The Role of Butyrate Producing Bacteria in the CpG Island Methylator Phenotype of Colorectal Cancer

Sarah E. Vititoe (sev2125)

Dr. Brent L. Williams, PhD (bw2101)

January 15, 2019

No IRB Approval Required

P9419 Master's Essay in Epidemiology I

Instructor (UNI): Dr. Larkin S. McReynolds (lsm34)

Table of Contents

SPECIAL CONSIDERATION NOTES

PERMISSION TO USE MY ASSIGNMENT AS AN EXAMPLE FOR FUTURE CLASSES

X I prefer not

___ You have my permission, but please make the file anonymous

___ You have my permission and I would like my name listed

ABSTRACT

Background. Cancer is the second leading cause of death in the United States, with colorectal cancer (CRC) contributing significantly to cancer incidence and mortality. The CpG Island Methylator Phenotype (CIMP) of CRCs are a subtype of CRCs that are marked by widespread hypermethylation of CpG islands and account for up to 15% of all sporadic colon cancers. The underlying mechanisms leading to these epigenetic modifications in CIMP CRC tumors remains unknown. Certain bacteria within the human gut microbiome could modify the cellular epigenetic landscape in the intestinal tract. Butyrate, a bacterial byproduct of dietary fiber and carbohydrate fermentation and a histone deacetylase inhibitor, could play a beneficial role in preventing CIMP CRC by regulating epigenetic changes in intestinal cells. Our investigation aims to identify CIMP-specific differences in relative abundance of butyrate-producing bacteria (BPB) found in tumor and normal-adjacent tissue biopsies of CRC patients to determine if BPB are deficient in individuals with the CIMP subtype of CRC.

Methods. In this study we analyzed paired tumor and normal adjacent tissues biopsied from 92 patients with CRC. We analyzed tissue samples for methylation of CIMP-specific markers and analyzed the tissue-associated microbiome with high throughput bacterial 16S rRNA gene sequencing. We compared differences in the tissue microbiome between patients with different levels of CIMP marker methylation, as well as with paired analyses between tumor and normal adjacent tissues within groups defined by the methylation status of CIMP markers. We quantified alpha and beta diversity in the microbiome to determine broad differences between CIMP groups. Regression analyses, linear discriminant analysis of effect size (LEfSe), and topological data analysis (TDA) were applied evaluate the relationships between CIMP phenotype, the total relative abundance of butyrate-producing bacteria, and specific butyrogenic genera.

Results. We found several factors in the microbiome that distinguished patients with high levels of CIMP-marker methylation from patients with low levels of CIMP-marker methylation, as well as factors that significantly distinguished CIMP-High tumors from CIMP-High normal adjacent tissues. Alpha diversity was significantly lower in tumor tissues compared to normal tissues of patients with CIMP marker methylation but not in patients without methylation. Cluster analyses revealed two distinct clusters, with one cluster (cluster 1) defined by normal constituents of the gut microbiome, including *Bacteroides* and common butyrate producers, and the other (cluster 2) defined by bacterial taxa that are not typically found at high abundance in the intestine, including *Fusobacterium*. CIMP-High samples and tumor tissues were more likely to be classified into cluster 2. Regression analysis revealed a significant inverse association between total BPB relative abundance and odds of CIMP CRC. LEfSe, TDA, and correlation analyses revealed inverse associations between specific butyrogenic genera, including *Blautia, Coprococcus,* and *Faecalibacterium*, and CIMP CRC. Our analyses also revealed decreases in BPB in tumor tissues relative to paired normal adjacent tissues, and that tumor tissues from CIMP-High patients had lower BPB relative abundance than CIMP-Low or Non-CIMP° tumors.

Conclusions. Our results show that the relative abundance of BPB is significantly decreased in CIMP-High tumor tissues, compared to patients with less or no CIMP marker methylation. While this decrease is among several factors that distinguish the microenvironment of tumor and normal tissues, our findings suggest that loss of BPBs may be an important factor associated with CIMP CRC.

I. INTRODUCTION

Cancer is the second leading cause of death in the United States, with colorectal cancer (CRC) contributing significantly to cancer incidence and mortality.[1] CRC is estimated to be the third leading type of new cancer cases and the third leading cause of cancer deaths in both men and women in the United States.[1] While overall incidence of CRC has declined between 2005 and 2014 in the United States due to increased screening and colonoscopies, incidence rates have increased by 2% a year for individuals under 55 years of age from the mid-1990s to 2014.[1]

A subgroup of normal colon tissues progresses to invasive carcinomas through the serrated neoplasia pathway (SNP). In the SNP, current evidence suggests that *BRAF* mutations are initiating events in the pathway, which are followed by widespread epigenetic changes in CpG islands, including promoter regions of tumor suppressor genes that promote tumorigenesis. Hypermethylation of CpG islands may contribute to progression of serrated polyps into carcinomas.[2] Cancers that develop through the SNP are characterized by defective deoxyribonucleic acid (DNA) mismatch repair (MMR) including loss of *MLH1,* and microsatellite instability (MSI), are commonly associated with *BRAF* mutations and CpG Island methylation, and are known as the CpG Island Methylator Phenotype (CIMP) of CRC.[2,3] This pathway is believed to be the major mechanism of tumorigenesis for sporadic MSI CRC. Such tumors account for up to 15% of sporadic colon cancers, and present with distinct clinical features including proximal location within the colon and an older age of the patient at onset.[2–6] Furthermore, recent literature has indicated that CIMP status in CRC patients is significantly associated with decreased disease-free survival and overall survival, compared to Non-CIMP CRC patients.[5] The underlying mechanisms leading to these epigenetic modifications in CIMP CRC tumors remains unknown.

Over 3.3 million non-redundant microbial genes have been discovered in the intestinal tract through Illumina-based metagenomics sequencing of fecal specimens, representing a community of bacteria that are collectively referred to as the gut microbiome.[7] Some evidence suggests that certain bacteria can impact the epigenome of the human host through DNA methylation and histone acetylation/deacetylation, and these changes may be either harmful or beneficial depending on the genes that undergo epigenetic modification.[8] Butyrate, a short-chain fatty acid (SCFA) produced by some bacteria as a byproduct of dietary fiber fermentation, plays an important role in maintaining homeostasis of the gut microbiome.[9] Butyrate has long been known to have antineoplastic effects, which may be, in part, mediated by its function as a histone deacetylase inhibitor (HDACi).[10] HDACs interact with the cellular DNA methylation machinery to affect chromatin modifications and control the transcriptional activation state of genes.[9] There is some evidence to suggest that butyrate may play a beneficial role in CRC prevention.[9] Butyrate-producing bacteria (BPB) are abundant in microbiomes of healthy individuals with sufficient dietary fiber intake.[11] However, the bacteria in the gut microbiome, including BPB, have been found in both human and primate studies to be highly susceptible to changes in diet.[12,13] This suggests that the health of the gut microbiome, as mediated by lifestyle factors, may influence the epigenome of the intestinal tract, may contribute to tumorigenesis along the SNP, and may potentially be associated with CIMP and MSI subtypes of CRC. However, the relationship between intestinal methylation of CIMP-specific markers and BPB has not been evaluated in CRCs.

Our investigation aims to identify CIMP-specific differences in relative abundance of BPB found in tumor and normal-adjacent tissue biopsies of CRC patients, as quantified by 16s rRNA gene high-throughput sequencing of the microbiome of tumor and normal adjacent tissue biopsies from CRC patients. We aim to compare patients with CIMP CRC tumors to patients with non-CIMP CRC tumors, as well as compare normal adjacent tissues of patients with CIMP CRC to normal adjacent tissues of

patients with non-CIMP CRC to elucidate differences in the microbial community between these groups of CRC patients.

We theorize that deficiencies of BPB in these tissues will be associated with CIMP-specific methylation patterns of CpG Islands. We will investigate whether BPB are deficient in individuals with CIMP CRC and whether this deficiency will be apparent in both their tumor and normal adjacent tissues. We will test both within-tumor-tissue and within-normal-adjacent-tissue comparisons to see if there is a decrease in BPB relative abundance associated with an increase in the number of positive methylation markers and their degree of methylation. Univariate tests for significance, regression models, Linear Discriminant Analysis Effect Size (LEfSe) and topological data analysis will be applied to assess the relationships among these factors.

II. METHODS

Dataset. Data for this thesis were derived from research conducted by Dr. Brent Williams and Mara Couto-Rodriguez and funded by the National Institute of Health's National Cancer Institute Award R01 CCA205028, entitled "The role of butyrate-producing bacteria in CIMP colorectal cancer tumorigenesis".

Study Sample. Tissue specimens were obtained from the Weill Cornell Colorectal Cancer Biobank. 184 tissue samples were collected (one sample of the tumor tissue, and one sample of the normal adjacent tissue each from 92 unique CRC patients). Samples were restricted to non-familial colon cancer probands. Tissues samples from atypical bowel prep and patients who had recently used antibiotics were excluded, since both of these criteria may result in variability or disruption in the gut microbiome **(Figure 1).** De-identified clinical information recorded by the biobank included gender, age at biopsy, Tumor/Lymph Node/Metastasis (TNM) cancer stage, tumor grade, microsatellite status, *MLH1* status, and other clinical variables.

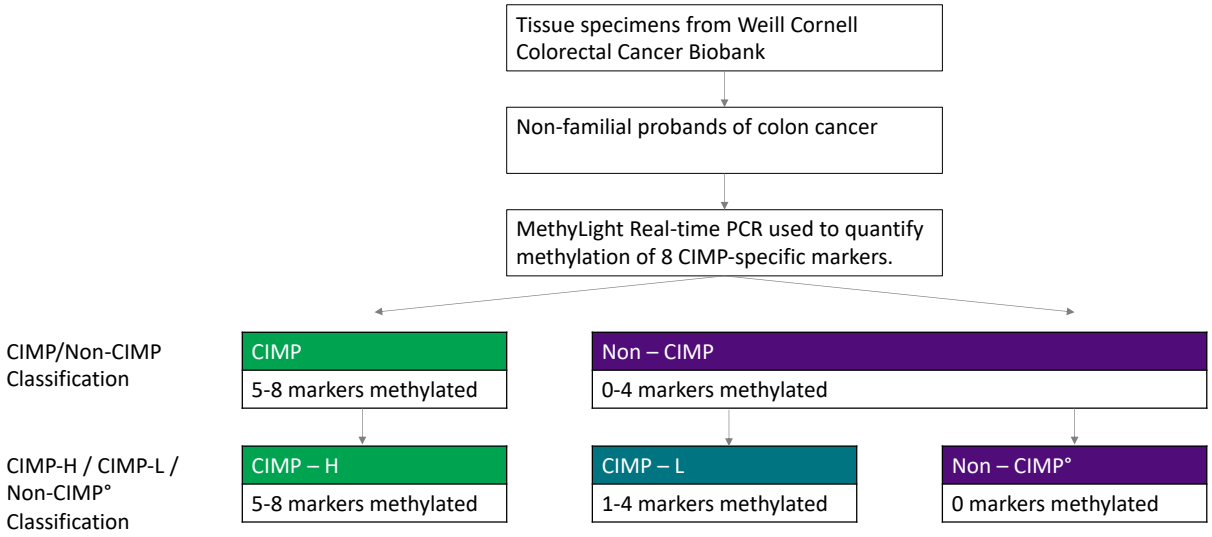


Figure 1: Study Sampling Scheme. *This figure also shows the distinction between the binary CIMP designation of CIMP/Non-CIMP and the categorical CIMP designation of CIMP-H, CIMP-L, and Non-CIMP°.*

DNA Extraction. AllPrep DNA/RNA Mini Kit (QIAGEN) was used to extract and purify DNA and RNA from each tissue sample.

Methylation Analysis. MethyLight real-time qPCR was performed on DNA obtained from both tumor and normal adjacent tissue samples once the biospecimens arrived at the Center for Infection and Immunity at Columbia University (CII). DNA from each sample was prepared for analysis with the EZ DNA Methylation Kit (Zymo Research). MethyLight Assays with the EpiTech MethyLight PCR Kit (QIAGEN) were used to determine the methylation status of each of eight CIMP-associated CpG islands (*CAGNA1G, CDKN2A, CRABP1, IGF2, MLH1, NEUROG1, RUNX3,* and *SOSC1*).[5,14] Markers with > 4% PMR (percent of methylation reference) were considered positive for methylation, as previously described.[5]

CIMP-Status. In the literature, there are two main classification systems used to categorize patients into subgroups based on the number of CIMP markers that are methylated in their tumor sample. Some

studies classify patients with five or more markers methylated as CIMP-High (CIMP-H), patients with one to four markers methylated as CIMP-Low (CIMP-L), and patients with zero markers methylated as Non-CIMP (Non-CIMP°). Other studies dichotomize patients by assigning those patients with five or more markers methylated as CIMP, and patients with four or fewer markers methylated as Non-CIMP. While all CIMP patients from the dichotomous designation are also CIMP-H in the CIMP-H/CIMP-L/Non-CIMP° designation, Non-CIMP patients from the CIMP/Non-CIMP system are further categorized into CIMP-L and Non-CIMP° in the CIMP-H/CIMP-L/Non-CIMP° classification scheme, depending on whether there are any markers methylated or no markers methylated in the sample. In this paper, we will distinguish between these two categorization systems using the naming convention of "Non-CIMP°" to indicate the group within the CIMP-H/CIMP-L/Non-CIMP° classification that has zero CIMP markers methylated, and "Non-CIMP" without the superscript to refer to the group within the dichotomous CIMP/Non-CIMP category that includes patients with anywhere from zero to four markers methylated.

***KRas/BRAF* Mutation Analyses.** PCR with Sanger Sequencing was performed on all tumor tissues. *BRAF V600E* primers were used to obtain *BRAF* mutation status, and *KRasG12-13* and *KRasQ61* primers were used to obtain *KRas* mutation status.[15–17]

Relative Abundance of Butyrate Producing Bacteria. The microbiome of tissue samples was evaluated by amplification of the bacterial v4 region of the 16S rRNA gene from tissue DNA and sequencing on an Illumina MiSeq. Resolved sequence variants (RSVs) and taxonomic analyses were carried out using Quantitative Insights into Microbial Ecology, version 2 (QIIME2).[18] Relative abundance for each bacterial taxa was calculated from raw abundance data. Our analysis included several bacterial genera known to be butyrate-producers. We included nine well known butyrate producing taxa in the human intestine, including one genera from the *Bacteroidia* class in the *Bacteroidetes* phylum (*Butyricimonas*), and several genera from the *Clostridia* class in the *Firmicutes* phylum (*Faecalibacterium,*

Roseburia, Coprococcus, Eubacterium, Blautia, Butyrivibrio, Anaerostipes, and *Pseudobutyrivibrio*).[10] Aggregated relative abundance of BPB was calculated by summing the relative abundances for each of these bacteria for each sample.

Confounder Selection. A recent systematic review and meta-analysis reported associations of CIMP-H CRC with female sex, older age, microsatellite instability (MSI), *BRAF* mutations, poor cellular differentiation, right-sided (proximal) tumor location, T3/T4 tumor staging, tumor-infiltrating lymphocytes (TIL), and high levels of *Fusobacterium nucleatum.*[6] As MSI, *BRAF* mutations, and poor cellular differentiation are initiating events or may be on the causal pathway between the tissue microenvironment and CIMP CRC, these variables were not considered as potential confounding variables, and were not controlled for in our analyses.

The physiology of the colorectum results in variation of the micro-environment from the cecum, located proximal to the small intestine and on the right side of the body, to the left-sided, distal colon and rectum. These variations can affect the growth and composition of the microbiome as location within the colon changes.[19] For example, one study found that CRC with high proportions of *Fusobacterium nucleatum* increased linearly along the length of the gastrointestinal tract, with the cecum having the highest proportion and the rectum having the lowest proportion of *Fusobacterium nucleatum*-high CRC.[20] This provides evidence to suggest that the relative abundance of other genera within the microbiome, including BPB may also vary by location within GI tract. Butyrate-producing bacteria have been shown to make up higher concentrations of the microbiome of the proximal colon, where there is an increased concentration of fermentable dietary fibers and decreased pH. As intestinal pH increases and the availability of fermentable fibers decreases towards the distal colon and rectum, the abundance of BPB decreases.[21,22] Thus, we hypothesized tumor location may confound the relationship between CIMP CRC and BPB, and adjusted for this variable in our statistical analyses.

The microbial community is also known to change as one ages. In elderly subjects, the gut microbiome is less diverse than the microbiome of younger subjects and is associated with a decrease of beneficial microorganisms including *Faecalibacterium prausnitzii* and *Clostridium* cluster XIVa, which contains members of the *Anaerostipes, Eubacterium, Butyrivibrio, Coprococcus,* and *Roseburia* genera, and a decrease in available SCFA.[10,23] We adjusted for age in our statistical models to account for the relationship between older age and CIMP, as well as older age and decreased BPB to avoid underestimating the relationship between BPB and CIMP.

The microbiome may also vary by sex in certain conditions, and CIMP has been reported to be more common in females.[6] One study found sex-specific differences in *Firmicutes/Bacteroidetes* ratio at the phyla level in addition to genus and species level differences, including a higher abundance of *Coprococcus catu*s in males compared to females.[24] Thus, patient sex was also selected as an important confounder to control for in our adjusted analyses.

PAM Cluster Analyses based on the Microbiota. The optimal number of clusters and membership of each sample into a cluster was determined with the *Partitioning Around Medoids (PAM)* clustering algorithm, based on bacterial genus-level relative abundances with the Bray-Curtis Dissimilarity Metric; the Silhouette Score was used to determine the optimal number of clusters.[25]

Statistical Analyses of Demographic and Clinical Variables. Demographic variables were analyzed to compare CIMP-H to CIMP-L patients, CIMP-H to Non-CIMP° patients, and CIMP-L to Non-CIMP° patients. Two-sided parametric hypothesis tests at α= 0.05 were conducted for all variables where cell counts were large enough to meet the assumptions of parametric methods. For MSI/MSS, tumor resection side, cancer stage, and *BRAF* mutation status, non-parametric tests were used since assumptions of sufficiently large sample size per cell count were violated. Age was analyzed as a

continuous variable. Sex (male vs. female), MSS status (MSI vs. MSS), resection side (left vs. right), and *BRAF/KRas* mutations (mutant vs wild type) were analyzed as dichotomous variables. Four patients were missing data on MSS/MSI status, five were missing data on resection side, and twenty were missing data on cancer stage, and were excluded from their respective univariate analyses.

Information on each patient's TNM staging was recorded and used to categorize patients by approximating their American Joint Committee on Cancer (AJCC) Cancer Stage.[26] Patients that were positive for metastasis, regardless of tumor or lymph node status, were coded as AJCC Stage IV. Patients that had negative or missing values for metastasis, but were positive for lymph nodes, were coded as AJCC Stage III. All other patients were coded as AJCC Stage I/II. Non-parametric tests were used to compare between CIMP-H, CIMP-L, and Non-CIMP° at α = 0.05, due to low prevalence of Stage IV cancers in the study sample.

The prevalence of positive methylation for each CIMP-associated marker was also calculated, and compared across CIMP-H, CIMP-L, and Non-CIMP° using non-parametric methods with α = 0.05, though only hypotheses tests comparing CIMP-H to CIMP-L were conducted, as Non-CIMP° patients were negative for all markers, by definition.

LEfSe (Linear discriminant analysis [LDA] effect size) for Identification of Biomarkers. LEfSe was used to examine the relationship between bacterial relative abundance and CIMP subgroups. LEfSe identifies genomic features that are differentially abundant between groups and is designed to aid in high-dimensional biomarker discovery and explanation while reducing the rate of false positives (Type I error).[27–29] Since LEfSe analyses have been shown to effectively reduce the false discovery rate, these analyses served as a gatekeeping procedure, where only taxa or genomic pathways identified by LEfSe as significant biomarkers of a group or subgroup of interest were analyzed further through additional

analyses including individual regression analyses that account for confounding variables, statistical tests that account for matching, and confirmatory topological data analyses. Comparisons for the purpose of gatekeeping included the following analyses 1) CIMP-H tumors versus CIMP-H normal tissues, 2) CIMP-L tumors versus CIMP-L normal tissues, 3) Non-CIMP° tumor tissues versus Non-CIMP° normal tissues, 4) CIMP-H tumors versus CIMP-L tumors versus Non-CIMP° tumors, and 5) CIMP-H normal tissues versus CIMP-L normal tissues versus Non-CIMP° normal tissues. For any genera determined to be significant in one or more of these comparisons, the likelihood of false rejection of the null hypothesis and false positives will only decrease when additional analyses that adjust for matching or confounding are conducted on these subsets of bacterial genera. LEfSe was also used to compare groups determined by PAM clustering analyses (cluster1 vs cluster2) in order to determine the genera driving separation between these clusters.

Regression Analyses. For each sample, the relative abundance of each of the ten BPB was summed to calculate a total relative abundance of BPB ("Total BPB"). Two separate multinomial logistic regression models, one for tumor tissues and one for normal adjacent tissues, were considered with CIMP-H, CIMP-L, and Non-CIMP° as the outcome and Total BPB as the predictor. However, due to limited sample size and non-significant differences in clinical and demographic characteristics aside from methylation status, we decided to collapse our categories into the binary outcome of CIMP and Non-CIMP for the purpose of our regression analyses.

Two separate crude binomial logistic regression models were run, one for tumor tissues and one for normal-adjacent tissues and were fit with CIMP status as the outcome and Total BPB as the predictor. Potential confounders were examined, and a final adjusted model was fit for each tissue type. Since the sample-size for the CIMP group was too small to be considered sufficiently powered for the fully adjusted model, we refit the fully adjusted model to 1000 bootstrapped samples and calculated the

mean and standard error of each coefficient from the bootstrapped samples to test the robustness of our regression model results. Additionally, we fit both a crude and adjusted relative risk regression model to estimate the relative risk of CIMP, since CIMP status was common in our study population. We also fit ten binary logistic regression models to test each of the ten individual BPB as predictors of CIMP status. All univariate and regression analyses were completed using R 3.5.1 "Feather Spray" and R Studio version 1.1.456.

Topological Data Analyses. Several topological models were created using Ayasdi, an enterprise-level machine learning platform.[30,31] The dataset used in this platform included 39 clinical and demographic variables, 16 methylation variables, PAM cluster membership, 3 alpha diversity metrics, bacterial relative abundance data at each taxonomic level (including 2669 resolved sequence variants [RSVs], 239 bacterial taxa at the genus level, 131 at the family level, 79 at the order level, 43 at the class level, and 24 at the phylum level), as well as PICRUSt data including 9 variables at level 1 (L1), 40 at level 2 (L2), and 281 at level 3 (L3). The first model (Topological Model 1/ TM1) was built using the top 100 genera with the highest average relative abundance across all 184 samples using the Euclidean Distance (L2) metric and Neighborhood 1 & 2 lenses at a resolution of 30 and a gain of 4.0 for each lens. The second set of models were built for each of the CIMP-H (Topological Model 2 /TM2, with 34 patients and 68 samples), CIMP-L (Topological Model 3/ TM3, with 38 patients and 76 samples), and Non-CIMP° (Topological Model 4/ TM4, with 20 patients and 40 samples) subsets, individually, using the top 100 genera with the highest average relative abundance. These networks were generated using the Euclidean Distance (L2) metric and Neighborhood 1 & 2 lenses at a resolution of 30 and a gain of 3.6 for each lens. Clinicopathological data, diversity metrics, methylation marker PMR, and functional genomic data, along with relative abundance data at other phylogenetic levels including phylum, class, order, family, and ribosomal sequence variants (RSV) were included in the dataset to allow the platform to

evaluate associations between the networks derived based on bacterial genus-level relative abundance and clinical and biomarker data, though the networks were not trained on these features of the data. Networks and subnetworks in each model were identified and tested for significance based on the Kolmogorov-Smirnov test, with p-values < 0.05 considered statistically significant.

In addition to topological networks, we performed supervised comparisons between CIMP-H and CIMP-L, CIMP-H and Non-CIMP°, and CIMP-L and Non-CIMP° to look for significant difference between subgroups. We repeated these comparisons twice, using only tumor tissues in one comparison, and only normal tissues in the other. Kolmogorov-Smirnov p-values < 0.05 were considered statistically significant for between-group comparisons. We also looked for differences between paired tumor and normal tissues from each patient within each of the three CIMP status designations, and considered Wilcoxon signed rank p-values < 0.05 as statistically significant for paired within-group comparisons.

III. RESULTS

Methylation Status. CIMP status (either CIMP or Non-CIMP) as predicted by positive MLH1 status was used to select samples in an effort to obtain a roughly even split between CIMP and Non-CIMP samples in this study. At the time of this analysis, tumor and normal tissue samples from 92 patients had been obtained from the Weill Cornell Gastrointestinal Cancer Biobank, 36 of which were classified as CIMP due to MLH1 methylation, 38 of which were classified as Non-CIMP due to lack of MLH1 methylation, and 18 of which had an unknown or missing CIMP classification by this assessment. Real-time MethyLight PCR was used to quantify the methylation of each of the 8 CpG island markers that are commonly used to determine CIMP status in tumor tissues from all 92 patients **(Figure 2).**

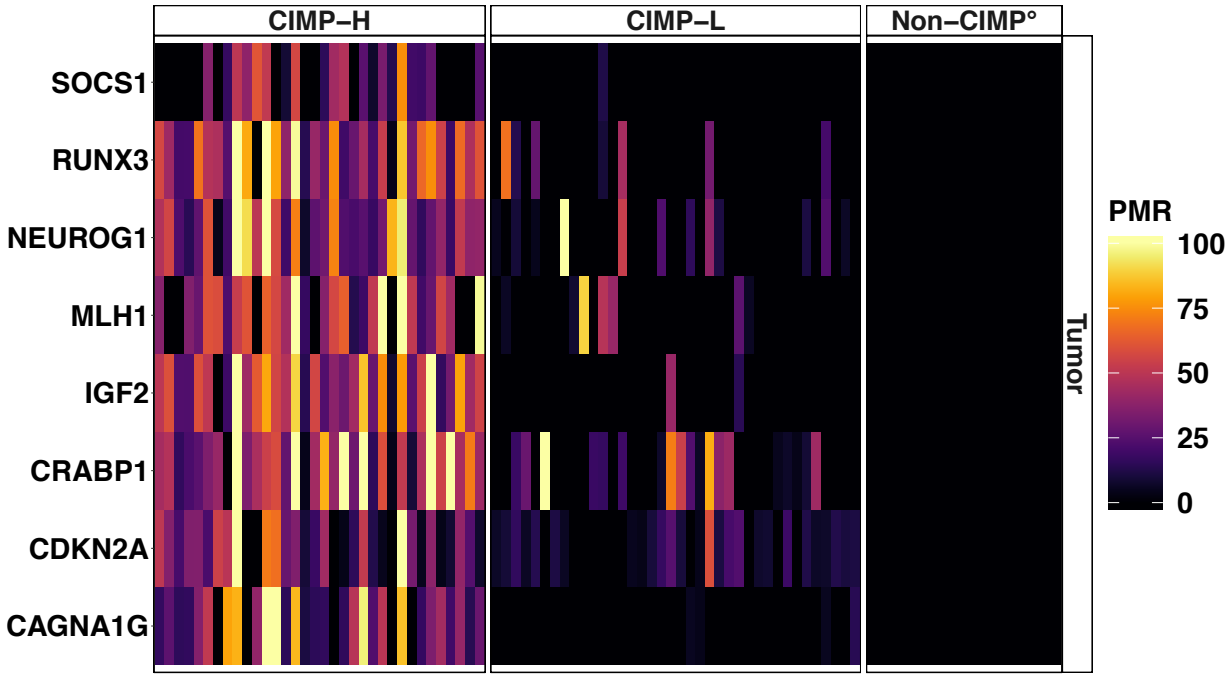


Figure 2: Methylation in Tumor Tissues. Percent of Methylation Marker (PMR) for each CIMP-specific methylation marker is shown for each subject's tumor tissue sample. Markers with PMR < 4% are considered to have 0% methylation.

According to our MethyLight analyses, 34 patients had five or more markers methylated and were classified as CIMP. The remaining 58 patients were determined to be Non-CIMP, with 38 patients classified as CIMP-L, and 20 patients classified as Non-CIMP°. Sensitivity and specificity analyses were run to determine the accuracy of the pathology-estimated CIMP status to the MethyLight status ascertained at the CII. Pathology estimation resulted in a sensitivity of 92.86% and a specificity of 78.26% after excluding samples for which CIMP classification was missing from pathology reports. Fisher's exact tests were used to test for differences in proportion of positive CIMP methylation markers between CIMP-H tumor tissues and CIMP-L tumor tissues. As expected, the proportion of positive methylation in tumor samples for each of the 8 markers was higher in the CIMP-H group compared to the CIMP-L group, and these differences were statistically significant for each marker **(Table 1).**

Table 1: Prevalence of Positive CIMP Markers in Tumor Tissues by CIMP Designation

	CIMP-H		**CIMP-L**		**Non-CIMP°**		**All Samples**	**CIMP-H vs CIMP-L**
	n	*%*	n	*%*	n	*%*	n	
CAGNA1G	29	*(85.29%)*	4	*(10.53%)*	0	*(0.00%)*	33	**p < 0.0001**
CDKN2A	30	*(88.24%)*	27	*(71.05%)*	0	*(0.00%)*	57	**p < 0.0001**
CRABP1	33	*(97.06%)*	19	*(50.00%)*	0	*(0.00%)*	52	**p < 0.0001**
IGF2	33	*(97.06%)*	2	*(5.26%)*	0	*(0.00%)*	35	**p < 0.0001**
MLH1	27	*(79.41%)*	7	*(18.42%)*	0	*(0.00%)*	34	**p < 0.0001**
NEUROG1	34	*(100.0%)*	12	*(31.58%)*	0	*(0.00%)*	46	**p < 0.0001**
RUNX3	33	*(97.06%)*	7	*(18.42%)*	0	*(0.00%)*	40	**p < 0.0001**
SOCS1	20	*(58.82%)*	1	*(2.63%)*	0	*(0.00%)*	21	**p < 0.0001**

P-values calculated using Fischer's Exact Test for Count Data, and adjusted for multiple testing using Benjamini & Hochberg methods.

Descriptive Statistics. Characteristics of the CIMP-H, CIMP-L, and Non-CIMP° groups are shown in **Table 2**. Of the tumor samples for CIMP-H patients, 82.35% exhibited microsatellite instability, compared to 21.05% in CIMP-L patients and 20.00% in Non-CIMP° patients. Of CIMP-H tumor samples, 85.29% carried *BRAF* mutations, compared to 0.00% in the CIMP-L tumor samples and 5.00% in the Non-CIMP° samples. For both MSI and *BRAF* mutations, the difference in proportions was statistically significant when comparing CIMP-H to CIMP-L patients, and when comparing CIMP-H to Non-CIMP° patients, but not when comparing CIMP-L to Non-CIMP°. While there appeared to be a higher prevalence of *KRas* mutations in CIMP-L and Non-CIMP° patients, only the difference between CIMP-H and CIMP-L were statistically significant. However, it should be noted that one CIMP-H patient carried both the *BRAF* and *KRas* mutations. When considering only patients with *KRas* mutations in the absence of any *BRAF* mutations, the difference between CIMP-H and Non-CIMP° was also significant.

Patients in the CIMP-H group were also significantly older at the time of tumor removal compared to patients in the CIMP-L group and compared to patients in the Non-CIMP° group, with an average age of 74.12 for CIMP-H patients, 64.74 for CIMP-L patients, and 63.55 for Non-CIMP° patients.

Table 2: Summary Statistics of Clinical and Demographic Patient Characteristics

Methylation markers	**CIMP-H 5-8**		**CIMP-L 1-4**		**Non-CIMP° *0***		**All Samples**	**CIMP-H vs CIMP-L**	**CIMP-H vs Non-CIMP°**	**CIMP-L vs Non-CIMP°**
Patients *(n)*	34		38		20		92			
Age[1] *(average)*	74.12		64.74		63.55		67.95	**p = 0.0046**	**p = 0.0040**	p = 0.7541
Sex [2] *(n, %)*								p = 0.7023	p = 0.0893	p = 0.2253
Females	23	*(67.65%)*	23	*(60.53%)*	8	*(40.00%)*	54			
Males	11	*(32.35%)*	15	*(39.47%)*	12	*(60.00%)*	38			
MSI/MSS[3] *(n, %)*								**p < 0.0001**	**p < 0.0001**	p = 1.0000
MSI	28	*(82.35%)*	8	*(21.05%)*	4	*(20.00%)*	40			
MSS	5	*(14.71%)*	30	*(78.95%)*	13	*(65.00%)*	48			
Missing	1	*(2.941%)*	0	*(0.00%)*	3	*(15.00%)*	4			
Resection Side[3] *(n, %)*								p = 0.5559	p = 0.4799	p = 1.0000
Left	5	*(14.71%)*	9	*(23.68%)*	5	*(25.00%)*	19			
Right	25	*(73.53%)*	29	*(76.32%)*	14	*(70.00%)*	68			
Missing	4	*(11.76%)*	0	*(0.00%)*	1	*(5.00%)*	5			
CRC Cancer Stage[3] *(n, %)*								p = 0.4034	p = 0.8726	p = 0.8206
Stage 1/ Stage 2	18	*(52.94%)*	15	*(39.47%)*	9	*(45.00%)*	42			
Stage 3	8	*(23.53%)*	10	*(26.32%)*	6	*(30.00%)*	24			
Stage 4	1	*(2.94%)*	4	*(10.53%)*	1	*(5.00%)*	6			
Missing	7	*(20.59%)*	9	*(23.68%)*	4	*(20.00%)*	20			
***BRAF* Mutation**[3] *(n, %)*								**p < 0.0001**	**p < 0.0001**	p = 0.3448
Mutant Type	29	*(85.29%)*	0	*(0.00%)*	1	*(5.00%)*	30			
Wild Type	5	*(14.71%)*	38	*(1.00%)*	19	*(95.00%)*	62			
***KRas* Mutation**[3] *(n, %)*								**p = 0.0146** **p* = 0.0030**	p = 0.0769 **p* = 0.0281**	p = 0.7833 p* = 0.7833
Mutant Type*	4*	*(11.76%)*	15	*(39.47%)*	7	*(35.00%)*	26			
Wild Type	30	*(88.24%)*	23	*(60.53%)*	13	*(65.00%)*	66			
Tumor Infiltrating Lymphocytes[3] *(n, %)*								p = 0.2758	**p = 0.0086**	p = 0.0827
Present	10	*(29.41%)*	7	*(18.42%)*	0	*(0.00%)*	17			
Absent	23	*(67.65%)*	31	*(81.58%)*	19	*(95.00%)*	73			
Missing	1	*(2.94%)*	0	*(0.00%)*	1	*(5.00%)*	2			

1. *Indicates that a Two-Sample Student's T-test was used to test for significance.*
2. *Indicates that a Pearson's Chi-Square Test was used to test for significance.*
3. *Indicates that a Fischer's Exact Test was used to test for significance.*

* *One CIMP-H patient had both KRAS and BRAF mutations. We also tested for significance for difference in proportions of patients with only KRAS mutations (i.e., excluding the patient with both KRAS and BRAF mutations). P-values for this test are designated with an asterisk (*) by their p-values.*

Differences in ages between CIMP-L and Non-CIMP° groups were not significant. Tumor Infiltrating Lymphocytes were significantly associated with CIMP-H compared to Non-CIMP° patients, but were not significant in other comparisons. We also examined differences in proportions between CIMP designations for sex, resection side, and cancer stage, but these results were not statistically significant between any groups **(Table 2).**

Alpha Diversity. The Shannon Diversity Index (Shannon), Faith's Phylogenetic Diversity (Faith PD), and Observed RSVs were calculated for each sample using ten iterations at a rarefaction depth of 2694 sequences. The ten iterations for each sample at each sequencing depth were averaged and plotted in rarefaction curves **(Supplemental Figure 1).** At the rarefaction depth of 2694, the median value of samples' average Faith PD index was 7.53 in normal tissues and 6.49 in tumor tissues, the median value of sample's average Shannon index was 4.60 in normal tissues and 3.93 in tumor tissues, and the median value of the sample's average Observed RSVs was 103.65 in normal tissues and 81.30 in tumor tissues.

When comparing samples of different CIMP groups, we only found a statistically significant difference in the Shannon Diversity Index, comparing CIMP-H tumor tissues to CIMP-L tumor tissues (Mann-Whitney p-value = 0.0435*)*; there was not a statistically significant difference in Shannon Diversity Index comparing CIMP-H tumor tissues to Non-CIMP° tumor tissues, nor was there a statistically significant difference between CIMP-H tumor tissues and CIMP-L tumor tissues for Observed RSVs or Faith PD. No other comparison was significant across any alpha diversity metric for any other between-CIMP group comparisons. Results of Observed RSVs comparisons can be visualized in **Figure 3A.**

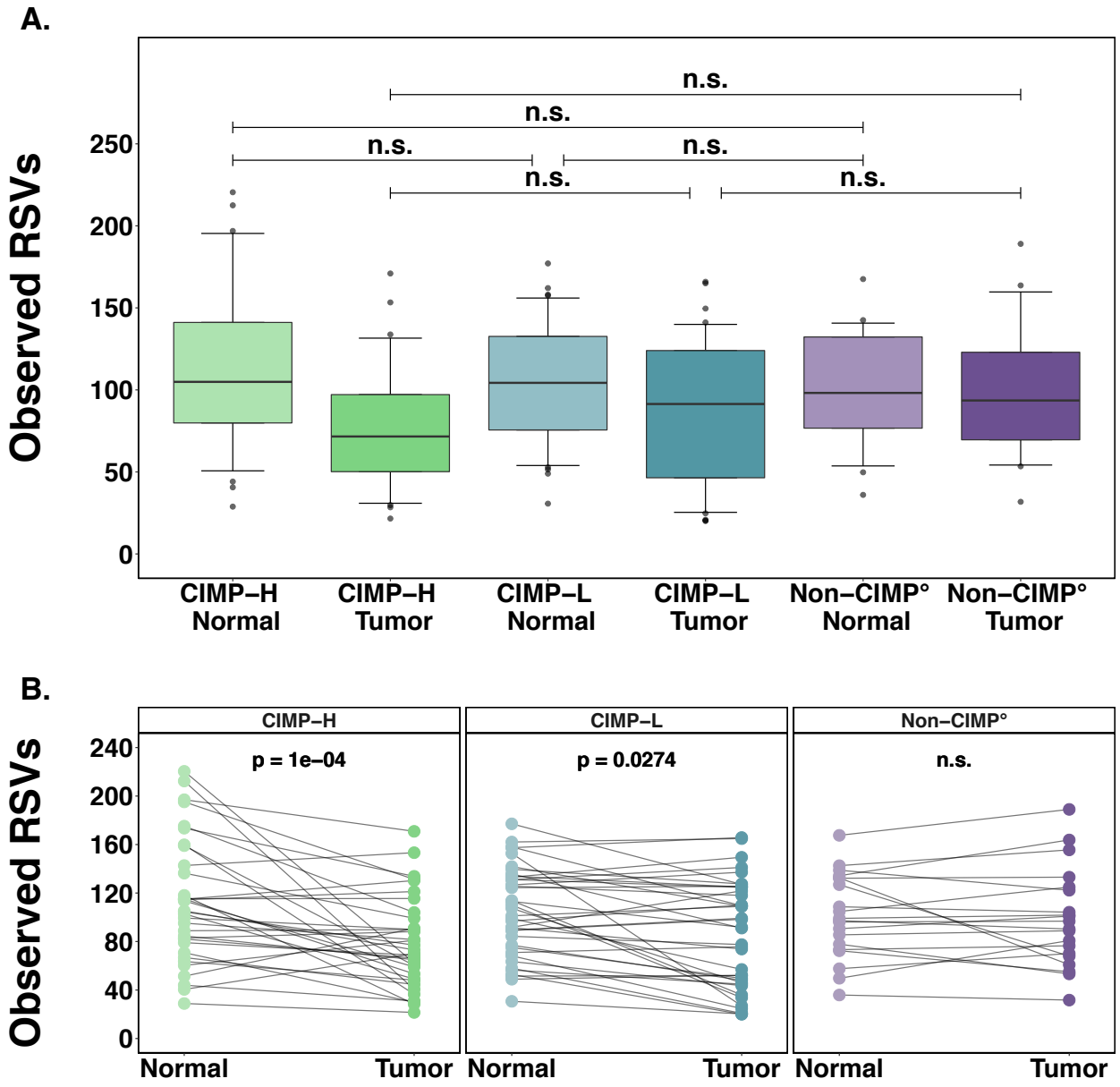


Figure 3: Within and Between Group Comparisons for Observed RSVs. *Figure 3A shows boxplots of observed RSVs by CIMP designation and tissue type. Whiskers represent the 10th to 90th percentile of data in that group, and outliers are represented by points. Mann-Whitney p-values are shown comparing between different groups of patients. Figure 3B shows spaghetti plots of matched tumor and normal tissue observed RSVs, stratified by CIMP designation. Wilcoxon paired p-values are shown.*

Interestingly, when comparing matched tumor and normal samples, our results differed by CIMP status. Among CIMP-H patients, normal tissues were significantly higher than tumor tissues for all three alpha diversity indices (Shannon Wilcoxon Signed-Rank p-value < 0.0001; Observed RSVs Wilcoxon Signed Rank p-value = 0.0001; Faith PD Wilcoxon Signed-Rank p-value = 0.0003), with the median difference of 33 fewer observed RSVs in tumor tissues, compared to normal tissues in CIMP-H patients. In CIMP-L tissues, normal tissues were also significantly higher than matched tumor tissues for all three alpha diversity metrics (Shannon Wilcoxon Signed Rank p-value = 0.0184; Observed RSVs Wilcoxon Signed Rank p-value = 0.0274 Faith PD Wilcoxon Signed Rank p-value = 0.0051). The median decrease in Observed RSVs from normal tissues to tumor tissues in CIMP-L patients was 13 taxa. However, for Non-CIMP° patients, there was no significant difference between tumor and normal samples for any of the alpha diversity indices (Shannon Wilcoxon Signed-Rank p-value = 0.2774; Observed RSVs, Wilcoxon Signed Rank p-value = 0.9854 Faith PD, Wilcoxon Signed Rank p-value = 0.9854). Results of the Observed RSVs paired analyses can be viewed in **Figure 3B.**

Beta Diversity. Principal Coordinate Analysis (PCoA) plots based on Bray-Curtis dissimilarly at the genus level, stratified by tumor and normal tissues, can be viewed in **Figure 4**. These plots show dissimilarity between each sample, where points closer to each other are more similar, while points further away from each other are dissimilar. Bray-Curtis dissimilarities between tumor tissues and normal tissues were significant overall (Permanova Pseudo-F = 2.60, p-value = 0.0010, q-value = 0.0010). Between group comparisons revealed that Bray-Curtis dissimilarity was significant between tumor tissues from CIMP-H and both CIMP-L and Non-CIMP°. Furthermore, marginal differences in beta-diversity were observed between CIMP-L and Non-CIMP° tumors. There was no statistically significant dissimilarity between the three CIMP groups in normal tissues **(Table 3)**.

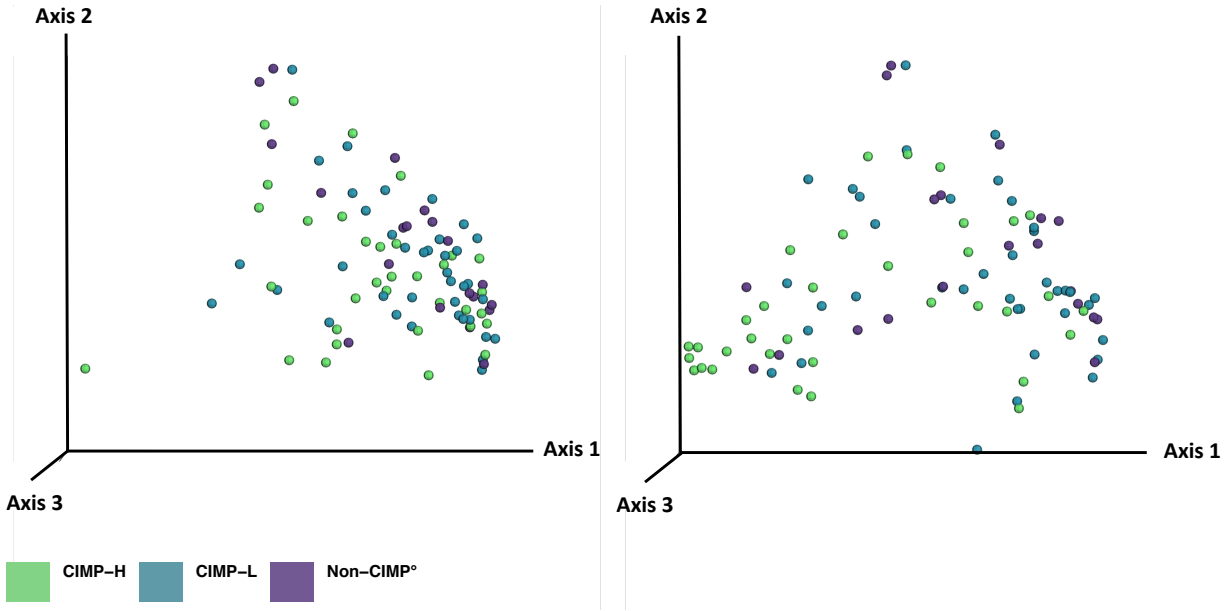


Figure 4: Bray-Curtis PCoA Plots. *The plot on the left shows only normal tissues, colored by CIMP designation. The plot on the right shows only tumor tissues, colored by CIMP designation. For both tumor and normal tissue plots displayed, principal coordinate analysis was run using all 184 samples at the genus level, using the Bray-Curtis Dissimilarity metric. Points that are plotted close together can be interpreted as similar to each other, while points plotted further away from each other are less similar. Axis 1 represents 22.67% of the variance among all samples. Axis 2 represents 10.29% of the variance among all samples. Axis 3 represents 8.22% of the variance among all samples.*

Table 3: Results of Bray-Curtis Pair-wise Permanova Analysis

	Tumor Tissues			Normal Tissues		
	Pseudo-F	**p-value**	**q-value**	**Pseudo-F**	**p-value**	**q-value**
CIMP-H vs CIMP-L	2.13	p = 0.0030	q = 0.0045	0.90	p = 0.5890	q = 0.5890
CIMP-H vs Non-CIMP°	2.42	p = 0.0020	q = 0.0045	1.11	p = 0.2940	q = 0.5610
CIMP-L vs Non-CIMP°	1.45	p = 0.0500	q =0.0500	1.03	p = 0.3740	q = 0.5610
Calculated using 999 permutations.						

PAM Clustering at the Genus Level. PAM clustering analysis based on the genus-level Bray-Curtis dissimilarity supported the existence of two clusters. Prevalence of cluster membership by CIMP designation, stratified by tissue type, can be viewed in **Table 4** and **Figure 5.** Overall, 79.89% of our 184

samples belonged to Cluster 1, while 20.11% of our 184 samples belonged to Cluster 2. PCoA plots in **Figure 5** show Bray-Curtis dissimilarities between PAM clusters.

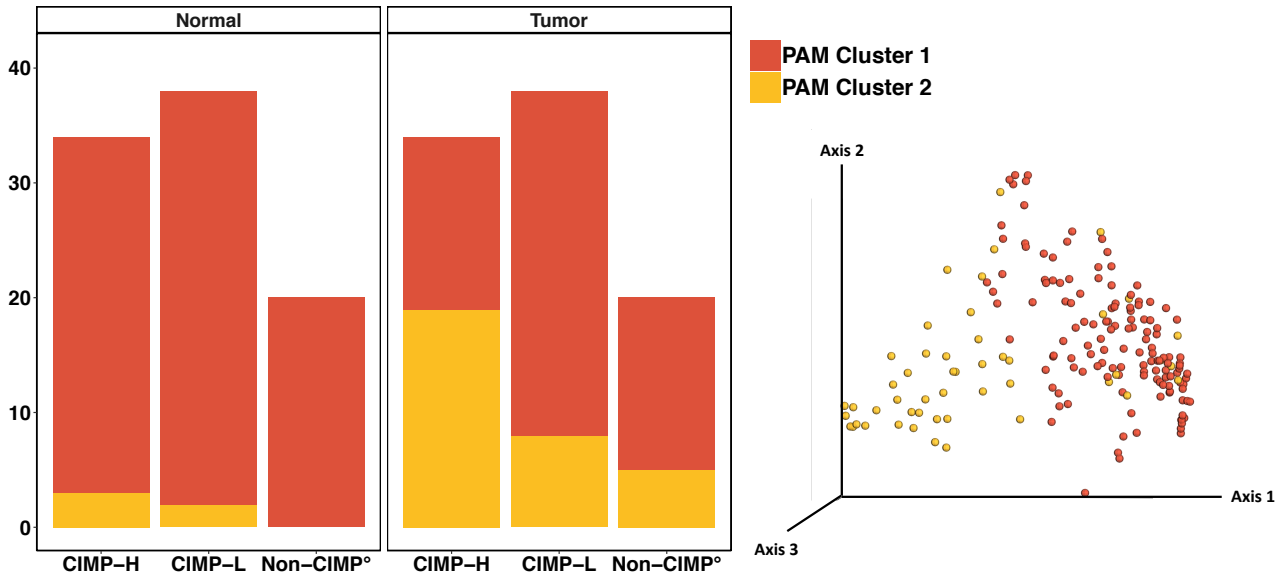


Figure 5: Results of PAM Cluster Analysis. *Clusters were created using Bray-Curtis Dissimilarity at the genus level, using the partitioning around medoids (PAM) clustering algorithm. The left plot shows membership into the two PAM cluster groups by CIMP designation, stratifying by normal and tumor tissue types. The bottom plot show the same Bray-Curtis PCoA plots as are shown in Figure 4, colored by cluster membership. Axis 1 represents 22.67% of the variance among all samples. Axis 2 represents 10.29% of the variance among all samples. Axis 3 represents 8.22% of the variance among all samples.*

Table 4: Results of PAM Cluster Analysis

	Tumor Tissues				**Normal Tissues**			
	Cluster 1		**Cluster 2**		**Cluster 1**		**Cluster 2**	
	n	(%)	n	(%)	n	(%)	n	(%)
Overall Prevalence, by Tissue Type	60	(65.22%)	32	(34.78%)	87	(94.57%)	5	(5.43%)
Prevalence by CIMP designation								
CIMP-H	15	(44.11%)	19	(55.88%)	31	(91.17%)	3	(8.82%)
CIMP-L	30	(78.94%)	8	(21.05%)	36	(94.74%)	2	(5.26%)
Non-CIMP°	15	(75.00%)	5	(25.00%)	20	(100.0%)	0	(0.00%)

Regression Analysis of PAM Clusters. Since the demographic characteristics of CIMP-L and Non-CIMP° patients were relatively similar and were not significantly different for any demographic variables or prevalence in PAM Cluster 1 or Cluster 2 in our univariate analyses, we used binary logistic regression, with CIMP status as the outcome, to model the relationship between Cluster 2 membership, relative to Cluster 1 membership, and CIMP status in our binary CIMP/Non-CIMP Designation. Separate models were run for tumor and normal tissues. Results of this model can be viewed in Model A of **Table 5.**

Table 5: Results of Regression Analyses with Bray-Curtis Cluster as the predictor of CIMP status (Cluster 2 vs Cluster 1 reference)

	Tumor Tissues			**Normal Tissues[2]**		
	Estimate and 95% CI		**p-value**	**Estimate and 95% CI**		**p-value**
Model A. Binary CIMP Logistic Regression[1]						
Crude Model[2]	OR = 4.39	(1.75, 10.96)	**p = 0.0016**	OR = 2.68	(0.29, 33.69)	p = 0.3547
Adjusted Model[3]	OR = 7.70	(2.53, 23.44)	**p = 0.0003**			
Bootstrapped Adjusted Model[4]	OR = 8.92	(2.47, 32.21)	-			

1. *CIMP status from CIMP/Non-CIMP classification is modeled as the outcome of interest.*
2. *Due to low cell counts, only crude Fisher's Exact test was run for Normal Tissues.*
3. *Model adjusted for age in years (continuous), sex (male or female), and resection side (left or right).*
4. *OR and CI from averaging beta coefficients of adjusted regression models from 1000 bootstrapped samples.*

Our crude model found that on average, the odds of CIMP classification was 4.39 times higher for tumor samples in Cluster 2 when compared to tumor samples in Cluster 1 (Crude OR 95% CI: 1.75, 10.96, p-value = 0.0016). After adjusting for age, sex, and resection type, the odds of CIMP classification was 7.70 times higher for tumor samples in Cluster 2, compared to tumor samples in Cluster 1 (Adj. OR 95% CI: 2.53, 23.44, p-value = 0.0003). Since our sample size is small enough that we may be underpowered to adjust for all three of these covariates, we bootstrapped 1000 samples with replacement from our study sample, ran our adjusted model on each of these samples, and calculated our average OR and bootstrapped standard error as a sensitivity test to see how robust our results were for our fully adjusted model. Our bootstrapped estimate for the odds ratio for CIMP was 8.92, comparing tumors in Cluster 2 to tumors in Cluster 1, and our 95% CI did not contain our null value of 1 (Bootstrapped OR 95% CI: 2.47, 32.21).

Due to low prevalence of samples belonging to Cluster 2 in our normal tissue subset, Fisher's Exact Test, rather than logistic regression, was used to calculate the crude OR. Normal tissue samples in Cluster 2 had 2.68 times the odds of being from patients with CIMP compared to Cluster 1, though this result did not reach significance (Crude OR 95% CI: 0.29, 33.69, p-value = 0.3547). We had insufficient power to calculate an adjusted OR for normal tissues with cluster membership as the predictor variable. Full results of this analysis can be viewed in **Table 5.**

LDA Effect Size Analysis (LEfSe) to Examine Differences in Bacterial Taxa between PAM Clusters. LDA Effect Size analysis was used to determine which genera were driving the separation of samples into these unsupervised PAM clusters. All genera that are significant markers of cluster membership can be viewed in **Figure 6.** The top genera that are associated with Cluster 1 included *Bacteroides, Ruminococcus,* and *Escherichia,* among other genera, including *Faecalibacterium, Blautia,* and *Roseburia,* which are BPB. In Cluster 2, we found that the top genera associated with Cluster 2 included *Fusobacterium, Leptotrichia, Alishewanella, Selenomonas, and Campylobacter.*

LDA Effect Size Analysis (LEfSe) to Examine Differences by CIMP Designation and Tissue Type

We also applied LEfSe analyses to determine which genera were markers of CIMP tumors, CIMP normal tissues, Non-CIMP tumors, and Non-CIMP normal tissues **(Figure 7).** *Fusobacterium, Campylobacter,* and *Leptotrichia* were all significant markers of the CIMP tumor group, while several bacterial genera, including *Bacteroides, Ruminococcus,* and some BPB including *Faecalibacterium, Blautia,* and *Coprococcus* were significant markers of the Non-CIMP normal tissue group. There were no significant markers that distinguished CIMP normal tissues or Non-CIMP tumor tissues from the other groups in this four-group comparison.

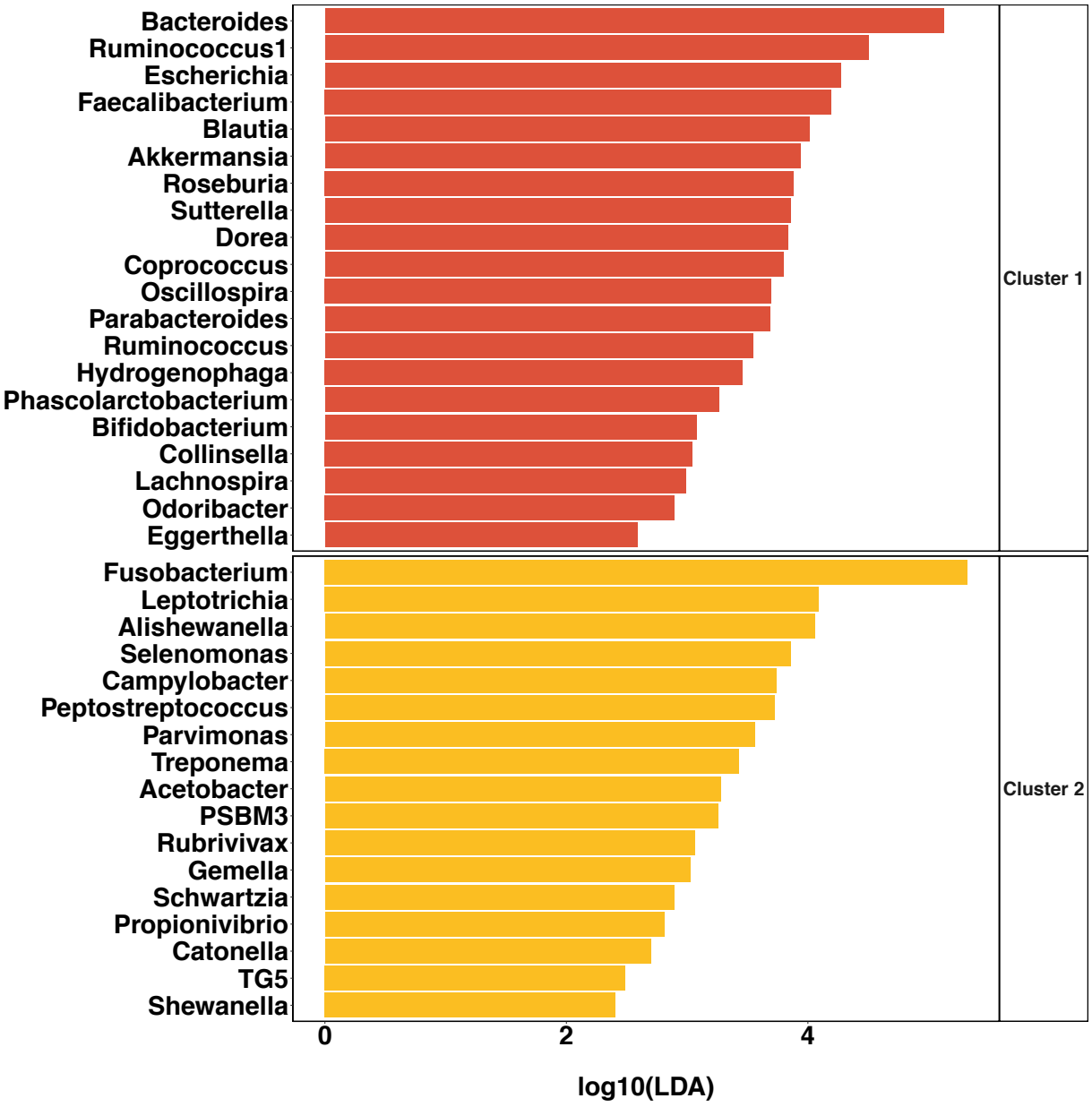


Figure 6: Results of the LEfSe Analysis at the genus level. *This figure shows all genera and their respective log10(LDA) scores for all genera that had log10(LDA) scores above 2, indicating they significantly distinguished between PAM Cluster 1 and PAM Cluster 2.*

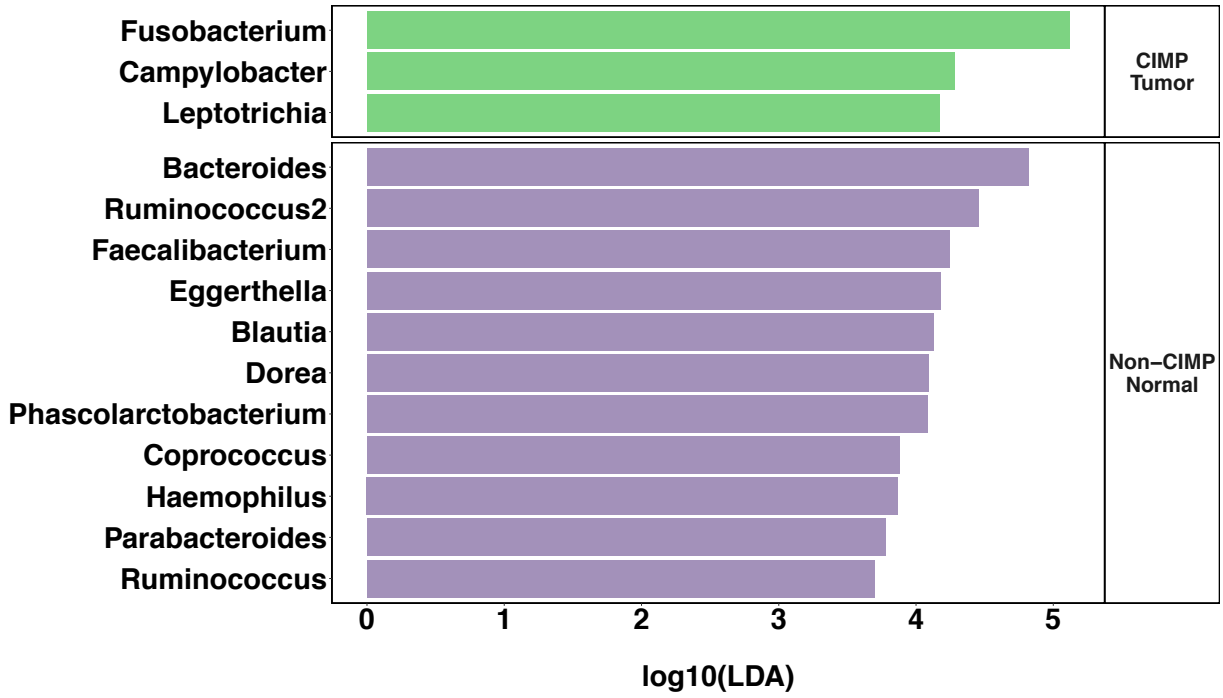


Figure 7: CIMP Tumor vs CIMP Normal vs Non-CIMP Tumor vs Non-CIMP Normal LEfSe Results. *This figure shows all genera and their respective LDA scores for all genera that had LDA scores above 2, indicating they significantly distinguished between CIMP tumors, CIMP normal tissues, Non-CIMP tumors and Non-CIMP normal tissues.*

To determine if the genera distinguishing tumors from normal tissues varies by CIMP subtypes, we ran LEfSe analysis to compare normal and tumor tissues in CIMP-H, CIMP-L, and Non-CIMP° subgroups. While the relative abundance of *Fusobacterium, Campylobacter,* and *Leptotrichia* were enriched in tumor tissues compared to normal tissues among CIMP-H patients, the levels of various bacteria, including Bacteroides and several BPBs had lower relative abundance in CIMP-H tumor tissues compared to CIMP-H normal tissue **(Figure 8)**. The relative abundance of *Fusobacterium* and *Sphingomonas* were enriched in tumor tissues compared to normal tissues in CIMP-L patients and fewer bacterial taxa were reduced in CIMP-L tumor tissue compared to CIMP-L normal tissue **(Figure 9).** Interestingly, in Non-CIMP° patients, there were no genera with log10(LDA) scores > 2 that distinguished tumor tissues from normal tissues.

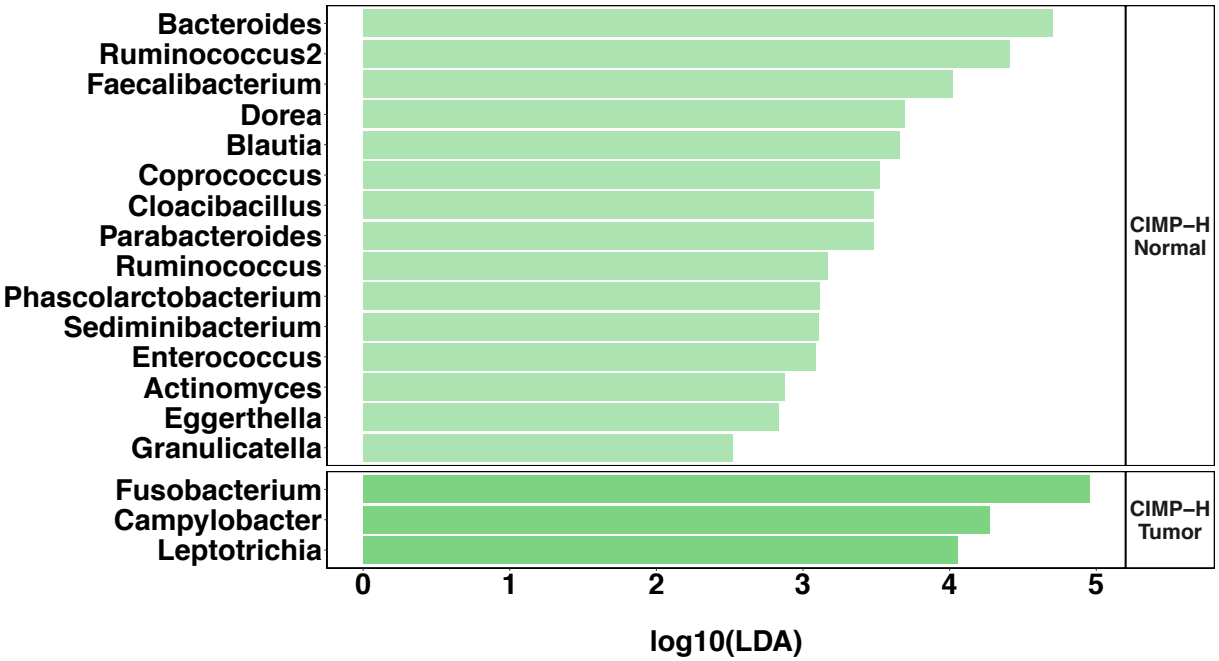


Figure 8: CIMP-H Tumor vs CIMP-H Normal LEfSe Results. *This figure shows all genera and their respective log10(LDA) scores for all genera that had log10(LDA) scores above 2, indicating they significantly distinguished between CIMP-H tumor and normal tissues.*

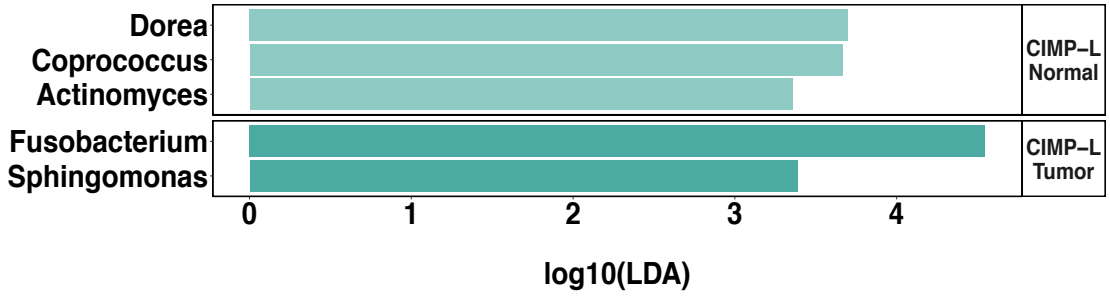


Figure 9: CIMP-L Tumor vs CIMP-L Normal LEfSe Results. *This figure shows all genera and their respective log10(LDA) scores for all genera that had log10(LDA) scores above 2, indicating they significantly distinguished between CIMP-L tumor and normal tissues.*

Bacterial Genera Identified by LDA Effect Size Analysis (LEfSe) Gatekeeping Methods. In addition to the bacterial genera identified as significant biomarkers in LEfSe analyses of CIMP-H tumor versus CIMP-H normal tissues **(Figure 8)**, CIMP-L tumor versus CIMP-L normal tissues **(Figure 9)**, and Non-CIMP° tumor tissues versus Non-CIMP° normal tissues (data not shown; no significant genera found), we also ran

LEfSe analyses comparing CIMP-H tumor versus CIMP-L tumor versus Non-CIMP° tumor tissues **(Supplemental Figure 2)**, and CIMP-H normal versus CIMP-L normal versus Non-CIMP° normal tissues **(Supplemental Figure 3)**. A full list of genera identified as significant biomarkers through LEfSe analyses can be viewed in **Table 6**.

Table 6: LEfSe Identified Biomarkers at the Genus Level

Genera				
Actinobacillus	**Colinsella**	**Fusobacterium**	**Pseudobutyrivibrio**	**Sphingomonas**
Actinomyces	**Corprobacillus**	**Granulicatella**	**Pseudoramibacter Eubacterium**	**Sutterella**
Alloscardovia	**Coprococcus**	**Haemophilus**	**Ruminococcus**	**Treponema**
Bacteroides	**Dorea**	**Leptotrichia**	**[Ruminococcus]**	
Blautia	**Eggerthella**	**Odoribacter**	**Schwartzia**	
Campylobacter	**Enterococcus**	**Parabacteroides**	**Sediminibacterium**	
Cloacibacillus	**Faecalibacterium**	**Phascolarctobacterium**	**Selenomonas**	

Paired Confirmatory Analyses for Tumor versus Normal Comparisons. Since LEfSe analyses do not take into consideration the paired nature of data in tumor versus normal comparisons, we ran Wilcoxon signed ranked tests for all genera identified through our gatekeeping procedures, listed in **Table 6**, to see if paired analyses corroborated our LEfSe results for our CIMP-H tumor versus CIMP-H normal tissue, CIMP-L tumor versus CIMP-L normal tissue, and Non-CIMP° tumor versus normal tissues. While there were several biomarkers LEfSe found that distinguished CIMP-H tumor and normal tissues that were also statistically significant in Wilcoxon tests after adjusting for multiple comparisons using the Benjamini-Hochberg Method (*Bacteroides, Blautia, Campylobacter, Coprococcus, Dorea, Eggerthella, Enterococcus, Faecalibacterium, Fusobacterium, Granulicatella, Leptotrichia, Parabacteroides, Phascolarctobacterium,* and *Ruminococcus),* only *Fusobacterium* significantly distinguished CIMP-L tumors from CIMP-L normal tissues. As in our LEfSe analysis, nothing significantly distinguished Non-CIMP° tumors from Non-CIMP° normal tissues. Full results of the Wilcoxon analysis can be viewed in **Supplemental Table 1.**

Correlations between Individual BPB and CIMP Markers. We used Spearman's R to determine if any correlations existed between individual CIMP markers and the 25 bacterial genera with the highest relative abundance across all sample types **(Figure 10).** In tumor tissues, *Blautia, Coprococcus, Dorea, Faecalibacterium,* and *Ruminococcus* were inversely correlated with several CIMP markers, while *Fusobacteria, Leptotrichia,* and *Selenomonas* were positively associated with several CIMP markers. Interestingly, the strongest correlation between any bacterial genera and any methylation marker was observed between *Blautia*, (a BPB and LEfSe-identified biomarker) and MLH1 methylation in tumor tissues (Spearman's R = -0.44). In fact, *Blautia*, was correlated with more methylation markers (7 of 8 markers) than any other genera. Two other important BPB in the intestine that were LEfSe-identified biomarkers, *Coprococcus* and *Faecalibacterium* were also inversely associated with multiple methylation markers (5 of 8 and 4 of 8 markers, respectively). CDKN2A was not significantly associated with any of the top 25 most abundant genera in tumor tissues, but was inversely associated with the BPBs *Blautia* and *Coprococcus*, as well as *Sutterella* in normal tissues.

Individual Genera of Butyrate-Producing Bacteria. As we have found that several important BPB were inversely associated with methylation status and were found to be significant biomarkers in our LEfSe analyses, we further investigated BPB taxa in this study. Average relative abundance of each genera of butyrate-producing bacteria per CIMP designation and tissue type can be viewed in **Figure 11**, along with test for significant differences among individual bacteria in **Table 7.** The mean relative abundance for *Anaerostipes, Butyrivibrio,* and *Pseudobutyrivibrio* were very low, with less than 0.01% average percent relative abundance for each CIMP designation in both tumor and normal tissues. For the remaining BPB, *Blautia, Butyricimonas, Coprococcus, Eubacterium, Faecalibacterium,* and *Roseburia,* CIMP-H had the lowest average relative abundance in both tumor and normal tissues compared to CIMP-L and Non-CIMP°, and these results were statistically significant.

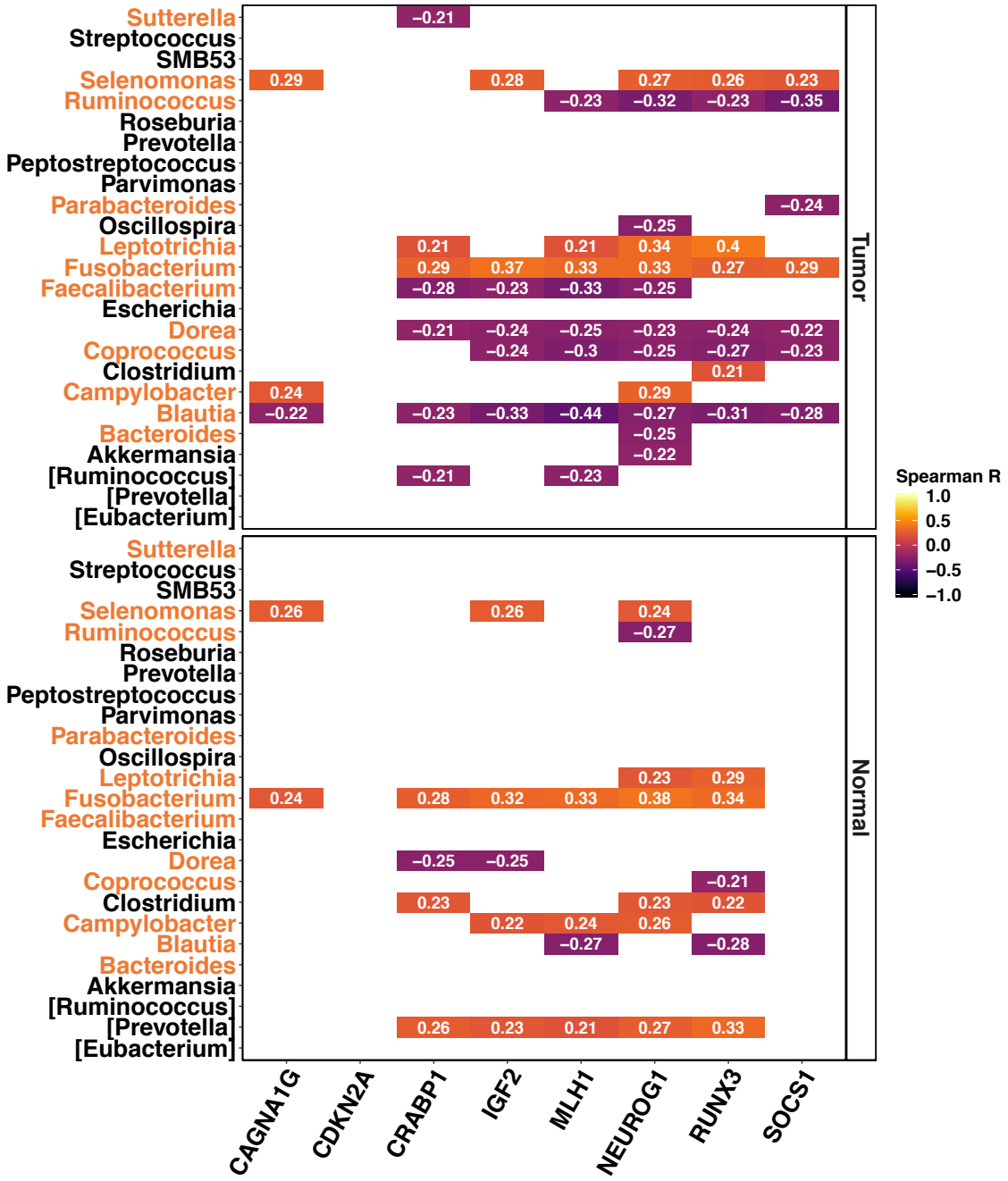


Figure 10: Correlations between Abundant Genera and CIMP Methylation Markers. Top plot shows correlations in normal tissues, while bottom plot shows correlations in tumor tissues. Correlations were calculated between the 25 genera with the highest relative abundance across all tissues and the PMR of each of the eight CIMP associated methylation markers in normal and tumor tissues, respectively. On the y-axis, bacterial genera identified by LEfSe gate-keeping methods as biomarkers are shown in orange, while bacterial species not identified by LEfSe are shown in black. Spearman correlation coefficients are displayed and colored according to their magnitude and direction. Only correlation coefficients significant at α = 0.05 are shown.

In tumor tissues, *Blautia* and *Eubacterium* were highest in average relative abundance in CIMP-L patients, while *Butyricimonas, Coprococcus, Faecalibacterium,* and *Roseburia* were highest in Non-CIMP° patients. In normal tissues, *Faecalibacterium* and *Eubacterium* had the highest average relative abundance in CIMP-L patients, while *Blautia, Butyricimonas, Coprococcus,* and *Roseburia* were highest in Non-CIMP° patients.

Aggregated Total Relative Abundance of BPB. When the relative abundance for all BPB was aggregated, the median total percent relative abundance of BPB in normal tissues for CIMP-H, CIMP-L, and Non-CIMP° patients was 6.65%, 9.68%, and 11.67%, respectively. In tumor tissues, the median total percent relative abundance for CIMP-H, CIMP-L, and Non-CIMP° patients was 1.82%, 4.68%, and 7.04%, respectively. Between group (CIMP-H, CIMP-L and Non-CIMP°) comparisons of cumulative percent relative abundance of all BPB genera did not reveal statistically significant differences in normal tissue; however, a trend was observed for CIMP-H normal tissue having lower total BPB percent relative abundance compared to CIMP-L and Non-CIMP° normal tissue **(Figure 12)**.

We did find a statistically significant difference between total BPB percent relative abundance comparing CIMP-H tumors to CIMP-L tumors and comparing CIMP-H tumors to Non-CIMP° tumors (**Figure 12).** We also found that there was a significant decrease in BPB percent relative abundance between paired tumor and normal samples for CIMP-H patients and for CIMP-L patients but not between the tumor and normal tissues in Non-CIMP° patients **(Figure 12).** These results indicate that total relative abundance of BPB in CIMP-H tumor is significantly reduced, compared to both tumor tissues in patients with fewer markers methylated, as well as compared to matched CIMP-H normal tissues. In fact, the normal tissues of CIMP-H patients had levels of BPB comparable to CIMP-L and Non-CIMP° tumor tissues. Our results also suggest that difference in total relative abundance of BPBs may be modified by CIMP marker methylation.

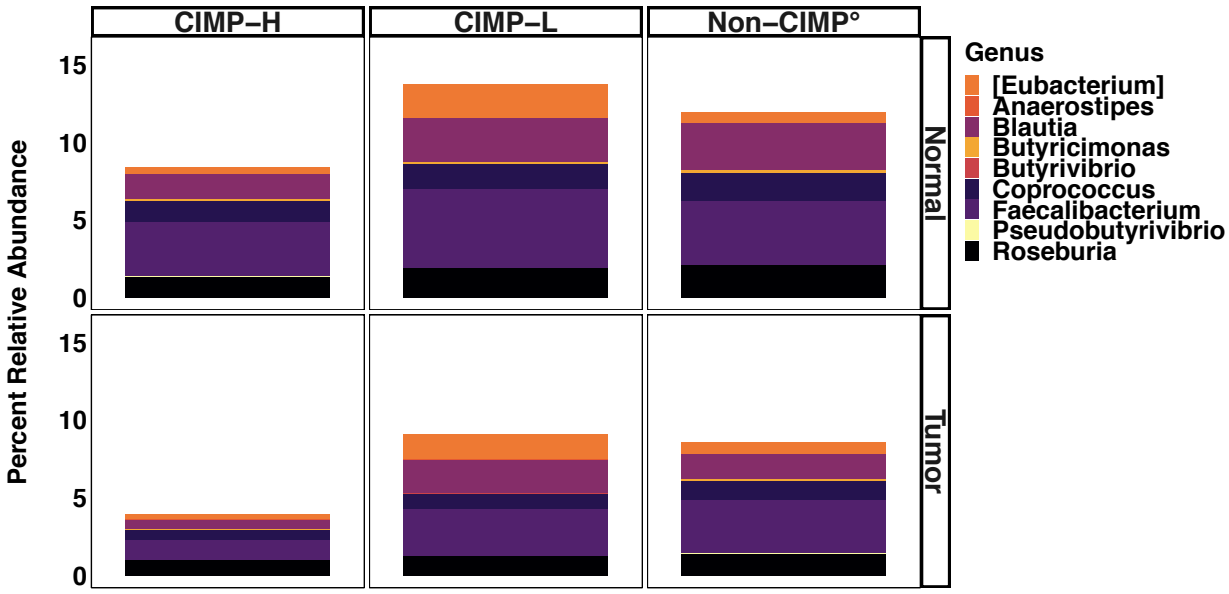


Figure 11: Average Total Relative Abundance of Butyrate Producing Bacteria. *Averages are calculated by CIMP designation, and stratified by tissue type, and displayed in percent relative abundance.*

Table 7: Average Percent Relative Abundance of Butyrate Producing Bacteria

	CIMP-H	CIMP-L	Non-CIMP°	CIMP-H vs CIMP-L	CIMP-H vs Non-CIMP°	CIMP-L vs Non-CIMP°
Tumor Tissues						
Anaerostipes	0.00%	0.00%	0.01%	**p < 0.0001**	**p < 0.0001**	**p < 0.0001**
Blautia	0.61%	2.13%	1.61%	**p < 0.0001**	**p < 0.0001**	**p < 0.0001**
Butyricimonas	0.03%	0.05%	0.10%	p = 0.0999	p = 0.999	**p = 0.0160**
Butyrivibrio	0.00%	0.00%	0.00%	**p < 0.0001**	**p < 0.0001**	**p < 0.0001**
Coprococcus	0.67%	1.00%	1.26%	**p < 0.0001**	**p < 0.0001**	**p < 0.0001**
Eubacterium	0.31%	1.60%	0.76%	**p = 0.0005**	**p = 0.0005**	**p = 0.0081**
Faecalibacterium	1.29%	3.02%	3.43%	**p < 0.0001**	**p < 0.0001**	**p < 0.0001**
Pseudobutyrivibrio	0.00%	0.00%	0.02%	**p < 0.0001**	**p < 0.0001**	**p < 0.0001**
Roseburia	1.02%	1.28%	1.43%	**p < 0.0001**	**p < 0.0001**	**p < 0.0001**
Normal Tissues						
Anaerostipes	0.01%	0.00%	0.01%	**p < 0.0001**	**p < 0.0001**	**p < 0.0001**
Blautia	1.65%	2.86%	3.05%	**p < 0.0001**	**p < 0.0001**	**p < 0.0001**
Butyricimonas	0.08%	0.11%	0.15%	p = 0.3879	p = 0.3879	p = 0.0984
Butyrivibrio	0.00%	0.00%	0.00%	**p < 0.0001**	**p < 0.0001**	**p < 0.0001**
Coprococcus	1.38%	1.63%	1.81%	**p < 0.0001**	**p < 0.0001**	**p < 0.0001**
Eubacterium	0.41%	2.14%	0.69%	**p = 0.0001**	**p = 0.0001**	**p = 0.0023**
Faecalibacterium	3.50%	5.04%	4.15%	**p < 0.0001**	**p < 0.0001**	**p < 0.0001**
Pseudobutyrivibrio	0.00%	0.00%	0.01%	**p < 0.0001**	**p < 0.0001**	**p < 0.0001**
Roseburia	1.38%	1.96%	2.10%	**p < 0.0001**	**p < 0.0001**	**p < 0.0001**

P-values calculated using Mann-Whitney test and adjusted for multiple comparisons using Benjamini & Hochberg methods.

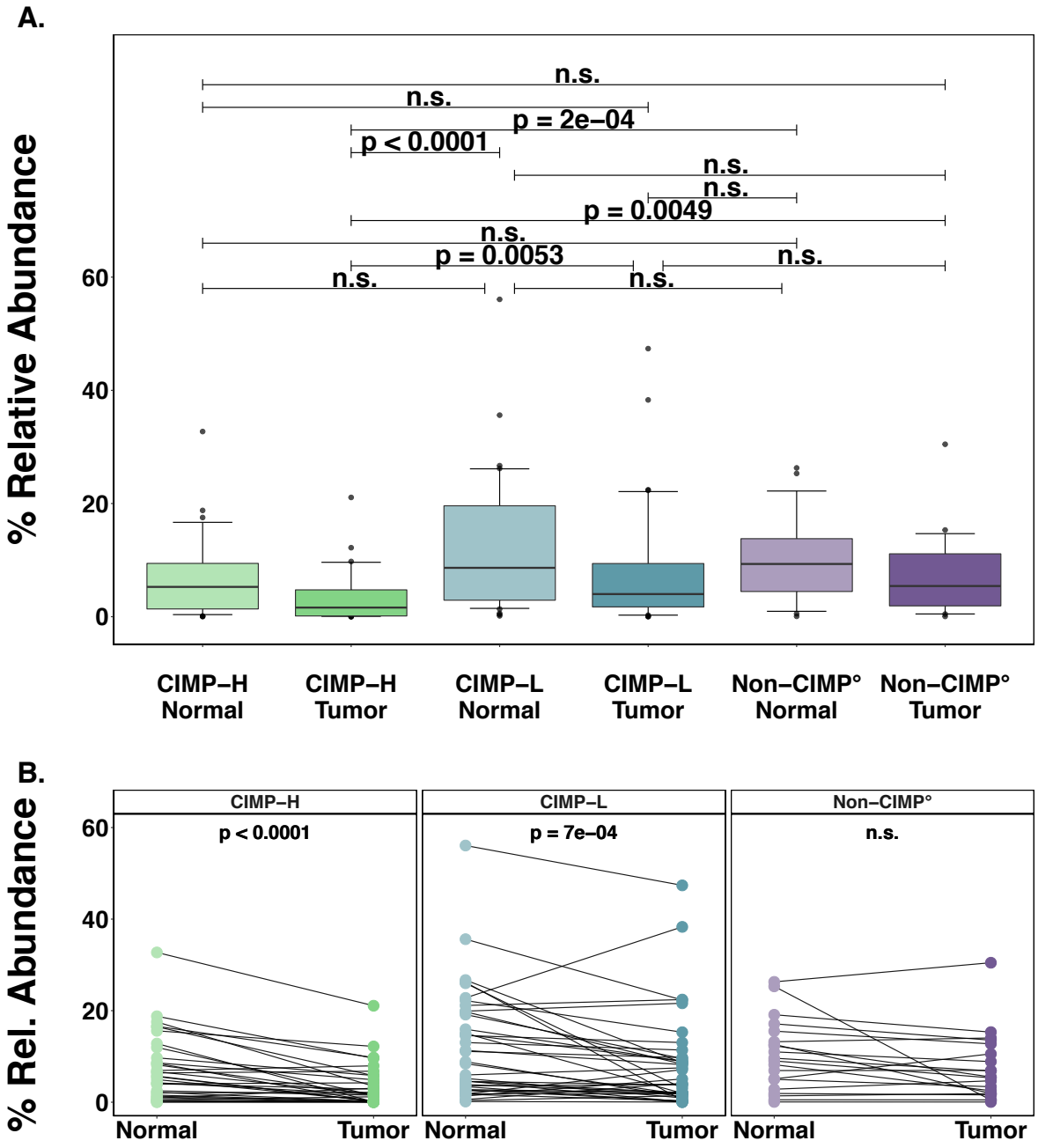


Figure 12: Within and Between Group Comparisons for Percent Relative Abundance of Butyrate-Producing Bacteria. *Figure 12A shows boxplots of Percent Relative Abundance of cumulative BPBs by CIMP designation and tissue type. Whiskers represent the 10th to 90th percentile of data in that group, and outliers are represented by points. Mann-Whitney p-values are shown for between group comparisons. Figure 12B shows spaghetti plots of percent relative abundance of cumulative BPBs for matched tumor and normal tissues for each subject, stratified by CIMP designation. Wilcoxon paired p-values are shown.*

Regression Analysis of BPB Relative Abundance as a Predictor of CIMP

Since the demographic characteristics of CIMP-L and Non-CIMP° patients were relatively similar, we used binary logistic regression, with CIMP status as the outcome, to model the relationship between total BPB relative abundance and CIMP. Separate models were run for tumor tissues and normal tissues. Full results of this model can be viewed in Model B of **Table 8.**

For every percentage increase in total percent relative abundance of BPB in their tumor tissue sample, the crude odds of a patient having CIMP subtype CRC decreases by 10.0% (Crude OR = 0.90, 95% CI: 0.82, 0.98, p-value = 0.013). When we adjusted for age in years, sex, and resection side, the odds of a patient having CIMP subtype CRC decreased by 12.0% for every percentage increase in total relative abundance in their tumor tissue sample (OR = 0.88, 95% CI: 0.80, 0.97, p-value = 0.0115). Since our sample size is small enough that we may be underpowered to adjust for all three of these covariates, we repeated our adjusted model on 1000 bootstrapped samples to test the robustness of our results. The results of our bootstrapped model were consistent with our crude and adjusted model, and the bootstrapped confidence interval did not contain the null value of 1.

In our model based on normal tissues, we found that for every one percentage increase in total relative abundance of BPB, the odds of a patient having CIMP subtype of CRC decreased by 5.0% (Crude OR = 0.95, 95% CI: 0.901, 0.997, p-value = 0.0383). After adjusting for age in years, sex, and resection side, we found that for every one percentage increase in total relative abundance of BPB, the odds of a patient having CIMP subtype of CRC decreased by 6.0%, but these results were not statistically significant (Adj. OR = 0.94, 95% CI: 0.89, 1.00, p-value = 0.0514). We used the same bootstrap methods to test the robustness of our results for our fully adjusted model in normal tissues, which resulted in a confidence interval that did include our null value of 1.

Since CIMP status is common in our sample, we also modeled the relationship between total BPB and CIMP status with a relative risk regression model to estimate our crude and adjusted RR, (Model C, **Table 8**). For every percentage increase in total relative abundance of BPB in a patient's tumor tissue sample, the risk of CIMP subtype of CRC decreased by 8.0% (RR = 0.92, 95% CI: 0.87, 0.98, p-value = 0.0095). After adjusting for dichotomized age, sex, and resection type, we found that for every percentage increase in relative abundance of BPB in tumor tissue samples, the risk of CIMP subtype of CRC decreased by 7.0% (RR = 0.93, 95% CI: (0.87, 0.99), p-value = 0.0158).

Using the relative abundance for normal tissues, we found that for every one percentage increase in total BPB relative abundance in normal tissues, the risk of CIMP subtype of CRC decreased by 4.0% (Crude RR = 0.96, 95% CI: 0.93, 1.00, p-value = 0.0355). However, after adjusting for dichotomized age, sex, and resection side, we found that for every percentage increase in total BPB relative abundance in normal tissues, the risk of CIMP subtype of CRC decreased by 3.0%, and this result was not statistically significant (Adj. RR = 0.97, 95% CI: 0.94, 1.01, p-value = 0.0968).

We also ran binary logistic regression models for each of the ten individual genera of BPB as the predictor variables of CIMP status (data not shown). A decrease in *Blautia* in tumor tissues was significantly associated with CIMP status in both the crude and adjusted models (β_{crude} p-value = 0.0049, β_{adj} p-value = 0.0027), but a decrease in *Blautia* in normal tissues was only significantly associated with CIMP in the crude model (β_{crude} p-value = 0.0190, β_{adj} p -value = 0.0742). A decrease in *Coprococcus* in tumor tissues was significantly associated with CIMP in the adjusted model, but not the crude model (β_{crude} p-value = 0.2364, β_{adj} p-value = 0.0216), and a decrease in *Faecalibacterium* in tumor tissues was associated with CIMP in the crude model, but not the adjusted model (β_{crude} p-value = 0.0405, β_{adj} p-value = 0.0557). No other BPB in either tumor or normal tissues were significantly

associated with CIMP in crude or adjusted regression models. Since all significant associations were among bacteria that were LEfSe-identified biomarkers, we did not adjust p-values for multiple testing.

Table 8: Results of Regression Analyses with Total Relative Abundance of Butyrate-Producing Bacteria as the predictor of CIMP status

	Tumor Tissues			**Normal Tissues**		
	Estimate and 95% CI		**p-value**	**Estimate and 95% CI**		**p-value**
Model B. Binary CIMP Logistic Regression						
Crude Model	OR = 0.90	(0.82, 0.98)	**p = 0.0126**	OR = 0.95	(0.90, 1.00)	**p = 0.0383**
Adjusted Model[2]	OR = 0.88	(0.78, 0.97)	**p = 0.0115**	OR = 0.94	(0.89, 1.00)	p = 0.0514
Bootstrapped Adj. Model[3]	OR = 0.86	(0.76, 0.98)	-	OR = 0.94	(0.888, 1.00)	-
Model C. Binary CIMP Rel. Risk Regression						
Crude Model	RR = 0.92	(0.87, 0.98)	**p = 0.0095**	RR = 0.96	(0.93, 1.00)	**p = 0.0355**
Adjusted Model[4]	RR = 0.93	(0.87, 0.99)	**p = 0.0158**	RR = 0.97	(0.94, 1.01)	p = 0.0968

1. *CIMP status is modeled as the outcome of interest.*
2. *Model adjusted for age in years (continuous), sex (male or female), and resection side (left or right).*
3. *OR and CI from averaging beta coefficients of adjusted regression models from 1000 bootstrapped samples.*
4. *Model adjusted for age (categorical, younger than 69 or 69 and older), sex (male or female) and resection side (left or right). Age was modeled as a categorical variable in order to make the relative risk regression model converge. 69 was chosen as the cutoff for age since it was the sample median.*

Unsupervised Topological Data Analysis with Ayasdi

The first topological model, TM1 can be viewed in **Figure 13**. This model shows a network derived from data for all of the top 100 genera with the highest average relative abundance, and includes all 184 samples. The model is colored to show a gradient between nodes that are highly associated with tumor samples (red nodes), and nodes that are highly associated with normal tissues (blue nodes). TM1 contains two outliers, and one network characterized by tight network in the center of the figure (TM1-E), with several subnetworks of nodes that are distinct from the central region of the network and also distinct from each other (TM1-A, TM1-B, TM1-C, and TM1-D). The five circled regions (labeled TM1-A, TM1-B, TM1-C, TM1-D, and TM1-E) were considered subnetworks and further analyzed to look for characteristics that distinguished these regions from the remaining model.

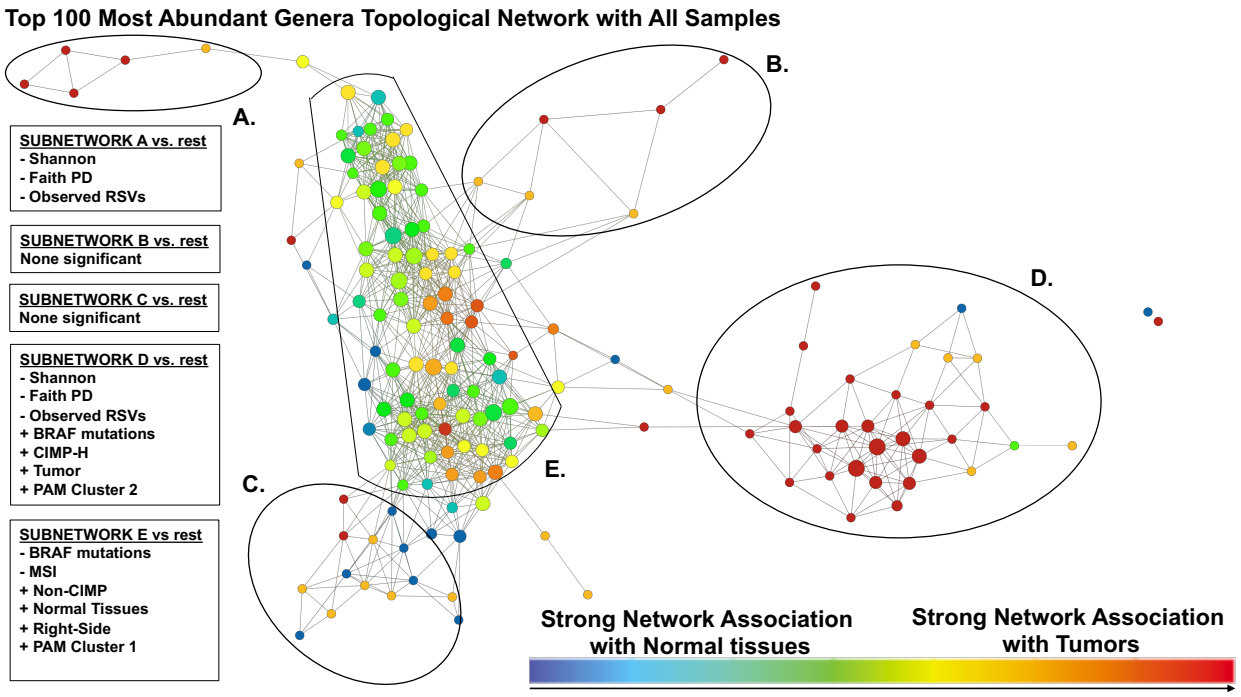


Figure 13: Ayasdi TDA Model 1 (TM1). *This model was created using Ayasdi Topological Data Analysis platform. The TM1 network was built on the top 100 most abundant bacterial genera across all samples, and includes all 184 samples in the dataset (tumor and normal tissues). Euclidean distance (L2) was applied as the metric, with Neighborhood Lenses 1 and 2 at a resolution of 30 and a gain of 4.0 for each lens. Nodes are connected by similarities, with tight networks (high degree of connections among nodes) indicating related data points, and loose networks (low degree of connections among nodes) indicating fewer relationships between data points. This model is colored so that red nodes indicate a strong association with tumor tissues, while blue nodes indicate a strong association with normal tissues. Subnetworks that were assessed are circled and labeled. We tested for significant differences in clinical and demographic differences between each subnetwork and the rest of the model. Significant results are listed on the left-hand side of the figure. A (+) indicated a significant positive relationship between the subnetwork and the variable, while a (-) indicates a significant inverse relationship between the subnetwork and the variable.*

We used the Ayasdi platform to test for significant difference in alpha diversity, clinical, and demographic characteristics between the subnetworks and the rest of the model. Significant results from this subnetwork analysis can be viewed in **Figure 13.** TM1-A had a statistically significant decrease in alpha diversity for all three metrics (Shannon, Faith PD, and Observed RSVs). TM1-D was positively associated with *BRAF* mutations, CIMP-H, and tumor tissues, and PAM Cluster 2, and showed a significant decrease in alpha diversity for all three metrics (Shannon, Faith PD, and Observed RSVs). Conversely, TM1-E was positively associated with PAM Cluster 1, Non-CIMP, normal tissues, right-side

resections, and MSS, and negatively associated with *BRAF* mutations. TM1-B and TM1-C were not significant for any alpha diversity, clinical, or pathological metrics.

In **Figure 14,** TDA network TM1 (Same model as Figure 13) is colored by the strength of associations for nodes within the network for microbiome variables of interest. In **Figure 14 A,** we see that TM-1 shows a strong separation between PAM Cluster 1 (red) and Cluster 2 (blue), with Cluster 2 association being strongest in TM1-D from Figure 13. We can also see that TM1-C and TM1-E are more strongly associated with the Firmicutes phylum, compared to the rest of the model **(Figure 14 B)**. We evaluated the relationship of each of the ten BPB of interest and found that each seemed to be highly associated with TM1-C and TM1-E and inversely associated with TM1-D. *Blautia* and *Faecalibacterium* can be viewed in **Figure 14 C** and **Figure 14 D**, respectively, though other BPB followed similar gradients. *Fusobacterium* follows a gradient of abundance, with the highest relative abundance in TM1-A and TM-D, and lower relative abundance in TM1-E **(Figure 14 E)**. Though TM1-A and TM1-D are both associated with *Fusobacterium,* we can see that TM1-A is more strongly associated with *Bacteroides* compared to TM1-D, which may explain why TM1-A seems to be associated with PAM Cluster 1 **(Figure 14 F).**

Bacterial genera that were significantly different in each subnetwork, compared to the rest of the model, are summarized in **Figure 15** below. Several BPB were found to be either associated with, or inversely associated with, one or more of the subnetworks in our model. *Blautia, Faecalibacterium,* and *Roseburia,* were found to be positively associated with TM1-E and negatively associated with TM1-A and TM1-D. *Coprococcus* was found to be positively associated with TM1-E and negatively associated with TM1-D. While TM1-E has a higher abundance of BPB, TM1-D is marked by *increased Campylobacter, Fusobacterium, Leptotrichia, Prevotella,* and *Selenomonas* in addition to the aforementioned decrease in certain BPB. **Supplemental Figure 4** shows all other taxonomic variables and **Supplemental Figure 5** shows all PICRUSt L3 functional variables that were statistically significant after adjusting for multiple

comparisons within at each taxonomic or PICRUSt level, along with resulting Kolmogorov-Smirnoff scores for comparisons between each subnetwork and the rest of the model.

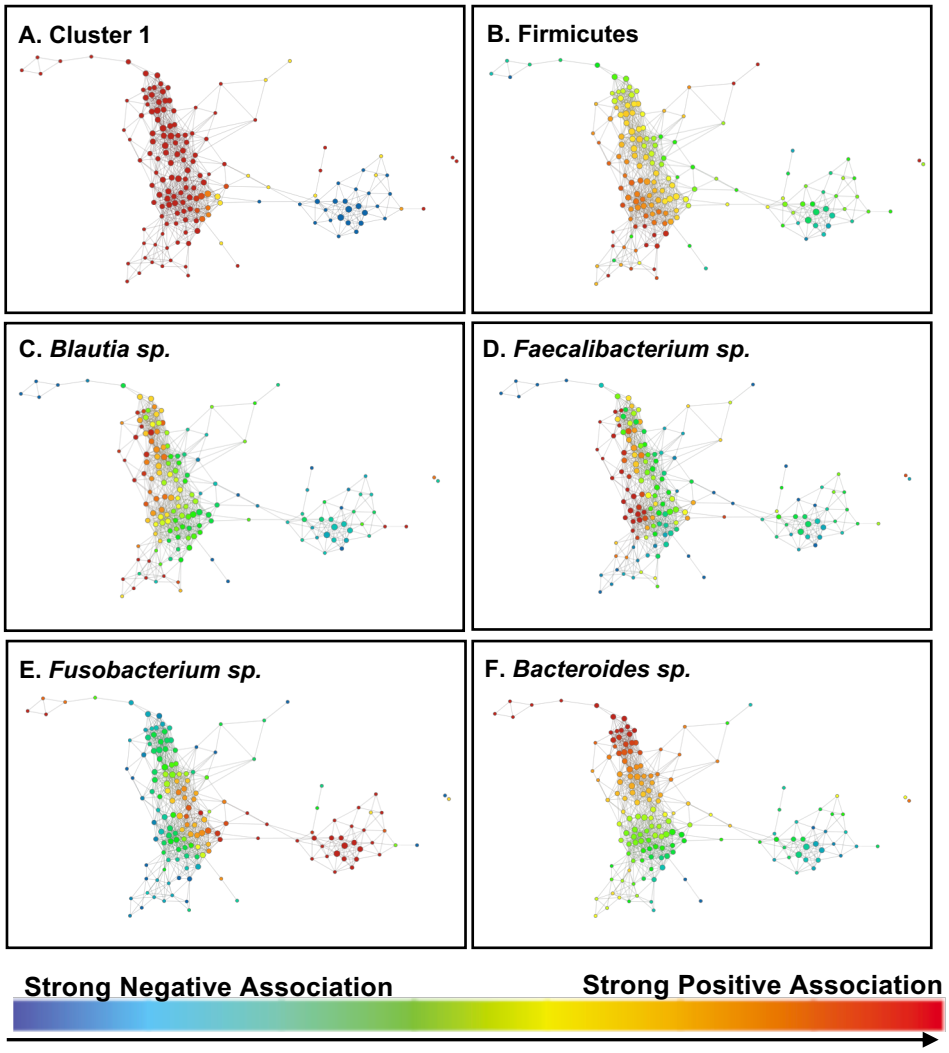


Figure 14: Ayasdi TDA Model 1 (TM1) colored by association with microbiome variables of interest. *The models in this figure are repetitions of TM1, with variation in coloring according to associations between the nodes and bacterial genera of interest. Additional information about TM1 can be viewed in Figure 13. Figure 14A is colored by PAM Cluster 1. Red nodes indicate strong associations with PAM Cluster 1, while blue nodes indicate a strong inverse association with PAM Cluster 1 (i.e., a strong association with PAM Cluster 2). Figure 14B is colored according to the Firmicutes phylum, which contains our ten BPB of interest. Red nodes indicate a strong positive association with Firmicutes. Figure 14C is colored so that red indicates a strong positive association with Blautia at the genus level, while blue indicates a strong inverse association. Figure 14D is colored so that red indicates as strong positive association with Faecalibacterium at the genus level, while blue indicates a strong inverse association. Figure 14E is colored so that red indicates a strong positive association with Fusobacterium at the genus level, while blue indicates a strong inverse association. Figure 14F is colored so that red indicates a strong positive association with Bacteroides at the genus level, while blue indicates a strong inverse association.*

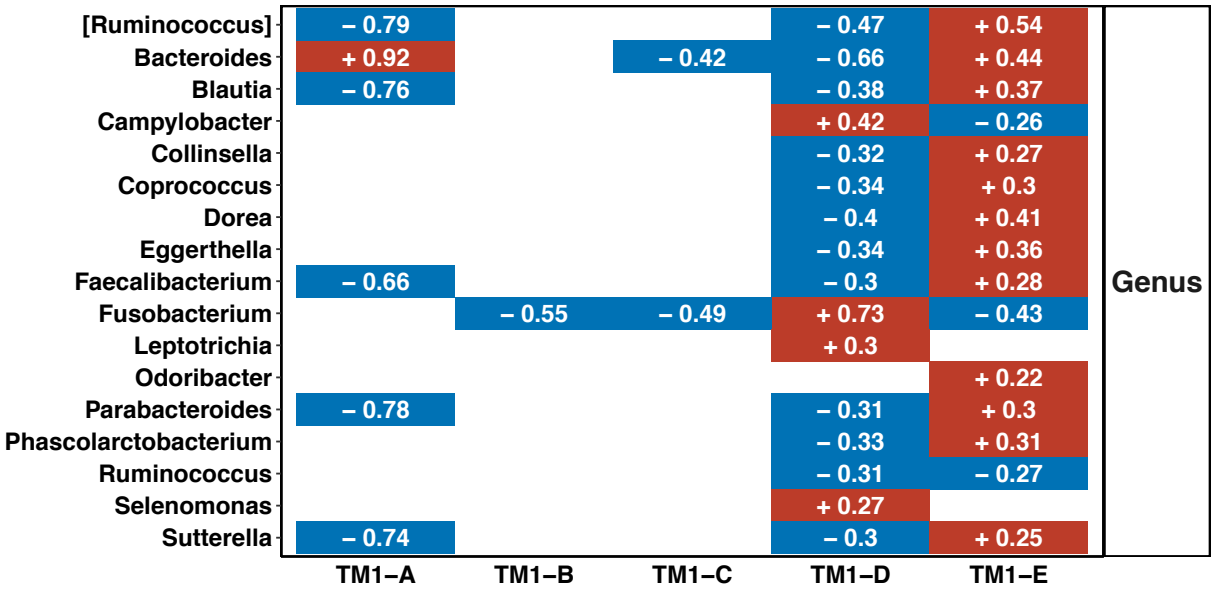


Figure 15: Genera that Significantly Distinguish Subnetworks in TM1. *For LEfSe-identified biomarkers, Kolmogorov-Smirnov test for significance was used to compare genera across each subnetwork to test for significance, and all genera that were significant at $\alpha = 0.05$ are shown. Since LEfSe analysis was used as a gatekeeper, additional adjustment for multiple testing was not done. Red indicates a positive association between the genera and the sub-network, while blue indicates an inverse association. Kolmogorov-Smirnov test statistics are displayed. For more information regarding TM1 and subnetworks, see Figures 13 & 14.*

We repeated our topological model based on the top 100 most relative abundant genera in CIMP-H, CIMP-L, and Non-CIMP° subsets to see if sub-setting by CIMP subtypes would result in separation of the model into distinct networks of tumor and normal tissues or by PAM clusters. The results of these models can be viewed in **Figure 16.** The Non-CIMP° topological model, TM4, model resulted in one network, while CIMP-L model, TM3, and CIMP-H model, TM2, each resulted in two networks. We analyzed the resulting networks in TM2 and TM3 to see if the model separated by tissue type or PAM cluster.

In our CIMP-H network, *Eggerthella, Bacteroides, Parabacteroides, [Ruminococcus], Blautia, Coprococcus, Dorea, Faecalibacterium, Hydrogenophaga*, and *Escherichia* were positively associated with CH2, and negatively associated with CH1, while *Fusobacterium* was positively associated with CH1, and negatively associated with CH2. CL1 was positively associated with *BRAF* mutations, PAM Cluster 2,

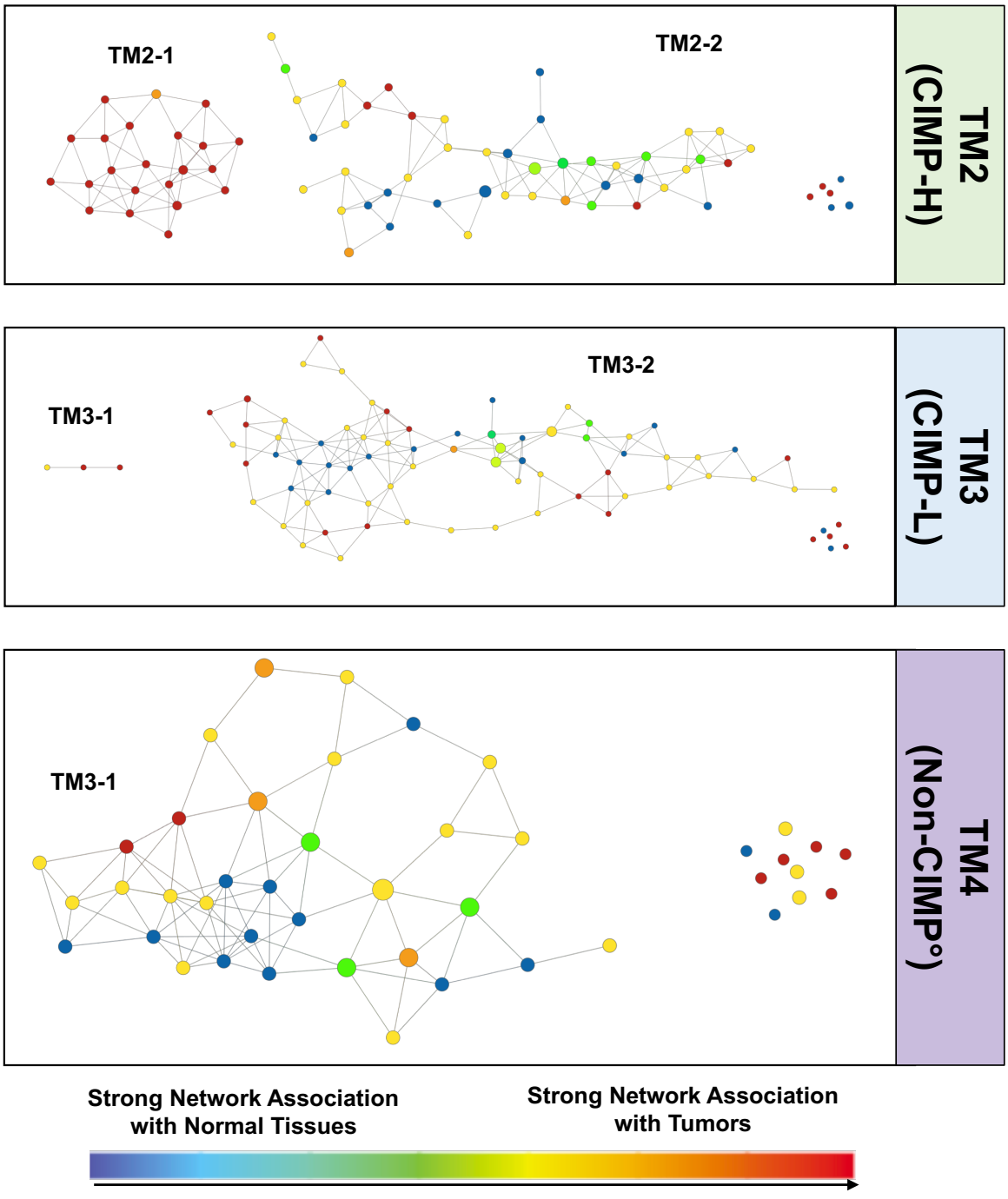


Figure 16: Ayasdi TDA Models 2, 3, and 4 (TM2, TM3, TM4). This model was created using Ayasdi Topological Data Analysis platform. TM2, TM3, and TM4 were derived from data for the top 100 most abundant genera across all samples, on samples in the CIMP-H group, CIMP-L group, and Non-CIMP° group, respectively. Euclidean distance (L2) was used, with Neighborhood Lenses 1 and 2 at a resolution of 30 and a gain of 3.6. Nodes are connected by similarities, with tight networking indicating related data points, and lose networks indicating fewer relationships between data points. This model is colored so that red nodes indicate a strong association with tumor tissues, while blue nodes indicate a strong association with normal tissues. Subnetworks were circled and labeled.

and tumor tissues as well as a lower alpha diversity for all three metrics (Shannon, Faith PD, and Observed RSVs). Meanwhile, CL2 was positively associated with PAM Cluster 1 and normal tissues, in addition to a higher alpha diversity for all three metrics (Shannon, Faith PD, and Observed RSVs).

In our CIMP-L model, CL1 was relatively small, with three nodes, while CL2 contained most of the samples. CL1 was associated with a decrease in relative abundance of Bacteroides and Parabacteroides, compared to the rest of the model, while CL2 was associated with an increase in relative abundance of *Bacteroides, Parabacteroides, Odoribacter, Faecalibacterium, Ruminococcus*, and *Fusobacterium,* and a decrease in relative abundance of *Ralstonia* and *Bilophila.* Neither network was significantly associated with tumor or normal tissues, nor were they significantly associated with any other clinical features, including PAM clusters.

While neither the Non-CIMP° nor the CIMP-L model separated by tissue type or PAM cluster, the CIMP-H model, showed clear separation into two networks, with one significantly associated with tumor tissue and PAM Cluster 2, and the other significantly associated with normal tissues and PAM Cluster 1.

Supervised Machine Learning with Ayasdi

We used Ayasdi to test for differences between variables in CIMP-H, CIMP-L, and Non-CIMP° groups, stratifying by tumor and normal tissues. Taxa that are significantly different for each phylogenetic level using the Kolmogorov-Smirnov p-value < 0.05 can be seen in **Figure 17**, along with their KS scores.

In tumor tissues, there were several taxa at each phylogenetic level that significantly distinguish between CIMP-H compared to CIMP-L patients, and CIMP-H compared to Non-CIMP° patients. Meanwhile, the differences between CIMP-L and Non-CIMP° seems to be driven by one bacterial genus, and four RSV-level differences. Many of these differences disappear when we instead compare groups

across their normal tissues. While there are fewer distinguishing taxa when comparing CIMP-H to CIMP-L normal tissues versus in tumor tissues, there are still a couple of taxa at each level that are statistically significant. There were no significant differences between CIMP-H to Non-CIMP° normal tissues, and only one bacterial genus that distinguished between CIMP-L and Non-CIMP° patients in normal tissues. Functional differences between CIMP groups based on PICRUSt analyses, stratified by tissue type, can be viewed in **Supplemental Figure 6.**

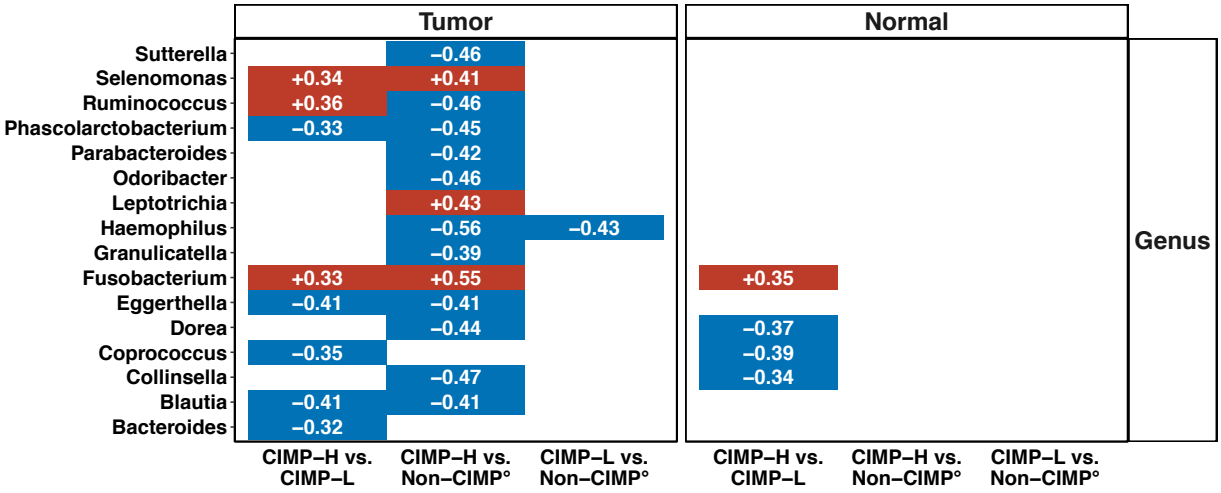


Figure 17: Significant differences at each phylogenetic level between CIMP groups, stratified by tissue type. *For LEfSe-identified biomarkers, Kolmogorov-Smirnov test for significance was used to compare genera across each subgroup to test for significance, and all genera that were significant at $\alpha = 0.05$ are shown. Since LEfSe analysis was used as a gatekeeper, additional adjustment for multiple testing was not done. Red indicates a positive association between the genera and the sub-group, while blue indicates an inverse association. Kolmogorov-Smirnov test statistics are displayed.*

We were also interested in assessing if the factors that distinguish tumor tissues from normal tissues in CIMP-H patients differed from factors that distinguish tumor tissues and normal tissues in CIMP-L and Non-CIMP° patients. **Figure 18** shows significant differences between tumor and normal tissues for each taxonomic level, stratified by CIMP subtype. Interestingly, there were several differences in CIMP-H patients' tissues that were not observed in CIMP-L or Non-CIMP° patients. While CIMP-H tumors were marked by increased *Fusobacterium* and *Campylobacter,* and a decrease in

Bacteroides, Bilophilia, Blautia, Coprococcus, Dorea, Eggerthella, Parabacteroides, and *Ruminococcus* at the genus level, compared to normal tissues, CIMP-L tumors were only distinguished from CIMP-L normal tissues by an increase in *Fusobacterium* and a decrease in *Coprococcus* and *Dorea.* There were no genera that significantly distinguished Non-CIMP° tumor from normal tissues.

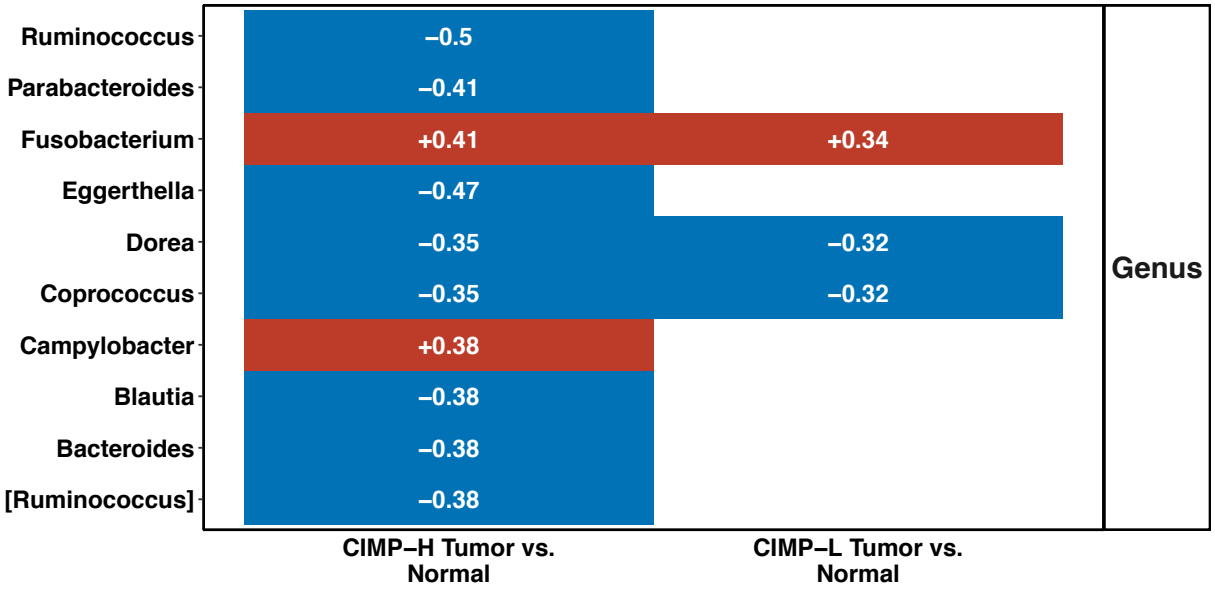


Figure 18: Significant differences at the genus level within CIMP groups. *For all LEfSe-identified biomarkers, Kolmogorov-Smirnov test for significance was used, and all genera that were significant at $\alpha = 0.05$ are shown. Red indicates a positive association between the genera and the sub-network, while blue indicates an inverse association. Kolmogorov-Smirnov test statistics are displayed.*

IV. DISCUSSION

In this study, we examined the relationship between the microbiome and subtypes of CRC. We were specifically interested in investigating the role of butyrate-producing bacteria and methylation of CIMP markers. Our research was unique in that we applied unsupervised topological data analysis to our data in addition to using traditional univariable, regression, clustering, and LDA Effect Size analyses. The use of several techniques allows us to corroborate findings between our different analyses.

Clinical and Demographic Features of CIMP Patients. We corroborated several associations found in existing literature between CIMP CRC and clinical, demographic, and molecular characteristics. In our sample of 92 patients, consisting of 34 CIMP-H patients, 38 CIMP-L patients, and 20 Non-CIMP° patients, we found that CIMP was associated with an older age at tumor resection, microsatellite instability, and an increased prevalence of *BRAF* mutations.[6] In contrast to current literature, there was no association between CIMP and female sex, resection side, or TNM stage.[6] However, in our cohort more females than males were in the CIMP-H and CIMP-L group and the majority of tumors were right-sided from each group. We did not find any significant differences in clinical or demographic features between CIMP-L and Non-CIMP° patients.

Decreased in Alpha Diversity of CIMP-H and CIMP-L Tumor Tissues. We found a significant loss in alpha diversity consistent across three different diversity metrics when comparing matched tumor tissues to normal tissues in CIMP-H and CIMP-L patients that was not apparent in Non-CIMP° patients. Diversity in the gut microbiome is important for promoting stability of beneficial microbiota and resilience of the microbiome against perturbations such as dietary changes, antibiotic administration, and invasion of pathogenic species.[32] Microbial diversity may be influenced by several factors, including environment, diet, antibiotic use, and early-life factors such as delivery mode and breastfeeding. Loss of diversity has been linked to diseases states including inflammatory bowel disease, *C. Difficile* Associated Diarrhea, and dysbiosis.[32]

Additionally, microbial diversity is decreased in elderly patients, compared to younger adults. Age-related reductions in alpha diversity are also characterized by shifts in types of bacteria that dominate the microbiome, including a decline in commensal microbiota, decreased availability of SCFAs, and a reduction in the abundance of Firmicutes, including bacteria in *Clostridium* cluster XIVa.[23]

Interestingly, since samples were matched by patient, these environmental and demographic factors cannot explain the differences in alpha diversity between tumor and normal tissues we observed in CIMP-H and CIMP-L, but not Non-CIMP° patients. These differences may be dependent on differences in the tumor microenvironment that support the growth of certain bacteria and prevent the growth of others. Our results, however, demonstrate that the effect on alpha-diversity is dependent on the subtype of CRC, where only those tumors with CIMP-associated methylation in CIMP marker genes exhibit this loss of diversity. Since the median difference in alpha diversity between tumor and normal tissues was greater in CIMP-H compared to CIMP-L tissues, this suggest that increasing degree of methylation may be associated with decreasing alpha diversity.

Tissue Enterotypes Driven by Bacteroides and Fusobacterium. We applied PAM clustering analysis on our microbiome data to classify our data into enterotypes based on the beta diversity within each tissue sample. Published literature on the microbiome has shown that enterotyping analysis of the gut microbiome normally results in two or three enterotypes, driven by *Bacteroides, Ruminococcus,* or *Prevotella.*[25] However, our analysis resulted in two enterotypes, with Cluster 1 driven by *Bacteroides,* and Cluster 2 driven by *Fusobacterium*. Cluster 1 contained several bacteria we'd expect to find in a normal gut microbiome, including *Bacteroides, Ruminococcus, Escherichia, Akkermansia, Bifidobacterium* and *Dorea* as well as butyrogenic bacteria, *Faecalibacterium, Blautia, Roseburia,* and *Coprococcus.*[33,34] Meanwhile, several bacteria in Cluster 2, including *Campylobacter, Cantonella , Fusobacterium, Gemella, Leptotrichia, Parvimonas, Selenomonas,* and *Treponema,* are more commonly associated with the oral microbiome.[35] Cluster 2 was significantly associated with CIMP status among tumor tissues, even after adjusting for age, sex, and resection side in our multivariable analysis, though Cluster 2 was not associated with CIMP in normal tissues. This association was corroborated by performing one versus all LEfSe analysis on CIMP versus Non-CIMP tumor and normal tissues, where we

found that genera that distinguished CIMP tumor tissues from the other samples were similar to the genera that distinguished Cluster 2, while genera that distinguished Non-CIMP normal tissues from the rest of the samples were similar to the genera that distinguished Cluster 1.

Topological data analysis using the top 100 genera revealed strong separation into subnetworks by PAM Cluster membership, as seen in TM1. The subnetwork most strongly associated with Cluster 2, TM1-D, was positively associated with *Selenomonas, Leptotrichia, Fusobacterium, Campylobacter,* and *Prevotella*. Similarly, the subnetwork most strongly associated with Cluster 1, TM1-E, was positively associated with *Ruminococcus, Bacteroides, Bilophia, Blautia, Colinsella, Dorea, Eggerthella, Escherichia, Lachnospira, Parabacteroides*, and *Phascolarctobacterium,* among other genera. Our TDA analysis also found that TM1-D was significantly associated with CIMP-H and tumor tissues, while TM1-E was significantly associated with Non-CIMP and normal tissues. There seemed to be strong agreement between associations with clusters and bacterial genera in both our LEfSe and our TDA analyses.

Genera that Distinguish Tumor Tissues from Normal Tissues vary by CIMP Designation. In our LEfSe analysis comparing CIMP-H tumor tissues to CIMP-H normal tissues, we found that the taxa that distinguished between the two tissue types were strongly driven by genera that significantly defined the PAM clusters. While CIMP-H tumor tissues were associated with higher levels of *Fusobacterium, Campylobacter,* and *Leptotrichia,* CIMP-H normal tissues were associated with *Bacteroides, Ruminococcus, Faecalibacterium, Dorea, Blautia,* and *Coprococcus,* among several other genera. In CIMP-L patients, difference between tissue types were driven by an increased association with *Fusobacterium* and *Sphingomonas* in tumor tissues, and an inverse association with *Dorea, Coprococcus,* and *Actinomyces.* LEfSe did not find any genera that distinguished the microbiome at the genus level in tumor tissues from the normal tissues of Non-CIMP° patients, though Wilcoxon paired tests showed a decrease in *Blautia* in tumor tissues of Non-CIMP° patients, compared to their normal tissues.

Our topological data analyses resulted in similar finding, in that our results showed that topological models of relative abundance data for the top 100 most abundant bacterial genera and stratified by CIMP type resulted in distinct separation of networks in CIMP-H samples that were strongly associated with PAM cluster membership and tissue type. CIMP-L and Non-CIMP° samples did not separate by cluster or tissue type. Our Ayasdi networks also found that increased relative abundance of *Fusobacterium* and decreased abundance of *Dorea* and *Coprococcus* distinguished tumor tissues from normal tissues in CIMP-H and CIMP-L patients. There were several additional distinguishing factors between tissue types unique to CIMP-H patients, including an increase of *Campylobacter* and a decrease of *Blautia.* Again, no genera distinguished tumor tissues from normal tissues in Non-CIMP° patients. These unique and consistent findings across our different analyses may suggest that either differences in microbiota between tumor and normal tissues may modify methylation status of CpG Islands, or that the differences in methylation profile between the tumor and normal tissues somehow modify the tissue microbiome.

Associations between CpG Island Methylation Markers and Butyrate Producing Bacteria.

While we discovered many characteristics of the microbiome that distinguished CIMP patients from Non-CIMP patients, we were able to show that total relative abundance of common intestinal butyrate producing bacteria in tumor tissues was inversely associated with CIMP subtype of CRC, even after adjusting for sex, resection side, and age. We also found a trending inverse association between total relative abundance of butyrate producing bacteria and CIMP CRC in normal tissues, though this association did not reach statistical significance. The total abundance of BPBs was also significantly lower in tumors in paired analyses of patient matched tumor and normal tissues, in CIMP-H and CIMP-L groups, while Non-CIMP° tissues showed only a trend of decreased BPB in tumors.

As mentioned above, there were a few genera of butyrate producers that were individually found to be significantly and inversely associated with CIMP CRC. All BPB genera with an average relative abundance above >0.01% were significantly associated with CIMP in our univariable analyses in both tumor tissues and normal tissues. *Faecalibacterium, Coprococcus,* and *Blautia* were also distinguishing features of PAM Cluster 1 in our LEfSe analyses. Our unsupervised topological model showed that the subnetwork enriched in Non-CIMP and normal tissues was significantly associated *Faecalibacterium, Coprococcus, Blautia,* and *Roseburia,* while the subnetwork enriched in CIMP-H and tumor tissues showed decreases in these four bacterial genera.

Moreover, we found significant correlations between relative abundance of *Blautia, Coprococcus,* and *Faecalibacterium,* in tumor tissues and methylation of CIMP-specific markers. Most notably, *Blautia* was negatively correlated with every marker with the exception of CDKN2A. Collectively, this evidence suggests that not only is total relative abundance of butyrate producing bacteria decreased in CIMP CRCs, but also that reduction of abundance of *Blautia, Coprococcus,* and *Faecalibacterium* specifically are associated methylation of CIMP-specific markers. However, it is important to note that while BPB are inversely associated with these methylation markers, other taxa, including Fusobacterium are positively associated. Thus, either the deficiency in BPB or the increase in Fusobacterium or other taxa could influence or be influenced by methylation.

The Impact of Decreased BPB and Increased *Fusobacterium* in Tumorigenesis of CIMP CRC. Our research showed that the overall relative abundance of butyrate producing bacteria, as well as the abundance of specific BPB, including *Blautia, Coprococcus,* and *Faecalibacterium*, are reduced in CIMP tumor tissues, compared to Non-CIMP tumor tissues. In contrast, there is a greater abundance of *Fusobacterium* and other microbes normally found in the oral microbiome in CIMP tumors. Other studies have reported an increased relative abundance of *Bacteroidaceae, Streptococcaeceae,*

Fusobacteriaceae, Peptostreptococcaceceae, Verillonellaceae, and *Pasteurellaceae,* and a decrease in *Lachnoospiraceae, Ruminococcaceae,* and *Lactobacillaceae* when comparing cancerous colorectal tissues to normal intestinal tissue.[32] Similarly, *Bifidobacterium, Faecalibacterium,* and *Blautia* were found to be reduced in the gut microbiome of CRC patients, compared to normal patients.[34] However, our study is unique in that we stratified by CIMP subtypes to perform subgroup analyses of the tumor environment of CIMP and Non-CIMP CRC patients.[34]

Butyrate producing bacteria in the Firmicutes phylum degrade dietary fibers, complex carbohydrates, and plant polysaccharides that human enzymes are unable to digest.[36] Butyrate regulates gene expression by acting as a histone deacetylase inhibitor (HDACi), and has been estimated to regulate up to 2% of the human transcriptome through this mechanism.[36] The antineoplastic effects of butyrate have been well established, and current literature suggests that the ability of butyrate to interact with chromatin modification machinery may provide a protective effect against colorectal cancer.[9] HDACs have been shown to have a silencing effect on tumor suppressor genes that play a critical role in tumorigenesis of the CIMP CRC pathway.[37] HDAC inhibitors may restore expression of tumor suppressor genes, induce cancer cell death, may sensitize cancer cells to treatment, and may have anti-inflammatory effects on the colonic tissue.[37]

While we found significant differences in relative abundance of BPB in tumor tissues between CIMP-H, CIMP-L, and Non-CIMP° patients, we did not find a significant association between the functional PICRUSt pathway, butanoate metabolism, and CIMP status. This may be due in part due to the fact that *Fusobacterium* can also synthesizes butyrate, though *Fusobacterium* does so by fermenting lysine, rather than through fermentation of dietary plant fibers and carbohydrates.[38] Since we did find that lysine degradation was significantly associated with our subnetwork in our unsupervised model that was most closely associated with CIMP-H tumor tissues, TM1, this may explain why we fail to detect a

difference in PICRUSt butanoate metabolism, despite the noted difference in abundance of BPBs between CIMP and Non-CIMP tumors.

Interestingly, in contrast to BPB in the Firmicutes cluster, *Fusobacterium* is known to be associated with oncogenesis and inflammation.[34] These opposing effects and opposing associations in our study may indicate that patients with reduced abundance of BPBs and increased abundance of *Fusobacterium* may be at a higher risk of acquiring epigenetic modifications that give rise to CIMP subtypes of CRC, while CRC patients without this microbial signature are less likely to acquire aberrant methylation profiles. If BPB is in fact important in the tumorigenesis of CIMP CRCs, this suggests that prebiotics that support the growth of these bacterial taxa may be beneficial to colonic health.

Study Strengths and Limitations. Our research had several limitations. A major limitation is that the progression of a colorectal polyp to a CIMP colorectal tumor happens gradually over time, as the CIMP-specific markers become silenced through hypermethylation. Definitions of CIMP-H, CIMP-L, and Non-CIMP° follow current conventions, and may imperfectly represent our data. For example, there may be patients in the CIMP-L or Non-CIMP° category that may have eventually gone on to have methylation for subsequent CIMP markers, had the biopsy on the CRC tumor been performed later in the development of the tumor. Furthermore, cutoffs between groups based on the number of methylation markers are not based on clinical relevance. There may not be a significant clinical difference between patients at the lower end of the CIMP-H group and the upper end of the CIMP-L group, as methylation progresses as a gradient from zero to eight methylated markers. Similarly, there may be significant clinical differences between patients at the upper end of our CIMP-L category and the lower end of the CIMP-L category.

Additionally, our study relied on a well-established panel of eight CIMP-associated markers. Global methylation patterns across the entire genome may have provided better categorization of CRC subtypes and may have provided stronger evidence for associations between individual bacterial taxa and either methylation patterns or individual CpG islands.

Another limitation is that this study relied on sequencing of the short V4 region of the bacterial 16S rRNA gene. While this is the state-of-the-art method for assessing the microbiome using the 16S rRNA gene, this short variable region of the 16S provides somewhat poor classification for many bacterial taxa at the species-level. For this reason, we have focused the majority of our analyses at the genus-level. While the genera that we evaluated are well-known BPB genera, some species that are butyrate producers, but could not be classified due to poor resolution of the V4 region, were undoubtedly missed. We cannot exclude the possibility that species-level assignment of bacterial taxa would have revealed additional associations, or that some genera may have been misclassified. However, since we have no evidence to suggest that differential misclassification of bacterial rRNA by CIMP marker profile occurred.

Our regression analyses may have been underpowered to detect a true relationship between BPB and CIMP designation when controlling for three confounding variables. We tested the robustness of our results for our logistic regressions using bootstrapped estimates and standard errors of our fully adjusted models; however, it would be incorrect to assume that testing the robustness of our results is sufficient to address the lack of power in our regression analyses. In our topological data analysis and LEfSe analyses, we analyzed differences within and between CIMP-H, CIMP-L, and Non-CIMP° patients. By splitting off patients with no methylation markers into their own category, our sample size was small for Non-CIMP° patients, and may have been underpowered to detect true associations within this CIMP designation, or between this category and others.

Another challenge with our regression analyses is that we cannot rule out the possibility of residual confounding in our analysis. For example, we were not able to control for other genera that many potentially confound the relationship between BPB and CIMP-H status in our regression analyses of total BPB since relative abundance of a confounding genera is not independent of our BPB variable. If a confounding genus was held constant in terms of absolute counts, while a decrease in BPB was observed, the relative abundance of the confounding variable would increase despite there being no change in the absolute abundance. Therefore, we are unable to conclude if the decrease in relative abundance of BPB in CIMP tumors compared to Non-CIMP tumors is driven by a true decrease in the abundance of BPB or by an increase in the abundance of other taxa, such as *Fusobacterium.* Future analyses of tumor samples will be used to quantify absolute counts of BPB, *Fusobacterium,* and other potential confounding genera so that we can more closely model the relationship between BPB and CIMP status after accounting for changes in the quantity of these other taxa.

Additionally, due to the high dimensionality of our dataset and our multifaceted approach to investigating our research question, we cannot exclude the possibility that one or more of the findings we found to be statistically significant was a false discovery. However, since we used LEfSe analyses, which has been shown to reduce false discovery rate in highly dimensional datasets, as a gatekeeping method, and applied adjustments for multiple corrections to our findings, it is unlikely that false discoveries occurred.

Finally, our study was cross section in design, and cannot be used to establish temporality or causality between the microbial environment and the methylation phenotype of these tissues. We are unable to determine whether the composition of the microbiome in the tissue preceded methylation, or if epigenetic changes impacted the makeup of the microbial community. It is possible that methylation events change the microenvironment in a way that makes it more suitable for tumor-associated bacteria

to colonize the tissue, though the bacteria did not cause the epigenetic changes. Even if we could have established temporality, we are unable to distinguish whether microbes that varied significantly among groups were “drivers”, meaning bacteria that directly caused changes in the epigenome or “passengers” that co-occur alongside “drivers”, but do not themselves cause any epigenetic modifications. Additional mechanistic studies in animal models, organoids or other culture systems may shed light on these associations.

REFERENCES/BIBLIOGRAPHY

1. Siegel, R. L., Miller, K. D. & Jemal, A. Cancer statistics, 2018. *CA. Cancer J. Clin.* **68**, 7–30 (2018).
2. O??Brien, M. J. *et al.* Comparison of Microsatellite Instability, CpG Island Methylation Phenotype, BRAF and KRAS Status in Serrated Polyps and Traditional Adenomas Indicates Separate Pathways to Distinct Colorectal Carcinoma End Points: *Am. J. Surg. Pathol.* **30**, 1491–1501 (2006).
3. Vaughn, C. P., Wilson, A. R. & Samowitz, W. S. Quantitative evaluation of CpG island methylation in hyperplastic polyps. *Mod. Pathol. Augusta* **23**, 151–6 (2010).
4. Chang, L., Chang, M., Chang, H. M. & Chang, F. Expending Role of Microsatellite Instability in Diagnosis and Treatment of Colorectal Cancers. *J. Gastrointest. Cancer* **48**, 305–313 (2017).
5. Cha, Y. *et al.* Adverse prognostic impact of the CpG island methylator phenotype in metastatic colorectal cancer. *Br. J. Cancer* **115**, 164–171 (2016).
6. Advani, S. M. *et al.* Clinical, Pathological, and Molecular Characteristics of CpG Island Methylator Phenotype in Colorectal Cancer: A Systematic Review and Meta-analysis. *Transl. Oncol.* **11**, 1188–1201 (2018).
7. Qin, J. *et al.* A human gut microbial gene catalogue established by metagenomic sequencing. *Nat. Lond.* **464**, 59–65 (2010).
8. Bierne, H., Hamon, M. & Cossart, P. Epigenetics and Bacterial Infections. *Cold Spring Harb. Perspect. Med.* **2**, a010272–a010272 (2012).
9. Leonel, A. J. & Alvarez-Leite, J. I. Butyrate: implications for intestinal function. *Curr. Opin. Clin. Nutr. Metab. Care* **15**, 474–479 (2012).
10. Pryde, S. E., Duncan, S. H., Hold, G. L., Stewart, C. S. & Flint, H. J. The microbiology of butyrate formation in the human colon. *FEMS Microbiol. Lett.* **217**, 133–139 (2002).

11. Hippe, B. *et al.* Quantification of butyryl CoA: acetate CoA-transferase genes reveals different butyrate production capacity in individuals according to diet and age. *FEMS Microbiol. Lett.* **316**, 130–135 (2011).

12. Hicks, A. L. *et al.* Gut microbiomes of wild great apes fluctuate seasonally in response to diet. *Nat. Commun.* **9**, (2018).

13. Walker, A. W. *et al.* Dominant and diet-responsive groups of bacteria within the human colonic microbiota. *ISME J. Lond.* **5**, 220–30 (2011).

14. Weisenberger, D. J. *et al.* CpG island methylator phenotype underlies sporadic microsatellite instability and is tightly associated with BRAF mutation in colorectal cancer. *Nat. Genet.* **38**, 787–793 (2006).

15. Jakubauskas, A. & Griskevicius, L. KRas and BRaf Mutational Status Analysis from Formalin-Fixed, Paraffin-Embedded Tissues Using Multiplex Polymerase Chain Reaction–Based Assay. *Arch Pathol Lab Med* **134**, 5 (2010).

16. Rosenberg, D. W. *et al.* Mutations in *BRAF* and *KRAS* Differentially Distinguish Serrated versus Non-Serrated Hyperplastic Aberrant Crypt Foci in Humans. *Cancer Res.* **67**, 3551–3554 (2007).

17. Zou, Y. *et al.* Mutational analysis of the RAS/RAF/MEK/ERK signaling pathway in 260 Han Chinese patients with cervical carcinoma. *Oncol. Lett.* **14**, 2427–2431 (2017).

18. Caporaso, J. G. *et al.* QIIME allows analysis of high-throughput community sequencing data. *Nat. Methods* **7**, 335–336 (2010).

19. Cremer, J., Arnoldini, M. & Hwa, T. Effect of water flow and chemical environment on microbiota growth and composition in the human colon. *Proc. Natl. Acad. Sci.* 201619598 (2017). doi:10.1073/pnas.1619598114

20. Mima, K. *et al.* Fusobacterium nucleatum in Colorectal Carcinoma Tissue According to Tumor Location. *Clin. Transl. Gastroenterol.* **7**, e200 (2016).

21. den Besten, G. *et al.* The role of short-chain fatty acids in the interplay between diet, gut microbiota, and host energy metabolism. *J. Lipid Res.* **54**, 2325–2340 (2013).

22. Walker, A. W., Duncan, S. H., Leitch, E. C. M., Child, M. W. & Flint, H. J. pH and Peptide Supply Can Radically Alter Bacterial Populations and Short-Chain Fatty Acid Ratios within Microbial Communities from the Human Colon. *Appl Env. Microbiol* **71**, 3692–3700 (2005).

23. Salazar, N., Valdés-Varela, L., González, S., Gueimonde, M. & de los Reyes-Gavilán, C. G. Nutrition and the gut microbiome in the elderly. *Gut Microbes* **8**, 82–97 (2016).

24. Haro, C. *et al.* Intestinal Microbiota Is Influenced by Gender and Body Mass Index. *PLoS ONE* **11**, (2016).

25. Arumugam, M. *et al.* Enterotypes of the human gut microbiome. *Nature* **473**, 174–180 (2011).

26. *AJCC cancer staging manual*. (Springer, 2010).

27. Segata, N. *et al.* Metagenomic biomarker discovery and explanation. *Genome Biol.* **12**, R60 (2011).

28. Westfall, P. H. & Krishen, A. Optimally weighted, fixed sequence and gatekeeper multiple testing procedures. *J. Stat. Plan. Inference* **99**, 25–40 (2001).

29. Dmitrienko, A., Offen, W. W. & Westfall, P. H. Gatekeeping strategies for clinical trials that do not require all primary effects to be significant. *Stat. Med.* **22**, 2387–2400 (2003).

30. Carlsson, G. Topology and data. *Bull. Am. Math. Soc.* **46**, 255–308 (2009).

31. Lum, P. Y. *et al.* Extracting insights from the shape of complex data using topology. *Sci. Rep.* **3**, (2013).

32. Lozupone, C. A., Stombaugh, J. I., Gordon, J. I., Jansson, J. K. & Knight, R. Diversity, stability and resilience of the human gut microbiota. *Nature* **489**, 220–230 (2012).

33. Schloissnig, S. *et al.* Genomic variation landscape of the human gut microbiome. *Nature* **493**, 45–50 (2013).

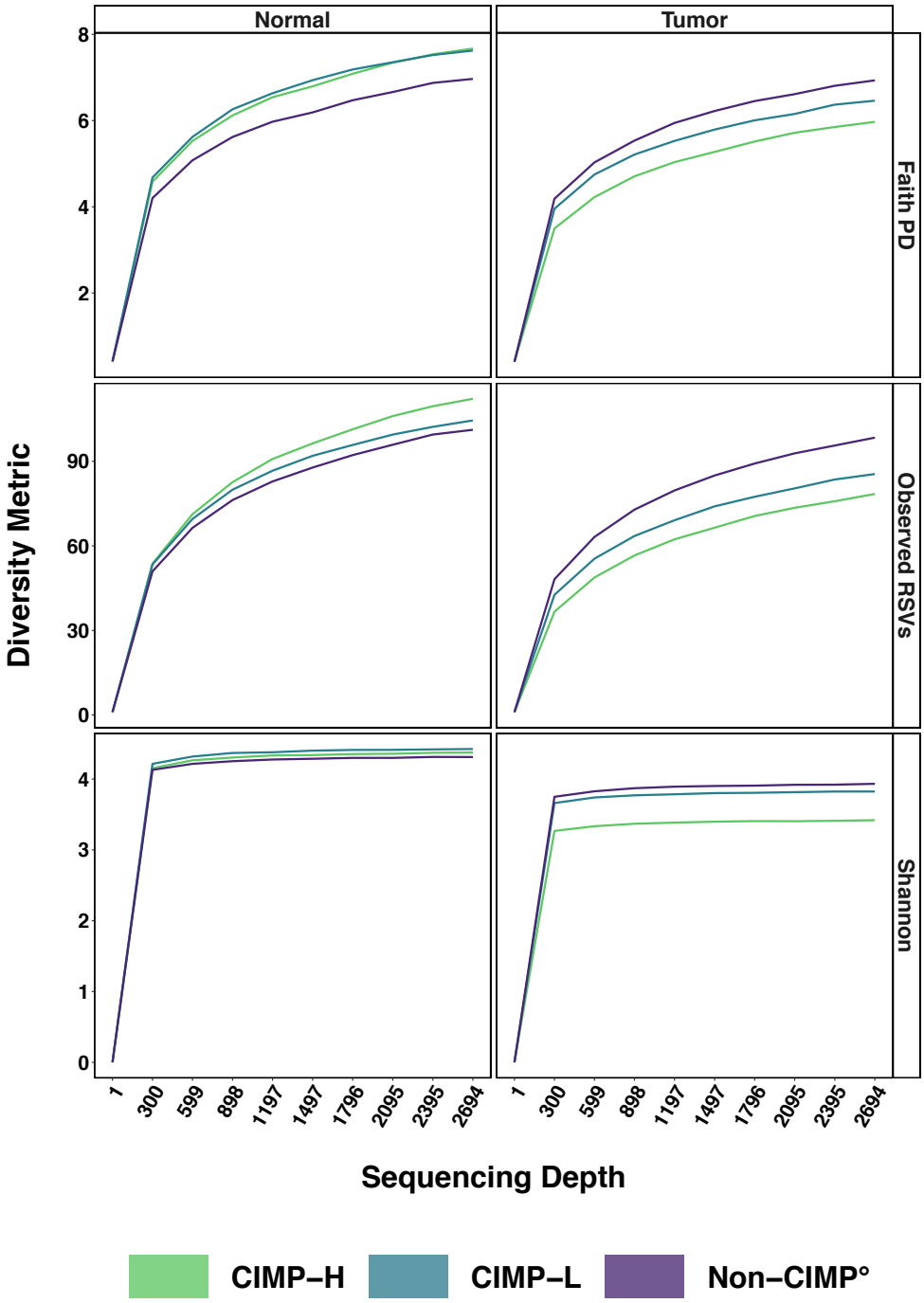
34. Cho, M., Carter, J., Harari, S. & Pei, Z. The Interrelationships of the Gut Microbiome and Inflammation in Colorectal Carcinogenesis. *Clin. Lab. Med.* **34**, 699–710 (2014).

35. Ahn, J. *et al.* Oral microbiome profiles: 16S rRNA pyrosequencing and microarray assay comparison. *PloS One* **6**, e22788 (2011).

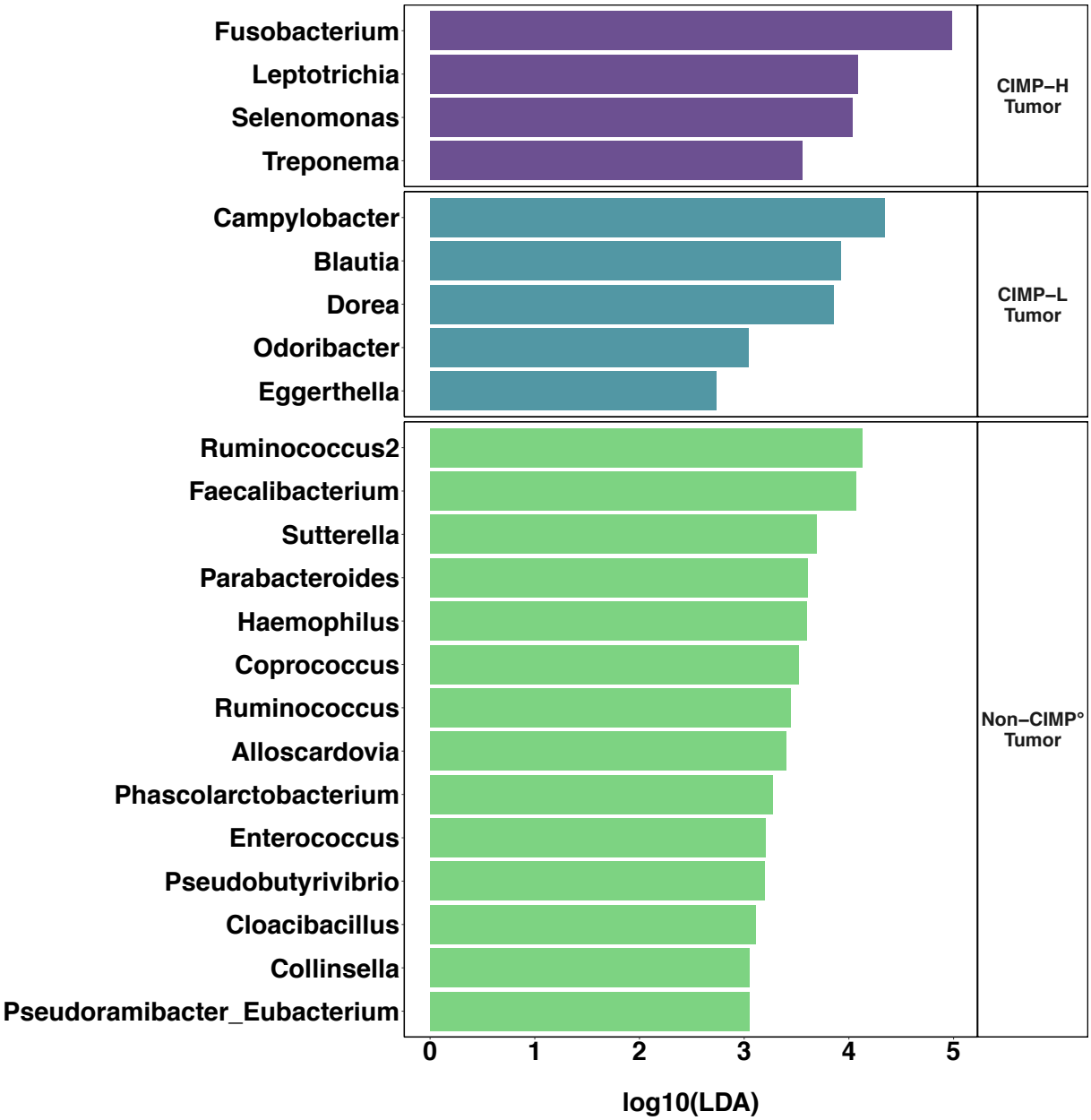
36. Tremaroli, V. & Bäckhed, F. Functional interactions between the gut microbiota and host metabolism. *Nat. Lond.* **489**, 242–9 (2012).

37. Lutz, L. *et al.* Histone modifiers and marks define heterogeneous groups of colorectal carcinomas and affect responses to HDAC inhibitors in vitro. *Am. J. Cancer Res.* **6**, 664–676 (2016).

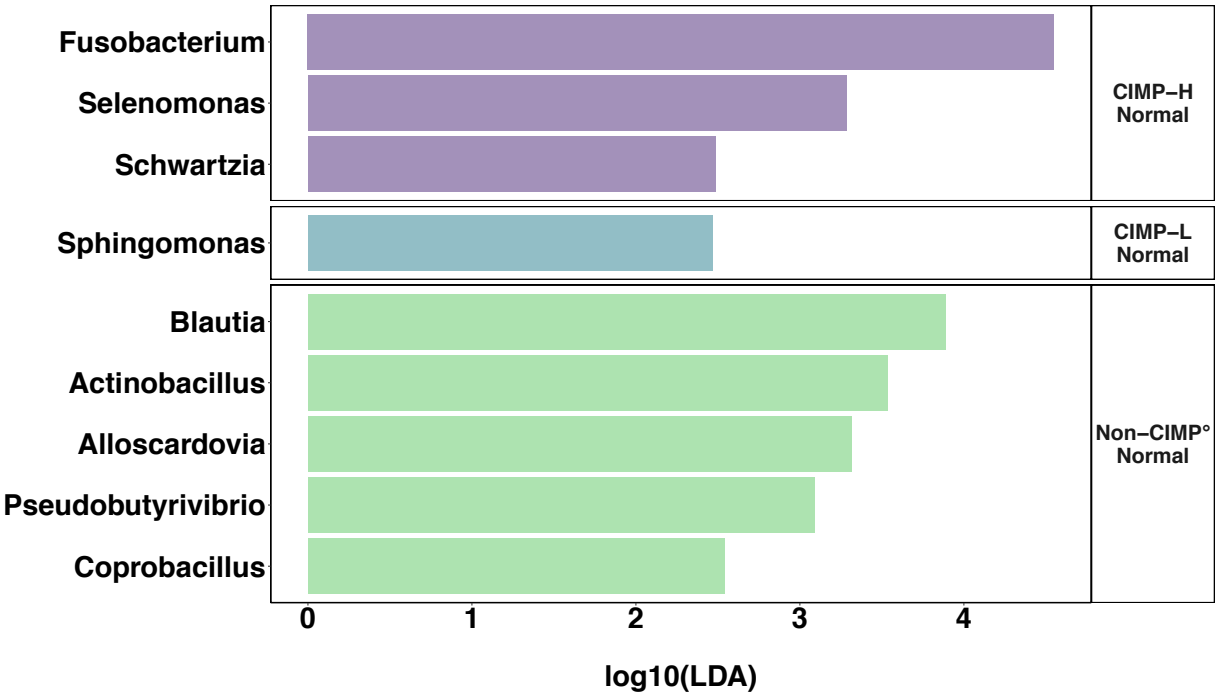
38. Kreimeyer, A. *et al.* Identification of the Last Unknown Genes in the Fermentation Pathway of Lysine. *J. Biol. Chem.* **282**, 7191–7197 (2007).



Supplemental Figure 1: Alpha Rarefaction by grouped by CIMP designation and stratified by tissue type. *Top two plots show normal and tumor tissue alpha rarefaction curves for Faith Phylogenetic Diversity Index. Middle two plots show normal and tumor tissue alpha rarefaction curves for Observed RSVs. Bottom two plots show normal and tumor tissue alpha rarefaction curved for Shannon Diversity Index. With rarefaction plots, we hope to see a plateau of diversity to indicate that additional reads would not be likely to find additional genera. While our Shannon rarefactions do show a plateau, our observed RSVs and Faith PD plots indicates that sequencing depth beyond 2694 may have resulted in finding additional genera in our samples.*



Supplemental Figure 2: CIMP-H Tumor vs CIMP-L Tumor vs Non-CIMP° Tumor Tissues LEfSe Results. *This figure shows all genera and their respective log10(LDA) scores for all genera that had log10(LDA) scores above 2, indicating they significantly distinguished between CIMP-H tumor, CIMP-L tumor, and Non-CIMP° tumor tissues.*

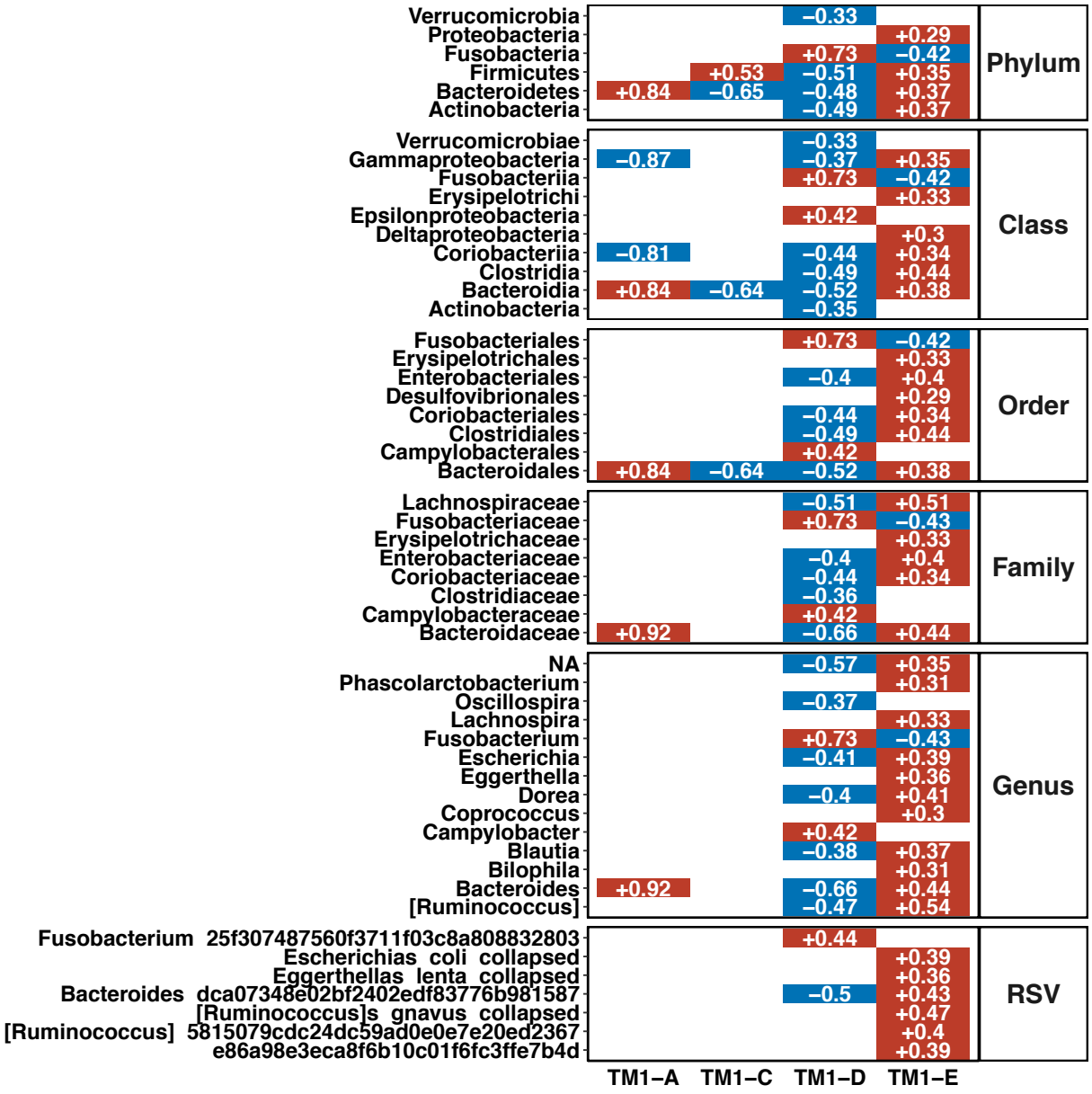


Supplemental Figure 3: CIMP-H Normal vs CIMP-L Normal vs Non-CIMP° Normal Tissues LEfSe Results. *This figure shows all genera and their respective log10(LDA) scores for all genera that had log10(LDA) scores above 2, indicating they significantly distinguished between CIMP-H normal, CIMP-L normal and Non-CIMP° normal tissues.*

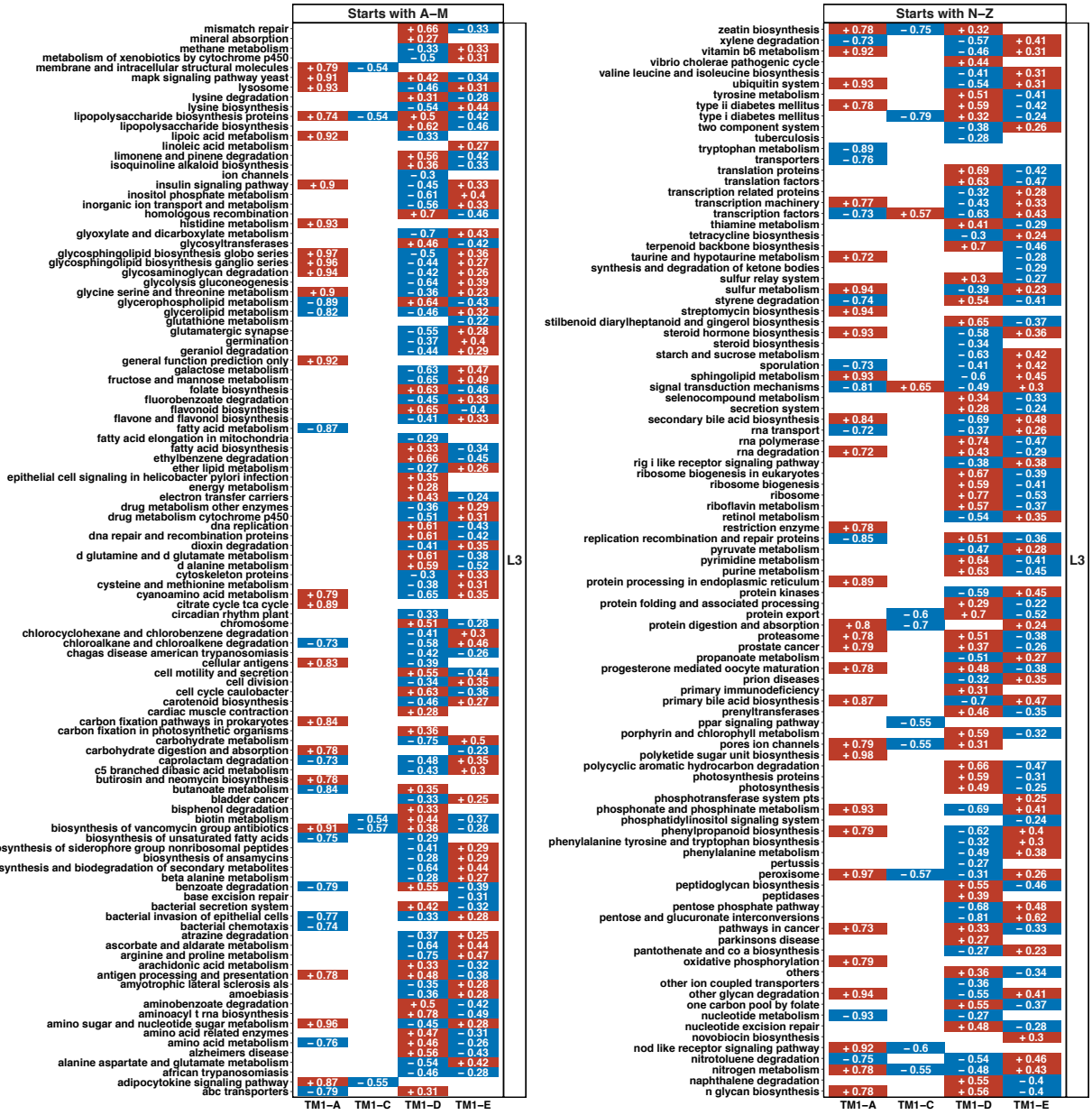
Supplemental Table 1: Wilcoxon Signed-Rank Test with Unadjusted and Adjusted Benjamini-Hochberg Correction for Multiple Tests Between Paired Tumor and Normal Tissues for Genera identified in LEfSe gatekeeping analyses[1]

	CIMP-H		CIMP-L		Non-CIMP°	
	Difference in Medians[2]	**p-value**	**Difference in Medians[2]**	**p-value**	**Difference in Medians[2]**	**p-value**
Actinobacillus	0.00	p = 1.0000	0.00	*NA*	0.00	p = 0.5847
Actinomyces	0.00	p = 0.0797	0.00	p = 0.1825	0.00	p = 0.8546
Alloscardovia	0.00	*NA*	0.00	*NA*	0.00	p = 0.9934
Bacteroides	**-15.55**	**p = 0.0015**	-6.59	p = 0.2646	-11.05	p = 0.8546
Blautia	**-1.05**	**p = 0.0004**	-0.65	p = 0.2915	-0.99	p = 0.5847
Campylobacter	**0.12**	**p = 0.0016**	0.01	p = 0.1444	0.00	p = 0.5847
Cloacibacillus	0.00	p = 0.1129	0.00	p = 0.6325	0.00	p = 0.9934
Collinsella	**-0.03**	**p = 0.0161**	-0.12	p = 0.1444	-0.11	p = 0.9934
Coprobacilus	**0.00**	**p = 0.0406**	0.00	p = 0.3078	0.00	p = 0.5847
Coprococcus	**-0.44**	**p = 0.0027**	-0.56	p = 0.0642	-0.12	p = 0.8546
Dorea	**-0.44**	**p = 0.0007**	-1.16	p = 0.1444	-0.49	p = 0.5846
Eggerthella	**-0.02**	**p = 0.0013**	-0.03	p = 0.4824	-0.02	p = 0.8546
Enterococcus	**0.00**	**p = 0.0339**	0.00	p = 0.6228	0.00	p = 0.9934
Faecalibacterium	**-1.25**	**p = 0.0002**	-2.53	p = 0.1857	-0.36	p = 0.8546
Fusobacterium	**13.41**	**p < 0.0001**	**4.15**	**p = 0.0083**	0.35	p = 0.5847
Granulicatella	**-0.01**	**p = 0.0231**	-0.01	p = 0.5480	0.00	p = 0.9934
Haemophilus	**-0.01**	**p = 0.0266**	-0.01	p = 0.5211	0.01	p = 0.5847
Leptotrichia	**0.25**	**p = 0.0003**	0.00	p = 0.1857	0.00	p = 0.5847
Odoribacter	**0.00**	**p = 0.0223**	-0.01	p = 0.7457	0.00	p = 0.9934
Parabacteroides	**-0.40**	**p = 0.0009**	-0.26	p = 0.1444	-0.07	p = 1.0000
Phascolarctobacterium	**-0.03**	**p = 0.0101**	-0.09	p = 0.4812	-0.09	p = 0.5847
Pseudobutyrivibrio	0.00	*NA*	0.00	*NA*	0.00	p = 0.9934
Ruminococcus	**-0.09**	**p = 0.0027**	-0.11	p = 0.1444	0.00	p = 0.9934
Schwartzia	0.00	p = 0.1293	0.00	p = 0.1870	0.00	p = 0.8546
Sediminibacterium	0.00	p = 0.1129	0.00	p = 0.4824	0.00	p = 1.0000
Selenomonas	**0.05**	**p = 0.0012**	0.00	p = 0.7086	0.00	p = 0.9934
Sphingomonas	0.00	p = 0.5850	0.00	p = 0.1444	0.00	p = 1.0000
Sutterella	**-0.18**	**p = 0.0223**	-0.18	p = 0.1444	-0.01	p = 1.0000
Treponema	0.00	p = 0.05728	0.00	p = 0.2646	0.00	p = 0.9934

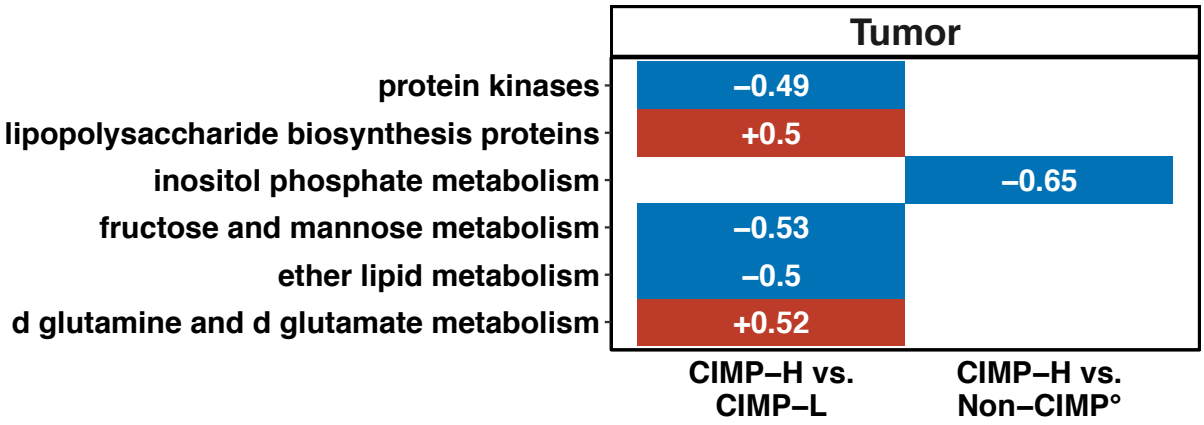
1. *Genera significant at p < 0.05 after adjustment for multiple corrections are shown in bold.*
2. *Difference in medians of percent relative abundance, Tumor – Normal.*



Supplemental Figure 4: Significant Taxa at Phylum, Class, Order, and Family Levels in TM1 Subnetworks. *Kolmogorov-Smirnov test for significance was used, and all genera that were significant at $\alpha = 0.05$ after adjusting for multiple testing with Benjamini Hochberg methods within each taxonomic level are shown. Red indicates a positive association between the genera and the sub-network, while blue indicates an inverse association. Kolmogorov-Smirnov test statistics are displayed. For more information regarding TM-1 and subnetworks, see Figure 13.*



Supplemental Figure 5: Significant PICRUSt L3 reads for TM-1. *Kolmogorov-Smirnov test for significance was used, and all genera that were significant at $\alpha = 0.05$ after Benjamini Hochberg adjustment for multiple tests are shown. Red indicates as positive association between the genera and the sub-network, while blue indicates an inverse association. Kolmogorov-Smirnov test statistics are displayed. For more information regarding TM-1 and subnetworks, see Figure 13.*



Supplemental Figure 6: Significant PICRUSt L3 reads for Supervised Between-Group Comparisons in Ayasdi, stratified by tissue type. *Kolmogorov-Smirnov test for significance was used, and all genera that were significant at $\alpha = 0.05$ are shown. Red indicates a positive association between the genera and the sub-network, while blue indicates an inverse association. Kolmogorov-Smirnov test statistics are displayed.*