The sequences in the Escapa et
218 al. database (28) were obtained mainly from GenBank, and we found that they contain
219 annotation errors that make it impossible to calculate the correct position of the primers
220 within each sequence in the case of a match.

We developed scripts in Python 3.9 (36) and Bash (37) to improve the Escapa et al. database (28). 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 (38) against a set of 16S rRNA gene sequences of *Escherichia coli*: three from GenBank and one from eHOMD (39). We installed clustal Omega in the local mode (40) to enable its use with Biopython (41). The default characteristics were employed to carry out the alignments. Fourth, all the gaps created by clustal Omega 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 Additional file 6.

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 presence of redundant information. b) Alignment of sequences with respect to *E. coli* and trimming of sequences below position 1 and above position 2000 indicated by *E. 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.

Creation of a 16S rRNA gene-sequence database of archaea

We searched the NCBI nucleotide database for the complete genomes of the archaeal species found in the human mouth. Along with a script developed in Python, these identifiers enabled us to download 193 genomes from RefSeq and eight from GenBank. The script was completed using a free downloadable module, search_16S.py (42), which is based on the algorithm created by Edgar (43). 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 (44). The module and integrating the "The Entrez Programming Utilities (E-utilities)" tool (45) into Biopython meant we could easily and automatically obtain and assign the complete taxonomic rank to the 16S rRNA genes. Biopython also enabled us to access the information of interest requested from the different NCBI databases, such as Taxonomy (46), RefSeq and Genbank.

Additionally, the 16S rRNA gene sequences of species without complete genome identifiers in RefSeq or GenBank were searched for in the aforementioned RefSeq archaeal database or, if not found, the Silva database (version 138) or the Genome Taxonomy Database (GTDB) (47). Finally, all the 16S rRNA gene sequences of the oral archaeal species were grouped into a single fasta file (Additional file 7).

These sequences were employed in BlastN against the NCBI non-redundant nucleotide database. 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. This contains 2842 sequences and all the ASVs presenting with a unique identifier with values between AS00001 and AS002842 (Additional file 8). 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 Additional file 9.

Coverage ratios of the 16S rRNA gene primers

Concept and definition of the coverage ratios calculated for the 16S rRNA gene primers

Two types of coverage were defined for the *in-silico* analysis. First, the coverage at the variant level (variant coverage=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 (48). The results were then stored in the Excel format with xlsxwriter (49), which is a Python package that allows the creation and formatting of xlsx files.

Selection of primer pairs and analysis of their coverage

All the information related to the coverage analysis of individual primers is included in the additional file 10.

The individual primers with a 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) 100 to 300 nucleotides; 2) 301 to 600; and 3) more than 600.

Primer pairs obtained were evaluated against both the bacteria and archaea databases. This step enabled us to determine whether a primer pair was bacteria-specific, archaea-specific, or suitable for both domains.

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 Additional files 11-13. 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 suitable for both domains, and the rest (i.e., 3837 and 489) were obtained exclusively when searching for bacterial- or archaeal-specific primer pairs.

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 the additional file 10.

Bacterial-specific primer pairs

The pair's analysis revealed that 3218 of the 3837 bacterial-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 first 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= 182 - 189). 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 second 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 bacterial-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 bacterial-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 the additional file 14.

Table 1. Selected primer pairs for detecting oral bacteria in different amplicon-length categories.

		Bacteria					Archaea				
LC (bp)	Primer pair	Gene region	SC (%)	Covered	Not covered	Mean length	Gene region	SC (%)	Covered	Not covered	Mean length
100-300	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
301-600	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
>600	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

bp= base pair; LC= length category; SC= coverage at the species level.
SC 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 *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. (50). The reference sequence utilised had 1542 bps distributed in 10 conserved (C1-C10) and nine hypervariable regions (V1-V9). The sequences of these selected primer pairs are described in additional file 15.

Additional file 16 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%.

Archaea-specific primer pairs

Of the 489 primer pairs that were specifically archaea-domain candidates, 359 simultaneously had a bacterial VC and a 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 bp 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 archaeal-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 second 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 bp category had a 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 on the basis of both their coverage and mean amplicon lengths. The VC results of these selected primers are detailed in the additional file 14.

Table 2. Selected primer pairs for detecting oral archaea in different amplicon-length categories.

		Bacteria					Archaea				
LC (bp)	Primer pair	Gene region	SC (%)	Covered	Not covered	Mean length	Gene region	SC (%)	Covered	Not covered	Mean length
100-300	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
301-600	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
>600	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

bp= base pair; LC= length category; SC= coverage at the species level.

SC 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 *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. (50). The reference sequence utilised had 1542 bps distributed in 10 conserved (C1-C10) and nine hypervariable regions (V1-V9). The sequences of these selected primer pairs are described in additional file 15.

Additional file 17 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%.

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 100-300 bp 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 bp 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); 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 bp 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 and 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 the additional file 14.

Table 3. Selected primer pairs for simultaneously detecting oral bacteria and archaea in different amplicon-length categories.

		Bacteria					Archaea				
LC (bp)	Primer pair	Gene region	SC (%)	Covered	Not covered	Mean length	Gene region	SC (%)	Covered	Not covered	Mean length
100-300	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
301-600	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
>600	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

Legend Table 3. bp= base pair; LC= length category; SC= coverage at the species level.
SC 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 *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. (50). The reference sequence utilised had 1542 bps distributed in 10 conserved (C1-C10) and nine hypervariable regions (V1-V9). The sequences of these selected primer pairs are described in additional file 15.

Finally, additional file 18 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.

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 microbiota inhabiting the human mouth, but also from an article containing primers used

515 in ecosystems as dissimilar as the marine, geothermal, human gut, or cattle gut (11). Thus,
516 numerous primers from diverse ecosystems were analysed to find those which performed
517 better in the oral cavity. Moreover, to perform the analysis, we improved an earlier
518 database of 16S rRNA gene sequences of oral bacteria (28) and created another from
519 scratch that contained sequences from archaeal species found in the oral cavity.
520 We identified a series of individual primers that performed well in the detection of oral
521 bacteria and/or archaea and combined them to create primer pairs. These were defined as
522 "bacteria-specific", "archaea-specific", or "bacterial and archaeal" based on the results
523 on their levels of coverage set out in the two databases. We also produced a series of
524 primer pairs that may be the most suitable combinations for use when sequencing the oral
525 ecosystem. These were classified according to the domain targeted, their mean amplicon-
526 length category, and the 16S rRNA gene region amplified.
527 **Comparative analysis of our coverage results of 16S rRNA gene primers with the**
528 **literature**
529 The investigation by Klindworth et al. (11) is perhaps the most comprehensive to date on
530 the coverage and phylum spectrum of 16S rRNA primers. These authors assessed 175
531 primers and 512 primer pairs *in silico* against the Silva non-redundant reference database
532 (version 108), producing a selection of those that performed best for bacteria and archaea.
533 Like us, this group organised the most suitable primer combinations for the different
534 sequencing technologies into three categories according to their amplicon length (100–
535 400, 400–1000, >1000 bp). They then re-evaluated their analysis using the Global Ocean
536 Sampling (GOS) dataset (51,52), which is limited to the marine habitat and examined
537 experimentally the primer pair that performed best (11).
538 We identified two investigations involving the oral ecosystem that used the Silva database
539 -versions 111 (14) or 132 (17)- to analyse the efficiency of 16S rRNA gene primers for

detecting the archaea diversity in oral samples (17), or for reconstructing the microbiome of ancient dental calculus specimens (14). 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 (18). The pair with the best results for classifying the foregut genes was also analysed against the RDP database. The numbers of 16S rRNA gene primer pairs evaluated in these three oral-related studies are substantially lower than in the present investigation: respectively, 12 individual primers combined to form 12 primer pairs (17); 25 individual primers combined into 14 pairs (14); and 14 individual primers grouped into 14 pairs (18). In our study, we analysed 369 distinct individual primers and 4482 different primer-pair combinations. On the other hand, two investigations used the Silva database (14,17) which has broad phylogenetic diversity and contains information applicable to many environments but also includes 16S rRNA gene sequences that are misannotated taxonomically (28). Specifically, comprehensive databases such as Silva, RDP or Greengenes have been estimated to have annotation error rates ranging from 10-17% (23), and their accuracy may also be reduced because they contain numerous sequences derived from some environments and only a few from others (24). 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. (18) 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, and to which we added sequences from our self-created archaeal dataset (bacteria: 223143 sequences, 769 species; archaea: 2842 sequences, 194 species).

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 (100-300 bps), KP_F049-KP_R033 (301-600 bps), KP_F056-KP_R074 (301-600 bps), KP_F033-KP_R060 (>600 bps) and KP_F047-KP_R053 (>600 bps) were similar to those of other studies, with differences no greater than 5.00% for both the bacteria and archaea domains (11). It should be noted that the latter, classified here as having a mean amplicon length > 600 bp but put in the medium-length category by Klindworth, had a lower bacterial coverage when analysed against the GOS database. This was also the case for KP_F056-KP_R074 (11). Moreover, the coverage values of the pairs KP_F077-KP_R071 (100-300 bp) and KP_F047-KP_R035 (301-600 bp) in other studies are similar to those in our research, but the archaeal coverage is notably higher -~31.00% (14) and ~83.00% (16,20) 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 (20). We, therefore, believe that KP_F047-KP_R035 has value for detecting archaea in environmental (16,20) or human gut (20) specimens, but not in samples from the oral cavity.

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 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).

589 *Archaea-specific primer pairs*

590 Two of the primer pairs selected herein to detect oral archaea species - KP_F018-
591 KP_R002 (100-300 bp) and KP_F020-KP_R013 (301-600 bp) - have previously been
592 described in other studies as the best options for targeting this domain (11). Nonetheless,
593 the archaeal SC values we obtained exceed Klindworth's by ~ 20.00% (Table 4).
594 Klindworth's group (11) also found that KP_F018-KP_R078 had the highest overall
595 archaeal coverage for the amplicons with a long mean length. However, they do not
596 recommend its use due to its low phylum spectrum. This combination produced bacterial
597 SC values of 0.00% and an archaeal SC of 93.30% when analysed against our database.
598 Although this is a good result, we prefer OP_F114-KP_R013, which achieves better
599 archaeal SC, or KP_F018-KP_R063, which produces both better archaeal SC and a
600 greater mean amplicon length (Table 4).

601 *Bacteria and archaea primer pairs*

602 KP_F020-KP_R032 and KP_F020-KP_R035 (100-300 bp), with bacterial and archaeal
603 coverage estimates >85.00% (mainly considering the Silva database), have been proposed
604 previously as suitable for the detection of both domains (11). As seen in table 4, the SC
605 values obtained herein, ≥95.00%, are better than those in other studies (11,17). OP_F066-
606 OP_F073 is also among the favoured primers in our study for the detection of bacteria
607 and archaea when using short amplicon lengths (SC= 98.31% and 92.78%, respectively),
608 achieving better coverage than in the research by Zhang et al. (SC= 88.90% and 75.30%,
609 respectively) (12). Meanwhile, although other studies' *in-silico* analyses of OP_F014-
610 OP_R014 have described it as a good primer pair for detecting the two domains (17,20),
611 it is not among our recommended primers, since others in the same length and gene-
612 region categories achieved better archaeal coverage. KP_F059-KP_R078 has been
613 proposed by Klindworth et al. (11) as suitable for use with both the bacteria and archaea

614 domains when employing medium mean amplicon lengths (608 bps). However, its length
615 was 622 bps in our study, meaning that it was included in the >600 bps group. In any
616 case, although our coverage values were higher than those obtained previously (SC=
617 93.63% and 96.39% *vs* 74.10 and 72.30%, respectively) (Table 4), other primer pairs
618 performed better in both categories as OP_F114-KP_R031 (301-600; bacterial SC=
619 95.71% and archaeal SC= 98.45%) and OP_F066-OP_R121 (>600; bacterial SC=
620 94.54% and archaeal SC= 96.91%).

621 **Non-covered species by the 16S rRNA gene primer pairs**

622 The *in-silico* analysis has enabled us to verify that, among the pairs achieving better
623 coverage, the species not covered by the primers targeting a particular region tend to be
624 covered by others relating to a different zone. In this sense, most of the species that were
625 not covered by the bacterial-specific primer pairs from regions 3-4 (100-300 bps), 3-5
626 (301-600 bps), or 3-7 (>600 bps) were by those from 5-7 and 6-7 (100-300 bps), 4-7 and
627 7-9 (301-600 bps), or 4-9 (>600 bps), and vice versa (Additional file 16). This was also
628 seen in the archaeal-specific primers, where taxa not detected by the pairs from regions 3
629 (100-300 bps), 3-5 (301-600 bps) or 3-6 (>600 bps) were covered by those from 5-6 (100-
630 300 bps), 3-6 and 5-9 (301-600 bps), or 3-9 and 5-9 (>600 bps), and vice versa (Additional
631 file 17). Lastly, the pairs for the two domains combined also demonstrated that species
632 not covered by primers from zones 4-5 (100-300 bps) or 3-5 (301-600 bps) were by those
633 from 5-6 (100-300 bps) or 4-6 (301-600 bps), and vice versa (Additional file 18). In the
634 combinations with mean amplicon lengths >600 bps, half the taxa that were not covered
635 using primers for amplifying region 3-9 were detected when targeting 5-9. However, in
636 this case, the opposite was not true.

637 There were exceptions to this general rule, which demonstrated that even for two primer
638 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 bacterial-specific pair OP_F101-OP_R030 covered 33 of the 66 species not detected by KP_F061-KP_R074 (6-7; 100-300 bps). Furthermore, several primers from region 4-7 (301-600 bps) covered more than half of the bacterial species not detected by KP_F051-KP_R053 and OP_F021-KP_R053 (Additional file 16). Meanwhile, the archaeal-specific primer KP_F016-KP_R002 covered almost all of the taxa not detected using KP_F018-KP_R002, KP_F018-KP_R003 and KP_F018-OP_R010 (3; 100-300 bps). In addition, KP_F016-KP_R032 was the only primer from region 3-5 (301-600 bps) able to identify six archaea that were not detected by the rest of the primers from the same region (Additional file 17).

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 the bacterial-specific or bacteria and archaea 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 (53). A further three bacteria commonly found in the healthy periodontium, *Leptotrichia bucalis* (54), *Leptotrichia hofstadii* (54) and *Rothia dentocariosa* (55,56), were also missed by some of the 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* (57), *Actinomyces israelii* (57), *Desulfomicrobium orale* (58), *Mogibacterium timidum* (59), *Solobacterium moorei* (60,61), *Treponema lecithinolyticum* (55,59,61) and *Treponema maltophilum* (62). A further *Actinomyces*

species, previously classified as *naeslundii* WVA 963 and now known as *johnsonii* (63), which has been encountered in both healthy and periodontitis sites (57), 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 (64), were also not covered. Meanwhile, the caries-associated bacterial species that were not detected by some of the primer pairs included *Bifidobacterium dentium* (65,66), *Lactobacillus reuteri* (67), *Leptotrichia buccalis* (68), *Parascardovia denticolens* (67,69), *Rothia dentocariosa* (70) and *Scardovia wiggsiae* (71) (Additional files 16 and 18).

The undetected archaeal species included *Methanobrevibacter gottschalkii*, *Methanopyrus kandleri*, *Nitrosoarchaeum limnia* and *Nitrososphaera evergladensis* (Additional files 17 and 18), which have been found, in order, in inflamed pulp tissue (72), periodontitis samples (73), endodontic infections (74) and ancient dental calculus (75,76). The rest of the non-detected archaea were extracted from the same publication (77) 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 study the oral microbiota.

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 microbiota 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 (Additional file 19).

Only 67 of these 206 pairs were evaluated in the present study. This means that at least one of 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 (Additional file 19), was developed by Caporaso et al. (78) 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 (Additional file 20). Numerous primer combinations in the same length category (100-300 bps) 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= 284-287 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 (Additional file 19) and is the pair proposed in the Illumina protocol for the preparation of the sequencing library (79). 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* (Additional file 20). Although it has been used extensively in oral microbiome studies, in our investigation other primers in the same length category (301-600 bps) 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 (Additional file 16).

Another widely used primer is 785F–1175R, which has been employed to amplify gene region 5-7 (Additional file 19). 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 (301-600 bps), was not among the best primers in our investigation (bacterial SC range= 89.60% - 79.97%; archaeal SC= 0.00%; mean amplicon length range= 408–411 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* (80-82), *Porphyromonas gingivalis* (53,80,82,83) and *Tannerella forsythia* (53,82,83) (Additional file 20). 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.

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 microbiome. 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 (84), in general terms, the longer the fragment sequenced, the lower the taxonomic level that can be achieved (12). Indeed, sequencing full-lengths, as is possible with PacBio, is regarded as the solution to the limitations of taxonomic classification (12). Nevertheless, Soergel et al. (24) 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 (12). Consequently, we agree that the choice of the target region is also an important factor (12,24).

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, 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 (*Candidatus Korarchaeum cryptofilum*, *Methanobrevibacter gottschalkii strain HO*, *Methanobrevibacter oralis*,

Methanobrevibacter thaueri strain CW and *Nitrosoarchaeum 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. The results of our *in-silico* analysis should also be confirmed in further experimental studies using omics techniques.

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 *Escherichia 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 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.

LIST OF ABBREVIATIONS (alphabetically ordered)

ASV(s): amplicon sequence variant(s)

bp(s): base pair(s)

C: conserved

eHOMD: extended human oral microbiome database

F: forward

GOS: global ocean sampling

GTDB: genome taxonomy database

JSON: javaScript object notation

KP: Klindworth primer

NCBI: National Centre for Biotechnology Information

NGS: next-generation sequencing

OP: oral primer

PacBio: Pacific Biosciences

PCR: polymerase chain reaction

PMID: PubMed unique identifier

R: reverse

RDP: ribosomal database project

rRNA: ribosomal RNA

SC: coverage at the species level

UI: unidentified

V: variable

VC: coverage at the variant level

DECLARATIONS

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Availability of data and material

Principal data generated or analysed during this study are included in this published article.

Competing interests

The authors have no competing interests to declare.

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) 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.

Authors' contributions

C. Balsa-Castro and I. Tomás contributed to the conception and design of the study; A. Regueira-Iglesias and T. Blanco-Pintos searched the articles in PubMed and selected those of interest, extracting the relevant information; L. Vázquez-González, N. Vila-

Blanco and M.J. Carreira developed the bioinformatics procedures for obtaining the analysis of the primer pairs; A. Regueira-Iglesias and C. Balsa-Castro made the graphs, tables and additional files; A. Regueira-Iglesias and I. Tomás wrote the manuscript; J. Tamames and M.J. Carreira carried out a critical review of the manuscript. All the authors approved the final version of the manuscript.

Acknowledgements

Not applicable.

FIGURES

Figure 1. Flowchart on the computational search of articles in PubMed and their analysis using text-mining techniques.

Legend Figure 1. 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 microbiota 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).

Figure 2. Processing of errors in annotations of oral bacterial and archaeal sequences.

Legend Figure 2. a) Unaligned sequences with missing information at the first and last positions of the 16S rRNA gene; and presence of redundant information. b) Alignment of sequences with respect to *E. coli* and trimming of sequences below position 1 and above position 2000 indicated by *E. 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.

869 **TABLES**

870 Table 4. Coverage findings described in the literature for the gene primer pairs analysed

871 in the present study.

Present study		Other studies	Results of the present study		Results of the other studies		Ref.
Primer pair	Mean length	Primer pair name	Bacterial SC (%)	Archaeal SC (%)	Bacterial coverage (%)	Archaeal coverage (%)	
KP_F044-KP_R023	100-300 bp	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	100-300 bp	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	100-300 bp	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	100-300 bp	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	100-300 bp	S-D-Bact-0341-a-S-17/ N/A	87.52	0.00	81.50	0.00	(12)
KP_F047-KP_R021	100-300 bp	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	100-300 bp	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	100-300 bp	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	100-300 bp	U341F – 534R	95.58	59.79	98.00	~ 91.00	(14)
KP_F078-OP_R010	100-300 bp	515F – 806 R (original)	95.32	62.89	86.80	52.90	(17)
		S-*-Univ-0515-a-S-19/ N/A			86.10	52.00	(12)
KP_F078-KP_R037	100-300 bp	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	100-300 bp	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	100-300 bp	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	100-300 bp	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	100-300 bp	515F – 806 R (modified)	95.32	88.14	87.70	85.70	(17)
		515F – 806R			96.20	96.39	(20)
OP_F066-OP_R073	100-300 bp	N/A/ N/A	98.31	92.78	88.90	75.30	(12)
KP_F044-KP_R032	301-600 bp	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	301-600 bp	S-D-Bact-0341-a-S-17/ N/A	93.89	0.00	83.30	0.10	(12)
KP_F047-KP_R035	301-600 bp	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	301-600 bp	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	301-600 bp	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	301-600 bp	N/A / N/A	91.68	1.03	86.50	0.50	(12)
KP_F020-KP_R013	301-600 bp	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	>600 bp	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	>600 bp	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	>600 bp	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	>600 bp	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	>600 bp	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	>600 bp	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	>600 bp	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. SC= coverage at the species level; Ref= references; a= Silva database; b= GOS database.

SUPPLEMENTARY INFORMATION

Additional file 1 (Additional_file_1.doc). List of words employed in the automated searches to identify the 16S rRNA gene primers used for detecting oral bacteria and oral archaea before sequencing.

Additional file 2 (Additional_file_2.doc). List of references from which a particular primer was initially obtained.

Additional file 3 (Additional_file_3.doc). List of references from which the archaeal species inhabiting different human-mouth niches were obtained.

Additional file 4 (Additional_file_4.xlsx). Forward and reverse 16S rRNA gene primers evaluated in the study and the sequence comparison used to detect repeats.

Additional file 5 (Additional_file_5.xlsx). List of archaeal species present in the human mouth and the PMID of the investigations from which they were obtained.

Additional file 6 (Additional_file_6.fa). Oral bacteria database of the 16S rRNA gene sequences used in the present study for the coverage analysis.

Additional file 7 (Additional_file_7.fa). 16S rRNA gene sequences from the oral archaea employed for the BlastN search against the NCBI non-redundant nucleotide database.

Additional file 8 (Additional_file_8.fa). Oral archaea database of the 16S rRNA gene sequences constructed by our group before alignment.

Additional file 9 (Additional_file_9.fa). Oral archaea database of the 16S rRNA gene sequences used in the present study for the coverage analysis.

Additional file 10 (Additional_file_10.doc). Information related to the coverage analysis of the 16S rRNA gene individual primers.

Additional file 11 (Additional_file_11.xlsx). Evaluation of individual primers against the oral bacteria database.

Additional file 12 (Additional_file_12.xlsx). Evaluation of individual primers against the oral archaea database.

Additional file 13 (Additional_file_13.xlsx). Evaluation of primer pairs against the oral bacteria and the oral archaea databases.

Additional file 14 (Additional_file_14.doc). Coverage at the variant level of the selected primer pairs for detecting oral bacteria or/and archaea.

Additional file 15 (Additional_file_15.xlsx). Sequences of the selected primer pairs for detecting oral bacteria or/and archaea.

Additional file 16 (Additional_file_16.xlsx). Bacterial species non covered by the selected bacterial-specific primer pairs.

Additional file 17 (Additional_file_17.xlsx). Archaeal species non covered by the selected archaeal-specific primer pairs.

Additional file 18 (Additional_file_18.xlsx). Species non covered by the selected primer pairs for detecting both bacteria and archaea.

Additional file 19 (Additional_file_19.xlsx). List of the primer pairs utilised in the reviewed oral microbiome studies through 16S rRNA gene sequencing.

Additional file 20 (Additional_file_20.xlsx). Oral bacterial species non covered by KP_F078-OP_R010, KP_F047-KP_R035 and OP_F009-OP_R029; and the oral archaea non covered by KP_F078-OP_R010.

REFERENCES

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

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

(3) Santos A, van Aerle R, Barrientos L, Martinez-Urtaza J. Computational methods for 16S metabarcoding studies using Nanopore sequencing data. Comput Struct Biotechnol J. 2020;18:296-305.

(4) Willis JR, Gabaldón T. The human oral microbiome in health and disease: from sequences to ecosystems. Microorganisms. 2020;8:308.

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

(6) Wade WG. The oral microbiome in health and disease. Pharmacol Res. 2013;69:137-43.

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

(8) de la Cuesta-Zuluaga J, Escobar JS. Considerations for optimizing microbiome analysis using a marker gene. Front Nutr. 2016;3:26.

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

(10) Mao DP, Zhou Q, Chen CY, Quan ZX. Coverage evaluation of universal bacterial primers using the metagenomic datasets. BMC Microbiol. 2012;12:66.

(11) 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;41:e1.

(12) 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;618:1254-67.

(13) 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;19:343-6.

(14) 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;5:16498.

(15) 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;11:487-94.

(16) 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;8:494.

(17) 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;10:2796.

(18) 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;16:4135-44.

(19) 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;24:812-22.

(20) 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;20:1558-71.

(21) 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;41:D590-6.

(22) 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. 2013;42:D633-42.

(23) Edgar R. Taxonomy annotation and guide tree errors in 16S rRNA databases. PeerJ. 2018;6:e5030.

(24) Soergel DA, Dey N, Knight R, Brenner SE. Selection of primers for optimal taxonomic classification of environmental 16S rRNA gene sequences. ISME J. 2012;6:1440-4.

(25) R Core Team. R: a language and environment for statistical computing. R package version 4.0.3. 2020. http://www.R-project.org/.

(26) Kovalchik S. RISmed: download content from NCBI databases. R package version 2.1.7. 2017. http://www.CRAN.R-project.org/.

(27) Feinerer I, Hornik K, Meyer D. Text mining infrastructure in R. J Stat Softw. 2008;25:1-54.

990 (28) Escapa IF, Huang Y, Chen T, Lin M, Kokaras A, Dewhirst FE, et al. Construction
991 of habitat-specific training sets to achieve species-level assignment in 16S rRNA gene
992 datasets. Microbiome. 2020;8:65.
993 (29) Escapa IF, Chen T, Huang Y, Gajare P, Dewhirst FE, Lemon KP. New insights into
994 human nostril microbiome from the expanded human oral microbiome database
995 (eHOMD): a resource for the microbiome of the human aerodigestive tract. mSystems.
996 2018;3:187.
997 (30) Morgulis A, Coulouris G, Raytselis Y, Madden TL, Agarwala R, Schäffer AA.
998 Database indexing for production MegaBLAST searches. Bioinformatics. 2008;24:1757-
999 64.
1000 (31) Chen Y, Ye W, Zhang Y, Xu Y. High speed BLASTN: an accelerated MegaBLAST
1001 search tool. Nucleic Acids Res. 2015;43:7762-8.
1002 (32) NCBI Resource Coordinators. Database resources of the National Center for
1003 Biotechnology Information. Nucleic Acids Res. 2016;44:7.
1004 (33) O'Leary NA, Wright MW, Brister JR, Ciufo S, Haddad D, McVeigh R, et al.
1005 Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and
1006 functional annotation. Nucleic Acids Res. 2016;44:733.
1007 (34) Clark K, Karsch-Mizrachi I, Lipman DJ, Ostell J, Sayers EW. GenBank. Nucleic
1008 Acids Res. 2016;44:D67-72.
1009 (35) Callahan BJ, McMurdie PJ, Rosen MJ, Han AW, Johnson AJ, Holmes SP. DADA2:
1010 High-resolution sample inference from Illumina amplicon data. Nat Methods.
1011 2016;13:581-3.
1012 (36) Python Software Foundation. Python. Version 3.9.0. 2020. http://www.python.org/.
1013 (37) GNU P. Free Software Foundation. Bash. Version 5.1. 2020. http://www.gnu.org/.

(38) 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;7:539.

(39) 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;2010:baq013.

(40) Conway Institute UCD Dublin. Clustal Omega installation instructions. 2018. http://www.clustal.org/omega/INSTALL.

(41) 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;25:1422-3.

(42) Lyalina S. Search 16S py algorithm. 2019. https://github.com/slyalina/search_16S_py.

(43) Edgar RC. SEARCH_16S: A new algorithm for identifying 16S ribosomal RNA genes in contigs an d chromosomes. bioRxiv. 2017;124131.

(44) National Center for Biotechnology Information. NCBI RefSeq Targeted Loci Project. Archaea FTP. 2008. ftp://ftp.ncbi.nlm.nih.gov/refseq/TargetedLoci/Archaea/.

(45) National Center for Biotechnology Information. Entrez Programming Utilities Help. 2010. https://www.ncbi.nlm.nih.gov/books/NBK25501/.

(46) 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;2020:baaa062

(47) 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;38:1079-86.

(48) Barnett M. regex. 2020. https://pypi.org/.

(49) McNamara J. xlsxwriter. 2013. https://xlsxwriter.readthedocs.io/.

(50) Baker GC, Smith JJ, Cowan DA. Review and re-analysis of domain-specific 16S primers. J Microbiol Methods. 2003;55:541-55.

(51) 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;5:e16.

(52) 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;5:e77.

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

(54) 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 Mi. 2019;9:347.

(55) 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;12:1814674.

(56) 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;91 Suppl 1:S56-67.

(57) Vielkind P, Jentsch H, Eschrich K, Rodloff AC, Stingu CS. Prevalence of *actinomyces* spp. in patients with chronic periodontitis. Int J Med Microbiol. 2015;305:682-8.

(58) 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;4:6602.

(59) 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;94:1425-31.

(60) 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;10:596490.

(61) Hiranmayi KV, Sirisha K, Ramoji Rao MV, Sudhakar P. Novel pathogens in periodontal microbiology. J Pharm Bioallied Sci. 2017;9:155-63.

(62) 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;46:745-52.

(63) 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;59:509-16.

(64) Bor B, Bedree JK, Shi W, McLean JS, He X. Saccharibacteria (TM7) in the human oral microbiome. J Dent Res. 2019;98:500-9.

(65) Mantzourani M, Fenlon M, Beighton D. Association between bifidobacteriaceae and the clinical severity of root caries lesions. Oral Microbiol Immunol. 2009;24:32-7.

(66) 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;43:308-13.

(67) 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;12:1830623.

(68) 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;doi: 10.1007/s00784-020-03624-7.

(69) 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;9:1270614.

(70) Jiang S, Gao X, Jin L, Lo EC. Salivary microbiome diversity in caries-free and caries-affected children. Int J Mol Sci. 2016;17:1978.

(71) 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;49:1464-74.

(72) Efenberger M, Agier J, Pawłowska E, Brzezińska-Błaszczyk E. Archaea prevalence in inflamed pulp tissues. Cent Eur J Immunol. 2015;40:194-200.

(73) 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;18:373-7.

(74) Keskin C, Demiryürek EÖ, Onuk EE. Pyrosequencing analysis of cryogenically ground samples from primary and secondary/persistent endodontic infections. J Endod. 2017;43:1309-16.

(75) Huynh HT, Verneau J, Levasseur A, Drancourt M, Aboudharam G. Bacteria and archaea paleomicrobiology of the dental calculus: a review. Mol Oral Microbiol. 2016;31:234-42.

(76) 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;6:25775.

(77) 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;7:3703-8.

(78) 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;108 Suppl 1:4516-22.

(79) Illumina Inc. 16S metagenomic sequencing library preparation. 2013. https://support.illumina.com/content/dam/illumina-support/documents/documentation/chemistry_documentation/16s/16s-metagenomic-library-prep-guide-15044223-b.pdf.

(80) 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;11:1653123.

(81) 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;42:1015-23.

(82) 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;5:9053.

(83) 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;12:e0182992.

(84) Dieffenbach CW, Lowe TM, Dveksler GS. General concepts for PCR primer design. PCR Methods Appl. 1993;3:30.

Figures

Fig.1

Purpose 1. To find 16s rRNA gene primers used to identify bacteria or archaea

a

STEP	DESCRIPTION	BACTERIA	ARCHAEA
1	No. of computational searches in PubMed performed:	2940	5796
2	No. of abstract 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

b

STEP	DESCRIPTION	ARCHAEA
1	No. of computational searches in PubMed performed:	276
2	No. of abstract 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. Legend Figure 1. 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 microbiota 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).

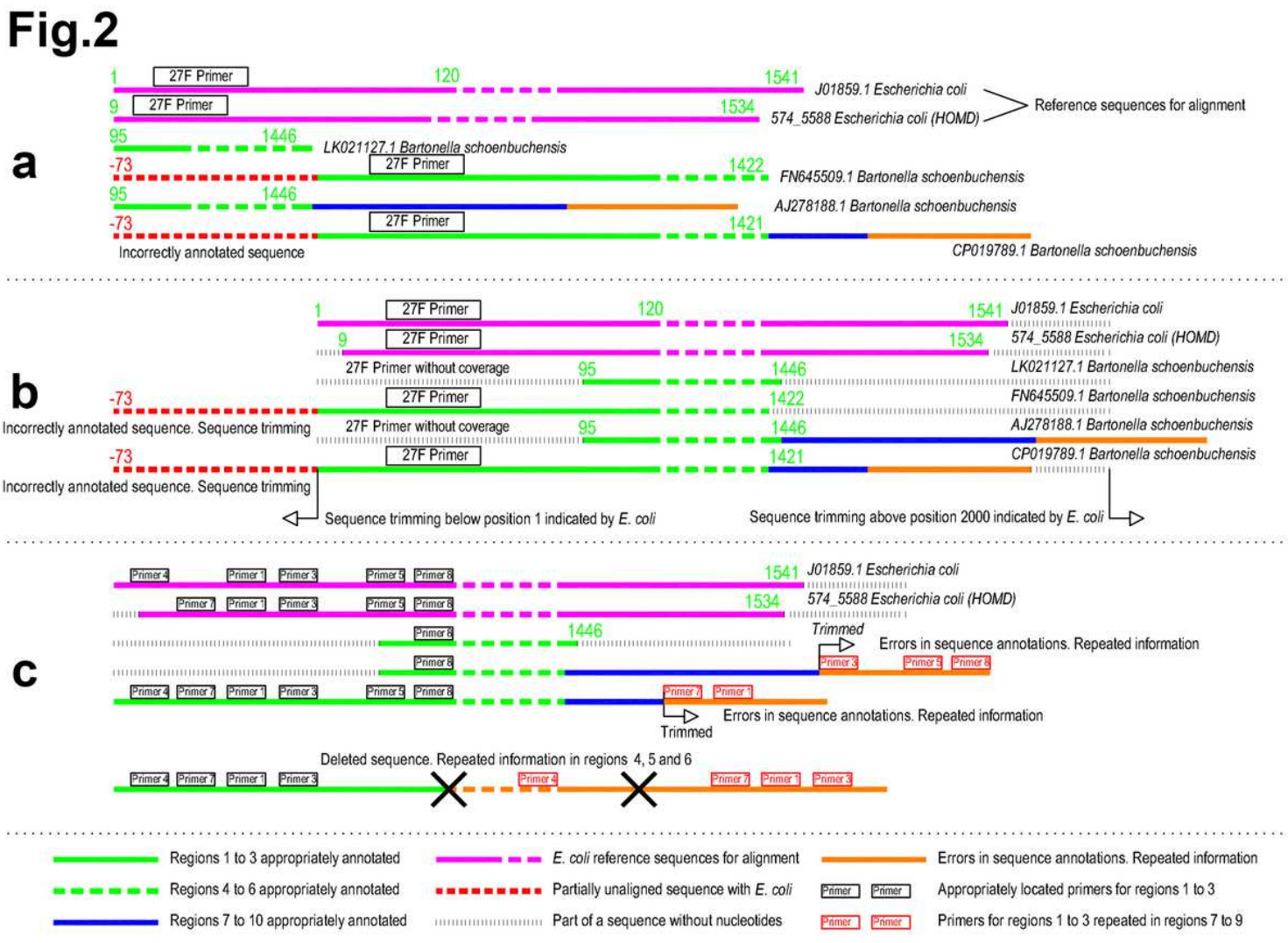


Figure 2

Processing of errors in annotations of oral bacterial and archaeal sequences. Legend Figure 2. a) Unaligned sequences with missing information at the first and last positions of the 16S rRNA gene; and presence of redundant information. b) Alignment of sequences with respect to E. coli and trimming of sequences below position 1 and above position 2000 indicated by E. 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.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- Additionalfile1.docx
- Additionalfile10.docx
- Additionalfile11.xlsx

- Additionalfile12.xlsx
- Additionalfile13.xlsx
- Additionalfile14.docx
- Additionalfile15.xls
- Additionalfile16.xlsx
- Additionalfile17.xlsx
- Additionalfile18.xlsx
- Additionalfile19.xlsx
- Additionalfile2.docx
- Additionalfile20.xlsx
- Additionalfile3.docx
- Additionalfile4.xlsx
- Additionalfile5.xlsx
- Additionalfile6.fa
- Additionalfile7.fa
- Additionalfile8.fa
- Additionalfile9.fa