FINDING ORAL CANCER BIOMARKERS USING BIG METAGENOMIC DATA ANALYSIS AND PIPELINE DEVELOPMENT FOR EARLY STAGE DETECTION BY MICROBIOME PROFILING

Neha Shrivastava[1], Nimisha Das[2], Sreejit Kar[3], Divyan Fernando[4]

1. Department of Life Sciences, CHRIST (Deemed to be University), Bangalore.
   neha.shrivastav@science.christuniversity.in
2. Department of Life Sciences, CHRIST (Deemed to be University), Bangalore.
   nimisha.das@science.christuniversity.in
3. Department of Life Sciences, CHRIST (Deemed to be University), Bangalore.
   sreejit.kar@science.christuniversity.in
4. Department of Life Sciences, CHRIST (Deemed to be University), Bangalore.
   divyan.fernando@science.christuniversity.in

Mentors:

1. **Roli Budhwar**
   Head – R&D, Bionivid Technology (P) Limited, Kasturi Nagar, Bangalore
   roli@bionivid.com
2. **Umme Salma M**
   Assistant Professor, Department of Computer Sciences, CHRIST (Deemed to be University), Bangalore.
   ummesalma.m@christuniversity.in
3. **Alok Malaviya***
   Associate Professor, Applied and Industrial Biotechnology Laboratory, Department of Life Sciences, CHRIST (Deemed to be University), Bangalore.
   Alokkumar.malaviya@christuniversity.in

ABSTRACT

Oral cancer is a disease that arises from both host genetics and environmental factors. Tobacco and alcohol consumption, betel quid chewing, and human papillomavirus infection are well-known risk factors. The incidence of oral cancer is increasing, and this disease continues to be a major global health problem. Oral cancer is one of the predominant cancer types in India and approximately 15% of the oral cancer are not associated with the abovementioned risk factors. Our oral cavity consists of plethora of microflora and are part of the tumour microenvironment. Past and ongoing studies reveals that there are various microbes and changes in different bacteria have been associated with several types of cancer.

The **objective of this study is** *to explore the Cancer-associated changes in the oral microbiome using the metagenome based biological big data available on public domain, using computational biology approaches*. To unravel the micro-biotic connections underlying oral squamous cell carcinoma (OSCC), cancer lesion samples and anatomically matched normal next generation sequenced samples will be obtained from the public domain (NCBI-SRA). We will be using Meta-genomic pipeline to comprehensively investigate bacterial community composition, abundance and identify associated bacterial biomarkers which will be helpful in better prognosis and diagnosis of the disease.

Key words: Biomarker, Oral Cancer, Big Data, Metagenome, Pipeline

INTRODUCTION

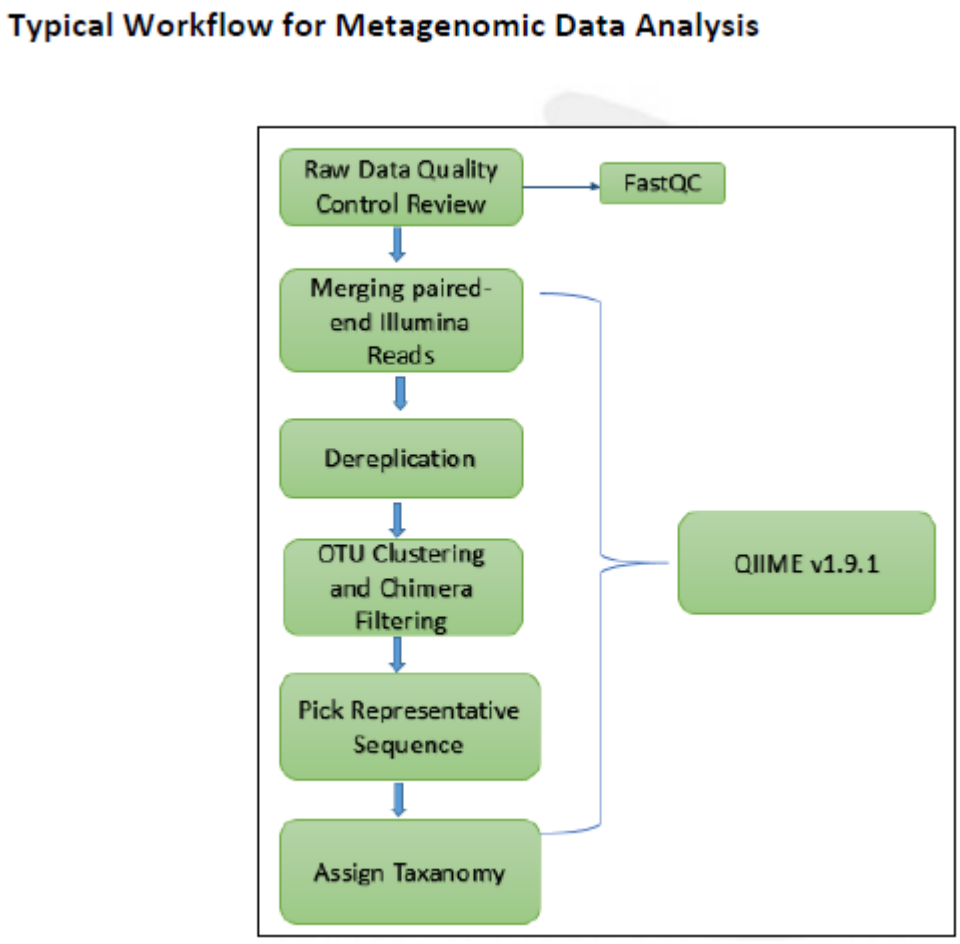
The human body is inhabited by over 100 trillion microbial cells living in symbiosis with their host [1]. Bacteria at certain body sites have long been believed to be involved in immune modulation, disease development, and health maintenance. The term microbiome was coined to describe *the collective genomes and gene products of all microbes residing within an organism* [2]. With the advent of high-throughput, next-generation sequencing (NGS), there has been a surge of interest in studying the human microbiome in the context of disease. Recent studies have demonstrated the importance of the gut microbiota in digestion, fat storage, angiogenesis, immune system development and responses, resistance to colonization, epithelial architecture [3,4], and dysbiosis, which is believed to contribute to the pathogenesis of local and systemic diseases, including inflammatory bowel disease, diabetes, and colorectal cancer [5]. Located at the beginning of the aerodigestive tract, approximately 700 prokaryote species have been detected in the human oral cavity. These species belong to 185 genera and 12 phyla, of which approximately 54% are officially named, 14% unnamed (but cultivated) and 32% known only as uncultivated phylotypes [6]. This oral bacterial flora plays an essential role in maintaining a normal oral physiological environment and is associated with host health [7]. In contrast to traditional views, recent analyses suggest the involvement of a consortium of microbes, rather than a single species, as causing disease [8], a phenomenon that has been well characterized for periodontal diseases [9].

Oral cancer, primarily oral squamous cell carcinoma (OSCC) deriving from the oral mucosa, is a disease that arises from both host genetic composition as well as environmental factors. Tobacco and alcohol consumption, betel quid chewing, and human papillomavirus infection are well-known risk factors [10]. The incidence of oral cancer is increasing, and this disease continues to be a major global health problem. Furthermore, approximately 15% of oral cancer cases cannot be attributed to the aforementioned major risk factors, resulting in the need to explore other potential risk factors [11]. A plethora of bacteria, the proverbial bacterial biofilm, coat each surface of the oral cavity [12], and groups inhabiting the mucosal surface might constitute the bulk of the tumor microenvironment. To date, various microbes and changes in different bacteria have been associated with several types of cancer [13]. Cancer-associated changes in the oral microbiome have been assessed in several early studies employing culture-based or molecular techniques [14-19] but a consensus has not been reached due to the limited number of strains/clones that it is feasible to test. However, the emergence of NGS allows microbial communities to be profiled at an unprecedented depth and coverage.

MATERIAL AND METHODS

Tumor samples and salivary rinses from HNSCC patients and normal controls were obtained from public domain database NCBI-SRA

https://www.ncbi.nlm.nih.gov/bioproject/?term=PRJNA321193



Bioinformatics Analysis:

Bioinformatics preprocessing steps will include quality filtering, error-correction, and chimera removal. Briefly, reads were de-multiplexed using 5' barcodes, trimmed of forward and reverse primer sequences, filtered for length and quality, and corrected for homopolymer errors. High quality reads were selected for analysis and reads with unknown bases ("N") were discarded. The resulting high-quality dataset was then screened for chimeric sequences and contaminant chloroplast DNA. Passing sequences will be characterized for diversity and taxonomic composition using the Quantitative Insights into Microbial Ecology (QIIME) suite, where all the beta and alpha diversity measure and significance tests were performed [20]. To begin, sequences were clustered into operational taxonomic units (OTUs) using UCLUST with a 97% identity threshold. Taxonomic assignment will be performed using the Ribosomal Database Project (RDP) classifier (trained by a customized version of the comprehensive Green Genes database, release v.13-05) with a minimum confidence threshold of 0.80. After considering the raw count data in full above, subsample analysis of each community will be performed to an equivalent depth. All results are based on the subsampled data, which mitigates biases due to differences in sampling depth. Additionally, the taxonomy of taxa at the genus-level for samples will be represented through pie charts.

Table 1 Selected Sample for analysis

Selected Samples for Analysis – ____ samples

Sr. No	Sample name	Total Reads

Figure 1 Example: Pie-Chart:

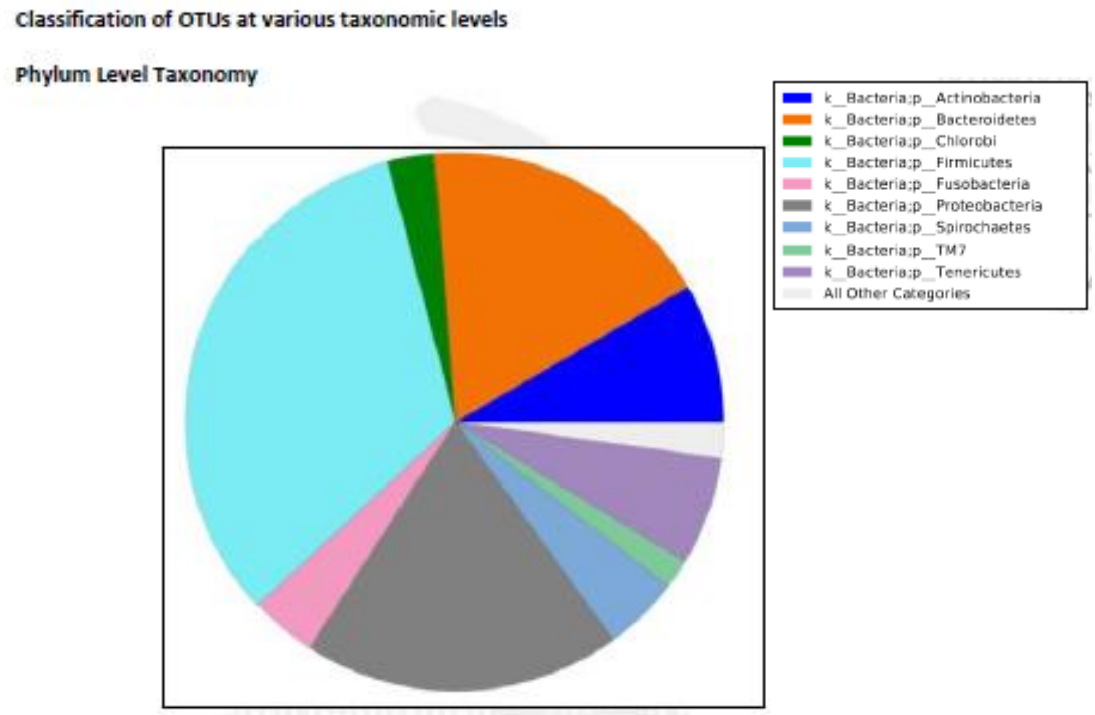
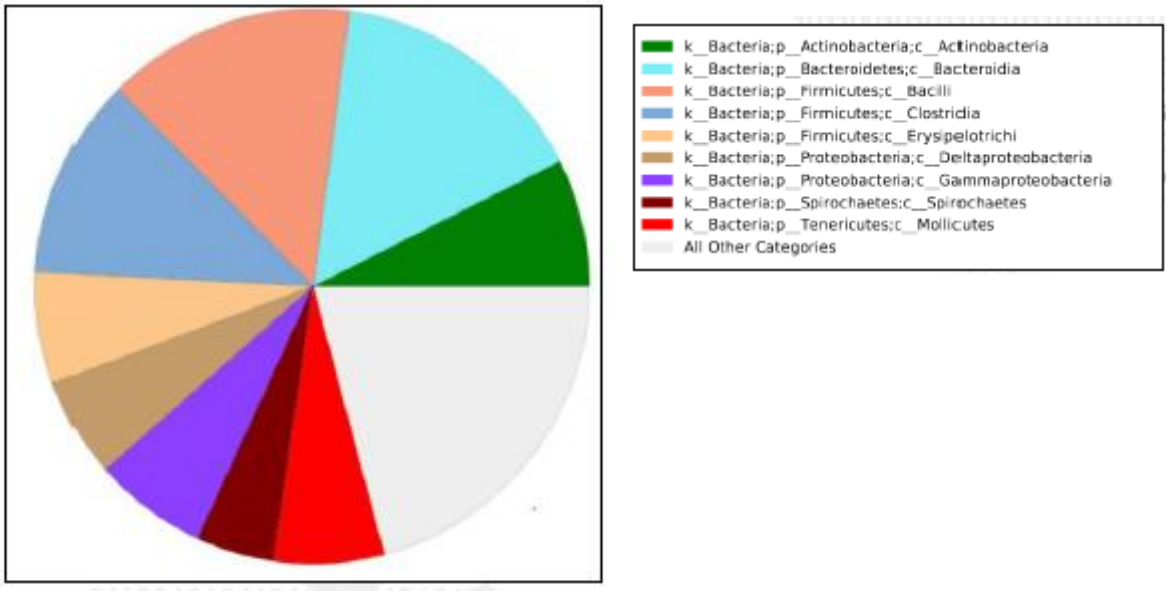


Figure 2 Example: Pie-Chart:



Example Alpha Diversity Table: Alpha Diversity was calculated for individual samples using QIIME.
Table 2: Alpha Diversity Results

	chao1	shannon	simpson	observed_otus
Sample1				
Sample2				
Sample3				
Sample4				
Sample5				
Sample6				
Sample7				
Sample8				
Sample9				
Sample10				

Example Rarefaction Curve:

Rarefaction Curve:-

Rarefaction allows the calculation of species richness for a given number of individual samples.

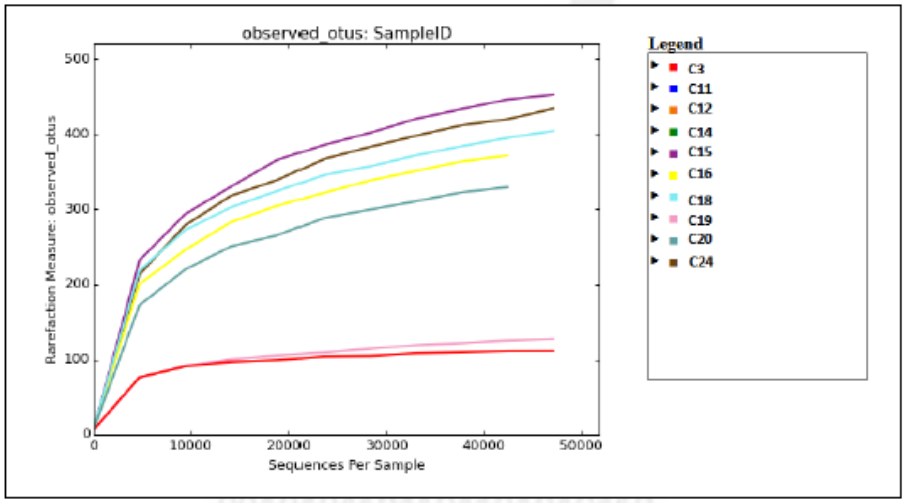


Figure 3 Refraction Curve

Example Table of Beta Diversity using QIIME.
Table 2: Beta Diversity Results

	Sample1	Sample2	Sample3	Sample4
Sample1				
Sample2				
Sample3				
Sample4				

Statistical analysis:

For each group comparison, significance tests will be computed including the maximum likelihood statistical significance tests that determine whether OTU presence/absence is associated with a category in the metadata. The goodness of fit or log-likelihood ratio parametric test (G-test) compares the ratio of the observed OTU frequencies in the sample groups to the expected frequencies based on the null hypothesis (all sample groups have equal OTU frequencies). QIIME [20] will be used to create all the heatmaps and estimate the following Alpha-diversity metrics: raw number of OTUs per sample, Chao1 estimator, Shannon entropy, Non-Metric dimensional scaling, and Bray-Curtis distance metric. The chao-1 index approach for richness will be used because it uses the numbers of singletons (OTUs with single appearance) and doubletons (OTUs that appeared twice) to estimate the number of missing species because missing species information is mostly concentrated on low frequency counts. Faith's phylogenetic diversity index (PD) estimates the relative feature diversity of any nominated set of species by the sum of the lengths of all phylogenetic branches required to span a given set of taxa on the phylogenetic tree. We will use linear discriminant analysis (LDA) with LefSe [21] an algorithm biomarker discovery that identifies taxa characterizing the differences between two metadata classes. It emphasizes statistical significance, biological consistency and effect relevance, allowing researchers to identify differentially abundant features that are also consistent with biologically meaningful categories (metadata), using non-parametric factorial Kruskal-Wallis (KW) sum-rank test, Wilcoxon rank-sum test and LDA. High LDA scores reflect significantly higher abundance of certain taxa.

Example PCoA PLOT generated by QIIME:

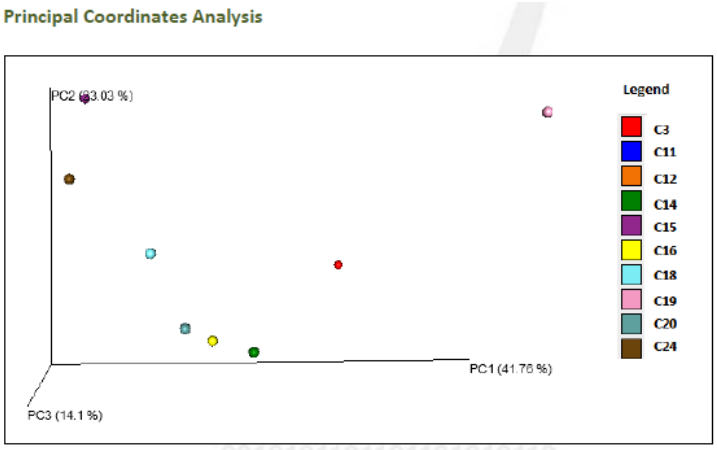


Figure 4: Example PCoA PLOT generated by QIIME

Example Heat map (Unsupervised hierarchical clustering)

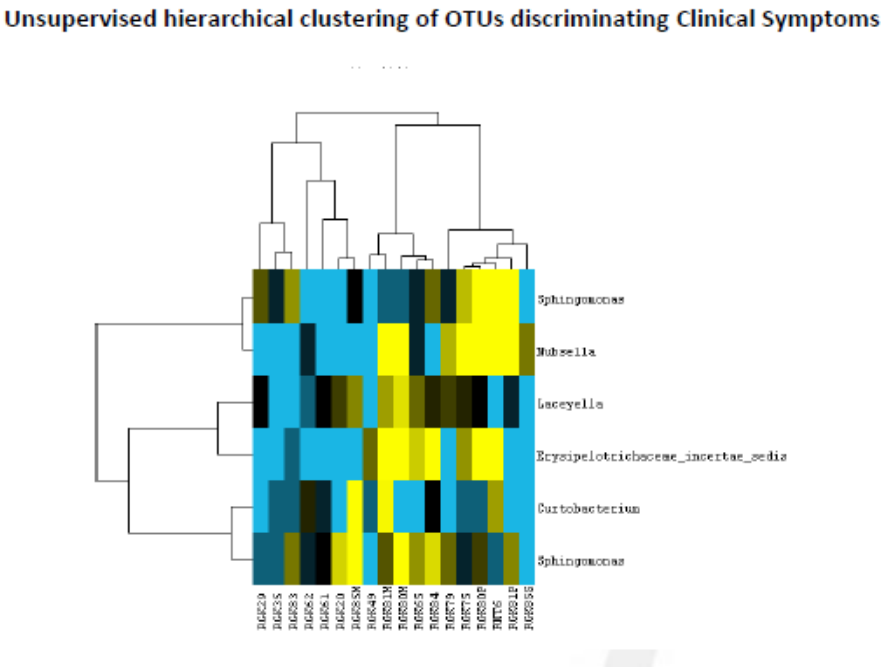


Figure 5 Heat Map

Example: Image of Lefse Analysis:

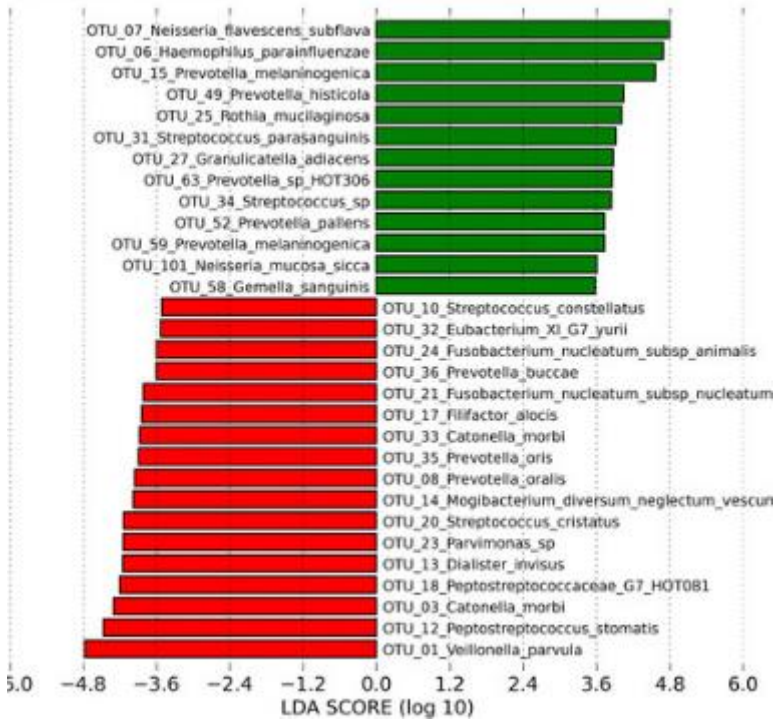


Figure 6: LDA Score

CONCLUSION

A biomarker or biological marker, according to the most recent definition [22], is a substance that is objectively measured and evaluated as an indicator of normal biologic processes, pathogenic processes, or pharmacologic responses to a therapeutic intervention. Because saliva are fluids easily collected and contain locally and systemically derived markers of periodontal disease, they may offer the basis for a patient-specific biomarker assessment for HNSCC as well. Due to the noninvasive and simple nature of their collection, analysis of saliva may be especially beneficial [23]. Our study will help us identify such biomarkers which will be helpful in the determination of oral cancer status and a means of monitoring response to treatment.

REFERENCES

1) Kudo, Y. et al. Oral environment and cancer. Genes Environ 38, 13 (2016).
2) Turnbaugh, P. J. et al. The human microbiome project: exploring the microbial part of ourselves in a changing world. Nature 449, 804–810 (2007).
3) Flint, H. J., Duncan, S. H., Scott, K. P. & Louis, P. Interactions and competition within the microbial community of the human colon: links between diet and health. Environ Microbiol 9, 1101–1111 (2007).
4) Burcelin, R. Gut microbiota and immune crosstalk in metabolic disease. Mol Metab 5, 771–781 (2016).
5) Allali, I. et al. Gut microbiome compositional and functional differences between tumor and non-tumor adjacent tissues from cohorts from the US and Spain. Gut Microbes 6, 161–172 (2015).
6) Dewhirst, F. E. et al. The human oral microbiome. J Bacteriol 192, 5002–5017 (2010).
7) Gholizadeh, P. et al. Role of oral microbiome on oral cancers, a review. Biomed Pharmacother 84, 552–558 (2016).
8) Darveau, R. P. Periodontitis: a polymicrobial disruption of host homeostasis. Nat Rev Microbiol 8, 481–490 (2010).
9) Chen, H. et al. A Filifactor alocis-centered co-occurrence group associates with periodontitis across different oral habitats. Sci Rep 5, 9053 (2015).
10) Ng, J. H., Iyer, N. G., Tan, M. H. & Edgren, G. Changing epidemiology of oral squamous cell carcinoma of the tongue: A global study. Head Neck 39, 297–304 (2017).
11) 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 (2016).
12) Aas, J. A., Paster, B. J., Stokes, L. N., Olsen, I. & Dewhirst, F. E. Defining the normal bacterial flora of the oral cavity. J Clin Microbiol 43, 5721–5732 (2005).
13) Bultman, S. J. Emerging roles of the microbiome in cancer. Carcinogenesis 35, 249–255 (2014).
14) Nagy, K. N., Sonkodi, I., Szoke, I., Nagy, E. & Newman, H. N. The microflora associated with human oral carcinomas. Oral Oncol34, 304–308 (1998).'

15) Hooper, S. J. et al. Viable bacteria present within oral squamous cell carcinoma tissue. J Clin Microbiol 44, 1719–1725 (2006).
16) Hooper, S. J. et al. A molecular analysis of the bacteria present within oral squamous cell carcinoma. J Med Microbiol 56, 1651–1659 (2007).
17) 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 (2012).
18) Bebek, G. et al. Microbiomic subprofiles and MDR1 promoter methylation in head and neck squamous cell carcinoma. Hum Mol Genet 21, 1557–1565 (2012).
19) Wang, H. et al. Microbiomic differences in tumor and paired-normal tissue in head and neck squamous cell carcinomas. 9, 14 (2017).
20) Caporaso, J. Gregory, et al. "QIIME allows analysis of high-throughput community sequencing data." Nature methods 7.5 , 335 (2010)
21) Huttenhower et al "Metagenomic biomarker discovery and explanation. "Genome Biol 12(6), (2011)
22) Biomarkers Definitions Working Group Biomarkers and surrogate endpoints: preferred definitions and conceptual framework. Clin Pharmacol Ther. 2001; 69(3):89–95.
23) Ferguson DB. Current diagnostic uses of saliva. J Dent Res. 1987; 66(2):420–4.