



is implicated in cancer progression and mitigation through their metabolites. For example, the absence of folate-producing *Bifidobacterium* and *Lactobacillus* was correlated with hypomethylation at regions of the p53 gene in colorectal cancer [11]. This research suggests that tissue-specific microbes may play a tumor suppressive role, specifically by causing or preventing the repair of DNA damage. Past microbiome studies have used The Cancer Genome Atlas (TCGA) data and machine learning to identify diagnostic microbial signatures in various cancers but have not elucidated how microbe composition varies in PTC and how microbes interact with pathways to influence cancer progression [10]. We hypothesize that the intratumor microbiome in thyroid carcinoma plays a critical role in determining the varying immune landscape across different types of PTC.

In this study, we characterized the abundance of microbes in PTC using RNA sequencing data obtained from TCGA. We identified microbes differentially abundant between males and females and between different PTC subtypes. To determine the mechanisms by which specific microbes may influence the immune landscape, we correlated microbe abundance to clinical variables, immune-associated (IA) gene expression, immune-cell expression, and cancer and immunological pathways. Finally, we correlated microbe abundance with known copy number variations (CNV) and mutations and methylation sites (Fig. 1). With this study, we aim to characterize the thyroid carcinoma intratumor microbiome and investigate microbiome's unique clinical significance in thyroid carcinoma prognosis.

## 2. Materials and methods

### 2.1. Data acquisition

Raw whole-transcriptome RNA-sequencing, Level 3 normalized mRNA expression read counts, and DNA methylation 450 k sequencing data for untreated, nonirradiated tumor and adjacent normal tissue were downloaded from the Genomic Data Commons (GDC) legacy archive for 563 thyroid carcinoma (THCA) patients (354 CPTC, 101 FVPTC, 35 TCPTC, 135 male, 366 female tumor samples) [12,13]. The GDC provides web-based access to data and metadata from cancer genomics studies aggregated in TCGA, and The Broad Institute developed a suite of tools and pipelines to assist in aggregating and analyzing various types of large-scale genomic and proteomic data. Clinical information for all patients was downloaded from Broad GDAC Firehose [14]. Genomic alteration information for each patient was obtained from the last analysis report (2016) of the Broad Institute TCGA Genome Data Analysis Center [15].

### 2.2. Extraction of microbial reads

RNA-sequencing data were filtered for bacterial reads via direct alignment to a reference library of bacterial sequences using the Pathoscope 2.0 software and the NCBI nucleotide database [16]. To account for the inherent compositionality of microbe data, we performed a ratio transformation of the microbe abundance data. Ratio transformations capture the relationships between samples in the data, regardless of whether the data are counts or proportions. Log-ratios make the dataset linearly related, so that log-ratio abundances of features can be compared to other samples in the data [17,18]. We used the R “compositions” package to perform the variance-stabilizing Aitchison's log transformation to eliminate negative correlation bias and account for large amounts of variability between samples.

### 2.3. Evaluation of contamination

Contamination of RNA-sequencing data may originate from appliance contamination or sequencing errors [19]. In order to account for contamination, we employed multiple contamination correction methods based on TCGA sampling and sequencing methodology [20]. First, scatterplots of microbial abundance versus date of sequencing were produced; potential contaminants were determined based on unrepresentative overabundance of bacteria, or overabundance solely across three (or less) consecutive dates. Second, boxplots of microbe abundance versus sequencing plates were produced to identify unrepresentative overabundance of bacteria, or overabundance of bacteria solely in samples sequenced on the same plate. Third, total abundance across all microbes was compared against the abundance of each individual microbe, per patient. As the total number of read counts for each sample increases, it is reasonable to assume that the abundance of any specific microbe will also increase if the microbe is not a contaminant. We identified potential contaminants in scatter plots of individual microbe abundance versus total abundance with a slope of zero (margin around 0 of  $\pm 0.1$ ).

### 2.4. Differential abundance between patient cohorts and correlation to MACIS (distant metastasis, patient age, completeness of resection, local invasion, and tumor Size) score and clinical variables

The Kruskal-Wallis test was used to determine differential abundance ( $p < 0.05$ ) and correlations between microbe abundance and lymph node metastasis, MACIS score, neoplasm cancer status, pathologic stage, pathologic tumor-node-metastasis stage, and residual tumor ( $p < 0.05$ ).

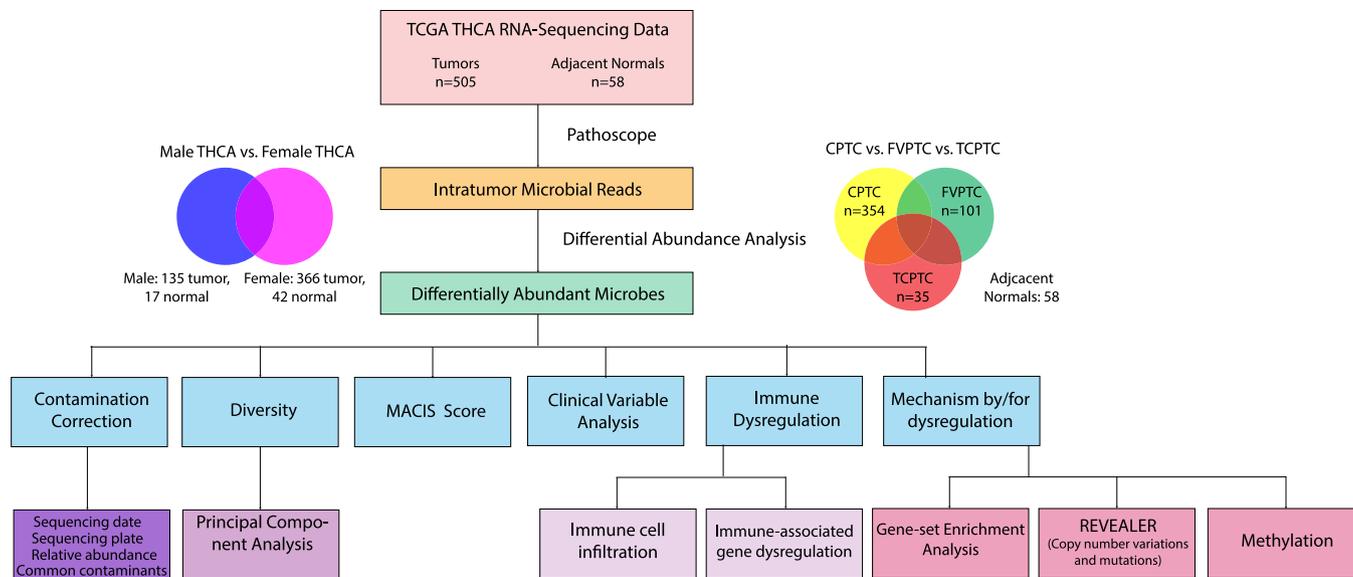
### 2.5. Correlation of microbial abundance to immune-associated elements, copy number variants and mutations, and methylation sites

Dysregulated immune-associated (IA) genes are defined as genes with altered expression between tumor and normal tissue. The genes were determined using edgeR analysis of mRNA expression data obtained from TCGA ( $FDR < 0.05$  and  $|\log \text{fold change}| > 1$ ) [13]. EdgeR offers highly accurate statistical methods for multi-group experiments [21,22]. Dysregulated IA genes were clustered using the Cytoscape Reactome FIViz software. The Kruskal-Wallis test was used to determine the presence of correlation between differentially abundant microbes and dysregulated IA genes ( $p < 0.01$ ). The p-value cut-off was reduced to 0.01 to focus on the most statistically significant correlations. The threshold was determined by p-value in order to evaluate effect size and assess the probability of the results supporting the hypothesis.

Using the software tool CIBERSORTx and mRNA expression data, we used gene expression profiles to estimate immune infiltration [23]. The input was a matrix of gene expression read counts by THCA samples, compiled from mRNA sequencing data [13]. CIBERSORTx software returned expression values for 22 immune cell types.

We conducted Gene Set Enrichment Analysis (GSEA) to correlate microbe abundance to known cancer pathways and immunological signatures. The GSEA software and signatures from the C2, C6, and C7 collections were downloaded from the Molecular Signatures Database [24]. Statistically significantly ( $p < 0.05$ ) enriched signatures were plotted.

The surface-level trends of mutation presence were analyzed by calculating the percentage of patients with each mutation, indicated by a binary value per mutation. The GDAC files were compiled into input files for the REVEALER (repeated evaluation of



**Fig. 1.** Schematic of project workflow and analyses. After obtaining sequencing data from The Cancer Genome Atlas (TCGA), we used the Pathoscope software to extract microbial read counts. We normalized the microbial counts using Aitchison’s log transformation. 1) We performed differential abundance to identify which microbes were significant differentially abundant in tumor vs. normal samples. The overlap in differentially abundant microbial species between various cohorts is demonstrated through venn diagrams: male versus female thyroid carcinoma and classical papillary thyroid carcinoma (CPTC) versus follicular variant papillary thyroid carcinoma (FVPTC) versus tall cell papillary thyroid carcinoma (TCPTC). 2) Three methods of contamination correction were used to identify potential contaminants among the significantly differentially abundant microbes we identified in Step 1. We performed correction by sequencing plate, sequencing data, and abundance in one individual sample vs. all samples. We also compiled a list of common contaminants from previously published research and eliminated these microbes from further study. 3) MACIS (distant Metastasis, patient Age, Completeness of resection, local Invasion, and tumor Size) score and clinical variable analysis of differentially abundant microbes were performed using the Kruskal-Wallis statistical test. Clinical variable data (pathologic stage, TNM stage) for each sample were obtained from the Broad Institute GDAC Firehose database. 4) Correlation of microbial abundance to immune-associated gene expression and immune cell infiltration: We performed gene differential expression analysis on Level 3 normalized mRNA expression read count data obtained from Genomic Data Commons legacy archive/TCGA. We then correlated microbe abundance with dysregulated immune-associated gene expression using Kruskal-Wallis and visualized and categorized the data using Cytoscape Reactome FIViz software. We used CIBERSORTx software to obtain immune cell expression correlated with microbe abundance. 5) We used Gene Set Enrichment Analysis to identify immunological and oncogenic signatures correlated with microbial abundance. 6) We used the Repeated Evaluation of Variables conditionAL Entropy and Redundancy (REVEALER) to identify chromosomal amplifications, deletions, and mutations correlated with microbial abundance. 7) We used DNA methylation 450k sequencing data and a modified workflow of methylationArrayAnalysis from Bioconductor to identify differentially methylated sites in thyroid carcinoma. We correlated differentially methylated sites with microbial abundance using the Kruskal-Wallis statistical test.

variables conditional entropy and redundancy) algorithm, which identifies sets of specific CNVs, including chromosomal deletions, amplifications, and alterations, and mutations that are most likely implicated in changes to the target expression profile. The target profile was identified as the expression of a single immune-associated gene. The REVEALER algorithm runs in multiple iterations in order to identify the most prominent genomic alterations. For our study, we set the maximum number of iterations to three. The algorithm also allows for the use of a seed, or a particular mutation of CNV gain or loss event that may account for target activity. Because we did not know the individual genetic alterations that were responsible for each IA genes’ dysregulation, the seed was set to null. Significant correlations were indicated by  $p < 0.05$  and  $CIC > 0.15$  [25].

Using a modified workflow of methylationArrayAnalysis from Bioconductor, we converted B-values to M-values and then performed probe-wise differential methylation analysis on the matrix of M-values in limma to obtain t-statistics and p-values for each CpG site [26,27]. The Kruskal-Wallis test was used to correlate microbe abundance to the extent of methylation.

2.6. Validation with external dataset

77 CPTC and 48 FVPTC samples were obtained from the European Nucleotide Archive (Project ID: PRJEB11591) [28]. Pathoscope 2.0 was used to align and extract microbial reads. Microbe abundance was plotted to compare relative abundance between tumor samples from TCGA and ENA.

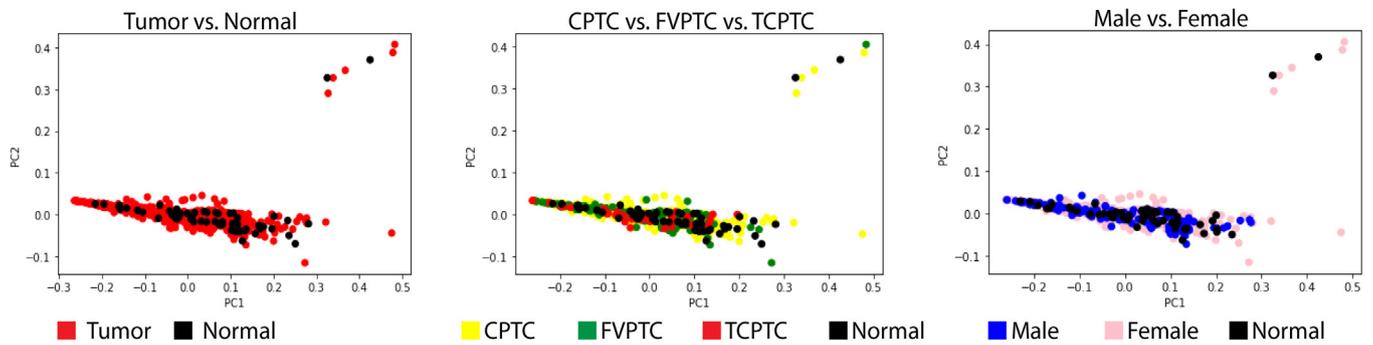
3. Results

3.1. Microbial landscapes of tumor tissue between genders are distinct

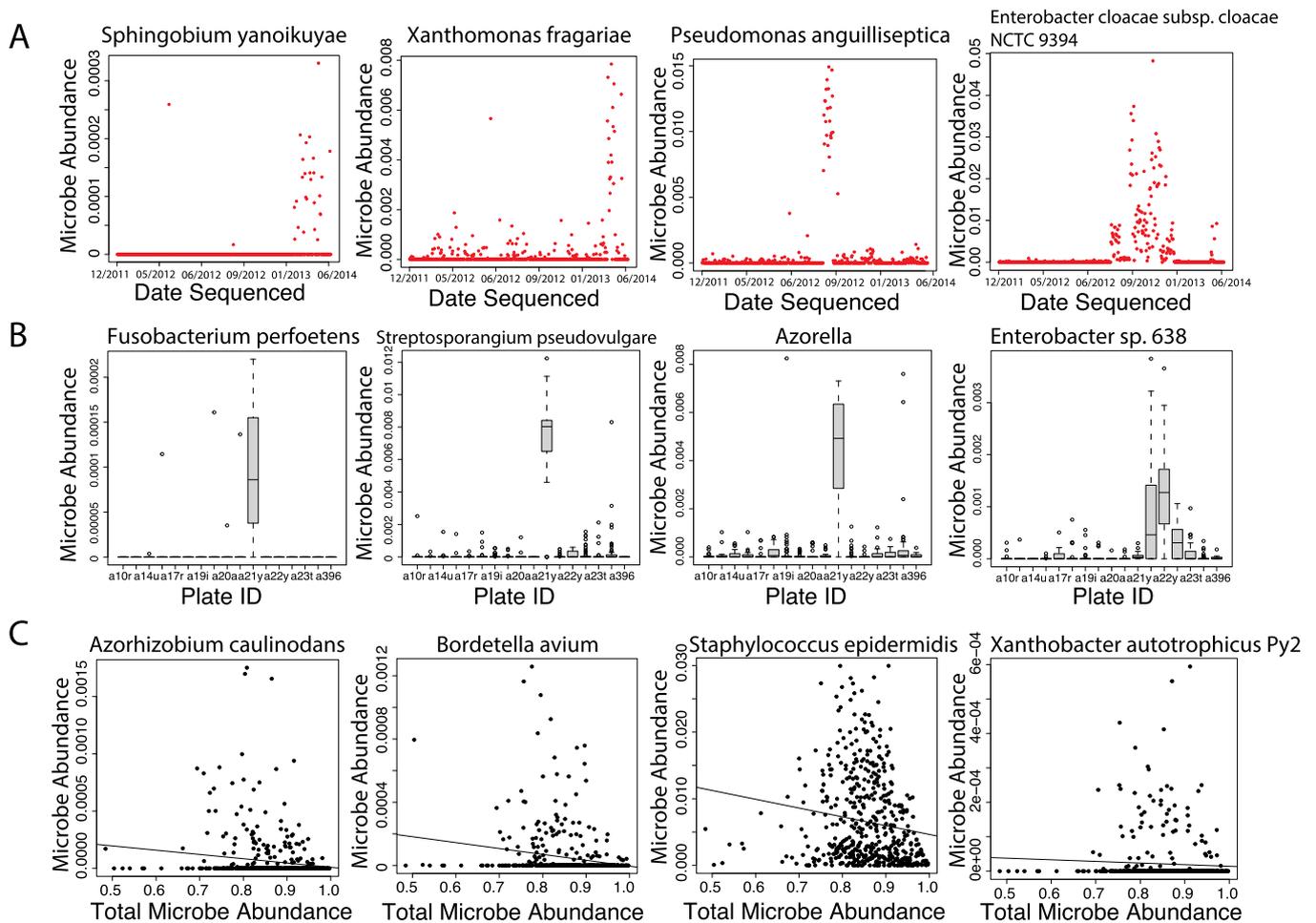
Principal component analysis (PCA) was conducted to measure heterogeneity in microbe species between tumor and normal samples, tumor samples of varied subtype, and tumor samples from different genders (Fig. 2). Overall, the principal component plots show major clustering and indicate that the microbe populations between the tumor samples are relatively homogenous. However, the third PCA plot contains a clustering of female patient tumor samples in the top right corner which makes clear that the microbial landscapes of tumor tissues from the different genders may be distinct. We found the microbes that contributed most to the top-right cluster to be cyanobacteria species.

3.2. Contamination correction

To verify that our measurement of intratumor microbial abundance was consistent across samples and not due to contamination, we conducted three methods of contamination correction. We identified 17 and 21 potential contaminants using correction by sequencing date and plate respectively (Fig. 3A, B). We compared individual microbial abundance to the abundance of all microbes to identify aberrant abundance potential contaminants (Fig. 3C). Finally, we compiled a list of 175 commonly known contaminants from studies of bacterial contamination from reagents and of the hospital microbiome [29,30]. In all, we identified 202



**Fig. 2.** Principal component analysis plots to visualize microbial landscape in the following comparisons: Tumor versus normal tissue, classical papillary thyroid carcinoma versus follicular variant papillary thyroid carcinoma versus tall cell papillary thyroid carcinoma, and males versus females.



**Fig. 3.** Contamination correction. (A) Potential microbial contaminants were identified by producing scatterplots of individual microbe abundance versus date of sequencing of sample. Independent clusters of high abundance across a few dates indicates that the microbe is not consistently differentially abundant across all samples. (B) Potential microbial contaminants were identified by producing boxplots of individual microbe abundance versus sequencing plate. Singular boxplots of high abundance in one sequencing plate indicates that the microbe is not consistently differentially abundant across all samples. Instead, the microbe appears to be differentially abundant due to contamination in one sequencing plate. (C) Read counts of total abundance across all microbes was compared against the abundance of each individual microbe, per patient. Potential contaminants were identified in scatter plots with a slope of zero. This method elicited zero potential contaminants, as all plots had a slope with magnitude significantly greater than zero.

potential contaminants to be eliminated from further investigation (Supplementary Table 1).

### 3.3. Differential abundance of intratumor microbes

We found 45 microbes in CPTC, 34 in FVPTC, and 33 in TCPTC to be differentially abundant between tumor and normal tissue.

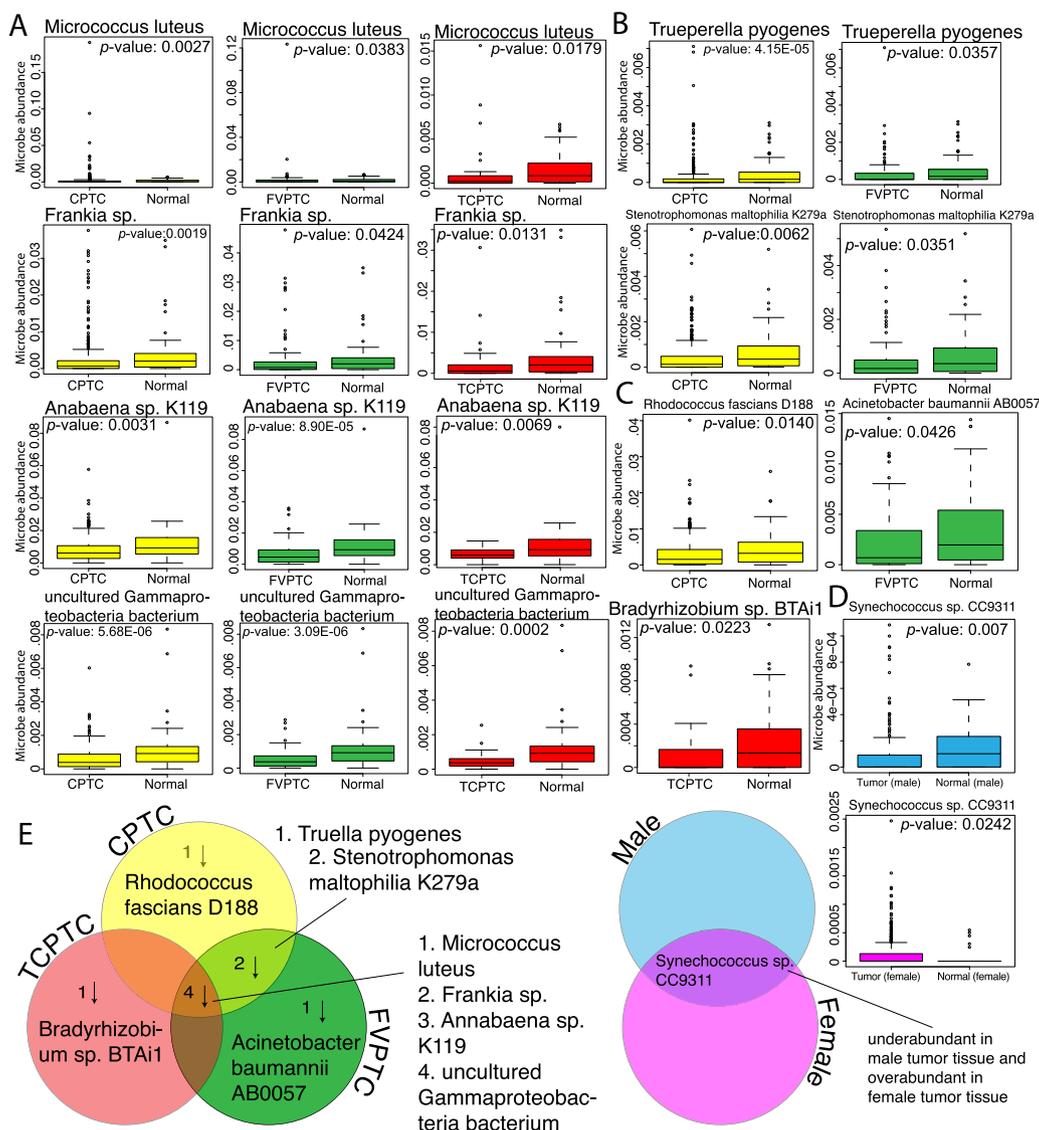
We found 33 microbes in male samples and 49 microbes in female samples to be differentially abundant between tumor and normal tissue. Of all the statistically significant correlations, we focused on investigating the strongest correlations ( $p < 0.05$ ) of microbes that have been previously discovered and studied in the context of cancer and inflammatory disease. *Micrococcus luteus*, *Frankia* sp., *Anabaena* sp. K119, and uncultured

*Gammaproteobacteria bacterium* were all similarly overabundant in normal tissue in CPTC, FVPTC, and TCPTC (Fig. 4A). *Trueperella pyogenes* and *Stenotrophomonas maltophilia* K279a were similarly dysregulated in CPTC and FVPTC (Fig. 4B). We also identified the strongest correlations of microbes uniquely differentially abundant in each subtype cohort (Fig. 4C). *Rhodococcus fascians* D188 was overabundant in normal samples in CPTC. *Acinetobacter baumannii* AB0057 was overabundant in normal samples in FVPTC. *Bradyrhizobium* sp. BTAi1 was overabundant in normal samples in TCPTC.

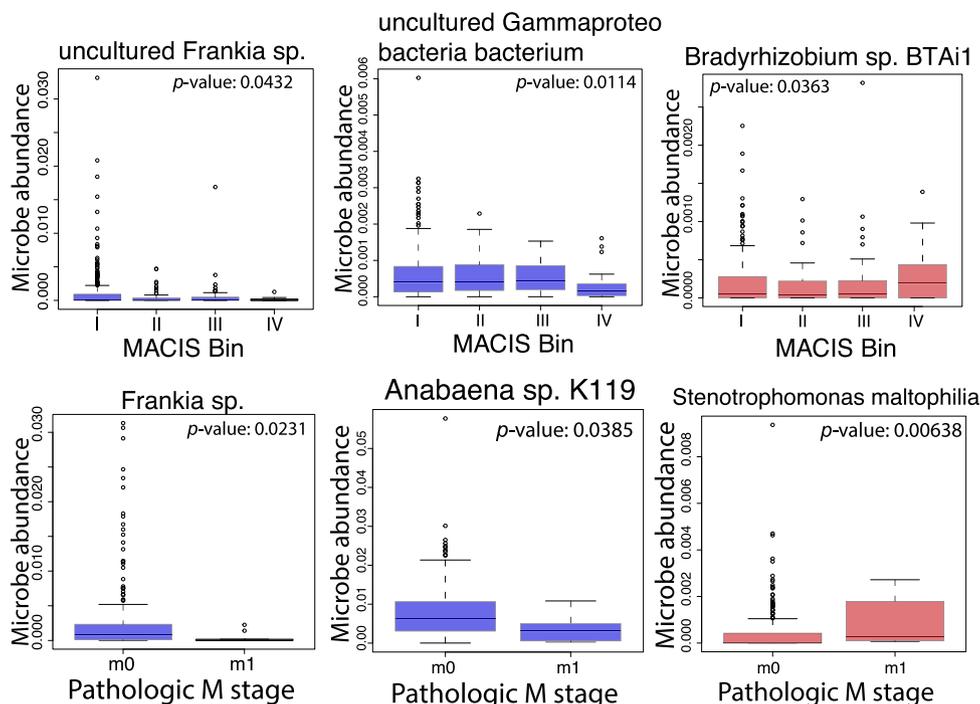
In males, *Synechococcus* sp. CC9311 was overabundant in the normal samples while in females, it was overabundant in tumor samples (Fig. 4D). Overall, normal tissue had significantly more differentially abundant microbes compared to tumor tissue samples (Fig. 4E).

### 3.4. Microbe abundance correlation with MACIS (distant metastasis, patient age, completeness of resection, local invasion, and tumor size) score and clinical variables

The majority of significant correlations between microbe abundance and clinical variables were negative, consistent with the finding that tumor tissue contained lower microbe abundance compared to adjacent normal tissue. *Frankia* sp. and *uncultured Gammaproteobacteria bacterium*, which are abundant in all PTC normal tissue samples, are correlated with lower MACIS score, while *Bradyrhizobium* sp. BTAi1, which is uniquely abundant in TCPTC normal tissue, is correlated with higher MACIS score. *Frankia* sp. and *Anabaena* sp. K119, both of which are overabundant in normal tissue samples of all PTC subtypes, are negatively correlated with Pathologic M stage, while *Stenotrophomonas maltophilia*,



**Fig. 4.** Differential abundance of microbes. (A) Microbes identified to be similarly differentially abundant across all three papillary thyroid carcinoma (PTC) subtypes. Yellow corresponds to classical PTC (CPTC) data, green corresponds to follicular variant PTC (FVPTC) data, and red corresponds to tall cell PTC (TCPTC) data. (B) Microbes identified to be similarly differentially abundant in CPTC and FVPTC. (C) Microbes identified to be uniquely differentially abundant in CPTC, FVPTC, and TCPTC. (D) *Synechococcus* sp. CC9311 was significantly differentially abundant in both male and female thyroid tumor samples. However, the microbe was absent in male tumor tissue versus normal tissue but overabundant in female tumor tissue versus normal tissue. (E) Venn diagram schematics summarizing microbes found to be differentially abundant in patient cohorts. The number indicates the number of most significant correlations, and the down arrow indicates that the microbe was sparse in tumor samples compared to normal samples. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)



**Fig. 5.** Correlation to clinical variables. Select boxplots of microbes significantly positively and negatively correlated with MACIS (distant Metastasis, patient Age, Completeness of resection, local Invasion, and tumor Size) prognostic score and pathological M stage. M0 indicates that the cancer has not spread to other parts of the body, while M1 indicates spread of cancer/distant metastases. All boxplots were produced using the Kruskal-Wallis test.

dysregulated in only CPTC and FVPTC, is positively correlated with pathologic M stage (Fig. 5).

### 3.5. Microbe abundance correlation to immune-associated genes

Dysregulated IA genes found by comparing gender and subtypes cohorts to normal samples were inputted into Reactome FIViz software and clustered into gene modules (Fig. 6A). We found that the greatest number of dysregulated IA genes with the greatest magnitude was correlated to Module 2 pathways in CPTC and FVPTC. However, in TCPTC, IA genes that were correlated to Module 3 pathways were the most dysregulated (Fig. 6B). IA gene dysregulation was comparable between male and female patient cohorts. We determined that CPTC had the greatest absolute number of correlations between microbe dysbiosis and IA gene dysregulation (Fig. 6C).

### 3.6. Microbe abundance correlation to immune cell infiltration

We found FVPTC samples to be strongly correlated to the expression of resting CD4 memory T-cells. In tumor versus normal samples, we observed higher expression of immune cells in normal tissue. Notably, follicular helper T-cells, resting dendritic cells, and activated mast cells were highly expressed in normal tissue in male patients, while naive B cells and resting CD4 memory T cells were highly expressed in normal tissue in female patients (Fig. 6D).

### 3.7. Microbe abundance correlation to cancer and immunologic pathways

By using GSEA to correlate microbe abundance to known cancer and immunologic pathways and signatures, we found that decreased abundance of microbes in CPTC and FVPTC tumor tissue was correlated with decreased expression of EIF4E\_UP and SRC\_UP. V1\_UP oncogenic signatures. In TCPTC, we found that abundance of

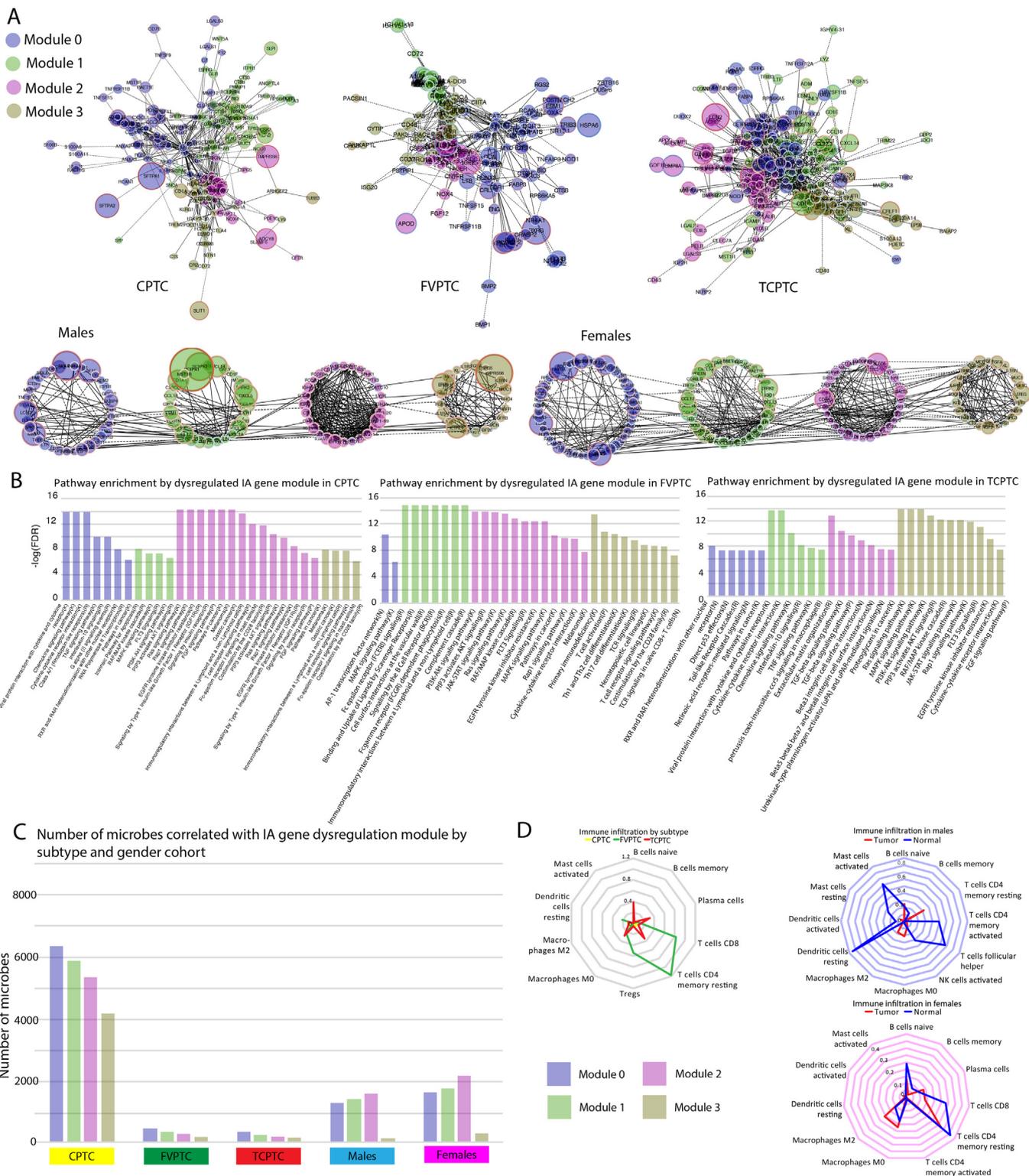
*Anabaena sp. K119* in tumor tissue was correlated with the upregulation of the EIF4E\_UP signature, which is consistent with the more aggressive nature of TCPTC. Among the gender cohorts, the majority of significantly correlated pathways were positively enriched with microbe abundance (Fig. 7A). Although *Synechococcus sp. CC9311* was overabundant in female but not in male tumor tissue, all overlapping pathways were similarly dysregulated in both gender cohorts (Fig. 7B). We found that the majority of most enriched pathways uniquely dysregulated in the different subtypes were related to cell growth (Fig. 7C). In males, the greatest number of uniquely dysregulated pathways belonged to the tumor suppression-related group, and in females, the greatest number of uniquely dysregulated pathways belonged to the DNA checkpoint and damage-related group (Fig. 7D).

### 3.8. Microbe abundance correlation to copy number variations and mutations

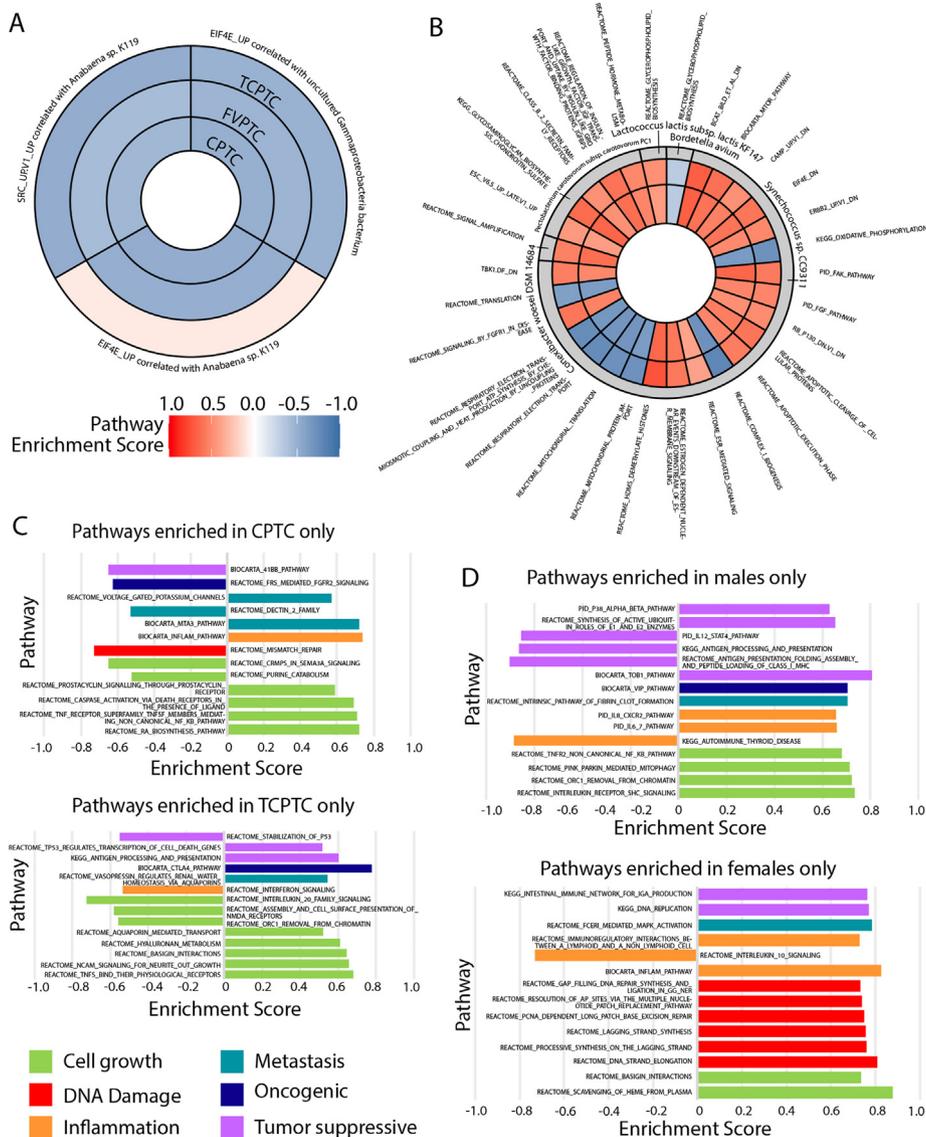
To further investigate how microbes may regulate gene expression, we used the REVEALER algorithm to correlate microbe abundance to CNV and mutation rates (Fig. 8A). Microbe abundance in FVPTC was correlated with the greatest number of chromosomal alterations and mutations. Between the male and female patient cohorts, microbe abundance in males was more frequently and strongly correlated to a greater number of CNV. The Kruskal-Wallis test confirmed that microbes dysregulated in TCPTC and male patients were correlated with higher expression of the BRAF V600E mutation, which is highly implicated in PTC progression (Fig. 8B).

### 3.9. Microbe abundance correlation to gene methylation

We found the most differentially methylated CpG sites between tumor and normal samples at promoter regions on chromosomes with the greatest number of copy number alterations that we



**Fig. 6.** Correlation to immune dysregulation (A) Dysregulated immune-associated (IA) genes of each subtype and gender cohort were clustered into modules by Cytoscape Reactome FIViz software. Module 0 corresponded to cytokine expression and signaling-related pathways, Modules 1, 2, and 3 describe signaling and cancer pathways that were grouped together according to strongest pairwise relationships of functional interactions, or interactions in which two proteins are involved in the same reaction as input, catalyst, activator, or inhibitor. (B) Pathway enrichment analysis using Reactome FI was performed to identify pathways most correlated with IA gene modules ( $p < 0.05$  and  $FDR < 0.05$ ) (C) Barplot summarizing the number of microbes (differentially abundant in each patient cohort) correlated with IA genes in each module. Differentially abundant microbes were correlated with dysregulated IA gene expression using the Kruskal-Wallis test. (D) Radar plots visualizing levels of immune cell infiltration, measured using CIBERSORTx software, in subtype and gender cohorts.



**Fig. 7.** Gene Set Enrichment Analysis (GSEA). (A) Heatmap of GSEA correlations associating microbe abundance to dysregulated pathways in the three subtype cohorts: classical papillary thyroid carcinoma, follicular variant papillary thyroid carcinoma, tall cell papillary thyroid carcinoma ( $p < 0.05$ ). Uncultured gammaproteobacteria bacterium and Anabaena sp. K119 were both less abundant in tumor samples. Red corresponds to upregulated pathways, and blue corresponds to downregulated pathways. (B) Heatmap of GSEA correlations associating microbe abundance to dysregulated pathways in both male and female patient cohorts ( $p < 0.05$ ). The innermost ring corresponds to enriched pathways in male tumor samples, and the middle ring corresponds to enriched pathways in female tumor samples. The outermost ring specifies the microbe correlated to each dysregulated pathway. All microbes, except for *Synechococcus* sp. CC9311, were less abundant in tumor samples versus normal samples in both cohorts. In male samples, *Synechococcus* sp. CC9311 was less abundant in tumor tissue than in normal tissue, while in females, *Synechococcus* sp. CC9311 was more abundant in tumor tissue than in normal tissue. (C) Barplots visualizing most enriched pathways uniquely dysregulated ( $p < 0.05$ ) in papillary thyroid cancer subtype cohorts. (D) Barplots visualizing most enriched pathways uniquely dysregulated ( $p < 0.05$ ) in gender cohorts. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

identified from our REVEALER analysis ( $p < 0.05$ ). The most significantly methylated sites occurred on chromosomes 1 and 17 and specifically at cell cycle related genes. For the majority of significant correlations, including 55 in CPTC, 17 in FVPTC, and 5 in TCPTC, lower microbe abundance was correlated with greater extent of methylation at known tumor suppressor genes including NEURL1B, POLE, and SRCIN1 [31–33] (Fig. 8C).

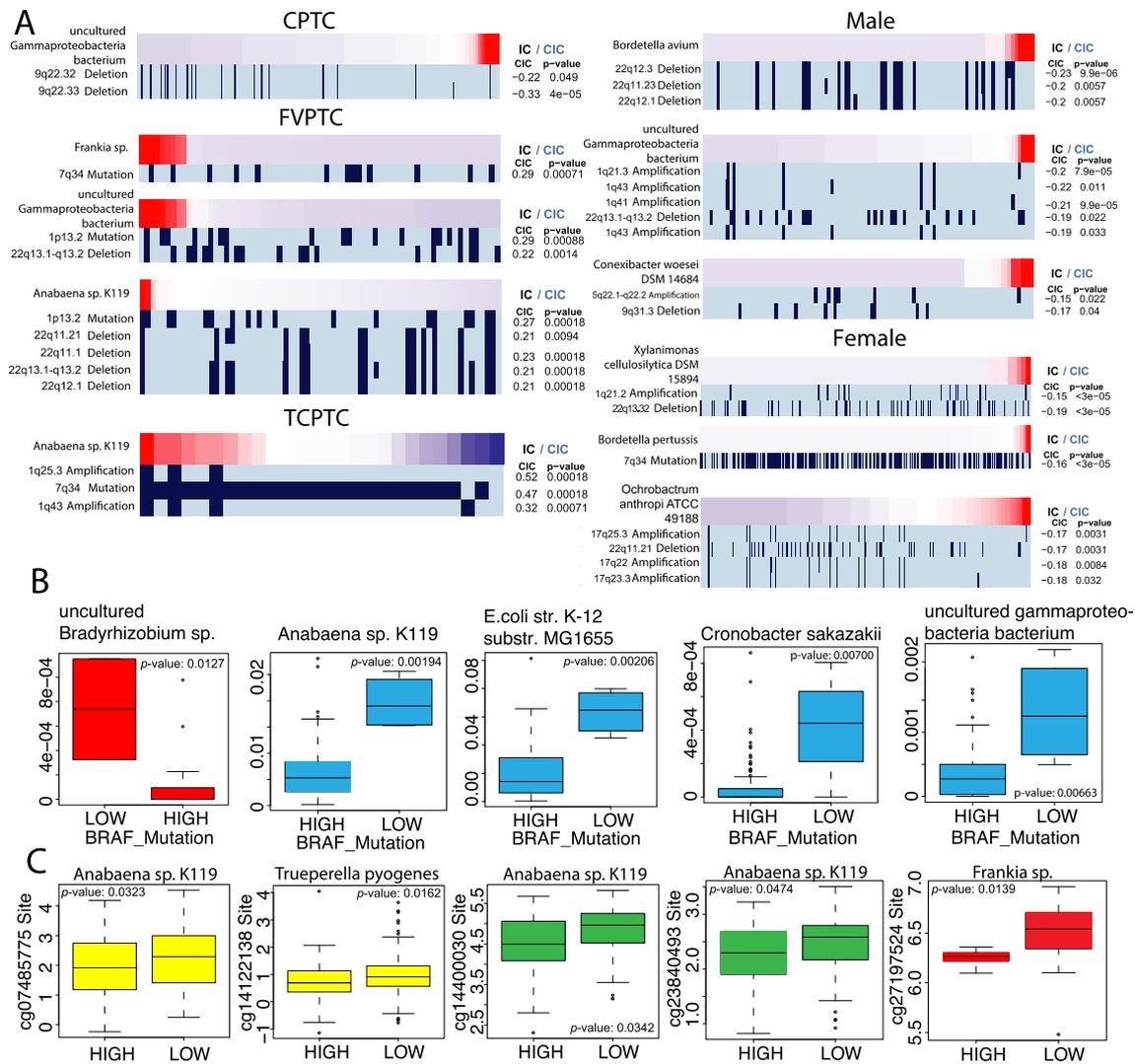
### 3.10. Validation with ENA dataset

All microbes identified from TCGA data and discussed in the manuscript were also present in samples from the ENA dataset (Fig. 9). Microbe abundance was comparable between the two groups of tumor samples from each dataset. The ENA dataset did

not contain normal samples, and further *in vitro* and *in vivo* studies are necessary to validate our findings.

## 4. Discussion

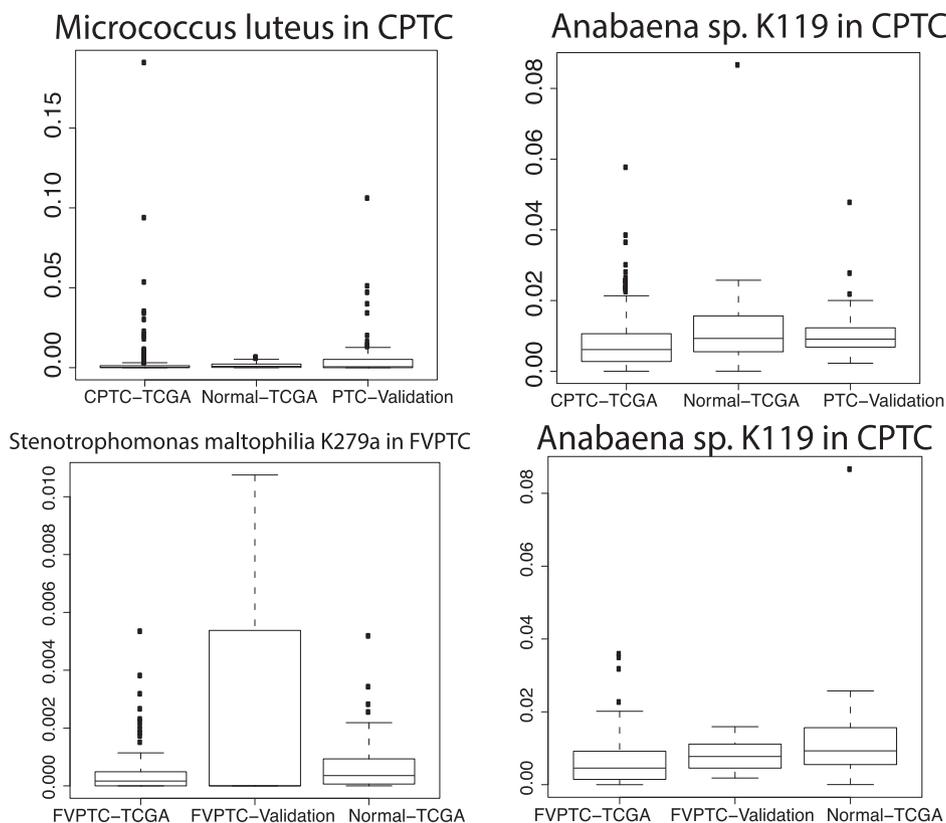
While recent studies have shown the influence of the intratumor microbiome on cancer development, the intratumor microbiome of papillary thyroid carcinoma has remained largely unexplored. Our results indicate that distinct differences in microbe abundance between tumor and normal samples can explain differences in prognosis between patients of varied PTC subtype and gender. PCA results revealed that *Chroococciopsis* sp. and other cyanobacteria species abundant in female tumor



**Fig. 8.** (A) Associations between genomic alterations and microbe abundance of differentially abundant microbes in each subtype and gender cohort represented by mutation heatmaps. Red on the left side of the heatmap indicates that the microbe is overabundant, while red on the right side of the heatmap indicates that the microbe is underabundant. The following rows below the heatmap specify CNV type/genetic variants (alteration, deletion, mutation) across all samples. Each bar represents the presence or absence of a CNV in one sample. Correlations were measured using the Repeated Evaluation of Variables conditionAL Entropy and Redundancy (REVEALER) algorithm (CIC > 0.15,  $p < 0.05$ ) (B) Boxplots correlating microbe abundance to BRAF V600E gene mutation, which is highly implicated in PTC. All boxplots were produced using the Kruskal-Wallis statistical test, and mutation expression data was obtained from the Broad Firehose database. Mutation expression was converted to binary and categorical variables of “HIGH” and “LOW” in order to conduct the Kruskal-Wallis statistical test. The median expression level for the BRAF\_Mutation across all samples was calculated. Values lower than the median value were assigned an expression level of “LOW,” and values above the median value were assigned an expression level of “HIGH.” The x-axis refers to mutation type and level, and the y-axis refers to numerical abundance values of each microbe. The red plot indicates that differential Bradyrhizobium abundance in TCPTC tissue is correlated with higher expression of the BRAF mutation. The blue plots indicate that lower microbe abundance in male tumor tissue is correlated with higher expression of the BRAF mutation. (C) Boxplots correlating microbe abundance with extent of methylation at CpG sites. In this analysis, microbe abundance was converted to a binary and categorical variable of “HIGH” and “LOW” in order to conduct the Kruskal-Wallis test. Microbe abundance was converted to binary and categorical variables of “HIGH” and “LOW” in order to conduct the Kruskal-Wallis statistical test. The median abundance level for each microbe across all samples was calculated. Values lower than the median value were assigned an abundance level of “LOW,” and values above the median value were assigned an abundance level of “HIGH.” The x-axis refers to microbe abundance level, and the y-axis refers to numerical extent of methylation at a particular CpG site. Overall, lower microbe abundance, as observed in PTC tissue, was correlated with higher levels of methylation of cell-cycle related genes and tumor suppressors, including NEURL1B (cg07485775), POLE (cg27197524), and SRCIN1 (cg14122138). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

and normal samples produce great amounts of phycobiliproteins, which possess high antioxidant capacity that prevents cancer cell proliferation [34]. Overall, PCA showed major clustering, which suggests that intratumor microbial populations across all patient samples are fairly predictive. This highlights the potential of using microbes as a diagnostic or treatment measure for PTC. *Micrococcus luteus*, which is overabundant in TCPTC tissue, has been previously shown to cause severe infection in immunocompromised patients [35] as well as anaphylactoid reactions and early death in mice [36]. Presence of *M. luteus* in TCPTC tissue can explain the aggressiveness of TCPTC. On the other hand, *Frankia* sp., which has exhib-

ited antimicrobial activity against phytopathogens, may act similarly against cancerous molecules in humans [37]. Among the species differentially abundant in only CPTC and FVPTC, it was recently found that presence of intratumor *Trueperella pyogenes* was significantly correlated with lower PSA levels in prostate cancer patients [38]. In addition, microbiota dysbiosis in CPTC was correlated with the greatest number of IA genes, potentially explaining why CPTC patients are low-risk, have superior prognosis, and respond better to immunotherapy compared to other PTC subtypes that are less immune-associated [39]. *Bradyrhizobium* sp. *BTAi1*, which we found to be uniquely dysregulated in TCPTC tissue



**Fig. 9.** Boxplots comparing microbe abundance between tumor samples from The Cancer Genome Atlas (TCGA) and a separate dataset with CPTC and FVPTC data obtained from the European Nucleotide Archive (ENA) (Project ID: PRJEB11591). “Normal-TCGA” refers to adjacent normal thyroid tissue matched to cancer tissue for corresponding patients.

and positively correlated with MACIS score, has been shown to be consistently overabundant in tumor tissue, especially in cervical cancer [40]. Significant difference of *Synechococcus* sp. *CC9311* abundance based on gender could explain differences in PTC progression between males and females.

Overall, the trend of overabundance of microbe species in normal samples strongly suggests that intratumor microbes in CPTC and FVPTC are critical for stimulating tumor-suppressive activity. Our analyses support the anti-tumor role of microbes by showing that microbes may regulate CD4 and helper T-cells to defend against tumor growth in FVPTC. However, the ubiquitous overexpression of immune cells may contribute to the aggressiveness of PTC by causing systemic inflammation, thereby promoting tumor growth, as seen in the male patient cohort and in the case of *M. luteus* and *Bradyrhizobium* sp. *BTAi1* abundance in TCPTC [41]. A recently published study found that microbiome-derived ligands and metabolites act directly on intestinal immune cells and travel to remote tissues via the systemic circulatory system to modulate immune cell expression. Some microbe species were found to prevent disease-promoting activity and limit the expansion of iNKT cells by producing sphingolipids. Other species were also found to promote CD4+ T cell differentiation and to balance Th1 and Th2 populations. On the other hand, dysbiosis of specific microbes and the overexpression of TGF- $\beta$  resulted in increased regulatory T-cell expression and expansion of proinflammatory Th17 cells [42]. These results serve to validate our own findings that microbes can regulate immune cell expression differently to mitigate or aggravate disease.

In order to clarify microbes’ putative roles established by past literature, we conducted pathway, mutation, and methylation correlational analyses to elucidate mechanisms by which these

microbes may regulate components of the immune landscape. We found microbes abundant in CPTC and FVPTC to be negatively correlated with oncogenic signatures, which is consistent with the two subtypes being less deadly than TCPTC and poorly differentiated thyroid cancer. We found microbe abundance in males to be negatively correlated with tumor suppressive pathways, which may explain worse prognosis in males, while microbe abundance in females was positively correlated with DNA damage. We found microbe abundance in males to correlate with a greater number of chromosomal alterations compared to female patient samples. Most significant methylation sites were found at cell-cycle and tumor suppressor related genes, and lower microbe abundance was correlated to higher levels of mutation, which suggests that higher microbe abundance, as observed in normal samples, may be necessary to properly regulate cell-cycle related genes.

To the best of our knowledge, we are the first to provide the largest comprehensive profiling of papillary thyroid carcinoma samples to date and show how the abundance of specific microbes may contribute to pathogenesis in different subtypes of PTC. We used numerous computational tools to rigorously collect and correlate microbial abundance data to cancer stage and severity, immune cell expression, cancer pathways, known cancer-related copy number variations and mutations, and methylation. Thorough integration and interpretation of our data suggests how specific microbes may work to mitigate tumor growth in CPTC and FVPTC or aggravate tumor growth in TCPTC. While we cannot definitively conclude if microbes directly cause differences in clinical outcome of thyroid cancer or if microbe abundance is a result of the hypoxic environment created by the tumor, previously published studies on the intratumor microbiome strongly suggest that microbes more likely contribute to pathogenesis than they are a mere result

of it. One study found that differences in the gut microbiome composition directly related to p53's tumor-suppressive effectiveness [43]. Another study found that microbiota contributed to colorectal cancer pathogenesis and genomic instability by secreting chemokines that recruit immunosuppressive myeloid-derived suppressor cells, tumor-associated macrophages, and tumor-associated neutrophils and by inhibiting the activity of cytotoxic activity of natural killer cells. The study also found some microbe species, including *Clostridia*, to play an anti-tumor role in colorectal cancer. These publications provide support to our own findings that specific microbe species can play an anti-tumor or pro-tumor role in thyroid cancer [44]. Nonetheless, since the results of our study are limited to correlative data, controlled *in vitro* and *in vivo* experiments are necessary to confirm our computational findings on the mechanisms by which specific microbes contribute to tumor growth or suppression in thyroid cancer subtypes.

## 5. Conclusion

In conclusion, our study significantly advances the understanding of the papillary thyroid carcinoma microbiome composition, its heterogeneity across samples of different PTC subtypes and patient gender, and its relationship with clinical and immunologic variables. Microbiota abundant in CPTC (*Frankia* sp. and *Trueperella pyogenes*) are highly correlated with immune-associated genes, suggesting that metabolite-gene interaction may result in better prognosis of CPTC compared to the more aggressive TCPTC phenotype. *M. luteus* and *Bradyrhizobium* sp. *BTAi1*, which are abundant in TCPTC, are positively correlated with uncontrolled cell growth pathways and oncogenic pathways (BIOCARTA\_CTLA4\_PATHWAY, p53 instability). Our results suggest that the overabundance of *M. luteus* and *Bradyrhizobium* sp. *BTAi1* in TCPTC and male samples causes dysregulation of these critical pathways, leading to high levels of mutation and causing greater cancer severity (strong positive correlation to MACIS score). While validation analysis showed similar levels of abundance of the same microbes that we identified from TCGA data, deeper sequencing, spatial examination of tissue, and controlled experiments are necessary to confirm our findings on the effects of PTC microbiota and microbial heterogeneity between samples. *In vitro* and *in vivo* experiments must also be conducted to elucidate how varied physiological conditions of the tumor microenvironment could possibly determine and regulate microbe abundance and activity and how microbes can be utilized as therapeutic agents and prebiotic or probiotic treatment options for PTC.

## Funding

UCSD Academic Senate, Grant, RG096651 to WMO.

## CRediT authorship contribution statement

**Aditi Gnanasekar:** Methodology, Formal analysis, Investigation, Writing - original draft, Writing - review & editing, Visualization. **Grant Castaneda:** Formal analysis, Investigation, Writing - review & editing. **Anjali Iyengar:** Formal analysis, Investigation, Writing - review & editing. **Shruti Magesh:** Formal analysis, Investigation, Writing - review & editing. **Daisy Perez:** Formal analysis, Investigation, Writing - review & editing. **Jaideep Chakladar:** Methodology, Writing - review & editing. **Wei Tse Li:** Methodology, Writing - review & editing. **Michael Bouvet:** Writing - review & editing. **Eric Y. Chang:** Writing - review & editing. **Weg M. Ongkeko:** Conceptualization, Resources, Writing - review & editing, Supervision, Project administration, Funding acquisition.

## 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.2021.03.032>.

## References

- [1] Thyroid Cancer - Statistics. (2020, February 28). Available at <https://www.cancer.net/cancer-types/thyroid-cancer/statistics> [Accessed October 09, 2020].
- [2] Yao, R, Chiu, CG, Strugnell SS, Gill S, Wiseman SM. Gender differences in thyroid cancer: a critical review; 2011. Available at: <https://www.medscape.com/viewarticle/739197> [Accessed October 09, 2020].
- [3] Rahbari R, Zhang L, Kebebew E. Thyroid cancer gender disparity. *Future Oncol* 2010;6(11):1771–9. <https://doi.org/10.2217/fon.10.127>.
- [4] Shi X, Liu R, Basolo F, Giannini R, et al. Differential clinicopathological risk and prognosis of major papillary thyroid cancer variants. *J Clin Endocrinol Metab* 2016;101(1):264–74. doi: 10.1210/jc.2015-2917. Epub 2015 Nov 3. PMID: 26529630; PMCID: PMC4701842.
- [5] Cabanillas ME, Hu MI, Durand J-B, Busaidy NL. Challenges associated with tyrosine kinase inhibitor therapy for metastatic thyroid cancer. *J Thyroid Res* 2011;2011:1–9. <https://doi.org/10.4061/2011/985780>.
- [6] Lirov R, Worden FP, Cohen MS. The treatment of advanced thyroid cancer in the age of novel targeted therapies. *Drugs* 2017;77(7):733–45. <https://doi.org/10.1007/s40265-017-0733-1>.
- [7] Sipsos JA, Shah MH. Thyroid cancer: emerging role for targeted therapies. *Ther Adv Med Oncol* 2010;2(1):3–16. <https://doi.org/10.1177/1758834009352667>.
- [8] Faugerat L, Pirson AS, Donckier J, et al. Refractory thyroid carcinoma: which systemic treatment to use? *Ther Adv Med Oncol* 2018;10:1758834017752853. Published 2018 Jan 23. doi:10.1177/1758834017752853.
- [9] Naoum GE, Morkos M, Kim B, Arafat W. Novel targeted therapies and immunotherapy for advanced thyroid cancers. *Mol Cancer* 2018;17(1). <https://doi.org/10.1186/s12943-018-0786-0>.
- [10] Poore GD, Kopylova E, Zhu Q, Carpenter C, Fraraccio S, Wandro S, et al. Microbiome analyses of blood and tissues suggest cancer diagnostic approach. *Nature* 2020;579(7800):567–74. <https://doi.org/10.1038/s41586-020-2095-1>.
- [11] Bahmani S, Azarpira N, Moazamian E. Anti-colon cancer activity of Bifidobacterium metabolites on colon cancer cell line SW742. *Turk J Gastroenterol* 2019;30(9):835–42. <https://doi.org/10.5152/tjg.201910.5152/tjg.2019.18451>.
- [12] National Cancer Institute GDC Legacy Archive. Available at: <https://portal.gdc.cancer.gov/legacy-archive/search/f> [Accessed August 01, 2020].
- [13] National Cancer Institute GDC Data Portal. Available at: <https://portal.gdc.cancer.gov/> [Accessed August 01, 2020].
- [14] Broad Institute Firehose Broad GDAC. Available at: <https://gdac.broadinstitute.org/> [Accessed August 01, 2020].
- [15] Broad Institute TCGA GDAC Analysis Overview. Available at: [http://gdac.broadinstitute.org/runs/analyses\\_latest/reports/](http://gdac.broadinstitute.org/runs/analyses_latest/reports/) [Accessed August 01, 2020].
- [16] NCBI Nucleotide. Available at: <https://www.ncbi.nlm.nih.gov/nucleotide/> [Accessed August 01, 2020].
- [17] Gloor GB, Macklaim JM, Pawlowsky-Glahn V, Egozcue JJ. Microbiome datasets are compositional: and this is not optional. *Front Microbiol* 2017 Nov;15(8):2224. <https://doi.org/10.3389/fmicb.2017.02224>. PMID: 29187837; PMCID: PMC5695134.
- [18] Pawlowsky-Glahn V, Egozcue JJ, Tolosana-Delgado R. *Modeling and analysis of compositional data*. London: John Wiley & Sons; 2015.
- [19] Nieuwenhuis TO, Yang SY, Verma RX, Pillalamarri V, Arking DE, Rosenberg AZ, et al. Consistent RNA sequencing contamination in GTEx and other data sets. *Nat Commun* 2020;11(1). <https://doi.org/10.1038/s41467-020-15821-9>.
- [20] Agrawal N, Akbani R, Aksoy BA, Ally A, Arachchi H, Asa S, et al. Integrated genomic characterization of papillary thyroid carcinoma. *Cell* 2014;159(3):676–90. <https://doi.org/10.1016/j.cell.2014.09.050>.
- [21] Robinson MD, McCarthy DJ, Smyth GK. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* 2010;26(1):139–40. <https://doi.org/10.1093/bioinformatics/btp616>.
- [22] McCarthy DJ, Chen Y, Smyth GK. Differential expression analysis of multifactor RNA-Seq experiments with respect to biological variation. *Nucleic Acids Res* 2012;40(10):4288–97. <https://doi.org/10.1093/nar/gks042>.
- [23] <https://cibersortx.stanford.edu/>.
- [24] <https://www.gsea-msigdb.org/gsea/datasets.jsp>.
- [25] Kim JW, Botvinnik OB, Abudayyeh O, Birger C, Rosenbluh J, Shrestha Y, et al. Characterizing genomic alterations in cancer by complementary functional associations. *Nat Biotechnol* 2016;34(5):539–46. <https://doi.org/10.1038/nbt.3527>.
- [26] Maintainer J. methylArrayAnalysis: a cross-package Bioconductor workflow for analysing methylation array data. R package version 1.14.0; 2020

- [27] Du P, Zhang X, Huang C-C, Jafari N, Kibbe WA, Hou L, et al. Comparison of beta-value and M-value methods for quantifying methylation levels by microarray analysis. *BMC Bioinf* 2010;11(1). <https://doi.org/10.1186/1471-2105-11-587>.
- [28] <https://www.ebi.ac.uk/ena/browser/view/PRJEB11591>.
- [29] Glassing A, Dowd SE, Galandiuk S, Davis B, Chiodini RJ. Inherent bacterial DNA contamination of extraction and sequencing reagents may affect interpretation of microbiota in low bacterial biomass samples. *GutPathog* 2016;8(1). <https://doi.org/10.1186/s13099-016-0103-7>.
- [30] Rampelotto PH, Sereia AFR, de Oliveira LFV, Margis R. Exploring the hospital microbiome by high-resolution 16S rRNA profiling. *Int J Mol Sci* 2019;20(12):3099. <https://doi.org/10.3390/ijms20123099>. PMID: 31242612; PMCID: PMC6696720.
- [31] Teider N, Scott DK, Neiss A, Weeraratne SD, Amani VM, Wang Y, Marquez VE, Cho YJ, Pomeroy SL. Neuralized1 causes apoptosis and downregulates Notch target genes in medulloblastoma. *Neuro Oncol* 2010;12(12):1244–56. Doi: 10.1093/neuonc/noq091. Epub 2010 Sep 16. PMID: 20847082; PMCID: PMC3018940.
- [32] Hansen MF, Johansen J, Bjørnevoll I, Sylvander AE, Steinsbekk KS, Sætrum P, et al. A novel POLE mutation associated with cancers of colon, pancreas, ovaries and small intestine. *Fam Cancer* 2015;14(3):437–48. <https://doi.org/10.1007/s10689-015-9803-2>.
- [33] Wang P, Wang H, Li X, Liu Y, Zhao C, Zhu D. SRCIN1 suppressed osteosarcoma cell proliferation and invasion. *PLoS One*. 2016;11(8):e0155518. Published 2016 Aug 11. Doi:10.1371/journal.pone.0155518.
- [34] <https://www.sciencedirect.com/science/article/pii/S1359511320307881>.
- [35] Ianniello NM, Andrade DC, Ivancic S, et al. Native valve infective endocarditis due to *Micrococcus luteus* in a non-Hodgkin's lymphoma patient. *IDCases*. 2019;18:e00657. Published 2019 Oct 14. doi:10.1016/j.idcr.2019.e00657.
- [36] Toshihiko M, Yoshihiro K, Haruhiko T. *Micrococcus luteus* cells and cell walls induce anaphylactoid reactions accompanied by early death and serum cytokines in mice primed with muramyl dipeptide. *FEMS Immunol Med Microbiol* 1997;17(1):49–55. <https://doi.org/10.1111/j.1574-695X.1997.tb00995.x>.
- [37] Marappa N, Ramachandran L, Dharumadurai D, et al. Plant growth-promoting active metabolites from *Frankia* spp. of Actinorhizal *Casuarina* spp. *Appl Biochem Biotechnol* 2020;191:74–91. <https://doi.org/10.1007/s12010-020-03243-8>.
- [38] Ma J, Gnanasekar A, Lee A, Li WT, Haas M, Wang-Rodriguez J, et al. Influence of intratumor microbiome on clinical outcome and immune processes in prostate cancer. *Cancers (Basel)* 2020;12(9):2524. <https://doi.org/10.3390/cancers12092524>.
- [39] Parhar R, Zou M, Al-Mohanna F, et al. IL-12 immunotherapy of *Braf*<sup>A600E</sup>-induced papillary thyroid cancer in a mouse model. *Lab Invest* 2016;96:89–97. <https://doi.org/10.1038/labinvest.2015.126>.
- [40] Rodriguez RM, Hernandez BY, Menor M, Deng Y, Khadka VS. The landscape of bacterial presence in tumor and adjacent normal tissue across 9 major cancer types using TCGA exome sequencing. *Comput Struct Biotechnol J* 2020;18:631–41. <https://doi.org/10.1016/j.csbi.2020.03.003>.
- [41] Coussens LM, Werb Z. Inflammation and cancer. *Nature* 2002;420(6917):860–7. <https://doi.org/10.1038/nature01322>.
- [42] Zheng D, Liwinski T, Elinav E. Interaction between microbiota and immunity in health and disease. *Cell Res* 2020;30(6):492–506. <https://doi.org/10.1038/s41422-020-0332-7>.
- [43] Kadosh E, Snir-Alkalay I, Venkatachalam A, May S, Lasry A, Elyada E, et al. The gut microbiome switches mutant p53 from tumour-suppressive to oncogenic. *Nature* 2020;586(7827):133–8. <https://doi.org/10.1038/s41586-020-2541-0>.
- [44] Janney A, Powrie F, Mann EH. Host–microbiota maladaptation in colorectal cancer. *Nature* 2020;585(7826):509–17. <https://doi.org/10.1038/s41586-020-2729-3>.