A Novel Microarray Design Strategy to Study Complex Bacterial Communities

Antoine Huyghe¹*, Patrice Francois¹, Yvan Charbonnier¹, Manuela Tangomo-Bento¹, Eve-Julie Bonetti¹, Bruce J. Paster², Ignacio Bolivar³, Denise Baratti-Mayer⁴, Didier Pittet⁵, Jacques Schrenzel¹ and the Geneva Study Group on Noma (GESNOMA)

¹ Genomic Research Laboratory, Infectious Diseases Service, University of Geneva Hospitals, Switzerland; ² The Forsyth Institute, Boston, MA, USA; ³ Institut für Angewandte Immunologie (IAI), Zuchwil, Switzerland; ⁴ Reconstructive and Plastic Surgery, Faculty of Medicine, Geneva, Switzerland; ⁵ Infection Control Program, Faculty of Medicine, Geneva, Switzerland

Running title: microarray for studying gingival flora composition

Keywords: microarray, complex flora, gingival samples, 16S rRNA gene, phylogenic composition.

* Corresponding author: Antoine Huyghe, Genomic Research Laboratory. Infectious Diseases Service, University of Geneva Hospitals, Micheli-du-Crest 24, 1211 Geneva 14, Geneva, Switzerland. Ph: +41(0)223729338, Fax: +41(0)223729830. email:antoine.huyghe@genomic.ch
ABSTRACT

Assessing bacterial flora composition appears of increasing importance to fields as diverse as physiology, development, medicine, epidemiology, environment and food industry. We report here the development and validation of an original microarray strategy that allows analyzing phylogenic composition of complex bacterial mixtures. The microarray contains ~9,500 feature elements targeting 16S rDNA specific regions. Probe design was performed by selecting oligonucleotide sequences specific to each node of the 7 levels of the bacterial phylogenetic tree (domain, phylum, class, order, family, genus and species). This approach, based on sequence information, allows analyzing bacterial contents of complex bacterial mixtures to detect known but also unknown microorganisms. The presence of unknown organisms can be suspected and mapped on the phylogenetic tree, indicating where to refine analysis. Initial proof of concept experiments were performed on oral bacterial communities. Our results show that this hierarchical approach is able to reveal minor changes (≤1%) in gingival flora content when comparing samples collected in individuals from similar geographical origin.
INTRODUCTION

Assessing bacterial flora composition appears of increasing importance to analyze progressive gut colonization after birth (18, 34), to unravel bacterial role in the epidemics of obesity in humans (44) or to better understand flora changes upon antibiotic selection pressure or in various disease states. The gingival flora was selected as a proof of concept for this study. It contains a wide variety of microorganisms, either commensals or potential pathogens but the vast majority remains simply uncultured or unknown. Complexity of this bacterial flora benefits from very diverse growth environments (O2, CO2, pH, accessibility to nutrients and epithelial debris, etc.) creating an array of anatomic niches for bacterial colonization. To illustrate this diversity, one estimates the number of bacterial species found in all oral niches to approximately of 500-700 (1, 35). Some of these bacteria are clearly associated with cavities, gingivitis and periodontitis, and might be implicated in less commonly encountered lesions such as noma (13, 14). Furthermore, particular bacterial species commonly found in the oral cavity can contribute to systemic diseases. Han et al. (20) showed that Fusobacterium nucleatum, associated with periodontal disease, can be implicated in preterm births. Similarly, the presence of Porphyromonas gingivalis has been documented to trigger inflammation in aortic endothelial cells (7, 42). In accordance with these observations, characterization of the bacterial oral flora appears essential for the study of poorly understood diseases, such as noma, and for the development of novel diagnostic approaches.

A classical way to characterize members of complex bacterial communities relies
on 16S rDNA sequence analysis. This target is particularly adapted to phylogenic studies as it contains highly conserved and variable moieties permitting reliable and detailed bacterial classification (12, 25, 37). In this approach, nucleic acids are directly extracted from samples without any prior cultivation; amplification is then performed using universal primers targeting conserved stretches of the 16S rDNA gene and identification is based on similarity with sequences deposited in public ribosomal gene databases. As the rate of nucleotide sequence change correlates with the evolutionary distance between organisms (8), sequence relatedness can be used for a phylogenetic approach. To analyze complex bacterial flora, large-scale cloning and sequencing of 16S rDNA targets can provide a detailed catalogue of bacterial flora from a representative sample. But despite its accuracy and potential for bacterial quantitation, this approach cannot be applied to compare large groups of samples exhibiting intra- and inter-variability, due to the very large number of experiments to perform.

Microarrays are frequently used to monitor gene regulation and expression on a genome-wide scale (6). Still, development of this technology for the comprehensive characterization of the content of complex microbial communities is still underway and different approaches have been proposed. Functional gene arrays (49) target enzymes involved in a peculiar metabolic process such as methane oxidation (2), nitrogen fixation (32, 43), sulfur metabolism or iron metabolism (50). While allowing characterization of key bacteria involved in a determined metabolism pathway, this approach is also useful to monitor the functional state of a microbial community under
different environmental conditions. Alternatively, profiling of prokaryotic populations has been achieved by using 16S rRNA gene microarrays with probes targeting bacterial groups such as *Cyanobacteria* (5), *Rhodocyclales* (31), or *Alphaproteobacteria* (40). Development of high density microarrays allowed extending the scope of phylogenetic oligoarrays to the whole bacterial kingdom (3, 33, 48). Although 16S rRNA microarrays do not appear optimal for discovering new taxa, this approach has permitted the detection of a broader bacterial diversity as compared to the use of clone libraries (10). The utilization of microarrays is appealing to evaluate samples containing complex flora. The design of oligonucleotide probes appears mandatory to resolve punctual sequence differences compared to PCR products that have been shown to exhibit poor performance for SNP or punctual mutation analysis (19, 27, 29, 38). Moreover oligonucleotide probes are more flexible and can be tailored to meet critical criteria such as sequence specificity and physicochemical properties.

To enable large-scale studies of complex bacterial flora composition in collections of samples, we developed an original oligonucleotide microarray design based on a phylogenetic approach. Microarray design was performed by selecting ~9,500 25 nt probes recognizing 16S rDNA targets that were specific to nodes matching the seven levels of the bacterial phylogenetic tree (domain, phylum, class, order, family, genus and species). While providing information on the taxonomic composition of microbial communities, this approach should also prove useful for detecting uncharacterized species or to detect over- or
under-representation of specific bacterial groups leading to imbalanced flora content. Our study shows that this hierarchical approach is able to reveal minor changes in microflora composition -as low as 1% of the global composition- when comparing two complex, but related, bacterial populations.

MATERIALS AND METHODS

Microarray design and manufacturing. In-house software was developed to produce the best probe set maximizing node coverage while respecting the number of available array features (MS in preparation). The input to this program is a list of the 194,696 bacterial small-subunit (SSU) rDNA sequences classified by the Ribosomal Database Project (RDP release 9.34) in 32 different taxonomic groups (9). The program then proceeds in 4 stages:

i) Each SSU rDNA sequence was scanned from its 5’- to 3’-end in order to extract all possible 25 nucleotide sub-sequences as a pool of candidate probes. Melting temperatures (Tm) were determined for each candidate probe using thermodynamic parameters based on a nearest-neighbor model (41). Candidate probes with a predicted Tm of 60 ± 5°C were stored in a hash table structure to eliminate duplicate sequences and to permit rapid data processing.

ii) The next stage consisted in assigning a node to each candidate probe. Probes matching several SSU rDNA sequences were assigned to the nearest parent node common to the referred sequences. For example, a probe matching both *Streptococcus* and *Lactococcus* genus was assigned to the *Streptococcaceae*
order (see ‘Probe C’ in Figure 1).

iii) Due to the limited number of microarray features (10,263) and also due to the large number of candidate probes (2,509,422), we decided to maximize node coverage by selecting only probes that provided substantial coverage (i.e. \( \geq 1.2\% \) coverage of all sequences of a given node, as determined empirically).

iv) Earlier steps provided a set of 8,195 probes describing phylogenetic classes from the domain to the species level. About 2.3 % of these probes presented one or more ambiguous nucleotides. Since most of these polymorphisms conveyed a significant contribution to phylogenetic coverage, we decided to consider all possible degenerated positions for this subset of targets, yielding to a final probe set of 9,477 probes resulting in a global coverage of 78.3%. To minimize steric hindrance, all 25-mer probes were poly(T)-tailed to reach an overall length of 60 nucleotides. Microarrays were manufactured by in situ synthesis (Agilent Technologies, Palo Alto, USA).

**Biological samples.** Two distinct bacterial mixtures were used to validate our microarray approach. We first generated a defined artificial sample (named sample A) using equal amounts of cRNA originating from three different organisms: i) *Streptococcus pyogenes* ATCC 12344, ii) *Fusobacterium necrophorum* ATCC 25556 and iii) *Chromobacterium violaceum* ATCC 12472. Artificial sample A (200 ng total cRNA) was compared to the same bacterial mixture previously spiked with 25% (i.e. 50 ng total RNA) *Escherichia coli* ATCC
25922, yielding to another 200 ng cRNA sample (spiked sample A).

Sterile endodontic paper points were used to collect gingival fluid from the dentogingival sulcus of two healthy European male subjects aged 28 and 41 years (samples B1 and B2, respectively). Samples were stored in RLT buffer (RNeasy Mini kit, Qiagen, Basel, Switzerland) at -80°C for subsequent analyses. 2µg cRNA of gingival samples B1 and B2 were then compared against their equivalents, but previously spiked with a lower concentration (1%) of *Fusobacterium necrogenes* ATCC 25556 (spiked sample B1 and spiked sample B2, respectively).

**RNA extraction and quantification.** To lyse cells, 100 mg of glass beads (diameter: 100µm; Schieritz & Hauenstein, Switzerland) were added to the samples. Volume was adjusted to 350 µl with RLT buffer and samples were vortexed for 1 minute. Total RNA was isolated and purified using the RNeasy Micro Kit (Qiagen) following manufacturer's instructions. Samples were lyophilized and dissolved in 5 µl sterile water. Total RNA quality was assessed using RNA Picochips on a BioAnalyzer 2100 (Agilent). RNA quantity was assessed by one-step RT-qPCR using 0.2 µM of primers Forward: GGCAAGCGTTATCCGGAATT, reverse: GTTTCCAATGACCCTCCACG (Invitrogen, Basel, Switzerland) and 0.1 µM of probe (probe: CCTACGCGCGCTTTACGCCCA, 5'-end coupled to FAM and 3'-end coupled to TAMRA, Eurogentec, Seraing, Belgium) designed in a highly conserved region of bacterial 16S rRNA gene allowing amplification of most bacterial 16S rRNA sequence. One-step RT-qPCR amplification (Invitrogen, final volume of 15 µl) was performed on a SDS 7700 (PE Biosystems, Santa Clara, CA, USA) using the following cycling procedure: t1, 20 min at 50°C; t2, 10 min at 94°C;
t3, 15 sec at 94°C; t4, 1 min at 60°C (t3 and t4 were repeated 40 times). Using this strategy, a positive fluorescent signal was obtained between cycles 21-30.

**RNA amplification.** The totality of purified RNA was subjected to *in vitro* transcription using the MessageAmp II-Bacteria kit (Ambion, Austin, TX, USA) according to manufacturer’s instructions. Amplified RNA was labeled during *in vitro* transcription in the presence of Cy-3 or Cy-5 cyanine dyes (Perkin-Elmer, Boston, MA, USA). Quality, quantity, amplification efficiency and dye incorporation were evaluated using the NanoDrop® ND-1000 Spectrophotometer (NanoDrop Technologies, Inc. Rockland, DE, USA) and the BioAnalyzer 2100 on RNA Nano 6000 chips (Agilent).

**Microarrays hybridization, scanning and analysis.** All samples were hybridized in duplicate. Cy5- and Cy3-labelled cRNAs were diluted in a total of 250 µl Agilent hybridization buffer, and hybridized at 60°C for 17 hours in a dedicated hybridization oven (Robbins Scientific, Sunnyvale, CA, USA). Slides were washed, dried under nitrogen flow, and scanned (Agilent) using 100% PMT power for both wavelengths. Image analysis and signal quantification were achieved using Feature Extraction software (version 6, Agilent). Probes exhibiting a non-uniform signal (i.e. pixel noise exceeding an established threshold) or mean signal values inferior to the corresponding background plus 2.6 standard deviations were excluded from subsequent analyses.
For spiking experiments, LOWESS (locally weighted linear regression) transformation was used to correct background subtracted signals for unequal dye incorporation, and a geometric mean was applied to average signals between duplicates. Statistical analysis consisted of a 2-tailed Student’s t-Test with a p-value tailored according to the relative spike abundance (p-value < 0.01 for sample A, 25% spiking; p-value <0.05 for samples B1 and B2, 1% spiking).

For assessing the bacterial subgingival flora on non-spiked samples, local background was subtracted from raw mean signal. Subsequently, probe signals were averaged among duplicates using a geometric mean.

**Comparison with 16S rDNA sequencing.** To compare obtained results with data generated by large-scale cloning and sequencing approaches in different populations of healthy volunteers, we used published data of Kroes (26) and Paster (35). Briefly, Paster et al. used clone library sequencing to investigate the bacterial diversity in the subgingival plaque of healthy subjects and subjects affected by various kinds of periodontitis and acute necrotizing ulcerative gingivitis. From the 5 libraries matching healthy subjects we chose 3 libraries showing the widest diversity in species and phylotypes in order to perform a direct comparison with our data. In the same way we used the data from the study of Kroes describing the bacterial diversity found in a single healthy human volunteer, likewise using clone library sequencing.
RESULTS

Microarray design

The final set of 9,477 phylogenic 25-mer probes covers 78.3% of the 194,696 SSU rDNA sequences listed in the RDP database, i.e. 1 single probe matches an average of 16 sequences. Node coverage at the phylum level, defined as the percentage of sequences covered by probes matching a given node, averages to 78.2% and ranges from 21% for the Chloroflexi phylum to 100% for Chrysiogenetes, Chlamydiae, Dictyoglomus and Incertae sedis BRC1 phyla (Figure 2). Nodes or phyla that might be particularly important in the investigation of oral diseases of unknown etiology - like noma (14, 36) - show appreciable coverage yields, such as Fusobacteria (99%), Spirochaetes (89%) and Bacteroidetes (including Prevotella, 89%) phyla.

Bacterial detection from a defined mixture. To assess the reliability of our approach, we first attempted to detect differences in bacterial composition by comparing a relatively simple mixture of three distinct bacterial species with its equivalent, but supplemented with another bacteria representing 25% of the total mixture. By performing duplicate experiments, we checked whether the spiked species, Escherichia coli, could be reliably detected from this defined sample and whether it would induce unexpected changes in the identification scheme. Probes showing a fold change >2 and a p-value <0.01 were considered as representing targets differentially present between compared samples. Among 9,477 probes present on our microarray, 6,991 probes (74%) yielded signals above
background values, but only 97 probes (1%) displayed statistical significance in spiked and non-spiked samples. The ‘volcano-plot’ depicted in Figure 3A shows p-values of the probes producing a fluorescent signal plotted against their fold change. Analysis reveals that 95 probes yielded statistically significant signals in the spiked mixture (upper-right side of panel A), while only 2 probes were detected as significant in the non-spiked sample (upper-left side). In the spiked sample, 56 probes out of 95 (59%) matched genus and family nodes that represent the *Gammaproteobacteria* class. The most significant feature is a probe matching the *Escherichia* genus with a fold change >300 (p-value = 0.004) which is the only probe at this level to match *Escherichia coli* ATCC 25922. The second most relevant feature (fold change >130, p-value = 0.0027) of this analysis is a probe matching the class of the *Gammaproteobacteria* that also includes the *Escherichia* genus. The third probe matching the *Alcalilimnicola* genus (*Gammaproteobacteria* class) shows a fold change of ~109 (p-value = 0.0004) and would not be expected, based on *in silico* predictions. Partial cross-hybridization could also explain why probes that do not directly match *Escherichia coli* ATCC 25922, but other nodes, such as *Shigella*, are revealed by our analysis. Note however that these two species displayed strongly homologous ribosomal gene sequences.

**Detecting bacterial changes from complex samples.** We then assessed the ability of our microarray to detect minor changes in the flora composition of clinical samples from healthy subjects. Nucleic acids were extracted from a complex
bacterial flora (i.e. subgingival samples) obtained from two healthy volunteers. Nucleic acids were labeled, hybridized and compared to the same samples spiked with 1% *Fusobacterium necrophorum* ATCC 25556. *F. necrophorum* was selected as a bacterium that does not belong to the normal oral flora (data not shown) i) to maximize the diversity of spiking material and ii) to assess whether lower detection sensitivity could be achieved. Total amounts of cRNA loaded on the microarray were adapted using 2 µg (instead of 200 ng) and a much lower percentage of spiked material (1% instead of 25%). We also adjusted the analysis parameters to compare samples using a 2-fold change and a p-value of 0.05 as the sample content is presumably highly similar. 27 probes revealed statistical significance in the spiked sample B1, while only 8 probes were detected in the spiked sample B2. In both analyses, the large majority of probes with a fold change >3 belonged to the *Fusobacterium* genus, the *Fusobacteriales* order or other closely phylogenetically related nodes (Table 1). In both samples, probes matching the *Fusobacterium* genus showed the most significant fold changes: ~106-fold for sample B1 and ~45-fold factor for sample B2. In contrast, no probes yielded statistically significant signals in the non-spiked samples, as illustrated by the upper left corners of the volcano plots (Figures 3B and 3C).

Figure 4 depicts probes that were detected as significant during this spiking experiment, using sample B2. For better readability, they are plotted on the actual phylogenetic tree trimmed from phyla where no significant signal was detected. Four probes (see Table 1, Sample B2) adequately refer to the expected nodes in the phylogeny of the spiked bacterium, *Fusobacterium necrophorum*. One other probe
points to a node identified as “Unclassified Fusobacterales” and shows a moderate increase in fold change (3.7). This observation can be explained by the strong homology of this probe with the sequence of the spiked target which is highly related. Only two other probes appear as unrelated to this phylogeny and map environmental and gut flora SSU sequences belonging to Melittangium and Saprospira genera (Table 1, Sample B2). Interestingly, these two probes revealed very high homology - position 23: GGAATCGCTAGTAATCGCAAAT(G/C)AG ; and positions 1, 9, 23 and 24: (G/T)TGCGTCC(T/C)ATTAGCTAGATGG(TA/AG)A, respectively - to the sequence of the spiked organism.

Defining the bacterial gingival flora in healthy subjects using phylogenic microarray.

We analyzed the bacterial composition of healthy subgingival samples. Analysis of sample B1 and B2 yielded in the identification of 31 and 114 probes, respectively, with a significant fluorescent level. 19 probes overlapped between the two samples. While all these probes matched different phylogenetic levels, a large majority matched at the genus level – such as Actinomyces, Bacteroides, Capnocytophaga, Gemella, Porphyromonas, Prevotella, Rickettsiella, Streptococcus and TM7, as reported in previous studies (1, 26, 30, 35). At the phylum level, both samples displayed approximately the same flora profile (Figure 5) encompassing Proteobacteria, Firmicutes, Actinobacteria and Bacteroidetes. Interestingly, the
*Bacteroidetes* phylum appeared more represented in sample B1 than in B2. Some rare and unexpected phyla were detected with our microarray approach, such as *Aquificae* and *Planctomycetes*. This part of the flora constituting a small proportion of the total bacterial content and was not identified in previous study as part of the gingival flora, but mainly found in environmental samples (11, 24, 47). We finally sought to compare the results of our approach with those of previously published studies (26, 35). We should emphasize here that these studies on the gingival flora were markedly different since volunteers were recruited in another continent (USA) and that the flora determination was performed by conventional generic amplification of 16S rRNA genes, then cloning and sequencing strategy. Yet, despite these caveats, we observed a remarkably similar composition across phyla and abundance thereof (Figure 5), strongly suggesting a rather common composition in the gingival flora between these subjects, at the phylum level. However, our data suggest that additional phyla were part of the normal oral flora.

**DISCUSSION**

Based on an original hierarchical phylogenetic design, the ~9,500 probe set on our oligoarray covers 78.3% of the 194,696 bacterial SSU rDNA sequences described in release 9.34 of RDP. Although phyla coverage ranges from 21% (*Chloroflexi*) to 100% (*Chlamydiae* e.g.), the design process was not intended to be limited to the study of specific phyla and thus should prove useful to study the whole eubacterial domain, which was the starting point of our strategy. This last point is of crucial importance since no specific phylum should have *a priori* more weight during the
profiling of any altered oral flora.

Potential applications of such microarray consist in detecting bacterial composition changes over time or across many related samples (e.g. healthy versus diseased or site, for the study of noma for example).

In this project, the limited sensitivity of microarray techniques required amplification of the starting nucleic acids. In addition, the amplification strategy proposed in our study appears robust and is potentially utilizable to other samples where amounts of starting material are strictly limited. Our strategy proved to be efficient in detecting minor differences (as low as 1% of *F. necrogenes*) in flora composition in controlled mixtures and, more importantly, in complex natural samples. Our approach was able to characterize the diversity of the gingival flora **down to the genus level** in 2 healthy European subjects. Results showed good congruence with previous studies performed in American subjects (26, 35) by using a cloning-sequencing strategy. This observation is noteworthy because this comparison involved volunteers originating from two distinct continents and using two markedly different methods. Our results suggest that different social and dietary habits have limited influence on gingival flora composition, as defined by our microarray strategy. This microarray approach revealed a broader diversity of microorganisms at the genus level than traditional clone libraries methods, a finding that has already been reported by DeSantis et al. (10). The underestimation of microbial diversity may be explained either by a cloning bias or by the paucity of clones sequenced. Moreover, as our strategy is based on 16S rRNA sequences, we
can hypothesize an additional bias due to sequence overrepresentation in public sequence databases (23). This bias can be expected on important human or animal pathogens as well as organisms of economical interest. Nonetheless, this phylogenetic approach not only allows monitoring bacterial population dynamics in different ecosystems, but also detecting potential pathogenic bacteria or bacterial taxons as recently illustrated in the study of airborne bacterial composition (4). In addition, the same strategy, as well as functional gene arrays, can be implemented to evaluate temporal evolution of complex bacterial communities such as in agricultural, environmental or human commensal ecosystems. Such applications include evaluating the impact of ecosystem modification on environmental flora (3) or the identification of specific bacterial population showing particular metabolic capacities, such as metabolism of toxic compounds (21, 39). In human medicine, a potential utilization would be the monitoring of flora composition alteration or bias composition, due to chemical, antimicrobial treatments (15) or due to metabolic dysfunction, or to understand its natural evolution during the lifecycle (34). This type of strategy represents serious advantages compared to culture or even other non-culture-based methodologies (22). The characteristics of sensitivity and specificity as well as the actual throughput are now compatible with real time monitoring or analysis (17).

In our case, the efficiency of our microarray could be improved in several ways. We could rely on novel 16S sequences retrieved from the latest
release of RDP as well as data emerging from various metagenomic projects (16, 45). Additionally, it would be useful to add probes targeting the Archaebacteria domain, since methanogenic Archaea have been recently reported in subgingival sites of patients suffering from periodontitis (28). Obviously, our design is dependent on the quality of the sequences retrieved by RDP. In this regard, short 16S rDNA sequences may be prone to misclassification at the genus level due to anomalies in the taxonomy and lack of data on short sequences (46). In addition, close sequence homology across different genera (e.g. Escherichia and Shigella) explains why specific probes can not always be designed. The potential of cross-hybridization by the probes should not be minimized, especially in the highly conserved 16S rDNA sequences. Whereas careful assessment of such cross-hybridization potential appears warranted for the “de novo” characterization of a microbial community, this is clearly less of a concern here because our approach was developed and tested to detect differences between related samples.

Finally, the arbitrary 2-fold change cut-off was selected as an empirical compromise between sensitivity and specificity. Further experimental determinations are now warranted to more precisely define this cut-off value and determine whether it can be applied to the whole range of fluorescence signals. It remains to be proven whether statistical approaches might prove more reliable in the long run.

Preliminary steps to the identification of uncharacterized bacterial species can be performed using our hierarchical approach, providing an alternative to classical
sequencing techniques. **However, this point remains to be experimentally proven.** Initially intended for the study of bacterial diversity in oral samples, our approach appears potentially useful for the study of various bacterial communities associated either with medical (intestinal or skin flora), environmental (soil, sludge, wastewaters, etc.), or food industry samples. Additionally, the same approach could be used to monitor the evolution of bacterial communities over time, replacing laborious techniques such as library cloning and sequencing.

This approach can be conveniently implemented to perform large-scale profiling studies of the oral bacterial flora. Experiments are underway for monitoring gingival flora in noma lesions from individual patients, and comparing them to matched healthy controls from the same geographical origin. **Again, implementing such novel microarray strategies for flora composition analyses require careful validation.**

**Acknowledgements.** This work was supported by a grant from the GESNOMA Foundation, a research program on the etiology of noma. PF and JS are supported by grants from the Swiss National Science Foundation 3100A0-112370/1 and COST C05.0103 (JS), 3100A0-116075/1 (PF), and NIH grant DE-11443 (BJP).
References


TABLE1. Listing of nodes identified as statistically significant in spiked samples B1 and B2.

### Sample B1

<table>
<thead>
<tr>
<th>Node</th>
<th>Description</th>
<th>Rank</th>
<th>Fold Change</th>
<th>P-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.21.1.1.1.1</td>
<td>Fusobacterium</td>
<td>Genus</td>
<td>106.23</td>
<td>0.037</td>
</tr>
<tr>
<td>1.21.1.1.1.1</td>
<td>Fusobacteriales</td>
<td>Order</td>
<td>10.15</td>
<td>0.018</td>
</tr>
<tr>
<td>1.21.1.1.1.1</td>
<td>Fusobacterium</td>
<td>Genus</td>
<td>5.84</td>
<td>0.024</td>
</tr>
<tr>
<td>1.21.1.1.1.1</td>
<td>Fusobacterium</td>
<td>Genus</td>
<td>5.26</td>
<td>0.019</td>
</tr>
<tr>
<td>1.21.1.1.1.3</td>
<td>Unclassified Fusobacteria</td>
<td>Unclassified</td>
<td>5.25</td>
<td>0.025</td>
</tr>
<tr>
<td>1.21.1.1.1.3</td>
<td>Unclassified Fusobacteria</td>
<td>Unclassified</td>
<td>5.06</td>
<td>0.035</td>
</tr>
<tr>
<td>1.21.1.1.1.2</td>
<td>Ilyobacter</td>
<td>Genus</td>
<td>3.97</td>
<td>0.010</td>
</tr>
<tr>
<td>1.21.1.1.1.2</td>
<td>Celobacterium</td>
<td>Genus</td>
<td>3.69</td>
<td>0.021</td>
</tr>
<tr>
<td>1.13.1</td>
<td>Clostridia</td>
<td>Class</td>
<td>3.66</td>
<td>0.007</td>
</tr>
<tr>
<td>1.21.1.1.1.3</td>
<td>Propionigenium</td>
<td>Genus</td>
<td>3.01</td>
<td>0.015</td>
</tr>
<tr>
<td>1.20.3.1.2.4</td>
<td>Unclassified Saprospiraceae</td>
<td>Unclassified</td>
<td>2.90</td>
<td>0.014</td>
</tr>
<tr>
<td>1.21.1.1.1.1</td>
<td>Fusobacterium</td>
<td>Genus</td>
<td>2.67</td>
<td>0.006</td>
</tr>
<tr>
<td>1.12.3.4.1.6</td>
<td>Rhodanobacter</td>
<td>Genus</td>
<td>2.60</td>
<td>0.037</td>
</tr>
<tr>
<td>1.20.2.1.3</td>
<td>Unclassified Flavobacteria</td>
<td>Unclassified</td>
<td>2.47</td>
<td>0.029</td>
</tr>
<tr>
<td>1.12.3.8.1.8</td>
<td>Unclassified Oceanospirilaceae</td>
<td>Unclassified</td>
<td>2.46</td>
<td>0.013</td>
</tr>
<tr>
<td>1.12.3.11.1.8</td>
<td>Psychromonas</td>
<td>Genus</td>
<td>2.21</td>
<td>0.039</td>
</tr>
<tr>
<td>1.14.1.4.26.1</td>
<td>Pseudonocardia</td>
<td>Genus</td>
<td>2.21</td>
<td>0.016</td>
</tr>
<tr>
<td>1.21.1.1.3</td>
<td>Unclassified Fusobacteria</td>
<td>Unclassified</td>
<td>2.16</td>
<td>0.002</td>
</tr>
<tr>
<td>1.12.2.6.1.6</td>
<td>Sterolbacterium</td>
<td>Genus</td>
<td>2.15</td>
<td>0.044</td>
</tr>
<tr>
<td>1.20.3.1.5.4</td>
<td>Unclassified Crenotrichaceae</td>
<td>Unclassified</td>
<td>2.14</td>
<td>0.012</td>
</tr>
<tr>
<td>1.21.1.1.3</td>
<td>Unclassified Fusobacteria</td>
<td>Unclassified</td>
<td>2.13</td>
<td>0.022</td>
</tr>
<tr>
<td>1.12.3.4.1.1</td>
<td>Xanthomonas</td>
<td>Genus</td>
<td>2.11</td>
<td>0.024</td>
</tr>
<tr>
<td>1.12.3.14.1.8</td>
<td>Unclassified Pasteurellaceae</td>
<td>Unclassified</td>
<td>2.06</td>
<td>0.019</td>
</tr>
<tr>
<td>1.12.1.3.1.4</td>
<td>Antarctobacter</td>
<td>Genus</td>
<td>2.04</td>
<td>0.019</td>
</tr>
<tr>
<td>1.25.1</td>
<td>BRC1</td>
<td>Genus</td>
<td>2.03</td>
<td>0.009</td>
</tr>
<tr>
<td>1.22.1.1.1.3</td>
<td>Unclassified Verrucomicrobiaceae</td>
<td>Unclassified</td>
<td>2.01</td>
<td>0.018</td>
</tr>
<tr>
<td>1.14.1.2.1.1</td>
<td>Rubrobacter</td>
<td>Genus</td>
<td>2.01</td>
<td>0.019</td>
</tr>
</tbody>
</table>

### Sample B2

<table>
<thead>
<tr>
<th>Node</th>
<th>Description</th>
<th>Rank</th>
<th>Fold Change</th>
<th>P-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.21.1.1.1.1</td>
<td>Fusobacterium</td>
<td>Genus</td>
<td>45.6</td>
<td>0.032</td>
</tr>
<tr>
<td>1.21.1.1.1.1</td>
<td>Fusobacteriales</td>
<td>Order</td>
<td>12.6</td>
<td>0.016</td>
</tr>
<tr>
<td>1.21.1.1.1.1</td>
<td>Fusobacterium</td>
<td>Genus</td>
<td>8.4</td>
<td>0.037</td>
</tr>
<tr>
<td>1.21.1.1.1.3</td>
<td>Unclassified Fusobacteria</td>
<td>Unclassified</td>
<td>3.7</td>
<td>0.049</td>
</tr>
<tr>
<td>1.12.4.6.1.3</td>
<td>Melittangium</td>
<td>Genus</td>
<td>3.1</td>
<td>0.046</td>
</tr>
<tr>
<td>1.20.3.1.2.4</td>
<td>Unclassified Saprospiraceae</td>
<td>Unclassified</td>
<td>3.0</td>
<td>0.031</td>
</tr>
<tr>
<td>1.21.1.1.1.2</td>
<td>Celobacterium</td>
<td>Genus</td>
<td>2.5</td>
<td>0.048</td>
</tr>
<tr>
<td>1.21.1.1.1.1</td>
<td>Fusobacterium</td>
<td>Genus</td>
<td>2.1</td>
<td>0.023</td>
</tr>
</tbody>
</table>
**Figure 1.** Schematic representation of the probe selection process on a small subset of 16S rDNA sequences. Candidate probes are compared to the library of 16S rDNA sequences and assigned to the most distal common node. For example, Probe A which is common to *Lactobacillus*, *Paralactobacillus*, *Pediococcus* and *Streptococcus* is assigned to the *Lactobacillales* order level. Probe C is assigned to the family level of the *Streptococcaceae* since it detects both, *Streptococcus* and *Lactococcus*. In contrast, Probe B is specific to some *Pediococcus* species and it is therefore assigned to the genus level.

**Figure 2.** Microarray coverage for the 32 phyla described in the Ribosomal Database Project (see Materials and Methods). Bars represent the percentage of strains detected within each phylum. The number of selected probes for providing this phylum coverage is specified to the right of the Figure.

**Figure 3.** Volcano plots display a summary of test statistics as a function of fluorescence intensity ratios (i.e. fold change). Each plot compares spiked versus non–spiked samples, using material of increasing bacterial complexity. Red dots in the upper-left or upper-right corners depict significant targets in the non-spiked sample or in the spiked sample, respectively. **Panel A:** Comparison of one defined bacterial mixture containing three bacterial species with/without a spike of 25% of the nucleic acid amounts (significance is defined as fold-change $\geq 2$ and $p \leq 0.01$).
Comparison of gingival flora from two healthy volunteers (B1: panel B; B2: panel C) with/without a spike of 1% of the nucleic acid amounts (significance is defined as fold-change ≥ 2 and p ≤0.05).

**Figure 4.** Phylogenetic tree representation of all phyla detected by statistical analysis (P<0.05) when comparing a gingival sample (B2) to its replicate spiked with 1% *Fusobacterium necrophorum* ATCC 25556. Red dots and lines depict nodes and branches where at least 1 probe yielded a statistically significant signal. For better readability, only the alpha Proteobacteria class is fully represented among the Proteobacteria phylum.

**Figure 5.** Cumulative prevalence of phylotypes identified using microarrays (samples B1 and B2) compared to previously published analyses of gingival flora. Figure depicts all phylotypes identified by any of these studies. Lib1, Lib2 and Lib3 describe libraries generated on 3 healthy subjects and described by Paster et al. (35) (Paster, personal communication).