TITLE: Impact of 16S rRNA gene redundancy and primer pair selection on the quantification and classification of oral microbiota in next-generation sequencing.

AUTHORS: Alba Regueira-Iglesias[1], Lara Vázquez-González L[2], Carlos Balsa-Castro[1], Triana Blanco-Pintos[1], Nicolás Vila-Blanco[2], Maria José Carreira[2], Inmaculada Tomás I[1].

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.

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

Vázquez-González L: laram.vazquez@usc.es

Balsa-Castro C: cbalsa@coitt.es

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

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

Carreira MJ: mariajose.carreira@usc.es

Tomás I: inmaculada.tomas@usc.es

CORRESPONDENCE

Inmaculada Tomás

School of Medicine and Dentistry. Universidade de Santiago de Compostela

15872 Santiago de Compostela, Spain

Tel: +34 981 563100 ext: 12377

Fax: +34 981 562226

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: The identification, at least at the species level, is highly desirable in 16S rRNA sequencing-based studies of the oral microbiota. However, no study in the oral microbiology field has examined the impact of which primer pair is selected to detect redundant and matching amplicons. Consequently, our aims 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 and classification of redundant amplicons and matching amplicons (MA) from different taxa.

Results: A total of 709 complete genomes (518 bacteria, 191 archaea) were downloaded from the NCBI database, and their complete 16S rRNA genes were extracted. 94.1% of oral bacteria and 52.59% of oral archaea had more than one 16S rRNA gene in their respective genomes. Next, 33 primer pairs identified in previous research and 6 commonly used in the literature were used against all the genomes to obtain amplicons. Between 46.67%-1.29% of the bacterial species and between 38.89%-4.65% of the archaeal species had MA, affecting relevant genera present in the oral environment such as *Actinomyces, Fusobacterium, Lactobacillus, Methanosarcina, Staphylococcus,* and *Streptococcus.* The best primer pairs were (the species coverage with no MA values, 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 matching amplicons 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 no 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 affects significantly diversity estimates and taxonomic classification, conditioning the comparability of oral microbiome studies using different primer pairs.

Keywords*:* 16S rRNA gene, gene variant, matching amplicon, oral microbiota, overestimation factor, primer, redundancy, sequence analysis

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 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 microbiota (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, this is only the case for those with high grades of dental pathology in *Fusobacterium nucleatum* subsp. *vicentii* (13). However, the taxonomic resolution at the species level could be affected by the presence of matching amplicons.

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 matching amplicons (MA) from different taxa.

MATERIALS AND METHODS

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 528 complete genomes were identified among 2074 on the eHOMD website.

The complete genomes indicated in the eHOMD, have one or more Genbank identifiers, which were used to access the complete sequences stored in the National Center for Biotechnology Information (NCBI) database (19). 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 identifiers, obtained as part of a previous investigation conducted by our research group (20), enabled us to access their complete sequences in the NCBI database.

Integrating the "Entrez Programming Utilities (E-utilities)" tool (21) in the Python script (22) allowed us to acquire the URLs needed to retrieve the information of interest from the various NCBI databases, including Taxonomy (23), RefSeq (24), and GenBank (25). The oral-bacteria and oral-archaea genomes were then downloaded, and finally, the taxonomy of each of them was obtained.

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 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 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 additional files 1 and 2, 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.

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 previous research by our group, as well as those used most in the oral-microbiome literature (20). 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 primer pairs (100-300 bps), medium primer pairs (301-600 bps), and large primer pairs (>600 bps) (20) (Additional files 3 and 4).

The direct and reverse sequences of each primer pair selected were used in combination with Python's regex module (30) 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 matching amplicons (SC-NMA). This coverage value was calculated as:

SC-NMA (%) = [(Number of species detected - Number of species with MA)/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 MA. 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 MA (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.

RESULTS

Number of intragenomic 16S rRNA genes in oral-bacteria and oral-archaea genomes

Table 1 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. The additional files 5 and 6 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 1. 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**	**Number of 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**	**Number of 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.

Evaluation of the primer pairs taken from our previous research and those used most in oral-microbiome studies

Tables 2 and 3 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 2. Size and number of 16S rRNA gene amplicons detected by the primer pairs in the oral-bacteria genomes.

			Superkingdom level			Species level		
ALC	**Bacterial-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 - 10	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 a previous investigation 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 selected in previous research (20). 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= large amplicon length, >600 base pairs; M= medium amplicon length, 301-600 base pairs; OP= oral primer; R= reverse; S= short amplicon length, 100-300 base pairs.

Table 3. Size and number of 16S rRNA gene amplicons detected by the primer pairs in the oral-archaea genomes.

			Superkingdom level			Species level		
ALC	**Archaeal-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	3	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}	3 – 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 a previous investigation 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 selected in previous research (20). 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= large amplicon length, >600 base pairs; M= medium amplicon length, 301-600 base pairs; OP= oral primer; R= reverse; S= short amplicon length, 100-300 base pairs.

Tables 4 and 5 show the percentages of detected taxa with and without matching amplicons 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, the short primer pairs 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 large 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 (large, SC-NMA= 93.55%, six MA, OF-MA= 1.06 for both) and OP_F053-KP_R020 (medium, 93.01%, six, 1.06). In contrast, the worst primer pair was OP_F066-KP_R040 (short, 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 (large, six MA, OF-MA= 1.11), 85.93% for KP_F022-KP_R063 (medium, eight, 1.14) to 69.63% for the OP_F066-KP_R013 (short, 35, 1.99). Interestingly, the large 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 MA, 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 (large, 12 MA, OF-MA= 1.08), 88.79% for OP_F114-KP_R031 (medium, 29, 1.26) to 54.21% for OP_F066-OP_R073 (short, 134, 3.45). In terms of overall SC-NMA, the second worst was KP_F078-OP_R010 (short, 66.67%, 48 MA, 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%). However, this primer pair is the most widely used in the literature to detect bacteria and archaea. Additional files 6-13 contain more detailed information on the results of MA- and MA-free species coverage and overabundance parameters (OF and OF-MA values) obtained by the primer pairs tested on the oral-bacteria and oral-archaea genomes.

Table 4. Detected taxa with and without matching amplicons and overabundance estimators obtained by the primer pairs tested on the oral-bacteria genomes.

ALC	Bacterial-specific primer pairs	Detected genomes (%)	Detected species (%)	Detected species with MA (%)	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 MA (%)**	**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 MA (%)	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= large amplicon length, >600 base pairs; M= medium amplicon length, 301-600 base pairs; MA= matching amplicons; OF= overestimation factor; OF-MA= overestimation factor with matching amplicons; OP= oral primer; R= reverse; S= short amplicon length, 100-300 base pairs; SC-NMA= species coverage with no matching amplicons.

Table 5. Detected taxa and overabundance estimators obtained by the primer pairs with the oral archaea genomes.

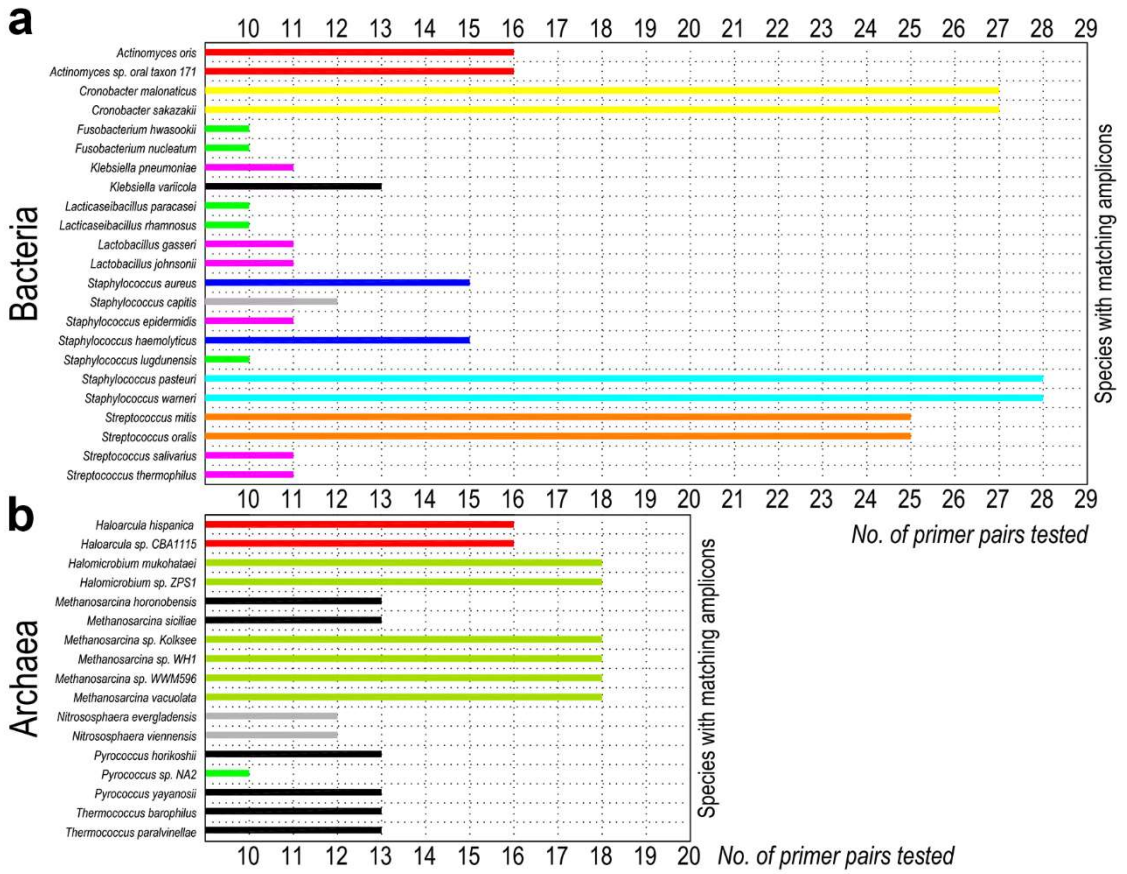
ALC	Archaeal- specific primer pairs	Detected genomes (%)	Detected species (%)	Detected species with MA (%)	SC-NMA (%)	OF	OF-MA
S	KP_F018-KP_R002	185 (96.86)	129 (95.56)	29 (22.48)	100 (74.07)	3.30	1.76
	OP_F066-KP_R013	184 (96.34)	129 (95.56)	35 (27.13)	94 (69.63)	4.02	1.99
M	KP_F018-KP_R032	186 (97.38)	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	1.16
	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 MA (%)**	**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 MA (%)**	**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= large amplicon length, >600 base pairs; M= medium amplicon length, 301-600 base pairs; MA= matching amplicons; OF= overestimation factor; OF-MA= overestimation factor with matching amplicons; OP= oral primer; R= reverse; S= short amplicon length, 100-300 base pairs; SC-NMA= species coverage with no matching amplicons.

Depending on the primer pair tested, between 46.67%-1.29% of the bacterial species and between 38.89%-4.65%. of the archaeal species had MA (Tables 4 and 5). Figure 1 shows the bacterial and archaeal species with MA 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*. Additional files 13-15 define in detail which species from the same or different genera shared MA, 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, universal, and the most widely used in the literature.

DISCUSSION

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 (5,7,25,31), the NCBI's microbial genome database (4,6,8,19), or the ribosomal RNA operon copy number database (rrnDB) (9,32,33). These *in silico* investigations extracted the gene sequences from the complete genomes through tools such as Kodon 2.0 (31,34) or RNAmmer (6,35), 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 (20); all these bacterial and archaeal genomes were downloaded from the NCBI website (19). 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 (7,31) 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) (6,8). 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%)(5,10). 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), *Fusobacterium nucleatum* (=4.25), *Filifactor alocis* (=4.00), *Porphyromonas gingivalis* (=4.00), *Tannerella forsythia* (=2.00), and *Treponema denticola* (=2.00) (36); the caries-associated bacteria *Streptococcus mutans* (=5.00) (37) and *Rothia dentocariosa* (=3.00) (38); and the commensal bacteria *Streptococcus mitis* (=4.00) (39) and *Streptococcus oralis* (=4.00) (40). Some archaeal species that can be found in healthy subjects or those with periodontitis (41) also had more than one gene/genome. These included: *Methanosphaera stadtmanae* (=4.00), *Methanosarcina mazei* (=3.00), *Methanococcus maripaludis* (=2.88), *Methanobrevibacter smithii* (=2.00), and *Sulfolobus acidocaldarius* (=2.00).

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 MA 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 (ASV) (7,42), 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 (20), 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 MA. 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 MA, and the OF caused by the presence of MA (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 large 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 MA 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 a previous study in which we used a larger number of oral taxa from the 16S rRNA gene sequences (20). Assessing the impact that MA could have on species abundance, large primer pairs had OF-MA values as low as 1.08 (e.g. with OP_F114_OP_R121), meaning that the small number of MA 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 MA.

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 MA 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, SC-NMA= 93.55%, OF-MA= 1.06), KP_F018-KP_R063 (for archaea, 89.63%, 1.11); and OP_F114_OP_R121 (for bacteria and archaea

jointly, 92.52%, 1.08). In consequence, we were thus able to demonstrate that sequencing longer fragments enable the identification of lower taxonomy levels (43), reducing the probability of overestimation and classification bias related to MA.

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, 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, 4-5), as described by Caporaso et al. (44). 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 the super-kingdoms of both the bacterial and archaeal domains. 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 (45), 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 MA.

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 MA is not a one-off issue, with 46.67%-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* (13,46) that had MA in at least 10 primer pairs (Figure 1). A 3.00%, 2.00%, 4.00%, 19.30%, 9.00% and 15.00% of all different bacterial or archaeal species with MA detected by all the primer pairs belonged, respectively, to such genera. Among them, there were health-associated species as *Streptococcus mitis* (39) and *Streptococcus oralis* (40), disease-associated taxa as *Fusobacterium nucleatum* (36) and *Streptococcus mutans* (37), or abundant in both states as *Methanosarcina vacuolata* (41); 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*, *Tannerella forsythia,* and *Treponema denticola* (36) also presented both intragenomic gene redundancy and MA.

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 AM 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 matching of amplicons 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.

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 MA, 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 matching amplicons 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.

LIST OF ABBREVIATIONS (alphabetically ordered)

A: archaea

ALC: amplicon length category

ASV(s): amplicon sequence variant(s)

B: bacteria

bp(s): base pair(s)

eHOMD: extended Human Oral Microbiome Database

F: forward

KP: Klindworth primer

L: amplicon large length, >600 base pairs

M: amplicon medium length, 301-600 base pairs

MA: matching amplicons

NCBI: National Center for Biotechnology Information

OF: overestimation factor

OF-MA: overestimation factor associated with matching amplicons

OP: oral primer

OTU(s): operational taxonomic unit(s)

R: reverse

rRNA: ribosomal RNA

rrnDB: ribosomal RNA operon copy number database

S: amplicon short length, 100-300 base pairs

SC-NMA: species coverage with no matching amplicons

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 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.

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; A. Regueira-Iglesias and C. Balsa-Castro made the graphs, tables and additional files; A. Regueira-Iglesias and I. Tomás wrote the manuscript; 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. Taxa with matching amplicons using at least 10 primer pairs: bacterial species (a) and archaeal species (b).

Legend Figure 1. To detect bacteria and archaea, 29 and 20 pairs of primers respectively were tested, including those specific, universal and the most widely used in the literature.

SUPPLEMENTARY INFORMATION

Additional file 1 (Additional_file_1.xlsx). NCBI taxonomy and identifiers of the bacterial genomes.

Additional file 2 (Additional_file_2.xlsx). NCBI taxonomy and identifiers of the archaeal genomes.

Additional file 3 (Additional_file_3.xlsx). Description of the sequences of the selected bacterial-specific, archaeal-specific, and bacterial and archaeal primer pairs.

Additional file 4 (Additional_file_4.xlsx). Description of the sequences of the most used primer pairs in the oral microbiome literature.

Additional file 5 (Additional_file_5.xlsx). Sizes of the bacterial genomes and genes, the number of genes/genome, and the number of gene variants/genome across eight taxonomic ranks.

Additional file 6 (Additional_file_6.xlsx). Sizes of the archaeal genomes and genes, the number of genes/genome, and the number of gene variants/genome across eight taxonomic ranks.

Additional file 7 (Additional_file_7.xlsx). Species with and without matching amplicons and coverage estimators for all primer pairs tested.

Additional file 8 (Additional_file_8.xlsx). Overestimation factor of bacterial species using the bacterial-specific primer pairs.

Additional file 9 (Additional_file_9.xlsx). Overestimation factor of archaeal species using the archaeal-specific primer pairs.

Additional file 10 (Additional_file_10.xlsx). Overestimation factor of the bacterial and archaeal species using the bacteria and archaea primer pairs.

Additional file 11 (Additional_file_11.xlsx). Overestimation factor based on matching amplicons of bacterial species using the bacterial-specific primer pairs.

Additional file 12 (Additional_file_12.xlsx). Overestimation factor based on matching amplicons of archaeal species using the archaeal-specific primer pairs

Additional file 13 (Additional_file_13.xlsx). Overestimation factor based on matching amplicons of bacterial and archaeal species using the bacteria and archaea primer pairs

Additional file 14 (Additional_file_14.xlsx). Bacterial species with matching amplicons using the bacterial-specific primer pairs.

Additional file 15 (Additional_file_15.xlsx). Archaeal species with matching amplicons using the archaeal-specific primer pairs.

Additional file 16 (Additional_file_16.xlsx). Bacterial and archaeal species with matching amplicons using the bacteria and archaea primer pairs.

REFERENCES

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

(2) Woese CR. Bacterial evolution. Microbiol Rev. 1987;51:221-271.

(3) del Rosario Rodicio M, del Carmen Mendoza M. Identificación bacteriana mediante secuenciación del ARNr 16S: fundamento, metodología y aplicaciones en microbiología clínica. Enferm Infecc Microbiol Clin. 2004;22:238-45.

(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;186:2629-35.

(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;76:3886-97.

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

(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:e57923.

(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;73:278-88.

(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;37:489.

(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;10:5029.

(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;10:e0130265.

(12) Duran-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.

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

(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;8:65.

(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;11:929.

(16) Camelo-Castillo A, Novoa L, Balsa-Castro C, Blanco J, Mira A, Tomás I. Relationship between periodontitis-associated subgingival microbiota and clinical inflammation by 16S pyrosequencing. J Clin Periodontol. 2015;42:1074-82.

(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;6:119.

(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;3:187.

(19) NCBI Resource Coordinators. Database resources of the National Center for Biotechnology Information. Nucleic Acids Res. 2016;44:7

(20) 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. Research Square. 2021; doi:10.21203/rs.3.rs-516961/v1

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

(22) Python Software Foundation. Python. Version 3.9.0. 2020. http://www.python.org/.

(23) 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.

(24) 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;44:733.

(25) Clark K, Karsch-Mizrachi I, Lipman DJ, Ostell J, Sayers EW. GenBank. Nucleic Acids Res. 2016;44:D67-72.

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

(27) Edgar RC. SEARCH_16S: A new algorithm for identifying 16S ribosomal RNA genes in contigs and chromosomes. bioRxiv. 2017:124131.

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

(29) McKinney W. Data Structures for Statistical Computing in Python. Proceedings of the 9th Python in Science Conference. Python in Science Conference. 2010; doi: 10.25080/Majora-92bf1922-00a.

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

(31) Coenye T, Vandamme P. Intragenomic heterogeneity between multiple 16S ribosomal RNA operons in sequenced bacterial genomes. FEMS Microbiol Lett. 2003;228:45-9.

(32) Klappenbach JA, Saxman PR, Cole JR, Schmidt TM. rrnDB: the ribosomal RNA operon copy number database. Nucleic Acids Res. 2001;29:181-4.

(33) 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;43:593.

(34) Applied Maths NV. Kodon 2.0. https://www.applied-maths.com.

(35) 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;35:3100-8.

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

(37) Abranches J, Zeng L, Kajfasz JK, Palmer SR, Chakraborty B, Wen ZT, et al. Biology of oral streptococci. Microbiol Spectr. 2018; doi:10.1128/microbiolspec.GPP3-0042-2018.

(38) 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.

(39) Mitchell J. *Streptococcus mitis*: walking the line between commensalism and pathogenesis. Mol Oral Microbiol. 2011;26:89-98.

(40) Thurnheer T, Belibasakis GN. *Streptococcus oralis* maintains homeostasis in oral biofilms by antagonizing the cariogenic pathogen *Streptococcus mutans*. Mol Oral Microbiol. 2018;33:234-9.

(41) 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.

(42) Schloss PD. Amplicon sequence variants artificially split bacterial genomes into separate clusters. bioRxiv. 2021; doi:10.1101/2021.02.26.433139.

(43) 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.

(44) 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.

(45) 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.

(46) Zhang Y, Wang X, Li H, Ni C, Du Z, Yan F. Human oral microbiota and its modulation for oral health. Biomed Pharmacother. 2018;99:883-93.