An international report on bacterial communities in esophageal squamous cell carcinoma

Jason Nomburg[1,2,3], Susan Bullman[4], Dariush Nasrollahzadeh[5,6], Eric A. Collisson[7,8], Behnoush Abedi-Ardekani[6], Larry O. Akoko[9], Joshua R. Atkins[6], Geoffrey C. Buckle[7,8], Satish Gopal[10], Nan Hu[11], Bongani Kaimila[12], Masoud Khoshnia[5], Reza Malekzadeh[5], Diana Menya[13], Blandina T. Mmbaga[14,15], Sarah Moody[16], Gift Mulima[17], Beatrice P. Mushi[9], Julius Mwaiselage[18], Ally Mwanga[9], Yulia Newton[19], Dianna L. Ng[7,20], Amie Radenbaugh[19], Deogratias S. Rwakatema[14,15], Msiba Selekwa[9], Joachim Schüz[21], Philip R. Taylor[11], Charles Vaske[19], Alisa Goldstein[11], Michael R. Stratton[16], Valerie McCormack[21], Paul Brennan[6], James A. DeCaprio[1,3,22], Matthew Meyerson[1,2,22,23*], Elia J. Mmbaga[9, 24*], Katherine Van Loon[7,8*]

1 Department of Medical Oncology, Dana-Farber Cancer Institute, Boston, MA

2 Broad Institute of MIT and Harvard, Cambridge, MA

3 Harvard Program in Virology, Harvard Medical School, Boston, MA

4 Fred Hutchinson Cancer Research Center, Seattle, Washington, USA

5 Digestive Oncology Research Center, Digestive Disease Research Institute, Tehran University of Medical Sciences, Shariati Hospital. Tehran Iran.

6 International Agency for Research on Cancer (IARC/WHO), Genomic Epidemiology Branch, Lyon, France

7 University of California, San Francisco (UCSF) Helen Diller Family Comprehensive Cancer Center, San Francisco, CA, USA

8 Division of Hematology/Oncology, Department of Medicine, UCSF, San Francisco, California, USA

9 Muhimbili University of Health and Allied Sciences, Dar es Salaam, Tanzania

10 University of North Carolina, Chapel Hill, North Carolina, USA

11 Division of Cancer Epidemiology and Genetics, National Cancer Institute, Bethesda, MD, USA

12 UNC Project - Lilongwe, Malawi

13 School of Public Health, Moi University, Eldoret, Kenya

14 Kilimanjaro Clinical Research Institute, Kilimanjaro Christian Medical Centre, Moshi, Tanzania

15 Kilimanjaro Christian Medical University College, Moshi, Tanzania

16 Cancer, Ageing and Somatic Mutation, Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus, Hinxton, Cambridgeshire, UK

17 Kamuzu Central Hospital, Lilongwe, Malawi

18 Ocean Road Cancer Institute, Dar es Salaam, Tanzania

19 NantOmics/NantHealth, Inc., El Segundo, California, USA

20 Department of Pathology, UCSF, San Francisco, CA, USA

21 International Agency for Research on Cancer (IARC/WHO), Environment and Lifestyle Epidemiology Branch, Lyon, France

22 Department of Medicine, Brigham and Women’s Hospital, Harvard Medical School, Boston, MA

23 Department of Genetics, Harvard Medical School, Boston, MA

24 Department of Community Medicine and Global Health, University of Oslo, Norway

***Correspondence to:**

Katherine Van Loon - Katherine.VanLoon@ucsf.edu

Elia J. Mmbaga - eliajelia@yahoo.co.uk

Matthew Meyerson - matthew_meyerson@dfci.harvard.edu

ABSTRACT

The incidence of esophageal squamous cell carcinoma (ESCC) is disproportionately high in the eastern corridor of Africa and parts of Asia. Emerging research has identified a potential association between poor oral health and ESCC. One proposed biological pathway linking poor oral health and ESCC involves the alteration of the microbiome. Thus, we performed an integrated analysis of four independent sequencing efforts of ESCC tumors from patients from high- and low-incidence regions of the world. Using whole genome sequencing (WGS) and RNA sequencing (RNAseq) of ESCC tumors and WGS of synchronous collections of saliva specimens from 61 patients in Tanzania, we identified a community of bacteria, including members of the genera *Fusobacterium, Selenomonas, Prevotella, Streptococcus, Porphyromonas, Veillonella,* and *Campylobacter*, present at high abundance in ESCC tumors. We then characterized the microbiome of 238 ESCC tumor specimens collected in two additional independent sequencing efforts consisting of patients from other high-ESCC incidence regions (Tanzania, Malawi, Kenya, Iran, China). This analysis revealed a similar tumor enrichment of the ESCC-associated bacterial community in these cancers. Because these genera are traditionally considered members of the oral microbiota, we explored if there is a relationship between the synchronous saliva and tumor microbiomes of ESCC patients in Tanzania. Comparative analyses revealed that paired saliva and tumor microbiomes are significantly similar with a specific enrichment of *Fusobacterium* and *Prevotella* in the tumor microbiome. Together, these data indicate that cancer-associated oral bacteria are associated with ESCC tumors at the time of diagnosis and support a model in which oral bacteria are present in high abundance in both saliva and

tumors of ESCC patients. Longitudinal studies of the pre-diagnostic oral microbiome are needed to investigate whether these cross-sectional similarities reflect temporal associations.

INTRODUCTION

Esophageal cancer is the sixth most common cause of cancer-related death worldwide (1).There are two histologic subtypes of esophageal cancer with distinct biological characteristics, geographic distributions, and risk factors (2). Esophageal adenocarcinoma is the most common histologic form of esophageal cancer in high-income countries and is associated with factors including gastroesophageal reflux disease, Barrett's esophagus, and obesity (3, 4). By contrast, esophageal squamous cell carcinoma (ESCC) represents more than 90% of worldwide esophageal cancer cases and is the dominant histology in low-resource settings. In particular, there are two main regions where ESCC is endemic: (1) the Asian esophageal cancer belt, extending from western/northern China to central and southeast Asia; and (2) the eastern corridor of Africa, extending from Ethiopia to South Africa (5, 6).

Emerging research has identified a possible association between poor oral health and ESCC. Studies from Asia, Europe, Latin America, Kenya, and Iran have reported associations of ESCC with poor oral hygiene, chronic periodontal disease, dental decay, and tooth loss (7-16). Recently, three parallel case-control studies in Kenya and Tanzania, conducted as part of the African Esophageal Cancer Consortium (AfrECC) and ESCCAPE (esccape.iarc.fr) collaborations, reported possible associations of poor or infrequent oral hygiene with increased risk for ESCC in East Africa (17-20).

Alterations of the oral microbiome due to poor oral health is one proposed biological pathway that could explain the link between oral health and ESCC. Many bacterial genera associated with gastrointestinal cancers contain species that are traditionally associated with healthy or diseased oral microbiomes. For example, *Helicobacter pylori* was discovered to be associated with gastric cancers and mucosa-associated lymphoid tissue (MALT) lymphomas, indirectly by promoting gastric inflammation and directly by influencing cellular signaling (21). Similarly, bacteria of the genera *Fusobacterium*, *Selenomonas*, and *Prevotella* are enriched in colorectal cancers (22-24) and can be visualized invasively within tumor tissue (25). *Fusobacterium*, in particular, has been reported to promote carcinogenesis through the selective expansion or inhibition of certain classes of immune cells (26) and may drive cellular proliferation by stimulating Wnt/β-catenin signaling (27, 28). Other bacterial genera such as *Porphyromonas, Campylobacter,* and *Streptococcus* have emerging associations with various human gastrointestinal cancers (29-35).

As part of ongoing investigation into the microbiome's association with ESCC, we performed an integrated analysis of four independent sequencing efforts including ESCC tumors from patients from both high- and low-incidence regions of the world. In addition, we investigated the relationship between the microbiomes of matched ESCC tumors and saliva specimens in a subset of ESCC cases.

RESULTS

Study Population

To evaluate the potential role of the host microbiota in ESCC, we investigated the microbiome of 299 ESCC specimens from patients in five different countries with a high incidence of ESCC. Specimens were collected through four independent sequencing efforts (**Figure 1A**). Specimens consisted of whole genome sequencing (WGS) and RNA sequencing (RNAseq) data from the tumor and saliva of 61 patients from Tanzania (the "MUHAS Tanzania" cohort) (36), RNAseq data from the tumors of 30 ESCC patients in Malawi (the "UNC Project – Malawi" cohort) (37), and WGS from 208 additional samples of tumors from patients in high ESCC incidence regions, including specimens from ESCC patients in Tanzania (n=18) and Kenya (n=64) that were collected in the ESCCAPE studies (esccape.iarc.fr) and specimens from ESCC patients in East Golestan, Iran (n=55) and Shanxi, China (n=71) that were sequenced as part of the Cancer Research UK Mutographs project ("Mutographs" cohorts) (38). In addition, we analyzed WGS data of ESCC from The Cancer Genome Atlas (39), which includes a small number of tumors from patients in low-incidence geographic regions including the United States (n=3), Ukraine (n=3), Vietnam (n=22), and Russia (n=8) (the "TCGA" cohort). Patient characteristics are shown in Table 1.

Bacterial populations are abundant and diverse in ESCC tumors

We used the metagenomic analysis tool GATK-PathSeq (40) to process the RNAseq and WGS data. GATK-PathSeq uses a sequential mapping strategy to assign reads to human and microbial reference genomes, resulting in detailed information on sequencing reads of human and microbial origin (**Figure S1A**). We likewise used GATK-PathSeq to process WGS data sets from 50 colon adenocarcinoma (COAD) specimens available from TCGA (41) for comparison, as there is strong evidence of microbial associations with COAD (22-25).

The bacterial burden of ESCC tumors ranged from 10 to 1000 bacterial reads per million human reads, similar to numbers observed in TCGA COAD (**Figure 1B**). Furthermore, the Shannon diversity of bacterial populations at the genus level ranged from 2 to 3 (**Figure 1C**). By comparison, ESCC-associated bacterial communities are as diverse or more diverse than TCGA COAD. At the phylum level, ESCC bacterial populations generally consist of *Firmicutes*, *Bacteroidetes*, *Proteobacteria*, *Actinobacteria*, and *Fusobacteria* (**Figure 1D, Figure S1B**). Of note, the higher than expected abundance of the phylum *Actinobacteria* specifically in the TCGA ESCC samples is attributable, in particular, to a very high abundance of the genus *Tetrasphaera* (**Figure S1C**). This is evidenced by a depressed Shannon diversity of *Actinobacteria* genera in these samples (**Figure S1D**) and may indicate contamination of the TCGA ESCC samples. *Actinobacteria* have been reported as a source of contaminating reads in TCGA gastrointestinal cancer samples (42).

Bacterial genera associated with carcinogenesis are observed at high relative abundance in ESCC tumors from Tanzania

To determine if bacteria with known associations with cancer are present in ESCC, we first analyzed the sequencing series of the 61 ESCC cases from the MUHAS Tanzania cohort with both WGS and RNAseq data. The paired WGS and RNAseq data from these tumors allowed investigation of bacterial communities at the DNA and RNA levels. Both WGS and RNAseq data revealed high relative abundance of bacterial genera previously associated with carcinogenesis in these ESCC tumors (**Figure 2A, 2B**). The high relative abundance of the *Fusobacterium* genus was particularly notable. Other bacterial genera of interest include *Streptococcus*, *Porphyromonas*, *Campylobacter*, *Prevotella*, *Veillonella*, and *Selenomonas*, many of which have been associated with gastrointestinal malignancies alongside or independently of *Fusobacterium* (25, 29, 32, 34, 43). The mean Jaccard similarity index between tumor RNAseq and WGS data from the same tumor is 0.54, greater than the average Jaccard similarity index of random RNAseq-WGS pairs (0.36), indicating that bacterial populations inferred from WGS and RNAseq data are generally consistent (**Figure 2C**).

Next, we attempted to determine if similar bacterial genera were also present in ESCC from patients in high-incidence countries beyond Tanzania. Investigation of RNA sequencing data from patients in Malawi, WGS data from patients in Kenya, China, and Iran, as well as from the independent ESCCAPE Tanzania patient group revealed pervasive evidence of similar bacterial genera in the tumors of these patients (**Figure 2D, Figure S2A**). To investigate if similar microorganisms were found in ESCC tumors from patients in low-incidence regions, we investigated WGS data from ESCC tumors originating from USA, Ukraine, Vietnam, and Russia that were available through TCGA. While the number of samples available from low-incidence regions is low and relies on a single sequencing effort, we found that the tumors of many of these patients contain similar bacterial genera (**Figure 2D, Figure S2A)**. Colon cancers from the TCGA COAD cohort revealed evidence of *Fusobacterium,* as expected; however, these COAD samples were notable for much lower relative abundance of the other genera of interest, when compared to ESCC tumors.

Evaluation of association between saliva and tumor microbiomes in ESCC patients from Tanzania

We next investigated the similarity between the saliva and tumor microbiomes of ESCC patients. Paired saliva samples were only available from patients in the MUHAS Tanzania cohort (N=45); these paired saliva specimens were analyzed to evaluate bacterial abundance as a proxy for the oral microbiome.

We first assessed the similarity between paired saliva and tumor microbiomes with the Bray Curtis similarity index (44). To avoid potential confounding due to low bacterial read counts in some tumor samples, we limited these analyses to the 21 tumor-saliva pairs that contain appreciable microbial sequencing depth (at least 10,000 bacterial reads each). We found that the saliva and tumor microbiomes from the same patient in the Tanzanian samples are significantly more similar than random saliva-tumor pairs (p=0.0003, Wilcoxon rank sum test) (**Figure 3A**). Next, we asked if there are bacterial genera whose relative abundance in the saliva correlates with their relative abundance in the tumor. For this analysis, we included only common-abundant bacterial genera with at least 1% relative abundance in at least three tumor-oral pairs. The relative abundance of four bacterial genera (*Fusobacterium*, *Veillonella*, *Streptococcus*, and *Porphyromonas*) are strongly correlated between tumor and saliva microbiomes, while other common-abundant bacterial genera were not (**Figure 3B**). To assess if any bacterial genera are preferentially enriched in the tumor microbiome relative to the saliva microbiome, we next calculated the difference in the relative abundance of the common-abundant bacterial genera between saliva-tumor pairs. Several genera including *Porphyromonas* and *Veillonella* were at higher relative abundance in the saliva, while *Prevotella* and *Fusobacterium* were enriched in the tumor microbiome (**Figure 3C**). Finally, the relative abundance of tumor-associated bacteria including *Fusobacterium*, *Prevotella*, *Selenomonas*, *Veillonella*, *Streptococcus*, and *Campylobacter* are strikingly similar between the microbiomes of tumor and oral pairs (**Figure 3D**). Altogether, these data support the hypothesis that there is an association between the oral and tumor microbiome of ESCC patients in Tanzania.

DISCUSSION

This report provides an analysis of bacterial communities present in ESCC tumors from nine countries from different regions of the world, analyzed in four independent sequencing efforts. We found traditionally oral, cancer-associated, bacterial genera in tumors from patients in Tanzania, Malawi, Kenya, China, and Iran. These results provide evidence that these bacterial genera may be associated with ESCC in high-incidence regions. We also identified similar bacterial genera in ESCC tumors from low-incidence regions, although this finding is based on a small sample size and only one sequencing cohort. Finally, in a sub-analysis of tumor and saliva pairs available from Tanzania, we demonstrated that the synchronous collected saliva and tumor microbiomes of ESCC patients are strikingly similar at the time of diagnosis; in particular, we identified a specific correlation between the saliva and tumor relative abundance of the bacterial genera *Fusobacterium*, *Veillonella*, *Streptococcus*, and

Porphyromonas, with *Prevotella* and *Fusobacterium* significantly enriched in the tumor microbiome.

Many of the bacterial genera identified in this study have been previously implicated in the carcinogenesis of gastrointestinal cancers. For example, studies have found that oral microbiota including *Fusobacterium*, *Prevotella*, *Selenomonas*, *Veillonella*, *Streptococcus*, and *Campylobacter* can be used to distinguish individuals with colorectal cancer from healthy controls (45), and that *Fusobacterium nucleatum* strains that colonize the oral cavity and tumors of patients with colorectal cancer are identical in some patients (46), raising the possibility that the oral cavity is a source of extra-oral cancer microbiota. Our group has previously shown that *Fusobacterium*, *Selenomonas*, and *Prevotella* can be visualized invasively within colorectal tumors and liver metastases (25). Fusobactium nucleatum has been previously identified in esophageal cancers and is associated with shorter survival (47). Members of the genus *Porphyromonas* have been previously observed invasively within ESCC tumors (29) and have been reported to promote oral squamous cell carcinoma through a variety of mechanisms (30, 31). *Campylobacter jejuni* has been reported to promote tumorigenesis in mice (32), and *Streptococcus* species have been identified in human esophageal cancers (33). In addition, the striking association of *Streptococcus bovis* with colorectal cancer has led to the recommendation that colonoscopy be performed upon detection of *Streptococcus bovis* bacteremia or endocarditis (34, 35). Oral commensal bacteria such as *Veillonella* species have been previously implicated in pathogenesis of lung cancer (43). A prospective cohort of American patients (48) and a study of Japanese patients (49) likewise found that oral microbiome composition reflects risk of esophageal cancers

We found that bacterial genera including *Fusobacterium*, *Prevotella*, *Selenomonas*, *Veillonella*, *Streptococcus*, and *Campylobacter* are pervasive in the microbiome of ESCC tumors from patients in high-incidence regions. Moreover, the bacterial composition of ESCC tumors is remarkably similar across countries in those high-incidence regions, raising the possibility that these bacterial genera may be involved in ESCC carcinogenesis or that they may colonize tumors as a result of the common clinical presentation of patients with severe dysphagia. Notably, there are several alternative hypotheses that warrant mention. For example, it is possible that the ESCC-associated bacterial genera simply represent common members of the esophageal microbiome (50) and that the microbial populations we observed in these cancers are not significantly different from those found in normal esophagus tissue. A limitation of our study is a lack of normal esophageal tissue from ESCC cases or healthy controls in these settings, which would allow us to address this possibility. Another possible explanation is that ESCC tumors provide a favorable niche in which these bacteria are sequestered and allowed to colonize due to the propensity of this disease to cause malignant obstruction. Thus, it is plausible that ESCC-associated bacteria are not necessarily promoting ESCC carcinogenesis but rather represent passengers resulting from the sequestration of oral secretions proximal to an obstructing tumor. While the previous association of these bacterial genera with other cancers is consistent with the hypothesis that they influence carcinogenesis of ESCC, future studies are necessary to identify which, if any, direct influences these bacterial genera have upon ESCC carcinogenesis. Nevertheless, even if these bacterial genera do not have a role in increasing ESCC risk, but arise at the time of disease onset, they may have an important role to play as part of a non-invasive early-detection biomarker. Finally, a concern of all microbiome analyses is that observed bacteria can be a consequence of contamination at some step between tumor harvest and sequencing. While some TCGA samples may be contaminated by *Actinobacteria* as previously noted, the presence of *Fusobacterium*, *Prevotella*, *Selenomonas*, *Veillonella*, *Streptococcus*, and *Campylobacter* in four independently collected cohorts indicates that these finding are unlikely due to contamination.

While this study focused on the presence of bacteria with ESCC in high-incidence regions, we found evidence of similar cancer-associated bacteria in tumors in patients from low-incidence regions (USA, Ukraine, Vietnam, and Russia). A limitation of this assessment is the small sample size (n=36) and reliance on a single TCGA cohort that likely contains contaminants (42). Regardless, this finding does not exclude the possibility that the microbiome could be a factor driving patterns of ESCC incidence. For example, it is possible that the prevalence of ESCC-associated bacteria in people could vary across regions, which in turn could drive these differing rates of ESCC incidence. This is an important topic for future study.

We found that the structure of synchronous paired tumor and oral microbiomes were strikingly similar. It is possible that this similarity is driven by transient contact of saliva and its associated microbiome with the tumor (e.g., during swallowing or tumor extraction). However, we found that only four of sixteen common-abundant bacterial genera correlate in abundance between the tumor and oral microbiomes, suggesting tumor-oral microbiome similarity is not driven exclusively by “in-trans” interactions between the saliva and tumor. We also found that genera including *Prevotella* and *Fusobacterium* are often specifically enriched in the tumor microbiome, supporting a model where specific oral bacterial preferentially colonize the tumor. A caveat of this study is that we infer oral bacterial populations from the saliva, despite diverse communities of bacteria throughout the oral cavity (51). However, we do observe *Fusobacterium* in the saliva despite its general association with periodontal plaques (52), suggesting saliva is capable of detecting periodontal pathogens. Additionally, because the samples studied here are from patients with late-stage disease, it is possible that tumor-induced changes to upper-gastrointestinal physiology and dysphagia symptom-induced major dietary changes could themselves alter the oral microbiomes of these patients. The previous findings from the ESCCAPE studies in Kenya and Tanzania (17, 19) which found strong associations with dental staining (ORs > 10) and for which photographic validation studies suggest that most dental staining was not fluorosis, also point to a recent build-up of chromogenic bacteria. Studies of the oral microbiome of patients at earlier stages of ESCC and in prospective studies are necessary to address this possibility. We restricted our analysis to 21 tumor-oral pairs that have a sufficient number of bacterial reads (at least 10,000). It is likely that excluded samples are not molecularly distinct from included samples but that the relatively low bacterial read counts in some tumors is simply reflective of low sequencing depth.

Our observation of similar tumor and saliva microbiomes in ESCC patients is especially notable considering emerging evidence linking periodontal disease and poor oral health with increased risk of various cancers (17, 53, 54). This raises several important open questions. It will be essential to determine if there is a difference in the oral prevalence of these identified cancer-associated bacteria between ESCC patients and non-patients earlier in the natural history of the disease, for example through comparisons of patients with esophageal squamous dysplasia and healthy controls. Because the prevalence of these bacteria may be associated with factors such as oral health, hygiene, and diet, studies of the impact of these factors on the oral microbiome in the general population would inform whether the oral microbiome is on a pathway linking oral hygiene to ESCC risk and may have a role in prevention.

In conclusion, we show that cancer-associated, traditionally-oral bacteria including the genera *Fusobacterium, Selenomonas, Prevotella, Streptococcus, Porphyromonas, Veillonella,* and *Campylobacter* are highly abundant within ESCC tumors from patients in high-ESCC incidence regions. We also show that there is a correlation between the genus composition of the saliva microbiome and the ESCC tumor microbiome of some ESCC patients. These findings will be foundational for future studies to understand if and how bacteria influence ESCC pathogenesis and to understand the role of the oral microbiome in this process. Finally, this study highlights the benefit of collaborative investigation to evaluate the international heterogeneity of this disease.

MATERIALS AND METHODS

Sample acquisition and sequencing

The sample acquisition and sequencing methods for the studies from the MUHAS Tanzania cohort (n=61) (36) and UNC Project - Malawi cohort (n=30) (11) have been previously described. Samples sequenced in the Mutographs study (n=210) (38) originated from patients in Golestan, Iran (n=55), ESCCAPE case-control studies in Tanzania (n=18) (19) and Kenya (n=64) (17), and patients in Shanxi, China (n=71). TCGA ESCC (n=36) and COAD samples (n=51) have been previously described (39, 41). The TCGA ESCC cohort includes tumors from patients in United States (n=3), Ukraine (n=3), Vietnam (n=22), and Russia (n=8), regions which have lower incidence of ESCC.

Metagenomic analysis

GATK-PathSeq (40) was used to conduct computational subtraction of human-mapping reads from input RNAseq and WGS datasets. GATK-PathSeq works by first mapping reads to a host reference database consisting of the human genome grch38 and various supplemental human reference sequences. Next, non-human reads are mapped against a comprehensive microbial database, and microbe read assignments are reported for further study. From the MUHAS Tanzania cohort, a total of 61 tumor

WGS samples, 45 saliva WGS samples, and 59 RNAseq samples were processed through GATK-PathSeq.

Bacterial abundance analyses and plotting were conducted in R (v3.5.1). To calculate relative abundance at a phylogenetic level (e.g., phylum or genus), GATK-PathSeq results were filtered for taxa at the level, and relative abundance was calculated for each taxon as follows: (# of taxon reads)/(total # reads at the selected phylogenetic level). The rows of all bacterial abundance heatmaps are arranged according to the mean abundance across all samples. The sample order of relative abundance stacked barplots were determined based on *Fusobacterium* genus relative abundance except where noted. In **Figure 2D**, if any cohort contained more than 50 samples, 50 samples were randomly selected for plotting. The distribution of relative abundances of genera of interest in all samples can be found in **Figure S2**, where width of each violin represents the relative distribution of observed bacterial relative abundance for all patients in each patient cohort.

Jaccard distance between RNAseq and WGS data from each ESCC tumor was calculated in R based on bacterial genera with at least 1% relative abundance. The qualitative Jaccard index was used in this case because the comparison was between DNA and RNA analytes which would not be expected to be quantitatively identical.

Tumor-saliva similarity

Only tumor-saliva pairs from the MUHAS Tanzania cohort with at least 10,000 reads mapped to the bacterial superkingdom were available for analysis. This resulted in a total of 21 tumor-oral pairs. Bray-Curtis dissimilarity metrics between tumor-oral pairs were calculated using the R package vegan (55). **Figure 3A** presents the Bray-Curtis *similarity* (1 – Bray-Curtis dissimilarity), for each tumor-oral pair.

To determine the correlation between the relative abundance of specific genera between tumor and saliva microbiomes, common-abundant genera that are at least 1% abundance in at least 3 tumor-oral pairs were identified. This resulted in the identification of 16 common-abundant genera. Correlations represent a two-sided Pearson correlation coefficient. To determine tumor-oral enrichment of common-abundant genera, the difference in relative abundance of each genus between each tumor-oral pair was plotted (**Figure 3C**). For the relative abundance bar plots of tumor-saliva pairs (**Figure 3D**), bacterial genera that had been highlighted in previous figures are labeled.

Code and processed data availability

All GATK-PathSeq output files and reproducible analysis and plotting R Notebooks are available.

Zenodo: https://doi.org/10.5281/zenodo.4750577

GitHub: https://github.com/jnoms/ESCC_microbiome

Furthermore, all analysis and figures can be automatically reproduced through a series of Google Colab documents.

Figure 1 and Supplementary Figure 1:
https://github.com/jnoms/ESCC_microbiome/blob/main/collab/Figure1.ipynb

Figure 2 and Supplementary Figure 2:
https://github.com/jnoms/ESCC_microbiome/blob/main/collab/Figure2.ipynb

Figure 3: https://github.com/jnoms/ESCC_microbiome/blob/main/collab/Figure3.ipynb

ABBREVIATIONS

AFRECC – African Esophageal Cancer Consortium

COAD – Colon adenocarcinoma

ESCA – Esophageal adenocarcinoma

ESCC – Esophageal squamous cell carcinoma

ESCCAPE – Esophageal Squamous Cell Carcinoma African Prevention Research

MUHAS – Muhimbili University of Health and Allied Sciences

RNAseq – RNA sequencing

TCGA – The Cancer Genome Atlas

WGS – Whole genome sequencing

ACKNOWLEDGMENTS

We thank Liu et al. (37) for providing access to sequencing data from ESCC patients in Malawi (dbGaP accession phs001448.v1.p1). Portions of this research were conducted using the O2 High Performance Computing Cluster, supported by the Research Computing Group at Harvard Medical School. We thank Aleksandar Kostic for his helpful discussions. We thank the teams from MUHAS-ORCI-UCSF Cancer Collaboration, UNC-Malawi Project, Mutographs, ESCCAPE, and TCGA for specimen and data collection.

CONFLICTS OF INTEREST

M.M. receives research support from Bayer, Novo, Ono, and Janssen, has patents licensed to Bayer and Labcorp, and is a consultant for Bayer, Interline and OrigiMed. J.A.D. receives research support from Constellation Pharmaceuticals and is a consultant to EMD Serono, Inc. and to Merck & Co. Inc. SB is a consultant for X-Biotix and BiomX. KVL receives research funding from Celgene Cancer Carelinks^{TM}. D.L.N. receives research funding from Cepheid, Inc.

FUNDING SOURCES

National Institutes of Health, National Cancer Institute Cancer Center Administrative Supplement to Promote Cancer Prevention and Control Research in Low and Middle Income Countries, A119617, [CA0082629] to K.V.L. S.B. received funding from NIH/NCI Grants: R00CA229984 and Cancer Center Support Grant P30 CA015704. EAC received funding from NIH/NCI Grants: U24CA210974, R01CA222862, R01CA227807, R01CA239604, R01CA230263. This work was supported in part by the US Public Health Service grants R35CA232128 and P01CA203655 to J.A.D. Content does not reflect the views of the National Cancer Institute or National Institutes of Health. This work was supported by a Cancer Grand Challenges OPTIMISTICC team award (C10674/A27140) to M.M. and a Cancer Grand Challenges Mutographs team award funded by Cancer Research UK [C98/A24032] to M.R.S. and P.B. R21CA191965 supported the collection of ESCCAPE Kenya tumors. The Internatinal Agency for Research on Cancer Section of Environment and Radiation supported the collection of ESCCAPE Tanzania tumors. This work was supported in part by the Intramural Research Program of the National Institutes of Health, National Cancer Institute, Division of Cancer Epidemiology and Genetics.

FIGURE LEGENDS

Figure 1. Microbiome structure and composition of ESCC tumors

A. Description of ESCC patients, and sample types, assessed in this study. TCGA – The Cancer Genome Atlas; ESCCAPE – Esophageal Squamous Cell Carcinoma African Prevention Research; Mutographs – Cancer Research UK Mutographs Project.

B. Bacterial burden of ESCC tumors for each patient cohort. Units are bacterial reads per million human reads as determined by GATK-PathSeq analysis. Each dot represents one sample. Analyte type (RNA or DNA) and tumor type (ESCC or COAD) are indicated by color.

C. Shannon diversity of ESCC tumors for each patient cohort. Shannon diversity was determined for each sample at the genus level based on genera that are at least 1% relative abundance. Each dot represents one sample. Analyte type (RNA or DNA) and tumor type (ESCC or COAD) are indicated by color.

D. Heatmap describing the relative abundance of the five top phyla sorted by average phylum relative abundance. Each column represents one sample. Rows represent the indicated phyla. Units are relative abundance. Samples from each cohort are WGS unless noted with “(RNA)”, in which case they are RNAseq.

Figure 2. Identification of bacterial genera associated with carcinogenesis

A. Bacterial genera relative abundance of WGS data from the MUHAS Tanzania cohort. Each column represents a single sample. Samples are ordered by decreasing Fusobacterium relative abundance. Units are relative abundance of bacterial genus-mapping reads. Color indicates the genus, and seven genera are specified. Only patients with GATK-PathSeq analysis from both RNAseq and WGS tumor data are plotted (n=59). Columns are ordered by decreasing relative abundance of *Fusobacterium* genus reads.

B. Bacterial genera relative abundance of RNAseq data from the MUHAS Tanzania cohort. Each column represents a single sample. Here, column order is dictated according to the patient order in Figure 2A. Units are relative abundance of bacterial genus-mapping reads. Color indicates the genus, and seven genera are specified. Only patients with GATK-PathSeq analysis from both RNAseq and WGS tumor data are plotted (n=59). Samples are ordered in the same order as Figure 2A, which is by *Fusobacterium* genus relative abundance in the WGS data.

C. Jaccard index between RNAseq and WGS data of tumors from the MUHAS Tanzania cohort. For the “Paired by Sample” column, Jaccard indices were calculated only between the WGS and RNAseq data from the same tumor (n=59 comparisons). For the “Random Pairs” column, Jaccard indices were calculated between all possible WGS-RNAseq pairs independent of patient of origin to represent the expected random distribution of Jaccard indices (n=3,481 comparisons). Jaccard index was calculated from relative abundance at the genus level based on genera that are at least 1% relative abundance. The width of the violin represents the relative proportion of comparisons with each Jaccard index, and lines indicate 25th, 50th, and 75th percentiles.

D. Bacterial genera relative abundance of the remaining patient cohorts, including RNAseq and WGS data as indicated. Each column represents a single sample. Samples are ordered by decreasing Fusobacterium relative abundance within each patient cohort. Units are relative abundance of bacterial genus-mapping reads. Color indicates the genus, and seven genera are specified. Here, if there were more than 50 samples in a patient cohort, 50 samples were randomly selected for visualization. USA – United States, UA – Ukraine, RU – Russia. All cohorts consist of WGS data, with the exception of the tumors from Malawi which are RNAseq. (Number of samples plotted: UNC Project - Malawi 30; ESCCAPE

Tanzania 18; ESCCAPE Kenya 50; Shanxi, China 50; Golestan, Iran 50; TCGA ESCC Vietnam 22; TCGA ESCC USA/UA/RU 14).

Figure 3. Association between synchronous saliva and tumor microbiomes in Tanzanian ESCC patients

A. Bray Curtis Similarity comparing tumor-saliva pairs from patients in the MUHAS Tanzania cohort. Analysis was restricted to the 21 tumor-saliva pairs that contained at least 10,000 bacterial reads. This analysis was conducted at the genus level and using relative abundance. For the “Paired by Patient” column, Bray Curtis Similarity was calculated only between the tumor and saliva WGS data from the same patient. For the “Random Pairs” column, Bray Curtis Similarity was calculated between all possible tumor-saliva pairs independent of patient of origin to represent the expected random distribution of Bray Curtis Similarity. (p=0.0003, Wilcoxon rank sum test).

B. Correlation between the relative abundance of common-abundant bacterial genera in paired saliva and tumor WGS data. Analysis was restricted to the 21 tumor-saliva pairs that contained at least 10,000 bacterial reads. Common-abundant bacterial genera are bacterial genera that are at least 1% abundance in at least 3 tumor-saliva pairs – 16 bacterial genera made this cutoff. Correlation represents a two-sided Pearson correlation. X-axis is the correlation coefficient, and Y axis is the correlation P-Value plotted on a log scale.

C. Enrichment of genera in the oral or tumor microbiome. Each row details one of the 16 common-abundant bacterial genera. Each row contains one data point per patient, for a total of 21 data points. The value of each point represents the difference in the relative abundance of the specified genus in the tumor and oral microbiomes of one patient, with positive values indicating a genus is at higher relative abundance in a patient's tumor. For example, if a genus is at a relative abundance of 0.7 (70%) in the tumor and 0.3 (30%) in the saliva of a patient, the plotted value for that genus and that patient is 0.4. Curves represent the distribution of this relative abundance difference across the tumor-oral pairs, with dots indicating individual tumor-oral pairs. Vertical red lines indicate quartiles.

D. Relative abundance bar charts of tumor-saliva pairs. Analysis was restricted to the 21 tumor-saliva pairs that contained at least 10,000 bacterial reads. Units are relative abundance of bacterial genus-mapping reads. Color indicates the genus, and seven genera are specified. (abbreviations: T – tumor, S – saliva).

Figure S1. GATK-PathSeq statistics and extended phyla and genera information

A. Boxplots indicating the number of GATK-PathSeq Human-mapped reads and GATK-PathSeq microbe-mapped reads for each patient cohort. Samples from each cohort are WGS unless noted with "(RNA)", in which case they are RNAseq.

B. Heatmap describing the relative abundance of the 15 top phyla sorted by average phylum relative abundance. Each column represents one sample. Rows represent the indicated phyla. Units are relative abundance. Samples from each cohort are WGS unless noted with "(RNA)", in which case they are RNAseq.

C. Heatmap describing the relative abundance of the 15 top genera sorted by average genera relative abundance. Each column represents one sample. Rows represent the indicated genera. Units are relative abundance. Samples from each cohort are WGS unless noted with “(RNA)”, in which case they are RNAseq.

D. Boxplot representing the Shannon diversity of genera that fall within the phylum *Actinobacteria* for each patient in each cohort. Samples from each cohort are WGS unless noted with “(RNA)”, in which case they are RNAseq.

Figure S2. Distribution of *Fusobacterium, Selenomonas, Prevotella, Streptococcus, Porphyromonas, Veillonella,* and *Campylobacter* relative abundance of genus reads for all samples in each study

A. The distribution of the relative abundance of genus-mapping reads for seven selected genera in all studies. The width of each violin represents the proportion of samples which have the indicated relative abundance of each genus. In contrast to **Figure 2D**, which only plots up to 50 samples per study, this plot includes all patients. Samples from each study are WGS unless noted with “(RNA)”, in which case they are RNAseq.

REFERENCES

1. 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: A Cancer Journal for Clinicians.n/a(n/a).
2. CC MRA, Dawsey S. Oesophageal cancer: A tale of two malignancies. World Cancer Report: Cancer Research for Cancer Prevention Lyon, France: International Agency for Research on Cancer Available from: http://publications.iarc.fr/586. 2020.
3. Coleman HG, Xie S-H, Lagergren J. The epidemiology of esophageal adenocarcinoma. Gastroenterology. 2018;154(2):390-405.
4. Rustgi AK, El-Serag HB. Esophageal carcinoma. New England Journal of Medicine. 2014;371(26):2499-509.
5. Arnold M, Soerjomataram I, Ferlay J, Forman D. Global incidence of oesophageal cancer by histological subtype in 2012. Gut. 2015;64(3):381-7.
6. Cheng ML, Zhang L, Borok M, Chokunonga E, Dzamamala C, Korir A, et al. The incidence of oesophageal cancer in Eastern Africa: identification of a new geographic hot spot? Cancer epidemiology. 2015;39(2):143-9.
7. Abnet CC, Kamangar F, Islami F, Nasrollahzadeh D, Brennan P, Aghcheli K, et al. Tooth loss and lack of regular oral hygiene are associated with higher risk of esophageal squamous cell carcinoma. Cancer Epidemiology and Prevention Biomarkers. 2008;17(11):3062-8.
8. Abnet CC, Qiao Y-L, Mark SD, Dong Z-W, Taylor PR, Dawsey SM. Prospective study of tooth loss and incident esophageal and gastric cancers in China. Cancer Causes & Control. 2001;12(9):847-54.
9. Dar N, Islami F, Bhat G, Shah I, Makhdoomi M, Iqbal B, et al. Poor oral hygiene and risk of esophageal squamous cell carcinoma in Kashmir. British journal of cancer. 2013;109(5):1367-72.
10. Chen X, Yuan Z, Lu M, Zhang Y, Jin L, Ye W. Poor oral health is associated with an increased risk of esophageal squamous cell carcinoma-a population-based case-control study in China. International journal of cancer. 2017;140(3):626-35.
11. Sato F, Oze I, Kawakita D, Yamamoto N, Ito H, Hosono S, et al. Inverse association between toothbrushing and upper aerodigestive tract cancer risk in a Japanese population. Head & neck. 2011;33(11):1628-37.
12. Liang H, Yang Z, Wang JB, Yu P, Fan JH, Qiao YL, et al. Association between oral leukoplakia and risk of upper gastrointestinal cancer death: a follow-up study of the Linxian general population trial. Thoracic cancer. 2017;8(6):642-8.
13. Guha N, Boffetta P, Wünsch Filho V, Eluf Neto J, Shangina O, Zaridze D, et al. Oral health and risk of squamous cell carcinoma of the head and neck and esophagus: results of two multicentric case-control studies. American journal of epidemiology. 2007;166(10):1159-73.
14. Chen Q-L, Zeng X-T, Luo Z-X, Duan X-L, Qin J, Leng W-D. Tooth loss is associated with increased risk of esophageal cancer: evidence from a meta-analysis with dose-response analysis. Scientific reports. 2016;6(1):1-7.
15. Sheikh M, Poustchi H, Pourshams A, Etemadi A, Islami F, Khoshnia M, et al. Individual and combined effects of environmental risk factors for esophageal cancer

based on results from the Golestan Cohort Study. Gastroenterology. 2019;156(5):1416-27.

16. Patel K, Wakhisi J, Mining S, Mwangi A, Patel R. Esophageal cancer, the topmost cancer at MTRH in the Rift Valley, Kenya, and its potential risk factors. International Scholarly Research Notices. 2013;2013.

17. Menya D, Maina SK, Kibosia C, Kigen N, Oduor M, Some F, et al. Dental fluorosis and oral health in the African Esophageal Cancer Corridor: Findings from the Kenya ESCCAPE case–control study and a pan-African perspective. International journal of cancer. 2019;145(1):99-109.

18. Mmbaga EJ, Mushi BP, Deardorff K, Mgisha W, Akoko LO, Paciorek A, et al. A Case–Control Study to Evaluate Environmental and Lifestyle Risk Factors for Esophageal Cancer in Tanzania. Cancer Epidemiology and Prevention Biomarkers. 2020.

19. Mmbaga BT, Mwasamwaja A, Mushi G, Mremi A, Nyakunga G, Kiwelu I, et al. Missing and decayed teeth, oral hygiene and dental staining in relation to esophageal cancer risk: ESCCAPE case-control study in Kilimanjaro, Tanzania. International journal of cancer. 2020.

20. Buckle GC, et al. Risk factors associated with early-onset esophageal cancer in Tanzania. (Under Review).

21. Ishaq S, Nunn L. Helicobacter pylori and gastric cancer: a state of the art review. Gastroenterology and hepatology from bed to bench. 2015;8(Suppl1):S6.

22. Kostic AD, Chun E, Robertson L, Glickman JN, Gallini CA, Michaud M, et al. Fusobacterium nucleatum potentiates intestinal tumorigenesis and modulates the tumor-immune microenvironment. Cell host & microbe. 2013;14(2):207-15.

23. Kostic AD, Gevers D, Pedamallu CS, Michaud M, Duke F, Earl AM, et al. Genomic analysis identifies association of Fusobacterium with colorectal carcinoma. Genome research. 2012;22(2):292-8.

24. Castellarin M, Warren RL, Freeman JD, Dreolini L, Krzywinski M, Strauss J, et al. Fusobacterium nucleatum infection is prevalent in human colorectal carcinoma. Genome research. 2012;22(2):299-306.

25. Bullman S, Pedamallu CS, Sicinska E, Clancy TE, Zhang X, Cai D, et al. Analysis of Fusobacterium persistence and antibiotic response in colorectal cancer. Science. 2017;358(6369):1443-8.

26. Gur C, Ibrahim Y, Isaacson B, Yamin R, Abed J, Gamliel M, et al. Binding of the Fap2 protein of Fusobacterium nucleatum to human inhibitory receptor TIGIT protects tumors from immune cell attack. Immunity. 2015;42(2):344-55.

27. Rubinstein MR, Baik JE, Lagana SM, Han RP, Raab WJ, Sahoo D, et al. Fusobacterium nucleatum promotes colorectal cancer by inducing Wnt/β-catenin modulator Annexin A1. EMBO reports. 2019;20(4):e47638.

28. Rubinstein MR, Wang X, Liu W, Hao Y, Cai G, Han YW. Fusobacterium nucleatum promotes colorectal carcinogenesis by modulating E-cadherin/β-catenin signaling via its FadA adhesin. Cell host & microbe. 2013;14(2):195-206.

29. Gao S, Li S, Ma Z, Liang S, Shan T, Zhang M, et al. Presence of Porphyromonas gingivalis in esophagus and its association with the clinicopathological characteristics and survival in patients with esophageal cancer. Infectious agents and cancer. 2016;11(1):3.

30. Whitmore SE, Lamont RJ. Oral bacteria and cancer. PLoS pathogens. 2014;10(3):e1003933.
31. Inaba H, Sugita H, Kuboniwa M, Iwai S, Hamada M, Noda T, et al. P orphyromonas gingivalis promotes invasion of oral squamous cell carcinoma through induction of pro MMP 9 and its activation. Cellular microbiology. 2014;16(1):131-45.
32. He Z, Gharaibeh RZ, Newsome RC, Pope JL, Dougherty MW, Tomkovich S, et al. Campylobacter jejuni promotes colorectal tumorigenesis through the action of cytolethal distending toxin. Gut. 2019;68(2):289-300.
33. Narikiyo M, Tanabe C, Yamada Y, Igaki H, Tachimori Y, Kato H, et al. Frequent and preferential infection of Treponema denticola, Streptococcus mitis, and Streptococcus anginosus in esophageal cancers. Cancer science. 2004;95(7):569-74.
34. Boleij A, Schaeps RM, Tjalsma H. Association between Streptococcus bovis and colon cancer. Journal of clinical microbiology. 2009;47(2):516-.
35. Ferrari A, Botrugno I, Bombelli E, Dominioni T, Cavazzi E, Dionigi P. Colonoscopy is mandatory after Streptococcus bovis endocarditis: a lesson still not learned. Case report. World journal of surgical oncology. 2008;6(1):49.
36. Van Loon K, et al. A Genomic Analysis of Esophageal Squamous Cell Carcinoma in Eastern Africa. (Under Review).
37. Liu W, Snell JM, Jeck WR, Hoadley KA, Wilkerson MD, Parker JS, et al. Subtyping sub-Saharan esophageal squamous cell carcinoma by comprehensive molecular analysis. JCI insight. 2016;1(16).
38. Moody S, Senkin S, Islam SMA, Wang J, Nasrollahzadeh D, Penha RCC, et al. Mutational signatures in esophageal squamous cell carcinoma from eight countries of varying incidence. medRxiv. 2021:2021.04.29.21255920.
39. Network CGAR. Integrated genomic characterization of oesophageal carcinoma. Nature. 2017;541(7636):169.
40. Walker MA, Pedamallu CS, Ojesina AI, Bullman S, Sharpe T, Whelan CW, et al. GATK PathSeq: a customizable computational tool for the discovery and identification of microbial sequences in libraries from eukaryotic hosts. Bioinformatics. 2018;34(24):4287-9.
41. Network CGA. Comprehensive molecular characterization of human colon and rectal cancer. Nature. 2012;487(7407):330.
42. Dohlman AB, Arguijo Mendoza D, Ding S, Gao M, Dressman H, Iliev ID, et al. The cancer microbiome atlas: a pan-cancer comparative analysis to distinguish tissue-resident microbiota from contaminants. Cell Host & Microbe. 2020.
43. Tsay J-CJ, Wu BG, Sulaiman I, Gershner K, Schluger R, Li Y, et al. Lower airway dysbiosis affects lung cancer progression. Cancer Discovery. 2020.
44. Ricotta C, Podani J. On some properties of the Bray-Curtis dissimilarity and their ecological meaning. Ecological Complexity. 2017;31:201-5.
45. Flemer B, Warren RD, Barrett MP, Cisek K, Das A, Jeffery IB, et al. The oral microbiota in colorectal cancer is distinctive and predictive. Gut. 2018;67(8):1454-63.
46. Komiya Y, Shimomura Y, Higurashi T, Sugi Y, Arimoto J, Umezawa S, et al. Patients with colorectal cancer have identical strains of Fusobacterium nucleatum in their colorectal cancer and oral cavity. Gut. 2019;68(7):1335-7.

47. Yamamura K, Baba Y, Nakagawa S, Mima K, Miyake K, Nakamura K, et al. Human microbiome Fusobacterium nucleatum in esophageal cancer tissue is associated with prognosis. Clinical Cancer Research. 2016;22(22):5574-81.

48. Peters BA, Wu J, Pei Z, Yang L, Purdue MP, Freedman ND, et al. Oral microbiome composition reflects prospective risk for esophageal cancers. Cancer research. 2017;77(23):6777-87.

49. Kawasaki M, Ikeda Y, Ikeda E, Takahashi M, Tanaka D, Nakajima Y, et al. Oral infectious bacteria in dental plaque and saliva as risk factors in patients with esophageal cancer. Cancer. 2021;127(4):512-9.

50. Corning B, Copland AP, Frye JW. The esophageal microbiome in health and disease. Current gastroenterology reports. 2018;20(8):1-7.

51. Dewhirst FE, Chen T, Izard J, Paster BJ, Tanner AC, Yu W-H, et al. The human oral microbiome. Journal of bacteriology. 2010;192(19):5002-17.

52. Signat B, Roques C, Poulet P, Duffaut D. Role of Fusobacterium nucleatum in periodontal health and disease. Curr Issues Mol Biol. 2011;13(2):25-36.

53. Michaud DS, Lu J, Peacock-Villada AY, Barber JR, Joshu CE, Prizment AE, et al. Periodontal disease assessed using clinical dental measurements and cancer risk in the ARIC study. JNCI: Journal of the National Cancer Institute. 2018;110(8):843-54.

54. Ahrens W, Pohlabeln H, Foraita R, Nelis M, Lagiou P, Lagiou A, et al. Oral health, dental care and mouthwash associated with upper aerodigestive tract cancer risk in Europe: the ARCAGE study. Oral oncology. 2014;50(6):616-25.

55. Oksanen J, Kindt R, Legendre P, O'Hara B, Stevens MHH, Oksanen MJ, et al. The vegan package. Community ecology package. 2007;10:631-7.

Figure 1

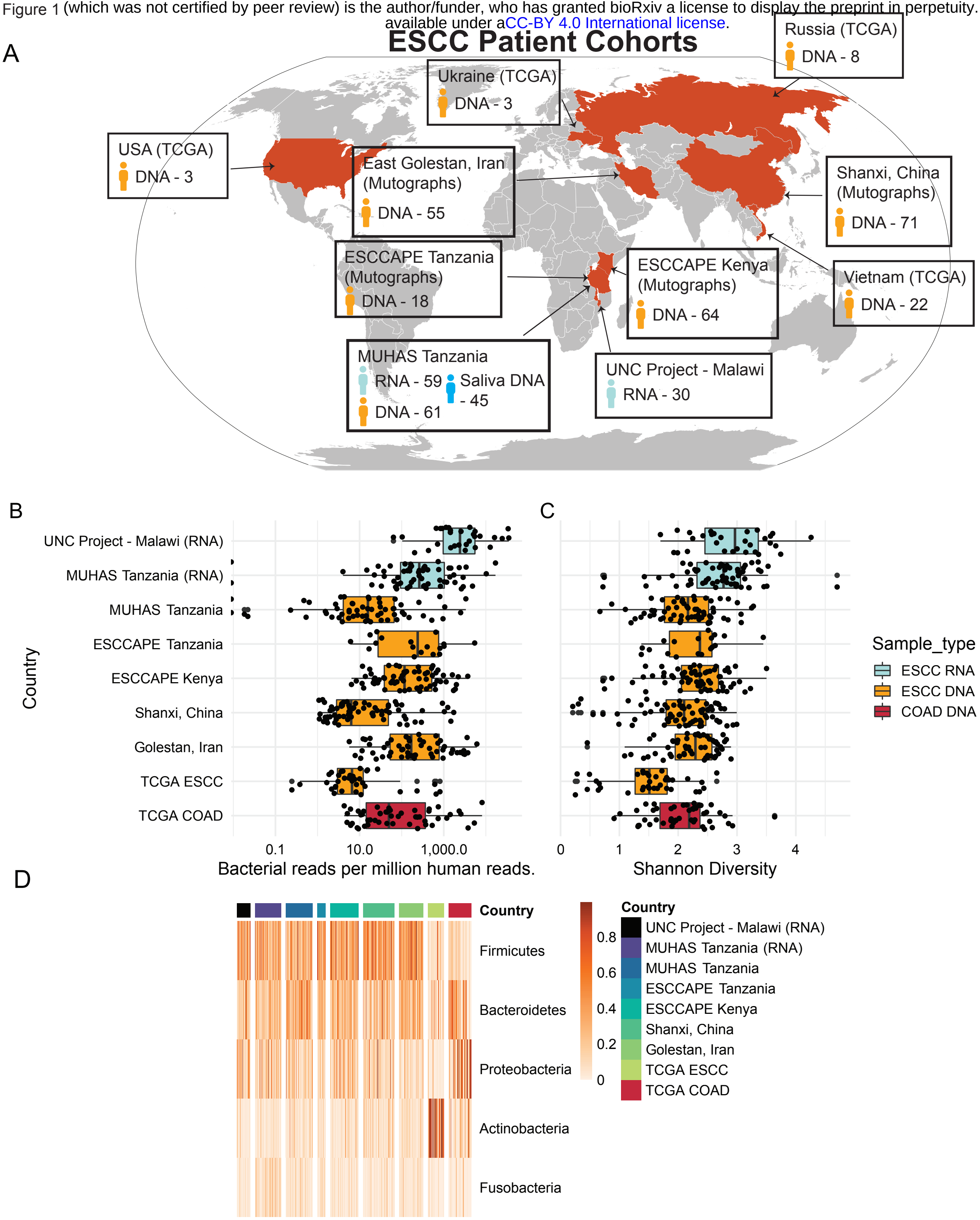

Figure 1. Microbiome structure and composition of ESCC tumors.

A. Description of ESCC patients, and sample types, assessed in this study. TCGA – The Cancer Genome Atlas; ESCCAPE – Esophageal Squamous Cell Carcinoma African Prevention Research; Mutographs – Cancer Research UK Mutographs Project.

B. Bacterial burden of ESCC tumors for each patient cohort. Units are bacterial reads per million human reads as determined by GATK-PathSeq analysis. Each dot represents one sample. Analyte type (RNA or DNA) and tumor type (ESCC or COAD) are indicated by color.

C. Shannon diversity of ESCC tumors for each patient cohort. Shannon diversity was determined for each sample at the genus level based on genera that are at least 1% relative abundance. Each dot represents one sample. Analyte type (RNA or DNA) and tumor type (ESCC or COAD) are indicated by color.

D. Heatmap describing the relative abundance of the five top phyla sorted by average phylum relative abundance. Each column represents one sample. Rows represent the indicated phyla. Units are relative abundance. Samples from each cohort are WGS unless noted with "(RNA)", in which case they are RNAseq.

Figure 2

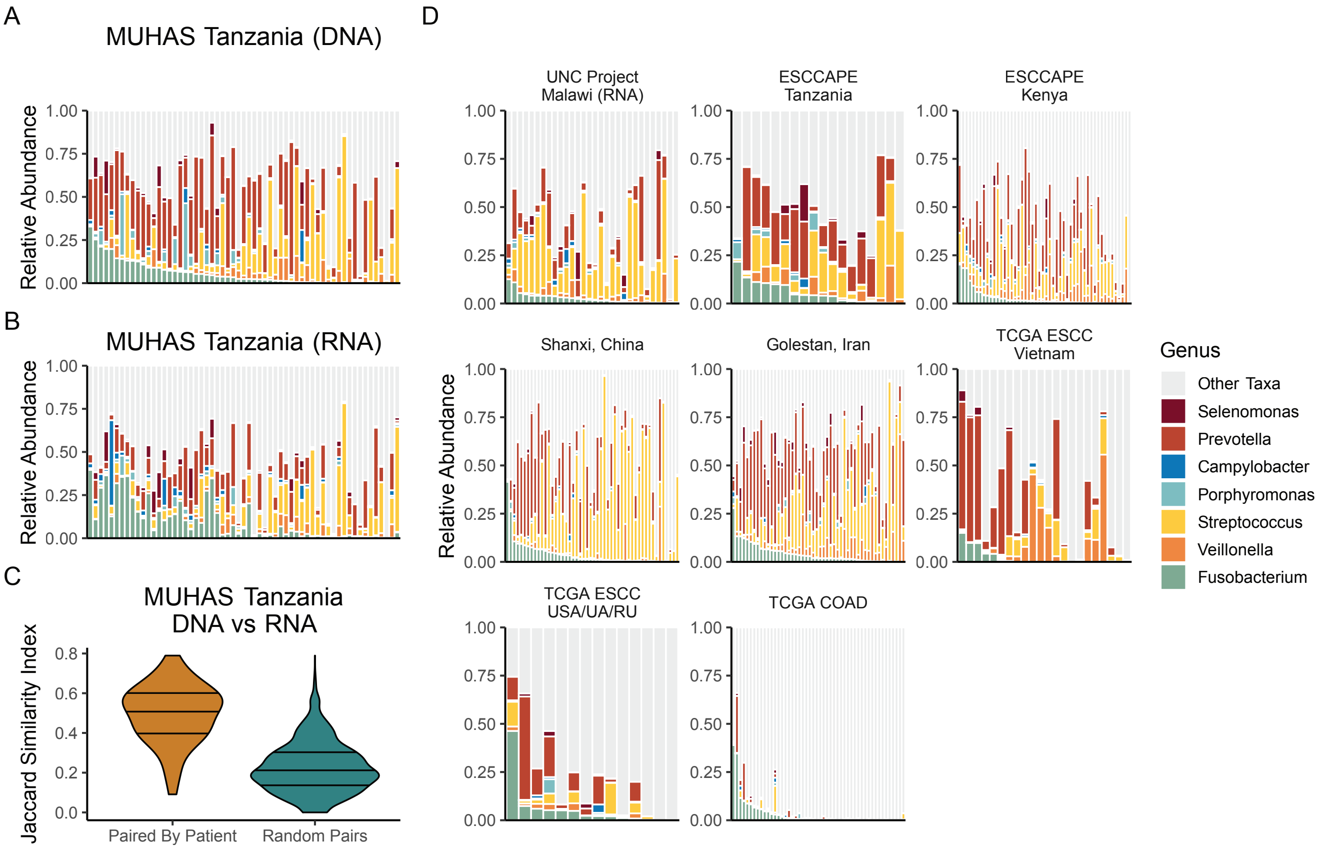

Figure 2. Identification of bacterial genera associated with carcinogenesis.

A. Bacterial genera relative abundance of WGS data from the MUHAS Tanzania cohort. Each column represents a single sample. Samples are ordered by decreasing Fusobacterium relative abundance. Units are relative abundance of bacterial genus-mapping reads. Color indicates the genus, and seven genera are specified. Only patients with GATK-PathSeq analysis from both RNAseq and WGS tumor data are plotted (n=59). Columns are ordered by decreasing relative abundance of Fusobacterium genus reads.

B. Bacterial genera relative abundance of RNAseq data from the MUHAS Tanzania cohort. Each column represents a single sample. Here, column order is dictated according to the patient order in Figure 2A. Units are relative abundance of bacterial genus-mapping reads. Color indicates the genus, and seven genera are specified. Only patients with GATK-PathSeq analysis from both RNAseq and WGS tumor data are plotted (n=59). Samples are ordered in the same order as Figure 2A, which is by Fusobacterium genus relative abundance in the WGS data.

C. Jaccard index between RNAseq and WGS data of tumors from the MUHAS Tanzania cohort. For the "Paired by Sample" column, Jaccard indices were calculated only between the WGS and RNAseq data from the same tumor (n=59 comparisons). For the "Random Pairs" column, Jaccard indices were calculated between all possible WGS-RNAseq pairs independent of patient of origin to represent the expected random distribution of Jaccard indices (n=3,481 comparisons). Jaccard index was calculated from relative abundance at the genus level based on genera that are at least 1% relative abundance. The width of the violin represents the relative proportion of comparisons with each Jaccard index, and lines indicate 25th, 50th, and 75th percentiles.

D. Bacterial genera relative abundance of the remaining patient cohorts, including RNAseq and WGS data as indicated. Each column represents a single sample. Samples are ordered by decreasing Fusobacterium relative abundance within each patient cohort. Units are relative abundance of bacterial genus-mapping reads. Color indicates the genus, and seven genera are specified. Here, if there were more than 50 samples in a patient cohort, 50 samples were randomly selected for visualization. USA – United States, UA – Ukraine, RU – Russia. All cohorts consist of WGS data, with the exception of the tumors from Malawi which are RNAseq. (Number of samples plotted: UNC Project - Malawi 30; ESCCAPE Tanzania 18; ESCCAPE Kenya 50; Shanxi, China 50; Golestan, Iran 50; TCGA ESCC Vietnam 22; TCGA ESCC USA/UA/RU 14).

Figure 3

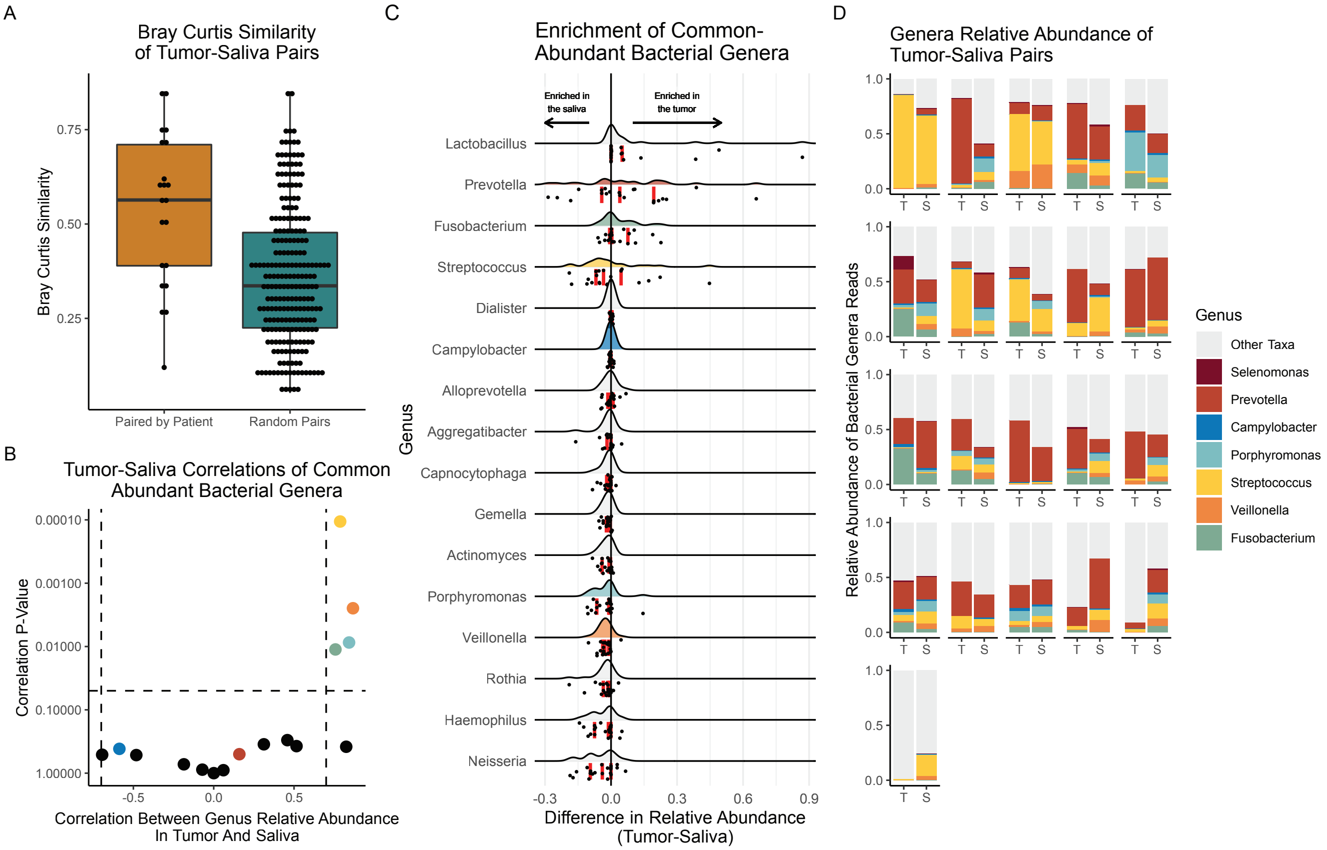

Figure 3. Association between synchronous saliva and tumor microbiomes in Tanzanian ESCC patients.

A. Bray Curtis Similarity comparing tumor-saliva pairs from patients in the MUHAS Tanzania cohort. Analysis was restricted to the 21 tumor-saliva pairs that contained at least 10,000 bacterial reads. This analysis was conducted at the genus level and using relative abundance. For the "Paired by Patient" column, Bray Curtis Similarity was calculated only between the tumor and saliva WGS data from the same patient. For the "Random Pairs" column, Bray Curtis Similarity was calculated between all possible tumor-saliva pairs independent of patient of origin to represent the expected random distribution of Bray Curtis Similarity. (p=0.0003, Wilcoxon rank sum test).

B. Correlation between the relative abundance of common-abundant bacterial genera in paired saliva and tumor WGS data. Analysis was restricted to the 21 tumor-saliva pairs that contained at least 10,000 bacterial reads. Common-abundant bacterial genera are bacterial genera that are at least 1% abundance in at least 3 tumor-saliva pairs – 16 bacterial genera made this cutoff. Correlation represents a two-sided Pearson correlation. X-axis is the correlation coefficient, and Y axis is the correlation P-Value plotted on a log scale.

C. Enrichment of genera in the oral or tumor microbiome. Each row details one of the 16 common-abundant bacterial genera. Each row contains one data point per patient, for a total of 21 data points. The value of each point represents the difference in the relative abundance of the specified genus in the tumor and oral microbiomes of one patient, with positive values indicating a genus is at higher relative abundance in a patient's tumor. For example, if a genus is at a relative abundance of 0.7 (70%) in the tumor and 0.3 (30%) in the saliva of a patient, the plotted value for that genus and that patient is 0.4. Curves represent the distribution of this relative abundance difference across the tumor-oral pairs, with dots indicating individual tumor-oral pairs. Vertical red lines indicate quartiles.

D. Relative abundance barcharts of tumor-saliva pairs. Analysis was restricted to the 21 tumor-saliva pairs that contained at least 10,000 bacterial eads. Units are relative abundance of bacterial genus-mapping reads. Color indicates the genus, and seven genera are specified. (Abbreviations: T – tumor, S – saliva.)

Figure S1

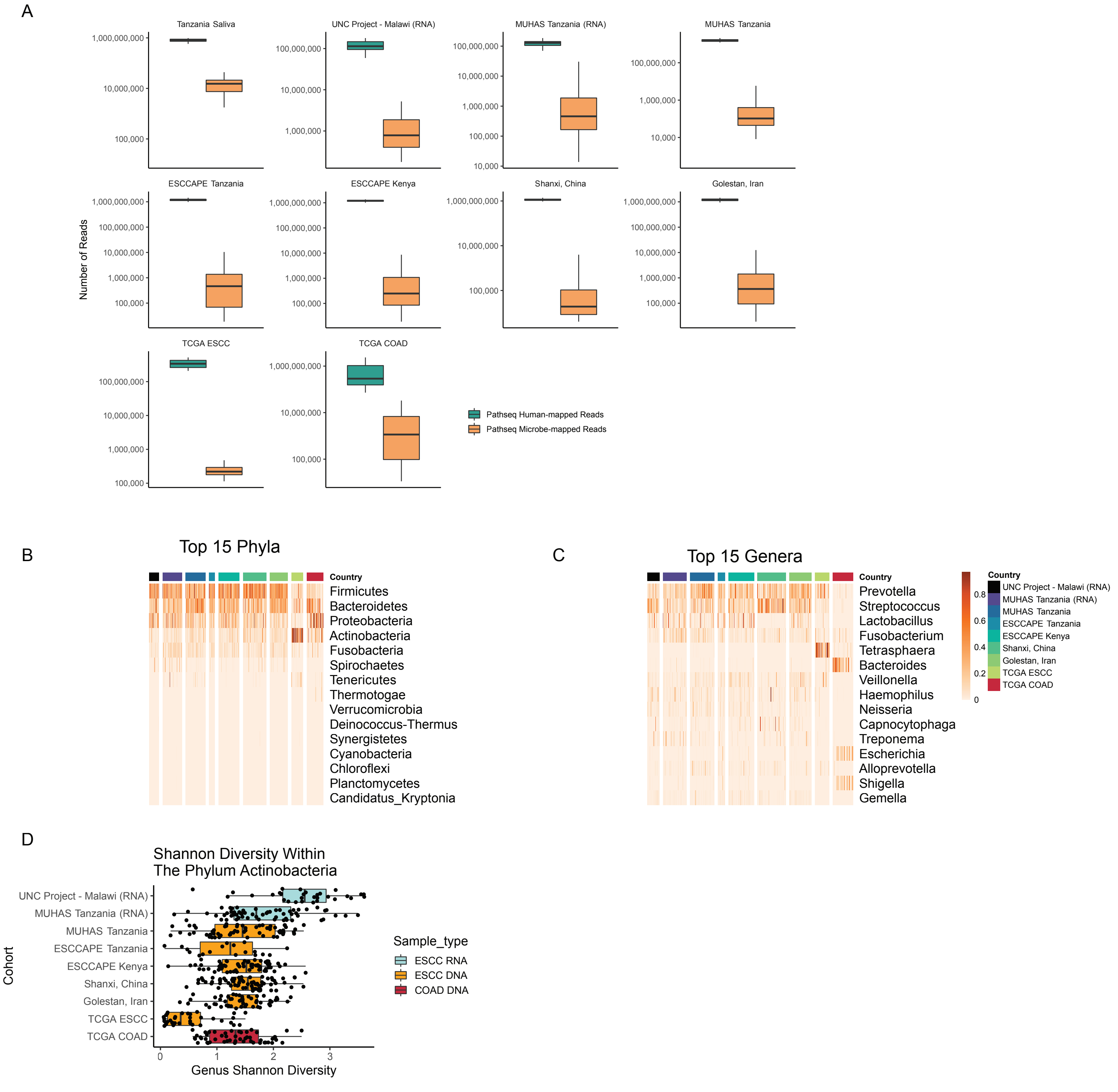

Figure S1. GATK-PathSeq statistics and extended phyla and genera information.

A. Boxplots indicating the number of GATK-PathSeq Human-mapped reads and GATK-PathSeq microbe-mapped reads for each patient cohort. Samples from each cohort are WGS unless noted with "(RNA)", in which case they are RNAseq.

B. Heatmap describing the relative abundance of the 15 top phyla sorted by average phylum relative abundance. Each column represents one sample. Rows represent the indicated phyla. Units are relative abundance. Samples from each cohort are WGS unless noted with "(RNA)", in which case they are RNAseq.

C. Heatmap describing the relative abundance of the 15 top genera sorted by average genera relative abundance. Each column represents one sample. Rows represent the indicated genera. Units are relative abundance. Samples from each cohort are WGS unless noted with "(RNA)", in which case they are RNAseq.

D. Boxplot representing the Shannon diversity of genera that fall within the phylum Actinobacteria for each patient in each cohort. Samples from each cohort are WGS unless noted with "(RNA)", in which case they are RNAseq.

Figure S2

A

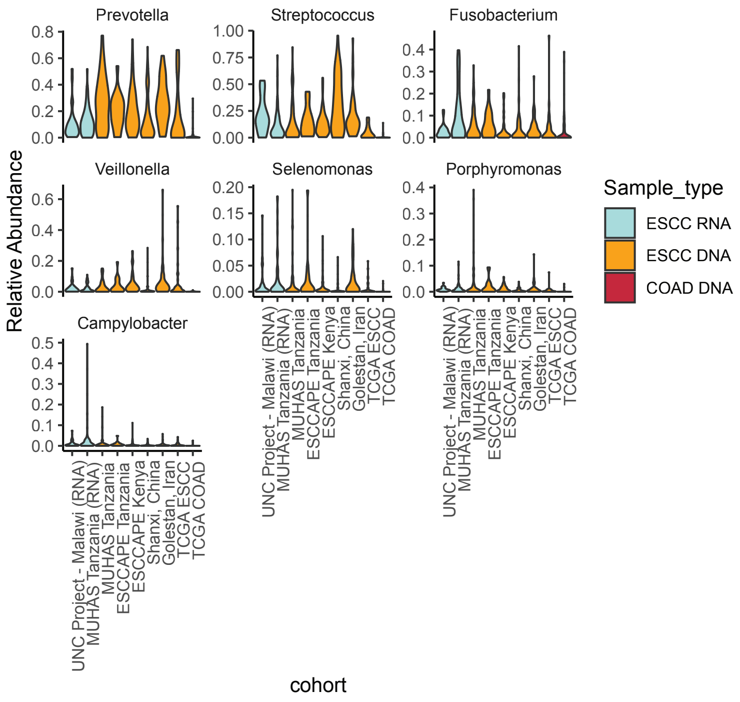

Figure S2. Distribution of Fusobacterium, Selenomonas, Prevotella, Streptococcus, Porphyromonas, Veillonella, and Campylobacter relative abundance of genus reads for all samples in each study.

A. The distribution of the relative abundance of genus-mapping reads for seven selected genera in all studies. The width of each violin represents the proportion of samples which have the indicated relative abundance of each genus. In contrast to Figure 2D, which only plots up to 50 samples per study, this plot includes all patients. Samples from each study are WGS unless noted with "(RNA)", in which case they are RNAseq.

TABLE 1						
Study	**Tanzania**	**Malawi****	**ESCCAPE Tanzania*****	**ESCCAPE Kenya*****	**East Golestan, Iran******	**Shanxi, China******
No. cases included	61	30	18	65	55	71
Demographics						
Median age (IQR)	49 (44-62)	56	65 (61-73)	64 (53, 71)	62 (54,73)	56 (50, 64)
% male	67%	45.8%	61%	68%	55%	56%
Status at diagnosis						
Weight (kg), median (IQR)			44 (40-52)	52 (46, 60)		
Body mass index (kg/m^2) median (IQR)			15.8 (15.4, 19.1)	19.5 (15.6, 22.0)		
Median months ill before coming to endoscopy (IQR)			2 (1, 6)	3 (2, 4.5)		
HIV status: Positive Negative Not known	2 (3.2%) 36 (59.0%) 23 (37.7%)	10 (16.9%) 44 (74.6%) 5 (8.5%)	1 (5%) 10 (56%) 7 (39%)	5 (8%) 48 (74%) 12 (18%)		
Key lifestyle habits						
N (%) ever tobacco users			11 (61%)	38 (58%)	17 (31%)	35 (49%)
N (%) who brush teeth daily: With toothbrush With stick			 12 (67%) 6 (33%)	 16 (25%)* 10 (15%)		
Median no missing teeth (IQR)			3 (1, 5)	4 (1, 8)		

*N=22 (34%) brush once per week or never, n=17 (26%) brush 2 to 6 times/week

**Indicates demographics are from the entire patient population, consisting of both included and unincluded patients.

***Indicates demographic percentages are from the entire patient population, with discrete counts scaled to the number of cases included.

****Indicates demographic information is exclusively for included patients.