Computational and Structural Biotechnology Journal 18 (2020) 631–641

COMPUTATIONAL AND STRUCTURAL BIOTECHNOLOGY JOURNAL

journal homepage: www.elsevier.com/locate/csbj

The landscape of bacterial presence in tumor and adjacent normal tissue across 9 major cancer types using TCGA exome sequencing

Rebecca M. Rodriguez [a,c], Brenda Y. Hernandez [b,c], Mark Menor [a], Youping Deng [a,*], Vedbar S. Khadka [a,*]

[a] Bioinformatics Core, Department of Quantitative Health Sciences, John A. Burns School of Medicine, University of Hawaii Mnoa, Honolulu, HI, United States
[b] Epidemiology, University of Hawaii Cancer Center, University of Hawaii, Honolulu, HI, United States
[c] Population Sciences in the Pacific Program-Cancer Epidemiology, University of Hawaii Cancer Center, Honolulu, HI, United States

ARTICLE INFO

Article history:
Received 1 October 2019
Received in revised form 2 March 2020
Accepted 6 March 2020
Available online 13 March 2020

Keywords:
Microbial landscape
Cancer microbiome
TCGA
Exome sequencing

ABSTRACT

Identification of microbial composition directly from tumor tissue permits studying the relationship between microbial changes and cancer pathogenesis. We interrogated bacterial presence in tumor and adjacent normal tissue strictly in pairs utilizing human whole exome sequencing to generate microbial profiles. Profiles were generated for 813 cases from stomach, liver, colon, rectal, lung, head & neck, cervical and bladder TCGA cohorts. Core microbiota examination revealed twelve taxa to be common across the nine cancer types at all classification levels. Paired analyses demonstrated significant differences in bacterial shifts between tumor and adjacent normal tissue across stomach, colon, lung squamous cell, and head & neck cohorts, whereas little or no differences were evident in liver, rectal, lung adenocarcinoma, cervical and bladder cancer cohorts in adjusted models. *Helicobacter pylori* in stomach and *Bacteroides vulgatus* in colon were found to be significantly higher in adjacent normal compared to tumor tissue after false discovery rate correction. Computational results were validated with tissue from an independent population by species-specific qPCR showing similar patterns of co-occurrence among *Fusobacterium nucleatum* and *Selenomonas sputigena* in gastric samples. This study demonstrates the ability to identify bacteria differential composition derived from human tissue whole exome sequences. Taken together our results suggest the microbial profiles shift with advanced disease and that the microbial composition of the adjacent tissue can be indicative of cancer stage disease progression.

1. Introduction

Bacteria have been associated with cancer progression exerting beneficial or detrimental effects depending on the time and site of the colonization [1,2]. Their highly site-specific colonization enables modulation of the tumor microenvironment [3,4]. Microbial-host dynamics can promote or inhibit host immune response [5,6]. These changes lead to the accumulation of insults and epigenetic changes that can change the course of a developing or established tumor [7]. Evidence has demonstrated that infection-associated cancer subtypes are molecularly distinct [8,9], which highlights the importance of microbial modulation within the tumor cells. These findings are significant and reveal important microbial patterns and mechanistic pathways in host-response to cancer. However, microbial composition in many of these studies has been derived from surrogate material like stool, saliva, or aspirate, rather than directly from the tumor and surrounding tissue.

The microbial presence within the tumor and adjacent tissue can inform disease progression and bacterial roles in cancer pathogenesis [10,30]. Recent studies suggest microbial presence information can be derived from human whole exome sequencing data [11], by computational subtraction, similar to transcriptomics or metagenomics methods. Bioinformatics tools facilitate profiling of tumor virome and bacteriome using human sequencing data in the context of cancer-associated pathogenesis [1216]. These methods have proven useful and are reason of much exploration. Most

Abbreviations: TCGA, The Cancer Genome Atlas; STAD, stomach adenocarcinoma; LIHC, liver hepatocellular carcinoma; COAD, colon adenocarcinoma; READ, rectal adenocarcinoma; COREAD, colon and rectal adenocarcinoma TCGA cohorts; LUSC, lung squamous cell carcinoma; LUAD, lung adenocarcinoma; HNSC, head & neck squamous cell carcinoma; CESC, cervical & endocervical squamous cell carcinomas; BLCA, bladder carcinoma; L2FC, log 2 fold change.

Corresponding authors.
E-mail addresses: dengy@hawaii.edu (Y. Deng), vedbar@hawaii.edu (V.S. Khadka).

https://doi.org/10.1016/j.csbj.2020.03.003

studies to date examining microbial composition (viral, bacterial, and other) using human sequencing data have done so using RNA sequencing data. For instance, Khoury et al. (2013) interrogated TCGA cancer cohorts to characterize viral DNA presence, and integration sites within tumor tissue. Khoury et al. described important HPV, HBV, and HHV-4 differential integration sites across TCGA cancer cohorts highlighting utility of RNA sequencing data for tumor virome characterization [17]. However, this study lacked validation in tumor tissue, either direct or cross-sectional.

Similarly, Tang et al. (2013) examined viral gene expression and host fusion building a viral expression map across TCGA cancer cohorts [18]. Salyakina et al. (2013) and Cao et al. (2016) also examined viral expression in tumor and normal specimens within TCGA cohorts across TCGA cohorts [19,20]. Cao et al. (2016) demonstrated the ability to identify associations between different viral strains and ethnic differences [20]. These works were all based on RNA sequencing derived pathogen information. Cantalupo et al. (2018) on the other hand, examined viral integration using RNA, whole exome and genome sequencing data across 22 of the TCGA cancer cohorts mapping viral prevalence differences and commonalities within the sample population [21]. None of these findings were experimentally validated.

Similar to viral profiling, Riley et al. (2013) examined bacterial DNA integration in 852 TCGA tumor and normal specimens [22]. They discovered significant bacterial gene integration within various TCGA cohorts. However, the highest integration rates were detected in groups for which no matched or paired normal sample data were available, including stomach adenocarcinoma and acute myeloid leukemia [22]. Robinson et al. (2017) later examined potential bacterial contamination across 5 TCGA cohorts including acute myeloid leukemia, breast, glioblastoma, ovarian and stomach adenocarcinomas using RNA sequencing data from paired tumor and adjacent normal samples [23]. Potential contaminants were present across all cohorts such as *Staphylococcus epidermis*, *Cutibacterium acnes*, and *Ralstonia* species after controlling for batch effects [23]. Like Riley (2013), Robinson et al. (2017) did not include experimental validation. On the other hand, Zhang et al. (2015) developed a workflow for the identification of low abundant microbial species using whole exome and RNA sequencing data derived from the Human Genome Project based on PathSeq [24]. Zhang offered experimental validation in gastric biopsies against TCGA whole genome sequencing data [25]. This study demonstrated the ability to identify low abundant microbes within the tumor.

Other studies examining tumor microbiota derived from human sequences have looked at mutation interaction and gene expression associations in one or a few cancers. Thompson et al. (2017) examined bacteria composition and gene expression profiles in TCGA breast cancer cohort [26]. Thompson found that bacteria presence correlates with genes that regulate tumor growth pathways. This study offered experimental validation via direct 16S rRNA sequencing of bacterial presence with samples from whom RNA sequencing data had initially been derived [26]. Greathouse et al. (2018) examined the lung microbiome association with TP53 mutation using 16S rRNA methods and confirmed findings with TCGA lung cancer data [27]. These studies highlight the feasibility of microbial profile identification and functional characterization from human sequencing data; however, the use of human RNA sequencing data may not be the best approach at characterizing bacterial signatures. Use of RNA sequencing could reflect cDNA library artifacts rather than actual RNA abundance [28]. Whole exome sequencing data, on the contrary, may provide a better picture because it represents protein-coding region of DNA, and is not subject to library artifacts [29].

To our knowledge, no studies have yet examined cross-cancer microbial composition differential profiling using whole exome sequencing data from tumor and adjacent normal in a strict paired design. Here we have interrogated tumor and adjacent normal tissue from paired solid cancers cases from TCGA. We generated bacterial composition across 9 cancer types (STAD: stomach adenocarcinoma, LIHC: liver hepatocellular carcinoma, COAD: colon adenocarcinoma, READ: rectal adenocarcinoma, LUSC: lung squamous cell carcinoma, LUAD: lung adenocarcinoma, HNSC: head & neck squamous cell carcinomas, CESC: cervical squamous cell carcinoma, and BLCA: bladder carcinoma) encompassing 3758 total tumor and adjacent normal sample files from 813 cases. We performed quantitative polymerase chain reaction (qPCR) in some selected differentially abundant taxa for validation on an independent sample population.

2. Methods

2.1. TCGA cancer database

We downloaded TCGA cancer types with whole exome sequencing case pairs meeting selection criteria. Cases were defined as solid tumor cancer types within TCGA that had human-aligned sequencing reads from exome sequencing in binary version of Sequence Alignment/Map (BAM) file format. The BAM files of primary tumor and adjacent solid tissue normal (paired cases) were selected at a 1:1 ratio along with available clinical data for the bioinformatics interrogation.

2.2. Computational framework for microbial detection

TCGA Level-1 data were used to derive microbial information. For microbiota identification, we used a computational pipeline designed to generate microbial profiles from human whole exome sequencing BAM files based on PathoScope 2.0 [15]. The pipeline pre-processed, quality filtered, and mapped all the BAM files using SAMtools and picard. Additional BLAST filtering step against hg38 reference genome was used to subtract any remaining human reads and remaining reads were simultaneously aligned using a custom library of known microbial genomes. Finally, the modified pipeline produced reports of quantified microbial proportions. Reports were used for bacterial differential analyses. Viral DNA detection, mainly HPV, HBV, and EBV, was used as internal pipeline validation by comparing viral detection rates with previous studies [1721]. Our pipeline was primarily designed to identify DNA sequences. So, RNA viruses like HCV were not detected. R-software was used for phylogenic classification and statistical analyses.

2.3. Core microbiota

Core taxa were defined as that identified at a minimum positive detection rate, present in the majority of the population, and shared between tumor and adjacent normal pairs with a minimum of 20% prevalence in each sample type. Identification of core microbiota was completed under study assumptions of relative abundance positivity threshold of 0.2% per microbe with a minimum prevalence in the population. Assuming that each identified taxa are present at least once in each sample with a minimum of 1 read, a positivity detection rate of 0.2% was deemed reasonable. Core taxa identification was verified using microbiome R-package (version 1.3.3) with default settings and detection and prevalence rates per study assumptions. Any species identified within each cohort

were then compared across all cohorts. Visualization of shared taxa was performed with UpSetR package (https://gehlenborglab.shinyapps.io/upsetr).

2.4. Diversity metrics and differential abundance analyses

Diversity measurements of alpha diversity (within sample diversity) and beta diversity (between samples) were completed for each cancer type. Mean differences of 15% are considered clinically relevant. Analyses were completed using R-software packages, phyloseq (v.1.25.3) and microbiome (v.1.3.3). Alpha (Shannon-Wiener Index, Simpson Index of Diversity and Fishers alpha) and Beta diversity were calculated using vegan R-package (v.2.5-3) and Microsoft Excel (v.2013). Measures were calculated at the taxonomy or operational taxonomic units (OTU) level and collated at the species level (by aggregating strains and subspecies of the same species). KRONA plots (https://github.com/marbl/Krona/wiki) were created for relative abundance visualization using Excel macro-enabled templates (Supplementary data FS2). Quantified proportions of bacteria and viruses generated from the bioinformatics pipeline were used to create plots. Including total per microbe read count, average reads per microbe, percent population prevalence, and relative abundance data. Differential relative abundance was determined using the Wilcoxon Signed Rank test within R-platform. Bacterial taxa with false discovery rate (FDR) adjusted p-value <0.05 were considered significant at the genus and species level. To determine the association between differences in relative abundance in tumor and its adjacent normal and clinical features, paired or unpaired *t*-test and analysis of variance (ANOVA) were used for two- and multi-group comparisons, respectively. Equivalent non-parametric tests were used for non-normally distributed data and to account for the compositional structure of microbial relative abundances. Chi-square test was used for categorical data. Linear regression models were used to adjust for clinical and demographical confounders.

2.5. PCR validation

We experimentally validated bioinformatics findings with de-identified archival tissue from the Hawaii Tumor Registry-Discard Residual Repository (RTR), a unique collection of formalin-fixed, paraffin-embedded (FFPE) tissue from cancer patients diagnosed within the catchment area of the Hawaii Tumor Registry. The Hawaii Tumor Registry is one of three population-based registries associated with the National Cancer Institute (NCI) and the Surveillance, Epidemiology, and End-Results (SEER) program. Archival tissue from a total of 85 paired cases from gastric (21), and colon (64) cancers were selected for validation. Specimen retrieval, cut & slide, sectioning, pathology review, and nucleic acid extraction were performed by the University of Hawaii Cancer Center Pathology Shared Resources. DNA was extracted from FFPE using Qiagen All Prep FFPET Kit (Qiagen, Valencia, CA) and quantified by NanoDrop spectrophotometer (Thermo-Scientific, Wilmington, DE). PCR reactions were completed using 30 ng of DNA for every 25 çl of reaction mix using commercially available species-specific primer-probe kits (Microbial DNA qPCR assay kits 330033, Qiagen, Valencia CA) per manufacturer's instructions under the following conditions: Activation: 10 min 95¯C, followed by 45 cycles of Denaturation and Annealing at 95¯C for 15 s and 60¯C for 2 min. Samples were tested in duplicates plus positive and negative controls. Discrepancies were resolved by repeat qPCR. Species-specific validation was performed for *Helicobacter pylori*, *Bacteroides vulgatus*, *Fusobacterium nucleatum*, and *Selenomonas sputigena*.

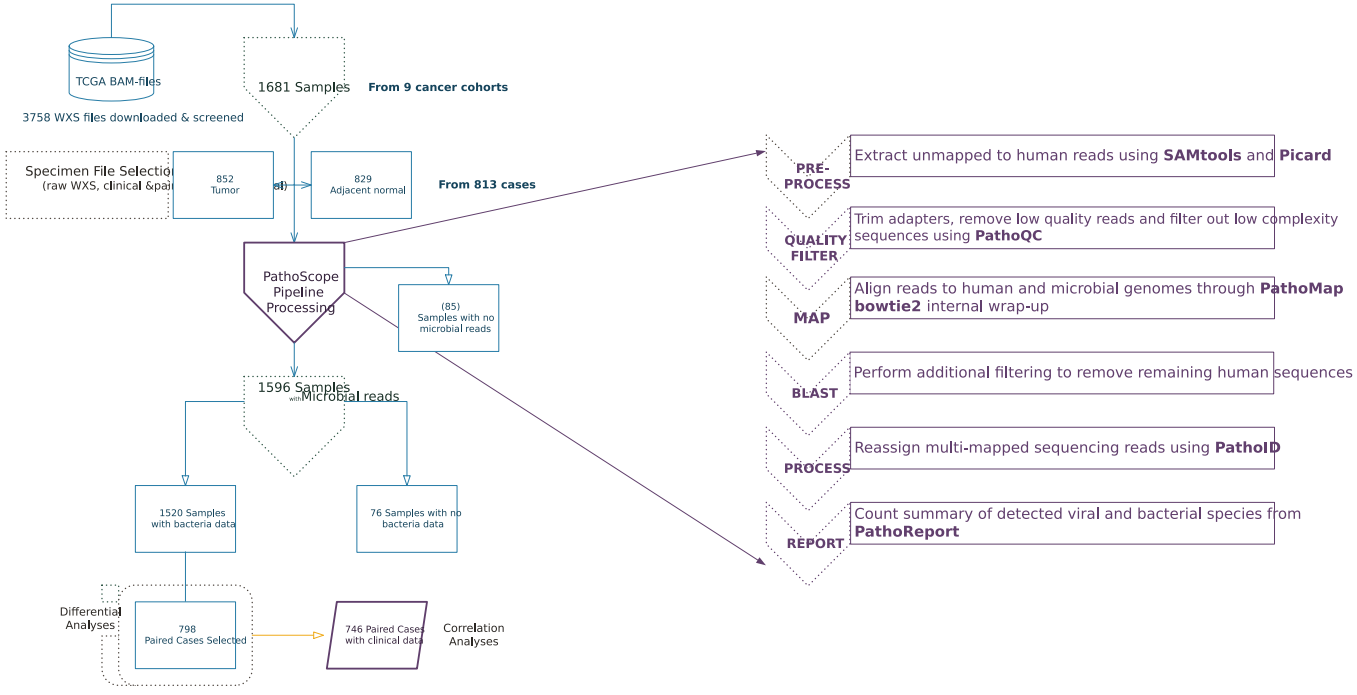

Fig. 1. Sample selection workflow and computational pipeline designed to extract microbial profiles based on PathoScope 2.0. Whole exome sequencing files (3758), from 813 cases were downloaded. From these cases, a total of 1681 sample sequences were processed through our modified pipeline (852 tumor and 829 adjacent normal). Bioinformatics pipeline includes additional filtering step described by Zhang et al. 2015. Additional filtering step completed against human reference genome (hg38) and simultaneously aligned to custom library of known microbial genomes. Relative abundance of PathoReport is calculated for each microbe based on normalized values in tumor and adjacent normal tissues. Detection of DNA viral sequences used as internal validation. Sample sequences without microbial reads in at least 1 pair tissue (tumor or adjacent normal) sample were removed. From PathoReport, strict one-to-one pairs with microbial reads were selected for bacterial differential analyses (a total of 1596 samples from 798 paired cases). Demographics and correlation analyses with clinical data were completed for cases with available data (746).

3. Results

3.1. Identification of microbial sequences in TCGA exome data

We generated microbial profiles for whole exome sequencing files from 1690 samples representing 813 paired cases across 9 TCGA cancer cohorts using a modified PathoScope 2.0 workflow (Fig. 1). Microbial reads were detected in 83% of the total samples screened (Supplementary data Table S1). Primary tumor and its paired adjacent normal with detected bacterial reads on either sample were selected at a 1:1 ratio for analyses (Supplementary data Table S2). We detected bacterial DNA presence in 94% of the primary tumors and 92% adjacent solid tissue normal samples, while viral DNA presence was 33% and 35%, respectively (Table 1). The highest proportion of viral DNA positivity was detected in colon and cervical cancers. Colorectal (COAD and READ) and HNSC cohorts were found to have the highest percentage of cases with bacterial DNA. The lowest proportion of samples with any bacterial DNA presence was observed in BLCA cancer cohort (76% of the samples).

3.2. Population characteristics

Out of 813 paired cases, only 746 cases had clinical data available and were used in association analyses (Table 2). Of the total, 69% of cases were White (independent of Hispanic origin), 9% were African American (independent of Hispanic origin), 4% Asian, and 1% were of other racial groups, aggregated to protect privacy. Age at diagnosis ranged from 20 to 90 years (mean˜ SD: 64˜ 11.9 yrs). There were some expected differences in age at diagnosis among the cancer cohorts with youngest population in CESC cohort (47˜ 13.5 yrs) and oldest in COAD cohort (71˜ 12.3 yrs). There was an 8% difference overall in the proportion of females to males (54% vs. 46% respectively). Approximately 48% of the tumors were classified as stage II or III.

3.3. Taxonomic composition

Our pipeline detected a total of 1,264,775 quality microbial reads representing 1353 unique bacteria taxa, from which 882 were shared across cancer cohorts for tumor and adjacent normal combined (Fig. 2). From these, 12 species were present in all cohorts including *Actinomyces oris*, *Bradyrhizobium* sp. BTAi1 *Bradyrhizobium* sp. ORS, *Cutibacterium acnes*, *Escherichia coli*, *Leptothix cholodnii*, *Neisseria sicca*, *Ralstonia insidiosa*, *Rhodopseudomonas palustris*, *Shingomonas melonis*, *Sphingomonas panacis* and *Bradyrhizobium diazoefficiens* (Table S3). Species from eight major phyla were found among all cancer cohorts with significant differential relative abundances (p-value <0.05). One of the critical findings was bacteria profile shifts observed in tumor compared to adjacent normal in STAD, COAD, CESC and BLCA cancer types. Meanwhile in LIHC, LUAD, and READ bacterial shifts were less apparent at the phylum level (Fig. 3). Taxa from Proteobacteria phylum were found in all nine cancer types. Firmicutes were higher in STAD tumor and BLCA adjacent normal compared to their paired corresponding paired tissues. Fusobacteria taxa were present at low levels across STAD, LUSC, HNSC, and COAD.

Bacteroidetes were highest in COAD compared to other cohorts. In STAD tumor we observed a 13% rise in the level of Bacteroidetes compared to its adjacent normal, and a 22% rise in the level of Firmicutes species which, composed nearly half of the total reads found in tumor while a 66% decrease was observed in the Proteobacteria like species in STAD tumor compared to its adjacent normal (Fig. 3; Table S4). In COAD, we detected a 31% decrease of Bacteroidetes and 33% increase in Proteobacteria in tumor tissue relative to adjacent normal. BLCA cohort cancer appeared to have the most considerable shift change in composition where the tumor was colonized almost entirely by Proteobacteria (96%) compared to adjacent normal which had a more diverse composition.

The distribution of the most abundant species in tumor and adjacent normal suggest that 24 species- *Bacillus subtilis*, *Cutibacterium acnes*, *Escherichia coli*, *Mycoplasma mycoides*, *Corynebacterium pseudotuberculosis*, *Ralstonia pickettii*, *Bacillus mycoides*, *Mitsuaria* sp. 7, *Streptomyces gilvosporeus*, *Bacteroides fragilis*, *Roseateles depolymerans*, *Psychromicrobium lacuslunae*, *Bacteroides thetaiotaomicron*, *Bacteroides dorei*, *Bacteroides ovatus*, *Bacteroides vulgatus*, *Bacteroides caecimuris*, *Alistipes finegoldii*, *Bradyrhizobium* sp. BTAi1, *Rothia mucilaginosa*, *Flavonifractor plautii*, *Arthrobacter* sp. IHBB 11108, *Sphingomonas koreensis,* and *Roseburia hominis* commonly co-occurred across cohorts (Table S4-B). Other species were found across cohorts with at least one read. Measures of total read absolute abundance, reads proportional relative abundance and percent prevalence provided different interpretation regarding the taxonomy compositional structure.

3.4. Core taxa characterization

To identify differences and commonalities of species shared between the tumor and its adjacent normal pair within each cohort and across cancer types, core microbiota were characterized

Table 1
Proportion of samples with microbial reads at any detection level.

TCGA Cohort	Samples with Bacteria in Tumor N (%)	Samples with Bacteria in Adjacent N (%)	Samples with Virus in Tumor N (%)	Samples with Virus in Adjacent N (%)
STAD (n = 176)	74/88 (84)	73/88 (83)	30/88 (34)	35/88 (40)
LIHC (n = 141)	68/71 (81)	66/70 (79)	13/71 (15)	17/70 (20)
COAD (n = 176)	88/88 (100)	88/88 (100)	88/88 (100)	88/88 (100)
READ (n = 36)	18/18 (100)	18/18 (100)	11/18 (61)	10/18 (56)
LUSC (n = 430)	211/221 (95)	209/221 (94)	63/221 (28)	81/221 (36)
LUAD (n = 393)	200/200 (100)	193/200 (97)	41/200 (21)	34/200 (17)
HNSC (n = 145)	72/72 (100)	71/73(97)	11/72 (16)	9/73 (13)
CESC (n = 16)	8/8 (100)	7/8 (88)	8/8 (100)	8/8 (100)
BLCA (n = 64)	28/35 (76)	28/29 (76)	2/35 (5)	1/29 (3)
Totals	767 (94)	753 (92)	267 (33)	283 (35)

Proportion of samples with microbial reads (bacterial and viral) prior to 1:1 pairing selection. There was no significant difference in the number of samples with microbial presence between tumor and its adjacent normal tissue for any cohort. Overall, we detected microbial reads in 94% of tumors and 92% of adjacent normal. We found DNA viral presence, mainly HHV-4, HPV and HBV in 33% of tumors and 35% of adjacent with cervical and colon cancers having 100% of samples with at least one viral read.

Table 2
Basic population demographic characteristics of 9 TCGA cancer cohorts.

	STAD N = 85	LIHC N = 81	COAD N = 88	READ N = 18	LUSC N = 221	LUAD N = 148	HNSC N = 69	CESC N = 8	BLCA N = 28	Totals N = 746
Race N (%)										
White	54 (64)	64 (79)	37 (42)	8 (44)	147 (67)	120 (8)	56 (81)	4 (50)	25 (89)	515 (69)
African American	3 (3)	6 (7)	7 (8)	1 (6)	16 (7)	23 (16)	9 (13)	1 (13)	2 (7)	68 (9)
Asian	16 (19)	7 (9)			3 (1)	2 (1)	1 (1)			29 (4)
Other Race			1 (1)			1 (1)		2 (25)		4 (1)
Not reported	12 (14)	4 (5)	43 (49)	9 (50)	55 (25)	2 (1)	3 (4)	1 (13)	1 (4)	130 (17)
Age at diagnosis										
Mean˜SD Range	67˜ 10.5 4188	64˜ 14.7 2086	71˜ 12.3 4090	63˜ 14.6 4090	68˜8.4 4085	65˜ 10.3 4087	63˜ 12.2 2688	47˜ 13.5 2269	69˜ 10.7 4890	64˜ 11.9 2090
Sex N (%)										
Male	48 (56)	46 (57)	47 (53)	7 (39)	64 (29)	63 (43)	48 (70)		19 (68)	342 (46)
female	37 (44)	35 (43)	41 (47)	11 (61)	157 (71)	85 (57)	21 (30)	8 (100)	9 (32)	404 (54)
Stage N (%)										
I	14 (17)	33 (41)	11 (13)	4 (22)	121 (55)	79 (53)	1 (1)	4 (50)	3 (11)	270 (36)
II-III	47 (55)	35 (43)	62 (70)	9 (50)	96 (43)	61 (41)	30 (43)	4 (50)	11 (39)	355 (48)
IV	8 (9)	3 (4)	14 (16)	4 (22)	3 (1)	6 (4)	38 (55)		14 (50)	90 (12)
No staging	16 (19)	10 (12)	1 (1)	1 (6)	1 (1)	2 (1)				31 (4)

Cases with paired tumor and adjacent tissue normal with available clinical data Clinical. Out of 798 cases, with bacterial presence, clinical data was available for 746. Largest fraction of cases without clinical data were from LUAD (n = 52). Other race include groups with 2 or less cases per group from Native American, Alaskan Native, Native Hawaiian or other Pacific Islander to maintain privacy. () indicates not available or not applicable

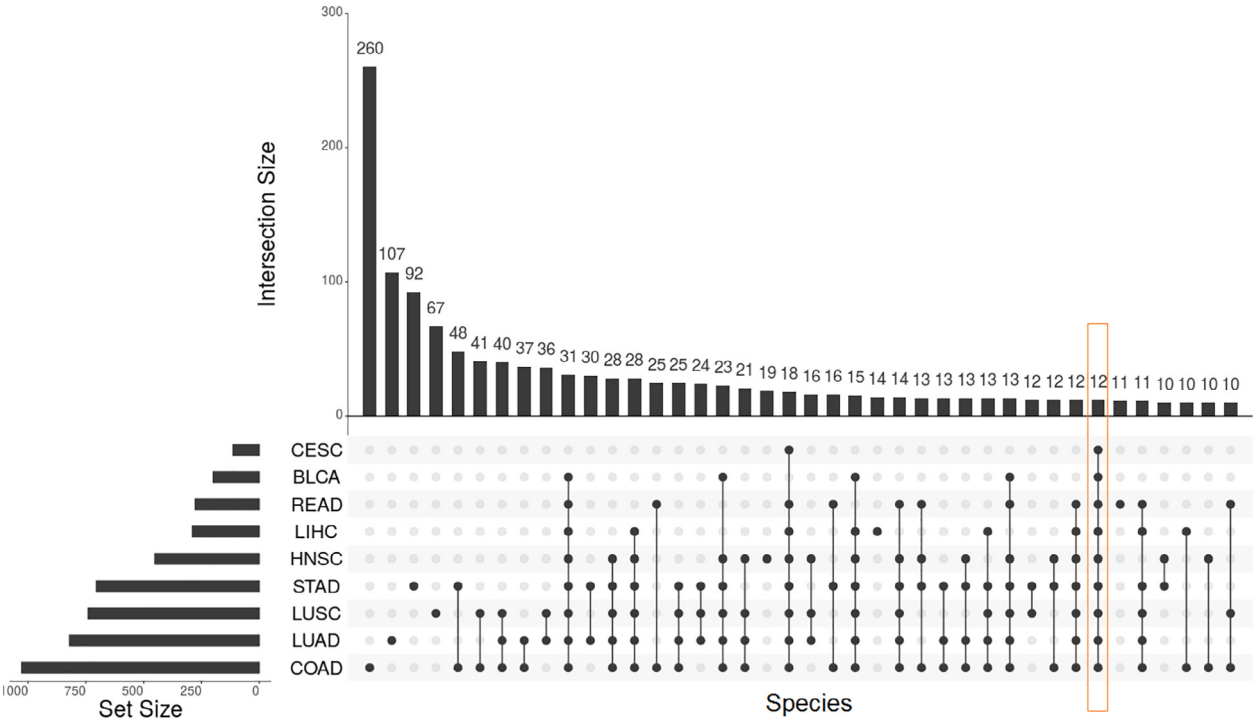

Fig. 2. Frequency of shared bacterial species across 9 TCGA cohorts. Compositional bar graph showing size of individual core taxonomies (left horizontal bars) and intersect of shared species (black dot and connecting lines) across cohorts with species frequency (vertical bars). Twelve taxa (yellow highlight), *Actinomyces oris, Bradyrhizobium sp.* BTAi1 *Bradyrhizobium sp.* ORS, *Cutibacterium acnes, Escherichia coli, Leptothix cholodnii, Neisseria sicca, Ralstonia insidiosa, Rhodopseudomonas palustris, Shingomonas melonis, Sphingomonas panacis and Bradyrhizobium diazoefficiens* were found to be shared across all nine cohorts at different rates, from which 3, *Bradyrhizobium sp.* BTAi1, *Cutibacterium acnes,* and *Escherichia coli,* were detected in both pairs tumor and adjacent normal tissue. Colon (COAD) had the greatest number of unique taxa (260 non-shared), while cervical (CESC) and bladder (BLCA) cancers had no unique species when comparing across cohorts.

(Table S5). Three species, *Escherichia coli*, *Cutibacterium acnes*, and *Bradyrhizobium* sp. BTAi1 were found to be present in all 9 cohorts. As part of taxa characterization we identified 12 species as core microbiota by study assumptions while *microbiome R-package* resulted in 24 core species (Table S5). *Bacillus subtilis* as the most frequent taxa in the population, present in 313 of the tumors and 351 of adjacent normal across five cohorts with a low proportion of reads and relative abundances. (Tables S4.1A-S4.3B). *Bradyrhizobium* like reads were 22 times higher in CESC tumor compared to its paired adjacent normal. Yet, *Bradyrhizobium* sp. BTAi1 relative abundances in core taxa calculations were lower in tumor than adjacent normal CESC samples. Similarly *Escherichia* spp. reads including *Escherichia coli*, *Escherichia fergusonii* and *Escherichia albertii* were detected in multiple cohorts. *Escherichia coli* relative abundances were also higher in CESC tumor. Overall, approximately 99% of the total reads detected in CESC were of viral origin

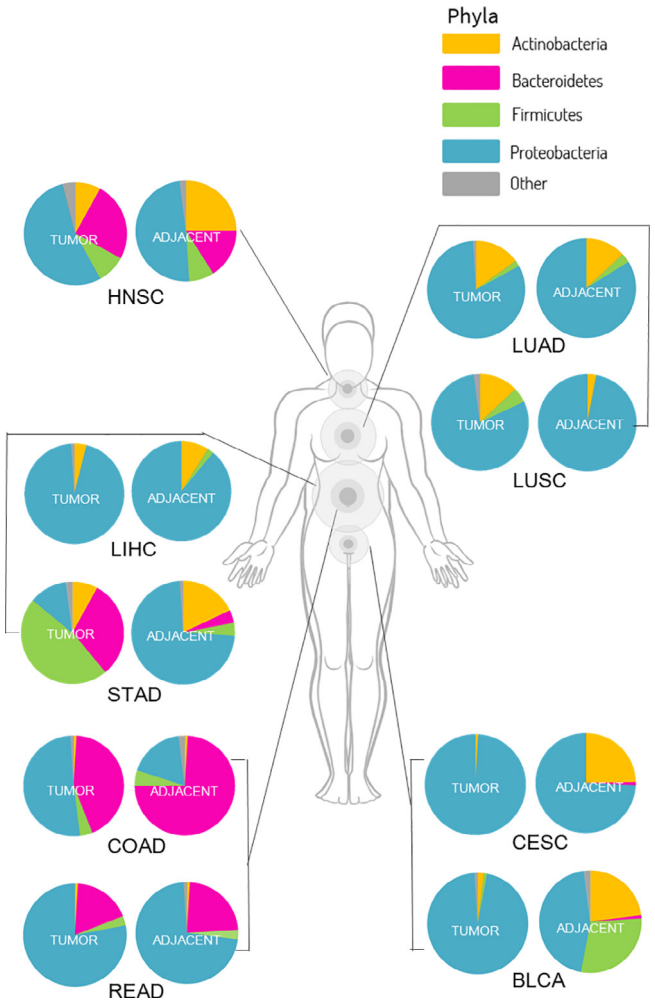

Fig. 3. The landscape of bacterial shift changes across 9 tumor types at the phylum level. Proportion of bacterial reads and compositional shifts at the phylum level and anatomical proximities per cancer type in tumor and adjacent normal tissues. Phyla >1% of the total reads per tissue is shown, phyla < 1% grouped as other and may include Verrucombia, Spirochaetes, Tenericutes, Fusobacteria, and Cyanobacteria primarily from colon, gastric and head and neck cancer cohorts. Significant shifts in bacterial composition are observed between the adjacent normal and tumor tissues that may indicate disease status or disease progression within the tumor microenvironment in the continuum of disease.

which may affect the impact the significance of these species as core taxa in CESC specifically.

3.5. Microbial diversity and cancer-specific findings

Species richness, the number of species per sample, was overall slightly higher in adjacent normal compared to tumor samples with an average of 358 species in tumor vs. 382 species in adjacent normal. Among STAD, and CESC and BLCA cohorts, richness was higher in tumor. Richness was higher in adjacent normal among COAD, LUSC and READ, while there were no differences noted in LIHC, LUAD, HNSC and cohorts at the threshold for >15% significant difference (Table S6). Diversity varied by age, sex, and histopathological staging at varying degrees across different cancer cohorts (Table S7) and is presented in detail in cancer specific findings. We compared the bacterial relative abundances and bacterial diversity in tumor and its paired adjacent solid tissue normal for each cancer type and across cancer groups as described in the methods section. All specimens with aligned bacterial reads were considered (Table S8).

3.5.1. Stomach

We examined 170 STAD paired primary tumor and adjacent normal sample WXS from 85 cases. Average read per sample was 360 in tumor and 107 in adjacent normal. The average numbers of species per sample in tumor samples were 24 compared to 19 in adjacent normal. There was a significant difference in the proportions of taxa in tumor compared to taxa numbers in adjacent normal independent of pairing (Fisher, p = 0.007). Four species, *Bacillus subtilis*, *Arthrobacter* sp. IHBB11108, *Cutibacterium acnes*, and *Mycoplasma mycoides* were found to be present in 25% or more of either sample type. *Selenomonas sputigena* was the most prevalent species in tumor samples with 13% of the total reads in tumor, while *Helicobacter pylori* strains made 60% of the total reads in adjacent normal. *Fusobacterium nucleatum* was detected in 9% (n = 8) of tumors with a median relative abundance of 0.001 (range 0.002, 0.25). (STAD Krona Plots Fig. S2). Differential relative abundance for 10 taxa representing 4 major phyla and 7 genus levels were found to be higher in tumor than in adjacent normal. However, this difference was not statistically significant (p = <0.05 FDR = 1). Presence of *Helicobacter pylori*, was found to be significantly higher in the adjacent normal compared to tumor tissue (log2fc = 4.8, p = <0.0001 FDR = 0.01) while *Veillonella parvula* was 16 times higher in tumor compared to adjacent normal (log2fc = 4.5, p = 0.03, FDR = 1) though not significant after multiple test correction. (Fig. S1, Table S8). Because gastric cancers molecular subtypes are associated with Epstein Barr virus, we evaluated the presence of HHV-4 as pipeline internal validation. We detected HHV-4 in 25 tumor and 25 adjacent normal samples. Status of HHV-4 did not differ within the paired sample population. However, the proportions of reads detected in tumor were significantly higher than those detected in adjacent normal tissue samples at a ratio of 102:1.

3.5.2. Liver

In LIHC *Escherichia coli* was the most abundant species, detected in 67% of cases (Table S5). Species richness was higher in males categorized at tumor stage I (estimate = 4.5˜ 2.1, p0.034). We found no difference in bacterial composition between tumor and adjacent normal in paired tests (Table S8) and no difference in alpha or beta diversity when stratifying by tumor stage and sex (Table S7). Given the viral etiology of LIHC, we evaluated HBV and HHV-4 viral reads as internal validation. HBV and HHV-4 represented <2% of the total tumor or adjacent normal reads, detected in 10 cases. The number of HBV reads in tumor samples were twice the number of reads in adjacent normal (LIHC Krona Plots-Fig. S2). HHV-4 was noted to be present only in tumor samples. Most samples with positive identification of HBV or HHV-4 did not have bacterial content. Of those with bacterial reads, most commonly co-occurred with Actinobacteria and Proteobacteria species (data not shown).

3.5.3. Colorectal cancers

A total of 212 samples from 88 colon (COAD) and 18 rectal (READ) paired cases were examined. In COAD, average read per sample mapping to bacterial genomes was 2006 reads in tumor compared to 3962 reads in adjacent normal (Table S2). Mean species per sample was significantly different in tumor compared to adjacent normal (mean difference = 12.3, p = 0.016, 95%CI 2.4, 22.2). For READ average read number per sample was 569 reads in tumor and 708 reads in adjacent normal, while the average number of species per sample was lower in tumor with 20 species per sample compared to 29 in adjacent normal (p = 0.008). Both colon and rectal cancers had a small proportion (<1%) of reads mapping to viral genomes across 7 families. A number of Torque-teno-virus like reads were also identified (Fig. S2).

The ratio of Proteobacteria to Bacteriodetes was significantly increased in colon tumor compared to its adjacent normal (log2

P/B tumor = 0.24, log2 P/B adjacent normal = 2.03). There was no difference in within sample diversity index or the evenness spread (t = 1.35, p = 0.18, 95%CI = 0.017, 0.005) by sample type. We wanted to know if differences existed when stratifying by sex, age at diagnosis, race and tumor stage. Intra-sample diversity measured by Shannon-Wiener diversity index did not differ by sex or age group (p = 0.46). *Bacteroides vulgatus* was found to be significantly different (log2fc = 0.8, p = <0.00001, FDR = 0.001) between sample types, however the log 2 fold change was negligible at 0.8 higher in adjacent normal compared to tumor (Table S8). Multiple studies have reported overabundance of *Fusobacterium nucleatum* in tumor tissue associated with colorectal cancer pathogenesis [3135]. Based on these reports we wanted to evaluate the presence of *Fusobacterium nucleatum* in our data set. Overall, Fusobacteria reads represented less than 1% of the total mapped reads in COAD and READ. Of these, 84% were identified as *Fusobacterium nucleatum*. We found that there were considerable differences between detected *Fusobacterium* reads in tumor and those detected in adjacent normal specimens within the COAD cohort. The relative abundance means within tumor and within adjacent normal samples differed significantly (tumor p = <0.0001 FDR = 0.002, adjacent normal p = 0.006 FDR = 0.05, respectively). However, in paired test by Wilcox Sign Rank, mean relative abundance differences were non-significant when comparing tumor to adjacent normal samples after multiple test correction (log2fc = 2.39, p = 0.003, FDR = 0.52). In READ, differential abundance of Fusobacteria reads were negligible.

3.5.4. Cancers of the lung

We analyzed sequencing files from 421 cases of lung squamous cell carcinoma (LUSC) and lung adenocarcinoma (LUAD). Of these, 23 cases in LUSC and 17 cases in LUAD had one or both samples without microbial reads or only viral reads. Clinical data was not available for 26% of LUAD cases (52 /200 cases). Stratified by sex and sample type, there were no significant differences in age at diagnosis, race, ethnicity, primary diagnosis, tumor stage, or survival days between the cohorts. In LUSC 43% were classified at stage II or III compared to 30% in LUAD.

In LUSC, a total of 9622 reads in primary tumor and 50,630 reads in adjacent normal were mapped and aligned, from which 51% of the tumor reads were identified as viral like reads compared to 15% in adjacent normal. In analyses of variance total number of unique species was significantly lower in LUSC tumor compared to the number of shared species and unique species in adjacent normal (F = 655.7, p = 0.0001) while no difference was observed in LUAD. Based on the significant number of reads mapped/aligned to viral reads in LUSC we wanted to explore the differences in distribution across tumor and adjacent normal. Total viral mapped reads were 34% higher in tumor compared to adjacent normal (Table S4). We found no statistically significant difference in species richness, within sample alpha diversity (measured by Shannon-Wiener Index), or the evenness spread (defined as alpha diversity / log normal (species richness)) in either cohort.

3.5.5. Cancers of the head & neck

We analyzed 69 squamous cell carcinoma of the head and neck region. There were significantly more males than females (70% vs 30%) in our subset of paired HNSC sample population. We found that proportion of bacteria reads accounted for more than 99% of the total reads. Relative abundance in tumor was not significantly different compared to adjacent normal. There were no differences in bacterial diversity means in paired tests comparing tumor to adjacent normal (p = 0.6). Slight differences were observed by anatomical site among HNSC larynx, LOP (lip, oral and pharynx overlap), and tongue (base and non-specified) when compared against floor of mouth within each tissue type when stratifying by sex and sample type. In analyses of variance, anatomical site was a predictor of alpha diversity (p < 0.001). When stratifying by sample type and sex, alpha diversity was slightly higher among females adjacent normal sample with significant differences by anatomical site (p = 0.002) after controlling for sample type, tumor stage, smoke, age, sex and race. Because the established association with HPV etiology, we examined viral read presence. Overall we detected viral reads in 14% of HNSC (20/143) samples, where HHV-4 and HPV were most common. HHV-4 was detected in approximately 13% of HNSC samples. Interestingly, specimens with positive detection of HPV reads were not found to be co-infected with HHV-4.

3.5.6. Cervical cancer

We analyzed 8 cervical squamous cell carcinoma specimens from TCGA CESC cohort. Contrary to HNSC, bacterial reads were less than 1% of the total aligned/mapped reads. In CESC cohort HPV and HHV-4 were found to co-occur in several samples. Based on casual, epidemiological and meta-analysis data (Zhu et al. 2016) linking *Chlamydia trachomatis* co-infection to susceptibility to cervical cancer after HPV infection, we evaluated co-occurrence of the bacterium and HPV. We found no evidence of *Chlamydia trachomatis* reads in our subset of CESC samples. No significant correlation was found between HPV status and diversity or bacterial abundance in CESC. There was no difference in CESC diversity index means between tumor and adjacent normal samples (p = 0.11).

3.5.7. Bladder cancer

We examined 850 files from 412 BLCA cases. From these, most were technical replicates, and a total of 56 paired tumor and adjacent normal samples from 28 cases were examined. The majority of the cases were male (68%), White (89%) and 92% non-Hispanic. Mean age at diagnosis was 69 years (10.7 SD) and 50% were classified at pathological tumor stage IV. Overall, Proteobacteria were the most abundant species, making 93% of the total reads with *Stenotrophomonas maltophilia* the most abundant species (61% of the total reads). However, prevalence within the sample population was low. We found no statistically significant differences between paired tumor and adjacent normal samples total number of reads, relative abundance, or positivity ratio. No taxa were found to be differentially abundant in paired analyses after multiple test correction. However, we note that *Cutibacterium acnes* reads were uniquely identified in tumor samples (p = 0.03 FDR = 1, L2FC = 3.1). Because the large number of non-paired files filtered out, we completed unpaired differential analyses (PathoStat-edgeR function). When considering all available tumor and adjacent normal files independent of 1:1 pairing, we found *Mathylobacterium radiotolerans*, *Pseudomonas aeruginosa* and *Pseudomonas putida* to be differentially abundant in tumor compared to normal after FDR multiple test correction including (p = <0.001, FDR = <0.05, LFC = 1.61, 2.99 and 1.65 respectively). *Stenotrophomonas maltophilia* was not identified as differentially abundant in paired or unpaired analyses. Presence of a combination of these species could be indicative of disease status or sepsis. In fully adjusted model, there are no significant differences in alpha diversity. No statistical significant differences were noted in paired tumor and adjacent normal analyses when stratifying by sex, race, anatomical site or tumor stage (data not shown).

3.6. Validation of bacterial species in gastric and colorectal cancers

Computational findings were validated with tissue from an independent population by species-specific qPCR. We examined gastric and colorectal adenocarcinoma samples on the basis of known or unknown infectious etiology. In TCGA STAD cohort, *Selenomonas sputigena* had the highest proportion of mapped reads

detected in tumor samples compared other species with a prevalence of 18% of tumor compared to 9% of adjacent normal samples. In the other hand, *Helicobacter pylori* was found to be differentially abundant between tumor and adjacent normal pairs with significant higher prevalence in adjacent tissue compared to tumor samples (n = 23 and n = 7, respectively), whereas *Fusobacterium nucleatum* was detected at very low abundance levels uniquely identified in 8% of tumor samples. In TCGA COREAD (colon and rectal cohorts combined) *Bacteroides vulgatus* was found in 70% of the samples with statistically significant differential abundance with higher prevalence in adjacent tissue compared to tumor (n = 80 and n = 68, respectively). *Fusobacterium nucleatum* was found in 24% (n = 25) of the COREAD cases (21 tumor and 10 adjacent normal). *Bacteroides vulgatus* and *Fusobacterium nucleatum* co-occurred in 76% of tumors and 90% of adjacent normal samples. We therefore wanted to validate detection of *Selenomonas sputigena*, *Fusobacterium nucleatum* and *Helicobacter pylori* in tumor and adjacent normal gastric specimens and presence of *Bacteroides vulgatus* and *Fusobacterium nucleatum* in colorectal specimens.

We selected 62 gastric and 162 colorectal samples from the Hawaii RTR based on paired tumor and adjacent normal tissue availability (Table 3A). From the 226 samples, 58 did not yield sufficient quality nucleic acid, 63 colorectal cases (52 colon and 11 rectal) and 21 gastric were tested by species specific qPCR. Population characteristics are summarized in Table 3B.

Overall, positive detection in Hawaii RTR gastric dataset was almost half that of TCGA-STAD dataset (27% vs 47%). Contrary to TCGA STAD, in RTR gastric samples, there was no association between bacterial presence and tissue type. One of the most significant findings from the qPCR examination of gastric tissue was detection of *Selenomonas sputigena* and *Fusobacterium nucleatum* with similar patterns of co-occurrence to TCGA STAD. *Selenomonas sputigena* and *Fusobacterium nucleatum* were detected in 10% (n = 2) of RTR tumor samples in the Hawaii RTR compared to 7% (n = 6) of the gastric tumor samples in TCGA STAD cohort. In RTR dataset, bacteria presence of *Helicobacter pylori*, *Fusobacterium nucleatum* and *Selenomonas sputigena* in tumor was associated with tumor stage and anatomical site, while differences in TCGA STAD were associated with race.

The Hawaii RTR colorectal cancer *Bacteroides vulgatus* was detected in 13% of cases (n = 8, 6 tumor and 5 adjacent normal) from which 88% were non White while *Fusobacterium nucleatum* was found in 21% of cases (n = 13, 11 tumor and 3 adjacent normal). *Bacteroides vulgatus* and *Fusobacterium nucleatum* co-occurred in 31% of positive tumor samples, while adjacent normal positive for *Bacteroides vulgatus* were not co-infected with *Fusobacterium nucleatum*. When looking at sample population composition comparing White vs non-White, we find that *Fusobacterium nucleatum* percent positivity was similar in TCGA cohorts and Hawaii RTR derived samples (17% vs 18% of respectively), while percent positivity for *Bacteroides vulgatus* was strikingly different (64% vs 5% respectively).

4. Discussion

Several studies have evaluated the viral composition in human tumors using unmapped to humans sequencing data; however, bacterial composition derived from human whole exome sequencing data is less explored. In this study we show differences in microbial composition between strict paired tumor and adjacent normal tissue samples across 9 TCGA cancer cohorts. Through our microbial detection methods, we showed significant differential bacterial abundance in stomach and colon adenocarcinomas. The role of *Helicobacter pylori* in stomach adenocarcinoma has been firmly established. We add potential interaction with the microbial community in the adjacent tissue as a sign of disease stage and cancer progression. Noteworthy consistent with the literature, in our study presence of *Helicobacter pylori* within TCGA gastric cohort was higher in the adjacent normal tissue samples (16 times higher) compared to their tumor pairs. Whereas in the tumor

Table 3A
Comparison of qPCR validation in gastric and colorectal cancers.

	RTR			TCGA		
	Tumor	Adjacent normal	Cases	Tumor	Adjacent normal	Cases
	Gastric (N = 21)			Gastric (N = 85)		
Helicobacter pylori	positive	positive	1	positive	positive	4
	positive	negative	1	positive	negative	3
RTR: P-value = 1	negative	positive	1	negative	positive	19
TCGA*: P-value* = 0.001	negative	negative	18	negative	negative	59
Fusobacterium nucleatum	positive	positive	2	positive	positive	0
	positive	negative	1	positive	negative	8
RTR: P-value = 1	negative	positive	0	negative	positive	0
TCGA*: P-value* = 0.01	negative	negative	18	negative	negative	77
Selenomonas sputigena (+)	positive	positive	0	positive	positive	3
	positive	negative	2	positive	negative	12
RTR: P-value = 0.479	negative	positive	0	negative	positive	5
TCGA*: P-value* = 0.146	negative	negative	19	negative	negative	65
	Colorectal (N = 63)			Colorectal (N = 106)		
Bacteroides vulgatus	positive	positive	3	positive	positive	62
	positive	negative	3	positive	negative	6
RTR: P-value = 1	negative	positive	2	negative	positive	18
TCGA*: P-value* = 0.025	negative	negative	55	negative	negative	20
Fusobacterium nucleatum	positive	positive	1	positive	positive	5
	positive	negative	10	positive	negative	15
RTR: P-value = 0.043	negative	positive	2	negative	positive	4
TCGA*: P-value* = 0.021	negative	negative	50	negative	negative	82

Bacteria presence counts for each taxa examined in tumor and adjacent normal gastric and colorectal cances in TCGA cohorts versus Hawaii RTR. *Selenomonas sputigena* and *Fusobacterium nucleatum* were found to co-occur in Hawaii RTR gastric cases in similar patterns to those observed in TCGA gastric cancer cohort. Percent positivity in White vs non-White for Fusobacterium nucleatum was similar in TCGA and Hawaii RTR (17% vs 18% of respectively), whereas percent positivity for Bacteroides vulgatus was strikingly different (64% vs 5% respectively). Possible explanation for differences observed could be due to other population characteristics (Table 3B), FFPE sample degradation and sample size. N = paired cases; P values were from McNemars test. Null hypothesis: there is no difference between tumor and adjacent normal due to specific microbial presence.

Table 3B
Population characteristics comparison between RTR and TCGA in gastric and colorectal cancers.

	RTR	TCGA
	Gastric (N = 21)	Gastric (N = 85)
Race other than White	95%	36%
Age > 60 years	76%	75%
Sex: Female	62%	44%
Diagnosis anatomical site	43% unspecified	31% antrum
Tumor classification	62% stage III	56% stage III
	Colorectal (N = 63)	Colorectal (N = 106)
Race other than White	73%	58%
Age > 60 years	38%	75%
Sex: Female	49%	49%
Diagnosis anatomical site	38% sigmoid/rectosigmoid region	34% unspecified colon/rectum
Tumor classification	31% stage II	40% stage II

Compared to TCGA gastric cancer subset, Hawaii RTR population characteristics were significantly different by race, sex and tumor site at initial diagnosis. While in colorectal cancer subset, differences existed in race, age at time of diagnosis, and tumor site at initial diagnosis. We believe differences in positivity ratios could be due population differences were in TCGA population is mostly White Eastern European compared to Hawaii RTR which is mostly Hawaiian and Asian ethnic subgroups.

samples we found significantly higher levels or exclusive presence of oral taxa including *Fusobacterium nucleatum*, *Veillionella parvula* and *Selenomonas sputigena*. Interestingly, *Selenomonas sputigena* has been detected in the tongue coatings of gastric cancer patients and identified as a potential biomarker [41]. This is the first time we note identification within the tumor tissue sequences of gastric patients. The clinical significance of these oral species and their interaction with the tumor microenvironment should be further explored. *Bacteroides vulgatus* is known to be one of the most numerically dominant species of the colonic microflora and thought to have beneficial pro-inflammatory immune response suppression effects [42]. Our findings of differential composition in colon cancer with significant overexpression in adjacent normal tissue could have important diagnostic and therapeutic implications. However, percent positivity within the Hawaii RTR population compared to colorectal TCGA cohorts was relatively low at 13% and we are hesitant to make conclusions at this point. Differences could be in part due to FFPE sample degradation or population composition differences. In liver hepatocellular carcinoma, we found no difference in bacteria composition within paired samples when comparing tumor to its adjacent normal tissue; however, there was a visible difference within tissue type in this cohort. In liver hepatocellular carcinoma, although the number of species per sample appeared to be similar for both tumor and its adjacent normal, diversity varied by stage, age at diagnosis and sex which could have potential clinical significance particularly when we seek to uncover targetable biomarkers to improve patient outcomes. Not surprisingly, consistent with previous reports, we identified HPV reads in all CESC paired samples examined [18,21]. While in cancers of the head and neck, HPV was detected in 5% of the samples. Previously reported the detection of HPV in head and neck cancers ranges from 20% to 21% [17,21,37]. Compared to previous head and neck studies reporting on whole exome sequencing data, we feel confident that our results are similar when considering the examined paired samples. Our study included 69 paired head and neck cases. From these seven samples (6 tumor, 1 adjacent solid tissue normal) corresponding to 6 cases (1 female, 5 male) were positive for HPV. We observed mild interaction between viral and bacterial presence suggestive of polymicrobial effects on disease stage. In CESC we found a (weak) negative correlation between HPV and *Bradyrhizobium* sp. which varied by pathological stage. We note that our sample size was small; perhaps correlation may be more prominent with increased sample size. Interestingly, Riley et al. reported that *Bradyrhizobium* like species including *Bradyrhizobium* BTAi1 were the most-common strain level operational taxonomic units found within the 1000 Genomes Project supporting lateral gene transfer (Riley, 2013). *Bradyrhizobium* sp. however, have been found to be common contaminants within the 1000 Genome Project and many high-throughput efforts [38]. *Bradyrhizobium diazoefficiens* and *Bradyrhizobium* BTAi1 are nitrogen fixating bacteria and their role in disease is currently unknown. Nevertheless, these species may have a potential role in cancer pathogenesis and cancer therapy by means of their Hsp70 family molecular chaperone protein interaction with p53 [39,40]. Contrary to those, in our study, microbial profiles were derived from a diverse population. Data has been collected at various Institutes and was sequenced at different Centers. Although water system or laboratory contaminants could be a source of *Bradyrhizobium* reads, strict use of paired samples and fold change analyses should assist with misidentification. In our data, presence varied among cohorts with highest total reads among CESC and COAD. We do point out that in our data, often both case pairs had bacterial reads for *Bradyrhizobium* like species. Presence in both tumor and its adjacent normal could be indicative of core taxa microbiota within the tumor microenvironment, contamination or laboratory artifacts. Riley, highlights that little is known about the composition of the human tumor microbiome and that although contamination can be suspected, the presence of the microbe may be due to diet and lifestyle differences in the population. Presence and clinical significance of these species should be further examined as identification of core microbiota is important to the understanding of the tumor microenvironment and the role bacteria play in cancer pathogenesis [36]

In our study we found that there are significant differences in diversity and composition of the tumor compared to adjacent normal across different cancer types with observable patterns when stratifying by age, sex, race, and tumor stage within each cohort. There were observable differences in clinical presentation among cases from different cohorts. Differences in clinical presentation among cancer patients may be explained by microbial abundance and diversity patterns and similarly can be the focus of future studies. Taxonomic composition was found to be similar to that previously reported in RNA-seq, whole genome or whole exome sequencing data [17,21,23,24]. We note that measures of relative abundance alone or total number of reads do not provide sufficient information regarding the compositional differences in the tumor microenvironment. A high read count with low relative abundances and vice versa, suggests that read counts could mask population prevalence. Measures of total reads, relative abundances, and percent prevalence in the population need to be taken into account for a more accurate description of the differences within and across cohorts. We point out that all three measures must be used for accurate characterization of the tumor microbiota with greater weight on percent population prevalence and relative

abundance when identifying clinically relevant taxonomy to avoid erroneous conclusions.

Our study is not without its limitations, the low reads relative to human sequences may not be sensitive to the magnitude of differential expression and it may be less powered because our paired analyses filtration resulted in a low number of cases analysed. We set limits to protect against this by not including any cancer cohorts with less than 15 specimens (smallest sample size CESC with 16 specimens). Our integrated analysis of exclusive one-to-one paired samples is not sensitive to tissue-specific baseline relative abundance or inherited 16S compositional assumptions. Our study is strengthen by each patient serving as their own control eliminating interacting and confounding factors. In this study we demonstrate the ability to identify the differential composition of bacterial species derived from human tissue whole exome sequencing data. This study highlights the importance of analysing adjacent tissue which can be indicative of cancer stage progression.

5. Conclusions

We conclude that identifying microbial composition in tumor and adjacent normal tissue, using whole exome sequencing data provides useful and comparative tool similar to transcriptome and metagenomic methods to study bacterial composition in cancer. Differences in bacterial composition and microbial interaction within the tumor microenvironment could be indicative of disease progression. Further qPCR validation of bacterial presence with tissue specimens from Hawaii Tumor Registry as an independent population strengthens our findings. We highlight co-occurrence of *Selenomonas sputigena* and *Fusobacterium nucleatum* in tumor tissue of stomach adenocarcinoma. *Selenomonas sputigena* has been identified in the tongue coating of gastric patients, but has not yet been identified in the tumor tissue. Similarly, *Bacteroides vulgatus* is believed to have protective anti-tumorigenic effects and one of the most commonly identified species from stool. Here we have identified both species, *Bacteroides vulgatus* predominantly in adjacent normal tissue and *Selenomonas sputigena* in tumor tissue which could have potential diagnostic and therapeutic implications. Additional studies are needed to better understand their roles in the tumor microenvironment. Future studies seeking to characterize the microbiota within the tumor microenvironment should consider examination of the adjacent tissue weighing prevalence within the population with equal weight to the total amount of reads detected. This will facilitate microbial functional predictions and distinguish between true presence and laboratory artifacts or possible contamination.

Acknowledgements

We thank the UH Cancer Center Pathology Lab for their support with FFPE tissue block nucleic acid extraction and PCR validation. The results published here are in part based upon data generated by the Cancer Genome Atlas (TCGA) managed by the National Cancer Institute (NCI) and National Human Genome Research Institute (NHGRI). Information about TCGA can be found at http://cancergenome.nih.gov.

Funding

This work was supported by Ola HAWAII, National Institute on Minority Health and Health Disparities (NIMHD) a component of the National Institute of Health (NIH) grant number 2U54MD007601-32 to VSK. The Bioinformatics Core is supported in part by NIH grant numbers P20GM103466, U54MD007584, and 5P30GM114737. The contents of this work are solely the responsibility of the authors and do not necessarily represent the views of NIMID or NIH.

Authors' contributions

RMR, VSK and YD designed the study; RMR and MM contributed to pipeline design. MM carried out computational microbial profile extraction. BYH managed FFPE sample extraction and qPCR validation. RMR and VSK analyzed the data and made figures. RMR wrote the original draft. All authors contributed to study conceptualization, reviewed, and editing of this manuscript. All authors approved the final version.

Ethics approval and consent to participate

All human data were handled in accordance with TCGA Data Use Certification Agreement and Data Access Request (DAR) 57292; Project-14778 (Y. Deng, PI), Request ID 57292-2 and 57292-4 (renewal 10/10/2018) for access phs000178 versions v9.p8 and v10.p8. Institutional Review Board approval was obtained from the University of Hawaii IRB. Paired tumor and adjacent normal FFPE samples from the Hawaii Tumor Registry-Discard Residual Repository (Hawaii RTR) were requested at a similar 1:1 ratio per case.

Competing interests

The authors declare that they have no competing interests.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.csbj.2020.03.003.

References

[1] Nauts HC. Bacteria and cancerantagonisms and benefits. Cancer Surv 1989;8:713–23.
[2] Schwabe RF, Jobin C. The microbiome and cancer. Nat Rev Cancer 2013;13:800–12. https://doi.org/10.1038/nrc3610.
[3] Mager D. Bacteria and cancer: cause, coincidence or cure? A review. J Transl Med 2006;4:14. https://doi.org/10.1186/1479-5876-4-14.
[4] Chang AH, Parsonnet J. Role of bacteria in oncogenesis. Clin Microbiol Rev 2010;23:837–57. https://doi.org/10.1128/CMR.00012-10.
[5] Paulos CM, Wrzesinski C, Kaiser A, Hinrichs CS, Chieppa M, Cassard L, et al. Microbial translocation augments the function of adoptively transferred self/tumor-specific CD8+ T cells via TLR4 signaling. J Clin Invest 2007;117:2197–204. https://doi.org/10.1172/JCI32205.
[6] Elinav E, Nowarski R, Thaiss CA, Hu B, Jin C, Flavell RA. Inflammation-induced cancer: crosstalk between tumours, immune cells and microorganisms. Nat Rev Cancer 2013;13:759–71. https://doi.org/10.1038/nrc3611.
[7] Hattori N, Ushijima T. Epigenetic impact of infection on carcinogenesis: mechanisms and applications. Genome Med 2016;8. https://doi.org/10.1186/s13073-016-0267-2.
[8] Cancer Genome Atlas Research Network. Comprehensive molecular characterization of gastric adenocarcinoma. Nature 2014;513:202–9. https://doi.org/10.1038/nature13480.
[9] The Cancer Genome Atlas Network. Comprehensive genomic characterization of head and neck squamous cell carcinomas. Nature 2015;517:576–82. https://doi.org/10.1038/nature14129.
[10] Thomas AM, Jesus EC, Lopes A, Aguiar S, Begnami MD, Rocha RM, et al. Tissue-associated bacterial alterations in rectal carcinoma patients revealed by 16S rRNA community profiling. Front Cell Infect Microbiol 2016;6. https://doi.org/10.3389/fcimb.2016.00179.
[11] Tae H, Karunasena E, Bavarva JH, McIver LJ, Garner HR. Large scale comparison of non-human sequences in human sequencing data. Genomics 2014;104:453–8. https://doi.org/10.1016/j.ygeno.2014.08.009.
[12] Kostic AD, Ojesina AI, Pedamallu CS, Jung J, Verhaak RGW, Getz G, et al. PathSeq: software to identify or discover microbes by deep sequencing of human tissue. Nat Biotechnol 2011;29:393–6. https://doi.org/10.1038/nbt.1868.
[13] Borozan I, Wilson S, Blanchette P, Laflamme P, Watt SN, Krzyzanowski PM, et al. CaPSID: a bioinformatics platform for computational pathogen sequence

identification in human genomes and transcriptomes. BMC Bioinf 2012;13 ():206. https://doi.org/10.1186/1471-2105-13-206.
[14] Naccache SN, Federman S, Veeraraghavan N, Zaharia M, Lee D, Samayoa E, et al. A cloud-compatible bioinformatics pipeline for ultrarapid pathogen identification from next-generation sequencing of clinical samples. Genome Res 2014;24:1180–92. https://doi.org/10.1101/gr.171934.113.
[15] Hong C, Manimaran S, Shen Y, Perez-Rogers JF, Byrd AL, Castro-Nallar E, et al. PathoScope 2.0: a complete computational framework for strain identification in environmental or clinical sequencing samples. Microbiome 2014;2:33. https://doi.org/10.1186/2049-2618-2-33.
[16] Chen Y, Yao H, Thompson EJ, Tannir NM, Weinstein JN, Su X. VirusSeq: software to identify viruses and their integration sites using next-generation sequencing of human cancer tissue. Bioinformatics 2013;29:266–7. https://doi.org/10.1093/bioinformatics/bts665.
[17] Khoury JD, Tannir NM, Williams MD, Chen Y, Yao H, Zhang J, et al. Landscape of DNA virus associations across human malignant cancers: analysis of 3,775 cases using RNA-Seq. J Virol 2013;87:8916–26. https://doi.org/10.1128/JVI.00340-13.
[18] Tang K-W, Alaei-Mahabadi B, Samuelsson T, Lindh M, Larsson E. The landscape of viral expression and host gene fusion and adaptation in human cancer. Nat Commun 2013;4:2513. https://doi.org/10.1038/ncomms3513.
[19] Salyakina D, Tsinoremas NF. Viral expression associated with gastrointestinal adenocarcinomas in TCGA high-throughput sequencing data. Hum Genomics 2013;7:23. https://doi.org/10.1186/1479-7364-7-23.
[20] Cao S, Wendl MC, Wyczalkowski MA, Wylie K, Ye K, Jayasinghe R, et al. Divergent viral presentation among human tumors and adjacent normal tissues. Sci Rep 2016;6:28294. https://doi.org/10.1038/srep28294.
[21] Cantalupo PG, Katz JP, Pipas JM. Viral sequences in human cancer. Virology 2018;513:208–16. https://doi.org/10.1016/j.virol.2017.10.017.
[22] Riley DR, Sieber KB, Robinson KM, White JR, Ganesan A, Nourbakhsh S, et al. Bacteria-human somatic cell lateral gene transfer is enriched in cancer samples. PLoS Comput Biol 2013;9:. https://doi.org/10.1371/journal.pcbi.1003107e1003107.
[23] Robinson KM, Crabtree J, Mattick JSA, Anderson KE, Dunning Hotopp JC. Distinguishing potential bacteria-tumor associations from contamination in a secondary data analysis of public cancer genome sequence data. Microbiome 2017;5:9. https://doi.org/10.1186/s40168-016-0224-8.
[24] Zhang C, Cleveland K, Schnoll-Sussman F, McClure B, Bigg M, Thakkar P, et al. Identification of low abundance microbiome in clinical samples using whole genome sequencing. Genome Biol 2015;16. https://doi.org/10.1186/s13059-015-0821-z.
[25] Huo Q, Zhang N, Yang Q. Epstein-Barr virus infection and sporadic breast cancer risk: a meta-analysis. PLoS ONE 2012;7:. https://doi.org/10.1371/journal.pone.0031656e31656.
[26] Thompson KJ, Ingle JN, Tang X, Chia N, Jeraldo PR, Walther-Antonio MR, et al. A comprehensive analysis of breast cancer microbiota and host gene expression. PLoS ONE 2017;12:. https://doi.org/10.1371/journal.pone.0188873e0188873.
[27] Greathouse KL, White JR, Vargas AJ, Bliskovsky VV, Beck JA, von Muhlinen N, et al. Interaction between the microbiome and TP53 in human lung cancer. Genome Biol 2018;19:123. https://doi.org/10.1186/s13059-018-1501-6.
[28] Wang Z, Gerstein M, Snyder M. RNA-Seq: a revolutionary tool for transcriptomics. Nat Rev Genet 2009;10:57–63. https://doi.org/10.1038/nrg2484.
[29] Choi M, Scholl UI, Ji W, Liu T, Tikhonova IR, Zumbo P, et al. Genetic diagnosis by whole exome capture and massively parallel DNA sequencing. Proc Natl Acad Sci USA 2009;106:19096–101. https://doi.org/10.1073/pnas.0910672106.
[30] Gopalakrishnan V, Spencer CN, Nezi L, Reuben A, Andrews MC, Karpinets TV, et al. Gut microbiome modulates response to anti-PD-1 immunotherapy in melanoma patients. Science 2018;359:97–103. https://doi.org/10.1126/science.aan4236.
[31] Castellarin M, Warren RL, Freeman JD, Dreolini L, Krzywinski M, Strauss J, et al. Fusobacterium nucleatum infection is prevalent in human colorectal carcinoma. Genome Res 2012;22:299–306. https://doi.org/10.1101/gr.126516.111.
[32] Kostic AD, Gevers D, Pedamallu CS, Michaud M, Duke F, Earl AM, et al. Genomic analysis identifies association of Fusobacterium with colorectal carcinoma. Genome Res 2012;22:292–8. https://doi.org/10.1101/gr.126573.111.
[33] 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;14:207–15. https://doi.org/10.1016/j.chom.2013.07.007.
[34] Warren RL, Freeman DJ, Pleasance S, Watson P, Moore RA, Cochrane K, et al. Co-occurrence of anaerobic bacteria in colorectal carcinomas. Microbiome 2013;1:16. https://doi.org/10.1186/2049-2618-1-16.
[35] Kumar A, Thotakura PL, Tiwary BK, Krishna R. Target identification in Fusobacterium nucleatum by subtractive genomics approach and enrichment analysis of host-pathogen protein-protein interactions. BMC Microbiol 2016;16:84. https://doi.org/10.1186/s12866-016-0700-0.
[36] Wang H, Funchain P, Bebek G, Altemus J, Zhang H, Niazi F, et al. Microbiomic differences in tumor and paired-normal tissue in head and neck squamous cell carcinomas. Genome Med 2017;9. https://doi.org/10.1186/s13073-017-0405-5.
[37] Hernandez BY, Goodman MT, Unger ER, Steinau M, Powers A, Lynch CF, et al. Human papillomavirus genotype prevalence in invasive penile cancers from a registry-based United States population. Front Oncol 2014;4:9. https://doi.org/10.3389/fonc.2014.00009.
[38] Laurence M, Hatzis C, Brash DE. Common contaminants in next-generation sequencing that hinder discovery of low-abundance microbes. PLoS ONE 2014;9:. https://doi.org/10.1371/journal.pone.0097876e97876.
[39] Deocaris CC, Widodo N, Ishii T, Kaul SC, Wadhwa R. Functional significance of minor structural and expression changes in stress chaperone mortalin. Ann N Y Acad Sci 2007;1119:165–75. https://doi.org/10.1196/annals.1404.007.
[40] Shevtsov M, Huile G, Multhoff G. Membrane heat shock protein 70: a theranostic target for cancer therapy. Philos Trans R Soc Lond B Biol Sci 2018;373. https://doi.org/10.1098/rstb.2016.0526.
[41] Xu J, Xiang C, Zhang C, Xu B, Wu J, Wang R, et al. Microbial biomarkers of common tongue coatings in patients with gastric cancer. Microb Pathog 2019;127:97–105. https://doi.org/10.1016/j.micpath.2018.11.051.
[42] Yoshida N, Emoto T, Yamashita T, Watanabe H, Hayashi T, Tabata T, et al. Bacteroides vulgatus and Bacteroides dorei reduce gut microbial lipopolysaccharide production and inhibit atherosclerosis. Circulation 2018;138(22):2486–98. https://doi.org/10.1161/CIRCULATIONAHA.118.033714.