BMC Microbiology

RESEARCH ARTICLE Open Access

Bacterial community structure alterations within the colorectal cancer gut microbiome

Mark Loftus[1], Sayf Al-Deen Hassouneh[1] and Shibu Yooseph[2*]

Abstract

Background: Colorectal cancer is a leading cause of cancer-related deaths worldwide. The human gut microbiome has become an active area of research for understanding the initiation, progression, and treatment of colorectal cancer. Despite multiple studies having found significant alterations in the carriage of specific bacteria within the gut microbiome of colorectal cancer patients, no single bacterium has been unequivocally connected to all cases. Whether alterations in species carriages are the cause or outcome of cancer formation is still unclear, but what is clear is that focus should be placed on understanding changes to the bacterial community structure within the cancer-associated gut microbiome.

Results: By applying a novel set of analyses on 252 previously published whole-genome shotgun sequenced fecal samples from healthy and late-stage colorectal cancer subjects, we identify taxonomic, functional, and structural changes within the cancer-associated human gut microbiome. Bacterial association networks constructed from these data exhibited widespread differences in the underlying bacterial community structure between healthy and colorectal cancer associated gut microbiomes. Within the cancer-associated ecosystem, bacterial species were found to form associations with other species that are taxonomically and functionally dissimilar to themselves, as well as form modules functionally geared towards potential changes in the tumor-associated ecosystem. Bacterial community profiling of these samples revealed a significant increase in species diversity within the cancer-associated gut microbiome, and an elevated relative abundance of species classified as originating from the oral microbiome including, but not limited to, *Fusobacterium nucleatum*, *Peptostreptococcus stomatis*, *Gemella morbillorum*, and *Parvimonas micra*. Differential abundance analyses of community functional capabilities revealed an elevation in functions linked to virulence factors and peptide degradation, and a reduction in functions involved in amino-acid biosynthesis within the colorectal cancer gut microbiome.

Conclusions: We utilize whole-genome shotgun sequenced fecal samples provided from a large cohort of late-stage colorectal cancer and healthy subjects to identify a number of potentially important taxonomic, functional, and structural alterations occurring within the colorectal cancer associated gut microbiome. Our analyses indicate that the cancer-associated ecosystem influences bacterial partner selection in the native microbiota, and we highlight specific oral bacteria and their associations as potentially relevant towards aiding tumor progression.

Keywords: Colorectal, Cancer, Microbiome, Metagenomics, Networks, Oral, Pathogens, Associations

* Correspondence: Shibu.Yooseph@ucf.edu
[2]Department of Computer Science, Genomics and Bioinformatics Cluster, University of Central Florida, Orlando, FL 32816, USA
Full list of author information is available at the end of the article

Background

The human gastrointestinal tract harbors a highly diverse community of bacterial cells thought to be in comparable abundance to those of its human host making it the largest and most complex community of bacteria found associating with the human body [1]. These bacteria are typically regarded as commensal, or symbiotic, in that they generally cause no harm and provide fundamental services for their host's nutrition and continued health. The most important of these services include the creation of metabolic by-products (short chain fatty acids, hormones, vitamins, etc.), aiding in proper intestinal tissue and immune system development and regulation, and protecting the gut from colonization by pathogenic organisms [2, 3]. Many diseases have been associated with the disruption of the gut microbiome's bacterial community, one of which is colorectal cancer (CRC) [4–7].

CRC is one of the leading causes of cancer-related deaths worldwide [8] and is characterized by the uncontrolled growth of epithelial cells within the colon or rectum. The transformation of epithelial cells from noncancerous to cancerous growth commonly begins with the formation of a polyp, which over a 10-to-20-year period may or may not progress to become an invasive cancer [9]. CRC initiation is understood as being the result of a combination of both genetic and environmental factors (diet, smoking, alcohol, etc.) [10–12], although the majority (around 75%) of CRC cases are spontaneous, with genetic risk factors being attributed to less than 10% of cases [13, 14]. Recently, there has been a surge in evidence supporting the hypothesis that the human gut microbiome plays a prominent role in relation to cancer initiation, progression, and in the efficacy of its treatment [7, 15–20]. One of the leading hypotheses is the "driver-passenger" model [17], which postulates that a "driver" bacterium such as *Fusobacterium nucleatum*, *Bacteroides fragilis*, or *Escherichia coli* promotes genomic instability (damage) to the DNA of epithelial cells, potentially through some virulence factor, which leads to cellular mutation and eventually tumor formation. Following tumor formation, the changes in micro-environmental conditions around the tumor mass (tumor microenvironment; TME) would optimize the growth of "passenger" microbes who are better suited to this niche facilitating their colonization, and eventual out-competing of the "driver" species as well as the native microbiota leading to a depletion in protective commensal species. These "passenger" microbes could either be pathogens that exist normally in the healthy gut microbiome in low abundance, or simply commensal bacteria that have acquired pathogenic characteristics due to the alteration in the local intestinal ecology. As of now, there is no consistent cancer-associated community profile that has been observed leaving researchers with limited understanding of the full extent the microbiota plays in CRC. Nevertheless, the modulation of the bacterial community within the cancer-associated gut microbiome is the next logical step in possible CRC treatment and prevention strategies.

To one day utilize the bacterial community toward these purposes, it is important to know more than which species are present or absent in the community during disease. We also need to understand how the associations between bacterial species have been affected. These associations are shaped by both direct and indirect interactions taking place in the community (e.g., cooperation or competition), and are important as they are the bedrock upon which the community services, as well as the structure and function, are founded on [21, 22]. In this study, we represent these associations using a weighted graph (network) in which a node denotes a bacterial species and a weighted edge between two nodes represents the strength of the association between the corresponding species. By using this framework, we can model the positive and negative associations between species, thereby shedding light on how cooperation and competition shape the structure of the bacterial community. Bacterial association networks are constructed from sample-taxa count matrices. A sample-taxa count matrix is commonly generated by sequencing the collected biological samples and determining the taxa (species) counts in each sample. However, DNA sequencing does not provide the absolute counts of these taxa within a sample, and instead provides only their relative abundances (i.e., compositional data) [23]. Due to this aspect, inferring associations between species is challenging, and using measures like correlation can produce misleading results when applied directly to compositional data [24]. With this limitation in mind, we applied a Gaussian Graphical Model (GGM) framework on Centered Log-Ratio (CLR) transformed sequence count data to model the conditional dependencies between species to construct association networks [25]. Prior studies that investigated the associations between bacteria within the CRC-associated gut microbiome have either not dealt appropriately with compositional data (for instance, application of correlation directly to untransformed data), or have utilized low taxonomic resolution data (16S rRNA data) which should be used cautiously to assign taxonomic classifications beneath genus-level [5, 26–30]. For the analysis presented here, we utilize 252 whole-genome shotgun (WGS) sequenced fecal samples provided by healthy and late-stage (stage III and IV) CRC subjects from a previously published study [31] to investigate bacterial associations at the species level [32]. The authors of that study originally performed metagenomic and metabolomic analyses to assess any taxonomic and functional differences of the gut microbiota, and metabolites, as well as find diagnostic markers for CRC. For their analyses, these researchers only focused on finding alterations of the microbiota pertaining to species currently known to be culturable and constructed bacterial association networks using correlation (Spearman's) at the genus-level. Our analysis framework and goals are different. For our study, we used a comprehensive collection of nearly eleven thousand bacterial strain reference genomes from NCBI's RefSeq database to calculate the genome relative abundance of bacterial species in each sample using an Expectation-Maximization (EM) algorithm. Subsequently, species were selected based on their prevalence, relative abundance, and feature importance, and were used to construct bacterial association networks using the graphical lasso (glasso) approach [33]. These networks were then analyzed to assess the differences in bacterial community structure between the healthy and late-stage CRC-associated gut microbiome. Taxonomic and functional analysis was performed to highlight differences in gut microbiome bacterial community functional capabilities and species carriages. Our results not only identify both individual and groups (modules) of species potentially capable of aiding tumor progression, but also shows how the bacterial community structure has dramatically altered in response to potential ecological changes occurring within the CRC-associated gut microbiome.

Results

Bacterial community taxonomic profiling

Following sample pre-processing (see methods), we computed the relative abundance of species within each sample using an EM-based method in order to construct a sample-taxa matrix (see methods). This sample-taxa matrix was then used to investigate the bacterial community diversity in the two sample groups (Healthy and late-stage CRC) by measuring the bacterial richness and Shannon index of each sample. Samples originating from the CRC group exhibited significantly greater diversity, both richness and Shannon index, (Mann-Whitney U test: MWU); Richness: MWU pvalue = 0.0005 and Shannon Index: MWU pvalue = 0.0009) compared to those of the Healthy group (Fig. 1a,b). Considering that measures of species diversity differ in their sensitivity to species evenness and richness [34], we additionally applied the Simpson index of diversity to compute species diversity within sample groups (see supplemental file). These results were congruent with our previous analyses showing a statistically significant (MWU: pvalue = 0.0238) higher species diversity in CRC samples compared to that found in Healthy samples (Supplemental 1). We next assessed the differences in bacterial community taxonomic profiles between the healthy and late-stage CRC-associated gut microbiomes. Prior to performing further analyses we applied a CLR-transformation to our sample-taxa matrix (see methods). Taxonomic profile-based sample ordination was carried out using Principal Components Analysis (PCA). The first two principal components explain only a small fraction of the total variance (PC1: 7.98%, PC2: 5.61%) (Supplemental 2), and the linear transformation based on PCA did not show evidence for separation of Healthy samples from CRC samples. However, we were able to distinguish between the two sample groups using a Random Forest Classifier (RFC) (AUC = 0.87) (Fig. 2a). While RFCs rank features (species) based on their importance, these default measures of importance are known to be biased and lead to the return of suboptimal predictor features [35]. To obtain statistical significance for

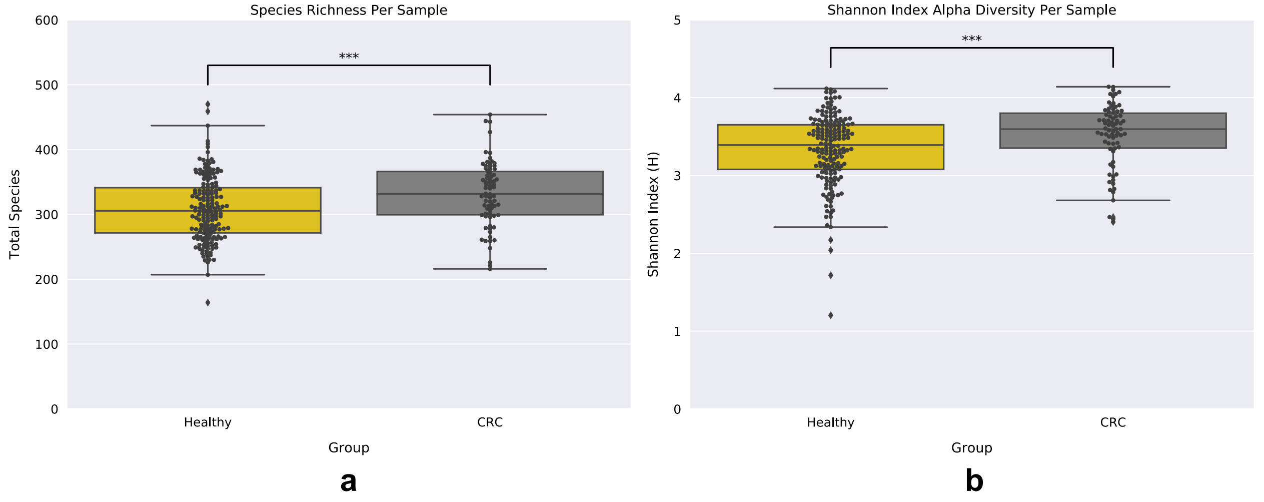

Fig. 1 Species Diversity Within Healthy and CRC Samples. **a**. Boxplot of sample species richness (total species) showing significantly greater species richness within the CRC sample group. **b**. Boxplot of sample Shannon diversity shows significantly greater species diversity within CRC sample group. Black dots represent individual samples and stars (***) denote statistical significance (MWU pvalue < 0.001)

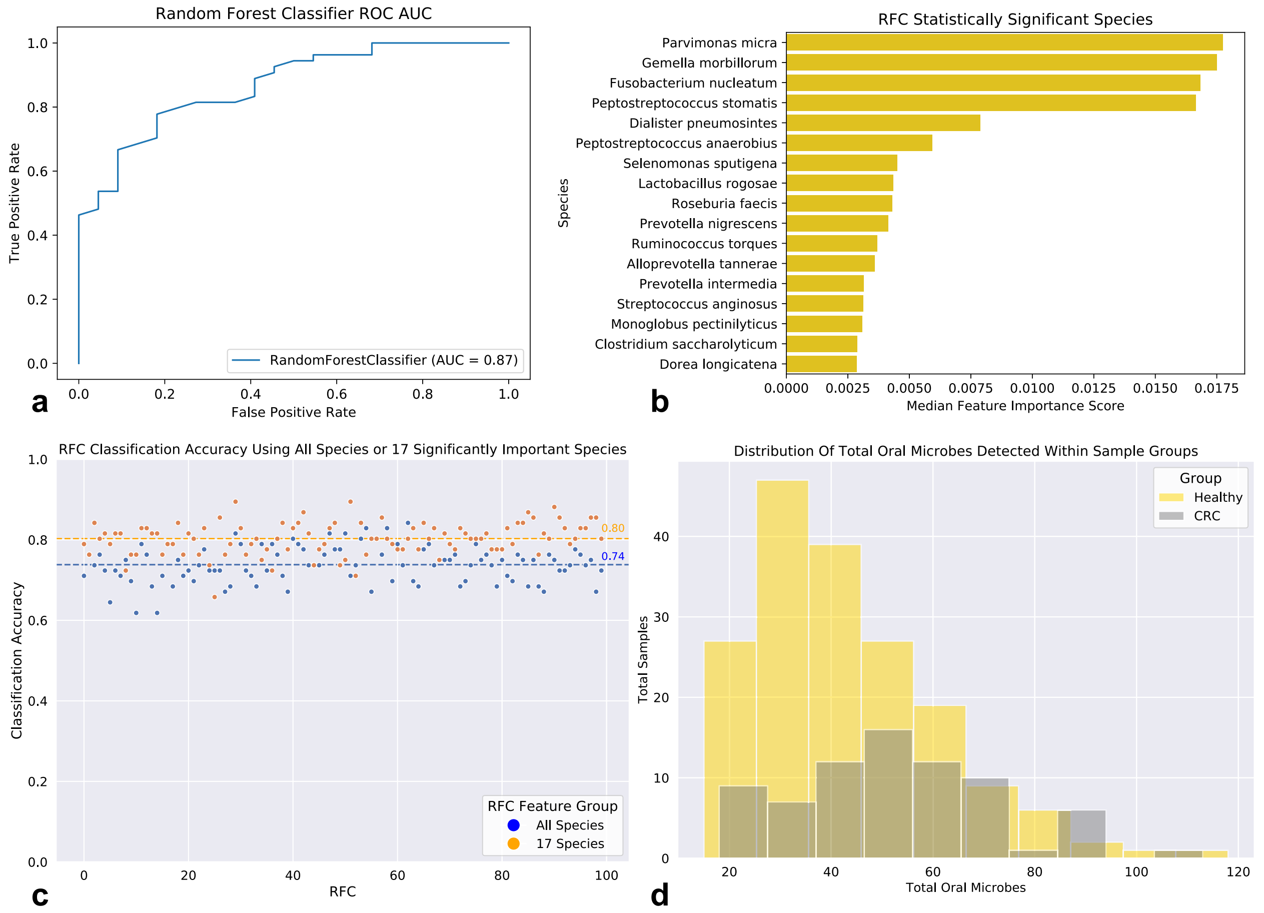

Fig. 2 Healthy and CRC Taxonomic Profiling. **a**. Random Forest Classifier (RFC) ROC showing an AUC = 0.87. **b**. Seventeen statistically significant (MWU-FDR qvalue< 0.05) 'important' species from 100 RFCs compared to a random feature. **c**. Classification accuracy of 100 RFCs using either all species or the 17 significantly 'important' species. **d**. Distribution of total oral microbes within Healthy and CRC sample groups. A significantly (MWU pvalue< 0.05) greater total population of oral microbes were found in the CRC-associated gut microbiome. Bacterial species were classified as oral microbes by using the expanded Human Oral Microbe Database (eHOMD)

species importances provided by the RFC we applied a technique where we included a "random" feature into our feature set (see methods). By using an ensemble of 100 RFCs we uncovered 17 bacterial species that were statistically (MWU and False Discovery Rate Multiple Testing Correction; MWU-FDR: qvalue< 0.05) more 'important' (deemed significantly 'important') than the random feature for distinguishing groups (Fig. 2b). We found that the accuracy classification score of 100 RFCs trained on the 17 significantly 'important' species was on average greater than that of the 100 RFCs trained on all species (All Species Mean Accuracy: 74%; 17 significantly 'important' Species Mean Accuracy: 80%) (Fig. 2c). We next performed species differential abundance analysis (see methods) which revealed 174 species significantly (MWU-FDR qvalue< 0.05) reduced in relative abundance, and 10 species significantly elevated in relative abundance within the CRC-associated gut microbiome compared to the Healthy gut microbiome. These 174 bacterial species are from a diverse background of 84 genera, although the largest fraction of species were from the genera *Enterobacter* (6.8%), *Klebsiella* (6.3%), *Streptococcus* (5.2%), *Lactobacillus* (5.1%), *Citrobacter* (4.6%), *Bifidobacterium* (4%), *Bacteroides* (3.4%), and *Clostridium* (3.4%) (Supplemental 3). The 10 species significantly elevated in relative abundance within CRC were: *Parvimonas micra* (qvalue = 3.09e-09), *Peptostreptococcus stomatis* (qvalue = 4.51e-08), *Gemella morbillorum* (qvalue = 4.55e-08), *Fusobacterium nucleatum* (qvalue = 1.08e-06), *Streptococcus anginosus* (qvalue = 1.13e-03), *Dialister pneumosintes* (qvalue = 1.37e-03), *Peptostreptococcus anaerobius* (qvalue = 4.74e-03), *Streptococcus sp. KCOM 2412* (*Streptococcus periodonticum*) (qvalue = 7.18e-03), *Ruminococcus torques* (qvalue = 1.55e-02), and *Filifactor alocis* (qvalue = 2.85e-02) (Supplemental 4a-c). Interestingly, many of the species that were deemed both significantly 'important' and elevated in relative abundance within CRC are also found in the oral microbiome and noted to be associated with oral diseases (periodontitis, periapical lesions, root canal infections, oral cancers, etc.) which have been associated with increased risks of CRC [36–45]. Subsequently, we utilized the expanded Human Oral Microbe Database (eHOMD) [46] to classify all oral species within our samples and found a significant increase in the total oral microbe population richness existing within the CRC-associated gut microbiome in comparison to that of the Healthy group (MWU: pvalue = 6.51e-05) (Fig. 2d).

Bacterial community functional profiling

To analyze the differences in community functional capabilities between the Healthy and CRC gut microbiomes we measured the relative abundance of protein families (TIGRFAMs [47]) and protein domains (Pfams [48]) within our WGS samples creating a sample-function matrix (see methods). A CLR-transformation was applied to this matrix and then PCA was performed. PCA showed evidence of inter-group clustering of samples (Healthy and CRC) and ultimately only explained a moderate variance (PC1: 27.19%, PC2: 4.33%) (Supplemental 5). Differential abundance analysis was performed using the CLR-transformed sample-function matrix which showed 12 Pfams (7 elevated and 5 reduced in CRC compared to Healthy) and two TIGR FAMs (1 elevated and 1 reduced in CRC compared to Healthy) to be statistically significantly (MWU-FDR: qvalue< 0.05) different in their relative abundance (Supplemental Table 1). Pfams that were significantly elevated within the CRC gut microbiome were linked to bacterial invasins and adhesins (ex: FadA), while those that were significantly reduced were tied to antibiotic resistance, bacteriophage maturation, and threonine biosynthesis. The single TIGRFAM significantly elevated in CRC was linked to proline iminopeptidase, while the only TIGRFAM significantly reduced was again linked to threonine biosynthesis.

Bacterial association networks

Species chosen for network construction were selected based on their prevalence, abundance, and 'importance'. First, the prevalence of each species was calculated across all samples within each group (Fig. 3a). The distributions of bacterial species prevalence counts within groups were found to exhibit a bi-modal distribution with one peak occurring at the 90% prevalence threshold. Going forward we refer to the species found above 90% sample prevalence within groups as the highly prevalent species (HPS). A large majority of species within each group's HPS were found to be shared (Healthy: 97% and CRC: 95%) (Fig. 3b). The five unique HPS in the Healthy group were: *Hespellia stercorisuis*, *Clostridium saccharolyticum*, *Monoglobus pectinilyticus*, *Streptococcus sp. oral taxon 431*, and *Odoribacter laneus*. The eight HPS unique to the CRC associated group were: *Intestinibacillus massiliensis*, *Prevotella copri*, *Haemophilus parainfluenzae*, *Ruminococcus bicirculans*, *Streptococcus mitis*, *Neglecta timonensis*, *Bifidobacterium catenulatum*, and *Anaerotignum neopropionicum*. Interestingly, *Streptococcus mitis* and *Haemophilus parainfluenzae* are both classified by the eHOMD as oral microbes. The relative abundances of HPS were found to account for the majority (Median = 82%) of a sample's total relative abundance (Fig. 3c). Moving forward we utilized the union of HPS within groups for network construction. In addition to these highly prevalent and abundant species we wanted to incorporate the species who were both deemed significantly 'important' by our RFCs and found in differential abundance. This led to the addition of 8 species (*Parvimonas micra*, *Peptostreptococcus stomatis*, *Gemella morbillorum*, *Fusobacterium nucleatum*, *Streptococcus anginosus*, *Dialister pneumosintes*, *Peptostreptococcus anaerobius*, and *Ruminococcus torques*) to our species group (165 species total) used in network construction. Bacterial association networks were then constructed from the CLR-transformed

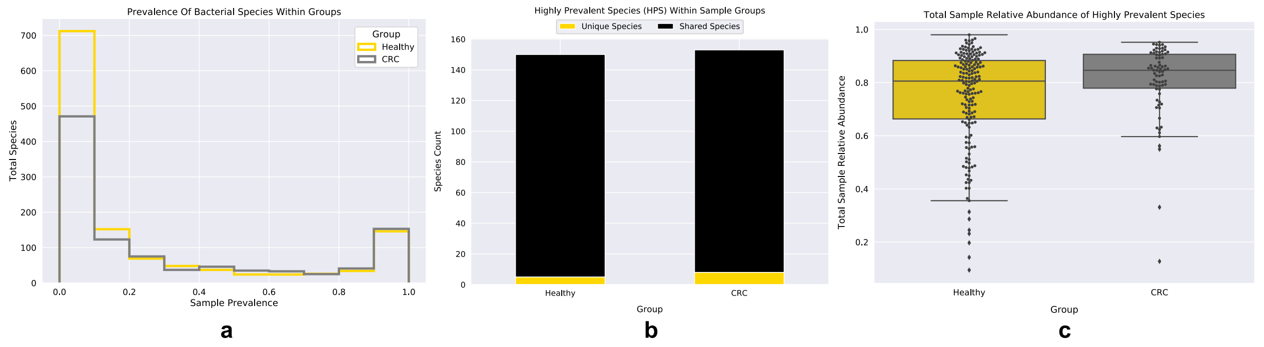

Fig. 3 Highly Prevalent and Abundant Species within Groups. **a**. Bi-modal distribution of species prevalence counts across samples within the Healthy and CRC sample groups. **b**. Stacked-bar plot showing the total unique and shared species of the highly prevalent (> 90% prevalence) species within the Healthy and CRC sample groups. **c**. Boxplot of the total sample relative abundance accounted for by the highly prevalent species within groups. Black dots represent individual samples

relative abundance of these selected species (see methods and supplemental information for additional information). Following network construction, we first checked our networks for non-randomness by comparing multiple network properties (average shortest path length, transitivity, and modularity) to those displayed from random networks (see methods). Compared to random networks, the Healthy and CRC networks both exhibited statistically significant (Monte Carlo Simulation; MCS) shorter average shortest path lengths (ASPL) (Healthy and CRC: MCS *p*value< 0.001), higher transitivity (Healthy and CRC: MCS *p*value< 0.001), and higher modularity (Healthy and CRC: MCS *p*value< 0.001) (Table 1). These results indicate that networks constructed displayed properties that were significantly non-random, and that species within networks: are connected to one another through short paths, have positive associations with the neighbors of their neighbors (friends of friends), and form modules (i.e. a group or cluster of species) that are characterized by the majority of associations occurring between species within the same module, and few associations existing with species outside the module.

Group networks contained similar distributions of association weights with positive associations being in greater abundance than negative associations (Fig. 4a). Notably, the CRC network contained a greater total of negative associations compared to that found in the Healthy network. Interestingly, 29% of these negative associations involved a species deemed as an oral microbe, whereas within the Healthy network zero negative associations were found to involve oral microbes. Surprisingly, the majority of associations found within networks were unique to that network (Healthy: 69%, CRC: 72%) (Fig. 4b). We hypothesized that this dramatic difference in community structure could reflect changes in the ecosystem and proceeded to analyze the taxonomic relationship between species within networks (see methods) (Fig. 4c). Both networks exhibited significantly (MCS pvalue< 0.05) more positive relationships between species within the same genera (Healthy: MCS pvalue = 0.00099, CRC: MCS pvalue = 0.00099) and family (Healthy: MCS pvalue = 0.00099, CRC: MCS *p*value = 0.00099) compared to those found in a random network (see methods). However, only within the Healthy network did species still have significantly more positive associations with other species from the same order more so than random (Healthy: MCS *p*value = 0.00099, CRC: MCS pvalue = 0.44). The CRC network also exhibited a larger abundance in taxonomically distant (outside phylum) relationships compared to the Healthy network (Healthy: 4%, CRC: 17%), although positive associations between taxonomically distant microbes were still significantly less in Healthy (Within Phylum: MCS pvalue = 0.00099, Outside Phylum: MCS *p*value = 0.00099) and CRC (Within Phylum: MCS pvalue = 0.00099, Outside Phylum: MCS *p*value = 0.00099) than random networks. We next examined the dissimilarity between functional profiles of associating species within the Healthy and CRC networks (Supplemental 6a-c). Interestingly, many of the bacterial associations that are unique to the CRC network were shown to be occurring between species that were functionally dissimilar to one another.

Considering that our networks exhibited high modularity, and that community functions in microbial environments are driven through polymicrobial synergy [49, 50], we applied a module detection algorithm to our networks, and proceeded to analyze the obtained species modules within our networks (see methods). We first started by comparing the potential functional capabilities of modules by constructing CLR-transformed module functional profiles (see methods). PCA of module functional (protein domain) profiles exhibited large variance (PC1: 33.73%, PC2: 14.53%), and modules appeared to form clusters which contained representation from both groups (Fig. 5a). To define clusters of modules, silhouette analysis was performed which estimated five clusters as the optimal K to use for K-means clustering (Fig. 5b). After module clusters were defined by K-means clustering (Fig. 5c), taxonomic analysis of these clusters was carried out. Across networks, modules that fell within the same cluster were found to be taxonomically similar, excluding cluster 1 and cluster 5 which exhibited a shift in species occupancy where some species found within cluster 1 in the Healthy network were shown to be within cluster 5 in the CRC network, and vice-versa (Supplemental 7a,b,c,g). However, both networks had strong agreement on the species found within clusters 2, 3, and 4. Species within cluster 2 were *only* 'pathobiont' (i.e., species that are generally not harmful but contain the capacity to cause disease under particular environmental conditions [51, 52]) oral microbes (Supplemental 7d), whereas cluster 3 was mainly Streptococcus species (Supplemental 7e), and cluster 4 predominantly Bacteroides species (Supplemental 7f). Subsequently, cluster functional analysis was performed to find protein

Table 1 Group Network Properties Compared to Random Networks

Network	Nodes	Edges	Density	ASPL	Transitivity	Modularity
CRC	165	324	0.024	*** 1.687	*** 0.379	*** 0.689
Healthy	165	292	0.022	*** 1.554	*** 0.453	*** 0.742

Network properties of Healthy and CRC networks. Both Healthy and CRC networks were found to exhibit significantly shorter Average Shortest Path Lengths (ASPL), higher Transitivity, and higher Modularity then 1000 random networks. Stars (***) denote statistical significance (Monte Carlo simulation pvalue < 0.001)

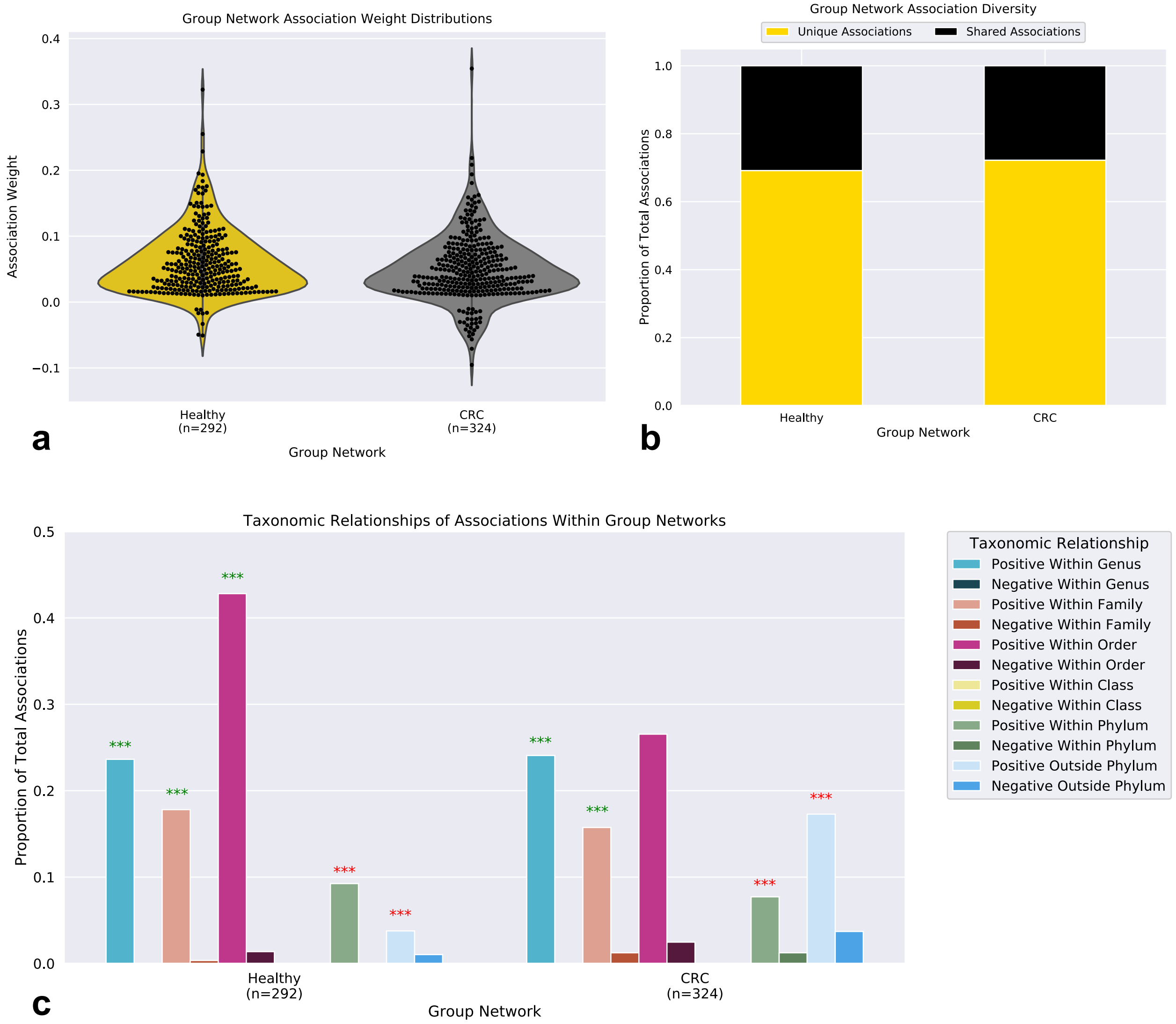

Fig. 4 Group Network Associations. **a**. Distribution of bacterial association weights within Healthy and CRC networks. **b**. Stacked-bar plot of the proportion of associations (edges) that are unique and shared between Healthy and CRC networks. **c**. Bar-plots representing the proportion of total associations within the lowest common taxonomic relationship between bacterial species. Stars (***) indicate statistical significance (Monte Carlo simulation pvalue < 0.001), star color (green or red) indicate higher or lower than that found in random networks, respectively

domains, as well as the main roles and sub roles of protein families, which made clusters functionally ‘distinct’ from one another (see methods) (Supplemental 8a-c). Functional capabilities (protein domains and protein family main/sub roles) distinguishing cluster 1 were linked to: cell surface adhesion, counter-conflict strategies, tyrosine recombinases, degradation of polysaccharides, glycosaminoglycan binding, tumor protease inhibition, peroxidase functions, carbohydrate/cellulose binding activities, and amino acid biosynthesis. Cluster 2’s distinguishing functions were linked to: adherence to host cells and extracellular matrix, cellular infection, collagen binding, complement resistance, ornithine/lysine/arginine decarboxylase (tissue putrefaction/polyamine synthesis/acidic environment resistance), metallopeptidases, type V secretion systems, ammonia production, and excretion of poisonous metal ions (copper efflux system), cell envelope, DNA metabolism, fatty acid and phospholipid metabolism, biosynthesis and degradation of surface polysaccharides and lipopolysaccharides. Cluster 3’s distinguishing functions were linked to: mucin binding, zinc scavenging/uptake, cell-surface adhesion, glucose binding/transport, and copper binding, protein and peptide fate/synthesis/secretion, degradation of polysaccharides/carbohydrates, organic alcohols, and acids. Cluster 4’s distinguishing functions were linked to: metal binding, diguanylate cyclase/phosphodiesterase, quorum sensing, carbohydrate-binding, and cysteine/papain proteases, nucleosides and nucleotides,

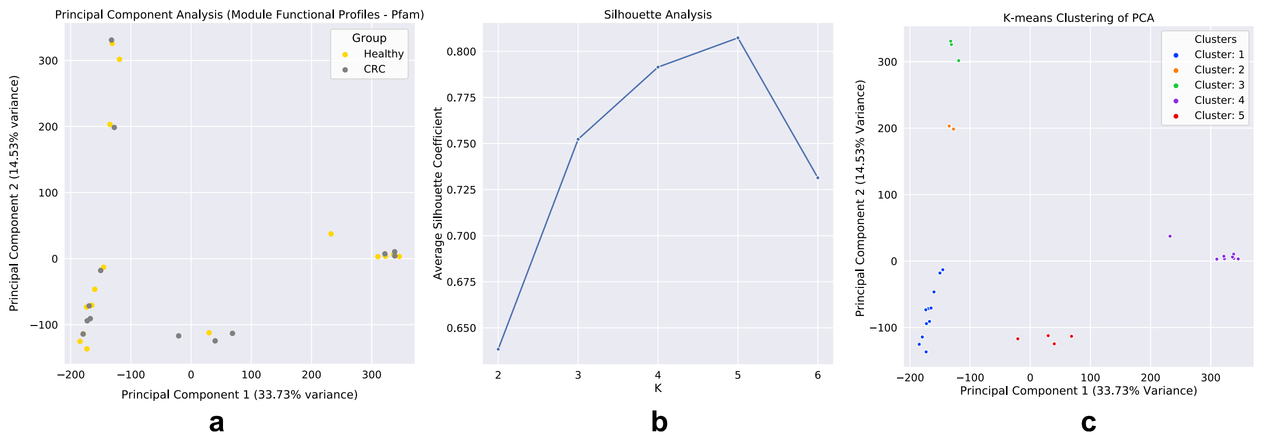

Fig. 5 Species Module Functional Clusters Within Networks. **a**. Principal component analysis of module functional profiles. Gold and grey dots represent individual modules from Healthy or CRC networks, respectively. **b.** Silhouette analysis showing K = 5 having the highest average silhouette coefficient. **c**. K-means clustering of the module functional profile PCA using K = 5

transport and binding proteins, TCA cycle, iron carrying, and the degradation and biosynthesis of surface polysaccharides. Lastly, cluster 5's distinguishing functions were linked to: aminopeptidases, tripartite tricarboxylate receptors, ethanolamine transportation, starch utilization, and xyloglucan/polysaccharide binding, energy metabolism, amino acids and amines, cation and iron compounds, electron transport, and the biosynthesis and degradation of surface polysaccharides and lipopolysaccharides. The abundance of species utilized for network construction found within each cluster was examined (Healthy: cluster 1 (33%), cluster 2 (2%), cluster 3 (5%), cluster 4 (26%), cluster 5 (7%), no cluster (27%); CRC: cluster 1 (19%), cluster 2 (3%), cluster 3 (3%), cluster 4 (12%), cluster 5 (30%), no cluster (33%)) (Fig. 6). Our findings showed that within the CRC network there was an increase in the total species found within a module of cluster type 2 and 5 and a reduction of species in cluster type 1, 3, and 4 compared to the Healthy. These results are also reflected in our findings of a statistically significant change in the total sample relative abundance that species within clusters accounted for between groups (Cluster 1: MWU pvalue = 4.29e-12; Cluster 2: MWU

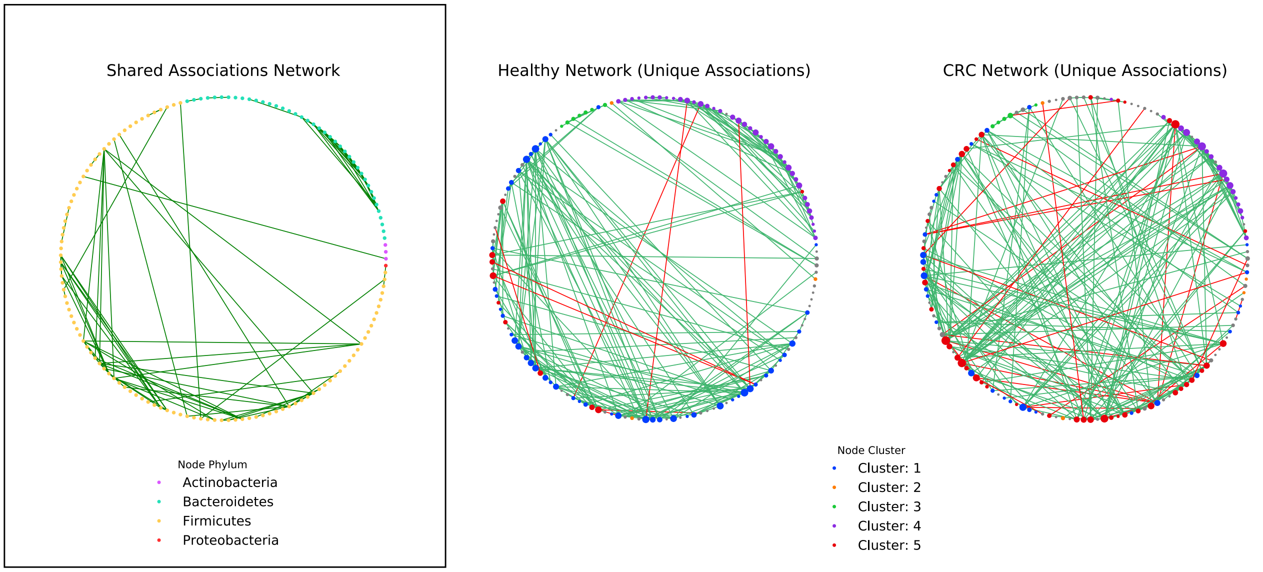

Fig. 6 Healthy and CRC Bacterial Association Networks. Bacterial association networks presented in a circular layout. Edge color (green or red) represent positive or negative associations, respectively. Far left network (Shared Associations Network) shows the associations (edges) found in both the Healthy and CRC network. Node color within that network represents the phylum of the species. The two networks on the right are displaying the associations unique only to the Healthy or CRC network. Node color within these networks represent the module cluster this species was found within. Node size is a function of the node's degree (total associations). For a list of species shown and not shown within networks see supplemental

pvalue = 3.16e-16; Cluster 3: MWU pvalue = 0.0002; Cluster 4: MWU pvalue = 2.62e-13; Cluster 5: MWU *p*value = 2.81e-29; No Cluster Species: MWU pvalue = 4.40e-17) (Fig. 7). Moreover, the majority of negative associations within networks (Healthy: 100%, CRC: 96%) were found to occur between species that occupy modules within different cluster types (Supplemental 9). Interestingly, only within the CRC network did an intra-cluster negative association arise between species of cluster 1 where a reduction in species membership and abundance was also exhibited.

Influential bacterial species within networks

Finally, we examined which species potentially have the greatest influence on the structure of our networks, and therefore possibly within the ecosystem as well, by identifying 'Hub' nodes. 'Hub' nodes are species with many associations that serve as a central point of connection between many other species [53, 54]. Most modules within networks (Healthy: 84.6%; CRC: 87.5%) were found to be disassortative with respect to node degree (Supplemental 10) suggesting that 'Hub' species existed within these modules [53]. We proceeded to identify 'Hub' species by selecting the species with the largest degree centrality within all modules exhibiting a degree assortativity below zero (see methods). In total, 22 unique 'Hub' species were identified, and of these 'Hubs' only two, *Bacteroides fluxus and Bacteroides pectinophilus,* were shared between Healthy and CRC networks. We noted that *Bacteroides fluxus* and *Bacteroides pectinophilus* also maintained their position as 'Hubs' within the same module cluster type (Cluster 4 and Cluster 1, respectively) across networks (Supplementary 11a,b). Interestingly, only within the CRC network were oral microbes, *Peptostreptococcus stomatis* and *Streptococcus parasanguinis*, designated as 'Hub' nodes. The module *Peptostreptococcus stomatis* is a 'Hub' within is particularly fascinating as it is the only CRC cluster 2, 'pathobiont' cluster, module where all species are both oral microbes (*Gemella morbillorum*, *Parvimonas micra*, and *Dialister pneumosintes*) and found to be significantly elevated in relative abundance. Moreover, *Anaerotruncus colihominis*, a 'Hub' species only within the Healthy network, was found to be negatively associated with *Gemella morbillorum* within this module in the CRC network (Supplementary 12).

Discussion

In this study, WGS data available from healthy and late-stage colorectal cancer subjects were utilized in conjunction with community profiling and network inference techniques to better understand the alterations in bacterial community ecology that have occurred within the late-stage cancer-associated human gut microbiome. Our study uncovered key distinctions in both the bacterial species and genomic functional capabilities that were different between the two communities, suggesting an overgrowth of potentially pathogenic species classified as oral microbes. We also observed a dramatic difference in bacterial community structure which we believe to be due to an alteration in bacteria partner selection in response to probable ecosystem changes occurring within the CRC-associated gut microbiome.

Our study showed that the CRC gut microbiome contained a significantly higher bacterial diversity. This higher diversity was *somewhat* unexpected since a high bacterial diversity is regularly associated with the healthy gut microbiome [55], and previous studies have described a lower diversity within the CRC gut microbiome [4, 26], although, these findings are still in contention as other studies have also found a higher bacterial richness [56]. In either case, this discrepancy in species diversity estimations between studies could be due to differences in the sequence data type (amplicon vs shotgun) used as 16S rRNA data is known to highly skew estimates of bacterial diversity [57]. We hypothesized that this higher species diversity was due to the formation (or expansion) of a bacterial niche in the cancer-associated ecosystem, most likely caused by the presence of the tumor mass. Any bacterial species existing closely to, or within, the tumor microenvironment (TME) niche would be exposed to a hostile environment characterized by low oxygen, high acidity, and an abundance of oxidative stressors [58, 59]. These environmental conditions are in part created by the altered metabolism of tumor cells which would lead to the reduction in the typical proteins, carbohydrates, and lipids available (nutrient scarcity) in the surrounding microenvironment [60–63]. Tumor cells will also scavenge for any additionally needed resources by degrading the extracellular matrix (ECM), and cannibalizing the surrounding necrotic intestinal tissue to fuel their metabolism [64]. These degradation products could provide certain microbiota capable of utilizing them a rich assortment of free resources including amino acids, membrane proteins, phospholipids, and some sugars. As our CRC samples were obtained from late-stage cancer subjects, this TME niche could be widespread across the colon having repercussions for even microbes not involved in the colonization of this niche. Our findings from using machine learning, differential abundance testing, and network inference point towards species capable of filling this niche, functions likely to promote its formation, and the potential impact that the creation of this niche has on the gut microbiota.

Species differential abundance testing between groups found 174 species significantly reduced and 10 species significantly elevated in relative abundance within the

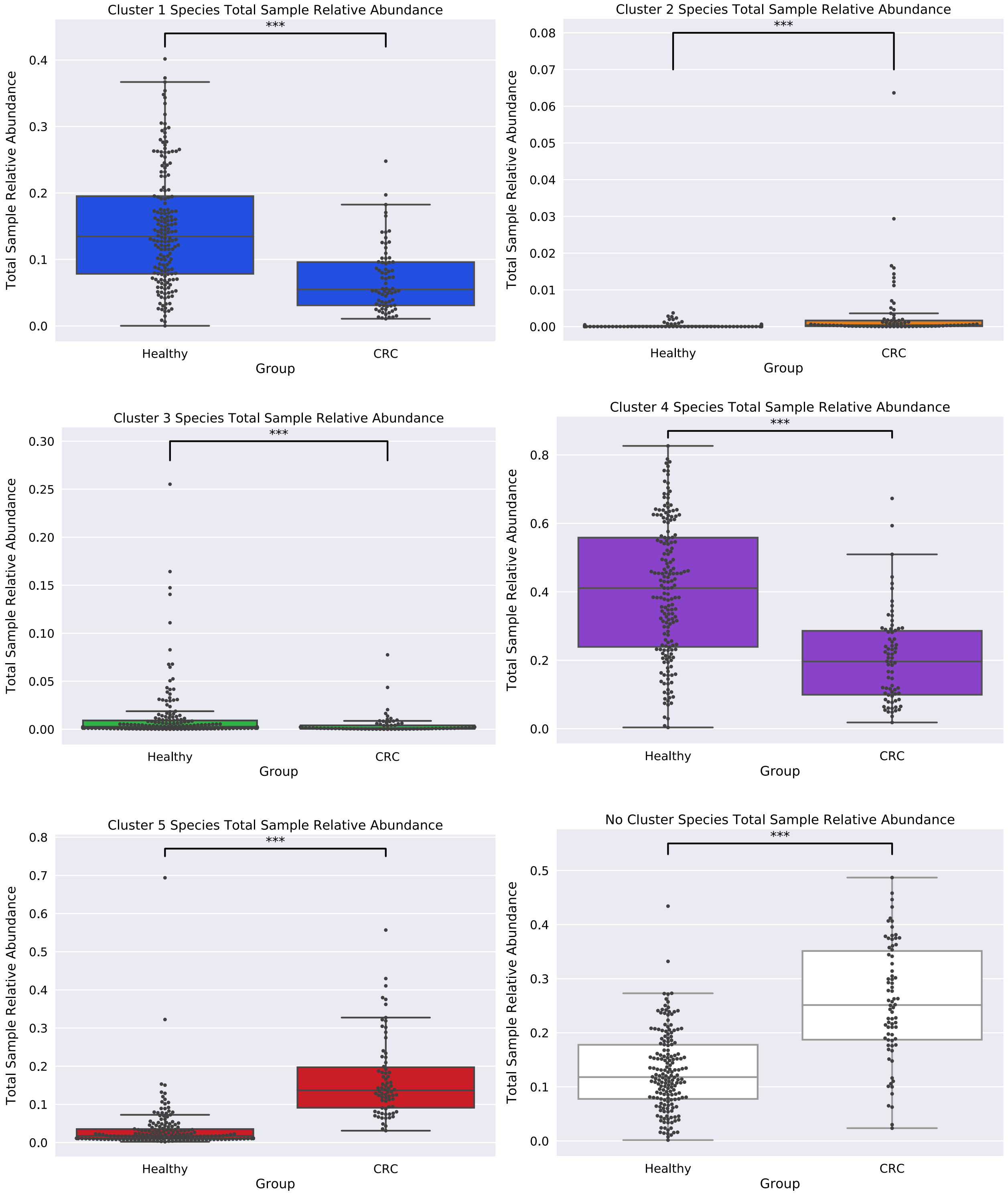

Fig. 7 Cluster Species Total Sample Relative Abundance. Boxplots of the total sample relative abundance that all species within each module cluster account for within groups. The species within module clusters 1, 4, and 3 account for a significantly greater total sample relative abundance within the Healthy network compared to the CRC network. The species within module clusters 2 and 5 and no cluster account for a significantly greater total sample relative abundance within the CRC network compared to the Healthy network. Stars (***) indicate statistical significance (MWU pvalue < 0.001)

CRC-associated gut microbiome compared to the Healthy gut microbiome. Of the 10 species, six (*Parvimonas micra*, *Peptostreptococcus stomatis*, *Gemella morbillorum*, *Fusobacterium nucleatum*, *Streptococcus anginosus*, and *Peptostreptococcus anaerobius*) were previously found elevated in relative abundance by the research study that generated the data analyzed here [31]. However, we additionally found *Dialister pneumosintes*, *Streptococcus sp. KCOM 2412* (*Streptococcus periodonticum*), *Ruminococcus torques*, and *Filifactor alocis* as being significantly elevated in relative abundance within the CRC sample group. This discrepancy in findings is most likely due to differences in both read mapping and species relative abundance calculations. That study mapped reads to the All-Species Living Tree Project (LTP) of the SILVA database [65] assigning taxonomy to the species which provided the lowest E-value, and calculated species relative abundances as the number of reads assigned to the species divided by the total number of aligned reads within the sample. In contrast, we mapped reads to a comprehensive collection of bacterial reference strain genomes downloaded from RefSeq [66], and calculated species relative abundances utilizing an accurate probabilistic framework [67]. To our knowledge, this is the first time *Filifactor alocis* has been shown to have elevated relative abundance within CRC. *Filifactor alocis*, previously known as *Fusobacterium alocis*, is a gram-positive obligate anaerobe that has routinely been discovered in periodontitis and endodontic infections and is described as an excellent marker organism for periodontal disease [40, 68, 69]. Interestingly, all 10 of the species found significantly elevated in relative abundance within CRC were classified as oral microbes, and despite normally existing within the Healthy gut microbiome these species are considered 'pathobionts' as they have numerous associations with infections [37, 70] and even CRC [56, 71–75]. Many of these species also have been previously shown to exist in close association with colonic tumor tissues [72] and possess the capability to colonize the TME niche as they are: anaerobic [76], regularly form biofilms together [39, 77], and exhibit asaccharolytic metabolism [76]. Since oral microbes exhibit an asaccharolytic metabolism they target peptides and amino acids for their digestion [76] and in doing so produce ammonia which would raise the local pH helping their colonization within the acidic TME. In this way, these species would be optimized for growth in the hostile TME niche. Outside of just these 10 oral species, we also uncovered a significantly higher richness of bacteria classified as oral microbes within the CRC gut microbiome. This finding suggests that oral microbes have become increasingly more capable of colonizing the gut within the CRC-associated ecosystem.

Interestingly, of the few bacterial community functions (Pfams and TIGRFAMS) found in differential abundance between the CRC and Healthy gut microbiomes, many could precipitate cancer progression, or aid in the colonization of the TME niche. Multiple protein functions found to be significantly reduced within the CRC gut microbiome were tied to threonine biosynthesis. Threonine is an essential amino acid; therefore, it must be provided exogenously from the gut microbiota's metabolism [78]. It is also an important amino acid in the production of short chain fatty acids (SCFAs) since it can be utilized for the formation of acetate, butyrate, or propionate [79]. Interestingly, of the 174 species found significantly reduced in CRC many are from genera (*Lactobacillus*, *Bacteroides*, *Bifidobacteria*, *Clostridium*, *Eubacterium*, etc.) shown to be linked to the production of SCFAs [80–82]. The reduction in the enzymatic capability to synthesize threonine could drive tumor progression as SCFAs (e.g., butyrate) have been shown to have anti-oncogenic and anti-inflammatory properties [83]. Of the functions found significantly elevated in relative abundance in the CRC gut microbiome many were tied to adhesins and invasins. These protein functions would allow bacteria to adhere to epithelial cells, especially those that are being sloughed off the intestinal wall, to gather nutrients. They would also assist in the invasion of the intestinal barrier which would drive inflammation and could cause DNA damage thereby inducing unwanted cellular mutation. For example, FadA, an adhesin found significantly elevated in relative abundance, is unique to the oral lineage of *Fusobacterium nucleatum's phylum (Fusobacteria)* and has previously been shown to promote binding and invasion into host epithelial cells [84], as well as driving cancer initiation [85, 86]. Additionally, we found a significantly elevated relative abundance of a protein function linked to proline iminopeptidase (PIP), an enzyme that catalyzes the release of proline residues from peptides. Proline is an important stress substrate in cancer metabolism as it is utilized in many critical functions related to apoptosis, autophagy, and nutrient/oxygen deprivation [87]. Tumor cells can harvest the proline they require by metabolizing collagen contained within the extracellular matrix (ECM), as nearly 25% of the collagen is proline [88]. Interestingly, in our study a few of the oral species found significantly elevated in relative abundance (*Peptostreptococcus stomatis*, *Gemella morbillorum*, *Parvimonas micra*, and *Dialister pneumosintes*) were shown to form a network module with the functionally distinct capability to bind and degrade collagen.

Only a few associations were shared between the bacterial association networks for the two sample groups which suggested there was a large difference in the bacterial community structure within Healthy and CRC-associated gut microbiomes. Part of the difference in community structure occurring within the CRC-associated gut microbiome is due to positive associations forming less between species that were taxonomically similar, and more between functionally dissimilar species compared to those found in the Healthy gut microbiome. Moreover, we found a greater number of negative associations within the CRC network, and in many of these negative associations an oral microbe was found to be involved, whereas, in the healthy network no such negative associations with oral microbes were occurring. This suggests that competitive exclusion between taxonomically and functionally similar species within the CRC-associated gut microbiome has increased, and oral microbes have become more competitive within this ecosystem. Additionally, as oral microbes are also found to be present within the Healthy gut microbiome, but negative associations against oral microbes were not, we hypothesized that the native microbiota has shifted towards utilizing similar resources to those targeted by oral microbes within the CRC gut microbiome. Our analysis of species modules within networks reflects this notion. Using PCA and K-means clustering, species modules within networks were found to fall into one of five distinct clusters depending on their functional capabilities. However, both Healthy and CRC networks contained representation (at least one module) within all clusters suggesting the niches that these clusters target are maintained across Healthy and CRC-associated gut microbiomes in some capacity. Yet, despite cluster retention, there was a dramatic shift in both the proportion of total species and the total sample relative abundance certain clusters accounted for within networks. For example, within the Healthy network we found clusters functionally geared towards amino acid biosynthesis, carbohydrate degradation, protein binding/ uptake, and tumor inhibition contained a greater number of species and represented a larger total sample relative abundance. Whereas, in the CRC network we observed a species shift towards forming modules functionally equipped for protein degradation, amino acid uptake, biosynthesis and degradation of surface polysaccharides and lipopolysaccharides, and ethanolamine utilization. Interestingly, *Klebsiella* species have been tied to ethanolamine usage in the healthy gut [89, 90] and were found in reduced relative abundance in the CRC gut microbiome, suggesting that these species were potentially outcompeted. In any case, this shift in species cluster membership and cluster total sample relative abundance suggests that the bacterial community structure has been reorganized to aid in the formation of modules of specific cluster types that contain functional capabilities better suited for life in the CRC-associated gut environment.

As mentioned previously, one module cluster (cluster 2) drew our attention as it was comprised solely of 'pathobiont' oral species and contained distinct functions which would allow these species to not only flourish within the TME niche but aid in cancer progression. These functions included: adherence to host cells and extracellular matrix, collagen-binding, complement resistance, ornithine/lysine/arginine decarboxylase (tissue putrefaction/ polyamine synthesis/acidic environment resistance), metallopeptidases, type V secretion systems, ammonia production, excretion of poisonous metal ions (copper efflux system), DNA metabolism, fatty acid and phospholipid metabolism, and biosynthesis and degradation of surface polysaccharides and lipopolysaccharides. Despite a module of this cluster type existing within the Healthy network, all species existing within the CRC module (*Peptostreptococcus stomatis*, *Gemella morbillorum*, *Parvimonas micra*, and *Dialister pneumosintes*) were found to be significantly elevated in relative abundance. It is also important to note that this module in the CRC network grew with the addition of another oral species, *Dialister pneumosintes*. Which suggests these oral species are indeed thriving in the CRC-associated gut microbiome and through their metabolic actions potentially driving tumor progression. It could be prudent to preemptively target *Peptostreptococcus stomatis* for elimination from the gut microbiome as it was the 'hub' species within the module. By doing so this could lead to the dissipation of the associations between these species and potentially dampen tumor progression. In any case, future in vivo studies should be performed to elucidate the extent that polymicrobial synergy between these species contributes to tumorigenesis.

Conclusion

In summary, our analysis of whole-genome shotgun sequenced fecal samples provided from a large cohort of late-stage colorectal cancer and healthy subjects revealed key differences in the bacterial community within Healthy and CRC-associated gut microbiomes. We showed a higher species diversity exists within the CRC-associated gut microbiome that is potentially due to the formation of a tumor-associated niche, and this niche is most likely occupied by species originating from the oral cavity. Moreover, we highlighted *Peptostreptococcus stomatis* as an influential 'hub' node within a 'pathobiont' oral species module where every species within the module were found in elevated relative abundance in CRC. Our results also indicated that tumor presence influences the reorganization of the native bacterial community structure to aid in the formation of modules that contain functional capabilities better suited for life in the CRC-associated gut environment.

Methods

Data acquisition and cohort description

For this study, 252 whole-genome shotgun sequenced fecal samples were retrieved from DDBJ Sequence Read Archive (DRA) under the bioproject ID PRJDB4176 [31]. The original study population of this cohort consisted of healthy and early/advanced colorectal cancer stage patients who were undergoing total colonoscopy at the National Cancer Center Hospital, Tokyo, Japan. Fecal samples were collected immediately following the first defecation after a bowel-cleansing agent was administered orally. Cancer patients who had or were thought to have hereditary disease, an inflammatory bowel disease, an abdominal surgery history, or whose stool samples were insufficient for data collection were excluded from the original study. Samples chosen to be utilized within this study came from 178 healthy and 74 late-stage (52 stage III / 22 stage IV) colorectal cancer (CRC) subjects. Sample groups had comparable male to female frequencies (Healthy: 56.18/43.82; CRC: 58.11/41.89) (Supplemental 13a) and subject ages (Healthy median age: 62; CRC median age: 61) (Supplemental 13b). For additional information on all samples used in this study see supplemental file.

Data pre-processing

Reads were trimmed with Trimmomatic [91] (version 0.36) utilizing a 4:15 sliding window approach where a read is clipped once the average quality score within a sliding window of 4 base pairs drops below a quality score of 15. Afterwards, reads from human origin were filtered by utilizing Bowtie2 [92] (version 5.4.0, --very-sensitive setting) and the GRCh38.p12 human genome [93].

Species level community taxonomic profiling

For bacterial community taxonomic profiling of WGS reads we elected to utilize a reference-based mapping approach. Sample reads were mapped to a reference database of 10,839 bacterial reference strain genomes downloaded from RefSeq [66] utilizing Bowtie2 (version 5.4.0, settings: --very-sensitive --reorder --mp 1,1 --rfg 1, 1 -k 1000 –score-min L,0,-0.1). In total over 3.5 billion (3,515,063,526) reads were mapped. Next, a probabilistic framework based on a mixture model [67, 94] was used to analyze the read mapping information to estimate the relative copy number of each reference genome in a sample. This framework used an Expectation-Maximization (EM) algorithm to optimize the log-likelihood function associated with the model. We have previously shown our EM algorithm to be highly accurate in its species relative abundance estimation capabilities [95]. Any bacterial strain found within a sample in less than 1e-5 relative abundance was considered to be noise and their abundance was dropped to 0. Bacterial strain-level assignments were rolled back to species-level assignments (by using accession and tax ids with NCBIs taxonomic assignments), and relative abundances were summed to produce bacterial species genome relative abundances. Principal components analysis was performed using Scikit-learn (version 0.23.2). Before PCA, species relative abundances within the sample-taxa matrix were first Centered Log-Ratio (CLR) transformed (all zero values were replaced with 1e-10 before transformation). CLR-transformation [24] is defined as:

$$\mathrm{clr}(\mathrm{x}) = \left[\ \ln\frac{\mathrm{x}_1}{\mathrm{g}(\mathrm{x})},\ \ln\frac{\mathrm{x}_2}{\mathrm{g}(\mathrm{x})}\ldots,\ \ \ln\frac{\mathrm{x}_D}{\mathrm{g}(\mathrm{x})}\right]$$

where (x) is the vector of species abundances within each sample and (D) is the total number of species. The geometric mean of vector (x) is defined as:

$$g(x) = \sqrt[D]{x_1 \times x_2 \times \ldots x_D}$$

Random Forest analysis

CLR-transformed species relative abundances were analyzed using the Random Forest Classifier (RFC) package from Scikit-learn [96]. Random forests were trained and tested with a 70% training and 30% testing sample split and 100 trees per forest. One-hundred RFCs were constructed in order to deem a species as significantly 'important'. First, a 'random' feature was created from randomly selected CLR-transformed species sample relative abundances to assist in the selection of significantly 'important' species, as default importance measurements from random forest classifiers are known to be biased [35]. Next, the importances (Gini importance) for each species provided from all 100 RFCs was compared to those of the 100 'random' feature importances. Only species with statistically significant higher 'importance' according to a Mann-Whitney U test and Benjamini-Hochberg (FDR) multiple testing correction (MWU-FDR: qvalue< 0.05) were deemed significantly 'important'. The AUC (Area Under the Receiver-Operator Curve) and Classification Accuracy (Jaccard index) were both utilized to measure the accuracy of trained forests. The AUC is an estimator of true and false positive prediction rates of our RFC, whereas the Classification Accuracy computes subset accuracy (where a prediction for a set of labels must *exactly* match those from the known true corresponding label set).

Bacterial species diversity analysis

To measure the diversity of species found within each sample, total bacterial richness (total species found in a sample) and the Shannon index [97] were calculated. The Shannon index is calculated as:

$$\text{Shannon Index (H)} = -\sum_{i=1}^{D} P_i \cdot \ln P_i$$

where (D) is the total number of species, and (P_i) is the proportion of that species within the sample.

Differential relative abundance of species

Species relative abundances within the sample-taxa matrix were first CLR-transformed (all zero values were replaced with 1e-10 before transformation). Mann-Whitney U test and FDR correction were utilized to test for significant species relative abundance differences between groups. Only species with a qvalue < 0.05 and a sample prevalence greater than 10% within at least one group were deemed truly differentially abundant.

Bacterial species functional profiles

Gene prediction was performed on all bacterial reference strain genomes utilizing Prodigal [98] (version 2.6.3). All protein sequence translations for genes output by prodigal were provided to InterProScan [99] (version 5.39–77.0) to find matches for protein domains and protein families against the Pfam [48] (version 32.0) and TIGRFAM [47] (version 15.0) databases, respectively. All Pfams and TIGRFAMS found within genomes were counted and then counts were normalized (by total) producing relative abundances. Species functional profiles were created separately for Healthy and CRC groups. This was performed by weighing strain functional profiles by strain average abundance within a group and then summing the strain functional profiles together, followed by re-normalization (by total).

Sample functional profiling

To explore the bacterial community functional capabilities contained within each sample, a simplified annotation format file (SAF) containing the bacterial chromosomal coordinates of features (either Pfams or TIGRFAMs) for all strains was created. Next, the SAF was provided to FeatureCounts [100] (Subread package 2.0.0) to find all features contained within the sample reads. Lastly, the counts of features were subsequently length normalized, summed, then re-normalized (by total) to create a sample functional profile. Function (Pfams or TIGRFAMS) relative abundances within the sample-function matrix were first CLR-transformed (all zero values were replaced with 1e-10 before transformation). Mann-Whitney U test and FDR correction were utilized to test for significant function relative abundance differences between groups. Only functions with a qvalue < 0.05 were deemed significantly different.

Species selection for association network construction

Species selected for network inference were either highly prevalent/abundant species (the union of species exhibiting > 90% sample prevalence within both groups), or species that were deemed as both significantly ‘important’ by random forests and differentially abundant. In total there were 165 species selected for network construction.

Bacterial association network inference

For each sample group, a bacterial association network was constructed from CLR-transformed sample-taxa matrix of that group using a Gaussian Graphical Model (GGM) framework. For each group, a sparse precision matrix (Ω) was computed using the huge [101] package in R, and this matrix formed the adjacency matrix of the association network. The stability approach to regularization (StARS) method [102] was utilized to choose the tuning parameter (ρ) in the l1-penalty model for sparse precision matrix estimation. The partial correlation matrix, *P,* was calculated as:

$$P_{[i,j]} = \frac{-\Omega_{[i,j]}}{\sqrt{\Omega_{[i,i]} \times \Omega_{[j,j]}}}$$

Finally, any associations below a magnitude of 0.01 within the partial correlation matrix was treated as statistical noise and removed.

Network topology comparison

For each network, the following properties were computed using NetworkX [103] (version 2.4): average shortest path length (ASPL), transitivity, and modularity. The ASPL (α) was calculated as:

$$\alpha = \Sigma_{s,t \in L} \frac{D[s,t]}{n(n-1)}$$

where (L) is the set of nodes in the graph (G), the shortest path between the nodes (s) and (t) is D [s,t], and (n) is the total number of nodes in (G). The transitivity (T) of a network was calculated as:

$$T = 3\frac{Total\ triangles}{Total\ triads}$$

in which triangles are considered a subset of three nodes within a network where each node is adjacent to all other nodes within the subset, and triads are connected triples (i.e. three nodes abc where edges (a,b) and (b,c) exist and the edge (a,c) can be present or absent). Transitivity is the fraction of all possible triangles present in the graph and is a measurement of node clustering. Finally, the modularity (Q) [104] of a network was calculated as:

$$Q = \sum_{c=1}^{n} \left[\frac{L_c}{m} - \left(\frac{k_c}{2m} \right)^2 \right]$$

where (n) is all modules of a graph partition, (c) is an individual module from the partition, (m) is the total number of edges of the graph (G), (L_c) is the total intra-module edges, and (k_c) is the sum of edges of all nodes within module (c). Networks were first partitioned into modules before modularity could be calculated (for module detection see Module Functional Profiles below). Monte Carlo simulations were utilized to test networks for non-randomness where 1000 random ($G_{n,p}$) networks were created, using NetworkX, and network properties (ASPL, transitivity, and modularity) were measured and used to produce pvalues for group network properties. For the creation of random networks (n) was equal to the group network of interest's node total and (p) the network density.

Taxonomic relationship analysis of species associations

For each association, the lowest common taxonomic relationship between species was characterized by using the NCBI taxonomic assignments. Monte Carlo simulations were utilized to test for significance and produce pvalues. First, 1000 random ($G_{n,m}$) networks were produced, using NetworkX, for comparison to each group network. Within these networks (n) was equal to the group network node total, and (m) the total edges (associations) within group networks. Next, species names and association weights from group networks were randomly assigned to nodes and edges within each random network. Lastly, the total of each lowest common taxonomic relationship between nodes in each random network were computed and compared to those found within group networks.

Module functional profiles

Species modules were first detected within networks utilizing an asynchronous label propagation algorithm [105] for module detection. The module detection algorithm was allowed to partition the graph into modules 100 times. The modules produced from the partition resulting in the highest 'performance' were kept for subsequent analyses. Performance (p) is calculated as:

$$p = \frac{a+b}{t}$$

where (a) is the total intra-module edges, (b) the total inter-module non-edges, and (t) is the total potential edges. Following module detection, module functional profiles were created by weighing the species functional profile (Pfam or TIGRFAM) of each species within a module by that species mean relative abundance within a group (Healthy or CRC), and then re-normalizing by total.

Module cluster functional analysis

Module functional profiles were CLR-transformed before PCA. To find clusters, modules were partitioned by performing K-means clustering, from Scikit-learn, on the PCA. Silhouette analysis, from Scikit-learn, was used to find the optimal K for K-means clustering. Silhouette coefficients (SC) range from [− 1,1] where a positive SC near 1 indicates that a module within our PCA is far away from neighboring clusters and a high average silhouette score is indicative of well-defined clusters. After clusters were defined, the distinct functionality of clusters was examined. First, PCA was run in a pairwise fashion on the modules from each cluster to find the most important functional features (Pfams or TIGRFA MS), which made a cluster distinct from every other cluster. Across all PCAs, the features which separated each cluster along the first principal component exhibiting an importance above a magnitude of 0.01 were noted and summed. Afterwards the top 100 TIGRFAMS with the highest importance within each cluster were selected, and the main and sub roles of each TIGRFAM elucidated. TIGRFAM main and sub role abundance importances were created by summing the importances of all TIGRFAMS that were assigned to that main and sub role then normalizing by total. Lastly, the top 10 Pfams with the highest total importance were utilized for a more in-depth inspection into a cluster's distinct functionality.

Node centrality 'hub' analysis

Degree centrality was used to find bacterial 'Hubs' within modules by choosing the species with the most associations (edges) within a module. Only 'Hubs' from module sub graphs that exhibited disassortative mixing in respect to degree (degree assortativity < 0), as measured by NetworkX, were selected for examination.

Statistical significance testing

A two-tailed nonparametric t-test (Mann-Whitney U test) [106] was used to compare groups for statistical significance. Benjamini-Hochberg (False discovery rate; FDR) [107] was used for multiple testing correction.

Abbreviations

CRC: Colorectal Cancer; TME: Tumor Microenvironment; RFC: Random Forest Classifier; eHOMD: expanded Human Oral Microbe Database; CLR: Centered Log-Ratio; EM: Expectation-Maximization.; GGM: Gaussian Graphical Model; MWU: Mann-Whitney U Test; FDR: False Discovery Rate

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12866-021-02153-x.

Additional file 1.

Acknowledgements

Not applicable

Authors' contributions

Study Design: ML, SH, SY. Data Analysis: ML. Manuscript Writing: ML, SH, SY. All author(s) have read and approved the final manuscript.

Funding

Not applicable

Availability of data and materials

All data and scripts used in this study can be found at https://github.com/Markloftus/CancerMicrobiome.

Declarations

Ethics approval and consent to participate

Not applicable

Consent for publication

Not applicable

Competing interests

The authors declare no competing interests.

Author details

[1]Burnett School of Biomedical Sciences, Genomics and Bioinformatics Cluster, University of Central Florida, Orlando 32816, FL, USA. [2]Department of Computer Science, Genomics and Bioinformatics Cluster, University of Central Florida, Orlando, FL 32816, USA.

Received: 23 December 2020 Accepted: 16 March 2021
Published online: 31 March 2021

References

1. Sender R, Fuchs S, Milo R. Revised estimates for the number of human and Bacteria cells in the body. PLoS Biol. 2016;14(8):e1002533. https://doi.org/10.1371/journal.pbio.1002533.
2. Kho ZY, Lal SK. The human gut microbiome – a potential controller of wellness and disease. Front Microbiol. 2018;9:1835. https://doi.org/10.3389/fmicb.2018.01835.
3. Thaiss CA, Zmora N, Levy M, Elinav E. The microbiome and innate immunity. Nature. 2016;535(7610):65–74. https://doi.org/10.1038/nature18847.
4. Ahn J, Sinha R, Pei Z, Dominianni C, Wu J, Shi J, Goedert JJ, Hayes RB, Yang L. Human gut microbiome and risk for colorectal Cancer. JNCI: Journal of the National Cancer Institute. 2013;105(24):1907–11. https://doi.org/10.1093/jnci/djt300.
5. Feng Q, Liang S, Jia H, Stadlmayr A, Tang L, Lan Z, Zhang D, Xia H, Xu X, Jie Z, Su L, Li X, Li X, Li J, Xiao L, Huber-Schönauer U, Niederseer D, Xu X, al-Aama JY, Yang H, Wang J, Kristiansen K, Arumugam M, Tilg H, Datz C, Wang J. Gut microbiome development along the colorectal adenoma–carcinoma sequence. Nat Commun. 2015;6(1):6528. https://doi.org/10.1038/ncomms7528.
6. Li S, Konstantinov SR, Smits R, Peppelenbosch MP. Bacterial biofilms in colorectal Cancer initiation and progression. Trends Mol Med. 2017;23(1):18–30. https://doi.org/10.1016/j.molmed.2016.11.004.
7. Kostic AD, Gevers D, Pedamallu CS, Michaud M, Duke F, Earl AM, Ojesina AI, Jung J, Bass AJ, Tabernero J, Baselga J, Liu C, Shivdasani RA, Ogino S, Birren BW, Huttenhower C, Garrett WS, Meyerson M. Genomic analysis identifies association of Fusobacterium with colorectal carcinoma. Genome Res. 2012;22(2):292–8. https://doi.org/10.1101/gr.126573.111.
8. Siegel RL, Miller KD, Goding Sauer A, Fedewa SA, Butterly LF, Anderson JC, Cercek A, Smith RA, Jemal A. Colorectal cancer statistics, 2020. CA A Cancer J Clin. 2020;70(3):145–64. https://doi.org/10.3322/caac.21601.
9. Jones S, Chen W. D., Parmigiani G, Diehl F, Beerenwinkel N, Antal T, et al. comparative lesion sequencing provides insights into tumor evolution. Proc Natl Acad Sci. 2008;105(11):4283–8. https://doi.org/10.1073/pnas.0712345105.
10. Fedirko V, Tramacere I, Bagnardi V, Rota M, Scotti L, Islami F, Negri E, Straif K, Romieu I, la Vecchia C, Boffetta P, Jenab M. Alcohol drinking and colorectal cancer risk: an overall and dose–response meta-analysis of published studies. Ann Oncol. 2011;22(9):1958–72. https://doi.org/10.1093/annonc/mdq653.
11. Thanikachalam K, Khan G. Colorectal Cancer and nutrition. Nutrients. 2019;11(1):164. https://doi.org/10.3390/nu11010164.
12. Botteri E, Iodice S, Bagnardi V, Raimondi S, Lowenfels AB, Maisonneuve P. Smoking and colorectal Cancer: a meta-analysis. JAMA. 2008;300(23):2765–78. https://doi.org/10.1001/jama.2008.839.
13. Fearon ER, Vogelstein B. A genetic model for colorectal tumorigenesis. Cell. 1990;61(5):759–67. https://doi.org/10.1016/0092-8674(90)90186-I.
14. Jasperson KW, Tuohy TM, Neklason DW, Burt RW. Hereditary and familial Colon Cancer. Gastroenterology. 2010;138(6):2044–58. https://doi.org/10.1053/j.gastro.2010.01.054.
15. Dzutsev A, Goldszmid RS, Viaud S, Zitvogel L, Trinchieri G. The role of the microbiota in inflammation, carcinogenesis, and cancer therapy. Eur J Immunol. 2015;45(1):17–31. https://doi.org/10.1002/eji.201444972.
16. Yu T, Guo F, Yu Y, Sun T, Ma D, Han J, et al. Fusobacterium nucleatum Promotes Chemoresistance to Colorectal Cancer by Modulating Autophagy. Cell. 2017;170(3):548–63 e16.
17. Tjalsma H, Boleij A, Marchesi JR, Dutilh BE. A bacterial driver–passenger model for colorectal cancer: beyond the usual suspects. Nat Rev Microbiol. 2012;10(8):575–82. https://doi.org/10.1038/nrmicro2819.
18. Sears CL, Pardoll DM. Perspective: alpha-bugs, their microbial partners, and the link to Colon Cancer. J Infect Dis. 2011;203(3):306–11. https://doi.org/10.1093/jinfdis/jiq061.
19. Gopalakrishnan V, Spencer CN, Nezi L, Reuben A, Andrews MC, Karpinets TV, Prieto PA, Vicente D, Hoffman K, Wei SC, Cogdill AP, Zhao L, Hudgens CW, Hutchinson DS, Manzo T, Petaccia de Macedo M, Cotechini T, Kumar T, Chen WS, Reddy SM, Szczepaniak Sloane R, Galloway-Pena J, Jiang H, Chen PL, Shpall EJ, Rezvani K, Alousi AM, Chemaly RF, Shelburne S, Vence LM, Okhuysen PC, Jensen VB, Swennes AG, McAllister F, Marcelo Riquelme Sanchez E, Zhang Y, le Chatelier E, Zitvogel L, Pons N, Austin-Breneman JL, Haydu LE, Burton EM, Gardner JM, Sirmans E, Hu J, Lazar AJ, Tsujikawa T, Diab A, Tawbi H, Glitza IC, Hwu WJ, Patel SP, Woodman SE, Amaria RN, Davies MA, Gershenwald JE, Hwu P, Lee JE, Zhang J, Coussens LM, Cooper ZA, Futreal PA, Daniel CR, Ajami NJ, Petrosino JF, Tetzlaff MT, Sharma P, Allison JP, Jenq RR, Wargo JA. Gut microbiome modulates response to anti–PD-1 immunotherapy in melanoma patients. Science. 2018;359(6371):97–103. https://doi.org/10.1126/science.aan4236.
20. Matson V, Fessler J, Bao R, Chongsuwat T, Zha Y, Alegre M-L, Luke JJ, Gajewski TF. The commensal microbiome is associated with anti–PD-1 efficacy in metastatic melanoma patients. Science. 2018;359(6371):104–8. https://doi.org/10.1126/science.aao3290.
21. Hibbing ME, Fuqua C, Parsek MR, Peterson SB. Bacterial competition: surviving and thriving in the microbial jungle. Nat Rev Microbiol. 2010;8(1):15–25. https://doi.org/10.1038/nrmicro2259.
22. Gould AL, Zhang V, Lamberti L, Jones EW, Obadia B, Korasidis N, Gavryushkin A, Carlson JM, Beerenwinkel N, Ludington WB. Microbiome interactions shape host fitness. Proc Natl Acad Sci U S A. 2018;115(51):E11951–60. https://doi.org/10.1073/pnas.1809349115.
23. Gloor GB, Macklaim JM, Pawlowsky-Glahn V, Egozcue JJ. Microbiome datasets are compositional: and this is not optional. Front Microbiol. 2017;8:2224. https://doi.org/10.3389/fmicb.2017.02224.
24. Aitchison J. The statistical analysis of compositional data. J R Stat Soc Ser B Methodol. 1982;44(2):139–77.
25. Kurtz ZD, Müller CL, Miraldi ER, Littman DR, Blaser MJ, Bonneau RA. Sparse and Compositionally Robust Inference of Microbial Ecological Networks. von Mering C, editor. PLOS Computational Biology. 2015 May 7;11(5):e1004226.
26. Ai D, Pan H, Li X, Gao Y, Liu G, Xia LC. Identifying gut microbiota associated with colorectal Cancer using a zero-inflated lognormal model. Front Microbiol. 2019;10:826. https://doi.org/10.3389/fmicb.2019.00826.
27. Liao H, Li C, Ai Y, Kou Y. The gut microbiome is more stable in males than in females during the development of colorectal cancer [Internet]. In Review; 2020 Jan [cited 2020 Dec 5]. Available from: https://www.researchsquare.com/article/rs-12211/v1

28. Flemer B, Lynch DB, Brown JMR, Jeffery IB, Ryan FJ, Claesson MJ, O'Riordain M, Shanahan F, O'Toole PW. Tumour-associated and non-tumour-associated microbiota in colorectal cancer. Gut. 2017;66(4):633–43. https://doi.org/10.1136/gutjnl-2015-309595.
29. Warren RL, Freeman DJ, Pleasance S, Watson P, Moore RA, Cochrane K, Allen-Vercoe E, Holt RA. Co-occurrence of anaerobic bacteria in colorectal carcinomas. Microbiome. 2013;1(1):16. https://doi.org/10.1186/2049-2618-1-16.
30. Edgar RC. Accuracy of taxonomy prediction for 16S rRNA and fungal ITS sequences. PeerJ. 2018;6:e4652. https://doi.org/10.7717/peerj.4652.
31. Yachida S, Mizutani S, Shiroma H, Shiba S, Nakajima T, Sakamoto T, Watanabe H, Masuda K, Nishimoto Y, Kubo M, Hosoda F, Rokutan H, Matsumoto M, Takamaru H, Yamada M, Matsuda T, Iwasaki M, Yamaji T, Yachida T, Soga T, Kurokawa K, Toyoda A, Ogura Y, Hayashi T, Hatakeyama M, Nakagama H, Saito Y, Fukuda S, Shibata T, Yamada T. Metagenomic and metabolomic analyses reveal distinct stage-specific phenotypes of the gut microbiota in colorectal cancer. Nat Med. 2019;25(6):968–76. https://doi.org/10.1038/s41591-019-0458-7.
32. Ranjan R, Rani A, Metwally A, McGee HS, Perkins DL. Analysis of the microbiome: advantages of whole genome shotgun versus 16S amplicon sequencing. Biochem Biophys Res Commun. 2016;469(4):967–77. https://doi.org/10.1016/j.bbrc.2015.12.083.
33. Friedman J, Hastie T, Tibshirani R. Sparse inverse covariance estimation with the graphical lasso. Biostatistics. 2008;9(3):432–41. https://doi.org/10.1093/biostatistics/kxm045.
34. Johnson KV-A, Burnet PWJ. Microbiome: should we diversify from diversity? Gut Microbes. 2016;7(6):455–8. https://doi.org/10.1080/19490976.2016.1241933.
35. Strobl C, Boulesteix A-L, Zeileis A, Hothorn T. Bias in random forest variable importance measures: illustrations, sources and a solution. BMC Bioinformatics. 2007;8(1):25. https://doi.org/10.1186/1471-2105-8-25.
36. Oswal P, Katti S, Joshi V, Shaikh H. Identification of Dialister pneumosintes in healthy and chronic periodontitis patients with type 2 diabetes mellitus and its correlation with the red complex bacteria. J Interdiscip Dentistry. 2020;10(1):17. https://doi.org/10.4103/jid.jid_4_19.
37. Contreras A, Doan N, Chen C, Rusitanonta T, Flynn MJ, Slots J. Importance of Dialister pneumosintes in human periodontitis: Dialister pneumosintes in periodontitis. Oral Microbiol Immunol. 2000;15(4):269–72. https://doi.org/10.1034/j.1399-302x.2000.150410.x.
38. Neilands J, Davies JR, Bikker FJ, Svensäter G. Parvimonas micra stimulates expression of gingipains from Porphyromonas gingivalis in multi-species communities. Anaerobe. 2019;55:54–60. https://doi.org/10.1016/j.anaerobe.2018.10.007.
39. Horiuchi A, Kokubu E, Warita T, Ishihara K. Synergistic biofilm formation by Parvimonas micra and Fusobacterium nucleatum. Anaerobe. 2020;62:102100. https://doi.org/10.1016/j.anaerobe.2019.102100.
40. Schlafer S, Riep B, Griffen AL, Petrich A, Hübner J, Berning M, Friedmann A, Göbel UB, Moter A. Filifactor alocis - involvement in periodontal biofilms. BMC Microbiol. 2010;10(1):66. https://doi.org/10.1186/1471-2180-10-66.
41. Chattopadhyay I, Verma M, Panda M. Role of Oral microbiome signatures in diagnosis and prognosis of Oral Cancer. Technol Cancer Res Treat. 2019;18:153303381986735. https://doi.org/10.1177/1533033819867354.
42. Momen-Heravi F, Babic A, Tworoger SS, Zhang L, Wu K, Smith-Warner SA, Ogino S, Chan AT, Meyerhardt J, Giovannucci E, Fuchs C, Cho E, Michaud DS, Stampfer MJ, Yu YH, Kim D, Zhang X. Periodontal disease, tooth loss and colorectal cancer risk: results from the nurses' health study: periodontal disease, tooth loss and colorectal cancer risk. Int J Cancer. 2017;140(3):646–52. https://doi.org/10.1002/ijc.30486.
43. Lee D, Jung KU, Kim HO, Kim H, Chun H-K. Association between oral health and colorectal adenoma in a screening population. Medicine. 2018;97(37):e12244. https://doi.org/10.1097/MD.0000000000012244.
44. Michaud DS, Fu Z, Shi J, Chung M. Periodontal disease, tooth loss, and Cancer risk. Epidemiol Rev. 2017;39(1):49–58. https://doi.org/10.1093/epirev/mxx006.
45. Lauritano D, FOCUS ON. Periodontal disease and colorectal carcinoma. ORL. 2017;10(3):229–33. https://doi.org/10.11138/orl/2017.10.3.229.
46. Escapa IF, Chen T, Huang Y, Gajare P, Dewhirst FE, Lemon KP. New Insights into Human Nostril Microbiome from the Expanded Human Oral Microbiome Database (eHOMD): a Resource for the Microbiome of the Human Aerodigestive Tract. Xu J, editor. mSystems. 2018 Dec 4;3(6):e00187–18, /msystems/3/6/msys.00187–18.atom.
47. Haft DH. TIGRFAMs: a protein family resource for the functional identification of proteins. Nucleic Acids Res. 2001;29(1):41–3. https://doi.org/10.1093/nar/29.1.41.
48. El-Gebali S, Mistry J, Bateman A, Eddy SR, Luciani A, Potter SC, et al. The Pfam protein families database in 2019. Nucleic Acids Res. 2019;47(D1):D427–32. https://doi.org/10.1093/nar/gky995.
49. Bottery MJ, Pitchford JW, Friman V-P. Ecology and evolution of antimicrobial resistance in bacterial communities. ISME J [Internet]. 2020 Nov 20 [cited 2020 Dec 5]; Available from: http://www.nature.com/articles/s41396-020-00832-7
50. D'Souza G, Shitut S, Preussger D, Yousif G, Waschina S, Kost C. Ecology and evolution of metabolic cross-feeding interactions in bacteria. Nat Prod Rep. 2018;35(5):455–88. https://doi.org/10.1039/C8NP00009C.
51. Round JL, Mazmanian SK. The gut microbiota shapes intestinal immune responses during health and disease. Nat Rev Immunol. 2009;9(5):313–23. https://doi.org/10.1038/nri2515.
52. Cerf-Bensussan N, Gaboriau-Routhiau V. The immune system and the gut microbiota: friends or foes? Nat Rev Immunol. 2010;10(10):735–44. https://doi.org/10.1038/nri2850.
53. Newman MEJ. Networks: an introduction. In Oxford University Press, Inc.; 2010. p. 168–234, DOI: https://doi.org/10.1093/acprof:oso/9780199206650.001.0001.
54. Steele JA, Countway PD, Xia L, Vigil PD, Beman JM, Kim DY, Chow CET, Sachdeva R, Jones AC, Schwalbach MS, Rose JM, Hewson I, Patel A, Sun F, Caron DA, Fuhrman JA. Marine bacterial, archaeal and protistan association networks reveal ecological linkages. ISME J. 2011;5(9):1414–25. https://doi.org/10.1038/ismej.2011.24.
55. Lozupone CA, Stombaugh JI, Gordon JI, Jansson JK, Knight R. Diversity, stability and resilience of the human gut microbiota. Nature. 2012;489(7415):220–30. https://doi.org/10.1038/nature11550.
56. Thomas AM, Manghi P, Asnicar F, Pasolli E, Armanini F, Zolfo M, Beghini F, Manara S, Karcher N, Pozzi C, Gandini S, Serrano D, Tarallo S, Francavilla A, Gallo G, Trompetto M, Ferrero G, Mizutani S, Shiroma H, Shiba S, Shibata T, Yachida S, Yamada T, Wirbel J, Schrotz-King P, Ulrich CM, Brenner H, Arumugam M, Bork P, Zeller G, Cordero F, Dias-Neto E, Setubal JC, Tett A, Pardini B, Rescigno M, Waldron L, Naccarati A, Segata N. Metagenomic analysis of colorectal cancer datasets identifies cross-cohort microbial diagnostic signatures and a link with choline degradation. Nat Med. 2019;25(4):667–78. https://doi.org/10.1038/s41591-019-0405-7.
57. Edgar RC. Accuracy of microbial community diversity estimated by closed- and open-reference OTUs. PeerJ. 2017;5:e3889. https://doi.org/10.7717/peerj.3889.
58. Harris AL. Hypoxia — a key regulatory factor in tumour growth. Nat Rev Cancer. 2002;2(1):38–47. https://doi.org/10.1038/nrc704.
59. Corbet C, Feron O. Tumour acidosis: from the passenger to the driver's seat. Nat Rev Cancer. 2017;17(10):577–93. https://doi.org/10.1038/nrc.2017.77.
60. Warburg O, Wind F, Negelein E. The metabolism of tumors in the body. J Gen Physiol. 1927;8(6):519–30. https://doi.org/10.1085/jgp.8.6.519.
61. Commisso C, Davidson SM, Soydaner-Azeloglu RG, Parker SJ, Kamphorst JJ, Hackett S, Grabocka E, Nofal M, Drebin JA, Thompson CB, Rabinowitz JD, Metallo CM, Vander Heiden MG, Bar-Sagi D. Macropinocytosis of protein is an amino acid supply route in Ras-transformed cells. Nature. 2013;497(7451):633–7. https://doi.org/10.1038/nature12138.
62. Kamphorst JJ, Nofal M, Commisso C, Hackett SR, Grabocka E, Heiden MGV, et al. Human pancreatic cancer tumors are nutrient poor and tumor cells actively scavenge extracellular protein 2016;20.
63. Beloribi-Djefaflia S. Lipid metabolic reprogramming in cancer cells. 2016;10.
64. Finicle BT, Jayashankar V, Edinger AL. Nutrient scavenging in cancer. Nat Rev Cancer. 2018;18(10):619–33. https://doi.org/10.1038/s41568-018-0048-x.
65. Yilmaz P, Parfrey LW, Yarza P, Gerken J, Pruesse E, Quast C, Schweer T, Peplies J, Ludwig W, Glöckner FO. The SILVA and "all-species living tree project (LTP)" taxonomic frameworks. Nucl Acids Res. 2014;42(D1):D643–8. https://doi.org/10.1093/nar/gkt1209.
66. O'Leary NA, Wright MW, Brister JR, Ciufo S, Haddad D, McVeigh R, et al. Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. Nucleic Acids Res. 2016;44(D1):D733–45. https://doi.org/10.1093/nar/gkv1189.
67. Loomba R, Seguritan V, Li W, Long T, Klitgord N, Bhatt A, et al. Gut Microbiome-Based Metagenomic Signature for Non-invasive Detection of Advanced Fibrosis in Human Nonalcoholic Fatty Liver Disease. Cell Metab. 2017;25(5):1054–62 e5.
68. Aruni AW, Mishra A, Dou Y, Chioma O, Hamilton BN, Fletcher HM. Filifactor alocis – a new emerging periodontal pathogen. Microbes Infect. 2015;17(7):517–30. https://doi.org/10.1016/j.micinf.2015.03.011.
69. Jalava J, Eerola E. Phylogenetic analysis of Fusobacterium alocis and Fusobacterium sulci based on 16S rRNA gene sequences: proposal of

Filifactor alocis (Cato, Moore and Moore) comb. nov. and Eubacterium sulci (Cato, Moore and Moore) comb. nov. Int J Syst Bacteriol. 1999 Oct;49 Pt 4: 1375–9.

70. Rousee JM, Bermond D, Piemont Y, Tournoud C, Heller R, Kehrli P, Harlay ML, Monteil H, Jaulhac B. Dialister pneumosintes associated with human brain abscesses. J Clin Microbiol. 2002;40(10):3871–3. https://doi.org/10.1128/JCM.40.10.3871-3873.2002.
71. Flemer B, Warren RD, Barrett MP, Cisek K, Das A, Jeffery IB, Hurley E, O'Riordain M, Shanahan F, O'Toole PW. The oral microbiota in colorectal cancer is distinctive and predictive. Gut. 2018;67(8):1454–63. https://doi.org/10.1136/gutjnl-2017-314814.
72. Mima K, Cao Y, Chan AT, Qian ZR, Nowak JA, Masugi Y, et al. Fusobacterium nucleatum in Colorectal Carcinoma Tissue According to Tumor Location: Clinical and Translational Gastroenterology 2016;7(11):e200, DOI: https://doi.org/10.1038/ctg.2016.53.
73. Drewes JL. High-resolution bacterial 16S rRNA gene profile meta-analysis and biofilm status reveal common colorectal cancer consortia. npj Biofilms and Microbiomes. 2017;12.
74. Dai Z, Coker OO, Nakatsu G, Wu WKK, Zhao L, Chen Z, Chan FKL, Kristiansen K, Sung JJY, Wong SH, Yu J. Multi-cohort analysis of colorectal cancer metagenome identified altered bacteria across populations and universal bacterial markers. Microbiome. 2018;6(1):70. https://doi.org/10.1186/s40168-018-0451-2.
75. Baxter NT, Ruffin MT, Rogers MAM, Schloss PD. Microbiota-based model improves the sensitivity of fecal immunochemical test for detecting colonic lesions. Genome Med. 2016;8(1):37. https://doi.org/10.1186/s13073-016-0290-3.
76. Takahashi N. Microbial ecosystem in the oral cavity: metabolic diversity in an ecological niche and its relationship with oral diseases. Int Congr Ser. 2005;1284:103–12. https://doi.org/10.1016/j.ics.2005.06.071.
77. Socransky SS, Haffajee AD. Dental biofilms: difficult therapeutic targets: Dental biofilms: difficult therapeutic targets. Periodontology 2000. 2002; 28(1):12–55.
78. Abubucker S, Segata N, Goll J, Schubert AM, Izard J, Cantarel BL, Rodriguez-Mueller B, Zucker J, Thiagarajan M, Henrissat B, White O, Kelley ST, Methé B, Schloss PD, Gevers D, Mitreva M, Huttenhower C. Metabolic reconstruction for metagenomic data and its application to the human microbiome. PLoS Comput Biol. 2012;8(6):e1002358. https://doi.org/10.1371/journal.pcbi.1002358.
79. Neis E, Dejong C, Rensen S. The role of microbial amino acid metabolism in host metabolism. Nutrients. 2015;7(4):2930–46. https://doi.org/10.3390/nu7042930.
80. Hopkins MJ, Englyst HN, Macfarlane S, Furrie E, Macfarlane GT, McBain AJ. Degradation of cross-linked and non-cross-linked Arabinoxylans by the intestinal microbiota in children. AEM. 2003;69(11):6354–60. https://doi.org/10.1128/AEM.69.11.6354-6360.2003.
81. Pessione E. Lactic acid bacteria contribution to gut microbiota complexity: lights and shadows. Front Cell Inf Microbio [Internet]. 2012 [cited 2020 Dec 8];2. Available from: http://journal.frontiersin.org/article/10.3389/fcimb.2012.00086/abstract
82. Koh A, De Vadder F, Kovatcheva-Datchary P, Bäckhed F. From dietary Fiber to host physiology: short-chain fatty acids as key bacterial metabolites. Cell. 2016;165(6):1332–45. https://doi.org/10.1016/j.cell.2016.05.041.
83. Canani RB. Potential beneficial effects of butyrate in intestinal and extraintestinal diseases. WJG. 2011;17(12):1519–28. https://doi.org/10.3748/wjg.v17.i12.1519.
84. Xu M, Yamada M, Li M, Liu H, Chen SG, Han YW. FadA from Fusobacterium nucleatum utilizes both secreted and nonsecreted forms for functional Oligomerization for attachment and invasion of host cells. J Biol Chem. 2007;282(34):25000–9. https://doi.org/10.1074/jbc.M611567200.
85. Rubinstein MR, Baik JE, Lagana SM, Han RP, Raab WJ, Sahoo D, et al. Fusobacterium nucleatum promotes colorectal cancer by inducing Wnt/β-catenin modulator Annexin A1. EMBO Rep [Internet]. 2019 Apr [cited 2020 Dec 6];20(4). Available from: https://onlinelibrary.wiley.com/doi/abs/10.15252/embr.201847638
86. Yang Y, Weng W, Peng J, Hong L, Yang L, Toiyama Y, et al. Fusobacterium nucleatum Increases Proliferation of Colorectal Cancer Cells and Tumor Development in Mice by Activating Toll-Like Receptor 4 Signaling to Nuclear Factor−κB, and Up-regulating Expression of MicroRNA-21. Gastroenterology. 2017 Mar;152(4):851–866.e24.
87. Phang JM, Liu W, Hancock CN, Fischer JW. Proline metabolism and cancer: emerging links to glutamine and collagen. Curr Opin Clin Nutr Metabolic Care. 2015;18(1):71–7. https://doi.org/10.1097/MCO.0000000000000121.
88. Dixit SN, Seyer JM, Kang AH. Covalent structure of collagen: amino-acid sequence of Chymotryptic peptides from the carboxyl-terminal region of alpha2-CB3 of Chick-skin collagen. Eur J Biochem. 1977;81(3):599–607. https://doi.org/10.1111/j.1432-1033.1977.tb11987.x.
89. Tsoy O, Ravcheev D, Mushegian A. Comparative genomics of ethanolamine utilization. JB. 2009;191(23):7157–64. https://doi.org/10.1128/JB.00838-09.
90. Garsin DA. Ethanolamine utilization in bacterial pathogens: roles and regulation. Nat Rev Microbiol. 2010;8(4):290–5. https://doi.org/10.1038/nrmicro2334.
91. Bolger AM, Lohse M, Usadel B. Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics. 2014;30(15):2114–20. https://doi.org/10.1093/bioinformatics/btu170.
92. Langmead B, Salzberg SL. Fast gapped-read alignment with bowtie 2. Nat Methods. 2012;9(4):357–9. https://doi.org/10.1038/nmeth.1923.
93. Genome Reference Consortium. The GRCh38.p12 Human Genome [Internet]. [cited 2018 Oct 1]. Available from: https://www.ncbi.nlm.nih.gov/assembly/GCF_000001405.38/
94. Xia LC, Cram JA, Chen T, Fuhrman JA, Sun F. Accurate Genome Relative Abundance Estimation Based on Shotgun Metagenomic Reads. Dias-Neto E, editor. PLoS ONE. 2011;6(12):e27992.
95. Loftus M, Hassouneh SA-D, Yooseph S. Bacterial associations in the healthy human gut microbiome across populations. Sci Rep. 2021;11(1):2828. https://doi.org/10.1038/s41598-021-82449-0.
96. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, et al. Scikit-learn: Machine Learning in Python. MACHINE LEARNING IN PYTHON. :6.
97. Shannon CE. A Mathematical Theory of Communication. :55.
98. Hyatt D, Chen G-L, LoCascio PF, Land ML, Larimer FW, Hauser LJ. Prodigal: prokaryotic gene recognition and translation initiation site identification. BMC Bioinformatics. 2010;11(1):119. https://doi.org/10.1186/1471-2105-11-119.
99. Jones P, Binns D, Chang H-Y, Fraser M, Li W, McAnulla C, McWilliam H, Maslen J, Mitchell A, Nuka G, Pesseat S, Quinn AF, Sangrador-Vegas A, Scheremetjew M, Yong SY, Lopez R, Hunter S. InterProScan 5: genome-scale protein function classification. Bioinformatics. 2014;30(9):1236–40. https://doi.org/10.1093/bioinformatics/btu031.
100. Liao Y, Smyth GK, Shi W. featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. Bioinformatics. 2014;30(7):923–30. https://doi.org/10.1093/bioinformatics/btt656.
101. Zhao T, Liu H, Roeder K. Wasserman L. The huge Package for High-dimensional Undirected Graph Estimation in R. 2016;6.
102. Liu H, Roeder K, Wasserman L. Stability Approach to Regularization Selection (StARS) for High Dimensional Graphical Models. arXiv:10063316 [stat] [Internet]. 2010 Jun 16 [cited 2020 Sep 9]; Available from: http://arxiv.org/abs/1006.3316
103. Hagberg AA, Schult DA, Swart PJ. Exploring network structure. Dynamics, and Function using NetworkX. 2008;6.
104. Clauset A, Newman MEJ, Moore C. Finding community structure in very large networks. Phys Rev E. 2004;70(6):066111. https://doi.org/10.1103/PhysRevE.70.066111.
105. Cordasco G, Gargano L. Community Detection via Semi-Synchronous Label Propagation Algorithms. arXiv:11034550 [physics] [Internet]. 2011 Mar 23 [cited 2020 Jul 31]; Available from: http://arxiv.org/abs/1103.4550
106. Mann HB, Whitney DR. On a test of whether one of two random variables is stochastically larger than the other. Ann Math Statist. 1947;18(1):50–60. https://doi.org/10.1214/aoms/1177730491.
107. Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. J R Stat Soc Ser B Methodol. 1995;57(1):289–300.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.