Computational and Structural Biotechnology Journal 20 (2022) 3068–3079

COMPUTATIONAL AND STRUCTURAL BIOTECHNOLOGY JOURNAL

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

Exploring the patient-microbiome interaction patterns for pan-cancer

Lan Zhao [a,*], William C.S. Cho [b], Jun-Li Luo [c,*]

[a] Department of Medicine, Stanford University School of Medicine, Stanford, CA 94305, United States
[b] Department of Clinical Oncology, Queen Elizabeth Hospital, Hong Kong, China
[c] The Cancer Research Institute, Hengyang Medical School, University of South China, Hengyang, 421001, Hunan, China

ARTICLE INFO

Article history:
Received 4 March 2022
Received in revised form 6 June 2022
Accepted 6 June 2022
Available online 8 June 2022

Keywords:
Cancer microbiome
Heterogeneity
Biclustering
Patient-microbe interaction
Microbial signature

ABSTRACT

Microbes play important roles in human health and disease. Immunocompromised cancer patients are more vulnerable to getting microbial infections. Regions of hypoxia and acidic tumor microenvironment shape the microbial community diversity and abundance. Each cancer has its own microbiome, making cancer-specific sets of microbiomes. High-throughput profiling technologies provide a culture-free approach for microbial profiling in tumor samples. Microbial compositional data was extracted and examined from the TCGA unmapped transcriptome data. Biclustering, correlation, and statistical analyses were performed to determine the seven patient-microbe interaction patterns. These two-dimensional patterns consist of a group of microbial species that show significant over-representation over the 7 pan-cancer subtypes (S1-S7), respectively. Approximately 60% of the untreated cancer patients have experienced tissue microbial composition and functional changes between subtypes and normal controls. Among these changes, subtype S5 had loss of microbial diversity as well as impaired immune functions. S1, S2, and S3 had been enriched with microbial signatures derived from the *Gammaproteobacteria, Actinobacteria* and *Betaproteobacteria,* respectively. Colorectal cancer (CRC) was largely composed of two subtypes, namely S4 and S6, driven by different microbial profiles. S4 patients had increased microbial load, and were enriched with CRC-related oncogenic pathways. S6 CRC together with other cancer patients, making up almost 40% of all cases were classified into the S6 subtype, which not only resembled the normal control's microbiota but also retained their original "normal-like" functions. Lastly, the S7 was a rare and understudied subtype. Our study investigated the pan-cancer heterogeneity at the microbial level. The identified seven pan-cancer subtypes with 424 subtype-specific microbial signatures will help us find new therapeutic targets and better treatment strategies for cancer patients.

1. Introduction

Human microbes are tiny microorganisms that live in and on us, which play important roles in both health and diseases. Humans rely on microorganisms to perform normal functions such as absorbing nutrients [1], strengthening immune system [2], and reducing the likelihood of serious microbe infections. Other microbes can be pathogenic and cause diseases in humans [3]. Humans get initial microbes from their mom at birth, and the microbial communities gradually grow to a mature, diverse, and stable status as people age and their interactions with the outside world. Trillions of microbes that inhabit the human guts, mouth, skin and other tissues as defined by the Human Microbiome Project (HMP) [4]. It was originally assumed that the internal organs such as the brain, heart, lungs, among others are microbe-free and considered sterile. Recent progress have led us to believe that microbial species including pathogens and commensals exist in human internal organs. For instance, blood-borne pathogens such as *Neisseria meningitidis* and *Streptococcus pneumoniae* can cross the blood-central nervous system (CNS) barriers and cause bacterial meningitis in newborns and adults [5]. Infective endocarditis is caused by *Staphylococcus aureus* infection of the inner layer of the heart [6]. Tuberculosis (TB), a highly communicable chronic lung disease, is caused by a bacterium called *Mycobacterium tuberculosis* (MTB). Furthermore, organ-specific commensals have been identified in studies of healthy controls in the absence of disease, which include the brain [7], lungs [8], pancreas [9], and more [10,11].

Microbes prefer to live in an environment which is suitable for their survival, in other words, environments help shape the diver-

* Corresponding authors.
E-mail addresses: lanzhao20140101@gmail.com (L. Zhao), jlluo@usc.edu.cn (J.-L. Luo).

https://doi.org/10.1016/j.csbj.2022.06.012

sity and abundance of the resident microbial communities. For example, obligate anaerobes bacteria including *Clostridiales* and many *Bacteroidetes* spp., are commonly found in the large intestine (colon) of humans where they can thrive without oxygen. Unlike most of the obligate anaerobes, facultative anaerobes such as *Lactobacilli* and *Streptococci* can grow with or without oxygen and are predominant in the small intestine [12]. Although lungs and many other non-gastrointestinal (GI) organs create inhospitable environments for bacteria with little nutrition, *Proteobacteria* are often overrepresented in several inflammatory-related extraintestinal diseases [13]. Microbes from the outside environments are constantly interacting with humans, and human microbiomes change in accordance with the times, places and health status, but remain stable over a period of time once established [4,14]. Environmental microbes from the soil, water, air, food, and many other sources can become part of the human microbiome. For instance, the soil/-root microbiota contribute to the human gut microbiome [15]. Apart from the skin, lungs are in close contact with the surrounding air which contains diverse microbial species [16]. Healthcare-associated pathogens such as *Klebsiella*, *Staphylococcus aureus*, and *Clostridium difficile (C. diff)* remodel the human microbiota across multiple body habitats [17].

Cancer is a complex and heterogeneous disease [18], which affects nearly all organ systems in the body. Both genetic and non-genetic factors contribute to cancer initiation and progression. For example, TP53 gene mutations increase the risk of developing a number of cancer types [19]. BRCA1 and BRCA2 mutations increase the breast cancer risk [20], despite of the fact that these mutations do not account for a significant proportion of the cases. Non-genetic factors such as exposure to carcinogens and lifestyle choices (i.e. smoking) strongly associated with cancer development and progression [21,22]. The idea that microbes can cause and promote the progression of cancer is not new and well established. More and more carcinogenic microbes have been identified, particularly the *Human papillomavirus* (HPV) has been implicated in up to 99% of all cervical cancers [23]. *H. pylori* accounts for more than 90% of gastric cancer cases [24]. *Hepatitis B virus* (HBV) is responsible for 56% of liver cancer [25]. Besides cancer-causing species, accumulated evidences have shown that dysbiosis (the disruption of a balanced microbiome) is associated with risk of multiple cancer types [26–30].

Recent advances in high-throughput profiling technologies have made it possible to generate vast amounts of sequencing data, which provides a culture-free approach for microbial profiling in tumor samples. The Cancer Genome Atlas (TCGA) has sequenced over 11,000 primary cancer cases from 33 most prevalent cancer types. Revealing the tumor heterogeneity at the microbiome level provides novel insights and targets for personalized treatment of cancer patients [31]. Although the Pan-Cancer Atlas has published over 27 papers on topics ranging from Pan-cancer subtyping (cell-of-origin patterns) to functional characterization (oncogenic processes, and signaling pathways), microbiome-based investigation of tumor heterogeneity hasn't been done before. In order to fill the gap, we re-analysed the TCGA whole-transcriptome sequencing data to determine the pan-cancer microbial subtypes and examine their interactions with the hosts.

2. Methods

2.1. TCGA transcriptome-based microbial data curation and preprocessing

Genomic Data Commons (GDC: https://gdc.cancer.gov/) generated BAM alignment files (harmonized TCGA GRCh38) from RNA-Seq were accessed on the Seven Bridges Platform by the Cancer Genomics Cloud (CGC: https://www.cancergenomicscloud.org/). Only non-Formalin-Fixed Paraffin-Embedded (FFPE) solid primary tumors (n = 9,232) and matched normal control samples (n = 720) were included in the study.

Non-human sequences were extracted from the BAM alignments using SAMtools (hosted on the CGC), and used as the inputs for microbial identification. Kraken2 [32], a taxonomic classification tool which relies on exact k-mer matches, was employed for estimating microbial species abundance in each extracted file. Kraken2 Database was built on NCBI RefSeq genomes (https://ftp.ncbi.nlm.nih.gov/genomes/refseq/; n = 52,127) of bacterial, archaeal, viral, and plasmid curated on 18 September 2019 (Zhao et al.) [33]. Five r4.8xlarge AWS EC2 instances were made to run in parallel for Kraken2 short-read taxonomy assignment.

Species-level microbial operational taxonomic units (OTUs) from Kraken2 reports were combined across the samples in a matrix, and all the following analyses were performed using R (4.0.2 version) and Python (3.7 version). The phyloseq R package [34] was used for diversity calculations and visualization with ggplot2 [35]. Rare microbial species that were not present in at least one read count in 0.1% of prevalence of the total samples were eliminated. In addition, to further minimize microbial contamination of tissue samples, over 60 known contaminant genera have been identified across multiple studies [36] were filtered out. The resulting OTU table was then normalized to median sequencing depth in log2 scale.

2.2. Identification of pan-cancer microbial biclusters

To select informative microbial species for pan-cancer classification, we assessed each species's ability to separate one cancer type from the others by calculating area under the curve (AUC). The retained 2,863 species with AUC greater than 0.6 were used as the features for unsupervised biclustering framework. cNMF [37], a biclustering-based Python package with Kullback-Leibler divergence (KL-divergence) was applied to identify the patient-microbe interaction patterns (biclusters). We varied the number of clusters K from 2 to 10, and set the number of iterations to 10. The value of K that resulted in the largest value in stability was chosen as the optimal number of biclusters. The threshold of 0.3 on average distance to KNN was used to filter out outliers, then a final consensus solution was reached among the replicates. 'Max' method defined by Carmona-Saez *et al* [38] was used for selecting the *meta*-microbe with the largest row feature scores as signatures for each bicluster.

The silhouette width [39] was used for sample cluster evaluations, and selecting the most coherent samples within each cluster. More specifically, the pairwise distances between samples using the Jaccard dissimilarity were calculated with the vegan R package [40]. The distances were then used for computing the silhouette width for each sample in the cluster (v2.1.0) R package. Samples with positive silhouette width were retained for further analysis.

2.3. Statistical analyses of microbial compositional data

Alpha diversity is to evaluate variance within a particular sample and beta diversity is to assess how different sample communities vary against each other. The Shannon index accounts for both abundance (richness) and evenness of the species present, which is used to characterize microbial alpha diversity in our study. Pairwise Wilcoxon rank-sum test was used for comparisons of microbial abundance between groups. Benjamini-Hochberg (BH) correction was applied to adjust for multiple comparisons. An adjusted p-value of 0.05 or lower was considered statistically significant.

Jaccard index was calculated to estimate the degree of dissimilarity between a pair of microbial communities, and used as a mea-

sure of beta diversity in this study. Permutational multivariate analysis of variance (PERMANOVA) [41] was performed on the distance matrix to compare distances, and tested using the adonis function with 999 permutations in the vegan R package [42]. Principal coordinates analysis (PCoA) was carried out to obtain principal coordinates, and visualized in a three-dimensional (3D) plot with the rgl package [43].

2.4. Microbial community correlation analyses

Pre-calculated TCGA enrichment scores (ESs) of 64 tumor-infiltrating cell components were downloaded from the xCell website (https://xcell.ucsf.edu/) [44]. For the human-mapped reads, gene quantification was performed by using HTSeq [45] converted to gene-level log2-transformed transcripts per million (TPM).

Specific microbial communities consist of a group of distinct microbial species enriched in each pan-cancer subtype. The average abundances for each microbial community across all patients were calculated to represent their enriched levels in each community. Spearman correlation analyses were applied to correlate the microbial community abundances to (human) genes and immune cell type characteristics, respectively. FDR corrected p-values<0.05 were considered significant. Overlapped significantly correlated genes across all communities were selected for heatmap visualization using the pheatmap package in R [46]. Pairwise correlation between microbial communities was plotted using the corrplot R package [47].

2.5. Functional enrichment and gene set enrichment analysis (GSEA)

Genes that were significantly correlated with the 7 microbial communities were functionally annotated using gene sets from KEGG and Gene Ontology (GO) via the online software Enrichr [48].

Gene expression fold changes between one subtype versus the remaining subtypes were calculated using the limma package [49], and employed for GSEA analysis. In our study, GSEA was conducted with the R package piano [50]. Annotated gene sets were downloaded from the MsigDB database (version 7.2 C2 and C5). We selected gene sets with the number of genes ranging from 10 to 500, and 1,000 permutations for significance tests. The other parameters were set as default. Top significantly enriched gene sets (FDR adjusted p-value < 0.05) in each subtype were ranked according to their enrichment scores, and selected for the p-value heatmap visualization. Selected gene sets were grouped in functional categories by hierarchical clustering analysis, and the heatmap was visualized by using the pheatmap package in R [51].

2.6. Survival and COX regression analysis

Kaplan-Meier survival analysis with log-rank p-values of<0.05 were considered statistically significant. Two clinical endpoints, namely overall survival (OS) and progression free survival (PFS) were evaluated by the Cox proportional hazard (PH) regression analysis, and were adjusted for confounders such as patient age, gender, tumor stage and grade. Forest plots were generated to estimate hazard ratios (HRs) and log-rank p-values for OS and PFS among pan-cancer subtypes and other factors by using the ggforest function in survminer R package [52].

3. Results

3.1. Identification of seven patient-microbe interaction patterns

Microbial reads were extracted from the TCGA RNA-Seq BAM alignment files and mapped against a standard Kraken2 database. 14,082 species-level microbial OTU assignments of each sample's Kraken2 outputs were combined into a single matrix. Patients' (n = 9,232) and tumor-adjacent normal control samples' (n = 720) metadata were queried via cBioPortal (https://www.cbioportal.org/; **Table S1**). 30 cancer types were analyzed, and the 720 normal controls matched to 20 different cancer types. The number of control samples per cancer range from 1 to 113 (average: 34). The remaining 10 cancer types lack control samples, this include ACC, CESC, LGG, GBM, DLBC, MESO, TGCT, UCS, OV, and SKCM. To make the tumor/normal comparison possible for the cancer types involved in our study, we combined all 720 adjacent normal samples together as a reference group.

We did filtering processes to remove species with low frequency (<0.1% prevalence) as well as potentially contaminated genera [35], which resulted in 7,075 species-level microbial OTUs in 9,232 patients (**Table S2**) available for the downstream analysis. 0 (100% prevalence), 16 (90% prevalence), and 46 (80% prevalence) substantial core microbial species [53] that shared by all or most patients were listed in the Table S3, with the majority of them annotated as free-living bacteria with unknown effect on human health and disease so far. 2,863 species with AUC values greater than 0.6 were subsequently selected as the features for unsupervised consensus non-negative matrix factorization (cNMF) analysis [37]. cNMF identified seven biclusters (patient-microbe interaction patterns) in the pan-cancer microbial data, as the value of K = 7 which resulted in the largest value in stability (Fig. 1**a**). Outliers which were above the 0.3 on average distance of K nearest neighbors (KNN) were filtered out (Fig. 1**b**), and a consensus NMF result was reached among the 10 replicates. The consensus matrix heatmap declared the existence of the 7 well-separated clusters of samples (Fig. 1**c**). Afterwards, a total of 424 microbial species were identified as the *meta*-microbe (signatures) for the 7 biclusters (**Tables S4**). Silhouette width analysis was subsequently performed to select the most stable samples within each cluster (Fig. 1**d**). The average silhouette width was 0.05, and samples with positive silhouette width (n = 6,612) were retained. The 6,612 patient's microbial abundance heatmap (with 424 microbial signatures) showed clear separation of the seven patient-microbe interaction patterns (P1-P7) (Fig. 1**e**).

3.2. Pan-cancer subtype and microbial signature distributions

The 6,612 patient's classification results were used to define the 7 pan-cancer subtypes (S1-S7, **Tables S5**). S6 was the largest subtype with 41.1% (n = 2,719) of all cancer cases. Followed by S2 (n = 1,402, 21.2%) and S5 (n = 1,237, 18.7%), making the second and third largest subtypes, respectively. S7 was a rare subtype, representing<1% (n = 63) of all cases (Table 1). Patient percentages from the S1, S3, and S4 subtypes each consisted of no more than 10% of all cases, and contained relatively fewer cancer types (Table 1**; Figure S1**). For instance, S4 had 533 patients accounting for 8.0% of all cases, with only 2 types of cancer in this subtype (Table 1**; Figure S1**). Unlike S4, the cancer types in S5 and S6 were broad and varied, spanning over 25 different cancer types, respectively (Table 1**; Figure S1**).

The number of microbial signatures in each pan-cancer subtype ranged from 34 (S6) to 108 (S3) with all major microbial phyla including *Actinobacteria, Proteobacteria, Firmicutes*, and *Bacteroidetes* (Table 1; **Figure S2**). Among the 424 microbial signatures, 10 were derived from the Archaea, 5 from the Viruses, and the remaining 409 species within the Bacteria Kingdom, respectively (**Tables S4**). 5 of the 10 Archaea were methanogenic archaea, which are presented naturally in the human intestinal tracts [54]. The remaining 5 Archaea were extremophile species that can survive in severe environments. There were 2 bacteriophages (*Geobacillus virus E2/E3*) in the 424 signatures, which have been

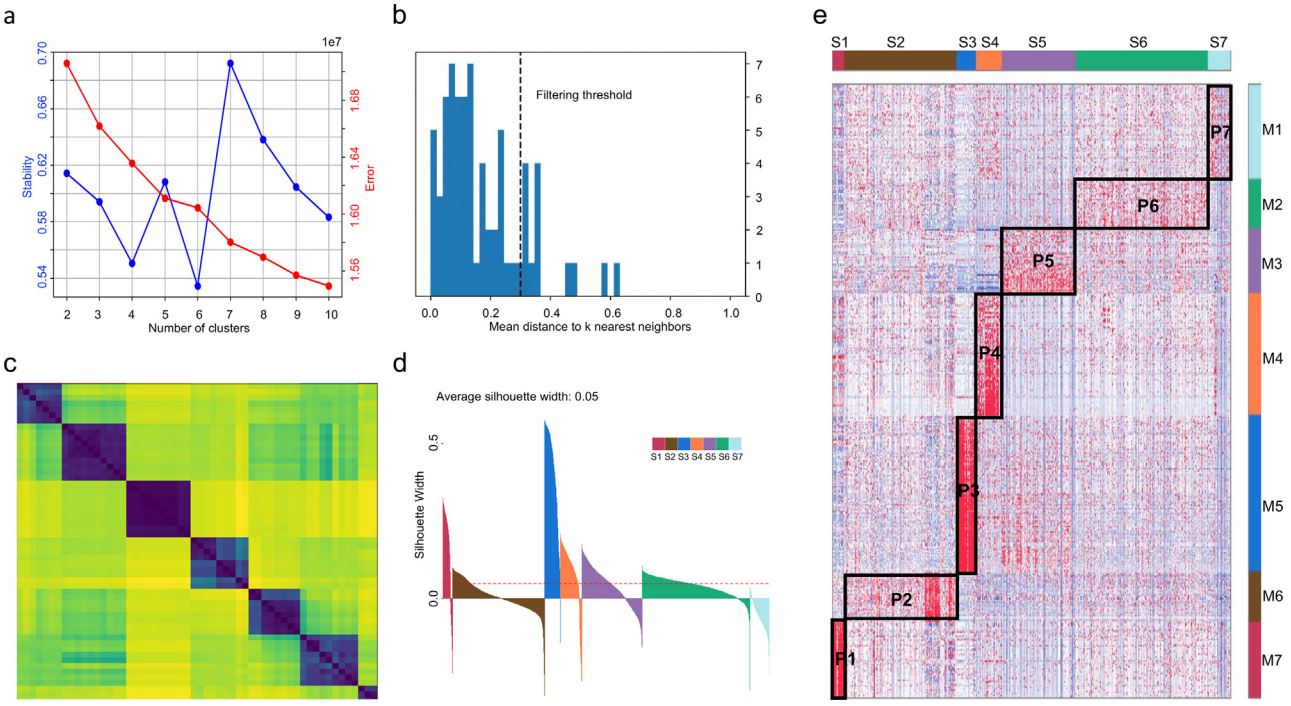


Fig. 1. Identification of seven patient-microbe interaction patterns. (a). The trade-off between stability (primary y-axis) and error (secondary y-axis) at each choice of K (x-axis). K = 7 was selected as the optimal cluster number as it reached the most stable solution and relatively lower error rate. (b). Set a density threshold of 0.3 on the histogram of average distances between clusters and their nearest neighbors. (c). The consensus matrix heatmap exhibited the clear separation of the seven clusters. (d). The Cluster Silhouette plot illustrated the Silhouette width of the seven clusters. Each cluster was shown in a different color. The average Silhouette width was 0.05, and samples with positive silhouette width were retained. (e). Heatmap of the relative abundance of the 424 signature microbes in the 6,612 patients. The 6,612 patients were classified into the seven subtypes (S1-S7), and the 424 subtype-specific microbial signatures were grouped into the seven corresponding communities (M1-M7) according to the consensus NMF biclustering. Microbial relative abundances were represented by different colors, red means higher values, and green for lower values. The heatmap displayed seven distinct and well-separated patient-microbe interaction patterns (P1-P7). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Table 1
The distribution and character of the pan-cancer subtypes and microbial signatures.

	Subtype	Number of components	Major types	Number of cancer/Class types	Character
Pattern1	S1	218	STAD (51.8%) ESCA (43.1%)	9	Upper GI-I
	M1	53	*Gammaproteobacteria* (92.4%)	5	
Pattern2	S2	1,402	OV (25.0%) STAD (14.1%) BRCA (12.4%) KIRC (9.1%)	27	Upper GI-II
	M2	35	*Actinobacteria* (82.8%)	6	
Pattern3	S3	440	BRCA (54.7%) LUSC (22.5%) KIRC (7.3%) STAD (6.1%)	13	Chest
	M3	108	*Betaproteobacteria* (93.5%)	2	
Pattern4	S4	533	UCEC (60.0%) CRC (40.0%)	2	Lower GI
	M4	84	*Gammaproteobacteria* (46.4%) *Actinobacteria* (13.1%) *Betaproteobacteria* (11.9%)	11	
Pattern5	S5	1,237	PRAD (11.7%) LIHC (9.0%) BLCA (7.9%) Adrenal gland tumors (5.6%)	27	Loss of microbiota diversity
	M5	45	*Alphaproteobacteria* (20.0%) *Actinobacteria* (13.3%)	16	
Pattern6	S6	2,719	LUAD (11.8%) LGG (9.8%) THCA (9.5%) HNSC (8.0%)	25	Normal-like
	M6	34	Many	19	
Pattern7	S7	63	THCA (30.2%) BRCA (12.7%)	13	Rare
	M7	65	*Actinobacteria* (36.9%)	18	
Total	Pan-cancer	6,612	NA	30	NA
	Microbial signature	424	NA	37	

recognised as the key players in shaping the bacterial communities in the human gut [55]. In addition, the *Picornavirales* has been detected in human fecal samples [56]. The other 2 viruses, namely *Deep-sea thermophilic phage D6E* and *Abelson murine leukemia virus,* have no direct associations with human malignancies. Interestingly, almost all the identified viruses and archaea, except for *Sulfodiicoccus acidiphilus*, were amongst the signatures of S5 and S6 (**Tables S4**).

In order to investigate if there were any relations between any of the 30 cancer types and 7 pan-cancer subtypes in the lists of significant microbes, we conducted an intersection and a correlation analysis. Significant microbial species for each cancer subtype

were identified (Table S4) based on the threshold of AUC greater than 0.6 as described in Methods. Cancer types such as ESCA, STAD, CHOL, OV, THYM, TGCT, MESO, and UCEC have higher numbers of significant species (range from 188 to 351), indicating more distinct microbial communities within these tumor microenvironments compared to other cancers. Significant microbial species identified from KIRC, HNSC, and LGG were generally low in numbers and AUC values (Table S4). The top 10 species with the highest AUC values were selected as the most significant microbes for each cancer type, respectively, and then collapsed to genus level. An UpSet plot (**Figure S3**) was generated to visualize the 37 sets (30 cancer types + 7 pan-cancer subtypes) and set intersections among different significant microbial lists using the ComplexUpset package [57] in R. Each bar in the bar chart shows a different combination (co-occurring) of intersected microbial genera. S3 and S2 have 57 and 13 unique microbial genus, making the largest and smallest numbers of set size among the 7 pan-cancer subtypes, respectively (**Figure S3**). The numbers of intersected microbial genera shown in the bar chart can be used to describe the similarities and differences among the sets. For example, possible strong associations were found between S5 and KIRC, as there were 10 shared genera (**Figure S3**). Associations may also exist among S5, BRCA, THYM, MESO, DLBC, and CHOL with 2 shared genera (**Figure S3**). No associations can be found between S1 and COAD indicated by zero shared genera (**Figure S3**). Correlation analysis using the Spearman method gave more detailed and informative results. For instance, S1 was strongly positively associated with ESCA (Figure S4). Cancer types such as OV, STAD, and BLCA were all positively correlated with S2 (Figure S4). S3 was mostly negatively correlated to different types of cancer, but positively related to THCA and BRCA (Figure S4). Similarly, S4 was only positively associated with UCEC, COAD, and READ (Figure S4). KIRC, PRAD, and many other cancer types were positively correlated with S5 (Figure S4). S6 was found to be positively associated with SKCM, LUSC, and HNSC, and negatively associated with cancer types such as MESO and TGCT (Figure S4). Finally, there were no clear correlations found between S7 and other groups (Figure S4).

3.3. Phylogenetic related bacteria enrichments in S1-3

There were 53 bacterial signatures in S1, with 92.4% of them (n = 49) belonging to the class of *Gammaproteobacteria* (Table 1**;** **Figure S2**). STAD (n = 113, 51.8%) and ESCA (n = 94, 43.1%) were the two most prevalent cancer types in this subtype (Table 1**;** **Figure S1**). Notable food- and water-borne pathogens belonging to the *Gammaproteobacteria* Class (e.g. *Salmonella, Yersinia*, and *Vibrio* spp.) were present in the S1 patients (Table S4).

The 35 bacterial signatures in S2 were significantly enriched in a sum of 1,402 cancer patients from 27 different cancer types (Fig. 1e; Table 1; adjusted $p < 0.05$; **Figure S5**). OV (n = 350, 25.0%), STAD (n = 198, 14.1%), BRCA (n = 174, 12.4%), and KIRC (n = 128, 9.1%) were the 4 major cancer types from this subtype (Table 1). More than 82.8% (n = 29) of the bacterial signatures in S2 were found in the class of *Actinobacteria* (especially in the genera of *Streptomyces* and *Micromonospora*). *Streptomyces* spp. were closely related species, and were known for their production of antibiotics and other secondary metabolites [58]. The pathogenic role of *Streptomyces* spp. remains unknown, although some of its members have been reported to cause invasive infections in immunocompromised patients [59–61]. The genus *Micromonospora* was capable of producing antimicrobial agents and other bioactive metabolites [62], and its pathogenic role in carcinogenesis has not been established.

A total of 440 cancer patients across 13 different cancer types have been classified into the S3 (Table 1**)**. The most prominent cancers in S3 were BRCA (n = 241, 54.7%), LUSC (n = 99, 22.5%), KIRC (n = 32, 7.3%), and STAD (n = 27, 6.1%) (Table 1**;** **Figure S1**). S3 contains 108 bacterial signatures that were exclusively derived from the Phylum of *Proteobacteria* (**Figure S2**), and were significantly highly enriched in these 440 patients (Fig. 1e; adjusted $p < 0.05$; **Figure S5**). More than 93.5% (n = 101) of the species were classified into the Class of *Betaproteobacteria*, and the remaining 7 were *Gammaproteobacteria*. Three orders including *Burkholderiales* (n = 71), *Rhodocyclales* (n = 13), and *Neisseriales* (n = 10) have dominated the *Betaproteobacteria* in the S3. Members of *Burkholderiales* were pathogens which were detected in cystic fibrosis patients' respiratory tract [63]. The order *Rhodocyclales* that was part of the lower respiratory tract microbiome, were observed to correlate with the inflammatory mediator IL-6 [64]. Few members of *Neisseriales* were identified to be human pathogens [65]. Collectively, more than 77% of the patients in the S3 had tumors either from the breast or chest, and were mostly enriched with harmful species from the class of *Betaproteobacteria.*

3.4. Microbial community diversity and abundance in pan-cancer subtypes

14,082 species-level microbial composition of 6,612 pan-cancer patients and 720 adjacent normal control samples were combined and processed as described previously, which resulted in a 7,534 species-level OTU table for the following microbial diversity and abundance analysis (**Table S6**).

The 7,534 species distributed primarily among the *Proteobacteria* (n = 1,569; 20.8%), *Actinobacteria* (n = 625, 8.3%), *Firmicutes* (n = 354, 4.7%), and *Bacteroidetes* (n = 261, 3.4%) phyla, as well as members of phages and viruses (n = 3,551; 47.1%). We subsequently compared the relative abundance (%) of the top 12 microbial phyla in the seven subtypes and normal control group. *Actinobacteria* and *Proteobacteria* were the two dominant phyla, representing together more than 70% of the microbiota among all groups (global-level; Fig. 2a**;** Table 2), and this percentage increased to 83% at the species-level (Fig. 2d**;** Table 2). The other major phyla includes *Firmicutes* and *Bacteroidetes*, along with bacteriophages (global-level: *Proteus phage VB_PmiS-Isfahan*; and signature-level: *Geobacillus virus E3***;** Table 2). Interestingly, the S5 subtype has the highest relative percentage of *Actinobacteria* but lowest of *Firmicutes* at both the global and signature-level (Table 2).

PCoA ordination of the 7,534 species further indicated the presence of the 7 subtypes, and significant differences were seen within the subtypes and between the normal control group (Fig. 2b; pairwise PERMANOVA, adjusted $p < 0.05$, Table S7). Pairwise Shannon diversity comparisons not only show that S5 patients had the significantly lowest overall microbial diversity as compared to the other groups (Fig. 2c; pairwise Wilcoxon test, adjusted $p < 0.05$, Table S7), but also indicated that only the S6 subtype has no differences between the normal group (Fig. 2c; pairwise Wilcoxon test, adjusted p greater than 0.05, Table S7). The above results were not only observed at the global-level with a total of 7,534 species but also at the signature-level with just 424 species (Fig. 2e-f**;** **Table S7**).

Taken together, we identified 7 pan-cancer subtypes which have significant microbial beta diversity (PCoA) differences within the subtypes and between the normal control group. S5 has reduced microbial alpha diversity and approximately 40% of the untreated cancer patients were classified into the S6 subtype, who have not experienced microbial composition changes compared to the adjacent normal control group. Furthermore, the identified 424 microbial signatures represent a snapshot of microbes that are unique to each pan-cancer subtypes.

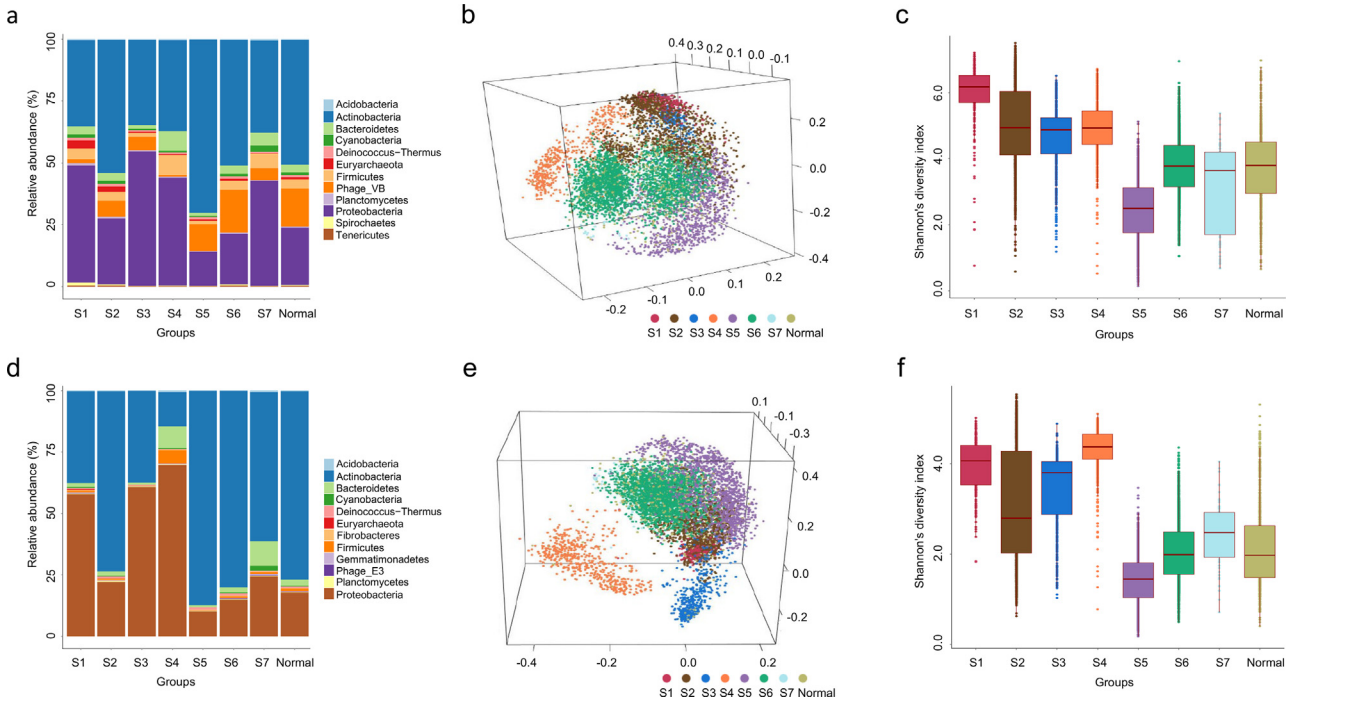


Fig. 2. Comparisons of different microbial community diversities and abundances. Taxa composition stacked bar plots illustrated the microbial relative abundance (%; in y-axis) of the top 12 phyla in the seven subtypes and normal control group (a. global-level; d. signature-level). 3D-view PCoA ordination plots based on Jaccard distances to display the species-level microbial beta diversity in different samples at the global-level (b) and signature-level (e), respectively. Each point represents a sample. A total of eight different colors were used to distinguish the seven subtypes and the normal control group. Shannon's alpha diversity index (in y-axis) of the seven subtypes and control group at the global-level (c) and signature-level (f), respectively.

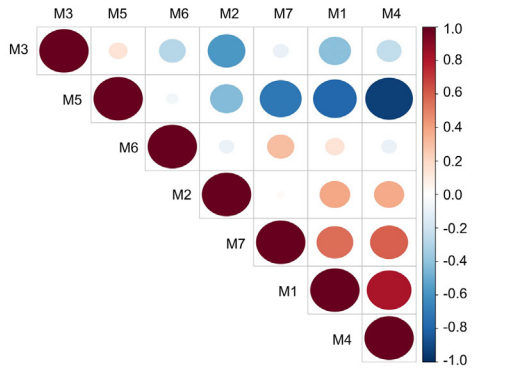
Table 2
Top 12 microbial phyla relative abundance in the pan-cancer subtypes.

			S1	S2	S3	S4	S5	S6	S7	S8
Common Phylum	*Proteobacteria*	Global	47.4	26.8	54.5	43.6	13.9	20.6	42.6	23.4
		Signature	57.9	22.2	60.8	69.7	10.1	14.8	24.4	17.8
	Actinobacteria	Global	35.0	54.0	34.7	37.0	70.2	51.1	37.4	50.6
		Signature	37.5	73.4	37.5	14.1	87.4	80.0	60.7	76.8
	Firmicutes	Global	4.4	3.7	1.7	8.3	1.4	3.7	6.1	3.9
		Signature	0.8	0.7	0.3	5.6	0.2	0.9	0.8	1.2
	Euryarchaeota	Global	3.3	2.2	0.5	0.3	0.5	0.9	0.1	0.9
		Signature	0.6	0.3	0.1	0.1	0.2	0.2	0.0	0.2
	Bacteroidetes	Global	3.3	3.2	1.4	7.8	1.3	3.2	5.2	3.1
		Signature	1.6	1.8	0.8	9.0	0.8	2.1	10.1	2.5
	Cyanobacteria	Global	1.4	1.2	0.5	0.8	0.4	1.0	2.5	1.0
		Signature	0.4	0.3	0.1	0.2	0.1	0.3	2.0	0.3
	Deinococcus-Thermus	Global	0.9	1.0	0.5	0.5	0.9	0.8	0.4	0.7
		Signature	0.3	0.5	0.2	0.3	0.8	0.8	0.2	0.5
	Planctomycetes	Global	0.7	0.6	0.3	0.4	0.1	0.3	0.2	0.3
		Signature	0.1	0.3	0.1	0.2	0.0	0.1	0.2	0.1
	Acidobacteria	Global	0.3	0.2	0.2	0.3	0.1	0.2	0.4	0.2
		Signature	0.3	0.2	0.1	0.6	0.0	0.1	0.6	0.2
	Phage_VB	Global	1.7	6.5	5.4	0.5	11.0	17.3	4.8	15.3
	Phage_E3	Signature	0.2	0.1	0.0	0.0	0.2	0.4	0.2	0.3
Distinct Phylum	*Spirochaetes*	Global	1.2	0.4	0.1	0.3	0.1	0.5	0.1	0.2
	Gemmatimonadetes	Signature	0.1	0.1	0.1	0.3	0.0	0.1	0.5	0.1
	Tenericutes	Global	0.4	0.4	0.1	0.2	0.1	0.4	0.1	0.3
	Fibrobacteres	Signature	0.1	0.1	0.0	0.1	0.1	0.0	0.1	0.0

3.5. Subtype-specific microbial communities correlation analyses

The average abundance of the microbial signatures in each microbial community were calculated and used in Spearman correlation analyses to assess associations among communities (M1-M7), microbial alpha diversity, as well as to evaluate the correlations between microbial communities and tumor-infiltrated immune/stromal cells (Fig. 3). M3 and M5 have minor positive, but negative correlations with the remaining 5 microbial communities (Fig. 3a). M1, M4, and M7 were positively associated, and the strongest positive correlation has been observed between M1 and M4 (Fig. 3a). The 7 microbial communities correlation with the alpha diversity indicated that M5 and M6 decreased, and the remaining 4 communities increased microbiome diversity (Fig. 3b). A more detailed and specific correlation analysis at the species level revealed that S5 subtype-specific microbial signatures such as *Streptomyces lividans*, MTB, and *Rhodobacter sphaeroides* were strongly negatively correlated with the alpha diversity index

a



b

	M1	M2	M3	M4	M5	M6	M7
Memory CD4 T	0	0	0	0	-0.21658	0	0
Mast cell	0	0	0	0	-0.26674	0	0
Natural killer T	0	0	0	0	0.205761	0	0
Osteoblast	0	0	0	0	0.225931	0	0
T helper type 2	0	0	0	0	-0.20387	0	0
Regulatory T	0	0	0	0	0	0.21817	0
Mesangial cell	0	0.236262	0	0	-0.21841	0	0
Microbial Div	0.40710	0.39217	0.03792	0.60224	-0.65505	-0.11451	0.21607

c

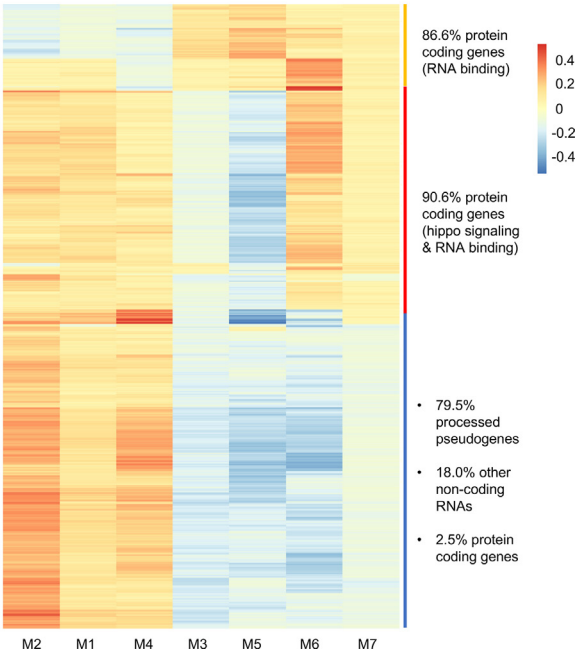


Fig. 3. Subtype-specific microbial communities' correlation with immune cells and human genes. (a). Correlogram displaying the Spearman's correlation for all pairs of microbial communities (M1-M7) comparisons. The area of the dots were proportional to their correlation coefficients, and the color indicated the strength of the correlation (red for positive, and blue for negative correlations). (b). Spearman's correlation coefficient table among abundance of the seven microbial communities with the seven selected significant tumor-infiltrating cell types and microbial alpha diversity (microbial div). (c). Spearman's correlation coefficients heatmap between the 696 genes (in rows) and the abundance of the seven microbial communities (in columns). Positive and negative correlations were shown in red and blue, respectively. The percentages of different categories of RNAs were shown on the right hand of the heatmap. Protein coding genes were used for functional enrichment analysis, and the enriched pathways were included in the bracket. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

(Table S4). *Actinobacteria bacterium IMCC25003, Candidatus Planktophila sulfonica, Rhodoluna lacicola,* among several others were strongly positively correlated with the alpha diversity index (Table S4). Interestingly, the most frequent positive and negative associations were seen in S4 and S5, respectively (Table S4).

Correlations between the 7 microbial communities and 64 tumor-infiltrating cell types identified key cellular components (including memory CD4 T, natural killer T, regulatory T, T helper type 2, osteoblasts, mast and mesangial cells) that were significantly associated with at least one microbial community (adjusted $p < 0.05$), although the correlation coefficients were generally small ($|r| < 0.3$; Fig. 3b). M5 was noticeable as it has correlated with more cellular components than other communities (Fig. 3b). For example, memory CD4 T, T helper type 2, mast and mesangial cells were significantly negatively correlated with the M5. Significant positive associations were found between M5 and natural killer T or osteoblasts (Fig. 3b). Additionally, the absolute correlation coefficients between tumor-infiltrating cells with the M1, M3, M4, and M7 were too small (<0.2) to consider separately (Fig. 3b).

A total of 696 genes were significantly correlated with the abundance of the 7 microbial communities (adjusted $p < 0.05$), with more than half of them (382; 54.9%) being non-coding RNAs (**Table S8**). Of note, the processed pseudogenes account for a significant proportion of the associated genes (302; 43.4%). A correlation heatmap was subsequently constructed based on the Spearman's correlation coefficients between the 696 genes and the 7 communities (Fig. 3c). More positive than negative correlations have been seen in M1, M2, and M4, which were opposite in correlation directions observed in M3 and M5. The number of genes which were either positively or negatively correlated with the M6 and M7 were almost equal (Fig. 3c**; Table S8**). Among the 355 genes which were negatively correlated with the M6, 282 were processed pseudogenes (**Table S8**). There were 64 other classes of non-coding RNAs, and only 9 of them were protein coding genes (**Table S8**). This group of genes were negatively correlated with the M5, M3, and M7, and positively associated with the M4, M2, and M1, respectively (Fig. 3c). A total of 341 genes were positively correlated with the M6, and 305 of them were protein coding genes. Functional enrichment analysis indicated that these genes were significantly involved in pathways such as hippo signaling, RNA and snRNA binding (**Table S9**). Additionally, 84 out of the 97 genes which were negatively correlated with the M4 were protein coding genes, and were significantly involved in 4 RNA/snRNA/protein binding pathways (**Table S9**).

In sum, the identified 7 pan-cancer microbial communities have connections to each other. As the species have the lowest alpla diversity and more immune cells were mostly negatively correlated with the M5, we speculate that dysbiosis may disrupt immune homeostasis. In regards to patient-microbe interactions, the 7 microbial communities were not only correlated with the patients' protein coding genes involved in the binding related pathways, but also correlated with many non-coding RNAs, especially with the processed pseudogenes.

3.6. CRC intratumor microbial heterogeneity

Colorectal cancer (CRC) is the third most prevalent and lethal type of cancer worldwide [66]. CRC is more common in developed countries, and is strongly linked to risk factors such as low-fiber and high-fat diet [67], lack of physical activity [68], alcohol-tobacco consumption, obesity, and dysbiosis [69]. Similar to what we found previously [31], CRC is largely composed of two subtypes including S4 and S6, characterized by distinct microbial profiles (Fig. 4).

The cancer types in S4 were exclusively from the UCEC and CRC (Table 1). More specifically, S4 contains 533 cancer patients with

60.0% of them (n = 320) were UCEC, and the remaining 40.0% were CRC (including 145 COAD, 42 READ, and 26 Rectosigmoid junction cancer; Table 1; **Table S5**). There were 84 microbial signatures (all bacteria) that significantly enriched in S4, and more than 46.4% (n = 39) of them were from the order of *Enterobacterales* (with the class of *Gammaproteobacteria;* Table 1; **Figure S2**). As more than 90% of the microbial signatures in S1 were classified as *Gammaproteobacteria* as well (Table 1), the phylogenetic species relatedness may explain the observed positive correlation between M1 and M4 (Fig. 3a). The other dominated species in S4 include members of *Bacteroidales* and *Clostridiales*, which were in high abundance and generally beneficial to human health (**Table S4**; Table 2). In addition, the highly enriched *Enterobacterales* were facultative anaerobes with the majority of them being pathogenic to humans. For example, *Shigella* spp. can infect and cause severe inflammation and dysentery in the human colon [70]. *Citrobacter* spp. were opportunistic intestinal and urinary tracts pathogens [71]. The other two pathogenic *Enterobacterales* spp. (*Salmonella enterica* and *Yersinia pestis*) have also been found to be enriched in S1 (Upper GI subtype).

A total of 4 archaea, 1 virus, and 29 bacterial species were signatures in S6. Unlike the taxa-dominated enrichment patterns in S1-4, there were no such patterns observed in the S6 subtype (**Figure S2**). S6 contains 25 different cancer types, and more than half of the CRC cases as mentioned above were classified into this subtype. There were altogether 428 CRC patients in either the S4 or S6 subtypes, and no significant OS and PFS differences were observed between them (Fig. 4b-c) Interestingly, the 245 S6 patients from the CRC had enriched similar microbial signatures with the normal control group (n = 720), indicating that the S6 had "normal-like" microbial communities and diversities (Fig. 4d-f).

3.7. Functional annotation and clinical significance of the pan-cancer subtypes

GSEA was performed to identify significant gene sets in each subtype. A total of 266 most highly enriched common gene sets from the seven pan-cancer subtypes (adjusted p-value < 0.05; **Table S10**) were selected. Except for the S2 subtype, subtype-specific gene sets were identified and shown in the p-value heatmap (Fig. 5a). For example, multiple immune-related gene sets were up-regulated in the S1 and S3, but were down-regulated in the S5 subtype. S3 subtype has extracellular matrix (ECM) and invasive-associated gene sets up-regulation (Fig. 5a). S5 has been enriched with lipoprotein assembly and metabolism-related gene sets. Moreover, cell cycle and CRC-related gene sets were enriched in the S4 subtype (Fig. 5a). Canonical cancer-related gene sets were enriched in the S6 subtype. Finally, cell signaling or cell communication pathways such as EGFR/TGFB/TNF signaling and fibronectin binding were up-regulated in the S7 (Fig. 5a).

The associations among pan-cancer subtypes and patients' clinical outcomes in terms of OS and PFS were assessed using the multivariable Cox PH regression analysis. The Cox model was adjusted for confounding factors such as age, gender, tumor stage and grade. Hazard ratio (HR) of age on OS and PFS were all close to 1, indicating that age's effect on survival was not clear (Fig. 5b-c). Compared to female patients, males have significantly lower survival rates (log-rank p-value < 0.05; HR greater than 1). Although many patients' tumor stage and grade information were not available, higher stage was basically a poor prognostic factor for both OS and PFS, and higher grade has generally worse prognosis compared to the lower grade cases (Fig. 5b-c). The "normal-like" subtype S6 was serving as the reference category in the multivariate Cox

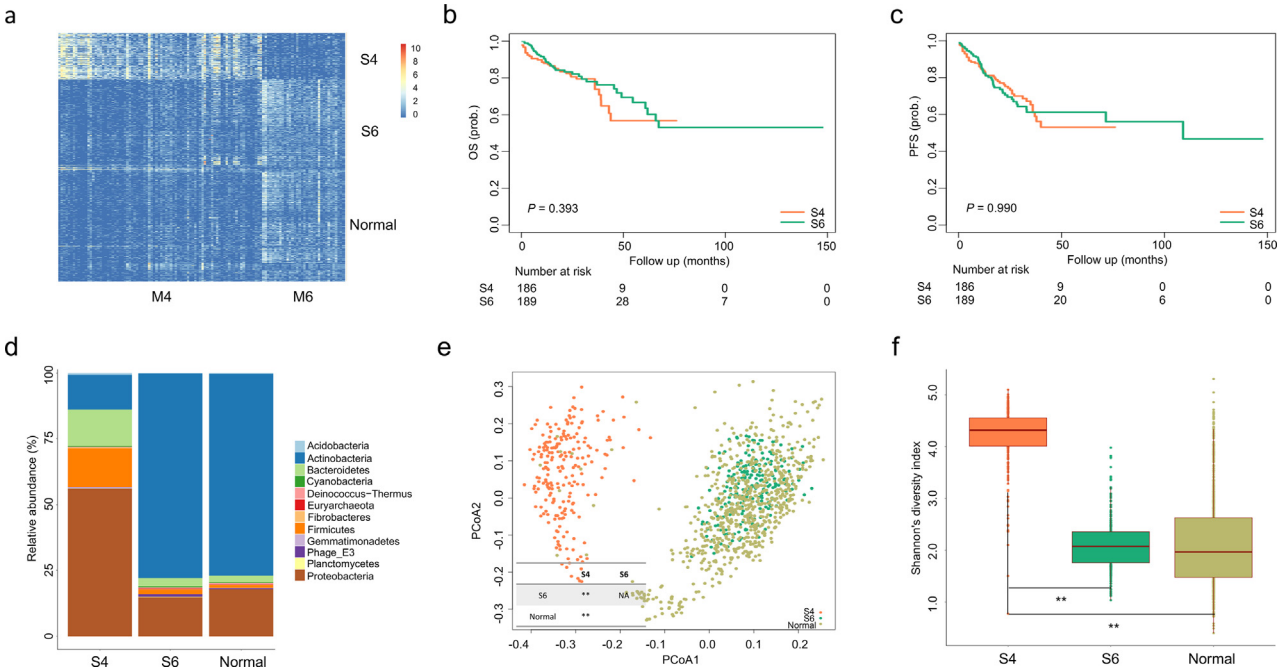


Fig. 4. Identification of two microbial subtypes of CRC. (a). Heatmap of the relative abundance of the M4 and M6 microbial signatures in the two CRC subtypes (S4 and S6) and normal control patients. Microbial relative abundances were represented by different colors, red means higher values, and green for lower values. The heatmap displayed two well-separated patient-microbe interaction patterns. Kaplan-Meier survival curves comparing OS (b) and PFS (c) of the two subtypes. The indicated p-values were obtained by the log-rank tests. (d). Taxa composition stacked bar plots illustrated the microbial relative abundance (%; in y-axis) of the top 12 phyla in the two subtypes and normal control group. (e). 2D-view PCoA ordination plot based on Jaccard distances to display the species-level microbial beta diversity in different samples (each point represents a sample). Pairwise PERMANOVA was used for testing the differences among groups. (f). Shannon's alpha diversity index in the three groups. Pairwise Wilcoxon tests (with BH correction) were used to compare the alpha diversity differences among groups. P-values<0.05 were considered significant, and asterisks denote significant levels (*: p < 0.05; **: p < 0.01). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

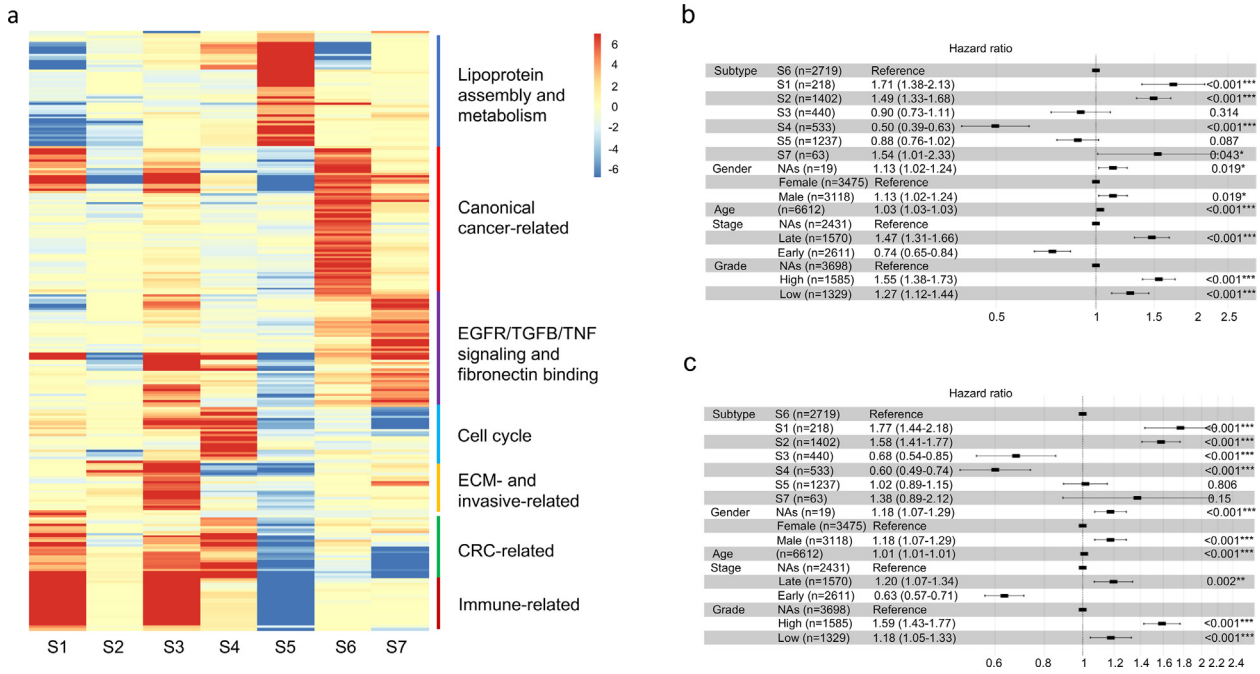


Fig. 5. Functional and clinical characterizations of the pan-cancer subtypes. (a). A p-value heatmap displaying the overlapped dysregulated gene sets (in rows) in the seven pan-cancer subtypes (in columns). Values in the heatmap equal to -log10 (FDR adjusted p-value). Red color indicates gene sets with up-regulations, and green with down-regulations. Gene sets were grouped into several different functional categories, and their functional names were shown on the right hand of the heatmap. Forest plots of Cox regression models for OS (b) and PFS (c), which illustrated the HRs, 95% CIs, and log-rank p-values for pan-cancer subtypes and confounder factors (age, gender, tumor stage, and grade). The p-value significance levels were labeled as follows: * $p < 0.05$, ** $p < 0.01$, and *** $p < 0.001$. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

model. S3 and S4 have favorable clinical outcomes, whereas S1, S2 and S7 predict poor OS and PFS. In addition, S5 was a good prognostic factor for OS, although not statistically significant (Fig. 5b-c).

4. Discussion

Pan-cancer analysis of genomes, transcriptomes, and beyond have identified subtype-specific genomic patterns, expression programs, distinct cellular identity and activities, which increased our understanding of tumor heterogeneity. Microorganisms live on and inside our bodies, and are increasingly being recognized for their roles in human health and disease. Cancer patients are immunocompromised and more vulnerable to getting microbial infections. However, identifying the cancer-associated microbiome and exploring the pan-cancer microbial heterogeneity are still in the early stages. Therefore, in our investigations, we sought to define the tumor microbiomes and did the first attempt of revealing the potential pan-cancer heterogeneity at the microbial level. Microbial compositional profiles were estimated from the unmapped transcriptome sequencing data. As the microbial community data share certain similar characteristics with single cell transcriptomic data, which is sparse and high-dimensional, we used a dimensionality reduction technique developed for single cell data (cNMF) to factorize the microbial compositional profiles. A total of 7 pan-cancer subtypes with distinct microbial community compositions and diversities were identified and characterized for their molecular and clinical significance. Each pan-cancer subtype was enriched for a group of microbial species that show significant over-representation over the 7 subtypes.

We found that pan-cancer subtype S1, S2, and S3 each individually enriched with phylogenetic related species More specifically, among the 53 bacteria signatures in S1, 49 (92.4%) of them belong to the class of *Gammaproteobacteria*. Likewise, more than 82.8% and 93.5% of microbial signatures identified from S2 and S3 were in the class of *Actinobacteria* and *Betaproteobacteria,* respectively.

Proteobacteria is a phylum of Gram-negative bacteria, with many of them having nitrogen fixation/metabolism properties. There are five major classes of the *Proteobacteria* including *Alpha-, Beta-, Gamma-, Delta-*, and *Epsilon-Proteobacteria*, which have been identified in various human body sites [4]. The *Alphaproteobacteria* is highly diverse and adaptable, which can survive with very few nutrients. *Alphaproteobacteria* has been identified to be the dominant phyla in the primate brain microbiome [72]. The class *Betaproteobacteria* is free-living aerobic and anaerobic bacteria. In our study, *Betaproteobacteria* was found to be dominant in chest-related cancers such as BRCA and LUSC. *Burkholderiales* is the most abundant taxon within the *Betaproteobacteria*, accounting for more than 65% of the microbial signatures in S3. Members of *Burkholderiales* including *Bordetella holmesii* and *pandoraea pulmonicola* are serious human pathogens [73,74] *Gammaproteobacteria* is a large class of microbes with many of them exist as commensals, and others are well-known human and animal pathogens including *Salmonella, Yersinia*, and *Vibrio*. *Salmonella* spp. are important causes of salmonellosis, characterized by inflammation of the intestine. Two species of *Salmonella*, namely *S. enterica* and *S. bongori* are in our 424 microbial list. *S. enterica* in S1, which is a life-threatening foodborne bacteria that poses a threat to human and animal health. *S. bongori* is associated with mild symptoms and classified into the S4. *Yersinia* spp. are responsible for a number of human diseases ranging from yersiniosis to plague. *Yersinia pestis*, the causative agent of bubonic plague (Black Death), has been detected in the presence of nucleic acid in S1 patients. *Vibrio* spp. are salt tolerant bacteria occurring naturally in the marine environment. The majority of *Vibrio* infections in humans are foodborne, coming from infected seafood. As around 95% of the cancer types in S1 were diagnosed with upper GI cancers, we speculate that the microbial

signatures in S1 are associated with upper GI cancer. The *Gammaproteobacteria* spp. also account for a significant proportion (greater than50%) of the microbial signatures in S4, which is a pan-cancer subtype dominated by a subset of lower GI cancer (CRC).

Actinobacteria was the second most predominant phylum after *Proteobacteria*, and was more abundant in the subtypes of S2,S5, S6, and S7. *Actinobacteria* share the similar morphological and functional properties of both bacteria and fungi, and most of them are aerobic spore-forming bacteria that play important roles in organic matter decomposition [58]. In our study, there were more than 25 different cancer types classified into the S2, S5, and S6, respectively. Take example of S2, almost all ovarian (OV, n = 350) and glioblastoma (GBM, n = 99), as well as 60% STAD (n = 198), 50% ESCA (n = 51), 50% KIRC (n = 128), 30% BRCA (n = 174), and other 21 different cancer types were included. S2 has 29 microbial signatures in the class *Actinobacteria*, and several closely related *Streptomyces* spp. dominated in this pan-cancer subtype. Soil-derived *Streptomyces* spp. have established symbiotic relationships with humans and were present in various body sites [75]. Most of these species were capable of producing antibiotics and other bioactive metabolites, making them promising candidates for inflammatory diseases and cancer [58,62,76]. Soil is part of human's natural habitat, which supports a variety of organisms and microorganisms [15,77]. Many phylogenetic and functional similarities have been found between the human intestinal microbial niche with the soil/root microbial ecosystems [15,78]. By identifying a group of human associated *Streptomyces* spp., our study helps to better understand the link between soil microbes and human health and diseases.

The patients classified into the S5 had the lowest microbial alpha diversity as compared to other subtypes. Although the 36 bacteria taxa distribution in S5 is wide and diverse, S5 patients had a significant amount of *Mycobacterium tuberculosis* (TB) enrichment (**Table S7**). Approximately one-quarter of the world population have latent MTB infection, according to the World Health Organization (WHO). Immunocompromised patients are at higher risk of activation of latent MTB infection to active TB disease [79,80]. Previous studies found that MTB infection was associated with decreased gut microbiome diversity in both mice [81] and humans [82]. In the present study, we observed the co-occurrence of TB enrichment and loss of microbiota diversity in a wide range of body sites (27 types of cancer involving different organs), and the dysbiosis disrupts the immune functions in the S5 patients. Additionally, one member of the *mycobacterium tuberculosis* complex (MTC): *M. kansasii*, had been highly enriched in the S5 as well. Overrepresentation of the MTC sequences in a subset of these patients may contribute to contamination [83]. Another potential contamination issue was found in CRC patients classified as S4, where it has higher microbial alpha diversity compared to the control group, and this is contrary to what we found previously using 16S rRNA amplicon sequencing datasets [31]. Thus, contamination issues should be a concern when working with unmapped RNA-Seq data, and a multi-analysis involving multiple independent cohorts is preferable.

In our study, S6 was one of the largest pan-cancer subtypes with more than 25 different cancer types, and S6-specific microbial signatures spanning all major phyla in bacteria. S6 accounts for approximately 40% of the untreated cancer patients who have not experienced microbial composition changes compared to the adjacent normal control group. In these "normal-like" patients, the overrepresented signatures as a whole microbial community were positively correlated with genes involved in hippo signaling, RNA and snRNA binding pathways, and were negatively correlated with many processed pseudogenes. Further studies are needed to understand the underlying mechanisms responsible for the functional differences between these coding and non-coding genes. Moreover, studying cancer microbial community structures and functional profiles will aid future works on designing more effective and targeted cancer therapies.

5. Conclusions

In conclusion, we did the first attempt of revealing the pan-cancer heterogeneity at the microbial level. A total of 7 pan-cancer subtypes (S1-S7) and 424 subtype-specific microbial signatures were identified and characterized for their functional role and clinical significance. Phylogenetic related bacteria signatures were overrepresented in S1, S2, and S3. CRC has been classified into the S4 and S6 subtypes. S4 contains many pathogenic *Enterobacterales* spp., and was enriched with cell cycle and CRC-related gene sets. S6 had "normal-like" features, and was one of the largest pan-cancer subtypes spanning more than 25 different cancer types. S5 patients' immune functions were impaired by the loss of microbial diversity. Lastly, the rare pan-cancer subtype S7 predicts poor survival and has cell to cell communication related gene sets upregulation. Our study not only examined and characterized the pan-cancer heterogeneity at the microbial level, but also provided promising therapeutic targets for cancer treatment.

Competing interests

The authors declare that they have no competing interests.

Funding

The study was supported by the grant from the National Natural Science Foundation of China (No. 81974469).

CRediT authorship contribution statement

Lan Zhao: Conceptualization, Methodology, Data curation, Visualization, Investigation, Writing – original draft. **William C.S. Cho:** . **Jun-Li Luo:** Supervision.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.csbj.2022.06.012.

References

[1] Rowland I, Gibson G, Heinken A, Scott K, Swann J, Thiele I, et al. Gut microbiota functions: metabolism of nutrients and other food components. Eur J Nutr 2018;57:1–24.
[2] Zheng D, Liwinski T, Elinav E. Interaction between microbiota and immunity in health and disease. Cell Res 2020;30:492–506.
[3] Scheld WM, Michael SW. Introduction to Microbial Disease. Goldman's Cecil Medicine 2012:1761–2. https://doi.org/10.1016/b978-1-4377-1604-7.00286-4.
[4] Human Microbiome Project Consortium. Structure, function and diversity of the healthy human microbiome. Nature 2012;486:207–14.
[5] Coureuil M, Lécuyer H, Bourdoulous S, Nassif X. A journey into the brain: insight into how bacterial pathogens cross blood–brain barriers. Nat Rev Microbiol 2017;15:149–59.
[6] Fowler Jr VG, Miro JM, Hoen B, Cabell CH, Abrutyn E, Rubinstein E, et al. Staphylococcus aureus endocarditis: a consequence of medical progress. JAMA 2005;293:3012–21.
[7] Emery DC, Shoemark DK, Batstone TE, Waterfall CM, Coghill JA, Cerajewska TL, et al. 16S rRNA Next Generation Sequencing Analysis Shows Bacteria in

Alzheimer's Post-Mortem Brain. Frontiers in Aging Neuroscience 2017;9. https://doi.org/10.3389/fnagi.2017.00195.
[8] Dickson RP, Erb-Downward JR, Freeman CM, McCloskey L, Falkowski NR, Huffnagle GB, et al. Bacterial Topography of the Healthy Human Lower Respiratory Tract. MBio 2017;8. https://doi.org/10.1128/mBio.02287-16.
[9] Pushalkar S, Hundeyin M, Daley D, Zambirinis CP, Kurz E, Mishra A, et al. The Pancreatic Cancer Microbiome Promotes Oncogenesis by Induction of Innate and Adaptive Immune Suppression. Cancer Discovery 2018;8:403–16. https://doi.org/10.1158/2159-8290.cd-17-1134.
[10] Meng S, Chen B, Yang J, Wang J, Zhu D, Meng Q, et al. Study of Microbiomes in Aseptically Collected Samples of Human Breast Tissue Using Needle Biopsy and the Potential Role of in situ Tissue Microbiomes for Promoting Malignancy. Frontiers in Oncology 2018;8. https://doi.org/10.3389/fonc.2018.00318.
[11] Leiby JS, McCormick K, Sherrill-Mix S, Clarke EL, Kessler LR, Taylor LJ, et al. Lack of detection of a human placenta microbiome in samples from preterm and term deliveries. Microbiome 2018;6. https://doi.org/10.1186/s40168-018-0575-4.
[12] Rastall RA. Bacteria in the gut: friends and foes and how to alter the balance. J Nutr 2004;134:2022S–6S.
[13] Rizzatti G, Lopetuso LR, Gibiino G, Binda C, Gasbarrini A. Proteobacteria: A Common Factor in Human Diseases. Biomed Res Int 2017;2017:9351507.
[14] Faith JJ, Guruge JL, Charbonneau M, Subramanian S, Seedorf H, Goodman AL, et al. The long-term stability of the human gut microbiota. Science 2013;341:1237439.
[15] Blum WEH, Zechmeister-Boltenstern S, Keiblinger KM. Does Soil Contribute to the Human Gut Microbiome? Microorganisms 2019;7. https://doi.org/10.3390/microorganisms7090287.
[16] Mitchell PD, O'Byrne PM. Biologics and the lung: TSLP and other epithelial cell-derived cytokines in asthma. Pharmacol Ther 2017;169:104–12.
[17] Haque M, Sartelli M, McKimm J, Abu BM. Health care-associated infections - an overview. Infect Drug Resist 2018;11:2321–33.
[18] Zhao L, Lee VHF, Ng MK, Yan H, Bijlsma MF. Molecular subtyping of cancer: current status and moving toward clinical applications. Brief Bioinform 2018. https://doi.org/10.1093/bib/bby026.
[19] Olivier M, Hollstein M, Hainaut P. TP53 mutations in human cancers: origins, consequences, and clinical use. Cold Spring Harb Perspect Biol 2010;2:a001008.
[20] Godet I, Gilkes DM. BRCA1 and BRCA2 mutations and treatment strategies for breast cancer. Integrative Cancer Science and Therapeutics 2017;4.
[21] Loeb LA, Harris CC. Advances in chemical carcinogenesis: a historical review and prospective. Cancer Res 2008;68:6863–72.
[22] Anand P, Kunnumakkara AB, Sundaram C, Harikumar KB, Tharakan ST, Lai OS, et al. Cancer is a preventable disease that requires major lifestyle changes. Pharm Res 2008;25:2097–116.
[23] Walboomers JM, Jacobs MV, Manos MM, Bosch FX, Kummer JA, Shah KV, et al. Human papillomavirus is a necessary cause of invasive cervical cancer worldwide. J Pathol 1999;189:12–9.
[24] Shiotani A, Cen P, Graham DY. Eradication of gastric cancer is now both possible and practical. Semin Cancer Biol 2013;23:492–501.
[25] Maucort-Boulch D, de Martel C, Franceschi S, Plummer M. Fraction and incidence of liver cancer attributable to hepatitis B and C viruses worldwide. Int J Cancer 2018;142:2471–7.
[26] Yu J, Feng Q, Wong SH, Zhang D, Liang QY, Qin Y, et al. Metagenomic analysis of faecal microbiome as a tool towards targeted non-invasive biomarkers for colorectal cancer. Gut 2017;66:70–8.
[27] Eslami-S Z, Majidzadeh-A K, Halvaei S, Babapirali F, Esmaeili R. Microbiome and Breast Cancer: New Role for an Ancient Population. Front Oncol 2020;10:120.
[28] Zhou B, Sun C, Huang J, Xia M, Guo E, Li N, et al. The biodiversity Composition of Microbiome in Ovarian Carcinoma Patients. Sci Rep 2019;9:1691.
[29] Rosa GL, La Rosa G, Gattuso G, Pedull E, Rapisarda E, Nicolosi D, et al. Association of oral dysbiosis with oral cancer development (Review). Oncology Letters 2020. https://doi.org/10.3892/ol.2020.11441.
[30] Wei M-Y, Shi S, Liang C, Meng Q-C, Hua J, Zhang Y-Y, et al. The microbiota and microbiome in pancreatic cancer: more influential than expected. Mol Cancer 2019;18:97.
[31] Zhao L, Cho WC, Nicolls MR. Colorectal Cancer-Associated Microbiome Patterns and Signatures. Front Genet 2021;12. https://doi.org/10.3389/fgene.2021.787176.
[32] Wood DE, Lu J, Langmead B. Improved metagenomic analysis with Kraken 2. Genome Biol 2019;20:257.
[33] Zhao Lan, Grimes Susan, Greer Stephanie, Kubit Matthew, Lee HoJoon, Nadauld Lincoln, et al. Characterization of the consensus mucosal microbiome of colorectal cancer. NAR cancer 2021;3(4):zcab049.
[34] McMurdie PJ, Holmes S. phyloseq: an R package for reproducible interactive analysis and graphics of microbiome census data. PLoS ONE 2013;8:e61217.
[35] Ginestet C. Ggplot2: Elegant graphics for data analysis. J R Stat Soc Ser A Stat Soc 2011;174:245–6.
[36] Eisenhofer R, Minich JJ, Marotz C, Cooper A, Knight R, Weyrich LS. Contamination in Low Microbial Biomass Microbiome Studies: Issues and Recommendations. Trends Microbiol 2019;27:105–17.
[37] Kotliar D, Veres A, Aurel Nagy M, Tabrizi S, Hodis E, Melton DA, et al. Identifying gene expression programs of cell-type identity and cellular activity with single-cell RNA-Seq. eLife 2019;8. https://doi.org/10.7554/elife.43803.
[38] Carmona-Saez P, Pascual-Marqui RD, Tirado F, Carazo JM, Pascual-Montano A. Biclustering of gene expression data by Non-smooth Non-negative Matrix Factorization. BMC Bioinf 2006;7:78.
[39] Rousseeuw PJ. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. J Comput Appl Math 1987;20:53–65.
[40] Oksanen J, Blanchet FG, Kindt R, Legendre P, Minchin PR, O'hara RB, et al. Package "vegan." Community Ecology Package, Version 2013;2:1–295.
[41] Anderson MJ. Permutational multivariate analysis of variance (PERMANOVA). Wiley StatsRef: Statistics Reference Online 2017:1–15. https://doi.org/10.1002/9781118445112.stat07841.
[42] Dixon P. VEGAN, a package of R functions for community ecology. J Veg Sci 2003;14:927–30.
[43] Adler D, Nenadic O, Zucchini W. Rgl: A r-library for 3d visualization with opengl. Proceedings of the 35th Symposium of the Interface: Computing Science and Statistics, Salt Lake City, vol. 35, 2003, p. 1–11.
[44] Aran D, Hu Z, Butte AJ. xCell: digitally portraying the tissue cellular heterogeneity landscape. Genome Biol 2017;18:220.
[45] Anders S, Pyl PT, Huber W. HTSeq—a Python framework to work with high-throughput sequencing data. Bioinformatics 2014;31:166–9.
[46] Kolde R. pheatmap: Pretty heatmaps [Software]. URL https://CRAN R-Project Org/package= Pheatmap 2015.
[47] Wei T, Simko V. R package "corrplot": Visualization of a Correlation Matrix (Version 0.84) 2017.
[48] Kuleshov MV, Jones MR, Rouillard AD, Fernandez NF, Duan Q, Wang Z, et al. Enrichr: a comprehensive gene set enrichment analysis web server 2016 update. Nucleic Acids Res 2016;44:W90–7.
[49] Ritchie ME, Phipson B, Wu D, Hu Y, Law CW, Shi W, et al. limma powers differential expression analyses for RNA-sequencing and microarray studies. Nucleic Acids Res 2015;43:e47.
[50] Väremo L, Nielsen J, Nookaew I. Enriching the gene set analysis of genome-wide data by incorporating directionality of gene expression and combining statistical hypotheses and methods. Nucleic Acids Res 2013;41:4378–91.
[51] Kolde R. Pheatmap: pretty heatmaps. R Package Version 2012;1.
[52] Kassambara A, Kosinski M, Biecek P, Fabian S. survminer: Drawing Survival Curves using'ggplot2'. R Package Version 2017(3):1.
[53] Hamady M, Knight R. Microbial community profiling for human microbiome projects: Tools, techniques, and challenges. Genome Res 2009;19:1141–52.
[54] Gaci N, Borrel G, Tottey W, O'Toole PW, Brugère J-F. Archaea and the human gut: new beginning of an old story. World J Gastroenterol 2014;20:16062–78.
[55] Sutton TDS, Hill C. Gut Bacteriophage: Current Understanding and Challenges. Front Endocrinol 2019;10:784.
[56] Munnink BBO, Cotten M, Deijs M, Jebbink MF, Bakker M, Farsani SMJ, et al. A novel genus in the order Picornavirales detected in human stool. J Gen Virol 2015;96:3440–3.
[57] Krassowski. ComplexUpset: Create Complex UpSet PlotsUsing ggplot2 Components. R Package Version 05 n.d.
[58] Ranjani A, Dhanasekaran D, Gopinath PM. An Introduction to Actinobacteria. Actinobacteria - Basics and Biotechnological Applications 2016. https://doi.org/10.5772/62329.
[59] Kapadia M, Rolston KVI, Han XY. Invasive Streptomyces infections: six cases and literature review. Am J Clin Pathol 2007;127:619–24.
[60] Rose CE, Brown JM, Fisher JF. Brain Abscess Caused by Streptomyces Infection following Penetration Trauma: Case Report and Results of Susceptibility Analysis of 92 Isolates of Streptomyces Species Submitted to the CDC from 2000 to 2004. J Clin Microbiol 2008;46:821–3. https://doi.org/10.1128/jcm.01132-07.
[61] Ariza-Prota MA, Pando-Sandoval A, Fole-Vázquez D, García-Clemente M, Budiño T, Casan P. Community-acquired bacteremic Streptomyces atratus pneumonia in an immunocompetent adult: a case report. J Med Case Rep 2015;9:262.
[62] Hifnawy MS, Fouda MM, Sayed AM, Mohammed R, Hassan HM, AbouZid SF, et al. The genus Micromonospora as a model microorganism for bioactive natural product discovery. RSC Adv 2020;10:20939–59.
[63] Voronina OL, Kunda MS, Ryzhova NN, Aksenova EI, Sharapova NE, Semenov AN, et al. On Burkholderiales order microorganisms and cystic fibrosis in Russia. BMC Genomics 2018;19:74.
[64] Li K-J, Chen Z-L, Huang Y, Zhang R, Luan X-Q, Lei T-T, et al. Dysbiosis of lower respiratory tract microbiome are associated with inflammation and microbial function variety. Respir Res 2019;20:272.
[65] Humbert MV, Christodoulides M. Atypical, Yet Not Infrequent, Infections with Neisseria Species. Pathogens 2019;9. https://doi.org/10.3390/pathogens9010010.
[66] Rawla P, Sunkara T, Barsouk A. Epidemiology of colorectal cancer: incidence, mortality, survival, and risk factors. Gastroenterology Review 2019;14:89–103. https://doi.org/10.5114/pg.2018.81072.
[67] Yang J, Yu J. The association of diet, gut microbiota and colorectal cancer: what we eat may imply what we get. Protein & Cell 2018;9:474–87. https://doi.org/10.1007/s13238-018-0543-6.
[68] Shaw E, Farris MS, Stone CR, Derksen JWG, Johnson R, Hilsden RJ, et al. Effects of physical activity on colorectal cancer risk among family history and body mass index subgroups: a systematic review and meta-analysis. BMC Cancer 2018;18:71.
[69] Sobhani I, Amiot A, Le Baleur Y, Levy M, Auriault M-L, Van Nhieu JT, et al. Microbial dysbiosis and colon carcinogenesis: could colon cancer be considered a bacteria-related disease? Therap Adv Gastroenterol 2013;6:215–29.

[70] Sansonetti P. Rupture, invasion and inflammatory destruction of the intestinal barrier by Shigella, making sense of prokaryote–eukaryote cross-talks. FEMS Microbiol Rev 2001;25:3–14. https://doi.org/10.1016/s0168-6445(00)00060-7.
[71] Ranjan KP, Ranjan N. Citrobacter: An emerging health care associated urinary pathogen. Urol Ann 2013;5:313–4.
[72] Branton WG, Ellestad KK, Maingat F, Wheatley BM, Rud E, Warren RL, et al. Brain microbial populations in HIV/AIDS: α-proteobacteria predominate independent of host immune status. PLoS ONE 2013;8:e54673.
[73] Pittet LF, Emonet S, Schrenzel J, Siegrist C-A, Posfay-Barbe KM. Bordetella holmesii: an under-recognised Bordetella species. Lancet Infect Dis 2014;14:510–9.
[74] Degand N, Lotte R, Decondé Le Butor C, Segonds C, Thouverez M, Ferroni A, et al. Epidemic spread of Pandoraea pulmonicola in a cystic fibrosis center. BMC Infect Dis 2015;15:583.
[75] Herbrík A, Corretto E, Chroňáková A, Langhansová H, Petrásková P, Petrásková P, et al. A Human Lung-Associated Streptomyces sp. TR1341 Produces Various Secondary Metabolites Responsible for Virulence, Cytotoxicity and Modulation of Immune Response. Prime Archives. Microbiology 2020. , https://doi.org/10.37247/pamic.1.2020.22.
[76] Bolourian A, Mojtahedi Z. Immunosuppressants produced by Streptomyces: evolution, hygiene hypothesis, tumour rapalog resistance and probiotics. Environ Microbiol Rep 2018;10:123–6.
[77] Singh BR, McLaughlin MJ, Brevik E. The Nexus of Soils, Plants. Catena Soil Sciences: Animals and Human Health; 2017.
[78] Hirt H. Healthy soils for healthy plants for healthy humans: How beneficial microbes in the soil, food and gut are interconnected and how agriculture can contribute to human health. EMBO Rep 2020:e51069.
[79] Sester M, van Leth F, Bruchfeld J, Bumbacea D, Cirillo DM, Dilektasli AG, et al. Risk assessment of tuberculosis in immunocompromised patients. A TBNET study. Am J Respir Crit Care Med 2014;190:1168–76.
[80] Kmeid J, Kulkarni PA, Batista MV, El Chaer F, Prayag A, Ariza-Heredia EJ, et al. Active Mycobacterium tuberculosis infection at a comprehensive cancer center, 2006–2014. BMC Infect Dis 2019;19:934.
[81] Winglee K, Eloe-Fadrosh E, Gupta S, Guo H, Fraser C, Bishai W. Aerosol Mycobacterium tuberculosis infection causes rapid loss of diversity in gut microbiota. PLoS ONE 2014;9:e97048.
[82] Hu Y, Feng Y, Wu J, Liu F, Zhang Z, Hao Y, et al. The Gut Microbiome Signatures Discriminate Healthy From Pulmonary Tuberculosis Patients. Front Cell Infect Microbiol 2019;9:90.
[83] 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.