

SCIENTIFIC REPORTS



OPEN

Inflammatory bacteriome featuring *Fusobacterium nucleatum* and *Pseudomonas aeruginosa* identified in association with oral squamous cell carcinoma

Nezar Noor Al-hebshi^{1,2}, Akram Thabet Nasher³, Mohamed Yousef Maryoud¹, Husham E. Homeida¹, Tsute Chen⁴, Ali Mohamed Idris^{1,5} & Newell W. Johnson⁶

Studies on the possible association between bacteria and oral squamous cell carcinoma (OSCC) remain inconclusive, largely due to methodological variations/limitations. The objective of this study was to characterize the species composition as well as functional potential of the bacteriome associated with OSCC. DNA obtained from 20 fresh OSCC biopsies (cases) and 20 deep-epithelium swabs (matched control subjects) was sequenced for the V1-V3 region using Illumina's 2 × 300 bp chemistry. High quality, non-chimeric merged reads were classified to species level using a prioritized BLASTN-algorithm. Downstream analyses were performed using QIIME, PICRUSt, and LEfSe. *Fusobacterium nucleatum subsp. polymorphum* was the most significantly overrepresented species in the tumors followed by *Pseudomonas aeruginosa* and *Campylobacter sp.* Oral taxon 44, while *Streptococcus mitis*, *Rothia mucilaginosa* and *Haemophilus parainfluenzae* were the most significantly abundant in the controls. Functional prediction showed that genes involved in bacterial mobility, flagellar assembly, bacterial chemotaxis and LPS synthesis were enriched in the tumors while those responsible for DNA repair and combination, purine metabolism, phenylalanine, tyrosine and tryptophan biosynthesis, ribosome biogenesis and glycolysis/gluconeogenesis were significantly associated with the controls. This is the first epidemiological evidence for association of *F. nucleatum* and *P. aeruginosa* with OSCC. Functionally, an "inflammatory bacteriome" is enriched in OSSC.

Oral cancer, predominantly squamous cell carcinoma (OSCC), is the 17th most common malignant neoplasm worldwide and the 8th in less developed regions¹; it continues to have poor prognosis with 5-year survival rates less than 50% in much of the world^{2,3}. OSCC has a number of established risk factors including use of various forms of tobacco, both smoked and smokeless, drinking alcohol, human papilloma virus (HPV) infections, nutrient deficiency, solar radiation and genetic predisposition^{4,5}. However, a significant proportion of OSCC (around 15%) is not explained by these risk factors⁶, suggesting existence of other as yet unidentified risk factors worth exploring.

Recently, there has been increasing interest in the possible role of bacteria in oral carcinogenesis, inspired by the established role of some bacteria in certain types of cancer such as that of *H. pylori* in gastric cancer⁷. A number of studies have been carried out to this effect using various methods ranging from cultivation to 16S rRNA gene sequencing^{8–20}. However, the results have been inconsistent across them. On the one hand, this is probably due to the significant methodological variations among these studies in terms of technology used for microbial

¹Department of Maxillofacial Surgery and Diagnostic Sciences, College of Dentistry, Jazan University, Jazan, Saudi Arabia. ²Kornberg School of Dentistry, Temple University, Philadelphia, PA, USA. ³Department of Oral and Maxillofacial Surgery, Faculty of Dentistry, Sana'a University, Yemen, Saudi Arabia. ⁴Department of Microbiology, Forsyth Institute, Cambridge, MA, USA. ⁵Substance Abuse Research Centre (SARC), Jazan University, Jazan, Saudi Arabia. ⁶Menzies Health Institute Queensland and School of Dentistry and Oral Health, Griffith University, Queensland, Australia. Correspondence and requests for materials should be addressed to N.N.A. (email: alhebshi@temple.edu)

Variable	Cases (n = 20)	Controls(n = 20)
Age (mean ± SD)	53.6 ± 10.4	52.3 ± 8.9
% males	50	50
Site: No. (%)		
Tongue	10 (50)	9 (45)
Gum	05 (25)	5 (25)
Floor of the mouth	04 (20)	5 (25)
Buccal	01 (05)	1 (05)
% Shammah (Arabian snuff) users	80	15

Table 1. Characteristics of the cases and control subjects included in the study (N = 40).

analysis, type of samples obtained (biopsy, surface swab or saliva) and selection of controls (self or other subject)⁷. On the other hand, it may well be that the microbial association with OSCC is at the level of the bacterial community's function, rather than at composition level. In other words, it may be that particular bacterial functions are associated with OSCC regardless of what species are contributing to them. In fact, a "core" functional bacteriome in the absence of a compositional one has been previously described for the gut²¹. So far, no attempt to perform functional bacteriome analysis has been made with respect to oral cancer.

The advent of next generation sequencing (NGS) has revolutionized the study of microbial communities. The 16S rRNA gene is typically targeted, enabling profiling of large number of samples at significant depth²² and thus detection of species which have very low abundance. Three studies have so far employed 16S rRNA-based NGS for profiling the bacteriome associated with OSCC^{11, 17, 19}. However, in addition to methodological variations that hinder direct comparison of the results, these studies have used the typical bioinformatic analysis pipeline that involves *de novo* clustering of sequences into operational taxonomic units (OTUs), then using a Bayesian classifier to taxonomically assign them. This approach limits classification of the majority of sequences to the genus level²³, rendering any associations identified of little biological significance, since specific species or even strains are usually involved in causing disease.

In a recent report, we have described a robust BLASTN-based algorithm that uses three well-curated sets of reference 16S rRNA gene sequences for classification of NGS reads to the species level, and pilot-tested it on 3 OSSC samples¹⁰. In the current study, we use NGS coupled with a modified version of the algorithm to profile the bacteriome within OSSC tissues in a full-scale study. In addition, we perform imputed functional analysis to predict bacterial genes and metabolic pathways associated with OSCC.

Methods

OSCC and control DNA samples. For cases, twenty samples were selected from an archive of anonymized, leftover DNA extracts obtained from fresh OSCC biopsies in a previous study²⁴. The biopsies had been collected by Dr. Akram Nasher between June 2009 and February 2011 in 2 major hospitals in Sana'a City, Yemen as detailed in the original study, and the DNA extracts, prepared from approximately 25 mg of tissue dissected from the body of the tumors, had been stored at -80°C since then. The selection was done so as to ensure proportional representation by gender and affected site.

Twenty healthy, gender- and aged-matched controls were recruited at the Faculty of Dentistry, Jazan University, in the South of Saudi Arabia (70 kilometers across the border with Yemen) between December 2014 and March 2015. Subjects with history of antibiotic intake in the last three months or a disease/condition known to modify oral microbial composition such as pregnancy, intake of contraceptive pills and diabetes, were excluded. Deep epithelium samples were obtained from anatomical sites matching those affected by the OSCC lesions in the cases as follows: a clean Catch-All Sample Collection swab (Epicenter, USA) was used to lightly swab the site to be sampled to remove the surface cells and adherent bacteria and then discarded. A second swab was then used to obtain deep epithelial cells by stroking with pressure 10 times in one direction, turning the swab 180° and stroking 10 times in the opposite direction. Each swab was placed in a sterile, DNase/RNase-free tube and stored at -20°C .

DNA extraction from the swabs was performed using the DDK DNA isolation kit (Isohelx, UK) according to the manufacturer's instructions. The quantity and quality of DNA, obtained from both the cases and controls, were assessed using the NanoDrop 2000 (ThermoFisher Scientific, USA) and Qubit[®] 2.0 Fluorimeter (Life Technologies, USA).

The study was conducted in compliance with the ethical guidelines of the Declaration of Helsinki and was approved by the biomedical research ethics committee at King Fahd University, Jazan, Saudi Arabia and an informed written consent was obtained from each of the controls. The clinical features of the cases and controls are presented in Table 1.

Amplicon library preparation and sequencing. Library preparation and sequencing were done at the Australian Centre for Ecogenomics as described previously²⁵. Briefly, the degenerate primers 27FYM²⁶ and 519R²⁷ were used to amplify the V1-3 region of the 16S rRNA gene using standard PCR conditions. The resultant PCR amplicons (~520 bp) were then purified, indexed with unique 8-base barcodes in a 2nd PCR and pooled together in equimolar concentrations. Finally, sequencing of the indexed library was performed employing the v3 2 × 300 bp chemistry on a MiSeq platform (Illumina, USA) according to the manufacturer's protocol.

Preprocessing of sequencing data. The raw data were submitted to Sequence Reads Archive (SRA) under project no. PRJNA352375 and preprocessed as described previously²⁵. Briefly, reads with primer mismatches were removed before the primer sequences were trimmed off. The software PEAR²⁸ was then employed to stitch paired sequences using the following parameters: minimum amplicon length = 432 bp; maximum amplicon lengths = 522 bp; and P-value = 0.001. Finally, the mothur software package version 1.38.1²⁹ was used to process the merged reads as follows: reads with ambiguous bases, with homopolymers >8 bases long, that did not achieve a sliding 50-nucleotide Q-score average of ≥ 35 or that poorly aligned to SILVA reference alignment³⁰ were filtered out; the remaining reads were checked for chimeras with Uchime³¹ using the self-reference approach³².

Compositional data analysis. The high-quality, non-chimeric sequences were classified to the species-level employing a combination of two BLASTN-based algorithms recently described^{10, 25}. Briefly, reads were individually BLASTN-searched against 4 sets of 16S rRNA reference sequences prioritized in the following order: The Human Oral Microbiome Database (HOMD) version 14.5 (available from <http://homd.org/>); a chimera-free version of the Human Oral Microbiome extended database (trusted-HOMDext)¹⁰; a modified version of the Greengene Gold set (modified-GGG)¹⁰; and NCBI's Microbial 16S set (August 2016 release downloaded from <ftp://ftp.ncbi.nlm.nih.gov/blast/db>). Matching was done at both alignment coverage (BLASTN alignment length/read length) and percent identity (matches/alignment length) of $\geq 98\%$. Matches, if any, were first ranked by relevance (e.g. hits from HOMD 14.5 were ranked first) and then by % identity and bit score. Reads were then classified to the species level based on taxonomy of the top hit reference sequence (i.e. the sequence with the highest % identity and bit score belonging to the highest priority reference set). Reads returning top hits belonging to multiple species underwent secondary *de novo* chimera checking using USEARCH³³ at a % identity cutoff of 98% and, if proved to be non-chimeric, were assigned multiple-species taxonomies. Reads with no matches at the specified criteria were subjected to the *de novo* chimera checking as above, and then to species-level *de novo* operational taxonomy unit (OTU) calling at 98% identity cutoff using USEARCH. Singleton OTUs were excluded and a representative sequence for each of the remaining OTUs was BLASTN-searched against the 4 reference sets again to determine the closest species for taxonomy assignment.

The QIIME (Quantitative Insights Into Microbial Ecology) software package version 1.9.1³⁴ was used to perform downstream analysis including subsampling to obtain an equal number of reads across the samples, generation of taxonomy plots/tables and rarefaction curves and calculation of species richness, coverage and a range of alpha and beta diversity indices. Comparison between samples in bacterial community membership and structure was performed with Principal Component Analysis (PCoA) based on binary and weighted Jaccard distance matrices. Detection of differentially abundant taxa between the cases and controls was done using Linear discriminant analysis Effect Size (LEfSe)³⁵ and G-test.

Functional prediction analysis. The reads were reclassified with mothur using Wang's method and Greengenes 97% OTUs (version 13.5) as reference. The reads were then assigned to OTUs based on their taxonomy (phyloTYPE command) and the generated file was converted into a BIOM (Biological Observation Matrix) table. The latter was then used as an input to PICRUSt (phylogenetic investigation of communities by reconstruction of unobserved states)³⁶, a bioinformatics resource for prediction of functional content of microbial communities by matching OTUs in the samples to reference OTUs with known/imputed gene content, normalizing for gene-copy number variations. The analysis was performed based on KEGG orthologs (KO) and pathways. Differences in genes and pathways between the cases and controls were explored using LEfSe.

Results

Sequencing and data processing statistics. The sequencing run generated 5,037,910 raw paired reads. Around 20% of these were identified with primer mismatches and removed. The majority of the remaining reads (89.9%) could be successfully stitched with PEAR. However, only 21.8% of the merged reads withstood the stringent quality-filtration strategy, which we previously found to reduce sequencing error rates by 10 fold²⁵. An additional 180,109 reads were filtered out in subsequent alignment and chimera checks, leaving a final of 611, 225 high-quality, non-chimeric merged reads with an average length of 477 bp. Around 94% of these reads were successfully classified to the species level; 172 reads were identified *de novo* as additional chimeras; 184 did not return BLASTN matches and 36,573 formed singleton OTUs and were thus excluded. The number of classified reads per sample ranged from 7791 to 28152 reads (14357 ± 4499 reads per subject). The results described below were based on a minimum read count per species (MC) of one. Results for higher MC cutoffs (10 and 100) can be found at ftp://www.homd.org/publication_data/20170317/qiime_results/index.html.

Bacteriome profile. A total of 1,118 species-level taxa, including 416 potentially novel species, belonging to 259 genera and 13 phyla were identified in the samples. The abundances and detection frequencies of these in each of the samples and across the study groups are presented in Supplementary Datasets 1–3. The number of species detected in the cases and controls was 795 and 746, respectively, with 423 species in common. Per sample, the number of species ranged from 53 to 254 and 79 to 245 for the cases and controls, respectively (average of 140 and 144 species per sample, respectively). Figure 1 displays the distribution of the phyla, top 15 genera and top 25 species detected. Overall, phyla Fusobacteria, Proteobacteria, Firmicutes, Bacteroidetes, Actinobacteria and genera *Fusobacterium*, *Streptococcus*, *Leptotrichia*, *Haemophilus*, *Prevotella*, *Rothia*, *Capnocytophaga*, *Campylobacter*, *Porphyromonas* and *Neisseria* accounted for the bulk of the bacteriome. At the species level, *Streptococcus mitis*, *Rothia mucilaginosa*, *Fusobacterium nucleatum subsp. polymorphum*, *Fusobacterium periodonticum*, *Haemophilus parainfluenzae*, *Prevotella melaninogenica*, *Leptotrichia sp. oral taxon 225*, *Neisseria flavescens|subflava* were overall the most predominant species.

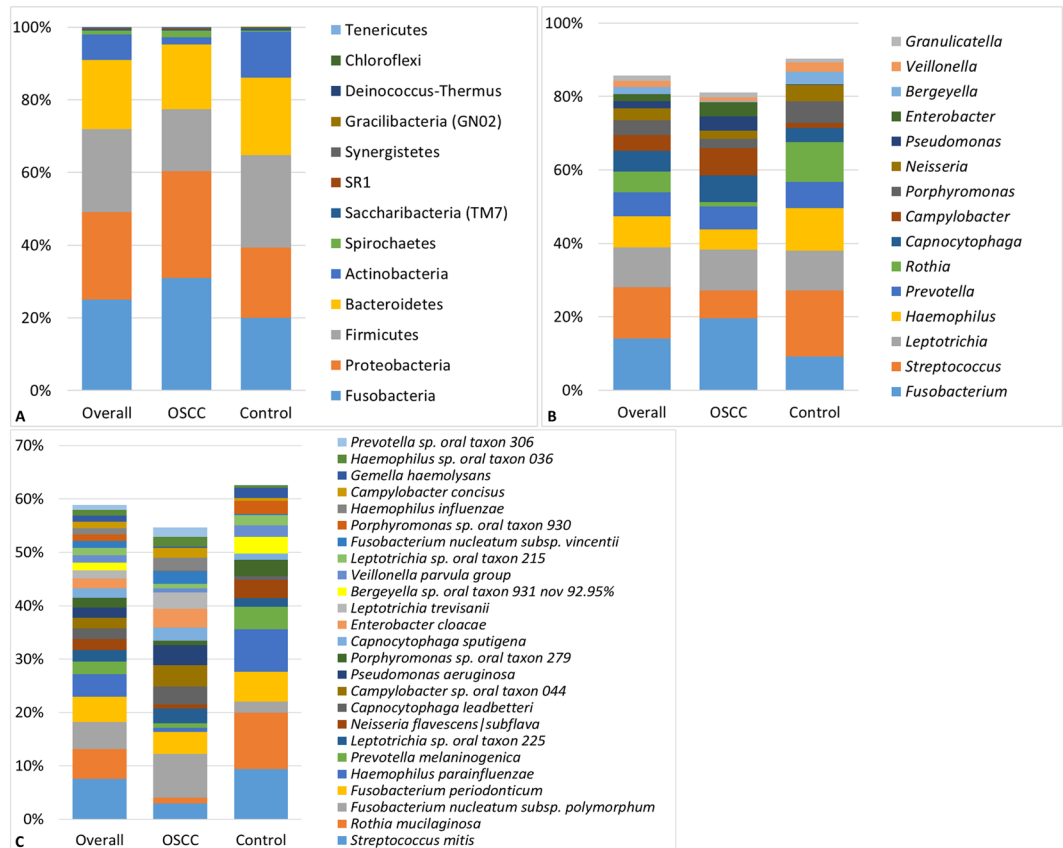


Figure 1. Bacteriome profile. Stacked bars showing the distribution of phyla, top 15 genera and top 25 species detected in the study population and groups.

Product type	Observed richness	Chao1	Shannon index	Good's coverage
OSCC	122.2 ± 49.9	145.7 ± 59.1	4.033 ± 0.939	0.997 ± 0.002
Control	128.2 ± 40.5	161.9 ± 47.6	3.876 ± 0.997	0.996 ± 0.001

Table 2. Species richness, α -diversity and coverage (mean ± SE) calculated from the rarefied biom.

The two study groups were comparable in terms of species richness, α -diversity and coverage (Table 2). Rarefaction indicated sufficient sequencing depth (Fig. 2). Figure 3 shows the separation between OSCC and control samples by PCoA based on community membership (mere presence/absence of species) and structure (taking into account their relative abundances), using binary and abundance weighted Jaccard indices, respectively. More variation could be explained based on community structure.

Differentially abundant taxa. The genera and species with significantly different abundance in the cases and controls are presented in Fig. 4. *Fusobacteria*, *Campylobacter* and *Pseudomonas* showed the strongest association with OSCC, while *Streptococcus*, *Rothia* and *Haemophilus* were the most overrepresented genera in the controls. Species-wise, *F. nucleatum subsp. polymorphum*, *Pseudomonas aeruginosa* and *Campylobacter sp.* Oral taxon 44 were the most significantly abundant in the tumors, while *S. mitis*, *R. mucilaginosus* and *H. parainfluenzae* were the most associated with the controls. The distribution of these 6 species in each of the study groups, overall and by sample collection site is presented in Supplementary Fig. 1. We elaborate here on *F. nucleatum subsp. polymorphum* and *P. aeruginosa*. The former accounted for more than 10% (and as high as 34%) of the reads in 7 (35%) of the OSCC samples but in only one control sample. Stratifying by sampling site, however, it maintained significant association only with tongue cancer. *P. aeruginosa* was identified in 70% of the tumors compared to only 15% in the controls; the abundance in the former reached 23% while it did not exceed 0.05% in the latter. After stratification, the association remained significant for cancer of the tongue and gum.

Species exclusively found in either groups at $\geq 15\%$ regardless of whether or not they were detected by LefSe as differentially abundant are listed in Supplementary Table 1. Among these were potentially pathogenic taxa e.g. *Haemophilus influenzae* (30%), *Staphylococcus aureus* (20%), *Bacteroides fragilis* (15%) and *Escherichia coli* (15%).

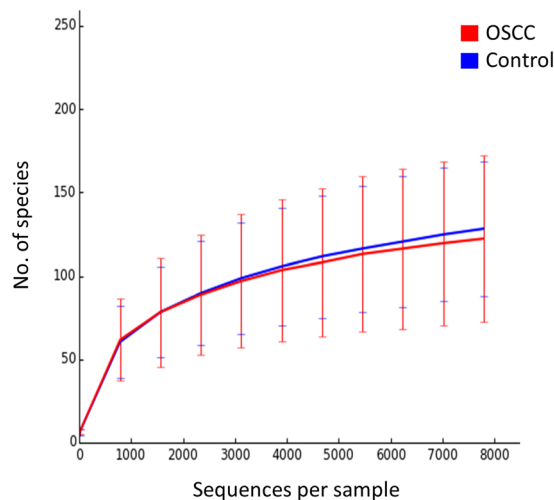


Figure 2. Rarefaction curves showing the number of observed species as a function of sequencing depth.

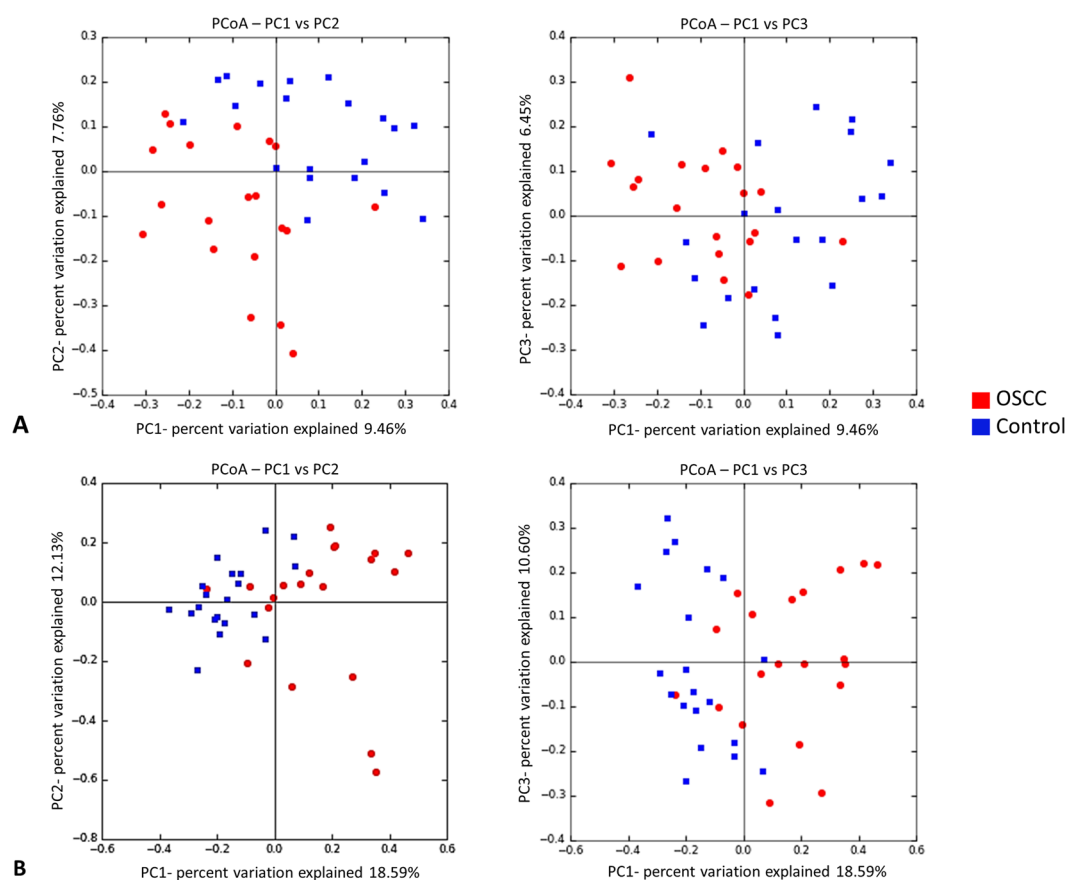


Figure 3. Principal Component Analysis. Clustering of the samples based on (A) binary Jaccard index (community membership) and (B) abundance weighted Jaccard index (community structure).

Differentially enriched genes and pathways. The microbial genes and pathways enriched in each of the study groups are shown in Fig. 5. At the gene level, genes encoding methyl accepting chemotaxis protein, restriction enzyme subunits and peptide nickel transport system permease and ATP binding proteins were enriched in the cases while those encoding antibiotic transport system permease and ATP binding proteins, 7,8-dihydro-8-oxoguanine-triphosphatase and ABC-2 type transport system permease and ATP binding proteins were the most overrepresented genes in the controls. At the pathway level, genes involved in bacterial mobility, flagellar assembly, bacterial chemotaxis and LPS synthesis were significantly more abundant in the tumor samples, while those involved in DNA repair and combination, purine metabolism, phenylalanine, tyrosine and tryptophan

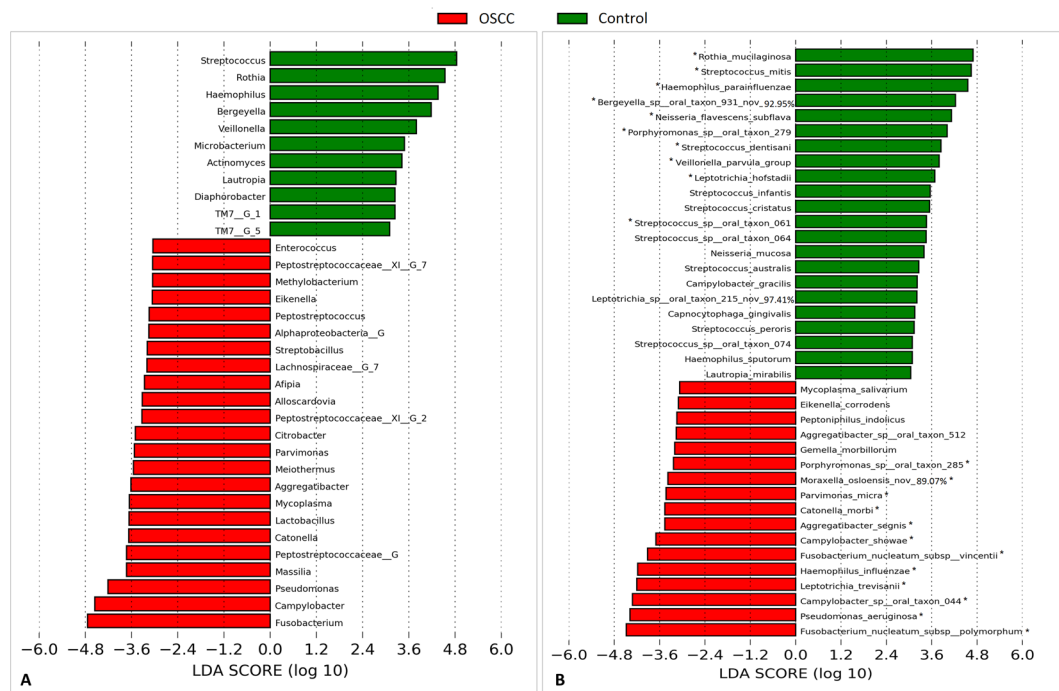


Figure 4. Differentially abundant taxa. Linear Discriminant Analysis Effect Size (LEfSe) analysis showing genera (A) and species (B) that were significantly differentially abundant between the cases and controls (LDA score ≥ 3). *The difference is also significant by G-test (False discovery rate = 0).

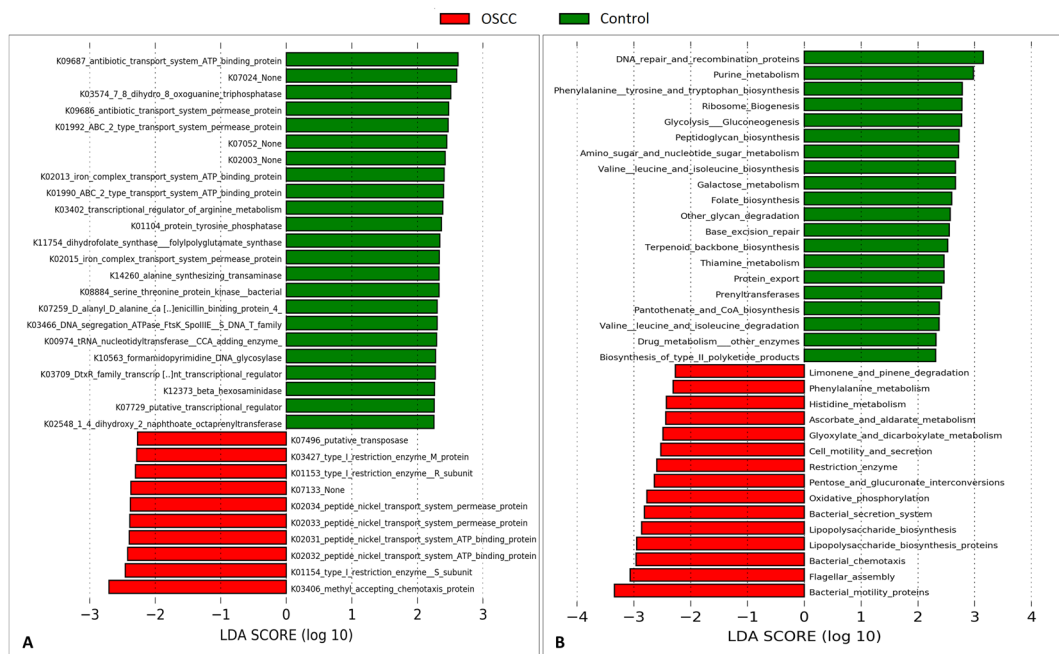


Figure 5. Differentially enriched functions. Linear Discriminant Analysis Effect Size (LEfSe) analysis showing genes (A) and pathways (B) that were significantly differentially enriched between the cases and controls (LDA score ≥ 2.25).

biosynthesis, ribosome biogenesis and glycolysis/gluconeogenesis were the most significantly associated with the controls.

Discussion

This is the first full-scale study to employ NGS for characterization of bacteria within OSCC tissues to the species level. It is also the first report on the functional potential of the bacteriome associated with OSCC. To maximize reliability of comparisons, the cases and controls were matched for gender, age and sampling site. Theoretically, tissue biopsies from the healthy subjects would have served the best control samples, but that was not possible due to ethical concerns. Instead, epithelium swabs were obtained, which may be viewed as an inevitable study limitation. Unlike previous studies, however, deep swabbing was performed so as to recover within-tissue rather than surface bacteria. Another limitation is that, due to the circumstances of the study, the controls had to be recruited in a setting and geographical location different from that of the cases. However, Jazan is located just across the Northern border of Yemen; in fact, this was a disputed region between Yemen and Saudi Arabia until 2001. In any case, there are no ethnic differences between the two countries. In addition, Jazan is culturally similar to Yemen. Of particular importance and relevance to the study is that OSCC is highly prevalent in both regions^{37,38} with the major risk factor in common: *shammah* use^{24,39,40}. Therefore, the difference in frequency of *shammah* use between the cases and controls is a reflection of the fact that *shammah* use is associated with OSCC, not because the controls were recruited in a different geographical location. However, the possibility remains that *shammah* use accounted, at least in part, for the differences in the bacteriome observed between the cases and controls. If so, it may be one mechanism by which *shammah* use contributes to oral carcinogenesis.

We exploited Illumina's 2 × 300 bp sequencing chemistry coupled with stringent read stitching and quality-filtering algorithms to generate high quality, full length V1-V3 reads (472–562 bp) and thus maximize the resolution and accuracy of species-level taxonomic assignment obtained with the prioritized BLASTN-based classification pipeline used. The advantages of using this classification algorithm over *de novo* OTU calling and the rationale of prioritizing the reference 16S rRNA sequence databases used have been discussed in previous reports^{10,41}. One study limitation is that only predictive functional analysis was performed using PICRUSt. Although PICRUSt has been demonstrated to produce accurate functional predictions when compared to whole metagenome sequencing³⁶, the latter remains the gold standard to confidently characterize the functional attributes of microbial communities.

The OSCC and control samples had similar species richness and α -diversity, which is consistent with previous reports in which tissue biopsies or swabs have been analyzed^{16,19}. In contrast, saliva samples obtained from OSCC subjects have been demonstrated in two studies to have significantly lower species richness and α -diversity than those obtained from control subjects^{11,17}, suggesting that salivary bacterial diversity may be used as a marker of OSCC risk. The average number of species per sample detected in this study is a bit higher than that found in our previous pilot study (142 vs. 118), obviously because of the higher sequencing depth here, but remains much less – and thus realistic – compared to the numbers reported in studies employing *de novo* OTU calling which is known to significantly inflate species richness⁴¹.

Many taxa were found to be differentially abundant between the cases and controls as identified by LEfSe and G-test. *Fusobacterium* was the most significantly abundant genus in the OSCC samples. Consistently, Nagy *et al.*¹⁵ and Schmidt *et al.*¹⁹ identified *Fusobacterium* at significantly higher levels in swabs of OSCC lesion surface compared to those of normal mucosa from the same patients. At the species level, however, the current study provides the first epidemiological evidence ever for association of *F. nucleatum* with OSCC, substantiating existing evidence on its carcinogenicity. *F. nucleatum* has been associated with colorectal carcinoma (CRC)^{42,43} and demonstrated to promote cellular proliferation and invasion in human epithelium and CRC cell lines^{44,45} and to enhance progression of OSCC and CRC in animal models^{46,47}. In this study, the association is specifically shown for *F. nucleatum subsp. polymorphum* and *F. nucleatum subsp. vincentii*, suggesting there may be differences in the carcinogenicity of this species at the subspecies level, a possibility never explored before.

For the first time, we here report an association between *P. aeruginosa* and OSCC. This species has not been linked in the literature to any cancer type. However, there is some recent evidence from *in vitro* studies to suggest a role in carcinogenesis⁴⁸. For example, *P. aeruginosa* has been demonstrated to trigger DNA breaks in epithelial cells⁴⁹, which could result in chromosomal instability. *P. aeruginosa* possesses structures, e.g. lipopolysaccharides (LPS) and flagella, and cytotoxins (e.g. ExoU) with potent proinflammatory activity that results in recruitment of neutrophils via activation of NF- κ B signaling pathway^{50,51}. This is relevant because inflammation is accepted to play an important role in carcinogenesis. Furthermore, *P. aeruginosa* secretes factor LasI that disrupts adherens junctions and reduces expression of E-cadherin, a molecule known to serve antagonistic function against cellular invasion and metastasis⁴⁸. Whether *P. aeruginosa* plays a role in initiation or/and progression of OSCC thus warrants further investigation.

Streptococcus and *Rothia* were the most significantly associated genera with the controls, which is consistent with findings from the study by Schmidt *et al.*¹⁹ in which surface swabs were analyzed. In contradiction, Pushalkar *et al.*¹⁷ and Guerrero-Preston *et al.*¹¹ found these genera to be more abundant in the saliva samples of OSCC. This, along with the differences in species richness and diversity for tissue biopsies and saliva described above, suggests that bacterial associations with OSCC dramatically differ by, and should thus be differently interpreted based on, the type of sample analyzed. In line with this, *S. mitis* was found here as well as in the study by Pushalkar *et al.*¹⁶ to be overrepresented in the control samples, while it was shown by Mager *et al.*¹³ to be more abundant in saliva samples from OSCC patients. *R. mucilaginosa* and *H. parainfluenzae* were among the top taxa showing association with health in this study. Consistently, Pushalkar *et al.* detected *R. mucilaginosa* much more frequently in their non-tumor samples. In addition, both species have been recently reported as members of the healthy core oral bacteriome⁴¹.

The bacteriome predicted functions found to be enriched in the OSCC samples in this study are strikingly similar to those identified very recently in association with chronic periodontitis⁵², emphasizing they are proinflammatory in nature. Indeed, bacterial flagella and LPS are potent inflammatory structures. The latter in particular has been found to induce cancer-promoting inflammatory reactions. For example, LPS has been demonstrated

to promote invasiveness of pancreatic cancer by activation of the TLR/MyD88/NF- κ B pathway⁵³, to facilitate lung metastasis in a breast cancer via the prostaglandin E2-EP2 pathway⁵⁴ and to increase liver metastasis of human CRC by stimulation of toll receptor TLR4⁵⁵. Flagella associated with *P. aeruginosa* are known to induce inflammation by activation of the NF- κ B⁴⁸; although, there is no evidence linking this to carcinogenesis directly, the possibility cannot be excluded. Bacterial chemotaxis also seems to play an important role in cancer-related inflammation. Studies on *H. pylori*, for example, show that mutants defective in chemotaxis induce less inflammation than the wild type⁵⁶. Overall, therefore, the bacteriome associated with OSCC can functionally be described as “inflammatory” which is a very important finding given the established role of inflammation in cancer. However, whole metagenome sequencing in a more extensive and independent functional study is required to confirm and explore these findings further.

In conclusion, a distinct bacteriome, compositionally and functionally, is associated with OSCC in these Yemeni patients. This study provides the first epidemiological evidence for association of *F. nucleatum* and *P. aeruginosa* with OSCC. It also suggests there may be some variation in carcinogenicity of *F. nucleatum* subspecies. At the functional level, the bacteriome enriched in OSCC can be described as “inflammatory”. Exploring the role of differentially abundant taxa and pathways identified in the development and/or progression of OSCC is warranted.

References

1. Ferlay, J. *et al.* GLOBOCAN 2012 v1.0, Cancer Incidence and Mortality Worldwide: IARC CancerBase No. 11 [Internet] <http://globocon.iarc.fr> (2013).
2. Wang, B., Zhang, S., Yue, K. & Wang, X. D. The recurrence and survival of oral squamous cell carcinoma: a report of 275 cases. *Chin J Cancer* **32**, 614–618, doi:10.5732/cjc.012.10219 (2013).
3. Sklenicka, S., Gardiner, S., Dierks, E. J., Potter, B. E. & Bell, R. B. Survival analysis and risk factors for recurrence in oral squamous cell carcinoma: does surgical salvage affect outcome? *J Oral Maxillofac Surg* **68**, 1270–1275, doi:10.1016/j.joms.2009.11.016 (2010).
4. Petti, S. Lifestyle risk factors for oral cancer. *Oral Oncol* **45**, 340–350, doi:10.1016/j.oraloncology.2008.05.018 (2009).
5. Gupta, B., Johnson, N. W. & Kumar, N. Global Epidemiology of Head and Neck Cancers: A Continuing Challenge. *Oncology* **91**, 13–23, doi:10.1159/000446117 (2016).
6. Chocolatewala, N., Chaturvedi, P. & Desale, R. The role of bacteria in oral cancer. *Indian J Med Paediatr Oncol* **31**, 126–131, doi:10.4103/0971-5851.76195 (2010).
7. Perera, M., Al-Hebshi, N. N., Speicher, D. J., Perera, I. & Johnson, N. W. Emerging role of bacteria in oral carcinogenesis: a review with special reference to perio-pathogenic bacteria. *J Oral Microbiol* **8**, 32762, doi:10.3402/jom.v8.32762 (2016).
8. Hooper, S. J. *et al.* Viable bacteria present within oral squamous cell carcinoma tissue. *J Clin Microbiol* **44**, 1719–1725, doi:10.1128/JCM.44.5.1719-1725.2006 (2006).
9. Hooper, S. J. *et al.* A molecular analysis of the bacteria present within oral squamous cell carcinoma. *J Med Microbiol* **56**, 1651–1659, doi:10.1099/jmm.0.46918-0 (2007).
10. Al-Hebshi, N. N., Nasher, A. T., Idris, A. M. & Chen, T. Robust species taxonomy assignment algorithm for 16S rRNA NGS reads: application to oral carcinoma samples. *J Oral Microbiol* **7**, 28934, doi:10.3402/jom.v7.28934 (2015).
11. Guerrero-Preston, R. *et al.* 16S rRNA amplicon sequencing identifies microbiota associated with oral cancer, Human Papilloma Virus infection and surgical treatment. *Oncotarget* (2016).
12. Katz, J., Onate, M. D., Pauley, K. M., Bhattacharyya, I. & Cha, S. Presence of Porphyromonas gingivalis in gingival squamous cell carcinoma. *Int J Oral Sci* **3**, 209–215, doi:10.4248/IJOS11075 (2011).
13. Mager, D. L. *et al.* The salivary microbiota as a diagnostic indicator of oral cancer: A descriptive, non-randomized study of cancer-free and oral squamous cell carcinoma subjects. *J Transl Med* **3**, 27, doi:10.1186/1479-5876-3-27 (2005).
14. Morita, E. *et al.* Different frequencies of Streptococcus anginosus infection in oral cancer and esophageal cancer. *Cancer Sci* **94**, 492–496, doi:10.1111/cas.2003.94.issue-6 (2003).
15. Nagy, K. N., Sonkodi, I., Szoke, L., Nagy, E. & Newman, H. N. The microflora associated with human oral carcinomas. *Oral Oncol* **34**, 304–308, doi:10.1016/S1368-8375(98)80012-2 (1998).
16. Pushalkar, S. *et al.* Comparison of oral microbiota in tumor and non-tumor tissues of patients with oral squamous cell carcinoma. *BMC Microbiol* **12**, 144, doi:10.1186/1471-2180-12-144 (2012).
17. Pushalkar, S. *et al.* Microbial diversity in saliva of oral squamous cell carcinoma. *FEMS Immunol Med Microbiol* **61**, 269–277, doi:10.1111/j.1574-695X.2010.00773.x (2011).
18. Sasaki, M. *et al.* Streptococcus anginosus infection in oral cancer and its infection route. *Oral Dis* **11**, 151–156, doi:10.1111/j.1601-0825.2005.01051.x (2005).
19. Schmidt, B. L. *et al.* Changes in abundance of oral microbiota associated with oral cancer. *PLoS One* **9**, e98741, doi:10.1371/journal.pone.0098741 (2014).
20. Tateda, M. *et al.* Streptococcus anginosus in head and neck squamous cell carcinoma: implication in carcinogenesis. *Int J Mol Med* **6**, 699–703, doi:10.3892/ijmm (2000).
21. Turnbaugh, P. J. *et al.* A core gut microbiome in obese and lean twins. *Nature* **457**, 480–484, doi:10.1038/nature07540 (2009).
22. Siqueira, J. F. Jr., Fouad, A. F. & Rocas, I. N. Pyrosequencing as a tool for better understanding of human microbiomes. *J Oral Microbiol* **4** (2012).
23. Schloss, P. D. & Westcott, S. L. Assessing and improving methods used in operational taxonomic unit-based approaches for 16S rRNA gene sequence analysis. *Appl Environ Microbiol* **77**, 3219–3226, doi:10.1128/AEM.02810-10 (2011).
24. Nasher, A. T., Al-Hebshi, N. N., Al-Moayad, E. E. & Suleiman, A. M. Viral infection and oral habits as risk factors for oral squamous cell carcinoma in Yemen: a case-control study. *Oral Surg Oral Med Oral Pathol Oral Radiol* **118**, 566–572 e561 (2014).
25. Al-Hebshi, N. N., Alharbi, F. A., Mahri, M. & Chen, T. Differences in the bacteriome of smokeless tobacco products with different oral carcinogenicity: compositional and predicted functional analysis. *Genes* **8**, 106, doi:10.3390/genes8040106 (2017).
26. Frank, J. A. *et al.* Critical evaluation of two primers commonly used for amplification of bacterial 16S rRNA genes. *Appl Environ Microbiol* **74**, 2461–2470, doi:10.1128/AEM.02272-07 (2008).
27. Lane, D. J. *et al.* Rapid determination of 16S ribosomal RNA sequences for phylogenetic analyses. *Proc Natl Acad Sci USA* **82**, 6955–6959, doi:10.1073/pnas.82.20.6955 (1985).
28. Zhang, J., Kobert, K., Flouri, T. & Stamatakis, A. PEAR: a fast and accurate Illumina Paired-End reAd mergeR. *Bioinformatics* **30**, 614–620, doi:10.1093/bioinformatics/btt593 (2014).
29. Schloss, P. D. *et al.* Introducing mothur: open-source, platform-independent, community-supported software for describing and comparing microbial communities. *Appl Environ Microbiol* **75**, 7537–7541, doi:10.1128/AEM.01541-09 (2009).

30. Pruesse, E. *et al.* SILVA: a comprehensive online resource for quality checked and aligned ribosomal RNA sequence data compatible with ARB. *Nucleic Acids Res* **35**, 7188–7196, doi:10.1093/nar/gkm864 (2007).
31. Edgar, R. C., Haas, B. J., Clemente, J. C., Quince, C. & Knight, R. UCHIME improves sensitivity and speed of chimera detection. *Bioinformatics* **27**, 2194–2200, doi:10.1093/bioinformatics/btr381 (2011).
32. Schloss, P. D., Gevers, D. & Westcott, S. L. Reducing the effects of PCR amplification and sequencing artifacts on 16S rRNA-based studies. *PLoS One* **6**, e27310, doi:10.1371/journal.pone.0027310 (2011).
33. Edgar, R. C. Search and clustering orders of magnitude faster than BLAST. *Bioinformatics* **26**, 2460–2461, doi:10.1093/bioinformatics/btq461 (2010).
34. Caporaso, J. G. *et al.* QIIME allows analysis of high-throughput community sequencing data. *Nat methods* **7**, 335–336, doi:10.1038/nmeth.f.303 (2010).
35. Segata, N. *et al.* Metagenomic biomarker discovery and explanation. *Genome Biol* **12**, R60, doi:10.1186/gb-2011-12-6-r60 (2011).
36. Langille, M. G. *et al.* Predictive functional profiling of microbial communities using 16S rRNA marker gene sequences. *Nat Biotechnol* **31**, 814–821, doi:10.1038/nbt.2676 (2013).
37. Sawair, F. A. *et al.* High relative frequency of oral squamous cell carcinoma in Yemen: qat and tobacco chewing as its aetiological background. *Int J Environ Health Res* **17**, 185–195, doi:10.1080/09603120701254813 (2007).
38. Brown, A., Ravichandran, K. & Warnakulasuriya, S. The unequal burden related to the risk of oral cancer in the different regions of the Kingdom of Saudi Arabia. *Community Dent Health* **23**, 101–106 (2006).
39. Salem, G., Juhl, R. & Schiodt, T. Oral malignant and premalignant changes in ‘Shamma’-users from the Gizan region, Saudi Arabia. *Acta Odontol Scand* **42**, 41–45, doi:10.3109/00016358409041130 (1984).
40. Allard, W. F., DeVol, E. B. & Te, O. B. Smokeless tobacco (shamma) and oral cancer in Saudi Arabia. *Community Dent Oral Epidemiol* **27**, 398–405, doi:10.1111/com.1999.27.issue-6 (1999).
41. Al-Hebshi, N. N., Abdulhaq, A., Albarrag, A., Basode, V. K. & Chen, T. Species-level core oral bacteriome identified by 16S rRNA pyrosequencing in a healthy young Arab population. *J Oral Microbiol* **8**, 31444, doi:10.3402/jom.v8.31444 (2016).
42. Kostic, A. D. *et al.* Genomic analysis identifies association of Fusobacterium with colorectal carcinoma. *Genome Res* **22**, 292–298, doi:10.1101/gr.126573.111 (2012).
43. Castellarin, M. *et al.* Fusobacterium nucleatum infection is prevalent in human colorectal carcinoma. *Genome Res* **22**, 299–306, doi:10.1101/gr.126516.111 (2012).
44. Rubinstein, M. R. *et al.* Fusobacterium nucleatum promotes colorectal carcinogenesis by modulating E-cadherin/ β -catenin signaling via its FadA adhesin. *Cell host & microbe* **14**, 195–206, doi:10.1016/j.chom.2013.07.012 (2013).
45. Uitto, V. J. *et al.* Fusobacterium nucleatum increases collagenase 3 production and migration of epithelial cells. *Infect Immun* **73**, 1171–1179, doi:10.1128/IAI.73.2.1171-1179.2005 (2005).
46. Kostic, A. D. *et al.* Fusobacterium nucleatum potentiates intestinal tumorigenesis and modulates the tumor-immune microenvironment. *Cell Host Microbe* **14**, 207–215, doi:10.1016/j.chom.2013.07.007 (2013).
47. Binder Gallimidi, A. *et al.* Periodontal pathogens Porphyromonas gingivalis and Fusobacterium nucleatum promote tumor progression in an oral-specific chemical carcinogenesis model. *Oncotarget* **6**, 22613–22623, doi:10.18632/oncotarget.4209 (2015).
48. Markou, P. & Apidianakis, Y. Pathogenesis of intestinal *Pseudomonas aeruginosa* infection in patients with cancer. *Front Cell Infect Microbiol* **3**, 115, doi:10.3389/fcimb.2013.00115 (2014).
49. Elsen, S., Collin-Faure, V., Gidrol, X. & Lemerrier, C. The opportunistic pathogen *Pseudomonas aeruginosa* activates the DNA double-strand break signaling and repair pathway in infected cells. *Cell Mol Life Sci* **70**, 4385–4397, doi:10.1007/s00018-013-1392-3 (2013).
50. Gellatly, S. L. & Hancock, R. E. *Pseudomonas aeruginosa*: new insights into pathogenesis and host defenses. *Pathog Dis* **67**, 159–173, doi:10.1111/2049-632X.12033 (2013).
51. de Lima, C. D. *et al.* ExoU activates NF- κ B and increases IL-8/KC secretion during *Pseudomonas aeruginosa* infection. *PLoS One* **7**, e41772, doi:10.1371/journal.pone.0041772 (2012).
52. Kirst, M. E. *et al.* Dysbiosis and alterations in predicted functions of the subgingival microbiome in chronic periodontitis. *Appl Environ Microbiol* **81**, 783–793, doi:10.1128/AEM.02712-14 (2015).
53. Ikebe, M. *et al.* Lipopolysaccharide (LPS) increases the invasive ability of pancreatic cancer cells through the TLR4/MyD88 signaling pathway. *J Surg Oncol* **100**, 725–731, doi:10.1002/jso.21392 (2009).
54. Li, S. *et al.* Lipopolysaccharide induces inflammation and facilitates lung metastasis in a breast cancer model via the prostaglandin E2-EP2 pathway. *Mol Med Rep* **11**, 4454–4462, doi:10.3892/mmr.2015.3258 (2015).
55. Hsu, R. Y. *et al.* LPS-induced TLR4 signaling in human colorectal cancer cells increases beta1 integrin-mediated cell adhesion and liver metastasis. *Cancer Res* **71**, 1989–1998, doi:10.1158/0008-5472.CAN-10-2833 (2011).
56. Williams, S. M. *et al.* Helicobacter pylori chemotaxis modulates inflammation and bacterium-gastric epithelium interactions in infected mice. *Infect Immun* **75**, 3747–3757, doi:10.1128/IAI.00082-07 (2007).

Acknowledgements

The study was funded by the Substance Abuse Research Center (SARC) at Jazan University, Saudi Arabia (grant no. 1010/2010).

Author Contributions

N.N.A. conceived the study, performed the preprocessing of the raw sequencing data, contributed to the development of the classification algorithm and wrote the first draft of the manuscript. A.M.I. and N.W.J. contributed to the study design and overall supervision of the research project. T.C. developed and ran the bioinformatic analysis pipeline. A.T.N. provided the OSCC DNA extracts and associated data. M.Y.M. recruited the control subjects and obtained samples from them. H.E.H. contributed to the laboratory work. All authors approved the final version of the manuscript.

Additional Information

Supplementary information accompanies this paper at doi:10.1038/s41598-017-02079-3

Competing Interests: The authors declare that they have no competing interests.

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.



Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2017