Integrative metatranscriptomic analysis reveals disease-specific microbiome-host interactions in oral squamous cell carcinoma

Vinay Jain[1,8], Divyashri Baraniya[1], Doaa E. El-Hadedy[1], Tsute Chen[4], Michael Slifker[5], Fadhl Alakwaa[6], Kathy Q. Cai[7], Kumaraswamy N. Chitrala[9], Christopher Fundakowski[3], Nezar N. Al-Hebshi[1,2*]

[1] Oral Microbiome Research Laboratory, Department of Oral Health Sciences, Maurice H. Kornberg School of Dentistry, Temple University, Philadelphia, PA, USA.
[2] Cancer Prevention and Control Program, Fox Chase Cancer Center, Temple University Health System, Philadelphia, USA
[3] Thomas Jefferson University, Philadelphia, PA, USA
[4] Department of Microbiology, Forsyth Institute, Cambridge, MA, USA.
[5] Biostatistics and Bioinformatics Facility, Fox Chase Cancer Center, Philadelphia, PA, USA.
[6] Department of Internal Medicine, Nephrology Division, University of Michigan, Ann Arbor, Michigan, USA.
[7] Histopathology Facility, Fox Chase Cancer Center, Philadelphia, PA, USA
[8] Low level Radiation Research Section, Radiation Biology & Health Sciences Division, Bhabha Atomic Research Centre, Mumbai, India.
[9] Fels Cancer Institute for Personalized Medicine, Lewis Katz School of Medicine, Temple University, Philadelphia, USA

List of abbreviations

OSCC: oral squamous cell carcinoma; TT: tumor tissue; ANT: adjacent normal tissue; OCT: optimal cutting temperature; LMD: laser microdissection; RIN: RNA integrity; qPCR: quantitative PCR; DEGs: differentially expressed genes; GSEA: Gene Set Enrichment Analysis; MSigDB: Molecular Signatures Database; EC: Enzyme classes; CLR: centered log-ratio; MaAsLin2: Microbiome Multivariable Associations with Linear Models; PCA: principle component analysis; FDR: false discovery rate; PERMANOVA: Permutational Multivariate Analysis of Variance; dbGaP: database of Genotypes and Phenotypes.

***Corresponding Authors**

Nezar Noor Al-Hebshi. Maurice H. Kornberg School of Dentistry, Temple University, Philadelphia, USA. Tel: +12157072091. E-mail address: alhebshi@temple.edu

Running title: Metatranscriptome of OSCC

Conflict of interest: The authors declare no potential conflicts of interest

Abstract

Studies on the microbiome of oral squamous cell carcinoma (OSCC) have been limited to 16S rRNA gene sequencing. Here, laser microdissection coupled with brute-force, deep metatranscriptome sequencing was employed to simultaneously characterize the microbiome and host transcriptomes and predict their interaction in OSCC. The analysis involved 20 HPV16/18-negative OSCC tumor/adjacent normal tissue pairs (TT and ANT) along with deep tongue scrapings from 20 matched healthy controls (HC). Standard bioinformatic tools coupled with in-house algorithms were used to map, analyze, and integrate microbial and host data. Host transcriptome analysis identified enrichment of known cancer-related gene sets, not only in TT vs. the ANT and HC, but also in the ANT vs. HC contrast, consistent with field cancerization. Microbial analysis identified a low abundance yet transcriptionally active, unique multi-kingdom microbiome in OSCC tissues predominated by bacteria and bacteriophages. HC showed a different taxonomic profile yet shared major microbial enzyme classes and pathways with TT/ANT, consistent with functional redundancy. Key taxa enriched in TT/ANT compared to HC were *Cutibacterium acnes*, *Malassezia restricta*, Human Herpes Virus 6B, and bacteriophage Yuavirus. Functionally, hyaluronate lyase was overexpressed by *C. acnes* in TT/ANT. Microbiome-host data integration revealed that OSCC-enriched taxa were associated with upregulation of proliferation-related pathways. In a preliminary *in vitro* validation experiment, infection of SCC25 oral cancer cells with *C. acnes* resulted in upregulation of MYC expression. The study provides a new insight into potential mechanisms by which the microbiome can contribute to oral carcinogenesis, which can be validated in future experimental studies.

Significance

Studies have shown that a distinct microbiome is associated with oral squamous cell carcinoma (OSCC), but how the microbiome functions within the tumor and interacts with the host cells remains unclear. By simultaneously characterizing the microbial and host transcriptomes in OSCC and control tissues, the study provides novel insights into microbiome-host interactions in OSCC which can be validated in future mechanistic studies.

Introduction

Oral squamous cell carcinoma (OSCC) is the predominant malignancy of the oral cavity with poor prognosis and a 5-year survival rate of less than 50% (1,2), resulting in more than 175,000 deaths annually (3). The tongue is the most affected subsite of the oral cavity (4). Use of various forms of tobacco and alcohol consumption are the major risk factors of OSCC, accounting for nearly 74% of cases in Western countries (5). A small fraction of OSCC cases (2-6%) will also possess high risk HPV strains though the potential causal role of HPV in OSCC has not been clearly demonstrated to the same extent as in the oropharynx (6,7). Recently, there has been increasing interest in the role of the microbiome in cancer in general including OSCC (8-11).

A plethora of studies have been carried out to characterize the microbiome associated with OSCC in a variety of samples including surface swabs, oral rinse, unstimulated saliva and tumor biopsies-- which we comprehensively reviewed elsewhere (8,9). While these studies demonstrate that OSCC-associated microbiome is significantly different from health-associated microbiome, the results do not reveal that a particular species or consortium is consistently enriched in OSCC samples across all patient cohorts. While these inconsistencies may have been a result of methodological variations among the studies, they can probably be explained by functional redundancy: the fact that different species can serve the same function within microbial communities (12). In fact, Tian et. al. (13) have recently shown that the gene composition and functional capacity of a microbiome is more conserved as compared to its taxonomic composition.

Consistently, we have recently identified different species in association with OSCC in two cohorts using 16S sequence-based compositional analysis; however, applying functional prediction analysis, we found pro-inflammatory microbial attributes to be enriched in both cohorts (14,15). Similarly, a pilot study by Yost et. al. (16) using metatranscriptomic approach (sequencing of mRNA transcripts from all organisms in a sample) revealed that OSCC-associated microbiomes have similar functional signatures despite

differences in their taxonomic composition. Together, these findings provide evidence for microbial functional redundancy and highlight the importance of functional analysis in assessing the role of the microbiome in OSCC.

Metatranscriptome analysis is one approach to study the functional activity of a microbiome (17). Compared to 16S rRNA gene sequencing, which has been the predominant microbiome analysis method so far, metatranscriptomics captures only viable, transcriptionally active species, allows identification of all types of microorganisms in the sample (bacteria, archaea, fungi and viruses) and provides higher taxonomic resolution. In addition, since samples will usually include host cells, metatranscriptomics provides an opportunity to simultaneously study the microbiome and host transcriptome and their potential interaction. To the best of our knowledge, the study by Yost. et. al. (16) is the only study so far that applied metatranscriptomics to OSCC. The study involved analysis of oral swabs collected from 4 OSCC patients and 4 healthy controls and was limited to assessment of the bacterial transcriptome; other microbial kingdoms and the host transcriptome were not evaluated.

In this first-of-kind study, we have employed laser microdissection coupled with metatranscriptome sequencing at unprecedented depth (brute-force deep sequencing) to characterize the composition and function of the multi-kingdom microbiome within OSCC tissues and its association with the transcriptional activity of the host to provide novel insights into microbiome-host interactions in OSCC.

Materials and methods

An overview of study design and workflow is given in **Figure 1**; the details are provided in the sections below. The study was approved by the Institutional Review Boards at Temple University (# 25808) and Thomas Jefferson University (#19D.270). The study was conducted in accordance with Declaration of Helsinki; a written informed consent was obtained from all prospectively recruited subjects.

Subject population and samples

Frozen OSCC tumor and adjacent normal tissue pairs (abbreviated thereafter as TT and ANT, respectively) were obtained from the Biosample Repository Facility at Fox Chase Cancer Centre and the Pathology Biorepository Shared Service at the University of Maryland, Baltimore. To minimize heterogeneity, the samples were restricted to cancer of the mobile tongue (ICD-10 code C02). Out of 50 tissue pairs initially obtained, only 20 pairs were found by histopathological evaluation to be suitable for microdissection (next section). All cases were HPV16/18-negative as confirmed by PCR, and all except one were treatment naïve (1 subject had received radiotherapy before resection). As an additional control group, deep epithelial tongue scrapings were obtained prospectively from 20 age-, race- and sex-matched healthy subjects (HC) given the following inclusion criteria: no evidence of malignancy and premalignant lesions, no signs of acute/chronic oral infections including severe gingivitis/periodontitis, no history of antibiotic/antifungal intake in the last 3 months and no history of endocarditis/valve issue, and no history of smoking. The tongue scrapings were collected from the dorsal surface, after drying with a cotton roll, using 10 heavy strokes with a 7 mm loop-type dermatological curette (Acuderm Inc, USA), which has been shown to capture sufficient samples for RNA analysis (18,19). The scrapings were immediately placed in RNAlater (ThermoFisher Scientific, USA) and stored at -80°C. Demographic details and clinical characteristics of the study subjects are shown in **Supplementary Table 1**.

Histopathological examination and laser microdissection

Frozen TT and ANT samples were embedded in optimal cutting temperature (OCT) medium and 8 µm-thick cryosections were cut using cryostat microtome. RNase free environment was maintained during all the steps. The sections were stained with hematoxylin and eosin for histopathologic evaluation and grading. Based on histopathologic review, 21-30 additional sections were cut for each tissue and placed on PEN membrane glass slides (ThermoFisher Scientific, USA) for laser microdissection (LMD). The sections were

processed and stained using Histogene Staining solution (ThermoFisher Scientific, USA) as per the manufacturer's instructions and sequentially dehydrated in 70%, 95% and 100% alcohol, before air drying for 5 minutes at room temperature. All the solutions were treated with 1x ProtectRNA^{TM} RNase inhibitor solution (Sigma Aldrich, USA) to prevent RNase contamination. LMD was performed using Leica LMD6500 gravity, contact-free collection system (Leica Microsystems, USA). The desired areas (tumor cells and adjacent normal epithelium) were carefully marked under 5x magnification and captured in RNAlater placed on the cap of 0.5 ml PCR tube. Between 3-6 sections were captured per cap and multiple tubes were used to collect tissue from each sample to minimize capture time and thus RNA degradation. The micro-dissected sections in RNAlater were stored at -80 º C until further processing. Representative images of micro-dissected tissues are shown in **Supplementary Figure 1**.

DNA and RNA extraction

DNA and RNA were extracted using AllPrep DNA/RNA Micro kit (Qiagen, USA), including a bead beating step to ensure lysis of microbial cells. Briefly, the tissue samples stored in RNAlater were thawed at 37℃, pelleted by adding equal volume of phosphate-buffered saline and spinning at 5000 rpm for 5 minutes, and resuspended in 600 µl of RLT plus lysis solution. The lysate was transferred into DNase/RNase free tubes containing 200 mg 100-micron zirconium beads (Molecular biology grade, PFMB 100-100-12, OPS diagnostics, USA), and bead beaten at 6m/s for 1 minute at room temperature using FastPrep FP100A instrument (MP Biomedicals, USA). The lysate was used to sequentially isolate DNA and RNA as per manufacturer's instructions. For RNA, in-column DNase treatment was done using RNase-Free DNase Set (Qiagen, USA). Aliquots of RNAlater were used as extraction negative control. The purity of RNA and DNA was assessed by measuring 260/280 ratio using Nanodrop (ThermoFisher Scientific, USA) and quantity was measured using Qubit RNA HS Assay Kit and Qubit dsDNA HS Assay Kit (ThermoFisher Scientific, USA), respectively. The RNA integrity (RIN) and size distribution was assessed using Agilent

RNA 6000 Pico Kit on Bioanalyzer 2100 (Agilent Technologies, CA). DNA and RNA concentrations and RIN numbers for the samples are presented in **Supplementary File 1.**

Determination of microbial kingdom loads

DNA isolated from the samples was used for determination of bacterial, archaeal and fungal loads by real-time quantitative PCR (qPCR). Universal primer pairs targeting bacteria (341F & R806), archaea (ARC344F & Arch806R) and fungi (ITS1-30F, ITS1-217R) (20,21) were used; the sequences are listed in **Supplementary Table 2**. Genomic DNA from *Haemophilus parainfluenzae* (NCTC 10665, Public Health England), *Methanobrevibacter oralis* (DSM 7256, DSMZ, Germany) and *Candida albicans (*CAI4 laboratory strain*)* was used as control for the bacterial, archaeal and fungal assays, respectively. The PCR efficiency for each primer pair was derived from the standard curve prepared with at least 5 serial dilutions of control DNA (**Supplementary Figure 2**). The PCR reaction mix (20 ul) contained 5 ng sample DNA, 1X PowerUp™ SYBR™ Green Master Mix (ThermoFisher Scientific, USA), 1 µM of each primer (for bacteria), 0.5 µM of each primer (for archaea) and 0.125 µM of each primer (for fungi). The cycling conditions were as follows: 50℃ for 2 min for UDG activation, 95℃ for 2 min for polymerase activation followed by 45 cycles of denaturation at 95℃ for 15 sec and annealing/extension at 60℃ (bacterial and archaeal primers) and 62℃ (fungal primers). All qPCR reactions were carried out in triplicates on a Quantstudio 3.0 thermal system (ThermoFisher Scientific, USA). No template, extraction and positive controls were added in each run. Abundance of each kingdom was normalized to human β-actin gene as described previously (22) .

rRNA depletion, library preparation and sequencing

Bacterial and human rRNA were depleted from total RNA using NEBNext® rRNA Depletion Kits E7850 and E7400 (New England Biolabs, USA), respectively. A cocktail of human and bacterial depletion solutions was used in the ratio of 2:1. Depleted RNA was purified using Agencourt RNA clean XP beads

(Beckman Coulter, USA) and used for preparation of RNA sequencing libraries using NEBNext® Ultra™ II Directional RNA Library Prep (New England Biolabs, USA) as per the manufacturer's protocol. The RNA fragmentation time was optimized and adjusted according to RIN no. Libraries were labelled with unique indexes for multiplexing using NEBNext® Multiplex Oligos for Illumina® (New England Biolabs, USA). The final libraries were quantified using Qubit dsDNA HS Assay Kit on Qubit 3.0 fluorimeter. The library quality and size distribution were assessed using Agilent High sensitivity DNA kit on Bioanalyzer 2100. Library concentrations are presented in **Supplementary File 1.** The 60 libraries were pooled in groups of 10 and sequenced using 2x100 CoolMPS chemistry on DNBSEQ-T7 platform (BGI, HongKong) with a target depth of 400 million paired reads per sample (brute-force deep sequencing).

Mapping of human and ribosomal sequences

Paired-end FASTQ files were quality checked using FastQC (RRID:SCR_014583 https://www.bioinformatics.babraham.ac.uk/projects/fastqc/) and MultiQC (23), and NEBNext adapter content was trimmed with Cutadapt (24). Trimmed FASTQ files were aligned separately to human reference GRCh38, the SILVA large subunit rRNA database (v138), and the human ribosomal DNA complete repeating unit U13369.1 using the STAR aligner (RRID:SCR_004463) (25). A custom script was used to identify reads aligning to various combinations of the 3 references and reads which were unmapped to any of the 3 references were saved as FASTQ files for mapping to microbial sequences (see below).

Host differential gene expression and pathway analysis

Reads aligning to GRCh38 were quantified at the gene-level using *htseq-count* [HTSeq] (mode = "union") (26). Differential expression analysis was carried out with *DESeq2* (RRID:SCR_000154) (27) for the following comparisons (each involving 40 samples): (1) TT vs. ANT (paired comparison), (2) TT vs. HC, and (3) ANT vs. HC. For each pairwise comparison, any gene with fewer than 50 total counts across the 40 samples was excluded prior to DESeq2 analysis. Subsequently, a ranked list of differentially expressed

genes (DEGS) for each comparison (ranked by Wald statistic generated by *DESeq2*) was used as input to Gene Set Enrichment Analysis (GSEA, *Preranked* method) (28) to explore whether phenotypic differences were significantly related to known functional gene sets. The hallmark gene set collections from Molecular Signatures Database (MSigDB) 7.1 was used for this purpose (29).

Microbiome data analysis

Reads not mapping to any of the three reference databases above were further processed with KneadData (30) to further remove any remaining human and rRNA sequences. KneadData-cleaned, unmapped reads were then analysed with HUMAnN 3.0 for microbial profiling (31). HUMAnN 3.0 uses MetaPhlAn 3.0 (32) for taxonomic profiling, including viruses. For functional profiling it uses Bowtie2 (33) for nucleotide-level searches against ChocoPhlAn 3 database (34) followed by translated search of yet unmapped reads against UniRef90 protein reference database (35) using DIAMOND (36). The individual sample gene lists generated are then combined and regrouped using different functional annotations. In this study we used enzyme classes (EC) and metabolic pathways based on MetaCyc database (RRID:SCR_007778) (37). Ultimately unmapped reads, and taxa/functional features present in ≤10% of the samples were excluded before the profiles were centered log-ratio (CLR) transformed (38) to account for compositionality of the data (39). MaAsLin2 (Microbiome Multivariable Associations with Linear Models) package in R (40) was employed to identify differentially abundant features between the groups (TT, ANT and HC), accounting for the paired nature of the comparison in the TT vs. ANT contrast; false discovery rate (FDR) according to Benjamini et al. (41) was set to 0.05. Principle Component Analysis (PCA) was performed using Phyloseq (42) and "microbiome" (43) R packages. Permutational Multivariate Analysis of Variance (PERMANOVA) test was performed using vegan package in R (44).

Microbiome-host data integration

Unsupervised Spearman rank correlation was performed between the host genes' expression levels (DESeq2 transformed) and CLR-transformed microbial features (genera and EC). Significant correlations were defined as those with absolute correlation coefficient (r_s) ≥ 0.6 and FDR ≤ 0.05. For each microbial feature, the correlating genes, ranked by r_s, were subjected to GSEA (28) to identify pathways potentially modulated by that feature. The results were visualized with Cytoscape Automation (45) with microbes and host pathways as nodes and enrichment scores as edges. We also employed MOFA (Multi Omics Factor Analysis) (46) as another method to integrate the host and microbiome data. MOFA reduces multiple high-dimensional data into a small number of factors that captures biological variation in the data, and then measure the contribution of omic sets to each factor, and ability of each factor to discriminate between the study subgroups.

Validation assays

For the host transcriptome, expression of 7 highly DEGs (MMP13, CA9, ROS1, KRT4, CRNN and SPRR3) was validated with quantitative, reverse-transcription PCR (qRT-PCR). The assays were performed using predesigned gene-specific primers/probe sets in customized TaqMan® Array Standard Plates (Thermofisher Scientific, USA). Specifically, EXPRESS One-Step Superscript™ qRT-PCR Kit (ThermoFisher Scientific, USA) was used in 20 µl reactions as recommended by the manufacturer, using 10 ng total RNA as template on Quantstudio 3.0 thermal cycler (Applied Biosystems, USA). The following cycling conditions were used: cDNA synthesis for 15min at 50℃, initial denaturation at 95℃ for 2min and 40 cycles of denaturation at 90℃ for 15 s and extension at 60℃ for 1min with data collection at the end of each extension step. All the genes were studied in triplicates and normalized with RPL30 as endogenous control.

16S rRNA gene-based microbial profiling

To supplement the information obtained from RNA sequencing, microbial composition was also assessed by 16S rRNA gene sequencing of the DNA extracts. Library preparation, sequencing and bioinformatics analysis were done as previously described (14).

Contamination control

Extraction negative controls were included for both RNA and 16S rRNA gene sequencing. In the former, no detectable levels of RNA were found using Qubit™ RNA High Sensitivity kit, so no further processing (i.e., library preparation) could be performed. For 16S rRNA gene sequencing, no amplifiable DNA was detected in the extraction negative controls by PCR, but they were still submitted for sequencing. We then performed manual filtration of probable contaminants based on 1) established knowledge of the microbial taxa typically found in the oral cavity (HOMD.org) 2) contaminants found in the sequenced negative control as well as those reported in the literature (as in Salter et al. 2014 (47)), and 3) comparing 16S and RNA sequencing profiles. All three points together were considered when deciding about taxa to filter out, i.e., they were not sequential filtering steps. For example, *Microbacterium* and *Sphingomonas* represented 27% and 18% of the reads in the 16S data from TT and ANT samples, respectively. These genera have been reported as negative control contaminants, are not typical members of the oral microbiome, and were not detected in the RNA sequencing data (i.e., do not represent live/transcriptionally active bacteria. Therefore, they were filtered out from the 16S profiles. In contrast, while genus *Streptococcus* has been reported as a negative control contaminant, it is also a major member of the oral microbiome, so it was not removed. Other known contaminants found in high abundance in the extraction negative controls by 16S sequencing included *Bradyrhizobium, Stenotrophomonas, Hyphomicrobium*, and *Escherichia*, and again were not observed in the RNA-based profiles. Together, the results provided a proof that the RNA sequencing data was free from DNA contamination.

Availability of data and materials

The data is made available for secondary use through the database of Genotypes and Phenotypes (dbGaP), accession number phs002678.v1.p1. However, due to consent constraints, the human sequences from the healthy controls' data were filtered out and cannot be shared for future research.

Results

Sequencing statistics

Ultra-deep sequencing of the 60 samples generated a total of ~ 23 billion 2x100 bp reads (~ 10TB of data), with an average of 398,569,093 ± 62,427,540 reads per sample. Per group, the average sequencing depth was 422,326,741 ± 70,776,305 for TT, 384,960,281 ± 55,224,182 for ANT, and 388,420,257 ± 56,083,238 for HC. At the Star aligner step, TT and ANT samples showed a similar pattern with ~97% of the reads mapping to GRCh38, ~0.7 % mapping to SILVA & U13369.1 combined, and ~2.0% remaining unmapped (**Supplementary Figure 3)**. In the HC group, on the other hand, only 23.4 % of the reads mapped to GRCh38, while 10% mapped to the rRNA databases and 67% remained unmapped. Processing the unmapped reads with kneadData removed an average of 84% and 9% of the reads from the TT/ANT and HC groups, respectively, as contaminating sequences (i.e., human and ribosomal sequences that did not meet STAR's mapping parameters). Eventually, an average of 1.3M and 250M reads from the TT/ANT and HC groups, respectively, were available as input for HUMAnN analysis. Detailed sequencing and mapping statistics are presented in **Supplementary file 2**.

Host transcriptome analysis confirms known cancer-associated genes and pathways and reveals field cancerization in tumor-adjacent normal tissue

A total of 410,397,495 (TT), 375,786,295 (ANT) and 91,631,306 (HC) reads mapped to the human reference GRCh38, which identified 54,221, 52,454 and 51,716 coding, non-coding and pseudogenes, respectively-56,962 in total. Filtration, normalization and differential gene expression analysis was carried using DESeq2 in three contrasts: TT vs. ANT, TT vs. HC, and ANT vs. HC. The output of these pairwise analyses, including fold-change, FDR, and Wald statistic are presented in **Supplementary File 3**. Applying cutoffs of ≥ 2.0-fold change and FDR ≤ 0.05 identified 8,640 DEGs in TT vs. ANT (5,186 upregulated and 3,454 downregulated; **Figure 2A**), 17,991 DEGs in TT vs. HC (10,400 upregulated and 7,591 downregulated;

Figure 2B) and 15,375 DEGs in ANT vs. HC (8,436 upregulated & 6,939 downregulated; **Figure 2C**). Principal component analysis (PCA) based on the top 1,500 most variable DEGs showed clear separation between the three groups (**Figure 2D**). Results of the validation qRT-PCR assays for six DEGs were consistent with those obtained with RNA-seq, although the latter tended to under-estimate the magnitude of fold change (**Supplementary Table 3).**

The details of GSEA results for all three contrasts are provided in **Supplementary File 4.** In the TT vs. ANT contrast, *INHBA* (Inhibin Subunit Beta A) stood out as the most significantly upregulated gene and contributed to enrichment of the epithelial-to-mesenchymal transition (EMT), inflammatory response and KRAS pathways (**Figure 2E)**. Other highly upregulated genes as part of these pathways were matrix metalloproteinase, collagen and growth factor genes (**Table 1**). Interferon alpha and gamma responses, IL-6/JAK/stat signaling, and angiogenesis were also upregulated in TT vs. ANT while oxidative phosphorylation, fatty acid metabolism, adipogenesis and P53 pathway were downregulated (**Figure 2E)**. The key genes involved in each of these pathways are presented in **Table 1**. Apart from the hallmark gene sets, it is worth mentioning that HOX genes belonging to all the 4 clusters (A, B, C, D), and the corresponding long non-coding RNAs (lncRNAs) were also highly upregulated in the TT vs. ANT contrast. Interestingly, proliferation-related gene sets E2F targets, MYC targets and G2M checkpoint were downregulated in the TT vs. ANT contrast. In contrast, they were the top upregulated pathways in the TT vs. HC as well as the ANT vs. HC comparisons (**Figure 2F &G)**. Largely, the same set of genes were involved in both comparisons (**Table 1**). EMT pathway was also upregulated in the two contrasts but didn't involve any MMPs as in the TT vs. ANT contrast. Instead, mainly genes encoding extracellular matrix proteins and adhesion molecules were upregulated (**Table 1**). In the other direction, genes involved in heme metabolism, P53 pathway and apoptosis were downregulated. Counterintuitively, TNF-alpha signaling via NFKB and inflammatory response pathway were also downregulated. Overall, the results from the ANT vs. HC contrast demonstrate presence of oncogenic changes in the ANT consistent with field cancerization.

As secondary analysis, we also compared within the OSCC cases between samples with and without lymph nodes involvement. At an FDR cutoff of ≤ 0.05, only 227 DEGs were identified (50 at fold change ≥ 2.0); using a more lenient cutoff (0.2) increased the number to 1,311 DEGs (**Supplementary File 5)**. At the gene level, *FDCSP* (cancer cell migration and invasion), *KIR2DL3* (immune response), *SLC30A10* (anti-apoptotic), *MAB21L2*, *PCDH9* (cell adhesion), *PEG3* (cell proliferation) were found to be highly expressed in the LN-positive samples. At the pathway level, MYC targets, E2F targets, MTORC1 and KRAS signaling were significantly enriched in LN-positive group (**Supplementary File 5, & Supplementary Figure 4)**.

Low abundance yet unique, transcriptionally active multi-kingdom microbiome in OSCC tissues

Loads of bacteria, fungi and archaea normalized to human beta-actin gene were determined by qPCR analysis of DNA extracted from the same cells as for RNA. Bacterial and fungal DNA were detected in all the samples; however, archaeal DNA was only detected in the HC (tongue scraping) group (**Figure 3A**). As expected, the microbial loads were far higher in the HC (~10^4 fold and ~15 fold for bacteria and fungi respectively). The abundance of bacterial DNA in adjacent normal tissue was marginally but significantly higher than tumor ($p \leq 0.001$); a similar trend was seen for fungal DNA, but the differences were not statistically significant.

Despite low abundance, analysis of non-human, non-ribosomal RNA sequences with HUMAnN 3.0 pipeline identified a transcriptionally active, multi-kingdom microbiome in the samples. In the taxonomic profiling step by MetaPhlAn 3.0, a total 14 phyla, 165 genera and 483 species were identified. The relative transcriptional abundances of each of these taxa in individual samples are presented in **Supplementary File 6.** The average taxonomic profiles based on the 20 most abundant phyla and genera in each group are presented in **Figures 3B & C;** the corresponding species-level profiles are presented in **Supplementary Figure 5**. In the HC samples, the microbial transcriptome was dominated by the bacterial phyla Firmicutes,

Proteobacteria, Bacteroidetes, Fusobacteria and Actinobacteria (in this order of abundance). The TT and ANT groups had similar profiles with Actinobacteria being the most transcriptionally abundant, followed by Firmicutes and viruses; the latter accounted for a significant proportion of the transcripts (12.2 % and 20.4 %, respectively, compared to 2.3% in the HC samples). Fungal transcripts were identified in all three groups at very low abundance. At the genus level, *Gemella* was the most transcriptionally abundant in the HC group, followed with other typical oral genera including *Prevotella*, *Neisseria*, *Campylobacter*, *Streptococcus*, *Fusobacterium* and *Veillonella*. On the other hand, while *Streptococcus*, *Gemella* and *Neisseria* were also among the top transcriptionally abundant taxa in the TT and ANT groups, *Cutibacterium* (predominantly *Cutibacterium acnes*) was the most abundant accounting for more than 20% of the transcripts on average (compared to < 1% abundance in the HC). Also, *Cloacibacterium* (predominantly *Cloacibacterium normanens*) and bacteriophage Siphoviridae (including Yuavirus) were among the most abundant taxa. Yuavirus was predominantly represented by bacteriophage alpha proteobacterium JL001. In PCA analysis (**Figure 3D**), the HC group clustered separately from the TT and ANT groups (P=0.001, PERMANOVA); however, the latter two groups did not differ and showed significant dispersion.

Following taxonomic profiling, the sequences were mapped to ChocoPhlAn 3 and UniRef90 databases which identified around an average of 23,556, 23,845 and 98,569 genes per sample in the TT, ANT, and HC groups, respectively. Rarefaction analysis of the number of microbial genes detected in each sample as a function of number of mapped reads (**Supplementary Figure 6)** revealed that all samples reached saturation with a Good's Coverage Index of > 99%. The gene lists were then regrouped and functionally annotated using MetaCyc database. The individual sample enzyme class (EC) and metabolic pathway profiles are presented in **Supplementary File 7.** The average functional profiles based on the top 20 ECs and top 20 pathways in each group are presented in **Figure 3E** and **Supplementary Figure 7**, respectively. Despite major differences in taxonomic profiles, the most abundant ECs and pathways were common to all 3 groups consistent with functional redundancy. Abundant ECs included those involved in DNA replication

and transcription (DNA polymerase, DNA helicase and RNA polymerase), response to oxidative stress (superoxide dismutase, peroxiredoxin and thioredoxin-disulfide reductase), and metabolism (e.g. adolase, phosphoglycerate kinase, dihydrolipoyl dehydrogenase and pyruvate kinase). At the pathway level, glycolysis, biosynthesis of nucleotides and biosynthesis of peptidoglycan were the dominant pathways. PCA analysis by ECs did not show separate clusters between the three groups, but the HC group formed a compact sub-cluster with little dispersion (**Supplementary Figure 8**).

For comparison and validation, we also performed 16S profiling on DNA extracted simultaneously with the RNA. Statistically, the 16S and RNA-seq bacterial taxonomic profiles were overall significantly different (P=01, PERNAMOVA) with the RNA-seq profiles showing more dispersion (**Supplementary Figure 9)**. Notably, *Gemella*, which was the most transcriptionally abundant genus in the HC group and third most abundant in the TT and ANT groups, did not show up among the 20 top abundant genera by 16S sequencing. Conversely, *Corynebacterium* was among the top genera in the TT/ANT samples but accounted for less than 0.5% of the transcripts in the RNA-seq data. Nevertheless, there were consistencies between the two methods, e.g., *Cutibacterium* (formerly *Propionibacterium*) was the most abundant genus in the TT and ANT groups in both methods.

Cutibacterium acnes, Malassezia restricta, Human Herpes Virus 6B, Nupapillomavirus, bacteriophages and hyaluronate lyase are key features enriched in OSSC tissues

Pairwise (TT vs. ANT, TT vs. HC, and ANT vs. HC) differential abundance analysis was performed with MaAsLin2 on CLR-transformed taxonomic (genus and species level) and functional profiles. The full results of this analysis in the form of lists of taxa and functional features and the corresponding coefficients and FDR values are provided in **Supplementary Files 8 & 9.** No significant differences in microbial profiles were found between the TT and ANT groups; however, the differences were dramatic for the TT vs. HC and ANT vs. HC contrasts, and for the most part, were similar between the two comparisons as seen in

Figure 4 (top differentially abundant genera and ECs) and **Supplementary Figure 10** (top differentially abundant species and pathways). Thirty-six bacterial genera that are commonly found in the oral cavity were transcriptionally more abundant in the tongue scrapings (HC group), including *Gemella*, *Prevotella*, *Neisseria*, *Campylobacter*, *Fusobacterium*, *Veillonella*, *Streptococcus*, *Haemophilus*, *Capnocytophaga*, *Tannerella*, *Actinomyces*, *Rothia* and *Porphyromonas* but *Stomatobaculum* was the most differentially abundant (**Figure 4A**). In the TT and ANT tissues, however, less common/typical oral bacteria were transcriptionally enriched including *Chlamydia*, *Moraxella*, *Enhydrobacter*, *Claocibacterium*, *Acinetobacter* and *Cutibacterium*. Of these, *Cutibacterium* (predominantly *C. acnes*) was the most abundant and thus chosen for validation by qPCR (**Supplementary Table 4**) which showed consistent results (**Supplementary Figure 11**). Besides bacteria, the fungus *Malassezia restricta* and several viruses were also transcriptionally more abundant in the TT and ANT groups. Enriched viruses can be grouped into human viruses (Roseolovirus represented by Human Herpes Virus 6B and Nupapillomavirus), bacteriophages (predominantly Siphoviridae, genus Yuavirus, species alpha proteobacterium JL001), plant viruses (e.g. Bromovirus) and retroviridae (**Supplementary Table 5**). The latter group was detected in very low abundance and included mainly Avian Endogenous Retrovirus EAV-HP.

Hierarchical clustering of samples by top differentially expressed ECs and pathways is presented in **Figure 4B and Supplementary Figure 10B**, respectively. Based on ECs, the analysis resulted in a cluster with all HC samples and two clusters with mixed TT and ANT samples, with the smaller of the two being closer to the HC cluster – roughly similar clusters were seen at the pathway level. Regardless of clustering, ECs that were overexpressed in most TT/ANT samples include methylmalonyl-CoA decarboxylase, trehalose-phosphatase, dimethyl-sulfide monooxygenase, malate synthase, triacylglycerol lipase, endoglycosylceramidase, proteasome endopeptidase complex, formimidoylglutamate deiminase and hyaluronate lyase (HL). Of these, we found the latter to be of potential relevance as it has hyaluronic acid degrading properties and can contribute to extracellular matrix breakdown and consequently facilitate tumor

invasion. Based on HUMAnN results, HL was exclusively contributed by *C. acnes* in the TT/ANT samples (**Figure 5**).

OSSC-associated microbial taxa potentially modulates host proliferation pathways

To predict potential microbiome-host interactions, we performed unsupervised correlation analysis between the microbial features and host genes and then, for each feature, performed GSEA analysis on significantly correlating genes to identify host pathways potentially modulated by that feature. The detailed outputs from these analyses are included in **Supplementary files 10-12**. Only OSCC-associated genera, including roseolovirus, *Cutibacterium*, retroviridae, *Chlamydia*, *Dermococcus*, Yuavirus showed substantial correlations (>3000 genes each) and resulted in significant gene set enrichment (**Figure 6A & B**). All these taxa showed association with upregulation of proliferation-related gene sets E2F targets and G2M checkpoint. Roseolovirus and *Cutibacterium* were also correlated with upregulation of MYC targets. As examples, the 10 most positively correlated and 10 most negatively correlated MsigDB genes with Roseolovirus, Cutibacterium and Yuavirus are presented in **Figure 6C-E**. Functionally, 3 ECs, namely guanosine phosphorylase, terephthalate 1,2-dioxygenase and nitrate reductase (NADH) were also associated with upregulation of proliferation-related pathways **Supplementary Figure 12A & B.**

Data integration with MOFA reduced the variation in host and microbiome data to 7 factors, of which two factors showed significant differences between the three groups **Supplementary Figure 12C & D**. Factor 1 accounted for ~ 20% of the variation and was equally contributed to by the host and microbiome; the differences between the groups were consistent those presented in Figures 3 and 4. However, Factor 2 was exclusively contributed to by the host and revealed similarities between TT and HC.

Cutibacterium acnes upregulates MYC expression in SCC25 oral cancer cells

To assess whether the observed correlations based on integrative data analysis could represent actual microbe-host cell interactions, we performed a preliminary *in vitro* validation experiment focusing on *C. acnes* and MYC gene. This pair was selected because 1) *C. acnes* was the bacteria with the highest number of significant correlations (see **Figure 6A**); 2) MYC is a key driver oncogene and was enriched as part of two of the pathways that showed association with *C. acnes*; 3) There was a strong correlation between the two (**Figure 7A**). The experiment was performed as previously described for other species (48). Briefly, *C. acnes* NCTC 737 (ATCC, USA) grown to mid-log phase was used to infect SCC25 cells (RRID: CVCL_1682, ATCC, USA) at multiplicity of infection (MOI) of 50, 100 or 200 for 24 hours, before the bacteria were washed and the cells used for RNA extraction. Measurement of MYC mRNA levels normalized to GADPH mRNA was performed using one-step q-RT-PCR. As shown in **Figure 7B**, infection with *C. acnes* resulted in upregulation of MYC expression by 1.25-1.5-fold, which was statistically significant at MOI of 200.

Discussion

Using ultra-deep metatranscriptomic analysis of micro-dissected cancerous and adjacent normal epithelium, we identified a low abundance, yet transcriptionally active, intra-tumoral multi-kingdom microbiome in OSCC. Laser microdissection has been widely used in OSCC host transcriptome studies, but, to our knowledge, this is the first time it is employed to study the microbiome associated with oral cancer. For global gene expression analysis, Illumina recommends a sequencing depth of 30–60 million reads per sample. In anticipation that microbial sequences would be concealed by the highly abundant host transcripts, we performed sequencing at unprecedented 400 million paired end reads/sample (brute-force deep sequencing) which enabled us to capture the microbial transcriptome that indeed turned out to be present in very low abundance as confirmed by q-PCR. Similarly, while adjacent normal tissue is an ideal control for analysis of the host transcriptome, we thought it may not be for that of the microbial metatranscriptome since the microbiome in normal tissue may be a continuum of that in the cancerous tissue. Therefore, we

also included tongue scrapings from matched healthy subjects as an additional control group; a dermatological curette was used to ensure deep epithelium sample is collected to make is as comparable as possible to the tumor samples.

Microbiome profiling identified novel findings including enrichment of *C. acnes, M. restricta, Human Herpes Virus 6B, Nupapillomavirus,* and *bacteriophages* in TT and ANT vs. HC. *C. acnes* was the most transcriptionally abundant species in the TT/ANT groups (~ 25 times higher than in the HC group). While *C. acnes* has not been implicated in oral cancer before, several studies have found it to be associated with prostate cancer (49-54). *C. acnes* is believed to contribute to prostate carcinogenesis through inducing chronic inflammation (55,56), so it may play a similar role in OSCC. In this study, we found the enzyme hyaluronate lyase (HL) to be exclusively expressed by *C. acnes* and to be significantly overexpressed in TT and ANT. HL degrades hyaluronic acid, an important component of the extracellular matrix of connective tissues. Two HLs have been characterized in *C. acnes* (57). In *Streptococcus pneumoniae*, HL is a known virulence factor involved in the spread of infection (58). Therefore, it is reasonable to hypothesize that *C. acnes* may contribute to tissue break down and thus invasion by cancer cells in OSCC via production of HLs. Further studies are required to test this hypothesis.

Other bacterial taxa less typically found in the oral cavity were associated with OSCC including known pathogens (*Moraxella catarrhalis*, and *Acinetobacter junii*) and species found typically in the skin (e.g *Enhydrobacter Aerococcus*) or in the gut (*Cloacibacterium normanense*); while it is not clear these may play a role in OSCC, some of these species (or sister species) were found to be enriched in colorectal cancer (59). Contrary to the literature, *Fusobacterium* was not found to be associated with OSCC in our data set, with the relative abundance being significantly higher in the tongue scrapings vs. the OSCC tissues; however, *Fusobacterium nucleatum* did tend to be higher in the TT vs. ANT groups (P=0.1).

Viral transcripts were also enriched in the TT/ANT samples, mostly bacteriophages belonging to genus *Yuavirus* and family *Siphoviridae*. Interestingly, *Siphoviridae* have been found to be the most abundant viruses associated with colorectal cancer (60). However, their role in cancer remains not known and merits further investigation. Apart from bacteriophages, human herpesvirus 6 (HHV6) was transcriptionally more abundant, actually exclusively found, in TT and ANT, which is not entirely novel, since HHV6 has been identified in association with several types of cancer, including OSCC (61). However, unlike other herpes viruses, such as EBV and HHV8, there is no direct evidence on carcinogenicity of HHV6 (61); it is hypothesized that HHP6 may have a contributory rather than direct oncogenic role (61). Nupapilloma virus was also significantly associated with TT/ANT. This virus is represented by one species, Nupapilloma virus 1 or HPV41 (https://www.hpvcenter.se/human_reference_clones/) which has been detected in some skin carcinomas and premalignant keratosis (62), but has never been implicated in oral cancer. Finally, a small number of sequences aligned to retroviridae primarily Avian endogenous retrovirus EAV HP, which was more abundant in the TT and ANT samples. This particular species shares sequence homology with another group of viruses, Avian Leukosis Virus Subgroup J, which are known to cause diverse avian tumors (63,64). Notably, human homolog of above virus, Human endogenous retroviruses (HERVs) are also strongly correlated with progression of multiple tumors including HNSCC (65,66). However, possible role of avian retroviruses in human tumor samples is not known.

In a previous study using ITS sequencing, *Candida albicans* was identified as the dominant species of a dysbiotic mycobiome associated with OSCC, while *Malassezia restricta* was found to be associated with health *(67)*.Similarly, a recent study on salivary mycobiome found a correlation between better overall survival and genus *Malassezia* abundance in OSCC patients (68). In contrast, in this study *C. albicans* was identified in only a single sample while *Malassezia restricta* was identified frequently and was transcriptionally more abundant in the OSCC samples. One possible explanation for this apparent contradiction is that previous studies were amplicon-based, i.e., the species identified may have not been

transcriptionally active. Indeed, in line with our findings, there is emerging evidence implicating *Malassezia* in inflammatory bowel disease as well as colorectal and pancreatic cancers (69). Consequently, the role of *Malassezia* in OSCC is worth further investigation.

In order to have a direct comparison between amplicon and RNA-seq based profiles, we performed 16S RNA gene sequencing on DNA obtained from the same cells on which metatranscriptomics was carried out. While largely the same major taxa were identified by both methods, the relative abundances/rank of these taxa varied between the two methods. For *Gemella* and *Corynebacterium*, the difference was drastic. while Gemella was the most transcriptionally active genus in HC and also among the top taxa in TT/ANT, it did not feature even in top 20 most abundant genera by 16s RNA sequencing. Conversely, *Corynebacterium* was among the most abundant genera in TT/ANT by 16S sequencing but found in very low abundance in the RNA-seq data. These finding demonstrate that more abundant genera may not necessarily be transcriptionally active and vice-versa.

In addition to taxonomic profiling, we also obtained functional profiles in terms of enzyme classes and metabolic pathways, for which we made a few important observations. First, despite differences in taxonomic profile, the major functional features were largely similar across the three groups, which substantiates evidence for microbial functional redundancy (13). Second, the top abundant enzyme classes and pathways were related to DNA replication and transcription, response to oxidative stress and metabolism, supporting presence of a viable and transcriptionally active microbiome. Thirdly, despite similarity in major functional groups, there were still significant difference between the TT/ANT and HC groups, including potentially relevant feature to OSCC such as HL as discussed above.

The host transcriptome in OSCC is well characterized as it has been comprehensively analyzed in several studies based on microarray and RNA-seq data sets available from the Cancer Genome Atlas (TCGA) and

Gene Expression Omnibus (GEO) (70,71). A detailed comparison with results from those studies here is not feasible and largely out of the scope of the paper. However, there are a few points to make. Overall, our results were consistent with the literature. For example, out of the top 25 upregulated genes and top 25 downregulated in head and neck cancer as per the TCGA project (lists available from the University of Alabama at Birmingham Cancer data analysis Portal; UALCAN (72)), 49 genes were also differentially expressed in the same direction in our data. Similarly, most of the protein coding genes, lncRNAs and hub genes identified as master regulators and potential biomarkers of OSCC in recent cross-database studies (70,73), were consistently up- or downregulated in our data.

A unique aspect of our study is that we also included epithelial tongue scrapings as matched controls from healthy individuals which allowed us to make 3 pairwise comparisons. While many of the DEGs identified in HC vs. TT and ANT may not be related to the cancer process (since the samples are coming from different subjects), GSEA showed that these DEGs were enriched in several cell proliferation and cancer progression associated pathways such as E2F targets, G2M checkpoint, epithelial mesenchymal transition, angiogenesis, DNA repair pathways, not only in the TT vs. HC contrast but also in the ANT vs. HC contrast. The latter is interesting and novel in that it indicates presence of oncogenic changes in normal adjacent epithelial tissue collected even they are not evident by histopathology evaluation, which is consistent with field cancerization (74). Understanding these potentially early oncogenic processes may have important implication for treatment of OSCC and prevention of its recurrence. Another unique aspect of our host transcriptome data is the unprecedented depth at which the samples were sequenced, which provides an opportunity for secondary analysis to identify rare transcripts and splice variant that could be playing a key role in oral carcinogenesis.

Finally, we performed integration of the microbiome and host transcriptome data to predict cancer-related host genes/pathways that are potentially modulated by the microbes. Given there were as many upregulated

genes and microbial taxa in the HC group as there was in the TT/ANT group, one would expect, based on pure statistical associations, to see more or less equal number of gene-microbe correlations for health-associated and OSCC-associated taxa. However, we observed far more significant correlations for OSCC-associated taxa. Furthermore, only genes correlating with OSCC-associated taxa resulted in significant pathway enrichment. Together, these observations indicate the correlations identified represent potential biological interaction, not just statistical associations. Several OSCC-associated taxa, including *Cutibacterium*, Yuavirus and Roseolovirus, showed significant correlations with more than 3,000 genes each, many of which belonged to the E2F targets, MYC targets and G2M check point gene sets suggesting these taxa may contribute to carcinogenesis through interaction with proliferation pathways. Since the results based on sequencing data are highly correlative and may not necessarily reflect actual biological interactions, we performed a preliminary *in vitro* validation study in which we showed that infection by *C. acnes* upregulated expression of the oncogene MYC in SCC25 oral cancer cells, suggesting that at least some of the observed correlations are biologically valid. We are currently developing a prioritization algorithm to identify microbe-gene candidates for further validation experiments.

The study has limitations to note. First, the sample size is small, so any generalization must be done with caution. Second, given the nature of the study (i.e., analysis of sequencing data), the results are purely correlative and should be viewed only as hypothesis-generating. Third, while no RNA was detected in negative extraction control, it should have still been included in library preparation and sequencing as done with 16S analysis, to provide additional contamination control. Additionally, the study would have benefited from also including a positive control (e.g. RNA/DNA extracted from a human cell line infected with a mock community). A fourth limitation is that although the controls were matched to the cases with respect to tumor site, age, sex and ethnicity, the two groups differed in terms of lifestyle factors (namely tobacco use and alcohol consumption) which may have confounded the results. Finally, while RNA-seq has the

advantage of studying the microbial transcriptional activity, it is limited to the expressed sequences and thus does not provide information about the full composition of the microbiome in the samples.

In conclusion, to the best of our knowledge, this is the first OSCC metatranscriptomic study where the host and microbiome transcriptomes are studied simultaneously. On the host side, the study did not only confirm known oral cancer-associated genes and pathways, but also provided evidence for field cancerization by showing oncogenic changes in the adjacent normal tissue. These genes can be used as potential diagnostic markers at early stages of carcinogenesis. On the microbial side, we identified a low abundance yet unique, transcriptionally active multi-kingdom microbiome in OSCC tissues. No differences in microbiome composition between tumor/normal pairs; but marked differences compared to healthy controls. Nevertheless, the major functional features were similar across the three groups (functional redundancy). *Cutibacterium acnes* along with its enzyme hyaluronate lyase in addition to *Malassezia restricta*, Human Herpes Virus 6B, Nupapilloma virus, bacteriophages were key features enriched in OSSC tissues and showed potential interactions with the host transcriptome through proliferation-related pathways, which requires further validation in future mechanistic studies. Overall, this work provides novel insights into microbiome-host interaction in OSCC and opens new avenues for future microbiome research.

Acknowledgements

This study was supported by the National Institute of Dental and Craniofacial Research (Grant R03DE028987). The publication fee of this manuscript has been covered by Dr. Cary R. Klimen Oral Health Sciences Research Program Fund.

References

1. Sklenicka S, Gardiner S, Dierks EJ, Potter BE, Bell RB. Survival analysis and risk factors for recurrence in oral squamous cell carcinoma: does surgical salvage affect outcome? J Oral Maxillofac Surg **2010**;68:1270-5
2. Wang B, Zhang S, Yue K, Wang XD. The recurrence and survival of oral squamous cell carcinoma: a report of 275 cases. Chin J Cancer **2013**;32:614-8
3. Sung H, Ferlay J, Siegel RL, Laversanne M, Soerjomataram I, Jemal A*, et al.* Global cancer statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA Cancer J Clin **2021**
4. Siegel RL, Miller KD, Jemal A. Cancer Statistics, 2017. CA Cancer J Clin **2017**;67:7-30
5. Petersen PE. Oral cancer prevention and control--the approach of the World Health Organization. Oral Oncol **2009**;45:454-60
6. Emmett S, Jenkins G, Boros S, Whiteman DC, Panizza B, Antonsson A. Low prevalence of human papillomavirus in oral cavity squamous cell carcinoma in Queensland, Australia. ANZ J Surg **2017**;87:714-9
7. Lingen MW, Xiao W, Schmitt A, Jiang B, Pickard R, Kreinbrink P*, et al.* Low etiologic fraction for high-risk human papillomavirus in oral cavity squamous cell carcinomas. Oral Oncol **2013**;49:1-8
8. Al-Hebshi NN, Borgnakke WS, Johnson NW. The Microbiome of Oral Squamous Cell Carcinomas: a Functional Perspective. Current Oral Health Reports **2019**;6:145-60
9. Perera M, Al-Hebshi NN, Speicher DJ, Perera I, Johnson NW. Emerging role of bacteria in oral carcinogenesis: a review with special reference to perio-pathogenic bacteria. J Oral Microbiol **2016**;8:32762
10. Oliva M, Mulet-Margalef N, Ochoa-De-Olza M, Napoli S, Mas J, Laquente B*, et al.* Tumor-Associated Microbiome: Where Do We Stand? Int J Mol Sci **2021**;22
11. Nejman D, Livyatan I, Fuks G, Gavert N, Zwang Y, Geller LT*, et al.* The human tumor microbiome is composed of tumor type-specific intracellular bacteria. Science **2020**;368:973-+
12. Tian L, Wu A-K, Friedman J, Waldor MK, Weiss ST, Liu Y-Y. Deciphering Functional Redundancy in the Human Microbiome. bioRxiv **2017**
13. Tian L, Wang XW, Wu AK, Fan YH, Friedman J, Dahlin A*, et al.* Deciphering functional redundancy in the human microbiome. Nat Commun **2020**;11
14. Al-Hebshi NN, Nasher AT, Maryoud MY, Homeida HE, Chen T, Idris AM*, et al.* Inflammatory bacteriome featuring Fusobacterium nucleatum and Pseudomonas aeruginosa identified in association with oral squamous cell carcinoma. Sci Rep **2017**;7:1834
15. Perera M, Al-hebshi NN, Perera I, Ipe D, Ulett GC, Speicher DJ*, et al.* Inflammatory Bacteriome and Oral Squamous Cell Carcinoma. Journal of Dental Research **2018**;in press
16. Yost S, Stashenko P, Choi Y, Kukuruzinska M, Genco CA, Salama A*, et al.* Increased virulence of the oral microbiome in oral squamous cell carcinoma revealed by metatranscriptome analyses. Int J Oral Sci **2018**;10:32
17. Shakya M, Lo CC, Chain PSG. Advances and Challenges in Metatranscriptomic Analysis. Front Genet **2019**;10:904
18. Prasad G, Seers C, Reynolds E, McCullough MJ. The assessment of the robustness of microRNAs from oral cytological scrapings. J Oral Pathol Med **2017**;46:359-64
19. Reboiras-Lopez MD, Perez-Sayans M, Somoza-Martin JM, Gayoso-Diz P, Barros-Angueira F, Gandara-Rey JM*, et al.* Comparison of the Cytobrush (R), dermatological curette and oral CDx (R) brush test as methods for obtaining samples of RNA for molecular analysis of oral cytology. Cytopathology **2012**;23:192-7
20. Usyk M, Zolnik CP, Patel H, Levi MH, Burk RD. Novel ITS1 Fungal Primers for Characterization of the Mycobiome. Msphere **2017**;2
21. Takahashi S, Tomita J, Nishioka K, Hisada T, Nishijima M. Development of a prokaryotic universal primer for simultaneous analysis of Bacteria and Archaea using next-generation sequencing. PLoS One **2014**;9:e105592
22. Pfaffl MW, Horgan GW, Dempfle L. Relative expression software tool (REST (c)) for group-wise comparison and statistical analysis of relative expression results in real-time PCR. Nucleic Acids Research **2002**;30

23. Ewels P, Magnusson M, Lundin S, Kaller M. MultiQC: summarize analysis results for multiple tools and samples in a single report. Bioinformatics **2016**;32:3047-8
24. Martin M. Cutadapt removes adapter sequences from high-throughput sequencing reads.
25. Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S*, et al.* STAR: ultrafast universal RNA-seq aligner. Bioinformatics **2013**;29:15-21
26. G Putri SA, PT Pyl, JE Pimanda, F Zanini. Analysing high-throughput sequencing data in Python with HTSeq 2.0. Bioinformatics **2022**
27. Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol **2014**;15:550
28. Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA*, et al.* Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. Proc Natl Acad Sci U S A **2005**;102:15545-50
29. Liberzon A, Birger C, Thorvaldsdottir H, Ghandi M, Mesirov JP, Tamayo P. The Molecular Signatures Database (MSigDB) hallmark gene set collection. Cell Syst **2015**;1:417-25
30. KneadData. <https://github.com/biobakery/kneaddata>.
31. Beghini F, McIver LJ, Blanco-Miguez A, Dubois L, Asnicar F, Maharjan S*, et al.* Integrating taxonomic, functional, and strain-level profiling of diverse microbial communities with bioBakery 3. Elife **2021**;10
32. Segata N, Waldron L, Ballarini A, Narasimhan V, Jousson O, Huttenhower C. Metagenomic microbial community profiling using unique clade-specific marker genes. Nature Methods **2012**;9:811-4
33. Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. Nat Methods **2012**;9:357-9
34. Franzosa EA, McIver LJ, Rahnavard G, Thompson LR, Schirmer M, Weingart G*, et al.* Species-level functional profiling of metagenomes and metatranscriptomes. Nat Methods **2018**;15:962-8
35. Suzek BE, Wang Y, Huang H, McGarvey PB, Wu CH, UniProt C. UniRef clusters: a comprehensive and scalable alternative for improving sequence similarity searches. Bioinformatics **2015**;31:926-32
36. Buchfink B, Xie C, Huson DH. Fast and sensitive protein alignment using DIAMOND. Nat Methods **2015**;12:59-60
37. Caspi R, Billington R, Ferrer L, Foerster H, Fulcher CA, Keseler IM*, et al.* The MetaCyc database of metabolic pathways and enzymes and the BioCyc collection of pathway/genome databases. Nucleic Acids Res **2016**;44:D471-80
38. Aitchison J. The Statistical Analysis of Compositional Data. Journal of the Royal Statistical Society Series B (Methodological) **1982**;44:139-77
39. Gloor GB, Macklaim JM, Pawlowsky-Glahn V, Egozcue JJ. Microbiome Datasets Are Compositional: And This Is Not Optional. Front Microbiol **2017**;8:2224
40. Mallick H, Rahnavard A, McIver LJ, Ma S, Zhang Y, Nguyen LH*, et al.* Multivariable association discovery in population-scale meta-omics studies. PLoS Comput Biol **2021**;17:e1009442
41. Benjamini Y, Krieger AM, Yekutieli D. Adaptive linear step-up procedures that control the false discovery rate. Biometrika **2006**;93:491-507
42. McMurdie PJ, Holmes S. phyloseq: an R package for reproducible interactive analysis and graphics of microbiome census data. PLoS One **2013**;8:e61217
43. Lahti L, Shetty S. Tools for microbiome analysis in R. https://github.com/microbiome/microbiome/2017.
44. Oksanen J, Blanchet FG, Friendly M, Kindt R, Legendre P, McGlinn D*, et al.* Community Ecology Package 2020.
45. Otasek D, Morris JH, Boucas J, Pico AR, Demchak B. Cytoscape Automation: empowering workflow-based network analysis. Genome Biol **2019**;20:185
46. Argelaguet R, Velten B, Arnol D, Dietrich S, Zenz T, Marioni JC*, et al.* Multi-Omics Factor Analysis-a framework for unsupervised integration of multi-omics data sets. Mol Syst Biol **2018**;14:e8124
47. Salter SJ, Cox MJ, Turek EM, Calus ST, Cookson WO, Moffatt MF*, et al.* Reagent and laboratory contamination can critically impact sequence-based microbiome analyses. BMC Biol **2014**;12:87

48. Baraniya D, Jain V, Lucarelli R, Tam V, Vanderveer L, Puri S*, et al.* Screening of Health-Associated Oral Bacteria for Anticancer Properties in vitro. Front Cell Infect Microbiol **2020**;10:575656
49. Davidsson S, Carlsson J, Greenberg L, Wijkander J, Soderquist B, Erlandsson A. Cutibacterium acnes Induces the Expression of Immunosuppressive Genes in Macrophages and is Associated with an Increase of Regulatory T-Cells in Prostate Cancer. Microbiol Spectr **2021**;9:e0149721
50. Bae Y, Ito T, Iida T, Uchida K, Sekine M, Nakajima Y*, et al.* Intracellular Propionibacterium acnes infection in glandular epithelium and stromal macrophages of the prostate with or without cancer. PLoS One **2014**;9:e90324
51. Davidsson S, Molling P, Rider JR, Unemo M, Karlsson MG, Carlsson J*, et al.* Frequency and typing of Propionibacterium acnes in prostate tissue obtained from men with and without prostate cancer. Infect Agent Cancer **2016**;11:26
52. Kakegawa T, Bae Y, Ito T, Uchida K, Sekine M, Nakajima Y*, et al.* Frequency of Propionibacterium acnes Infection in Prostate Glands with Negative Biopsy Results Is an Independent Risk Factor for Prostate Cancer in Patients with Increased Serum PSA Titers. PLoS One **2017**;12:e0169984
53. Severi G, Shannon BA, Hoang HN, Baglietto L, English DR, Hopper JL*, et al.* Plasma concentration of Propionibacterium acnes antibodies and prostate cancer risk: results from an Australian population-based case-control study. Br J Cancer **2010**;103:411-5
54. Shannon BA, Garrett KL, Cohen RJ. Links between Propionibacterium acnes and prostate cancer. Future Oncol **2006**;2:225-32
55. Shinohara DB, Vaghasia AM, Yu SH, Mak TN, Bruggemann H, Nelson WG*, et al.* A mouse model of chronic prostatic inflammation using a human prostate cancer-derived isolate of Propionibacterium acnes. Prostate **2013**;73:1007-15
56. Che B, Zhang W, Xu S, Yin J, He J, Huang T*, et al.* Prostate Microbiota and Prostate Cancer: A New Trend in Treatment. Front Oncol **2021**;11:805459
57. Nazipi S, Stodkilde-Jorgensen K, Scavenius C, Bruggemann H. The Skin Bacterium Propionibacterium acnes Employs Two Variants of Hyaluronate Lyase with Distinct Properties. Microorganisms **2017**;5
58. Li S, Kelly SJ, Lamani E, Ferraroni M, Jedrzejas MJ. Structural basis of hyaluronan degradation by Streptococcus pneumoniae hyaluronate lyase. EMBO J **2000**;19:1228-40
59. Yang Y, Li L, Xu C, Wang Y, Wang Z, Chen M*, et al.* Cross-talk between the gut microbiota and monocyte-like macrophages mediates an inflammatory response to promote colitis-associated tumourigenesis. Gut **2020**
60. Hannigan GD, Duhaime MB, Ruffin MTt, Koumpouras CC, Schloss PD. Diagnostic Potential and Interactive Dynamics of the Colorectal Cancer Virome. mBio **2018**;9
61. Eliassen E, Lum E, Pritchett J, Ongradi J, Krueger G, Crawford JR*, et al.* Human Herpesvirus 6 and Malignancy: A Review. Front Oncol **2018**;8:512
62. Hirt L, Hirsch-Behnam A, de Villiers EM. Nucleotide sequence of human papillomavirus (HPV) type 41: an unusual HPV type without a typical E2 binding site consensus sequence. Virus Res **1991**;18:179-89
63. Lin W, Xu Z, Yan Y, Zhang H, Li H, Chen W*, et al.* Avian Leukosis Virus Subgroup J Attenuates Type I Interferon Production Through Blocking IkappaB Phosphorylation. Front Microbiol **2018**;9:1089
64. Sacco MA, Flannery DM, Howes K, Venugopal K. Avian endogenous retrovirus EAV-HP shares regions of identity with avian leukosis virus subgroup J and the avian retrotransposon ART-CH. J Virol **2000**;74:1296-306
65. Gonzalez-Cao M, Iduma P, Karachaliou N, Santarpia M, Blanco J, Rosell R. Human endogenous retroviruses and cancer. Cancer Biol Med **2016**;13:483-8
66. Kolbe AR, Bendall ML, Pearson AT, Paul D, Nixon DF, Perez-Losada M*, et al.* Human Endogenous Retrovirus Expression Is Associated with Head and Neck Cancer and Differential Survival. Viruses **2020**;12
67. Perera M, Al-hebshi N, Perera I, Ipe D, Ulett G, Speicher D*, et al.* A Dysbiotic Mycobiome Dominated by Candida albicans is Identified within Oral Squamous Cell Carcinomas. Journal of Oral Microbiology **2017, in press.**;10:1385369

68. Mohamed N, Litlekalsoy J, Ahmed IA, Martinsen EMH, Furriol J, Javier-Lopez R*, et al.* Analysis of Salivary Mycobiome in a Cohort of Oral Squamous Cell Carcinoma Patients From Sudan Identifies Higher Salivary Carriage of Malassezia as an Independent and Favorable Predictor of Overall Survival. Front Cell Infect Microbiol **2021**;11:673465
69. Spatz M, Richard ML. Overview of the Potential Role of Malassezia in Gut Health and Disease. Front Cell Infect Microbiol **2020**;10:201
70. Huang GZ, Wu QQ, Zheng ZN, Shao TR, Lv XZ. Identification of Candidate Biomarkers and Analysis of Prognostic Values in Oral Squamous Cell Carcinoma. Front Oncol **2019**;9:1054
71. De Cecco L, Nicolau M, Giannoccaro M, Daidone MG, Bossi P, Locati L*, et al.* Head and neck cancer subtypes with biological and clinical relevance: Meta-analysis of gene-expression data. Oncotarget **2015**;6:9627-42
72. Chandrashekar DS, Bashel B, Balasubramanya SAH, Creighton CJ, Ponce-Rodriguez I, Chakravarthi B*, et al.* UALCAN: A Portal for Facilitating Tumor Subgroup Gene Expression and Survival Analyses. Neoplasia **2017**;19:649-58
73. Yang W, Zhou W, Zhao X, Wang X, Duan L, Li Y*, et al.* Prognostic biomarkers and therapeutic targets in oral squamous cell carcinoma: a study based on cross-database analysis. Hereditas **2021**;158:15
74. Willenbrink TJ, Ruiz ES, Cornejo CM, Schmults CD, Arron ST, Jambusaria-Pahlajani A. Field cancerization: Definition, epidemiology, risk factors, and outcomes. J Am Acad Dermatol **2020**;83:709-17

Table 1. Key genes within significantly enriched host pathways

Tumor tissues compared to the adjacent normal tissues	
Pathway	**Genes**
Upregulated	
Epithelial-to-mesenchymal transition	*INHBA* Matrix metalloproteinases: *MMP1*, *MMP2*, *MMP3*, *MMP9*, *MMP10*, *MMP11* and *MMP14* Collagen genes: *COL4A1*, *COL4A2*, *COL11A1*, *COL12A1* and *CTHRC1* Growth factors: *TGFBI*, *IGF2* and *CSF2* Others: *FSTL3* (Follistatin-like 3) and *ITGA5* (Integrin subunit α5)
Inflammatory response	
KRAS signaling	
Interferon alpha and gamma responses	Interferon-induced proteins: *IFI27*, *IFIT1*, *IFI35*, *IFIT3* and *ISG15*
IL-6/JAK/stat pathway	CXCL genes, *IL-6*, *STAT1*, *TGFBI* and *CSF2*
Angiogenesis	*VAV2*, *SPP1*, *COL5A2*, *PDGFA* and *POSTN*
Others	HOX genes (clusters A, B, C, D) and (e.g. HOTAIR)
Down-regulated	
Oxidative phosphorylation	*MPC1*, *ETFDH*, *VDAC2*, *ACADSB*, *ACAA1*, *ACADVL*, *RETSAT*
Fatty acid metabolism	*ADH7*, HPGD, CBR3, ALDH3A1, ALDH3A2
Adipogenesis	*CYP4B1*, *EPHX2*, *PPARG*, *ELOVL6*, *MGLL*
P53 pathway	*KLF4*, GLS2, *BAIAP2*, *CDKN2AIP*, and *EPS8L2*
Tumor/Adjacent normal compared to healthy controls	
Pathway	**Genes**
Upregulated	
E2F targets	*STAG*, *PRIM2*, *SMC6*, *MRE11*, *PRKDC*, *CHEK2* and *BRCA2*
MYC targets	*CBX3*, *DDX18*, *HSPD1*, *NOP56*, *MCM4* and *NOLC*
G2M checkpoint genes	*STAG1*, *CDC27*, *SRSF10*, *MYC*, *WRN*, *CENPE* and *MNAT1*
Epithelial-to-mesenchymal transition	Extracellular matrix proteins and adhesion molecules*: DST*, *COL4A1*, *CDH6*, *PLOD3*, *GEM*, *LAMC1*, *MYLK*, *COLGALT1* and *LAMA3*
Down-regulated	
Heme metabolism	*FBXO34*, *HBB*, *BPGM*, *RHCE*, *ADIPOR1*, and *LMO2*
P53 pathway	*MXD1*, *TGFA*, *CDKN2AIP*, *HMOX1*, *SAT1*, *FOXO3*
Apoptosis	*EMP1*, *SQSTM1*, *HMOX1*, *BCL2L1*, *H1-0*, *IL18*, *CDKN1A* and *BCL10*
TNF-alpha signaling via NFKB	*DUSP5*, *TNIP1*, *MXD1*, *IL23A*, *MAP2K3*, *IL1A*, *IL-1B* and *TNF*
Inflammatory response pathway	*MXD1*, *FFAR2*, *IRAK2*, *IL1A*, *SPHK1*, *RAF1* and *CXCL8*

Figure Legends

Figure 1. A flow chart of study design and procedures. Left: Study groups, sampling, sample processing, qPCR, RNA library preparation and sequencing. Right: Bioinformatic analysis pipeline used to map, analyze and integrate the host and microbiome sequencing data.

Figure 2. Host transcriptome. Sequences were mapped to human reference GRCh38 and quantified at the gene-level using STAR aligner and HTSeq, respectively. *DESeq2* was used to identify differential expressed genes (DEGs) which are depicted in volcano plots for **(A)** Tumor vs. adjacent normal (paired comparison), **(B)** Tumor vs. healthy control, and **(C)** Adjacent normal vs. healthy control—the color code in the volcano plots corresponds to the fold change and FDR cutoffs (2 and 0.05 respectively). (**D**) A PCA plot based on the most variable 1,500 DEGs, generated using "plotPCA" function from DESeq2 R/Bioconductor package. Lists of DEGs pre-ranked by Wald statistic were used as input for Gene Set Enrichment Analysis (GSEA) to identify upregulated and downregulated pathways in each contrast (**E-G**) based on Hallmark gene sets (MSigDB 7.1).

Figure 3. Microbial loads and transcriptional profiles. DNA extracted from the same cells as for RNA was used to determine bacterial, fungal and archaeal loads in the samples relative to human actin-β gene by q-PCR (**A**). Non-human, non-ribosomal RNA sequences were used as input for HUMAnN 3.0 for microbial taxonomic and functional profiling. (**B**) Microbial phyla and (**C**) top 20 genera identified in each of the study groups ranked by their average transcriptional relative abundances. (**D**) A PCA plot based on CLR-transformed genus-level profiles (created with Phyoseq and microbiome R packages). (**E**) Top 20 expressed enzyme classes ranked by their average transcriptional abundances (CLR transformed counts) in each sample type.

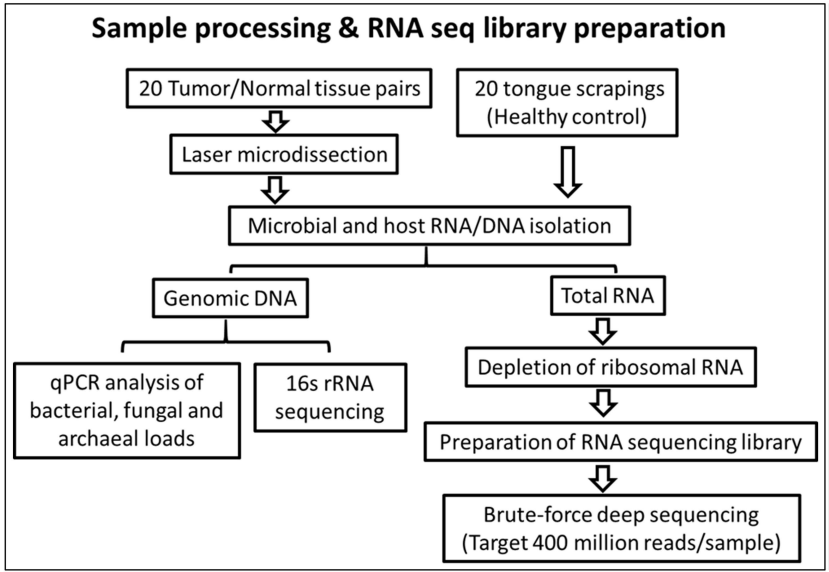
Figure 4. Differentially abundant microbial features. Taxonomic and functional profiles obtained with HUMAnN 3.0 were CLR transformed and differential abundance analysis was performed with MaAsLin2 setting FDR cutoff to 0.05. (**A**) Bar plots of the top differentially abundant genera in the tumor vs. control and adjacent normal vs. healthy control contrasts. (**B**) A heat map showing clustering of samples based on top differentially abundant enzyme classes. The plots were created with ggplot2 and pheatmap R packages.

Figure 5. Hyaluronate lyase expression by *Cutibacterium acnes*. Relative abundance of hyaluronate lyase transcripts in individual samples. The transcripts were exclusively contributed by *Cutibacterium acnes* based on HUMAnN 3.0 results.

Figure 6. Microbiome-host data integration. Unsupervised Spearman rank correlation was performed between appropriately transformed host gene and microbial genus counts. Significant correlations were defined as those with absolute correlation coefficient (r_s) ≥ 0.6 and FDR ≤ 0.05. For each microbial feature, the correlating genes, ranked by r_s, were subjected to GSEA. (**A**) genera that correlated with > 1,000 host genes and whether GSEA turned significant results for each genus. (**B**) interaction of selected genera with the host pathways based on GSEA results. Red edges denote activation while blue edges denote inhibition. For each edge, the number of genes involved is displayed. (**C-E**) MSigDB genes with the highest correlations with Roseolovirus, *Cutibacterium* and Yuavirus, respectively. Red edges denote positive correlation while blue edges denote inhibition. For each edge, r_s is displayed.

Figure 7. Preliminary *in vitro* experimental validation in SCC25 cells. (A) *Cutibacterium acnes* and the MYC gene were chosen for this experiment based on their strong correlation in the sequencing data. r_s, Spearman correlation coefficient; CLR, centered log-ratio; Rlog, regularized log transformation defined in the DESeq2 package**. (B)** *C. ances* NCTC 737 grown to mid-log phase was co-cultured with SCC25 cells at multiplicity of infection (MOI) of 50, 100 and 200 for 24 hours. Levels of MYC mRNA were measured by qRT-PCR, normalized to GADPH and relative to the non-infected cells (controls). * Statistically significant, Welch-corrected t-test.

Figure 1



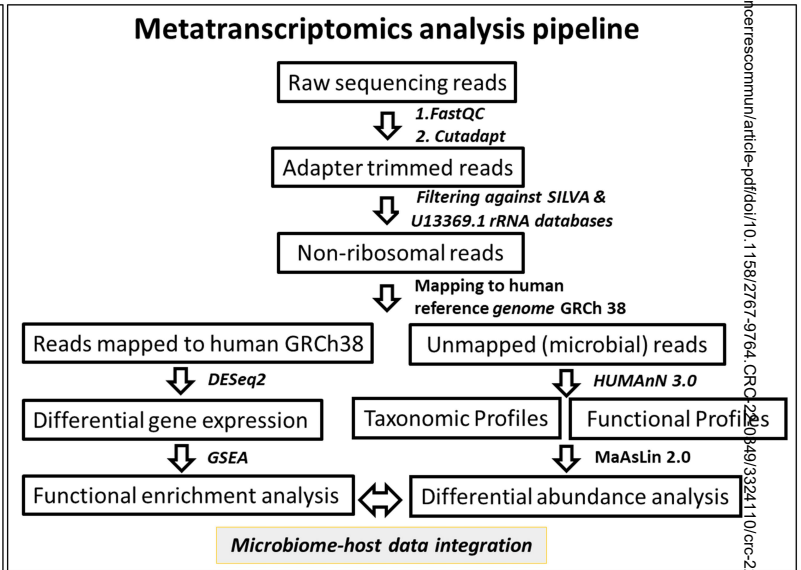


Figure 2

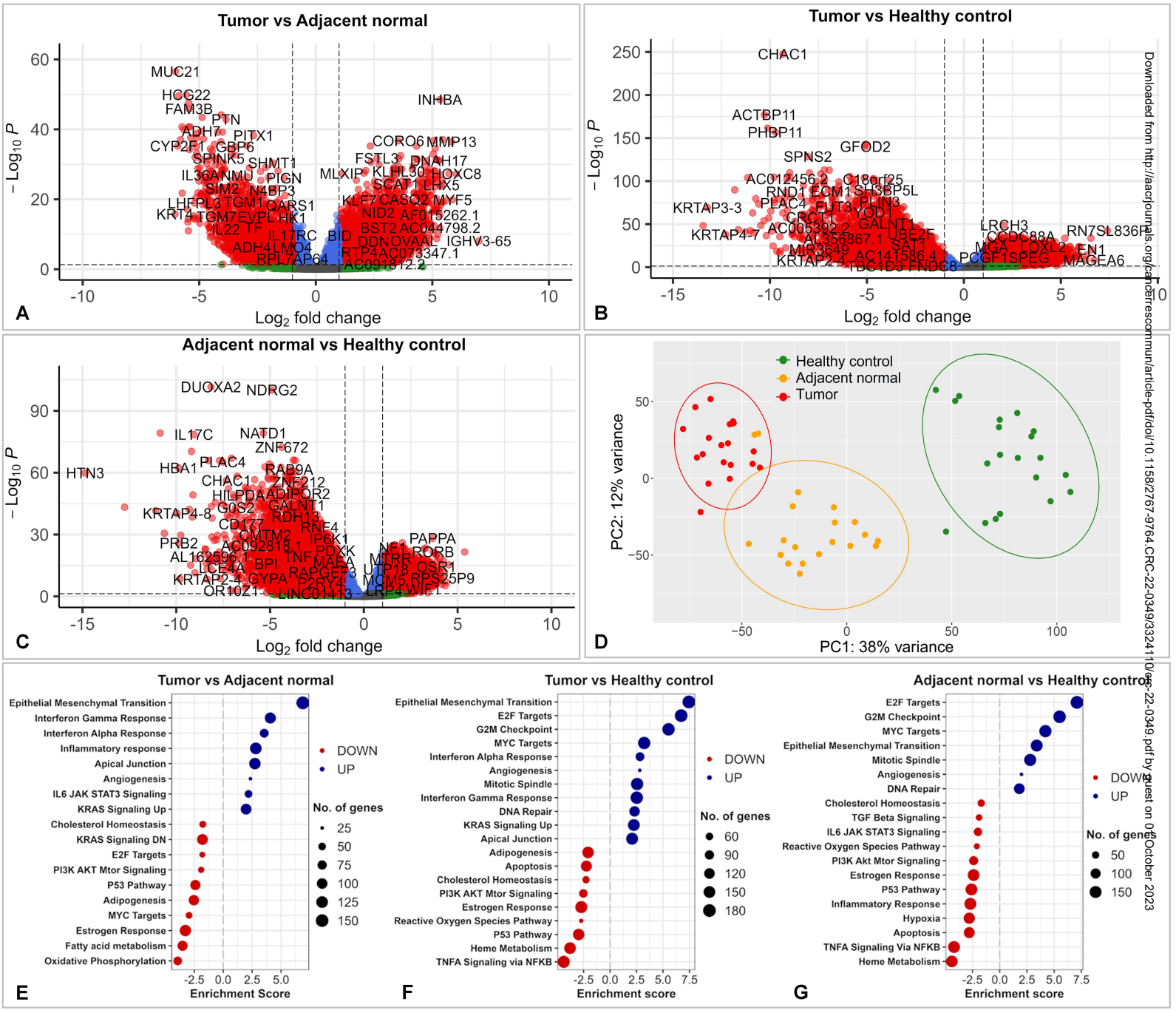


Figure 3

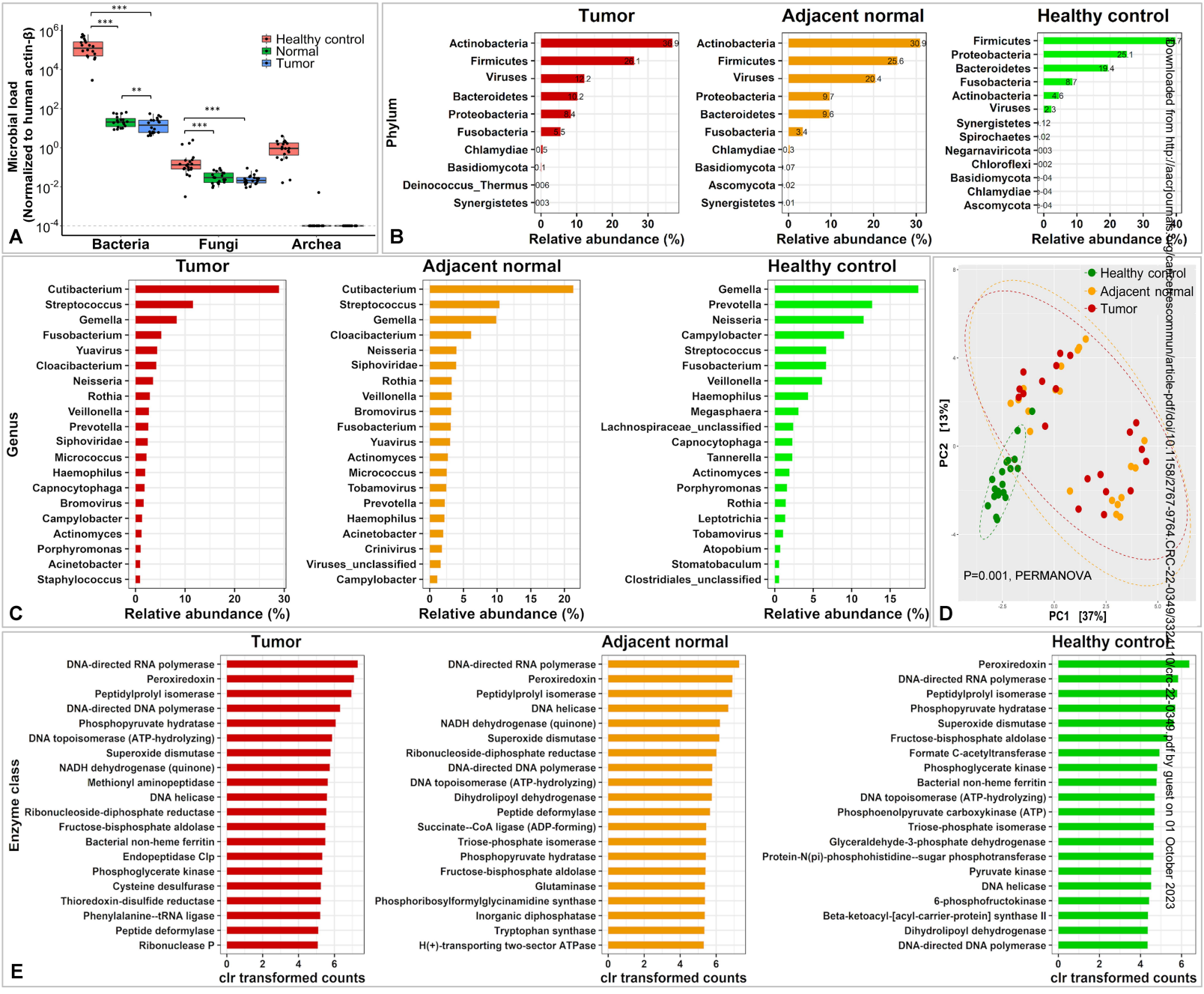


Figure 4

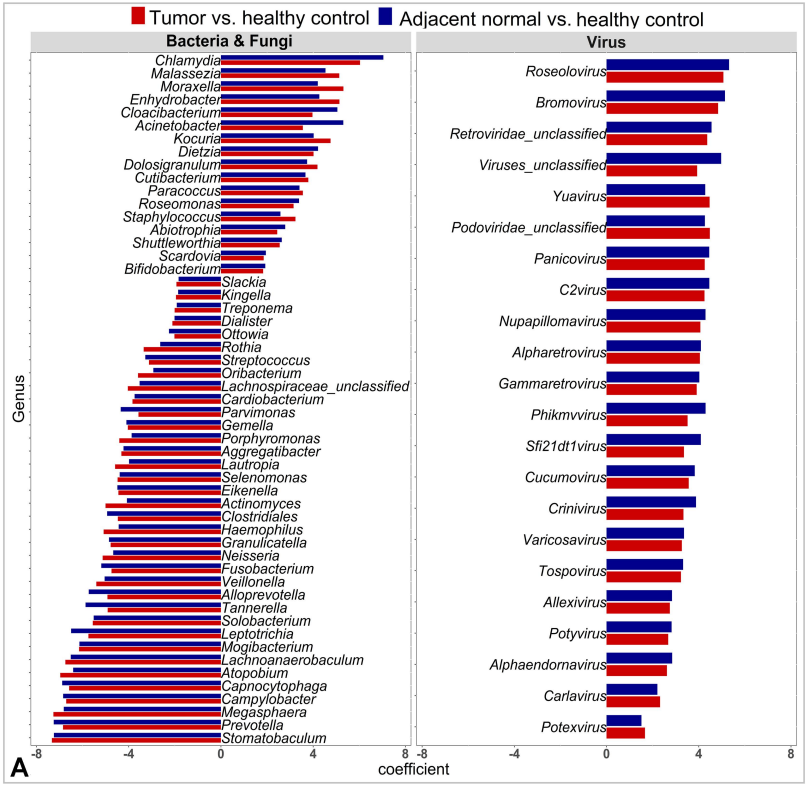


Figure 5

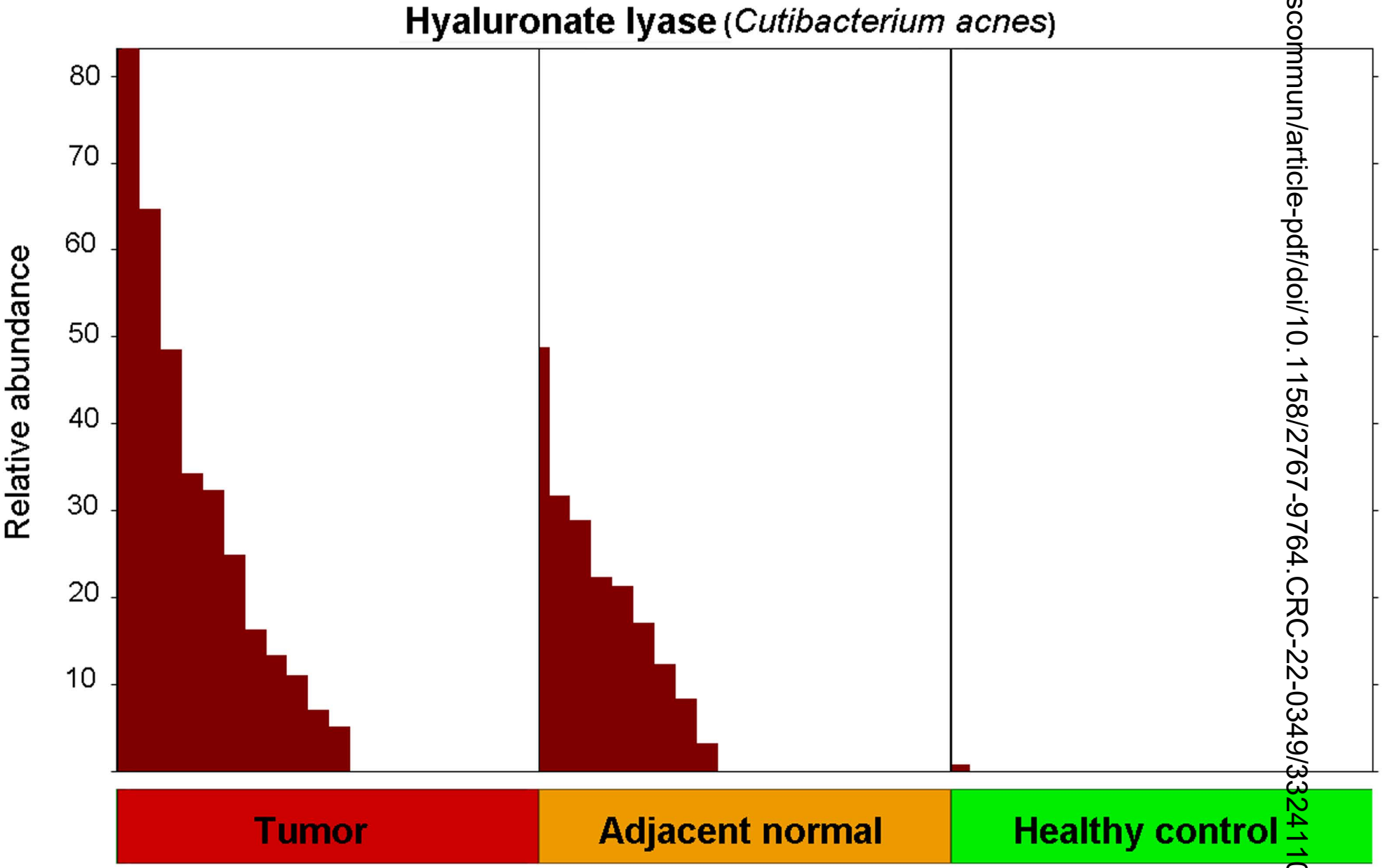


Figure 6

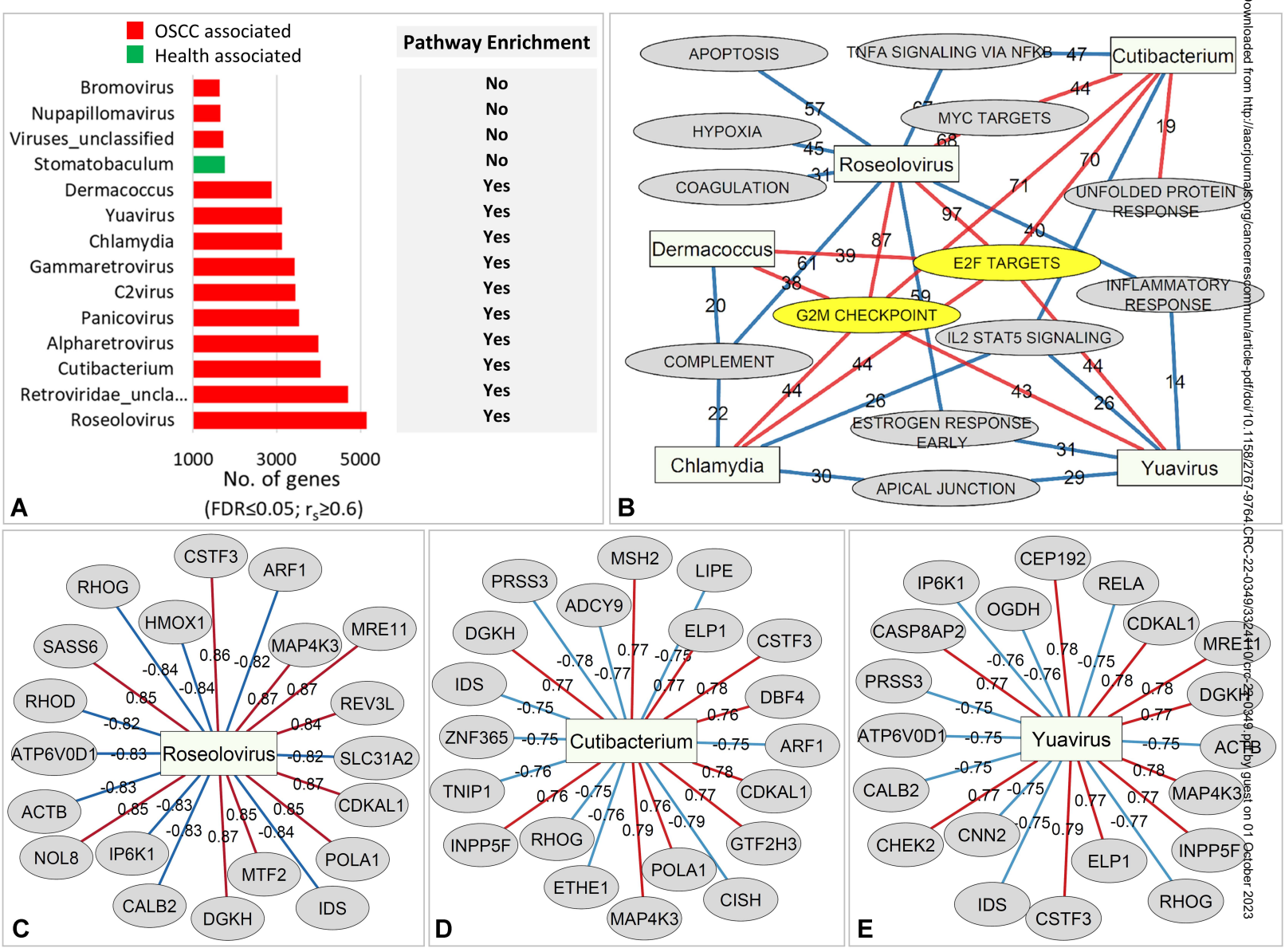

OSCC associated
Health associated
Pathway Enrichment
Bromovirus
Nupapillomavirus
Viruses_unclassified
Stomatobaculum
Dermacoccus
Yuavirus
Chlamydia
Gammaretrovirus
C2virus
Panicovirus
Alpharetrovirus
Cutibacterium
Retroviridae_uncla...
Roseolovirus
No
No
No
No
Yes
Yes
Yes
Yes
Yes
Yes
Yes
Yes
Yes
Yes
1000
3000
5000
No. of genes
(FDR≤0.05; r_s≥0.6)
A
APOPTOSIS
TNFA SIGNALING VIA NFKB
Cutibacterium
HYPOXIA
MYC TARGETS
Roseolovirus
COAGULATION
UNFOLDED PROTEIN RESPONSE
Dermacoccus
E2F TARGETS
INFLAMMATORY RESPONSE
G2M CHECKPOINT
IL2 STAT5 SIGNALING
COMPLEMENT
ESTROGEN RESPONSE EARLY
Chlamydia
Yuavirus
APICAL JUNCTION
B
CSTF3
ARF1
RHOG
HMOX1
MRE11
SASS6
MAP4K3
RHOD
REV3L
ATP6V0D1
Roseolovirus
SLC31A2
ACTB
CDKAL1
NOL8
IP6K1
MTF2
POLA1
CALB2
DGKH
IDS
C
MSH2
LIPE
PRSS3
ADCY9
ELP1
DGKH
CSTF3
IDS
DBF4
ZNF365
Cutibacterium
ARF1
TNIP1
CDKAL1
INPP5F
RHOG
GTF2H3
ETHE1
POLA1
CISH
MAP4K3
D
CEP192
IP6K1
RELA
OGDH
CDKAL1
CASP8AP2
MRE11
PRSS3
DGKH
ATP6V0D1
Yuavirus
ACTB
CALB2
MAP4K3
CHEK2
CNN2
INPP5F
ELP1
IDS
RHOG
CSTF3
E
Downloaded from http://aacrjournals.org/cancerresearchcommun/article-pdf/doi/10.1158/2767-9764.CRC-22-0349/3324410/crc-22-0349.pdf by guest on 01 October 2023

Figure 7