The salivary metatranscriptome as an accurate diagnostic indicator of oral cancer

Guruduth Banavar (guru@viome.com)
Viome https://orcid.org/0000-0002-4170-7841

Oyetunji Ogundijo
Viome https://orcid.org/0000-0001-7581-0436

Ryan Toma
Viome https://orcid.org/0000-0003-4156-8316

Sathyapriya Rajagopal
Viome

Yenkai Lim
University of Queensland

Kai Dun Tang
Queensland University of Technology

Francine Camacho
Viome https://orcid.org/0000-0002-2817-8252

Pedro Torres
Viome https://orcid.org/0000-0002-0606-8482

Stephanie Gline
Viome

Matthew Parks
Viome https://orcid.org/0000-0003-4283-9675

Liz Kenny
Royal Brisbane and Women's Hospital

Nevenka Dimitrova
New York Medical College

Ally Perlina
Viome

Hal Tily
Viome

Salomon Amar
NYMC https://orcid.org/0000-0002-0017-5930

Momchilo Vuyisich
Viome

Chamindie Punyadeera

https://orcid.org/0000-0001-9039-8259

Article

Keywords: oral cancer, diagnostics, machine learning model

Posted Date: August 19th, 2020

DOI: https://doi.org/10.21203/rs.3.rs-55052/v1

Version of Record: A version of this preprint was published at npj Genomic Medicine on December 1st, 2021. See the published version at https://doi.org/10.1038/s41525-021-00257-x.

The salivary metatranscriptome as an accurate diagnostic indicator of oral cancer

Guruduth Banavar[1,*], Oyetunji Ogundijo[1], Ryan Toma[1], Sathyapriya Rajagopal[1], Yen Kai Lim[2,3], Kai Tang[2,3], Francine Camacho[1], Pedro J. Torres[1], Stephanie Gline[1], Matthew Parks[1], Liz Kenny[4], Nevenka Dimitrova[5], Ally Perlina[1], Hal Tily[1], Salomon Amar[5], Momchilo Vuyisich[1], Chamindie Punyadeera[2,3,*]

[1] Viome Research Institute, Viome Inc, Bellevue, WA / Los Alamos, NM / New York, NY / San Diego, CA, USA
[2] The Saliva and Liquid Biopsy Translational Laboratory, Institute of Health and Biomedical Innovation, Queensland University of Technology, Brisbane, QLD, Australia
[3] The Translational Research Institute, Woolloongabba, Brisbane, QLD, Australia
[4] The School of Medicine, University of Queensland, Royal Brisbane and Women's Hospital, Brisbane, QLD, Australia
[5] New York Medical College, Valhalla, NY, USA
* Corresponding authors: G.B. guru@viome.com and C.P. chamindie.punyadeera@qut.edu.au

Abstract

Despite advances in cancer treatment, the five-year mortality rate for oral cancers (OC) is 40%, mainly due to the lack of early diagnostics. To advance early diagnostics for high-risk and average-risk populations, we developed and evaluated machine-learning (ML) classifiers using metatranscriptomic data from saliva samples (n=433) collected from oral premalignant disorders (OPMD), OC patients (n=71) and normal controls (n=171). Our diagnostic classifiers yielded a receiver operating characteristics (ROC) area under the curve (AUC) up to 0.9, sensitivity up to 83% (92.3% for stage 1 cancer) and specificity up to 97.9%. Our metatranscriptomic signature incorporates both taxonomic and functional microbiome features, and reveals a number of previously known and novel taxa and functional pathways associated with OC. For the first time, we demonstrate the potential clinical utility of an AI/ML model for diagnosing OC early, opening a new era of non-invasive diagnostics, enabling early intervention and improved patient outcomes.

Introduction

Oral cancer (OC) is a major subtype of head and neck cancers (HNC) [HNC Guide]. Worldwide, there are an estimated 350,000 to 400,000 new cases of OC each year, and more than 150,000 deaths [World Health Organization]. In the US in 2020, it is estimated that 53,500 people (~71% male) will be newly diagnosed, and that there will be 10,860 deaths (~73% male) from OC. That amounts to 145 new cases diagnosed every day, and one person dying from OC every hour. The overall 5-year survival rate for people with OC is 40% and this figure has not improved in the past 40 years, resulting in more cancer deaths when compared to melanoma and cervical cancer in the USA [Schmidt 2014]. However, if diagnosed at an early stage, the overall 5-year survival rate is 84%. Unfortunately, with today's practices, only 29% of patients are diagnosed at an early stage.

The cost effectiveness of targeted screening/early diagnostic approaches has been supported by the results from a simulation model study [Dedhia 2011]. Currently, OC is hard to detect in the early stages because of the lack of effective early diagnostic tools, resulting in late diagnosis, leading to poor prognosis and low survival rates [Brocklehurst 2013; Asio 2018], with a significant impact on the healthcare system. Major risk factors for the development of OC are excessive tobacco smoking, alcohol consumption, and in Asia, betel nut chewing. Tobacco use can include consuming tobacco products by smoking, chewing, vaping, etc. OC risk increases with age or a history of tobacco use [ACS; Morse 2007], and the increase becomes more rapid after 50 years of age [NIH Dental Research]. Only 2-4% of OC cases are associated with human papillomavirus (HPV) infection. Additionally, OC commonly occurs in people without a history of tobacco use or alcohol consumption, which argues that additional environmental factors may lead to the development of OC.

Existing microbiological literature has established a significant correlation between changes in the microbiome and cancer phenotypes [Elinav 2019; Poore 2020]. Perhaps the best-known association is of bacteria (*Helicobacter pylori*) causing gastric ulcers that progress into gastric

cancer. In the last decade, multiple microbiome studies using biopsies, tissue samples, and deep epithelial swabs taken from OC patients have shown associations of certain microbes with the development of OC. In previous studies, although there were significant methodological variations in terms of type of samples, technologies used for microbial analysis (16S rRNA gene sequencing or shotgun DNA analysis), design and inclusion criteria, some overlaps were observed at high taxonomic levels. More recently, the notion has emerged that the microbial association with OC is at the level of the microbial community's function, rather than at its composition [Al-hebshi 2017]. Most intriguingly, recent evidence raises the possibility that changes in salivary microbiome composition may have potential as biomarkers for detecting HNCs [Krishnan 2017; Lim 2018; Lee 2017; Zhang 2020; LaRosa 2020].

Visual and tactile screening, followed by laboratory testing and clinical assessment remain the backbone of the current clinical standard of care. A simpler alternative would be measurements made from saliva samples. Saliva specimen collection is non-invasive, straightforward, safe, painless; patients can collect samples themselves. As with home-based stool sample collection, we imagine that removing the need for professional healthcare personnel for sample collection could lead to greater potential for access as well as patient compliance compared to blood-based methods [Salazar 2014]. Saliva is also a more stable and a less complex matrix compared to blood and as such, is ideal for broad use [Tang 2019]. Despite all of these advantages of the use of saliva, an accurate method of profiling the microbiome changes in saliva samples as an early diagnostic indicator has not been developed to date that could generate the much needed clinical impact in this prevalent and deadly disease.

Our overarching aim is to develop a simple, non-invasive, and scalable method, with a classification algorithm that can be used as an early diagnostic tool to address an urgent unmet clinical need. We hypothesized that combining salivary microbial transcriptome (metatranscriptome) profiling using next-generation sequencing (NGS) technology with machine learning (AI/ML) would allow us to develop a classifier that could accurately discriminate premalignant/OC cases from normal healthy controls. We have developed and validated both state-of-the-art techniques for achieving accuracy and robustness in our OC classifier: (1) NGS metatranscriptomic analysis, which captures the microbial activity (RNA) within the saliva sample in high resolution, and accurately identifies both the microbial taxonomies as well as the microbial functions [Hatch 2019], and (2) analytical discovery of the metatranscriptomic signature associated with OC, using a model trained from a machine learning algorithm.

To achieve the above objectives, we collected 433 saliva samples and meta-data from 242 unique individuals, and divided these samples into the cohorts described in Table 1 below. Using these cohorts, we developed and evaluated classifiers for two scenarios:

1. Screening for OC or OPMD within the high-risk population, i.e., 50 years or older, OR with a history of tobacco use
2. Screening for OC only within the average-risk population, i.e., general population across all backgrounds

While we provide the results for both scenarios, we highlight the high-risk OC+OPMD screening scenario in the rest of the paper, since this represents the largest unmet clinical need. Based on our analysis and results across cohorts, the findings from this study provide the foundation for a large multi-center clinical trial to validate the effectiveness of the diagnostic classifier on the populations of interest.

Methods

Study cohorts[1]: The goal of this study was to evaluate diagnostic performance of a novel liquid biopsy on both a high-risk as well as an average-risk population. Table 1 summarizes the participants in the cohorts used in this study: Cohorts A & B represent the high risk population, i.e., people aged 50 years or older OR with a history of tobacco use, so a 55 year old never-smoker and a 25 year old smoker would both belong to this cohort.

- The goal of Cohort A (high-risk OC+OPMD discovery cohort) was to support the primary use case of this study -- to develop a machine-learned classifier for early diagnosis in the high-risk cohort, to analyze the features in the raw data (Figure 1), evaluate the classifier performance (Figure 2 and Table 2), and summarize the metatranscriptomic signature (Figure 3). For this objective, we included both OC and OPMD patients within the positive "cases" category, as one would expect in a clinical early detection or screening test.

[1] This study was approved by the Queensland University of Technology and University of Queensland Medical Ethical Institutional Boards (HREC no.: 1400000617 and HREC no.: 2017000662 respectively) and the Royal Brisbane and Women's Hospital (HREC no.: HREC/12/QPAH/381) Ethics Review Board. Written informed consent was obtained from all participants and all of the methods in this study were performed in accordance with the relevant guidelines and regulations.

Table 1: Study cohorts. High-risk population is 50 yrs or older OR history of smoking (current or past smoker). Average-risk population is the general population across all backgrounds and histories.

	A: High-risk OC+OPMD discovery cohort	B: High-risk OC+OPMD cross-validation (A+27 samples)	C: Average-risk OC-only (OC subset of A + 7 avg-risk)	D: Average-risk technical validation	Total unique across all cohorts
Number of participants Controls Cases	117 59 58	144 75 69	99 49 50	91 91 n/a	242 171 71
Number of samples	117	144	99	282	433
Cases Pre-malignant Malignant	58 10 48	69 14 55	50 n/a 50	n/a n/a n/a	71 14 57
Sex (% female) Controls Cases	37.6 54.2 20.7	37.5 50.7 23.2	40.4 57.1 24	38.7 38.7 n/a	38.8
Age (y) mean ± std Controls Cases	60.2 ± 11.3 56.3 ± 10 64.1 ± 11.4	61.4 ± 11.4 58.5 ± 11 64.5 ± 11.1	59.7 ± 12.6 56 ± 10.8 63.3 ± 13.3	22.6 ± 10.5 22.6 ± 10.5 n/a	37.2 ± 21.7

- The goal of Cohort B (high-risk cross-validation cohort) was to evaluate the performance of our approach by including an additional 27 samples on top of Cohort A.
- The goal of Cohort C (average risk OC-only) was to develop and evaluate a classifier for a broad general population, and with only OC cases (i.e., without the pre-malignant OPMD cases).
- The goal of Cohort D was to perform a technical validation using samples from "presumed normal" individuals from the general population, and to determine whether external interference factors influenced the metatranscriptomic analysis.

For Cohorts A, B, and C, we recruited 71 newly-diagnosed treatment-naive patients with OC and OPMD, and collected a saliva sample from each of them at baseline. In addition, we collected 362 saliva samples from 171 non-diseased individuals across all cohorts shown. The exact inclusion and exclusion criteria are described in the Supplementary Material. Based on histopathological reports, the clinical stages of patients with OC were classified based on the cancer staging system of the American Joint Committee on Cancer [AJCC]. All patients in Cohorts A, B, and C were HPV negative based on a PCR-based test of their saliva [Tang 2020].

Sample collection and laboratory analysis. Laboratory analysis of the saliva samples was similar to the metatranscriptomic method designed for large-scale population analysis of stool samples as described previously [Hatch 2019] (summarized in Figure S1 in Supplementary Material), and included sample collection, ambient temperature sample preservation, total RNA extraction, physical removal of ribosomal RNAs (rRNAs), preparation of directional Illumina libraries, and Illumina sequencing. The stability of the RNA stabilizer was tested for up to 28 days at ambient temperature, including shipping. (More details in Figures S1, S2, S3 in Supplementary Material.)

Bioinformatics processing. Paired-end reads were mapped to a catalog of 53,660 microbial genome assemblies spanning archaea, bacteria, fungi, protozoa, and viruses. Strain-level relative activities were computed from mapped reads via the expectation-maximization (EM) algorithm [Dempster 1977]. Relative activities at other levels of the taxonomic tree were then computed by aggregation according to the taxonomic rank. Relative activities for biological functions were computed by mapping paired-end reads to a catalog of 52,324,420 microbial genes, quantifying gene-level relative activities with the EM algorithm, and then aggregating gene-level activity by

KEGG Ortholog (KO) annotation [Kanehisa 2000]. The identified and quantified active microbial species and KOs for each sample were then provided to the OC classifier. (More details in Supplementary Material.)

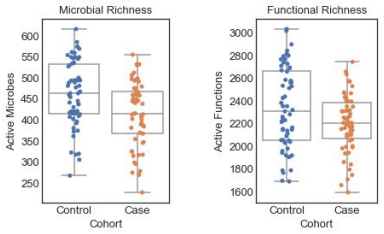
Descriptive statistical analysis. Standard statistical analysis was initially performed to analyze the differential expression of active microbes and active functions between the 58 cases and the 59 healthy controls in Cohort A (Figure 1). The data was transformed using the centered log ratio transformation (CLR) (Aitchison, 1986) after imputation of zero values using multiplicative replacement (Martín-Fernández et al., 2003). We used the Mann-Whitney U (MWU) test ($p < 0.05$) and at least 2 fold difference in means (0.69 in CLR space). It is important to note that this is a descriptive statistical test to analyze features independently for differential expression without taking into account the interactions among features, and is thus not suitable for the machine learning classification method (below).

Mapping KOs to functional categories for presentation. For Figure 1, the Python module "Bio.KEGG" was used to take as input the KO name and return KO hierarchy at three different levels (level-1 to level-3). For Figure 3, Viome Functional Categories (VFCs), each KO and taxa feature from the ML model was analyzed in the context of expert-assessed directional pathway mechanisms or biologically characterized taxonomic microbial groups (see Supplementary Material). Subsequently, the VFCs were summarized into broader biological themes based on literature and their relevance to carcinogenesis or OC progression as described in the Discussion and Supplementary Material.

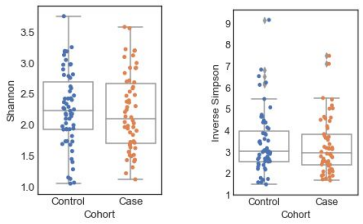
Machine learning (ML) classifier development and cross-validation. The OC binary classifier was trained using the appropriate number of samples from the population in the cohorts described in Table 1. Each sample was annotated as a case (OC or OPMD) or control. The molecular data (microbial species and KOs) derived from the metatranscriptomic analysis of the saliva samples, were used as input features for training. For this study, we chose a logistic regression (LR) model since it performs well and is easily interpretable. In particular, we used l_2 regularized logistic regression with a regularization parameter of 1, implemented in scikit-learn (Pedregosa et al, 2011). This choice was motivated by low model complexity as protection against overfitting.

We used "leave one out cross-validation" (LOOCV) to validate both feature selection and model performance. It is conventionally held that in k-fold validation, as k approaches N (i.e., approaches LOOCV), estimator variance decreases due to increasing number of observations and aggregation over a greater number of folds, but increases due to increasing nonindependence of the data comprising each fold (e.g. [Hastie 2001]). Due to the small sample size relative to the number of active microbes and functions, we took precautions to ensure that the features we present are robust to random variation in the data. The following procedure was used:

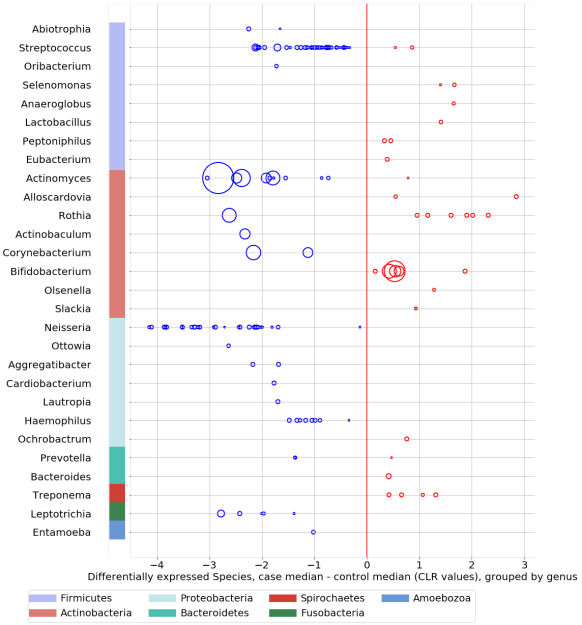
- To begin with, the features in our molecular data consist of all detected active microbes (1587 species) and functions (4932 KOs).
- Data was transformed using the CLR method [Aitchison 1986]. Features with variance less than 25th percentile of the variances of all features were removed as part of data pre-processing and 533 active microbes and 2216 active functions were used for the remaining analysis.
- For each fold of the LOOCV method, we performed feature selection as follows. Bootstrap sampling of each training set 1000 times provided the sampling distribution of all LR coefficients. We considered features where the 95% CI of this distribution did not cross zero to be significant at $p<.05$, and used these to estimate the model in each iteration of the LOOCV procedure.
- To obtain a final model for the purposes of follow-on validation or clinical use, we fit an LR model with the 348 features (101 active species and 247 KOs) at the *intersection* of the models built in each fold of cross validation. This is a conservative choice made to select the features consistently selected across cross-validation, and therefore reduce overfitting. We call these 348 features used in the final ML model the *metatranscriptomic signature* of OC.



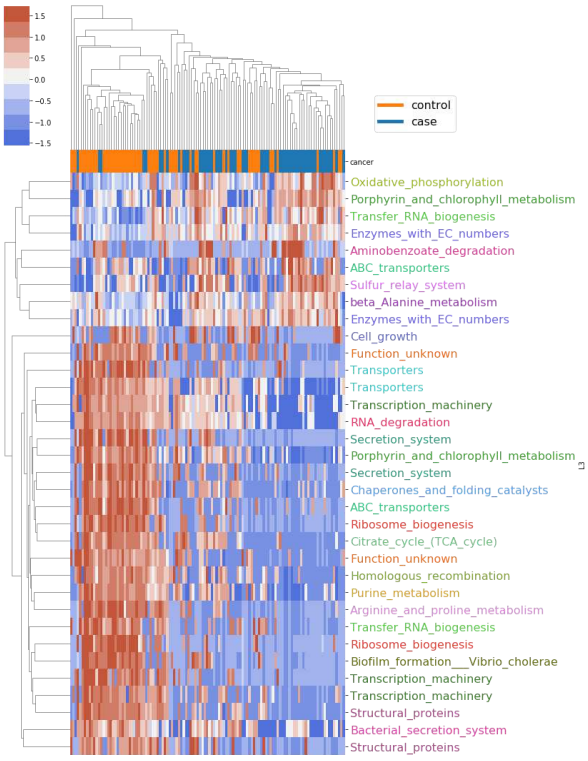
(a)



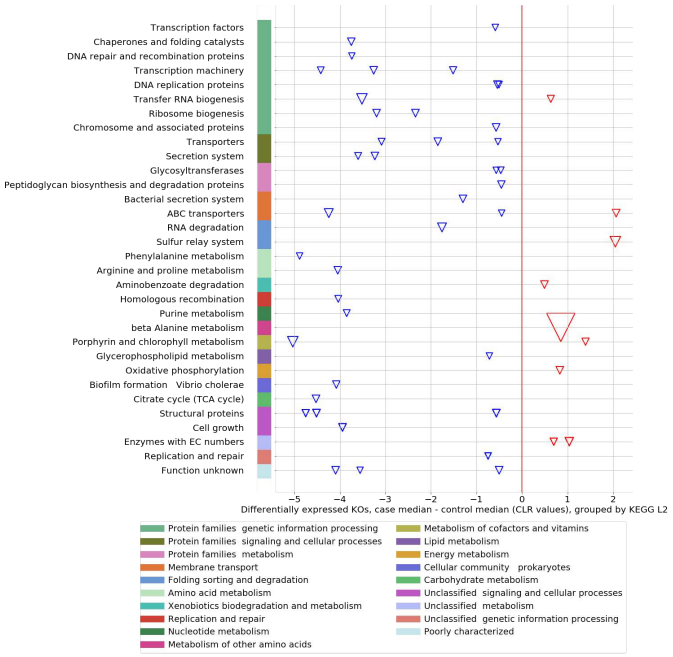
(b)



(c)



(e)



(d)

Figure 1. Descriptive statistics of salivary metatranscriptome of the high-risk population (Cohort A in Table 1). (a) Species richness; control median 463, case median 415 and function richness; control median 2306, case median 2205 (b) Shannon diversity index; control mean 2.25, case mean 2.20; and Inverse Simpson diversity index; control mean 3.41, case mean 3.26 (c) Using Mann-Whitney U tests and at least 2 fold difference in means (0.69 in CLR space), 139 differentially expressed species (at $p < 0.05$) up- or down-regulated in cases relative to controls, organized by genus and phylum (median difference in CLR values); the size of the bubble is inversely proportional to the p-value (d) Using Mann-Whitney U tests and at least 2 fold difference in means (0.69 in CLR space), 49 differentially expressed KOs (at $p<0.05$) up- or down-regulated in cases relative to controls, organized by KEGG level-3 and level-2 functional groups; the size of each triangle is inversely proportional to its p-value (e) Clustermap using Euclidean distance of CLR transformed sum(transcripts per million) data for active function (KO) features significant by Mann-Whitney U tests. Features are shown with corrected p-values < 0.01 and median CLR differences between the cohorts of greater than 0 or less than -1. KOs are color coded by their KEGG level-3 functional group.

Results

Descriptive statistics. Figure 1 summarizes a set of descriptive statistics to show the differences in active species and KOs between the 58 cases and 59 controls in our study. Across all samples used in this study, we detected a wide range of unique active microbes (1587 active species, sample mean 438 ± StD 81) and unique active functions (4932 KEGG Orthologs or KOs, sample mean 2270 ± 314). As shown in Figure 1(a), we observed a lower richness in cases compared to controls, both in terms of active species and active KOs. However, as shown in Figure 1(b), we do not see a statistically significant difference in diversity indices, such as the shannon index or the inverse simpson diversity index between cases and controls. Figure 1(c) shows the up/down regulation of the 139 statistically significant differentially active species between cases and controls, grouped into 28 genera ($p<0.05$, Mann-Whitney U test). We observed a downward shift, i.e. that 75.5% of the species are downregulated in cases compared to controls. For example, out of the 41 differentially active species from the *Streptococcus* genus, 39 were down-regulated, most of them within 2 units on the centered-log-ratio (CLR) scale; the 27 species from the *Neisseria* genus were all down-regulated in cases, with many of them at 4 units on the CLR scale. In contrast, 6 species from the *Rothia* genus were up-regulated in cases compared to controls. Figure 1(d) shows the up/down regulation of the 49 statistically significant differentially expressed KOs between cases and controls, grouped into 30+ KEGG level-3 functional groups ($p<0.05$, Mann-Whitney U test). We observed that most of the microbial functions were downregulated in cases (81.6%), compared to controls. Finally, Figure 1(e) shows a visible distinction between cases and controls using only the differential expression of functions.

It is important to note that Figure 1 presents a *descriptive* statistical analysis of the 58 cases versus 59 normal controls. The differential expression of individual features taken one-at-a-time without interactions provides a level of insight into the raw data, but may not necessarily result in the highest performing diagnostic model. In the section below, we demonstrate that a linear regression machine learning approach provides a significantly higher diagnostic performance, as shown in Figure 2(c).

Predictive performance of the machine-learned (ML) classifier. Figure 2 depicts the clinical diagnostic performance of our trained classifier within the discovery dataset (Cohort A, n=117 in Table 1), using the LOOCV method described earlier. For each incoming validation sample, the trained model outputs a probability that the input sample belongs to the OC/OPMD class (cases). When this probability is above the clinical decision threshold of 0.5, the sample is classified as OC/OPMD (case), otherwise Not-OC/OPMD (control). We used the default probability value of 0.5 for the clinical decision threshold, since it minimizes loss on the training data, and has the advantage that it balances sensitivity and specificity in general. Figure 2(a) shows the probabilities output by our model for all samples in cross validation. Our classifier results are bimodal with good separation of cases and controls, and most data points have predicted probability close to 0 or 1 with very few near the clinical decision threshold. The sensitivity and specificity tradeoff with 95% confidence interval is shown in Figure 2(b). At the clinical decision threshold of 0.5, the sensitivity is 0.81 and specificity is 0.85. Finally, Figure 2(c) shows that our classifier has an ROC AUC of 0.87. Note that a classifier constructed using only the differentially expressed features shown in Figure 1 (139 taxa and 49 KOs) performs at ROC AUC of 0.76 (shown by the orange line in Figure 2c).

Figure 2(d) illustrates that gender does not overly bias our classifier. Figure 2(e) shows that smoking history also does not bias our classifier. It detects non-smokers who have cancer, and it detects smokers and ex-smokers who do not have cancer. Figure 2(f) shows PCA clustering analysis of samples using the top 100 model features, which shows that non-cancer samples are clustered together, providing evidence that the signature is relatively stable. In addition, nine volunteers (from Cohort D in Table 1) provided saliva samples with different potential interferants, such as chewing gum, chewing tobacco, and brushing teeth. Figure 2(g) shows that the probability output of the classifier does not change based on the presence of an interfering substance, showing that our cancer classifier is robust. Taken together, this data demonstrates that our model's performance and robustness is state of the art in the field [Martin 2015; Zhang 2020].

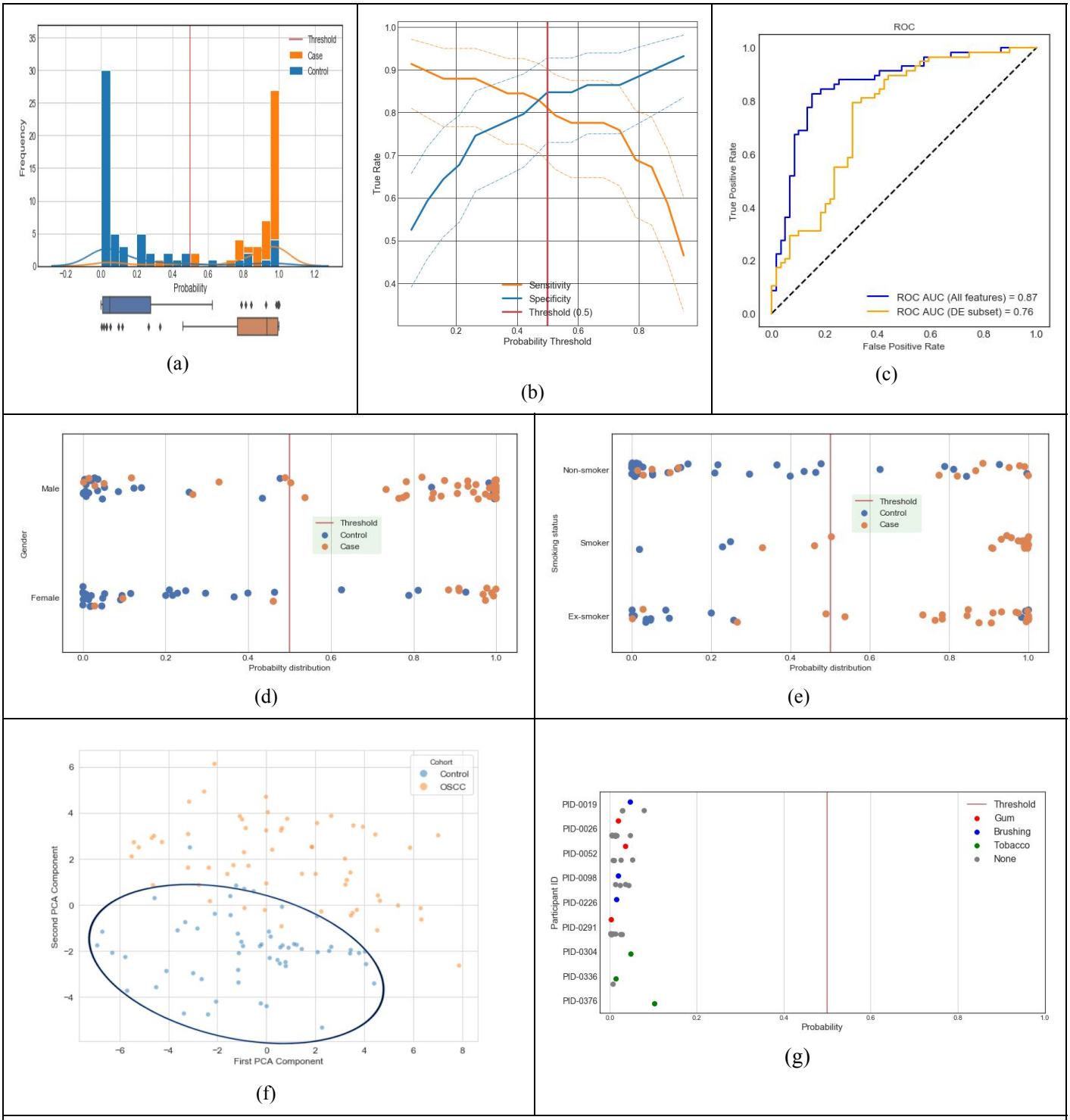


Figure 2. Predictive performance of machine-learnt classifier trained with discovery dataset (Cohort A in Table 1). (a) Distribution of classifier output probabilities across the sample set (b) Sensitivity & specificity tradeoff with 95% confidence interval computed using the Clopper-Pearson method; at the default decision boundary of 0.5, sensitivity is 0.81 and specificity is 0.85. (c) ROC AUC of the classifier using the LOOCV method is 0.87 (blue curve); using differentially expressed features only is 0.76 (orange curve) (d) Classifier probabilities separated by gender (e) Classifier probabilities separated by smoking status (e) PCA analysis using top 100 features (PC1 and PC2 capture 10.2% and 6.3% of the total variation, respectively.) (f) Probability of cancer output from the classifier for control samples with and without interference from chewing gum, chewing tobacco, and brushing teeth.

Table 2: Model performance for cohorts described in Table 1. For sensitivity and specificity, we used the standard default clinical decision threshold of prediction probability=0.5. Technical validation for the average-risk cohort D was performed using the model developed for Cohort A.

	A: High-risk OC+OPMD	B: High-risk CV OC+OPMD	C: Average-risk OC-only	D: Average-risk Technical validation
ROC AUC	0.87	0.87	0.9	n/a
Sensitivity	81%	83%	76%	n/a
Specificity	85%	79%	88%	97.9%
True positives by stage				
OPMD	7/10	11/14	n/a	n/a
Stage 1	12/13	11/14	12/14	n/a
Stage 2	11/16	12/17	11/16	n/a
Stage 3	1/2	2/2	2/3	n/a
Stage 4	13/14	18/19	10/14	n/a

Table 2 gives a summary of the classifier performance for all cohorts in Table 1. The larger high-risk cross-validation Cohort B resulted in a similar performance as Cohort A. The cross-validation performance for a model trained with Cohort C is higher than Cohorts A & B since Cohort C consists of only OC cases without any of the OPMD cases. Cohort D was evaluated with the diagnostic model developed for the primary use case presented in this paper for Cohort A, with the purpose of ensuring that this model is still able to correctly classify a general population, which was confirmed with a specificity of 97.9% (276 true negatives and 6 false positives). Additional details are in the Supplementary Material.

Metatranscriptomic signature from the ML classifier. Figure 3 depicts the details of the features that drive our predictive model. As described earlier, our "metatranscriptomic signature" consists of 348 features (101 active species and 247 active functions) from the intersection of models built in each fold of cross validation. Here, we introduce a curated set of pathway and taxa categories called 'Viome Functional Categories' (VFCs) that group all the features into 9 major biological themes comprising 36 functional categories. These VFCs shed light on some of the microbial activities and biological pathway mechanisms that are known to be associated with oral carcinogenesis. For example, the functional category "Opportunistic Microbial Activities" consists of 3 features (1 taxon and 2 KOs) with a negative effect in the classifier, and 9 features (5 KOs and 4 taxa) with a positive effect. A brief description of the VFCs, the themes and the features (taxa and KOs) constituting the themes are provided in the Supplementary Material.

The functional categories and features within 'ProInflammatory Activities promoting Carcinogenesis', 'Hydrogen Sulfide Production', 'Cancer-specific Energy Metabolism and Utilization', 'Lack of Protective or Detox Mechanisms', 'Reduced Microbial Nitrate Utilization', 'Protein Fermentation', and 'Toxicity Burden', are more direct in terms of their association in oral carcinogenesis. This can be seen by the presence of a greater number of features (taxa and KO) that have a positive effect from the model in the 'Opportunistic Microbial Activities' and 'Hydrogen Sulfide Production' and 'Production of Carcinogenic Exotoxins' themes. These themes have already been implied in oral carcinogenesis [Bouza 2017; Chen 1988]. Amongst the other functional categories, the features are more associated with general Oral Microbiome related Activities and are more predictive of controls. The 'Oral Commensals, Dental Plaque Microbes' constitute microbes such as commensal *Streptococcus* sp. that are important in maintaining oral commensalism and microbiome balance [Kaci 2014]. The functional categories that harbor many features have a negative effect from the model include pathways supporting normal cellular metabolism such as 'Carbohydrate Metabolism and Transport Pathways', 'Amino Acid Production and Transport Pathways', 'Microbial Heat and Osmolarity Induced Stress Pathways' and themes involved in cell growth, such as 'Ribosome Biogenesis' and 'Cell Wall and Sporulation'.

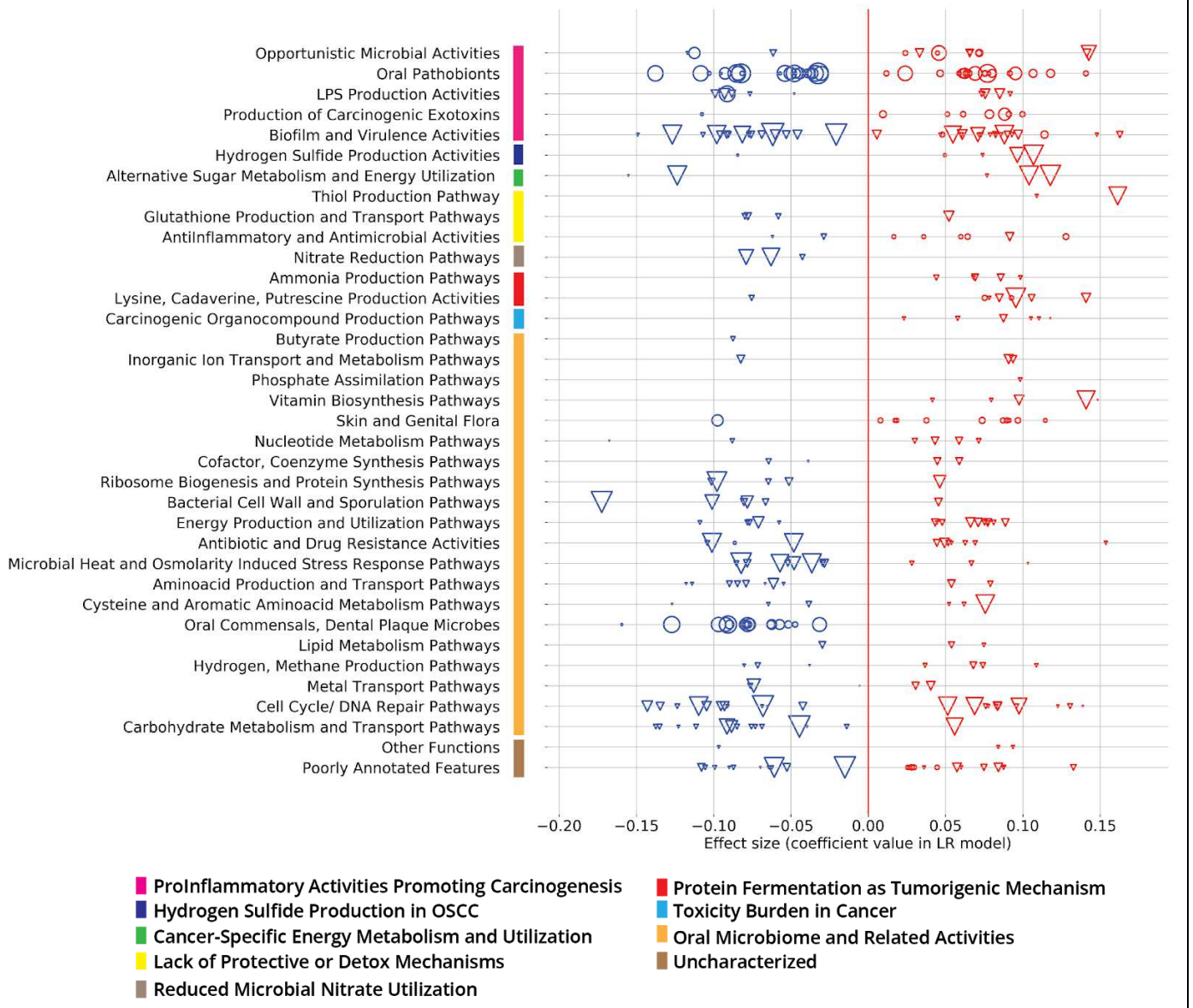


Figure 3: Oral metatranscriptomic signature from the ML classifier trained with Cohort A from Table 1: effect sizes of 101 active species (circles) and 247 active KOs (triangles), grouped into curated Viome Functional Categories (VFC, see Supplementary Material); size of circles or triangles is proportional to the CLR median difference between cases and controls

Discussion

The five-year overall survival rate for all OC in the U.S. is 84%, but drops to 39%-65% when diagnosed at an advanced stage (percentage dependent on location and extent of metastasis) [Cancer Facts and Figures 2019]. Visual and tactile screenings are the foundation of the current standard of care, usually performed by dental hygienists and primary care physicians, which while being quick and easy, are subjective (e.g., verbal questions about symptoms) and prone to a high number of false negatives and false positives [LeHew 2010]. Research has shown some of the reasons for late diagnosis, which is a layered and complicated problem, including under-utilization of dental and primary care, lack of and poor quality of screening in individuals at a higher risk of developing OC and who do not seek general care, and especially, the fact that in the earliest, most treatable stages, many OC show few symptoms and may not be visible [Pitiphat 2002; Peacock 2008; LeHew 2010; Panzarella 2014; El-hakim 2016; Rodriguez-archilla 2017]. Several adjunct diagnostic tools are available to aid providers in identification and diagnosis of OCs [Charanya 2016], such as brush cytology [Abdulhameed 2018; CDx Diagnostics 2020], toluidine blue staining [Pallagatti 2013], and

light-based visual detection systems [Nagi 2016]. The use of these tools varies among providers, and currently, none of the available tools has been studied sufficiently to prove that their use improves the sensitivity and specificity of the current standard of care physical exam [Lingen 2008; Giovannacci 2016]. The work presented in this paper addresses these issues in the standard of care. We introduce a novel method which has non-invasive and easy sample collection using saliva rinse, coupled with an objective and robust classification algorithm with high sensitivity and specificity to distinguish between control samples and oral cancer samples.

There has been interest in investigating either individual bacteria or shifts in microbiome composition and their potential association with different stages of cancer development, since the classification of *Helicobacter pylori* as a causative agent for stomach cancers. In addition, there have been many published studies on the potential association between changes in the microbiome (mainly at the metagenomics level) and cancer. Even though microorganisms have been implicated in 15.4% of human malignancies, there is a dearth of knowledge regarding the role of bacteria in the development and progression of OC. Conventional differential expression analysis reported by existing studies [Guerrero-Preston 2016] shows statistical differences in microbial features between cases and controls, but no study has yet presented a microbiome-based predictive classifier using a non-invasive sampling method. Furthermore, while the majority of microbiome studies to date have focused on microbial taxonomy (due mostly to the limitation of DNA sequencing), we used a combined taxonomic and functional analysis (metatranscriptomics) and demonstrate that microbial functions make important contributions to our model. This is not unexpected, since the biological activity (of mechanistic relevance to OC biology) is the result of active gene expression, and not just genetic potential encoded by DNA.

In this study, we have used both taxonomic profiling and functional profiling to develop a diagnostic classifier based on AI/ML using salivary metatranscriptomic data. We have detected a wider range of unique active microbes (1587 active species, sample mean 438 StD 81) and unique active functions (4932 KEGG Orthologs or KOs, sample mean 2270 ± 314) than previous studies, making it feasible to comprehensively profile bacterial functions (KOs). Our AI/ML diagnostic classifier is effective in identifying individuals who are at high risk of developing OC, starting with pre-malignant lesions / OPMD (Cohort A in Table 1), which is the largest unmet clinical need in this space. For this cohort, cross-validation of our diagnostic classifier yielded an ROC AUC of 0.87, sensitivity of 0.81 and specificity of 0.85. For a more narrow use case such as Cohort C which includes only OC cases, our ML model achieves ROC AUC over 0.9. A secondary technical validation using 91 healthy individuals (Cohort D) yielded a sensitivity of 97.9%. To the best of our knowledge, our classifier has the best diagnostic performance published currently.

We have observed a lower richness, both in terms of active species and active KOs in saliva samples analysed from cases compared to controls (Figure 1a), corroborating with a previous study by [Guerrero-Preston 2016] using salivary metagenomic analysis. In contrast, another study revealed much greater diversity of bacterial communities in OC samples [Zhao 2017]. Our signature has 10 out of 11 genera common with Yang's work [Yang 2018], with *Streptococcus* at the top in both, showing a high degree of concordance. Our high-throughput metatranscriptomic technology can detect features (strain level taxa as well as KOs for functional activity) at a much finer granularity compared with 16S techniques used in Yang's work (or the work pertaining to most other academic research today) [Hatch2019]. Nevertheless, this level of concordance with prior work is highly encouraging. We have also detected at the genus level high amounts of periodontal bacteria *Fusobacterium*, *Prevotella* and *Porphyromonas* in saliva samples from OC and OPMD, confirming previous findings. Furthermore, we believe that our model is specific to OC and does not overlap with other common conditions such as canker sores, since there is negligible overlap (2 species) between the features of our signature and the microbial signature discovered by Kim [Kim 2016].

Among the ProInflammatory Activities promoting carcinogenesis, we identified several species of pathobionts from *Porphyromonas*, *Treponema*, *Fusobacterium*, and *Streptococcus* genera and their raffinose, stachyose, and melibiose transporters, as previously reported [Alanazi 2018; Nagata 2011; Conrads 2014]. This theme also captured two *Porphyromonas* species and one microbial KO shown to produce proinflammatory mediators [Utispan 2018; Goncalves 2016] and eight KOs that are involved in biofilm formation and virulence [Li 2008; Matilla 2018]. Protein Fermentation and polyamine metabolism is known to be associated with tumorigenesis by mediating oxidative damage to the host cells [Goodwin 2011], we report protein fermentation and ammonia-producing KOs as predictors of OC [Palmer 2009; Moreno-Sanchez 2020; Spinelli 2017]). Five toxin-generating KOs that produce benzaldehyde,

arsenite, and other carcinogenic metabolites also contribute to the pathogenesis of OC [Bouza 2017; Chen 1988; Hughes 2002].

Species-level taxonomic classifications were essential for identifying relevant taxa that are predictive of the phenotype. This is clearly depicted in Figures 1(c) and 3, where several genera contain multiple species and that make opposite contributions to the model. This is an important observation, as there are many literature reports that show genera as contributing to a phenotype. In reality, that finding may be driven by certain species within the genera, but other species may have the opposite effect. Therefore, genus-level analysis can lead to false results of a test, depending on the specific species present in a sample.

Our approach improves on previous functional methods by revealing not simply differential expression and functional categorization, but more importantly, mechanisms that integratively connect predictive gene-encoded active functions along with active microbes to relevant biological themes characteristic of OC. Understanding the systems biology level perspective revealed by our ML model can take us one step closer to developing not only diagnostic but also future therapeutic strategies to address this disease.

Ideally, the diagnostic classifier developed in this study would be used clinically as an early detection / screening tool for a high-risk population (adults of either sex 50 years or older OR those with a history of tobacco use). A positive result may indicate the presence of either OPMD or OC and should be followed by, for instance, a detailed physical examination and/or a biopsy by an appropriate medical practitioner (dental surgeon, ENT specialist, etc). Due to the simple, efficient and non-invasive nature of the saliva collection procedure, it is unlikely that such a prediction model will cause any potential adverse effects. The primary risk associated with this prediction model is the possibility of a false prediction (i.e. a false positive or a false negative result). All positive test results will need to be followed by a physical examination of the patient. In a situation where the device produces a false negative result, there is a chance that a case of OC could go undetected, but this risk is no greater than what exists under current standard of care (visual/tactile examination by a medical practitioner).

Overall, we believe that the AI/ML-based diagnostic classifier developed and validated in this study opens a new era of non-invasive diagnostics, enabling early intervention and improving patient outcomes, while significantly reducing healthcare costs.

Conclusions

The main contribution of this paper is a diagnostic system that addresses an unmet clinical need for early detection of oral cancer (including pre-malignant cases) in high-risk populations (people 50 years or older OR with a history of tobacco use). Our system uses (a) a simple, non-invasive, saliva sample (b) high throughput NGS metatranscriptomic lab analysis, and (c) a machine-learned diagnostic classifier that accurately discriminates between cases and controls. We show that this system can identify high-risk OPMD/OC patients vs. normal healthy controls with ROC AUC of 0.87. When restricted only to OC patients at average risk, our classifier achieves ROC AUC over 0.9. For the first time, we demonstrate a system that effectively improves upon the current standard of care globally, opening a new era of non-invasive diagnostics, enabling early intervention and improving patient outcomes.

Our method is based on extracting high-resolution metatranscriptomic (RNA) functional and taxonomic features from saliva samples (rather than genus-level 16S or metagenomic/DNA features), which represents gene expression of active microbial functions in the sample. Second, rather than performing a differential expression analysis of each feature as in most current literature, we perform a machine-learning analysis that captures the inter-dependencies among the thousands of features within the processes, and allows us to *predict* the probability of a cancer signature in a sample. This allows us to identify and connect the most important predictive features that represent active microbial functions along with active microbes to relevant biological themes characteristic of oral cancer. While the results in this discovery study are encouraging, we recognize the limitations of the number of samples in the current study, and plan to perform a large multi-site study to validate the signature on a broader scale.

Finally, once an early diagnostic test is available at scale, we can routinely improve the accuracy of our test as we collect more “real world evidence” to further train our machine learning models. This enables *de novo* discoveries that will have a great impact and open a new era of precision-medicine.

Acknowledgements
CP is funded by the Cancer Australia grant (APP1145657) and the Garnett Passé and Rodney Williams Foundation.

Author contributions

G.B., C.P., and M.V. designed the study and wrote the manuscript. C.P. coordinated the sample and clínical data acquisition. R.T., Y.K.L., and K.T. collected samples and managed the sample logistics. R.T. and M.V. performed the lab analysis. F.C., P.J.T., and M.P. developed the bioinformatics pipeline. O.O. developed the machine learning models, generated visualizations along with S.G., and contributed to writing. H.T. and N.D. guided the data analysis and contributed to writing. A.P. and S.R. interpreted the features and contributed to writing. L.K. recruited the patients and provided clinical data. S.A. interpreted the clinical data and contributed to writing. G.B. guided the analysis and coordinated the project.

Conflict of interest statement

Several of the authors are employees of Viome Inc, a commercial for-profit company. For the other authors there is no conflict of interest to be best of our knowledge.

References

[Abdulhameed 2018] Abdulhameed, H.A., Kujan, O. and Farah, C.S. The utility of oral brush cytology in the early detection of oral cancer and oral potentially malignant disorders: a systematic review. Journal of Oral Pathology \& Medicine. 47(2): 104--116. https://doi.org/10.1111/jop.12660

[ACS] Risk Factors for Oral Cavity and Oropharyngeal Cancers

[Aitchison 1986] Aitchison, J. The Statistical Analysis of Compositional Data. New York: Chapman and Hall. 416p

[AJCC] The 8th edition of the American Joint Committee on Cancer/Union for International Cancer Control (AJCC/UICC) tumour-node-metastasis (TNM) staging system.

[Al-hebshi 2017] Al-hebshi, N.N., Nasher, A.T., Maryoud, M.Y., Homeida, H.E., Chen, T., Idris, A.M. and Johnson, N.W. Inflammatory bacteriome featuring Fusobacterium nucleatum and Pseudomonas aeruginosa identified in association with oral squamous cell carcinoma. Scientific reports, 7(1): 1--10. https://doi.org/10.1038/s41598-017-02079-3.

[Al-Ibrahim 1990] Al-Ibrahim, M.S. and Gross, J.Y. Tobacco Use. Chapter 40 in Clinical Methods: The History, Physical, and Laboratory Examinations. 3rd edition. by Walker HK, Hall WD, Hurst JW, editors. Boston: Butterworths; 1990 https://www.ncbi.nlm.nih.gov/books/NBK362/

[Alanazi 2018] Sultan Ali S. Alanazi, Khalid Tawfik A. Alduaiji, Bharathraj Shetty, Hamad Awadh Alrashedi, B.L. Guruprasanna Acharya, Sajith Vellappally, Darshan Devang Divakar, Pathogenic features of Streptococcus mutans isolated from dental prosthesis patients and diagnosed cancer patients with dental prosthesis, Microbial Pathogenesis, Volume 116, 2018, Pages 356-361, ISSN 0882-4010, https://doi.org/10.1016/j.micpath.2018.01.037.

[Asio 2018] Asio, J., Kamulegeya, A., & Banura, C. (2018). Survival and associated factors among patients with oral squamous cell carcinoma (OSCC) in Mulago hospital, Kampala, Uganda. Cancers of the Head & Neck, 3(1), 9. https://doi.org/10.1186/s41199-018-0036-6

[Belazi 2004] Belazi M, Velegraki A, Koussidou-Eremondi T, et al. Oral Candida isolates in patients undergoing radiotherapy for head and neck cancer: prevalence, azole susceptibility profiles and response to antifungal treatment. Oral Microbiol Immunol. 2004;19(6):347-351.

[Bouza 2017] Bouza M, Gonzalez-Soto J, Pereiro R, de Vicente JC, Sanz-Medel A. Exhaled breath and oral cavity VOCs as potential biomarkers in oral cancer patients. J Breath Res. 2017;11(1):016015. Published 2017 Mar 1. doi:10.1088/1752-7163/aa5e76

[Brocklehurst 2013] Brocklehurst, P., Kujan, O., O'Malley, L.A., Lucy, A., Ogden, G., Shepherd, S. and Glenny, A. Screening programmes for the early detection and prevention of oral cancer. Cochrane database of systematic reviews. 11. https://doi.org/10.1002/14651858.CD004150.pub4

[Cancer Facts and Figures 2019] American Cancer Society. Cancer Treatment & Survivorship Facts & Figures 2019-2021. Atlanta, Ga: American Cancer Society; 2019. https://www.cancer.org/content/dam/cancer-org/research/cancer-facts-and-statistics/cancer-treatment-and-survivorship-facts-and-figures/cancer-treatment-and-survivorship-facts-and-figures-2019-2021.pdf

[CDx Diagnostics 2020] CDx Diagnostics, Inc. What is the OralCDx Brushtest? CDx Diagnostics, 2020 https://www.cdxdiagnostics.com/what-is-oralcdx-brushtest/.

[Charanya 2016] Charanya, D., Raghupathy, L.P., Farzana, A.F., Murugan, R., Krishnaraj, R. and Kalarani, G. 2016. Adjunctive aids for the detection of oral premalignancy. Journal of pharmacy & bioallied sciences 8(Suppl 1), pp. S13–S19. https://dx.doi.org/10.4103%2F0975-7406.191942

[Chen 1988] Chien-Jen Chen, Tsung-Li Kuo, Meei-Maan Wu, Arsenic and cancers 1988, 331, P414-415. DOI:https://doi.org/10.1016/S0140-6736(88)91207-X

[Conrads 2014] Georg Conrads, Johannes J. de Soet, Lifu Song, Karsten Henne, Helena Sztajer, Irene Wagner-Döbler & An-Ping Zeng (2014) Comparing the cariogenic species Streptococcus sobrinus and S. mutans on whole genome level, Journal of Oral Microbiology, 6:1, DOI: 10.3402/jom.v6.26189.

[Dedhia 2011] Dedhia, R.C., Smith, K.J., Johnson, J.T. and Roberts, M. The cost-effectiveness of community-based screening for oral cancer in high-risk males in the United States: a Markov decision analysis approach. The Laryngoscope. 121(5): 952--960. https://doi.org/10.1002/lary.21412

[Elinav 2019] Elinav, E., Garrett, W.S., Trinchieri, G. et al. The cancer microbiome. Nat Rev Cancer 19, 371–376 (2019). https://doi.org/10.1038/s41568-019-0155-3

[El-hakim 2016] EL-Hakim, I. 2016. Delay in oral cancer diagnosis: Who is to blame and are we doing enough? Saudi

Journal of Oral Sciences, 3(1), 56. https://doi.org/10.4103/1658-6816.174339.

[Giovannacci 2016] Giovannacci, I., Vescovi, P., Manfredi, M., & Meleti, M. (2016, May 1). Non-invasive visual tools for diagnosis of oral cancer and dysplasia: A systematic review. *Medicina Oral Patologia Oral y Cirugia Bucal*, Vol. 21, pp. e305–e315. https://doi.org/10.4317/medoral.20996.

[Goncalves 2016] Goncalves, M. et al. Effect of LPS on the Viability and Proliferation of Human Oral and Esophageal Cancer Cell Lines. Braz. arch. biol. technol. [online]. 2016, vol.59, e16150485. Epub Mar 22, 2016. ISSN 1678-4324. https://doi.org/10.1590/1678-4324-2016150485.

Goodwin, Andrew C., Christina E. Destefano Shields, Shaoguang Wu, David L. Huso, XinQun Wu, Tracy R. Murray-Stewart, Amy Hacker-Prietz, et al. "Polyamine Catabolism Contributes to Enterotoxigenic Bacteroides Fragilis</Em>-Induced Colon Tumorigenesis." *Proceedings of the National Academy of Sciences* 108, no. 37 (September 13, 2011): 15354. https://doi.org/10.1073/pnas.1010203108.

[Guerrero-Preston 2016] Guerrero-Preston, R., Godoy-Vitorino, F., Jedlicka, A., Rodr{\'\i}guez-Hilario, A., Gonz{\'a}lez, H., Bondy, J., Lawson, F., Folawiyo, O., Michailidi, C., Dziedzic, A. and others. 16S rRNA amplicon sequencing identifies microbiota associated with oral cancer, human papilloma virus infection and surgical treatment. Oncotarget 7(32): 51320. https://doi.org/10.18632/oncotarget.9710

[Hastie 2001] Friedman, Jerome, Trevor Hastie, and Robert Tibshirani. The elements of statistical learning. Vol. 1. No. 10. New York: Springer series in statistics, 2001.

[Hatch 2019] Hatch A, Horne J, Toma R, Twibell BL, Somerville KM, Pelle B, Canfield KP, Genkin M, Banavar G, Perlina A, Messier H, Klitgord N, Vuyisich M. A robust metatranscriptomic technology for population-scale studies of diet, gut microbiome, and human health. International Journal of Genomics, Volume 2019, Article ID 1718741. https://doi.org/10.1155/2019/1718741

[HNC Guide] Head and Neck Cancer Guide. https://headandneckcancerguide.org/adults/cancer-diagnosis-treatments/surgery-and-rehabilitation/cancer-removal-surgeries/mandibulectomy/

[Hughes 2002] Michael F Hughes, Arsenic toxicity and potential mechanisms of action, Toxicology Letters, 133, 2002, 1-16, ISSN 0378-4274, https://doi.org/10.1016/S0378-4274(02)00084-X.

[Kaci 2014] Kaci, Ghalia, Denise Goudercourt, Véronique Dennin, Bruno Pot, Joël Doré, S. Dusko Ehrlich, Pierre Renault, Hervé M. Blottière, Catherine Daniel, and Christine Delorme. "Anti-Inflammatory Properties of Streptococcus Salivarius, a Commensal Bacterium of the Oral Cavity and Digestive Tract." Applied and Environmental Microbiology 80, no. 3 (February 2014): 928–34. https://doi.org/10.1128/AEM.03133-13.

[Kim 2016] Kim, S.M. Human papilloma virus in oral cancer. Journal of the Korean Association of Oral and Maxillofacial Surgeons. 42(6): 327--336. https://doi.org/10.5125/jkaoms.2016.42.6.327

[Krishnan 2017] Krishnan, K., Chen, T., & Paster, B. J. (2017). A practical guide to the oral microbiome and its relation to health and disease. Oral diseases, 23(3), 276–286. https://doi.org/10.1111/odi.12509

[LaRosa 2020] La Rosa, G., Gattuso, G., Pedullà, E., Rapisarda, E., Nicolosi, D., & Salmeri, M. (2020). Association of oral dysbiosis with oral cancer development. Oncology letters, 19(4), 3045–3058. https://doi.org/10.3892/ol.2020.11441

[Lee 2017] Lee, W., Chen, H., Yang, S. et al. Bacterial alterations in salivary microbiota and their association in oral cancer. Sci Rep 7, 16540 (2017). https://doi.org/10.1038/s41598-017-16418-x

[Lehew 2010] Lehew C.W., Epstein, J.B., Joel, B., Kaste, L.M., Linda, M. and Choi, Y. Assessing oral cancer early detection: clarifying dentists' practices. Journal of public health dentistry. 70(2): 93--100.

[Li 2008] Li M, Wang C, Feng Y, et al. SalK/SalR, a two-component signal transduction system, is essential for full virulence of highly invasive Streptococcus suis serotype 2. PLoS One. 2008;3(5):e2080. Published 2008 May 7. doi:10.1371/journal.pone.0002080 https://doi.org/10.1111/j.1752-7325.2009.00148.x

[Lim 2018] Lim Y.K. and Punyadeera. A pilot study to investigate the feasibility of transporting saliva samples at room temperature with MAWI Cell Stabilization buffer. Cogent Biology. 4(1): 1470895. 10.1080/23312025.2018.1470895

[Lingen 2008] Lingen, M. W., Kalmar, J. R., Karrison, T., & Speight, P. M. (2008, January). Critical evaluation of diagnostic aids for the detection of oral cancer. Oral Oncology, Vol. 44, pp. 10–22. https://doi.org/10.1016/j.oraloncology.2007.06.011

[Martin 2015] Martin, J.L., Gottehrer, N., Zalesin, H., Hoff, P.T., Shaw, M., Clarkson, J.H.W., Pam Haan, B.S.N., Vartanian, M., Terry McLeod, B.S.N. and Swanick, S.M. Evaluation of salivary transcriptome markers for the early detection of oral squamous cell cancer in a prospective blinded trial. SaT. 14: 14--0. https://www.ncbi.nlm.nih.gov/pubmed/26053640

[Matilla 2018] Miguel A Matilla, Tino Krell, The effect of bacterial chemotaxis on host infection and pathogenicity, FEMS Microbiology Reviews, Volume 42, Issue 1, January 2018, fux052, https://doi.org/10.1093/femsre/fux052

[Moreno-Sanchez 2020] Moreno-Sánchez Rafael, Marín-Hernández Álvaro, Gallardo-Pérez Juan C., Pacheco-Velázquez Silvia C., Robledo-Cadena Diana X., Padilla-Flores Joaquín Alberto, Saavedra Emma, Rodríguez-Enríquez Sara, Physiological Role of Glutamate Dehydrogenase in Cancer Cells, Frontiers in Oncology, 10, 2020, 429. DOI=10.3389/fonc.2020.00429

[Morse 2007] Morse, D. E., Psoter, W. J., Cleveland, D., Cohen, D., Mohit-Tabatabai, M., Kosis, D. L., & Eisenberg, E. (2007). Smoking and drinking in relation to oral cancer and oral epithelial dysplasia. Cancer causes & control : CCC, 18(9), 919–929.

https://doi.org/10.1007/s10552-007-9026-4
[Nagata 2011] Nagata E, de Toledo A, Oho T. Invasion of human aortic endothelial cells by oral viridans group streptococci and induction of inflammatory cytokine production. *Mol Oral Microbiol*. 2011; 26(1):78-88. doi:10.1111/j.2041-1014.2010.00597.x
[Nagi 2016] Nagi, R., Reddy-Kantharaj, Y., Rakesh, N., Janardhan-Reddy, S. and Sahu, S. Efficacy of light based detection systems for early detection of oral cancer and oral potentially malignant disorders: systematic review. Medicina oral, patologia oral y cirugia bucal. 21(4): e447.
https://doi.org/10.4317/medoral.21104
[NIH Dental Research] Oral Cancer Incidence by Age, Race, and Gender | Data & Statistics
[Pallagatti 2013] Pallagatti, S., Sheikh, S., Aggarwal, A., Gupta, D., Singh, R., Handa, R., Kaur, S., and Mago, J. Toluidine blue staining as an adjunctive tool for early diagnosis of dysplastic changes in the oral mucosa. Journal of clinical and experimental dentistry. 5(4): e187.
https://doi.org/10.4317/jced.51121
[Palmer 2009] Andrew J. Palmer, Radiah A. Ghani, Navneet Kaur, Otto Phanstiel, Heather M. Wallace; A putrescine–anthracene conjugate: a paradigm for selective drug delivery. Biochem J 15 December 2009; 424 (3): 431–438. doi: https://doi.org/10.1042/BJ20090815
[Pan 2014] Pan J, Zhao J, Jiang N. Oral cavity infection: an adverse effect after the treatment of oral cancer in aged individuals. J Appl Oral Sci. 2014;22(4):261-267. doi:10.1590/1678-775720130546
[Panzarella 2014] Panzarella, V., Pizzo, G., Calvino, F., Compilato, D., Colella, G., & Campisi, G. (2014). Diagnostic delay in oral squamous cell carcinoma: The role of cognitive and psychological variables. International Journal of Oral Science, 6(1), 39–45.
https://doi.org/10.1038/ijos.2013.88.
[Patel 2017] Patel, S., Ansari, J., Meram, A., Abdulsattar, J., Cotelingam, J., Coppola, D., Ghali, G. and Shackelford, R. (2017). Increased nicotinamide phosphoribosyltransferase and cystathionine-beta-synthase in oral cavity squamous cell carcinomas. Int J Clin Exp Pathol, 10(1), 702–707.
[Peacock 2008] Peacock, Z. S., Pogrel, M. A., & Schmidt, B. L. (2008). Exploring the reasons for delay in treatment of oral cancer. Journal of the American Dental Association, 139(10), 1346–1352.
https://doi.org/10.14219/jada.archive.2008.0046.
[Pitiphat 2002] Pitiphat, W., Diehl, S. R., Laskaris, G., Cartsos, V., Douglass, C. W., & Zavras, A. I. (2002). Factors associated with delay in the diagnosis of oral cancer. Journal of Dental Research, 81(3), 192–197.
https://doi.org/10.1177/154405910208100310
[Poore 2020] Poore, G.D., Kopylova, E., Zhu, Q. et al. Microbiome analyses of blood and tissues suggest cancer diagnostic approach. Nature 579, 567–574 (2020).
https://doi.org/10.1038/s41586-020-2095-1
[Rodriguez-archilla 2017] Rodriguez-Archilla, A. 2017. Diagnostic Delay in Oral Cancer. In Ann Clin Res Trials (Vol. 2).
https://scientonline.org/open-access/diagnostic-delay-in-oral-cancer.pdf
[Salazar 2014] A novel saliva-based microRNA biomarker panel to detect head and neck cancers. C Salazar, R Nagadia, P Pandit, J Cooper-White, N Banerjee, N Dimitrova, et al. Cellular Oncology 37 (5), 331-338. DOI: 10.1007/s13402-014-0188-2
[Schmidt 2014] Brian L. Schmidt, Justin Kuczynski, Aditi Bhattacharya, Bing Huey, Patricia M. Corby, Erica L. S. Queiroz, Kira Nightingale, A. Ross Kerr, Mark D. DeLacure, Ratna Veeramachaneni, Adam B. Olshen, Donna G. Albertson, Muy-Teck Teh. Changes in Abundance of Oral Microbiota Associated with Oral Cancer. PLOS One, June 2, 2014
https://doi.org/10.1371/journal.pone.0098741
[Spinelli 2017] Spinelli JB, Yoon H, Ringel AE, Jeanfavre S, Clish CB, Haigis MC. Metabolic recycling of ammonia via glutamate dehydrogenase supports breast cancer biomass. Science. 2017;358(6365):941-946. doi:10.1126/science.aam9305
[Tang 2020 Oral HPV16 prevalence in oral potentially malignant disorders and oral cavity cancers. KD Tang, L Menezes, K Baeten, LJ Walsh, B Whitfield, MD Batstone, C Punyadeera, et al. Biomolecules 10 (2), 223. DOI: 10.3390/biom10020223
[Tang 2019] High-risk human papillomavirus detection in oropharyngeal cancers: Comparison of saliva sampling methods. KD Tang, L Kenny, IH Frazer, C Punyadeera. Head & neck 41 (5), 1484-1489. DOI: 10.1002/hed.25578
[Utispan 2018] Kusumawadee Utispan, Kamolparn Pugdee, Sittichai Koontongkaew, Porphyromonas gingivalis lipopolysaccharide-induced macrophages modulate proliferation and invasion of head and neck cancer cell lines, Biomedicine & Pharmacotherapy, Volume 101, 2018, Pages 988-995, ISSN 0753-3322, https://doi.org/10.1016/j.biopha.2018.03.033.
[Yang 2018] Yang, C., Yeh, Y., Yu, H., Chin, C., Hsu, C., Liu, H. and others. Oral microbiota community dynamics associated with oral squamous cell carcinoma staging. Frontiers in microbiology. 9, 862. https://doi.10.3389/fmicb.2018.00862
[Yost 2018] Yost, S., Stashenko, P., Choi, Y., Kukuruzinska, M., Genco, C.A., Salama, A. and others. Increased virulence of the oral microbiome in oral squamous cell carcinoma revealed by metatranscriptome analyses. International journal of oral science. 10 (4), 1 -- 10.
https://doi.org/10.1038/s41368-018-0037-7
[World Health Organization]
https://www.who.int/cancer/prevention/diagnosis-screening/oral-cancer/en/
[Zhang 2020] Zhang Ling, Liu Yuan, Zheng Hua Jun, Zhang Chen Ping. *The Oral Microbiota May Have Influence on Oral Cancer.* Frontiers in Cellular and Infection Microbiology Vol 9, Jan 2020.
https://www.frontiersin.org/article/10.3389/fcimb.2019.00476
[Zhao 2017] Zhao Hongsen, Chu Min, Huang Zhengwei, Yang Xi, Ran Shujun, Hu Bin, Zhang Chenping, Liang Jingping. *Variations in oral microbiota associated with oral cancer.* Scientific Reports.7(1):1-10. https://doi.org/10.1038/s41598-017-11779-9

Supplementary Material

Study inclusion and exclusion criteria. All participants were aged 18 years or older, able to speak and read English, and were willing and able to follow study instructions. We excluded people who were pregnant, had known active infections and/or under antibiotics. All participants were not taking any local and/or systemic antibiotics prior to sample collection at point of diagnosis. For the normal healthy controls, their state of health was assessed using a survey questionnaire on recent history of alcohol or drug abuse or other medical condition; no prior individual history of any cancer (acceptable if family history of cancer); and no previous irradiation to head and neck region.

Clinical labels. Clinical assessments for the OC patients were performed using standard of care biopsies and histopathology evaluations. 45 OSCC samples were provided TNM codes as specified by [AJCC]. These TNM codes were mapped to 13 Stage I, 16 Stage II, 2 Stage III, and 14 Stage IV samples. The OPMD samples captured conditions such as: Epithelial hyperplasia with hyperkeratosis and mild dysplasia, fibroepithelial hyperplasia with hyperkeratosis, mild epithelial dysplasia, mild patchy lichenoid inflammatory change, mild lichenoid dysplasia, lichenoid reaction, hyperplastic squamous mucosa with hyper and parakeratosis, acanthosis associated with lichenoid inflammatory changes, mild non-specific chronic inflammation and overlying parakeratosis, oral lichen planus, and verrucous leukoplakia.

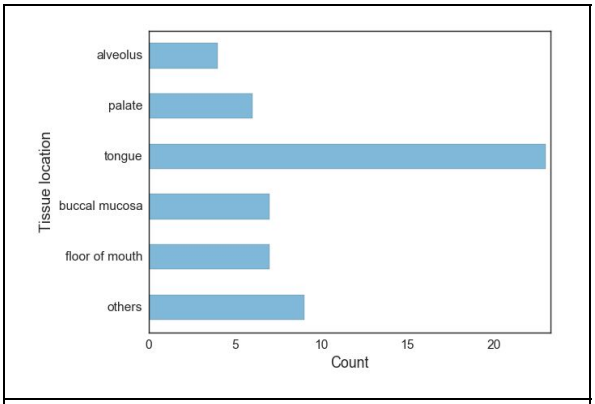


Figure S0: Details of the discovery cohort.

Saliva sample collection and processing. Saliva samples were collected from all participants at a resting stage. Participants were asked to refrain from eating and drinking for 1-hour prior to the collection of saliva, with the exemption of drinking plain water to ensure they are fully hydrated. Prior sample collection, bottled water was given to participants to rinse their mouth. During saliva collection/expectoration, participants sat comfortably in an upright position with the head slightly tilted forward so that saliva pools to the front of the mouth. The participants were asked to pool saliva (head tilted slightly down) in the mouth for about 2-5 minutes, and expectorate into a specimen collection cup (at least 1-5 ml of saliva) as per our previous (Lim et al 2017, Ovchinnikov et al 2014). Collection was done under the supervision / assistance of trained staff. All specimens were preserved using the Viome RNA stabilizer [Hatch 2019], transported back to the laboratory, and stored at -80 °C until further use.

For NGS analysis, a saliva specimen is lysed using bead beating in a chemical denaturant; total RNA is extracted from clarified lysate; DNA is removed using DNase; Bacterial and human rRNAs are physically removed from the specimen using a subtractive hybridization method. Biotinylated DNA probes complementary to rRNAs are hybridized to the total RNA and removed using streptavidin magnetic beads. The remaining RNAs are converted into Illumina sequencing libraries. Each specimen is tagged with 11 bp dual unique molecular barcodes; libraries are pooled; the concentration of library pool is determined and library pools are sequenced on Illumina NovaSeq 6000 to produce sequencing data.

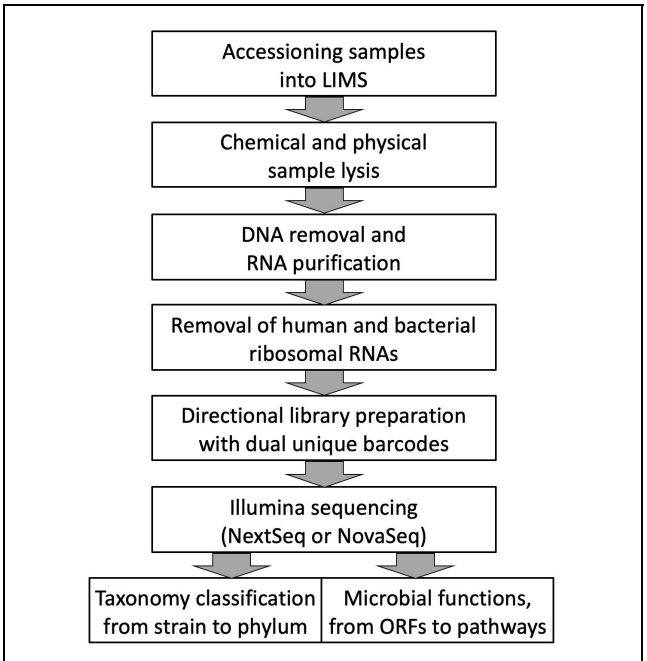
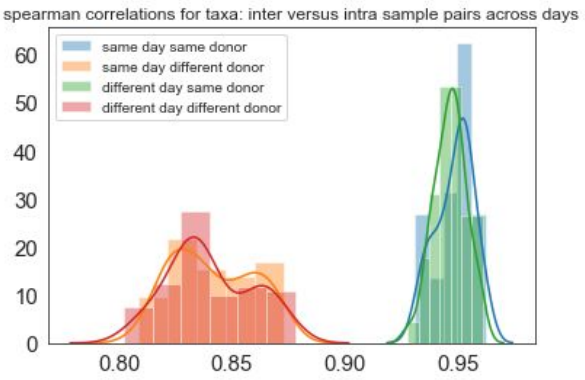
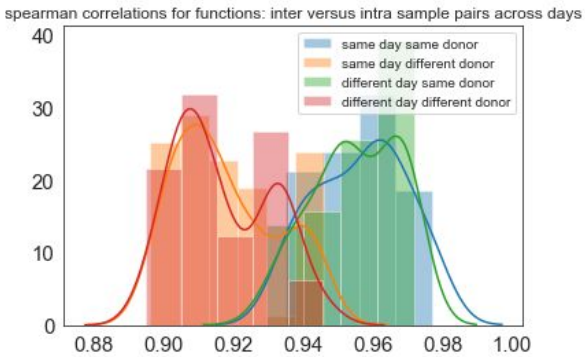


Figure S1: Saliva sample processing procedure for NGS workflow



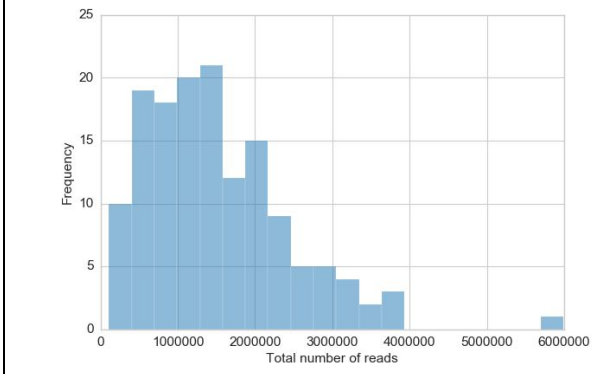
(a)



(b)

Figure S2. Distribution of Spearman correlations for inter and intra donor saliva sample pairs processed concurrently and independently (a) taxonomy data (b) functional data.

Robustness of lab assay. Figure S2 provides a high-level summary of the robustness of our lab process. For this evaluation, we took technical replicates from three saliva donors collected and processed both immediately and stored for 7 days, sequenced in separate batches. We then looked at Spearman correlations between all sample pair combinations spanning donors, sequencing batches and storage conditions for both active microbial and functional data. We find very high correlations across all sample pairs from the same donor regardless of storage and sequencing batch (see overlap in blue and green distributions in Figure S2, mean at 0.96). Additionally, inter-donor sample pairs have lower Spearman correlations which is expected due to biological variation, however, there is no distinction between storage or sequencing batch (see overlap in red and orange distributions in Figure S2, mean at 0.91).



(a)

Figure S3: Distribution of the number of reads

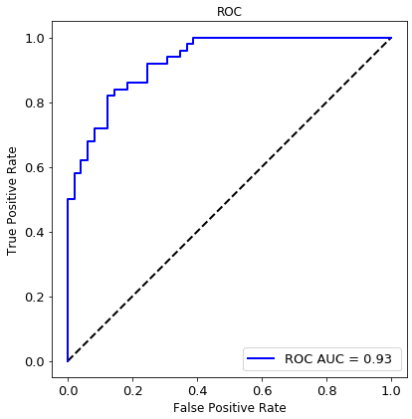
Metatranscriptomic data. The saliva samples were processed through our lab and bioinformatics pipelines, from where we obtained high-resolution metatranscriptomic data of the oral microbiome: these data include 1) active microbes identified against Viome's taxonomic catalog, and their relative activities are calculated at three different taxonomic ranks (genus, species, and strain); and 2) active gene-encoded functions, which are functional ortholog assignments (KEGG Orthologs or KOs) annotated for all sequencing reads aligned to the gene catalog (IGC) of the human microbiome and the KEGG databases. In Figure S3(a), we show the distribution of the total number of RNA reads in all the samples in our pilot study. On average, each sample has 1.5 million reads mapping to mRNA. In total, our molecular data consists of 1587 active microbes and 4932 active functions, a total of 6,519 features. On average, each sample has 444 active microbes and 2299 active functional assignments (Figure S3 (b) & (c)).

We have used the above method/platform to process more than 120,000 samples in our CLIA certified lab to establish that it is robust, accurate, inexpensive, and scalable.

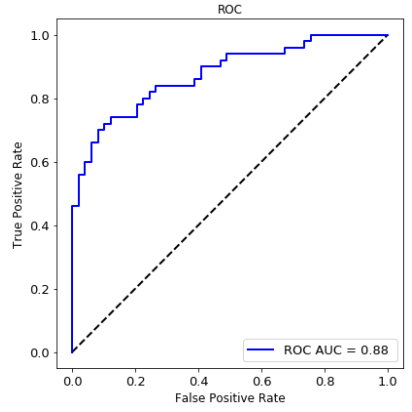
Machine learning methods for additional cohorts

Here we present results of additional analyses performed on Cohort C (average-risk OC-only), which has significant overlaps with Cohort A (92 samples from A that were OC-only, plus 7 additional OC patients who were younger than 50 and with no history of tobacco). In addition to using the entire set of features (taxa and KOs) for modeling purposes as presented in Table 2, we also developed models by separating the taxa & KO features. Figure S4 (a) and (b) shows that the ROC AUC for using taxa and KOs separately

as features are 0.93 and 0.88, respectively. To err on the conservative side, we only presented in Table 2 the model using both taxa & KO features, which performs at 0.9 ROC AUC (which is between the other two values above).



(a)



(b)

Figure S4: ROC AUC of the classifier for Cohort C using (a) only taxa as the features and (b) only the KOs as the features.

Viome Functional Categories (VFC)

Transcriptomic data support the concept that functional, rather than compositional, properties of oral bacterial communities have more relevance to cancer development. We have built an annotation system that integrates both taxonomic abundances and the functional expression profiles from KOs into higher order biological themes that are relevant to the Oral Cancer phenotype and Oral microbiome in general. We call these biological themes as `Viome Functional Categories (VFC)'. The VFC are unique, highly curated themes that take into account the direction of association of taxa and KO features from the OSCC predictive model discussed in the paper and provide mechanistic insights into Oral Carcinogenesis. For instance, the functions or pathways resulting in the production or utilization of a specific metabolite like Hydrogen sulfide or carcinogens could be attributed from the curated VFCs. We report a total of 36 VFC that could be grouped into 9 major biological themes relevant to Oral Cancer and Oral microbiome below.

The theme `**ProInflammatory Activities Promoting Carcinogenesis**` provides evidence of a modified polymicrobial synergy and dysbiosis model for bacterial involvement in OC. The following three VFC provide details about the mechanism that induce inflammation and thereby favor carcinogenesis. Here, we report some of the features that are predictive of OSCC and shed light on some of the mechanisms in oral dysbiosis and periodontal conditions that mediate oral carcinogenesis.

1. Opportunistic Microbial Activities and Oral Pathobionts: The opportunists like "Porphyromonas", "Fusobacterium" and Oral Pathobionts (commensal-derived opportunistic pathogens) such as "Streptococcus sp.", "Gemella sp." are known to mediate oral dysbiosis and lead to subsequent periodontal conditions that might be conducive of OC. These organisms share the ability to attach and invade oral epithelial cells, and communicate with the host epithelium, and ultimately acquire phenotypes associated with cancer such as inhibition of apoptosis (Perara et al 2016), increased proliferation, and increased migration of epithelial cells (Karpin ́ski 2019). Additionally, emerging properties of structured bacterial communities may increase oncogenic potential, and consortia of P. gingivalis and F. nucleatum are synergistically pathogenic within in vivo OC models Among the pathogens positively associated with OSCC from the model are Porphyromonas, Treponema and Fusobacterium and have higher abundances in oral swabs of patients with oral cancer. (Chattopadhyay et al 2019).
2. LPS Production Activities: Bacterial outer membrane lipopolysaccharides are entities that mediate proinflammatory immune response and inflammation host cells. LPS regulates gene expression of pro-inflammatory cytokines through activation of toll-like receptor 4 (TLR4) via NF-kB (Márcia et al 2016). The 'O antigens', an extremely polymorphic polysaccharide binds to LipidA to form the LPS outer-membrane of Gram-negative bacteria thereby imparting antigenic specificity to

the organism. For instance, LPS from Porphyromonas, a positively associated taxa from the OSCC model, is known to activate macrophages and increase NO production of cancer cell lines (Utispan 2018). Furthermore, a functional KO implied in LPS production is positively associated from the OSCC model.

3. Biofilm and Virulence Pathways: The OSCC model predicts a number of functional features associated with bacterial virulence promoting inflammation and positively associated with OC. For instance, sugar transport and chemotaxis associated KOs from oral microbes that are deterministic of virulence and pathogenesis (Matilla 2018) are predicted. Many lytic enzymes, cell wall synthesis associated transporter and phospholipase are the other virulence determining functional KOs that are found as predictive of OSCC from the model.

AntiInflammatory and Antimicrobial Pathways: The commensal bacteria Streptococcus sp. establishes in the human oral cavity a few hours after birth and remains there as a predominant commensal and as a primary colonizer of biofilms. Upon strong adhesion mediated by the glycosylated surface-exposed proteins Streptococcus sp. promotes innate immunity by suppressing proinflammatory cascades as well as by producing anti-microbial substances like bacteriocins that antagonizes the virulent streptococci involved in tooth decay or pharyngitis or pathogens involved in periodontitis (Kaci et al 2014). Similarly, Streptococcus sp. 2, also an early colonial member of oral biofilm produces H2O2 to inhibit the growth of competitors, like the mutans streptococci, as well as strict anaerobic middle and later colonizers of the dental biofilm. Interestingly, Veillonella species, possess a putative catalase gene that mediates resistance to the Streptococcus sp.2 thereby enabling direct physical interaction (co-aggregate) with Streptococcus sp.2 as well as Fusobacterium sp. that are late colonizers of biofilm (Zhou 2017). It is interesting to note that Fusobacterium is a positive predictor of OC while Streptococcus sp.1 is negatively associated. Furthermore, the model captures functional determinant of antimicrobial resistance gene and catalase as positive predictors of OSCC.

Hydrogen sulfide (H2S), a gaseous transmitter, is associated with oral periodontitis and is one of the main causes of halitosis and is generally associated with many oral diseases including OC (Zhang et al 2016). Hydrogen sulfide production pathways including enzymes that produce H2S are increased in different human malignancies. The expression of both enzymes and cellular H2S levels increase tumor survival and promote tumor dedifferentiation [Patel 2017]. Among the taxa, members of the Streptococcus group, Fusobacterium and Porphyromonas, some of the known producers of oral H2S are in turn also predicted from the model to be cancer specific. The model predicts three H2S producing KOs are also positively associated with OC.

Cancer-Specific Energy Metabolism and Utilization: In cancer cells, the Pentose Phosphate Pathway (PPP) together with glycolysis, coordinates glucose flux and supports the cellular biogenesis of macromolecules such as lipids and DNA for energy production. An increased PPP flux in human cancer cells is indicative of its role in meeting the bioenergetic demands of cancer cell proliferation and contribution to the Warburg effect. (Jianrong 2015). Enzymes involved in pentose interconversion, as well in pentose-5P production, are positively associated features from the model suggest microbial dysregulation of PPP flux in human cancer cells.

Lack of Protective or Detox mechanisms: Detoxification mechanisms are essential for multitude of cellular processes, including cell differentiation, proliferation, and apoptosis, and disturbances in their homeostasis are implicated in the etiology and/or progression of a number of human diseases, including cancer, diseases of aging, inflammatory, immune, metabolic, and neurodegenerative diseases With the advent of cancer, a number of protective and detoxifying mechanisms are dysregulated in the cell in response to combat intracellular and extracellular stress. From the model, we see an upregulation of thiol based deconjugation functions, to be positively associated to cancer. Low Molecular weight (LMV) thiols are produced by gram-positive firmicutes that function in protecting cells against reactive oxygen species (ROS) and reactive electrophilic species, antibiotics, alkylating agents, as well as heavy metals (Chandrangsu 2018). On the other hand, microbial glutathione mediated stress response is negatively associated in the model. Thus, a preferential microbial thiol based detoxification of ROS and reactive electrophilic species is known to be associated with OC from our model. Along with these, the antibacterial as well as AntiInflammatory functions such as catalase and butyrate production are downregulated and are found to be negatively associated with cancer.

Protein fermentation as a tumorigenic mechanism: Protein fermentation results in the accumulation of by-products that are resourceful for the cancer cells hence is a favorable environment as a tumor promoting microenvironment. Polyamines such as putrescine,

cadaverine and spermidine are products of microbial protein fermentation are essential for normal cell growth, and their depletion results in cytostasis. Polyamine metabolism is frequently dysregulated in cancer and elevated polyamine levels are necessary for transformation and tumor progression (Murray-Stewart, et al 2016). For instance, the spermidine is needed as a precursor of hypusine (a post-translational addition to eukaryotic initiation factor 5A isoform 1 (eIF5A) that is necessary to prevent ribosomal stalling in the translation of mRNAs encoding polyproline tracts and certain other amino acid combinations. The MYC oncogene plays a role in hypusine formation by driving the transcription of the gene encoding ornithine decarboxylase (ODC) and indirectly increasing the availability of spermidine for hypusine synthesis (Park et al 2010, Casero Jr. et al 2018). A deoxyhypusine synthase requiring spermidine is identified as a positively associated feature from the model. The cancer cells tend to accumulate increased concentrations of polyamines through increased uptake via their Polyamine Transport System (PTS) (Palmer et al 2009). With increased microbial protein breakdown, cadaverine transport systems transport cadaverine into the host cell and promote carcinogenesis and such a polyamine antiporter is identified as positively associated with cancer from the model. The cellular protein degradation produces ammonia as a by-product which is recycled into central amino acid metabolism to maximize nitrogen utilization (Moreno-Sánchez et al 2020). Increased microbial ammonia production is noted from KOs such as glutamate dehydrogenase associated with OSCC from the model.

Benzaldehyde, arsenite, and other carcinogenic toxins: The exposure to synthetic chemicals such as dyes, organopesticides and pharmaceuticals increases the toxicity burden of cells that elevates the cancer causing potential in general. A feature that contributes to the production of benzaldehyde is detected as the top second feature from the predictive model of OSCC. Benzaldehyde is a potential biomarker for OSCC in breath test (Bouza et al 2017). Further, traces of fluorobenzoate metabolism and acetaldehyde production KOs are also observed to be predictive of oral cancer. Exposure to metallic arsenic is toxic to the cells and the extent of arsenic toxicity is dependent on its oxidative state (Hughes 2002, Chen et al 1988). Arsenite transporters are positive predictors of OSCC from the model.

While the above tumor promoting functions are all positively associated with the OSCC, a host of taxa and related activities are also detected as predictive of cancer. These include the **Skin and genital microbes** and several pathway functions such as Inorganic Ion Transport Pathways, Amino acid production and Vitamin Biosynthesis pathways and Cofactor and coenzyme synthesis. Amongst the most prominent negative associated features include the Oral Commensal and plaque microbes such as Streptococcus as well as several pathways such as Energy production, Cell wall biosynthesis and sporulation, Antibiotic resistance, Microbial heat and osmolarity mediated stress which are related to Common oral microbiome related functions not necessarily implicated in cancer. Several pathways such as Cell cycle and DNA repair and Carbohydrate metabolism and transport pathways are found to be less predictive as the features are found to be predictive of both cancer as well as controls.

Supplementary References

- Aas, Jørn A., Bruce J. Paster, Lauren N. Stokes, Ingar Olsen, and Floyd E. Dewhirst. “Defining the Normal Bacterial Flora of the Oral Cavity.” Journal of Clinical Microbiology 43, no. 11 (November 2005): 5721–32. https://doi.org/10.1128/JCM.43.11.5721-5732.2005
- Mikkelsen, L., E. Theilade, and K. Poulsen. “Abiotrophia Species in Early Dental Plaque.” Oral Microbiology and Immunology 15, no. 4 (2000): 263–68. https://doi.org/10.1034/j.1399-302x.2000.150409.x.
- Robertson, D., and A. J. Smith. “The Microbiology of the Acute Dental Abscess.” Journal of Medical Microbiology 58, no. Pt 2 (February 2009): 155–62. https://doi.org/10.1099/jmm.0.003517-0.
- Cargill, James S., Katharine S. Scott, Deborah Gascoyne-Binzi, and Jonathan A. T. Sandoe. “Granulicatella Infection: Diagnosis and Management.” Journal of Medical Microbiology 61, no. 6 (June 1, 2012): 755–61. https://doi.org/10.1099/jmm.0.039693-0.
- Conrads, Georg, Johannes J. de Soet, Lifu Song, Karsten Henne, Helena Sztajer, Irene Wagner-Döbler, and An-Ping Zeng. “Comparing the Cariogenic Species Streptococcus Sobrinus and S. Mutans on Whole Genome Level.” Journal of Oral Microbiology 6, no. 1 (January 2014): 26189. https://doi.org/10.3402/jom.v6.26189.
- Johansson, I., E. Witkowska, B. Kaveh, P. Lif Holgerson, and A.C.R. Tanner. “The Microbiome in Populations with a Low and High Prevalence of Caries.” Journal of Dental Research 95, no. 1 (January 2016): 80–86. https://doi.org/10.1177/0022034515609554.
- Fragkou, S., C. Balasouli, O. Tsuzukibashi, A. Argyropoulou, G. Menexes, N. Kotsanos, and S. Kalfas. “Streptococcus Mutans, Streptococcus Sobrinus and Candida Albicans in Oral Samples from Caries-Free and Caries-Active Children.” European Archives of Paediatric Dentistry 17, no. 5 (October 2016): 367–75. https://doi.org/10.1007/s40368-016-0239-7.
- Vielkind, Paul, Holger Jentsch, Klaus Eschrich, Arne C. Rodloff, and Catalina-Suzana Stingu. “Prevalence of Actinomyces Spp. in Patients with Chronic Periodontitis.” International Journal of Medical

Microbiology: IJMM 305, no. 7 (October 2015): 682–88. https://doi.org/10.1016/j.ijmm.2015.08.018.

- Vieira Colombo, Ana Paula, Clarissa Bichara Magalhães, Fátima Aparecida Rocha Resende Hartenbach, Renata Martins do Souto, and Carina Maciel da Silva-Boghossian. "Periodontal-Disease-Associated Biofilm: A Reservoir for Pathogens of Medical Importance." Microbial Pathogenesis 94 (May 2016): 27–34. https://doi.org/10.1016/j.micpath.2015.09.009.
- Kovacs, C. J., R. C. Faustoferri, and R. G. Quivey. "RgpF Is Required for Maintenance of Stress Tolerance and Virulence in Streptococcus Mutans." Journal of Bacteriology 199, no. 24 (15 2017). https://doi.org/10.1128/JB.00497-17.
- Alanazi, Sultan Ali S., Khalid Tawfik A. Alduaiji, Bharathraj Shetty, Hamad Awadh Alrashedi, B. L. Guruprasanna Acharya, Sajith Vellappally, and Darshan Devang Divakar. "Pathogenic Features of Streptococcus Mutans Isolated from Dental Prosthesis Patients and Diagnosed Cancer Patients with Dental Prosthesis." Microbial Pathogenesis 116 (March 2018): 356–61. https://doi.org/10.1016/j.micpath.2018.01.037.
- Schmidt, Brian L., Justin Kuczynski, Aditi Bhattacharya, Bing Huey, Patricia M. Corby, Erica L. S. Queiroz, Kira Nightingale, et al. "Changes in Abundance of Oral Microbiota Associated with Oral Cancer." PLoS ONE 9, no. 6 (June 2, 2014). https://doi.org/10.1371/journal.pone.0098741.
- Chattopadhyay, Indranil, Mukesh Verma, and Madhusmita Panda. "Role of Oral Microbiome Signatures in Diagnosis and Prognosis of Oral Cancer." Technology in Cancer Research & Treatment 18 (01 2019): 1533033819867354. https://doi.org/10.1177/1533033819867354.
- Karpiński, Tomasz M. "Role of Oral Microbiota in Cancer Development." Microorganisms 7, no. 1 (January 13, 2019). https://doi.org/10.3390/microorganisms7010020.
- Jin, Lin, and Yanhong Zhou. "Crucial Role of the Pentose Phosphate Pathway in Malignant Tumors." Oncology Letters 17, no. 5 (May 2019): 4213–21. https://doi.org/10.3892/ol.2019.10112.
- Giacomini, Isabella, Eugenio Ragazzi, Gianfranco Pasut, and Monica Montopoli. "The Pentose Phosphate Pathway and Its Involvement in Cisplatin Resistance." International Journal of Molecular Sciences 21, no. 3 (January 31, 2020). https://doi.org/10.3390/ijms21030937.
- Li, Ming, Changjun Wang, Youjun Feng, Xiuzhen Pan, Gong Cheng, Jing Wang, Junchao Ge, et al. "SalK/SalR, a Two-Component Signal Transduction System, Is Essential for Full Virulence of Highly Invasive Streptococcus Suis Serotype 2." PLOS ONE 3, no. 5 (May 7, 2008): e2080. https://doi.org/10.1371/journal.pone.0002080.
- "Characterization of a Copper Resistance and Transport System in Streptococcus Mutans | Semantic Scholar." Accessed May 8, 2020. https://www.semanticscholar.org/paper/Characterization-of-a-Copper-Resistance-and-System-Singh/7cb9013d53b09f7c7159cc880ffef6464efed57f.
- Samanovic, Marie I., Chen Ding, Dennis J. Thiele, and K. Heran Darwin. "Copper in Microbial Pathogenesis: Meddling with the Metal." Cell Host & Microbe 11, no. 2 (February 16, 2012): 106–15. https://doi.org/10.1016/j.chom.2012.01.009.
- Vats, Neeraj Kumar, and Song F. Lee. "Characterization of a Copper-Transport Operon, CopYAZ, from Streptococcus Mutans." Microbiology, 2001. https://doi.org/10.1099/00221287-147-3-653.
- Garcia, S. S., Q. Du, and H. Wu. "Streptococcus Mutans Copper Chaperone, CopZ, Is Critical for Biofilm Formation and Competitiveness." Molecular Oral Microbiology 31, no. 6 (December 2016): 515–25. https://doi.org/10.1111/omi.12150.
- Palmer, Andrew J., Radiah A. Ghani, Navneet Kaur, Otto Phanstiel, and Heather M. Wallace. "A Putrescine-Anthracene Conjugate: A Paradigm for Selective Drug Delivery." The Biochemical Journal 424, no. 3 (December 10, 2009): 431–38. https://doi.org/10.1042/BJ20090815.
- Bouza, M., J. Gonzalez-Soto, R. Pereiro, J. C. de Vicente, and A. Sanz-Medel. "Exhaled Breath and Oral Cavity VOCs as Potential Biomarkers in Oral Cancer Patients." Journal of Breath Research 11, no. 1 (01 2017): 016015. https://doi.org/10.1088/1752-7163/aa5e76.
- Nagata, E., A. de Toledo, and T. Oho. "Invasion of Human Aortic Endothelial Cells by Oral Viridans Group Streptococci and Induction of Inflammatory Cytokine Production." Molecular Oral Microbiology 26, no. 1 (February 2011): 78–88. https://doi.org/10.1111/j.2041-1014.2010.00597.x.
- Kaci, Ghalia, Denise Goudercourt, Véronique Dennin, Bruno Pot, Joël Doré, S. Dusko Ehrlich, Pierre Renault, Hervé M. Blottière, Catherine Daniel, and Christine Delorme. "Anti-Inflammatory Properties of Streptococcus Salivarius, a Commensal Bacterium of the Oral Cavity and Digestive Tract." Applied and Environmental Microbiology 80, no. 3 (February 2014): 928–34. https://doi.org/10.1128/AEM.03133-13.
- Lu, Jianrong, Ming Tan, and Qingsong Cai. "The Warburg Effect in Tumor Progression: Mitochondrial Oxidative Metabolism as an Anti-Metastasis Mechanism." Cancer Letters 356, no. 2 Pt A (January 28, 2015): 156–64. https://doi.org/10.1016/j.canlet.2014.04.001.
- Gonçalves, Márcia, Ángelica Regina Cappellari, André Avelino dos Santos Junior, Fernanda Olicheski de Marchi, Fernanda Souza Macchi, Krist Helen Antunes, Ana Paula Duarte de Souza, et al. "Effect of LPS on the Viability and Proliferation of Human Oral and Esophageal Cancer Cell Lines." Brazilian Archives of Biology and Technology 59 (2016). https://doi.org/10.1590/1678-4324-2016150485.
- Utispan, Kusumawadee, Kamolparn Pugdee, and Sittichai Koontongkaew. "Porphyromonas Gingivalis Lipopolysaccharide-Induced Macrophages Modulate Proliferation and Invasion of Head and Neck Cancer Cell Lines." Biomedicine & Pharmacotherapy 101 (May 1, 2018): 988–95. https://doi.org/10.1016/j.biopha.2018.03.033.

- Raja, Manoj, Fajar Ummer, and C.P Dhivakar. "Aggregatibacter Actinomycetemcomitans – A Tooth Killer?" Journal of Clinical and Diagnostic Research : JCDR 8, no. 8 (August 2014): ZE13–16. https://doi.org/10.7860/JCDR/2014/9845.4766.
- Zhou, Peng, Xiaoli Li, I.-Hsiu Huang, and Fengxia Qi. "Veillonella Catalase Protects the Growth of Fusobacterium Nucleatum in Microaerophilic and Streptococcus Gordonii-Resident Environments." Applied and Environmental Microbiology 83, no. 19 (01 2017). https://doi.org/10.1128/AEM.01079-17.
- Patel, Stavan, Junaid Ansari, Andrew Meram, Jehan Abdulsattar, James Cotelingam, Ghali Ghali, and Rodney Shackelford. "Increased Nicotinamide Phosphoribosyltransferase and Cystathionine-Beta-Synthase in Oral Cavity Squamous Cell Carcinomas," n.d., 6.
- Zhang, Shuai, Huan Bian, Xiaoxu Li, Huanhuan Wu, Qingwei Bi, Yingbin Yan, and Yixiang Wang. "Hydrogen Sulfide Promotes Cell Proliferation of Oral Cancer through Activation of the COX2/AKT/ERK1/2 Axis." Oncology Reports 35, no. 5 (May 2016): 2825–32. https://doi.org/10.3892/or.2016.4691.
- Moreno-Sánchez, Rafael, Álvaro Marín-Hernández, Juan C. Gallardo-Pérez, Silvia C. Pacheco-Velázquez, Diana X. Robledo-Cadena, Joaquín Alberto Padilla-Flores, Emma Saavedra, and Sara Rodríguez-Enríquez. "Physiological Role of Glutamate Dehydrogenase in Cancer Cells." Frontiers in Oncology 10 (April 9, 2020). https://doi.org/10.3389/fonc.2020.00429.
- Spinelli, Jessica B., Haejin Yoon, Alison E. Ringel, Sarah Jeanfavre, Clary B. Clish, and Marcia C. Haigis. "Metabolic Recycling of Ammonia via Glutamate Dehydrogenase Supports Breast Cancer Biomass." Science (New York, N.Y.) 358, no. 6365 (November 17, 2017): 941–46. https://doi.org/10.1126/science.aam9305.
- Chen, C. J., T. L. Kuo, and M. M. Wu. "Arsenic and Cancers." Lancet (London, England) 1, no. 8582 (February 20, 1988): 414–15. https://doi.org/10.1016/s0140-6736(88)91207-x.
- Hughes, Michael F. "Arsenic Toxicity and Potential Mechanisms of Action." Toxicology Letters 133, no. 1 (July 7, 2002): 1–16. https://doi.org/10.1016/s0378-4274(02)00084-x.
- Belazi, M., A. Velegraki, T. Koussidou-Eremondi, D. Andreadis, S. Hini, G. Arsenis, C. Eliopoulou, E. Destouni, and D. Antoniades. "Oral Candida Isolates in Patients Undergoing Radiotherapy for Head and Neck Cancer: Prevalence, Azole Susceptibility Profiles and Response to Antifungal Treatment." Oral Microbiology and Immunology 19, no. 6 (December 2004): 347–51. https://doi.org/10.1111/j.1399-302x.2004.00165.x.
- PAN, Jie, Jun ZHAO, and Ning JIANG. "Oral Cavity Infection: An Adverse Effect after the Treatment of Oral Cancer in Aged Individuals." Journal of Applied Oral Science 22, no. 4 (2014): 261–67. https://doi.org/10.1590/1678-775720130546.

Figures

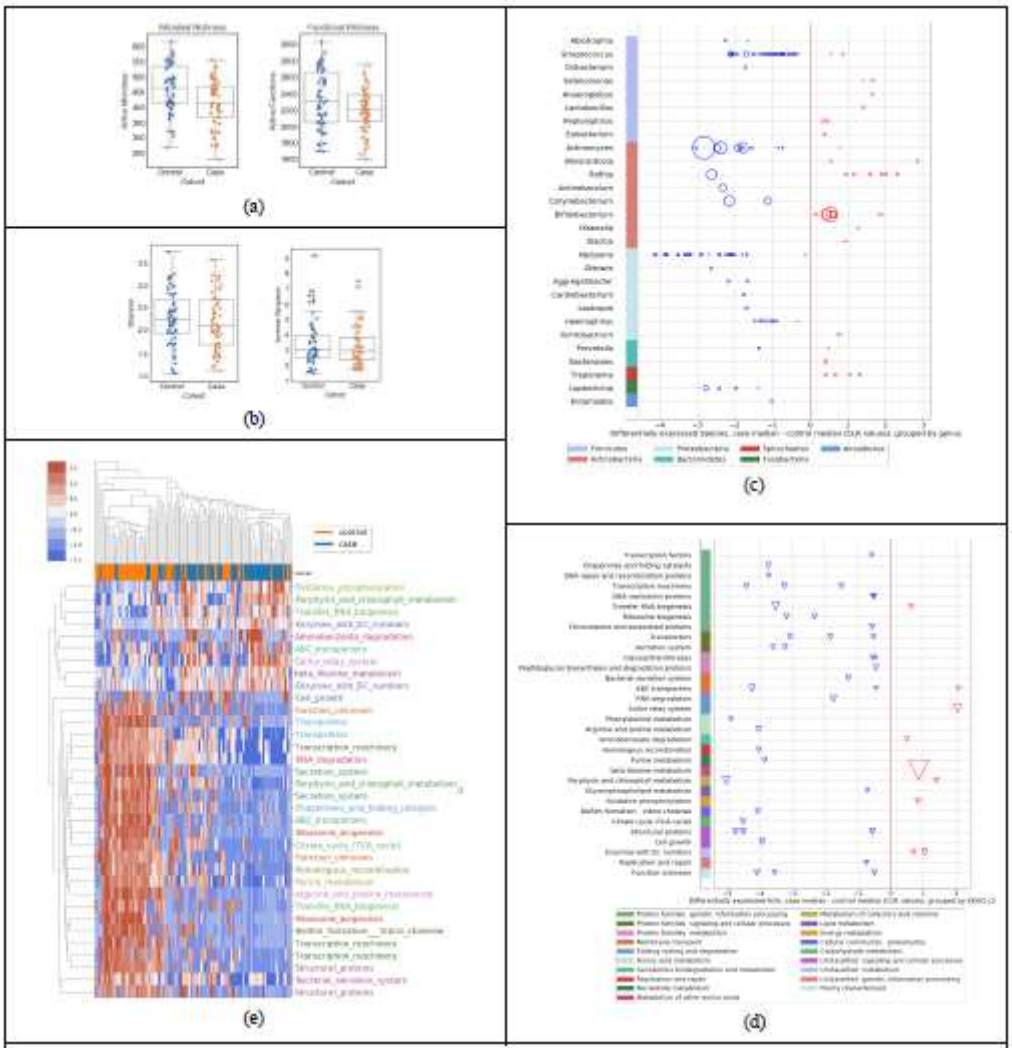


Figure 1

Descriptive statistics of salivary metatranscriptome of the high-risk population (Cohort A in Table 1). (a) Species richness; control median 463, case median 415 and function richness; control median 2306, case median 2205 (b) Shannon diversity index; control mean 2.25, case mean 2.20; and Inverse Simpson diversity index; control mean 3.41, case mean 3.26 (c) Using Mann-Whitney U tests and at least 2 fold difference in means (0.69 in CLR space), 139 differentially expressed species (at $p < 0.05$) up- or down-regulated in cases relative to controls, organized by genus and phylum (median difference in CLR values); the size of the bubble is inversely proportional to the p-value (d) Using Mann-Whitney U tests and at least 2 fold difference in means (0.69 in CLR space), 49 differentially expressed KOs (at $p<0.05$) up- or down-regulated in cases relative to controls, organized by KEGG level-3 and level-2 functional groups; the size of each triangle is inversely proportional to its p-value (e) Clustermap using Euclidean distance of CLR transformed sum(transcripts per million) data for active function (KO) features significant by Mann-Whitney U tests. Features are shown with corrected p-values < 0.01 and median CLR differences between the cohorts of greater than 0 or less than -1. KOs are color coded by their KEGG level-3 functional group.

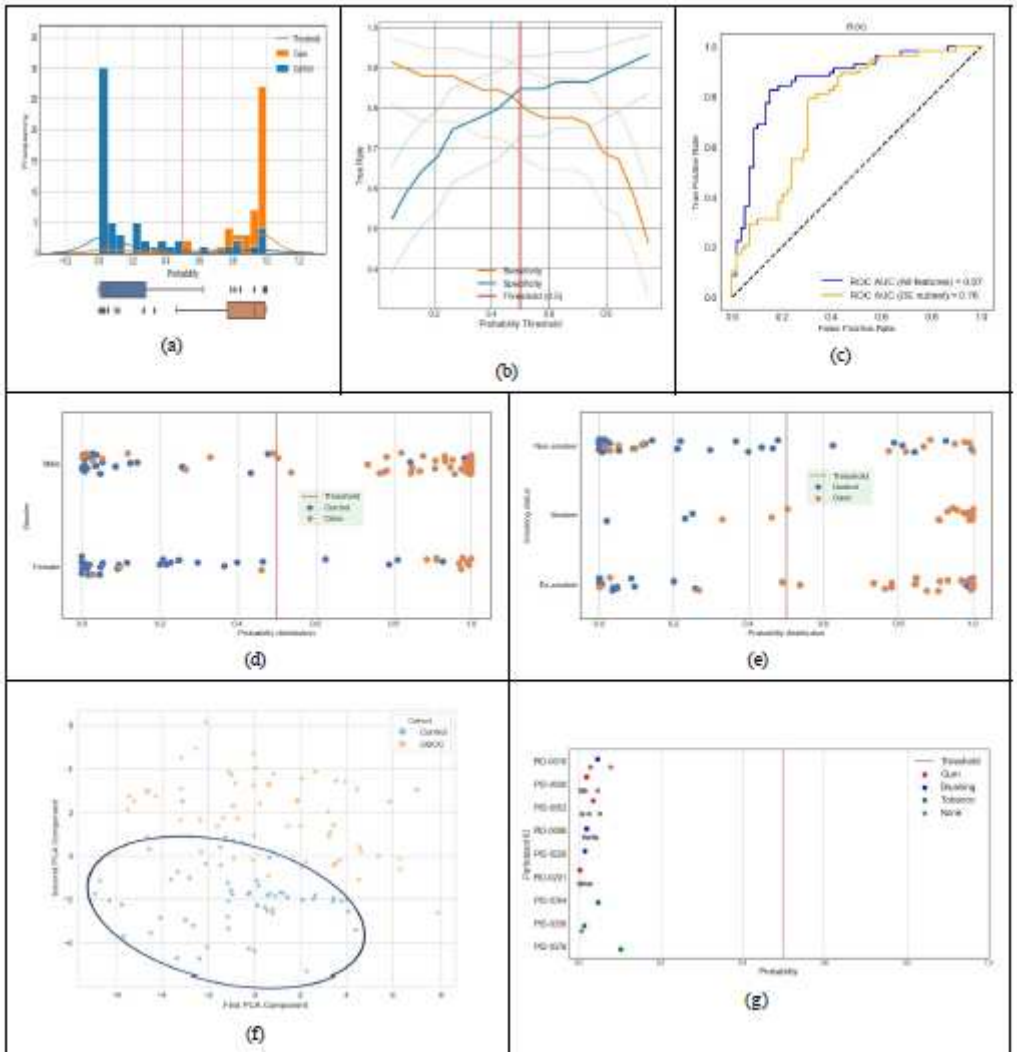


Figure 2

Predictive performance of machine-learnt classifier trained with discovery dataset (Cohort A in Table 1). (a) Distribution of classifier output probabilities across the sample set (b) Sensitivity & specificity tradeoff with 95% confidence interval computed using the Clopper-Pearson method; at the default decision boundary of 0.5, sensitivity is 0.81 and specificity is 0.85. (c) ROC AUC of the classifier using the LOOCV method is 0.87 (blue curve); using differentially expressed features only is 0.76 (orange curve) (d) Classifier probabilities separated by gender (e) Classifier probabilities separated by smoking status (e) PCA analysis using top 100 features (PC1 and PC2 capture 10.2% and 6.3% of the total variation, respectively.) (f) Probability of cancer output from the classifier for control samples with and without interference from chewing gum, chewing tobacco, and brushing teeth.

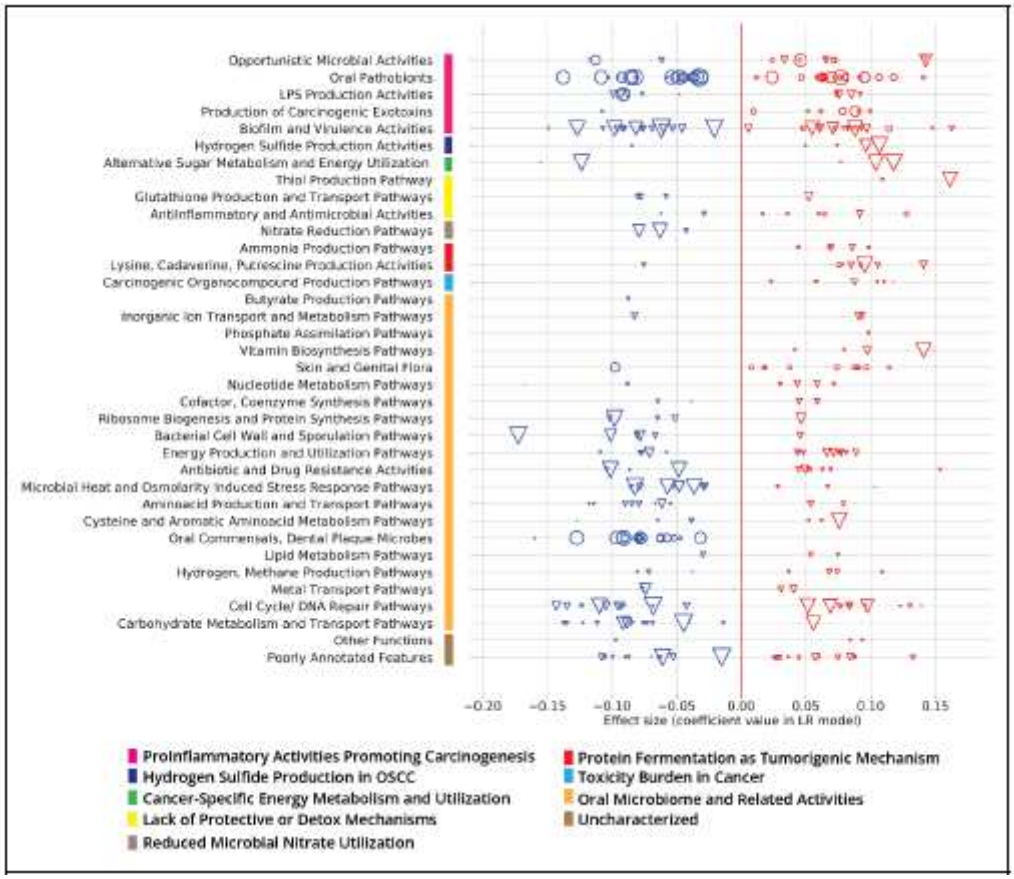


Figure 3

Oral metatranscriptomic signature from the ML classifier trained with Cohort A from Table 1: effect sizes of 101 active species (circles) and 247 active KOs (triangles), grouped into curated Viome Functional Categories (VFC, see Supplementary Material); size of circles or triangles is proportional to the CLR median difference between cases and controls

Supplementary Files

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

- FigS0.png
- FigS1.png
- FigS2.png
- FigS3.png
- FigS4.png