Tumor Microbial Biodiversity and Microsatellite Instability in Colorectal Cancer

Calen Peter Mendall

A thesis

submitted in partial fulfillment of the

requirements for the degree of

Master of Science

University of Washington

2025

Committee:

Amanda I. Phipps

Meredith A. J. Hullar

Program Authorized to Offer Degree:

Epidemiology

University of Washington

Abstract

Tumor Microbial Biodiversity and Microsatellite Instability in Colorectal Cancer

Calen Peter Mendall

Chair of the Supervisory Committee

Amanda I. Phipps

Epidemiology

Background: There is growing interest in characterizing how the gut microbiome is related to colorectal cancer (CRC) progression and etiology after several species have been found to be strongly enriched in certain molecular subtypes of CRC tumors. DNA mismatch repair deficiency in CRC is associated with a favorable prognosis and positive response to immunotherapy, and it can be characterized by the presence of tumor microsatellite instability (MSI). Here, we explore the relationship between MSI status and tumor microbial biodiversity.

Methods: The Seattle site of the Colon Cancer Family Registry (SCCFR) recruited patients diagnosed with incident CRC from 1998 – 2007. Tumor tissue samples were assessed for MSI status and targeted sequencing of the V3-V4 hypervariable region of the prokaryotic *16S rRNA* gene was performed. We used an adaptive test of alpha diversity (aMiAD) that incorporates a set of abundance-based and phylogenetic alpha diversity metrics to estimate the association of microbial biodiversity with MSI status. Differential abundance analysis was performed using ANCOM-BC to identify specific enriched genera in tumor tissue

based on MSI status. The alpha diversity analyses and differential abundance analyses were both adjusted for age, sex, smoking history, and tumor location in the subset of the SCCFR patients with complete profiling of all measures (N = 632).

Results: The adaptive aMiAD effect estimate was -1.08 ($p = 0.29$), indicating a non-statistically significant negative association of alpha diversity on MSI status, with MSI-high tumors having lower estimated alpha diversity. We identified 20 differentially abundant genera in MSI tumors, with 8 enriched genera and 12 depleted genera in MSI-high tumors. The genera that were most strongly enriched in MSI-high tumors were *Gemella* and *Lawsonella*, while the most strongly depleted genera in MSI-high tumors were *Sporolactobacillaceae* and *Cloacibacterium*.

Conclusions: We failed to detect an association between any of the individual alpha diversity measures or the adaptive alpha diversity measure with MSI status, though all measures concordantly estimated a depletion of alpha diversity in the MSI-high tumors. We find evidence of differential abundance of certain genera dependent on MSI status, including several genera that have not been previously identified as associated with MSI status.

INTRODUCTION

Colorectal cancer (CRC) is a heterogeneous disease with several well-characterized molecular phenotypes and subtypes, which have been suggested to vary in their etiology and exhibit differing clinical features and prognoses.[1] Microsatellite instability (MSI) is a characteristic marker of the mismatch repair deficient (MMR) CRC phenotype, where deleterious mutations or epigenetic silencing of MMR genes results in non-functional or unexpressed proteins responsible for correcting DNA replication errors.[1,2] High levels of MSI (MSI-high) in a tumor is present in an estimated 12 - 17% of CRC cases;[2] it is generally associated with a more favorable prognosis and responsiveness to certain immunotherapy treatments.[1,2] Despite having a hereditary component, sporadic pathogenesis of MMR deficiency occurs in 9 - 14% of CRC cases,[2] making it important to understand the factors leading to sporadic MMR deficient CRC.

The microbiome present in CRC tumors can differ in composition compared to the microbiome in normal colorectal tissue.[3] Dysbiosis of the colorectal tissue-associated microbiome is anticipated to have a causal role in pathogenesis of CRC.[4,5] At the genera level, *Bacteroides, Escherichia, Fusobacterium,* and *Salmonella* are associated with and enriched in CRC tumor tissue.[3] Furthermore, several bacterial species are enriched in CRC tumor tissue;[3,6,7] *Fusobacterium nucleatum* (*F. nucleatum*), in particular, is strongly associated with CRC, where it is suspected to induce inflammation in colorectal tissue.[4,8] Inflammation of colorectal tissue appears to play an important role in the development of sporadic MMR-deficient (i.e., MSI-high) CRC.[2] A strong relationship between *F. nucleatum* colonization and CRC has been identified by numerous groups,[8,9] with additional associations detected particularly between *F. nucleatum* and the MMR-deficient (MSI-high) phenotype.[10,11] *F. nucleatum* colonization is associated with a worse prognosis,[11,12] while the MSI-high phenotype is generally associated with a favorable prognosis;[1,2] thus, understanding the factors that contribute to this contrast remains important.

While certain individual bacterial species are likely to play a role in shaping CRC phenotype and prognosis, aspects of the overall composition of the microbiome in CRC may also be relevant. Alpha diversity summarizes how taxonomically varied and balanced a microbial community is within an ecological niche of interest.[13] Measuring alpha diversity can help us understand whether there are broad compositional differences in the tumor microbiome of MSI CRC.[4] There are a number of commonly used alpha diversity metrics that prioritize different taxa based on relative abundance in the community and phylogenetic relationships between taxa. Without *a priori* reasons to select one measure over others, it is common to use multiple measures. However, it is challenging to interpret results when the measures

disagree. Adaptive tests help resolve this by providing valid tests and effect estimates that compare a set of pre-specified alpha diversity metrics and maintain control of type I error rates.[13]

Associations between microbial biodiversity and MSI-high CRC have been identified previously, with higher biodiversity found in MMR-deficient (MSI-high) CRC tumors compared to tumors with proficient MMR (MSS/MSI-low, reflecting intact MMR systems) and genera specific shifts in composition found.[14–17] The generalizability of these results remains uncertain since the estimated effect depended on the choice of alpha diversity metric and not all previous works adjusted for important confounding variables.[14–16] Additionally, these previous works only considered relative abundance-based alpha diversity metrics; phylogeny based metrics may capture important features of community structure in the tumor microbiome which goes unmeasured with purely abundance based alpha diversity metrics.[18]

Here, we explore the association between tumor alpha diversity and MSI in CRC using an adaptive measure of alpha diversity (aMiAD) that incorporates both abundance-based alpha diversity metrics and phylogeny-based metrics,[13] and adjust for important confounding measures. We also explore whether *F. nucleatum* positivity in tumor tissue affects the presence of an association between alpha diversity and MSI in CRC based on the presence of the *F. nucleatum nusG* gene by ddPCR. Finally, we searched for specific genera that were differentially abundant in MSI-high tumors compared to MSS/MSI-low tumors to identify taxa that may be driving associations between alpha diversity and MSI status.

METHODS

Study Population

Data for this investigation comes from the Seattle site of the Colon Cancer Family Registry (SCCFR).[19] The SCCFR conducted population-based recruitment of participants with incident diagnoses of invasive CRC, identified through the Seattle-Puget Sound registry site of the Surveillance, Epidemiology, and End Results (SEER) cancer registry network. Recruitment was performed in two phases; in Phase I, recruited CRC cases resided in select Washington counties and were aged 20 to 74 years old with diagnosis dates between January 1998 and June 2002. In Phase II, recruited incident CRC cases were all younger than 50 years old at the time of diagnosis, were from an extended set of Washington counties, and were diagnosed between April 2002 to July 2007. All participants gave informed consent for the collection and analysis of their biospecimens, medical reports, and questionnaire responses. Review and approval of this study was performed and given by the Institutional Review Board at the Fred Hutchinson Cancer Center.

Biospecimen and Data Collection

Participants completed a questionnaire at enrollment that collected basic demographic information, family history of cancer, limited aspects of diet, and exposure to potential risk factors of CRC. Biospecimens were collected from tumor tissue, and were stored in formalin-fixed paraffin-embedded (FFPE) blocks. Tissue samples were assessed for several molecular markers, including MSI status, somatic mutations in *KRAS* and *BRAF*, and the CpG island methylator phenotype (CIMP), as described elsewhere.[20] For a subset of SCCFR case participants (N = 898), characterization of the tumor-associated microbiome was also conducted, including sequencing of the bacterial *16S rRNA* gene, targeting the V3-V4 hypervariable region, using microbial DNA extracted from tumor biopsy samples with a modified extraction protocol for FFPE samples.[21] Processing of *16S rRNA* gene sequencing and classification of amplicon sequencing variants (ASVs) with SILVA taxonomy was performed at the Fred Hutchinson Cancer Center.[22] Additionally, targeted ddPCR of the *F. nucleatum* specific homolog of the *nusG* gene was used to identify whether tissue samples were positive for the species.[23]

Inclusion criteria

All participants needed complete classification of MSI status and to have passed *16S rRNA* gene sequencing quality control for inclusion in the analysis (N = 632). For analyses involving a participant's tumor *F. nucleatum* positivity status, successful classification of *F. nucleatum* positivity status by targeted ddPCR of the *nusG* gene was also required (N = 627).

Measures

Tumor-specific bacterial alpha diversity was measured from the tissue samples of CRC tumors using the processed *16S rRNA* gene sequencing. The primary analysis was performed using an adaptive measure of diversity (aMiAD), which incorporates six measures of alpha diversity: observed Richness,[24] the Shannon index,[25] the Simpson index,[26] Phylogenetic Diversity (PD),[27] Phylogenetic Entropy (PE),[28] and Phylogenetic Quadratic Entropy (PQE),[29] to estimate the magnitude and direction of association between alpha diversity and MSI in tumors. Different taxa are prioritized depending on selection of alpha diversity metric: the Simpson index up-weights the most relatively abundant taxa, the Shannon index moderately weights each taxa by their relative abundance, and Richness gives equal weight all taxa.[13] PD, PE, and PQE are analogous to Richness, the Shannon index, and the Simpson index, respectively, and incorporate

the relatedness of taxa using phylogenetic branch lengths in addition to relative abundances, where more distantly related taxa receive higher weights.[13,28]

16S rRNA gene sequencing was performed at the Molecular Research LP (Shallowater, TX); sequence pre-processing was performed at the Fred Hutchinson Cancer Research Center. In brief, the sequencing pre-processing was performed with USEARCH-UNOISE3[30,31] (usearch v11.0.667_i86linux64) and QIIME 2[32] to generate ASVs. ASVs are unique, denoised amplicon sequences that (ideally) represent true variation in the the *16S rRNA* gene at a high degree of resolution, enabling granular distinctions in taxa present in a sample.[33] ASVs were filtered by abundance at a threshold of 0.00001 following previous recommendations.[34] ASVs were taxonomically classified using the classify-sklearn module with SILVA 138 taxonomy.[35] An initial phylogeny was generated in QIIME 2 with MAFFT[36] for multiple sequence alignment and fasttree2[37] for phylogeny construction. Sample contamination was addressed with SCRuB[38] (v0.0.1) and a final ASV filter was applied at the genus level (>0.0009 abundance). Additional genus-level filtering was conducted based on both passing LOD and LOQ levels (or just LOD levels if the genera was previously observed in CRC tumors),[39,40] as well as by excluding probable contaminants,[40–42] leaving a total of 48 genera in the final analysis. 1403 ASVs passed filtering and were kept in the generation of alpha diversity measures and the final phylogeny used in this study. To generate the phylogeny-based alpha diversity metrics, prepared *16S rRNA* gene sequences were aligned with the Super5 algorithm using MUSCLE[43] (v5.3.osxarm64). A phylogeny was generated from aligned sequences using VeryFastTree[44] (v4.0.4). A representative *16S rRNA* gene sequence from the gut archaea *Methanobrevibacter intestini* was included in the alignment and clustering steps (RefSeq: PQ_670969.2)[45] to root the phylogeny. Phyloseq[46] (v1.50.0) and Enteropart[47] (v1.6-16) were used to calculate the individual non-phylogenetic and phylogenetic alpha diversity metrics, respectively.

The presence of the MSI-high phenotype in CRC tumors was dichotomized as MSS or MSI-low versus MSI-high. Categorization of MSI status was defined by the proportion of unstable loci from a targeted sequencing panel of 10 markers or based on immunohistochemical (IHC) staining of four MMR proteins (MLH1, MSH2, MSH6, and PMS2) as previously described.[19]

Characterization of tumor *F. nucleatum* positivity was assessed by ddPCR of the *F. nucleatum nusG* gene.[23] To be classified as *F. nucleatum* positive, a sample needed to have at least 4.1 copies/10 ng tissue of the *Fusobacterium nusG* present to exceed the limit of quantitation for the assay. For analyses examining *F. nucleatum* as an effect modifier of the relationship between microbial biodiversity and MSI-high CRC, *F. nucleatum* positivity was treated as a binary measure of either present or absent in

tumor and was paired with the Shannon index, since the Shannon index offers a balanced weighting of relative abundance compared to Richness and the Simpson index.[13]

Several measures were controlled for as important confounders of the relationship between alpha diversity and MSI status in CRC: age at diagnosis, self-reported sex, tumor location, and smoking status two years prior to a CRC diagnosis were adjusted for in the analyses. Tumor location was classified based on the assigned ICD-O-3 codes as 'left-sided' (C18.6, C18.7, and C18.5), 'right-sided' (C18.0, C18.2, C18.3, and C18.4), or 'rectal' (C19 and C20). Smoking history was classified based on responses to two survey questions: "Have you ever smoked at least one cigarette a day for 3 months or longer?" and "About two years ago, were you still smoking at least one cigarette a day?". Individuals who had smoked for 3 months or longer and were smoking about two years ago as of the baseline questionnaire were classified as 'currently smokes'. Individuals who had ever smoked for 3 months or longer, but who were not smoking about two years ago were classified as 'formerly smoked'. Individuals who had never smoked cigarettes for 3 or more months were classified as 'never smoked'.

Statistical methods

To preserve power and reduce bias in the *F. nucleatum* effect modification models, missing covariate values were imputed using the expectation-maximization with bootstrapping algorithm from the Amelia package[48] (v1.8.3) in R. We generated 10 imputed datasets, only imputing values for age at diagnosis, sex, smoking status, and tumor location. Regression was run independently on each imputed dataset, and then results were pooled with the Mice package[49] (v3.17.0) using the mean coefficient value for final point estimates and Rubin's rules[50] to estimate variance. Complete-cases were used in all other models.

The aMiAD methodology estimates standardized effect scores (MiDivES) and p-values for each individual alpha diversity measure using a residual permutation method to generate null models.[13] The measure with the strongest support is selected for estimating the adaptive effect score; to control type I error rates, the adaptive microbial diversity effect score (aMiDivES) and related p-value are re-estimated using a further set of permutations. This approach yields valid estimates of both the effect direction and magnitude of the individual alpha diversity measures and for the adaptive measure. We applied the aMiAD method[13] (v.2.0) using multiple logistic regression of six alpha diversity measures (observed Richness,[24] Shannon index,[25] Simpson index,[26] PD,[27] PE,[28] and PQE[29]) on dichotomous MSI status and adjusted for age at diagnosis, sex, tumor location, and smoking status to measure the association between alpha diversity and MSI CRC. We used 5000 permutations to estimate the test statistics for this measure.

We fit additional logistic regression models using the Shannon index in relation to MSI status and adjusted for age at diagnosis, sex, tumor location, and smoking status; these models also included tumor *F. nucleatum* positivity (by ddPCR) with and without an interaction term between alpha diversity and tumor *F. nucleatum* positivity. A Wald test[51] was used to estimate whether there was effect modification of the association between alpha diversity and MSI status by tumor *F. nucleatum* positivity.

We used Analysis of Compositions of Microbiomes with Bias Correction (ANCOM-BC, v2.8.1)[52] to assess whether genera were differentially abundant between MSI-high and MSI-low/MSS. We selected ANCOM-BC due to its strong control of false discovery rate (FDR) and relatively strong power.[52,53] Furthermore, the ANCOM-BC methodology allows for control over the inclusion of structural zeros in the analysis and provides sensitivity analyses to determine whether results are robust to pseudocount variation. We adjusted for age at diagnosis, sex, smoking history, and tumor location in analyses. Multiple comparison correction was performed using the Benjamini-Hochberg procedure.[54] We used a prevalence cut-off of 0.05 for inclusion in the analysis and allowed structural zeros to be included using both identification criteria described by the authors.[52] With this prevalence cut-off, only 45 genera were considered in the analysis.

Sensitivity Analyses

To further examine how specific alpha diversity metrics were associated with MSI status, we fit logistic regression models on the multiple imputed data, and individually modelled each alpha diversity metric on MSI status adjusted for age at diagnosis, sex, tumor location, and smoking status. To allow for more direct comparisons of individual alpha diversity metrics on a common scale,[55–57] the alpha diversity metrics were transformed to effective number of species (ENS) values and were fit in logistic regression models adjusting for the same covariates as above. The phylogeny-based diversity metrics were transformed to ENS values with a standardized phylogeny with height of 1 (agnostic to the true scale of the phylogeny). An ENS value corresponds to the estimated number of equally distantly related taxa of equal abundance that would have generated the same alpha diversity value as was observed in the data for that measure.[55–57] We also re-conducted the aMiAD analysis with these transformed ENS values. To further explore effect modification by *F. nucleatum* on the relationship between alpha diversity and MSI status, we fit additional logistic regression models of the Shannon index on MSI status; these models were fit separately on the subset of individuals with *F. nucleatum* positive tumors and the subset of individuals with *F. nucleatum* negative tumors. ANCOM-BC analysis was repeated at the family level using the same parameters as the

main analysis. Since *F. nucleatum* has a well-described enrichment in MSI-high tumor tissues[8,10,11] and some of the *Fusobacterium* ASVs were annotated as a species other than *F. nucleatum*, we ran the ANCOM-BC analysis on *F. nucleatum* specific ASVs. We used BLASTn[58,59] (Command line v2.16.0, megablast; word size 28, using the 16S_ribosomal_RNA database) to identify and subset *Fusobacterium* ASVs to ASVs with the top hit matching one of the 4 present *F. nucleatum* subspecies: *F. nucleatum nucleatum, F. animalis, F. polymorphum,* and *F. vincentii*.[60]

All analyses were performed in R[61] (version 4.4.0) unless otherwise specified.

RESULTS

Of the 632 participants who met inclusion criteria, 125 were classified as MSI-high. The overall cohort was close to evenly distributed by sex (52% female), though the MSI-high group had a higher proportion of female participants (67%) than the MSS/MSI-low group (49%) (**Table 1**). The majority of participants were White (79%) or reported two or more races (13%), with no other racial groups being represented at more than 5% of the overall study or within either MSI status group. Right-sided tumors were the most frequently observed overall (46%) and, as expected, the proportion of participants with right-sided tumors differed markedly by MSI status (86% of MSI-high participants vs. 36% of MSS/MSI-low participants). Most participants were diagnosed with regional stage CRC (58%), while 12% were diagnosed with distant stage disease. Participants with MSI-high tumors were more likely to be current smokers two years prior to CRC diagnosis (14% vs. 10% in those with MSS/MSI-low tumors and 11% overall). The prevalence of *F. nucleatum* positivity was markedly different by MSI status, with 46% vs. 17% of cases with MSI-high vs. MSI-low/MSS tumors, respectively, exhibiting *F. nucleatum* positivity based on ddPCR. Overall, Shannon diversity was slightly higher in participants with MSI-low/MSS CRC compared to individuals with MSI-high CRC, with medians of 2.13 and 1.98, respectively (**Supp. Table 1**).

All aMiAD effect scores of alpha diversity showed a negative association with MSI status, where lower alpha diversity was associated with a higher odds of having MSI-high CRC after adjusting for sex, age at diagnosis, smoking history, and tumor location (**Figure 1**). The magnitude of these associations was the greatest for Richness and Shannon diversity (MiDivES -1.51 and -1.57, respectively). The adaptive effect score, aMiDivES, was -1.08; however, there was no evidence of a statistically significant association between any of the individual alpha diversity measures nor the adaptive aMiAD effect score and MSI status (aMiAD p = 0.295) (**Figure 1**). Similarly, sensitivity analyses of individual models for each of the alpha diversity measures on the ENS scale and as nominal alpha diversity values showed negative

associations with MSI status across all measures, though no individual alpha diversity measures were statistically significant on either scale (**Supp. Figure 1)**. All aMiAD effect scores remained negative when fit with the alpha diversity measures on the ENS scale, but notably, the strength of the associations increased for all individual measures that were functionally different (i.e. richness and PD do not change on ENS scale as parameterized here). In particular, the Simpson index decreased to an effect score of -2.0 from -0.8 (**Supp. Figure 2**).

In the model of the Shannon index on MSI status that included a tumor *F. nucleatum* positivity term, we found a strong association between *F. nucleatum* positivity and MSI status, with *F. nucleatum* positive tumors estimated to have 3.03 (95% CI 1.87 – 4.92) times the odds of MSI-high status compared to *F. nucleatum* negative tumors (**Figure 2**). However, in models with the inclusion of interaction between the Shannon index and *F. nucleatum* status, we found no statistically significant evidence for effect modification by *F. nucleatum* positivity on the association between the Shannon index on MSI status ($p = 0.504$) (**Figure 2**). Sensitivity analysis of models fit on participant subsets by *F. nucleatum* positivity showed an attenuated estimate of the association between Shannon index and MSI status in *F. nucleatum* positive tumors (**Supp. Figure 3**).

Differential abundance analysis of 45 genera by MSI status with ANCOM-BC identified a set of 20 genera that were statistically significant after controlling FDR and adjusting for sex, age at diagnosis, smoking history, and tumor location. There were 8 genera that were enriched in MSI-high tumors: *Gemella, Lawsonella, Akkermansia, Campylobacter, Alloprevotella, Porphyromonas, Intestinibacter,* and *Granulicatella*, none of which were robust to sensitivity analysis varying the pseudocounts (**Figure 3**). A further 12 genera were enriched in MSS/MSI-low tumors, *Roseburia, Neisseriaceae, Ruminococcus gnavus, Faecalibacterium, Leptotrichia, Blautia, Escherichia-Shigella, Peptoclostridium, Veillonella, Enterococcus, Sporolactobacillaceae,* and *Cloacibacterium* (**Figure 3**). Of these, only *Escherichia-Shigella* was robust to the pseudocount sensitivity analysis. To see how sensitive these results were to the level of taxonomic resolution, we performed ANCOM-BC at the family level and observed consistent enrichment of the family-level taxa containing the genera identified as differentially abundant. There was mixed representation of the families expected based on the genera level analysis; the families *Lawsonellaceae, Prevotellaceae, Clostridiaceae, Lachnospiraceae, Neisseriaceae,* and *Synergistota* were not differentially abundant as would be expected based genera level analysis (**Supp. Figure 4**). Interestingly, the family *Weeksellaceae* was depleted in MSI-high tumors though none of its members were significantly associated at the genera level (**Supp. Figure 4**). We repeated ANCOM-BC at the genera level with only *F. nucleatum* specific ASVs included in the *Fusobacterium* genus; in this updated

analysis, *Fusobacterium* was found to be both statistically significantly enriched in MSI-high tumors and robust to pseudocount sensitivity analysis (**Supp. Figure 5**).

DISCUSSION

There has been growing interest in exploring the effect of the gut microbiome on CRC etiology and progression, but the relationship between MSI status and the ensemble CRC tumor microbial community has gone largely unexplored to date. We explored the relationship between the MSI phenotype and CRC tumor microbiome alpha diversity among CRC patients using an adaptive alpha diversity measure that incorporates several potentially important features of community structure. We then searched for specific genera that were differentially abundant by MSI phenotype. While none of the individual alpha diversity measures nor the adaptive measure were statistically significantly associated with MSI status in main analyses, higher alpha diversity was associated with a lower odds of MSI-high CRC consistently across all individual measures and in the adaptive measure. Furthermore, 20 genera were identified as differentially abundant in tumor tissue based on MSI status after adjustment for important confounding factors, though only *Escherichia-Shigella* was robust to sensitivity analysis, where it was found to be enriched in MSS/MSI-low tumors. *Fusobacterium* was found to be enriched in MSI-high tumors and robust to sensitivity analysis only after filtering *Fusobacterium* ASVs to those identified as *F. nucleatum* and its subspecies. Specific microbes are strongly associated with MSI status and have been suggested to play a causal role in CRC development,[62] so further work is needed to understand which aspects of the microbial community may affect the development of specific CRC phenotypes.

A number of previous studies have explored the relationship between MSI status and the specific members of the CRC tumor microbiome,[8–11,17,63] but exploration into the relationship between alpha diversity and MSI status has been limited.[14,16] It has been previously reported that tumor alpha diversity is elevated in MSI-high tumors compared to MSS/MSI-low tumors across several alpha diversity measures,[14] while other work has reported no statistically significant associations between MSI status and tumor alpha diversity in CRC tumors.[16] In contrast, our results are generally consistent with a null or negative association between alpha diversity and odds of MSI-high tumors, since the Shannon index and Simpson diversity tended towards a negative association with MSI-high tumors on the ENS scale. There are a number of possible causes for an observed lack of a strong association, for example, the *16S rRNA* gene sequencing was performed on formalin-fixed paraffin-embedded tumor samples after a minimum of 12 years of storage which has been shown to decrease DNA quality and alpha diversity over time.[64] To our knowledge, our study is the first to examine the association between tumor alpha diversity and MSI

status using phylogeny-based measures of alpha diversity, which may be important if greater phylogenetic distances between community taxa are strongly associated with MSI status. For example, if the MSI-high tumor microenvironment has greater variation in metabolic resources available than MSS/MSI-low tumors, then we might expect to find greater phylogeny-based diversity as more distantly related species are more likely to be able to functionally utilize those additional resources. We did not find evidence for this phylogeny-based effect, though we do see that prioritization of the most abundant taxa in a community tended to produce greater point estimates than abundance agnostic alpha diversity measures (i.e. Richness, PD) using alpha diversity measures on the ENS scale. This may indicate that the more abundant taxa are the drivers of any true association between alpha diversity and and MSI status.

Transforming alpha diversity measures to ENS values can provide useful information about how different features of community structure affect associations with alpha diversity, but are infrequently used in the CRC literature. Comparing measures of alpha diversity is challenging because most measures are on vastly different scales, and thus, direct comparisons of their actual values is uninformative and it is unclear how much differences in their effect size are due to their scale versus prioritization of different community features (e.g. abundance). ENS values allow for the direct comparisons of alpha diversity measures by transforming them to a common scale that can be framed as the number of equally common, equally distantly related taxa required to yield the same original value of the input alpha diversity metric.[56,65] For example, a community may have a Shannon index value of 2.5 (ENS = ~12.2) and a Simpson index value of 0.9 (ENS = 10), but it's unclear which is more diverse with their nominal values. Furthermore, a Shannon index value of 2.5 (ENS = ~12.2) or Simpson index value of 0.9 (ENS =10) can be generated by many different communities with differing numbers of species and in different proportions, but if taxa are equally common it would necessarily take ~12.2 species to observe a Shannon index value of 2.5 and 10 species to observe a Simpson index value 0.9. Given that we don't generally observe a truly even (or equally distantly related) distribution of species in a community, the ENS value for Shannon diversity and Simpson diversity will be different for each community/sample, so differences in ENS value are due to differences in how abundance is prioritized by each measure. This becomes informative in interpreting effect sizes and direction; in our study we see that prioritization of the most abundant taxa in a community tended to strengthen the association between alpha diversity and MSI status across both non-phylogenetic and phylogenetic alpha diversity measures. The transformation to ENS values also made it clear that the Shannon index and Simpson index had similar effect estimates for their association with MSI status, which was not apparent in their original forms. Furthermore, ENS values are interpretable on a linear scale that lends itself to more naturalistic interpretations of effect estimates. Phylogeny-based alpha diversity measures and ENS values are valuable tools for exploring

how the microbiome is related to different disease features and help explain what microbial community features drive those relationships.

There have been efforts to characterize differentially abundant taxa in CRC tumors based on MSI status; Jin et al.[14] identified 17 genera that were differentially abundant using *16S rRNA* gene sequencing data. The current study and Jin et al. find statistically significant enrichment of *Akkermansia* in MSI-high tumors compared to MSS/MSI-low tumors; results were conflicting between these studies for the direction of enrichment for *Faecalibacterium* and *Roseburia*, where in the current study both they were both depleted in MSI-high tumors. Notably, Jin et al. found enrichment of *Fusobacterium* in MSI-high tumors, whereas in our main analysis it was not significantly associated with MSI-status, though the log-fold change estimate was consistent with an enrichment in MSI-high tumors in main analyses. Purcell et al.[17] found enrichment of *Fusobacterium hwasookii* and *Porphyromonas gingivalis* in tumors with gene expression profiles aligning with MSI-high tissues, supporting the finding of enrichment of *Porphyromonas* in MSI-high tumors from the current study. Other studies have performed much more targeted differential abundance analysis of microbes by MSI status,[15,16] with a particular emphasis on *Fusobacterium*.[10,11,63] The lack of a strong association of *Fusobacterium* with MSI status in the current study's ANCOM-BC analysis was likely a consequence of heterogeneity in the species represented in the ASVs classified as *Fusobacterium*: some of the *Fusobacterium* ASVs were classified as *Fusobacterium mortiferum* and *Fusobacterium necrophorum*, or lacked deeper resolution, so we re-ran ANCOM-BC at the genus level with only *Fusobacterium* ASVs that could be mapped to *F. nucleatum* or its subspecies. In this analysis, *Fusobacterium* were robustly enriched in the MSI-high tumors. Our analysis of effect modification of Shannon diversity on MSI status by *F. nucleatum* tumor positivity does support an association between *F. nucleatum* positivity and increased odds of MSI-high CRC, at least in the absence of interaction between diversity and *F. nucleatum* positivity.

There is a growing body of literature identifying associations between CRC and colonization of the intestinal tract with specific microbes. While recent literature has looked at whether specific bacteria are associated with MSI status, the majority of studies have not used community measures of diversity to assess the microbiome's relationship with CRC development and MMR phenotype. Even among those that have, there has been inconsistent evidence of an association. While our results do not strongly support an association between alpha diversity and MSI status, they do present some interesting areas for future work. Given the increased strength of association between MSI status and alpha diversity for measures that prioritize the more abundant taxa, it is likely that the highly abundant taxa on tumors drive any true associations and examining the role of these microbes may be informative. We also identified

novel genera that were differentially abundant by MSI status, including *Lawsonella, Gemella Sporolactobacillaceae,* and *Cloacibacterium,* which warrants further research into the role of these microbes. We highlight the utility of phylogeny-based alpha diversity measures and ENS values in examining associations with disease and advocate for greater uptake of these tools in future research on the CRC microbiome.

This study has several key strengths. The inclusion of phylogenetic measures of alpha diversity incorporated valuable information about the plausible drivers of any observed association. We were also able to adjust for important confounders of the relationship between alpha diversity and MSI status, which has not been addressed in all previous works. This study also has important limitations. Samples used to characterize MSI status and alpha diversity were collected at the same time, so we cannot address whether associations between these attributes are causally linked nor the direction of any causal effects. There may be additional unmeasured confounding of these results; in particular, alcohol consumption, diet, and racial identity (as a proxy for racialized disparities) were not included in analyses which may induce biased estimates. We also did not account for the inherent uncertainty in estimating alpha diversity measures which can induce bias in effect estimates.[66] Furthermore, while *16S rRNA* gene sequencing is fast and affordable, the resolution of taxonomic classification is broadly limited to the genus level. As a result, the associations measured are aggregated across all species, subspecies, and strains, which may attenuate measured effects if sub-taxa act in opposing directions or have no effect. For example, *F. nucleatum* has repeatedly been associated with the MSI-high phenotype, yet we failed to detect a statistically significant enrichment of *F. nucleatum* in MSI-high tumors with ANCOM-BC until the Fusobacterium genus was depleted of non *F. nucleatum* ASVs.

CONCLUSIONS

We explored the relationship between CRC tumor microbial diversity and MSI status using CRC tumor tissue samples assayed for markers of CRC phenotype and profiled the tumor microbiome by sequencing the prokaryotic *16S rRNA* gene. While we did not find statistically significant evidence of an association between a set of alpha diversity measures and MSI status, we did identify 20 differentially abundant genera by MSI status, with several novel genera identified as differentially abundant. Future work should examine these genera for specific species that may have roles in the progression and etiology of CRC. Furthermore, the relationship between alpha diversity and MSI status remains unclear, so future work should seek to elucidate this relationship.

REFERENCES

1. Battaglin F, Naseem M, Lenz HJ, Salem ME. Microsatellite Instability in Colorectal Cancer: Overview of Its Clinical Significance and Novel Perspectives. *Clin Adv Hematol Oncol HO*. 2018;16(11):735-745. Accessed September 13, 2024. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7493692/

2. Boland CR, Goel A. Microsatellite Instability in Colorectal Cancer. *Gastroenterology*. 2010;138(6):2073-2087.e3. doi:10.1053/j.gastro.2009.12.064

3. Hernández-Luna MA, López-Briones S, Luria-Pérez R. The Four Horsemen in Colon Cancer. *J Oncol*. 2019;2019(1):5636272. doi:10.1155/2019/5636272

4. Sillo TO, Beggs AD, Middleton G, Akingboye A. The Gut Microbiome, Microsatellite Status and the Response to Immunotherapy in Colorectal Cancer. *Int J Mol Sci*. 2023;24(6):5767. doi:10.3390/ijms24065767

5. Murphy CL, Barrett M, Pellanda P, et al. Mapping the colorectal tumor microbiota. *Gut Microbes*. 2021;13(1). doi:10.1080/19490976.2021.1920657

6. Flemer B, Lynch DB, Brown JMR, et al. Tumour-associated and non-tumour-associated microbiota in colorectal cancer. *Gut*. 2017;66(4):633-643. doi:10.1136/gutjnl-2015-309595

7. Ai D, Pan H, Li X, Gao Y, Liu G, Xia LC. Identifying Gut Microbiota Associated With Colorectal Cancer Using a Zero-Inflated Lognormal Model. *Front Microbiol*. 2019;10:826. doi:10.3389/fmicb.2019.00826

8. Okita Y, Koi M, Takeda K, et al. Fusobacterium nucleatum infection correlates with two types of microsatellite alterations in colorectal cancer and triggers DNA damage. *Gut Pathog*. 2020;12(1):46. doi:10.1186/s13099-020-00384-3

9. Sun CH, Li BB, Wang B, et al. The role of Fusobacterium nucleatum in colorectal cancer: from carcinogenesis to clinical management. *Chronic Dis Transl Med*. 2019;5(3):178. doi:10.1016/j.cdtm.2019.09.001

10. Ono T, Yamaguchi T, Takao M, Kojika E, Iijima T, Horiguchi S ichiro. Fusobacterium nucleatum load in MSI colorectal cancer subtypes. *Int J Clin Oncol*. 2022;27(10):1580-1588. doi:10.1007/s10147-022-02218-5

11. Mima K, Nishihara R, Qian ZR, et al. Fusobacterium nucleatum in colorectal carcinoma tissue and patient prognosis. *Gut*. 2016;65(12):1973-1980. doi:10.1136/gutjnl-2015-310101

12. Borozan I, Zaidi SH, Harrison TA, et al. Molecular and Pathology Features of Colorectal Tumors and Patient Outcomes Are Associated with Fusobacterium nucleatum and Its Subspecies animalis. *Cancer Epidemiol Biomark Prev Publ Am Assoc Cancer Res Cosponsored Am Soc Prev Oncol*. 2022;31(1):210-220. doi:10.1158/1055-9965.EPI-21-0463

13. Koh H. An adaptive microbiome α-diversity-based association analysis method. *Sci Rep*. 2018;8(1):18026. doi:10.1038/s41598-018-36355-7

14. Jin M, Wu J, Shi L, et al. Gut microbiota distinct between colorectal cancers with deficient and proficient mismatch repair: A study of 230 CRC patients. *Front Microbiol*. 2022;13:993285. doi:10.3389/fmicb.2022.993285

15. Hale VL, Jeraldo P, Chen J, et al. Distinct microbes, metabolites, and ecologies define the microbiome in deficient and proficient mismatch repair colorectal cancers. *Genome Med*. 2018;10(1):78. doi:10.1186/s13073-018-0586-6

16. Byrd DA, Fan W, Greathouse KL, Wu MC, Xie H, Wang X. The intratumor microbiome is associated with microsatellite instability. *JNCI J Natl Cancer Inst*. 2023;115(8):989. doi:10.1093/jnci/djad083

17. Purcell RV, Visnovska M, Biggs PJ, Schmeier S, Frizelle FA. Distinct gut microbiome patterns associate with consensus molecular subtypes of colorectal cancer. *Sci Rep*. 2017;7(1):11590. doi:10.1038/s41598-017-11237-6

18. McCoy CO, Matsen FA. Abundance-weighted phylogenetic diversity measures distinguish microbial community states and are robust to sampling depth. *PeerJ*. 2013;1:e157. doi:10.7717/peerj.157

19. Newcomb PA, Baron J, Cotterchio M, et al. Colon Cancer Family Registry: An International Resource for Studies of the Genetic Epidemiology of Colon Cancer. *Cancer Epidemiol Biomarkers Prev*. 2007;16(11):2331-2343. doi:10.1158/1055-9965.EPI-07-0648

20. Phipps AI, Limburg PJ, Baron JA, et al. Association between molecular subtypes of colorectal cancer and patient survival. *Gastroenterology*. 2015;148(1):77-87.e2. doi:10.1053/j.gastro.2014.09.038

21. Flores Bueso Y, Walker SP, Tangney M. Characterization of FFPE-induced bacterial DNA damage and development of a repair method. *Biol Methods Protoc*. 2020;5(1):bpaa015. doi:10.1093/biomethods/bpaa015

22. Hullar MAJ, Jenkins IC, Randolph TW, et al. Associations of the gut microbiome with hepatic adiposity in the Multiethnic Cohort Adiposity Phenotype Study. *Gut Microbes*. 2021;13(1):1965463. doi:10.1080/19490976.2021.1965463

23. Hullar MAJ, Kahsai OJ, Hill C, et al. Highly sensitive DNA testing of Fusobacterium nucleatum (Fn) in colorectal tumors. *Cancer Epidemiol Biomarkers Prev*. Published online May 15, 2025. doi:10.1158/1055-9965.EPI-24-1020

24. McIntosh RP. An Index of Diversity and the Relation of Certain Concepts to Diversity. *Ecology*. 1967;48(3):392-404. doi:10.2307/1932674

25. Shannon CE. A Mathematical Theory of Communication. *Bell Syst Tech J*. 1948;27(3):379-423. doi:10.1002/j.1538-7305.1948.tb01338.x

26. Simpson EH. Measurement of Diversity. *Nature*. 1949;163(4148):688-688. doi:10.1038/163688a0

27. Faith DP. Conservation evaluation and phylogenetic diversity. *Biol Conserv*. 1992;61(1):1-10. doi:10.1016/0006-3207(92)91201-3

28. Allen B, Kon M, Bar‐Yam Y. A New Phylogenetic Diversity Measure Generalizing the Shannon Index and Its Application to Phyllostomid Bats. *Am Nat*. 2009;174(2):236-243. doi:10.1086/600101

29. Rao CR. Diversity and dissimilarity coefficients: A unified approach. *Theor Popul Biol*. 1982;21(1):24-43. doi:10.1016/0040-5809(82)90004-1

30. Edgar RC. Search and clustering orders of magnitude faster than BLAST. *Bioinformatics*. 2010;26(19):2460-2461. doi:10.1093/bioinformatics/btq461

31. Edgar RC. UNOISE2: improved error-correction for Illumina 16S and ITS amplicon sequencing. Published online October 15, 2016:081257. doi:10.1101/081257

32. Bolyen E, Rideout JR, Dillon MR, et al. Reproducible, interactive, scalable and extensible microbiome data science using QIIME 2. *Nat Biotechnol*. 2019;37(8):852-857. doi:10.1038/s41587-019-0209-9

33. Callahan BJ, McMurdie PJ, Holmes SP. Exact sequence variants should replace operational taxonomic units in marker-gene data analysis. *ISME J*. 2017;11(12):2639-2643. doi:10.1038/ismej.2017.119

34. Prodan A, Tremaroli V, Brolin H, Zwinderman AH, Nieuwdorp M, Levin E. Comparing bioinformatic pipelines for microbial 16S rRNA amplicon sequencing. *PLOS ONE*. 2020;15(1):e0227434. doi:10.1371/journal.pone.0227434

35. Quast C, Pruesse E, Yilmaz P, et al. The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. *Nucleic Acids Res*. 2013;41(Database issue):D590-596. doi:10.1093/nar/gks1219

36. Katoh K, Misawa K, Kuma K ichi, Miyata T. MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucleic Acids Res*. 2002;30(14):3059-3066. doi:10.1093/nar/gkf436

37. Price MN, Dehal PS, Arkin AP. FastTree 2 – Approximately Maximum-Likelihood Trees for Large Alignments. *PLOS ONE*. 2010;5(3):e9490. doi:10.1371/journal.pone.0009490

38. Austin GI, Park H, Meydan Y, et al. Contamination source modeling with SCRuB improves cancer phenotype prediction from microbiome data. *Nat Biotechnol*. 2023;41(12):1820-1828. doi:10.1038/s41587-023-01696-w

39. Galeano Niño JL, Wu H, LaCourse KD, et al. Effect of the intratumoral microbiota on spatial and cellular heterogeneity in cancer. *Nature*. 2022;611(7937):810-817. doi:10.1038/s41586-022-05435-0

40. Nejman D, Livyatan I, Fuks G, et al. The human tumor microbiome is composed of tumor type–specific intracellular bacteria. *Science*. 2020;368(6494):973-980. doi:10.1126/science.aay9189

41. Glassing A, Dowd SE, Galandiuk S, Davis B, Chiodini RJ. Inherent bacterial DNA contamination of extraction and sequencing reagents may affect interpretation of microbiota in low bacterial biomass samples. *Gut Pathog*. 2016;8:24. doi:10.1186/s13099-016-0103-7

42. Salter SJ, Cox MJ, Turek EM, et al. Reagent and laboratory contamination can critically impact sequence-based microbiome analyses. *BMC Biol*. 2014;12:87. doi:10.1186/s12915-014-0087-z

43. Edgar RC. Muscle5: High-accuracy alignment ensembles enable unbiased assessments of sequence homology and phylogeny. *Nat Commun*. 2022;13(1):6968. doi:10.1038/s41467-022-34630-w

44. Piñeiro C, Abuín JM, Pichel JC. Very Fast Tree: speeding up the estimation of phylogenies for large alignments through parallelization and vectorization strategies. *Bioinformatics*. 2020;36(17):4658-4659. doi:10.1093/bioinformatics/btaa582

45. Weinberger V, Mohammadzadeh R, Blohs M, et al. Expanding the cultivable human archaeome: Methanobrevibacter intestini sp. nov. and strain Methanobrevibacter smithii "GRAZ-2" from human faeces. *Int J Syst Evol Microbiol*. 2025;75(4). doi:10.1099/ijsem.0.006751

46. McMurdie PJ, Holmes S. phyloseq: An R Package for Reproducible Interactive Analysis and Graphics of Microbiome Census Data. *PLOS ONE*. 2013;8(4):e61217. doi:10.1371/journal.pone.0061217

47. Marcon E, Hérault B. entropart: An R Package to Measure and Partition Diversity. *J Stat Softw*. 2015;67:1-26. doi:10.18637/jss.v067.i08

48. Honaker J, King G, Blackwell M. Amelia II: A Program for Missing Data. *J Stat Softw*. 2011;45:1-47. doi:10.18637/jss.v045.i07

49. Buuren S van, Groothuis-Oudshoorn K. mice: Multivariate Imputation by Chained Equations in R. *J Stat Softw*. 2011;45:1-67. doi:10.18637/jss.v045.i03

50. Rubin DB. *Multiple Imputation for Nonresponse in Surveys.* John Wiley & Sons Inc.; 1987. http://dx.doi.org/10.1002/9780470316696

51. Wald A. Tests of Statistical Hypotheses Concerning Several Parameters When the Number of Observations is Large. *Trans Am Math Soc*. 1943;54(3):426-482. doi:10.2307/1990256

52. Lin H, Peddada SD. Analysis of compositions of microbiomes with bias correction. *Nat Commun*. 2020;11(1):3514. doi:10.1038/s41467-020-17041-7

53. Nearing JT, Douglas GM, Hayes MG, et al. Microbiome differential abundance methods produce different results across 38 datasets. *Nat Commun*. 2022;13(1):342. doi:10.1038/s41467-022-28034-z

54. Benjamini Y, Hochberg Y. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *J R Stat Soc Ser B Methodol*. 1995;57(1):289-300. doi:10.1111/j.2517-6161.1995.tb02031.x

55. Chao A, Chiu CH, Jost L. Phylogenetic diversity measures based on Hill numbers. *Philos Trans R Soc B Biol Sci*. 2010;365(1558):3599-3609. doi:10.1098/rstb.2010.0272

56. Chao A, Chiu CH, Jost L. Phylogenetic Diversity Measures and Their Decomposition: A Framework Based on Hill Numbers. In: Pellens R, Grandcolas P, eds. *Biodiversity Conservation and Phylogenetic Systematics: Preserving Our Evolutionary Heritage in an Extinction Crisis*. Springer International Publishing; 2016:141-172. doi:10.1007/978-3-319-22461-9_8

57. Ricotta C, Szeidl L. Diversity partitioning of Rao's quadratic entropy. *Theor Popul Biol*. 2009;76(4):299-302. doi:10.1016/j.tpb.2009.10.001

58. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. *J Mol Biol*. 1990;215(3):403-410. doi:10.1016/S0022-2836(05)80360-2

59. Morgulis A, Coulouris G, Raytselis Y, Madden TL, Agarwala R, Schäffer AA. Database indexing for production MegaBLAST searches. *Bioinforma Oxf Engl*. 2008;24(16):1757-1764. doi:10.1093/bioinformatics/btn322

60. Zepeda-Rivera M, Minot SS, Bouzek H, et al. A distinct Fusobacterium nucleatum clade dominates the colorectal cancer niche. *Nature*. 2024;628(8007):424-432. doi:10.1038/s41586-024-07182-w

61. R: The R Project for Statistical Computing. Accessed June 7, 2025. https://www.r-project.org/

62. Wu N, Feng YQ, Lyu N, Wang D, Yu WD, Hu YF. Fusobacterium nucleatum promotes colon cancer progression by changing the mucosal microbiota and colon transcriptome in a mouse model. *World J Gastroenterol*. 2022;28(18):1981-1995. doi:10.3748/wjg.v28.i18.1981

63. Tahara T, Yamamoto E, Suzuki H, et al. Fusobacterium in colonic flora and molecular features of colorectal carcinoma. *Cancer Res*. 2014;74(5):1311-1318. doi:10.1158/0008-5472.CAN-13-1865

64. Pinto-Ribeiro I, Ferreira RM, Pereira-Marques J, et al. Evaluation of the Use of Formalin-Fixed and Paraffin-Embedded Archive Gastric Tissues for Microbiota Characterization Using Next-Generation Sequencing. *Int J Mol Sci*. 2020;21(3):1096. doi:10.3390/ijms21031096

65. Jost L. Entropy and diversity. *Oikos*. 2006;113(2):363-375. doi:10.1111/j.2006.0030-1299.14714.x

66. Willis AD. Rarefaction, Alpha Diversity, and Statistics. *Front Microbiol*. 2019;10:2407. doi:10.3389/fmicb.2019.02407

TABLES AND FIGURES

Table 1: Summary of study participant characteristics by MSI status and overall.

Characteristic	MSI-high N = 125[1]	MSS/MSI-low N = 507[1]	Overall N = 632[1]
Sex			
Female	84 (67%)	246 (49%)	330 (52%)
Male	41 (33%)	261 (51%)	302 (48%)
Age	63 (47, 69)	59 (49, 67)	60 (48, 67)
Race			
American Indian/Alaska Native	1 (0.9%)	1 (0.2%)	2 (0.3%)
Asian	2 (1.7%)	16 (3.3%)	18 (3.0%)
Black or African American	1 (0.9%)	21 (4.3%)	22 (3.7%)
Native Hawaiian or Other Pacific Islander	1 (0.9%)	4 (0.8%)	5 (0.8%)
White	90 (78%)	384 (80%)	474 (79%)
More Than One Race	20 (17%)	57 (12%)	77 (13%)
Unknown	10	24	34
Tumor Site			
Left-sided	12 (9.6%)	157 (31%)	169 (27%)
Rectal	3 (2.4%)	157 (31%)	160 (25%)
Right-sided	107 (86%)	183 (36%)	290 (46%)
Unknown	3 (2.4%)	10 (2.0%)	13 (2.1%)
F. nucleatum status			
Positive	56 (45%)	88 (17%)	144 (23%)
Negative, or below LOQ	67 (54%)	416 (82%)	483 (76%)
Unknown	2 (1.6%)	3 (0.6%)	5 (0.8%)
Smoking history			
Current	18 (14%)	52 (10%)	70 (11%)
Former	58 (46%)	236 (47%)	294 (47%)
Never	49 (39%)	219 (43%)	268 (42%)
Stage			
Distant	5 (4.0%)	67 (13%)	72 (11%)
Local	45 (36%)	143 (28%)	188 (30%)
Regional	73 (58%)	283 (56%)	356 (56%)
Unknown	2 (1.6%)	14 (2.8%)	16 (2.5%)

[1] n (%); Median (Q1, Q3)

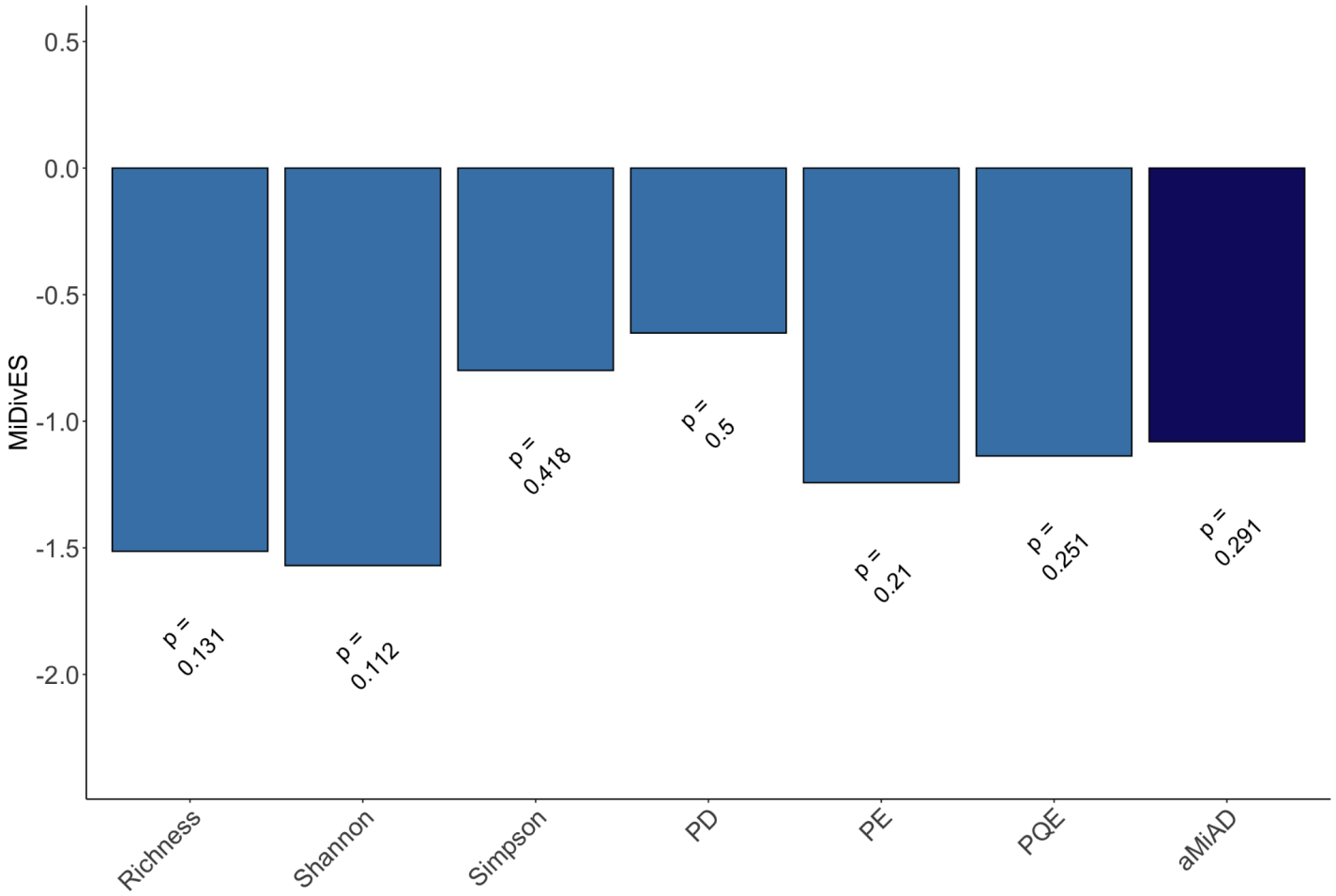


Figure 1. Individual effect score estimates (MiDivES: light blue) from aMiAD analysis of the 6 alpha diversity metrics and the effect estimate of the adaptive measure aMiAD (aMiDvES: dark blue). Effect score estimates below 0 represent negative associations of alpha diversity with MSI status, with the magnitude of the result indicating the relative strengths of the associations.

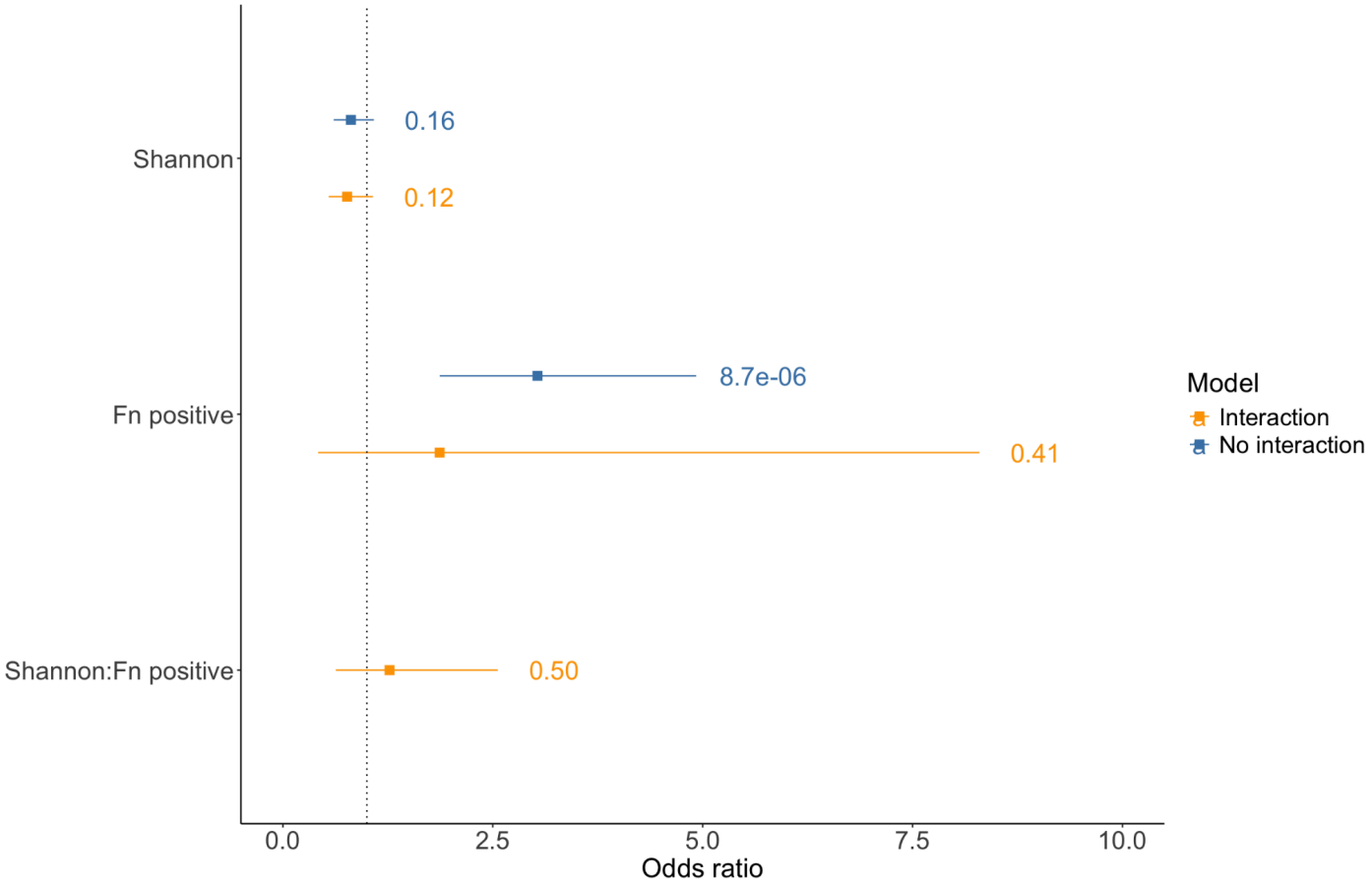


Figure 2. Forest plots of estimated odds ratios of the Shannon index and *F. nucleatum* (Fn) positivity from ddPCR on MSI status from models with and without interaction between the two measures. Both models were otherwise adjusted for age at diagnosis, sex, smoking history, sex, and tumor location. P-values displayed next to their respective estimate.

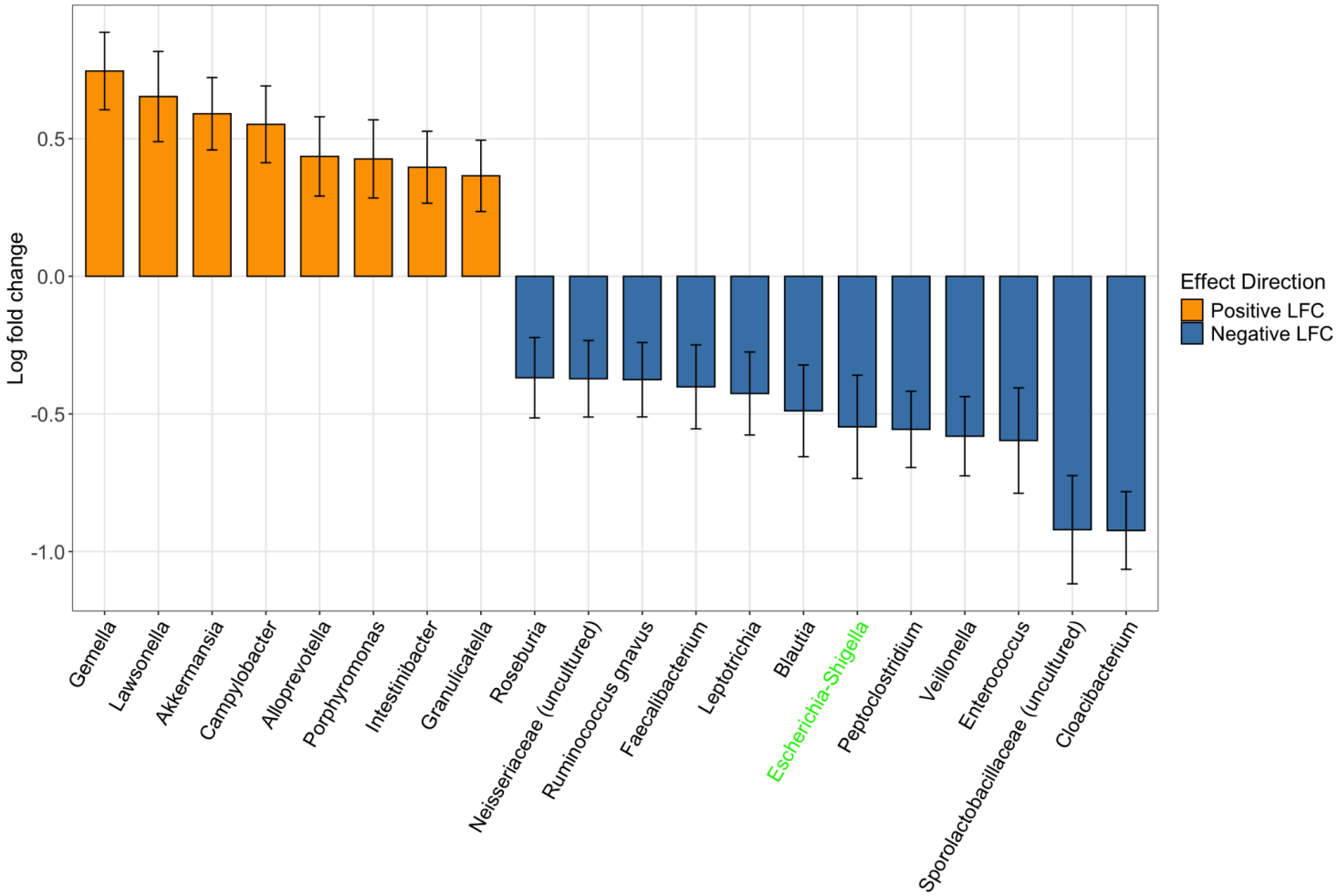


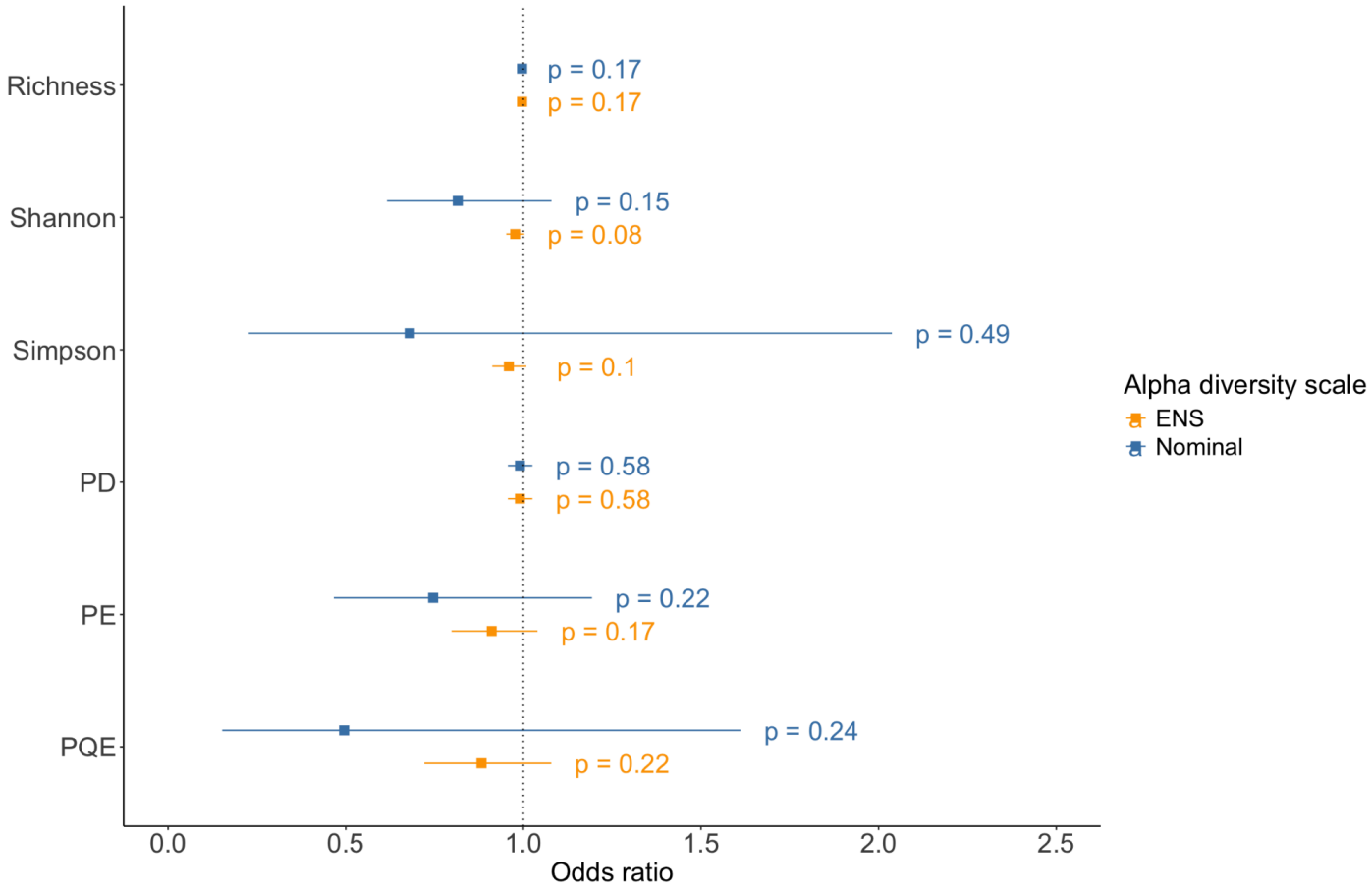
Figure 3. Differential abundance of taxa at the genera level using ANCOM-BC2, adjusted for age at diagnosis, sex, smoking history, and tumor location. Log-fold change in relative abundance plotted by the different genera that were statistically significant after FDR correction. Genera that were enriched in MSI-high tumors are in orange and genera that were enriched in MSS/MSI-low tumors are blue. Genera labels highlighted in green were robust to sensitivity analysis of variation in pseudocounts.

SUPPLEMENTAL MATERIALS

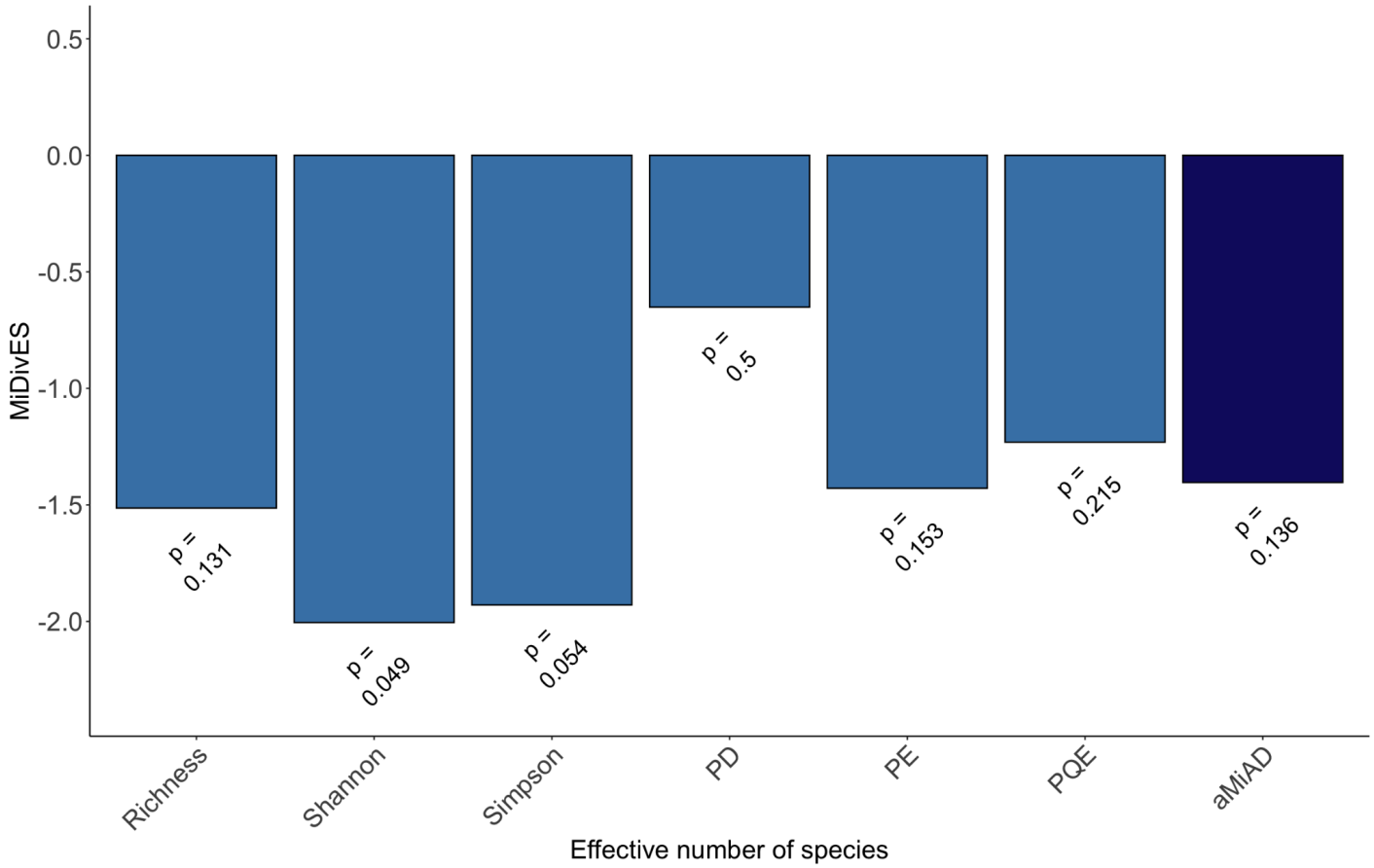
Supplemental Table 1. Summary of Shannon diversity across different participant characteristics by MSI status and overall.

	Shannon Index		
	MSI-high 1.98 (1.55-2.45)[1]	**MSS/MSI-low 2.13 (1.66-2.55)[1]**	**Overall 2.12 (1.64-2.53)[1]**
Sex			
Female	2.03 (1.57-2.45)	2.11 (1.62-2.58)	2.08 (1.61-2.50)
Male	1.94 (1.54-2.43)	2.15 (1.70-2.54)	2.13 (1.68-2.54)
Race			
American Indian/Alaska Native	1.97 (1.97-1.97)	1.81 (1.81-1.81)	1.89 (1.85-1.93)
Asian	1.80 (1.67-1.93)	2.03 (1.34-2.59)	2.00 (1.34-2.41)
Black or African American	1.71 (1.71-1.71)	2.02 (1.39-2.33)	1.98 (1.46-2.32)
Native Hawaiian or Other Pacific Islander	1.13 (1.13-1.13)	2.14 (2.04-2.23)	2.08 (1.92-2.20)
White	1.99 (1.50-2.47)	2.16 (1.69-2.59)	2.14 (1.67-2.56)
More Than One Race	2.06 (1.75-2.41)	1.86 (1.60-2.33)	1.93 (1.60-2.39)
Unknown	1.91 (1.80-2.41)	2.32 (1.65-2.87)	2.19 (1.69-2.57)
Tumor Site			
Left-sided	2.09 (1.43-2.68)	2.06 (1.64-2.49)	2.06 (1.63-2.49)
Rectal	2.29 (2.17-2.38)	2.13 (1.69-2.60)	2.13 (1.71-2.58)
Right-sided	1.96 (1.57-2.43)	2.16 (1.65-2.55)	2.11 (1.60-2.50)
Unknown	1.89 (1.36-2.30)	2.25 (1.86-2.54)	2.17 (1.80-2.54)
F. nucleatum status			
Positive	1.98 (1.59-2.46)	2.24 (1.72-2.55)	2.15 (1.69-2.54)
Negative, or below LOQ	1.97 (1.47-2.41)	2.12 (1.64-2.54)	2.10 (1.62-2.51)
Unknown	2.99 (2.62-3.37)	2.02 (1.87-2.76)	2.24 (2.02-3.50)
Smoking History			
Current	2.12 (1.97-2.28)	2.22 (1.87-2.58)	2.22 (1.87-2.55)
Former	1.82 (1.38-2.27)	2.11 (1.62-2.44)	2.03 (1.57-2.43)
Never	2.13 (1.69-2.50)	2.12 (1.66-2.63)	2.12 (1.67-2.59)
Stage			
Distant	1.85 (1.37-2.29)	2.22 (1.69-2.75)	2.22 (1.68-2.70)
Local	2.00 (1.54-2.45)	2.06 (1.62-2.57)	2.04 (1.59-2.51)
Regional	1.97 (1.58-2.40)	2.15 (1.70-2.52)	2.13 (1.66-2.50)
Unknown	1.77 (1.29-2.24)	2.12 (1.41-2.19)	2.12 (1.29-2.23)

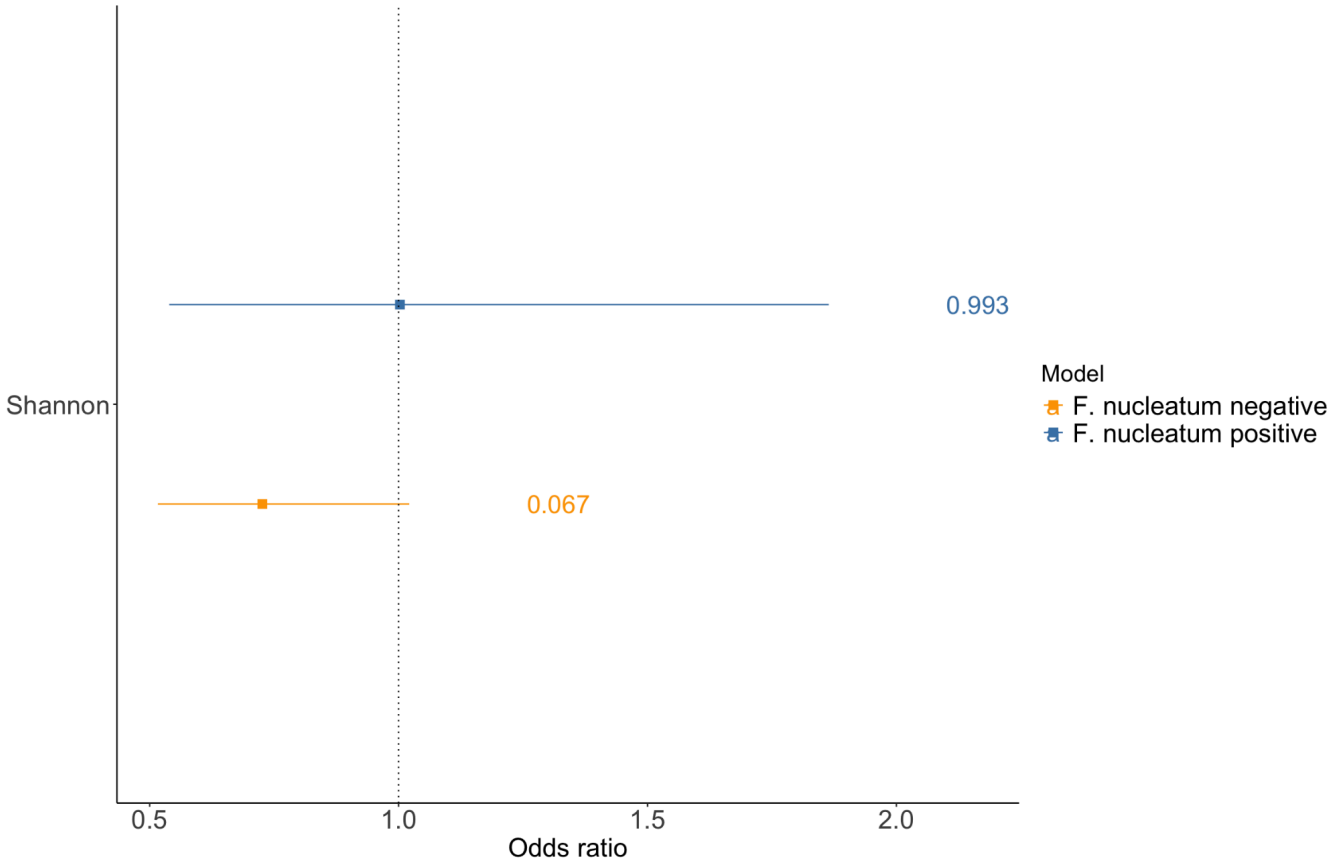
[1] Median (Q1, Q3)



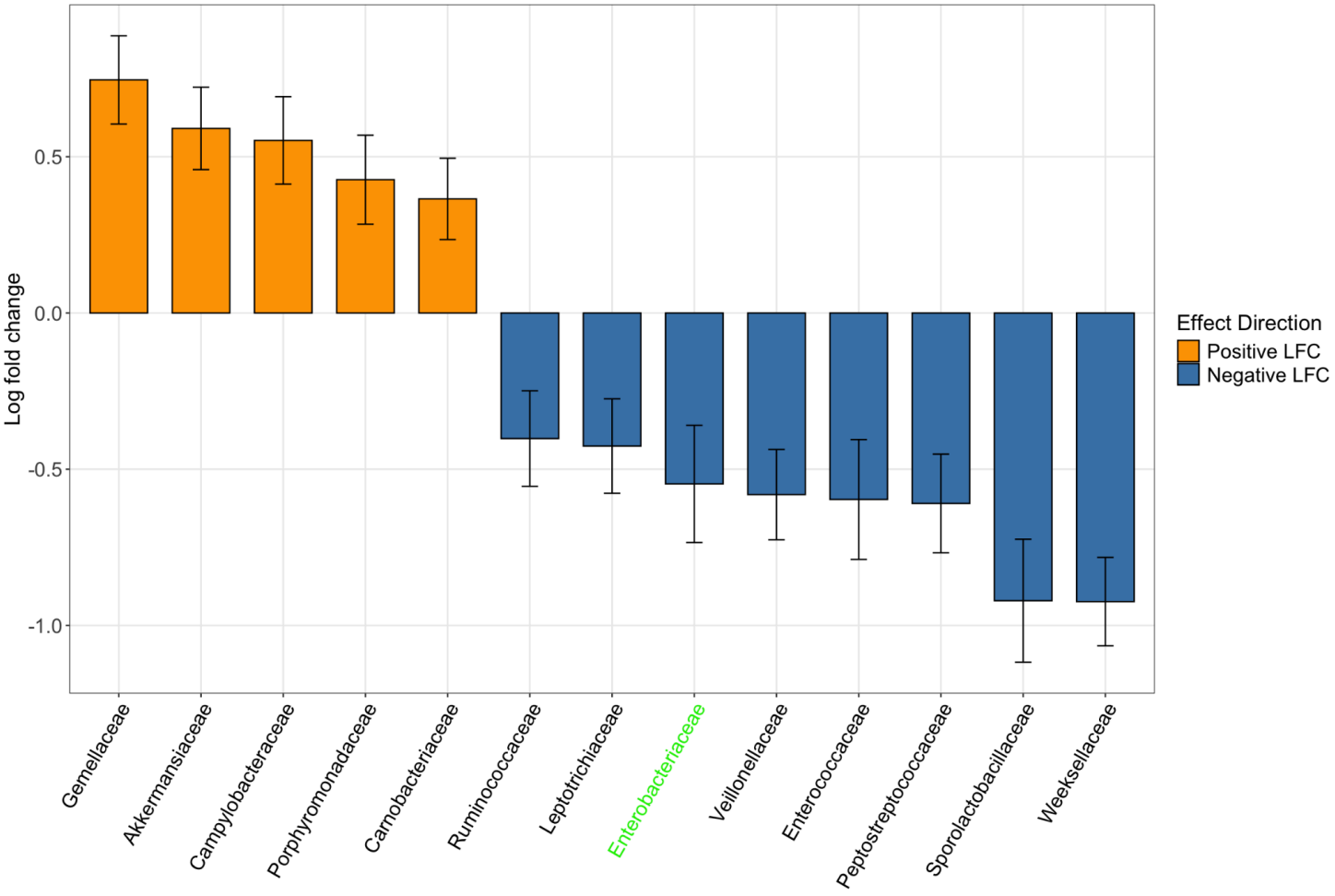
Supplemental Figure 1. Logistic regression results of alpha diversity metrics on odds of having MSI-high CRC. Analysis adjusted for sex, age at diagnosis, smoking history, and tumor location. Models fit for both the nominal alpha diversity metrics and as the effective number of species (here, ASVs). P-values for each estimate presented along each estimate. Note that Richness and PD are unchanged on the ENS scale as formulated here.



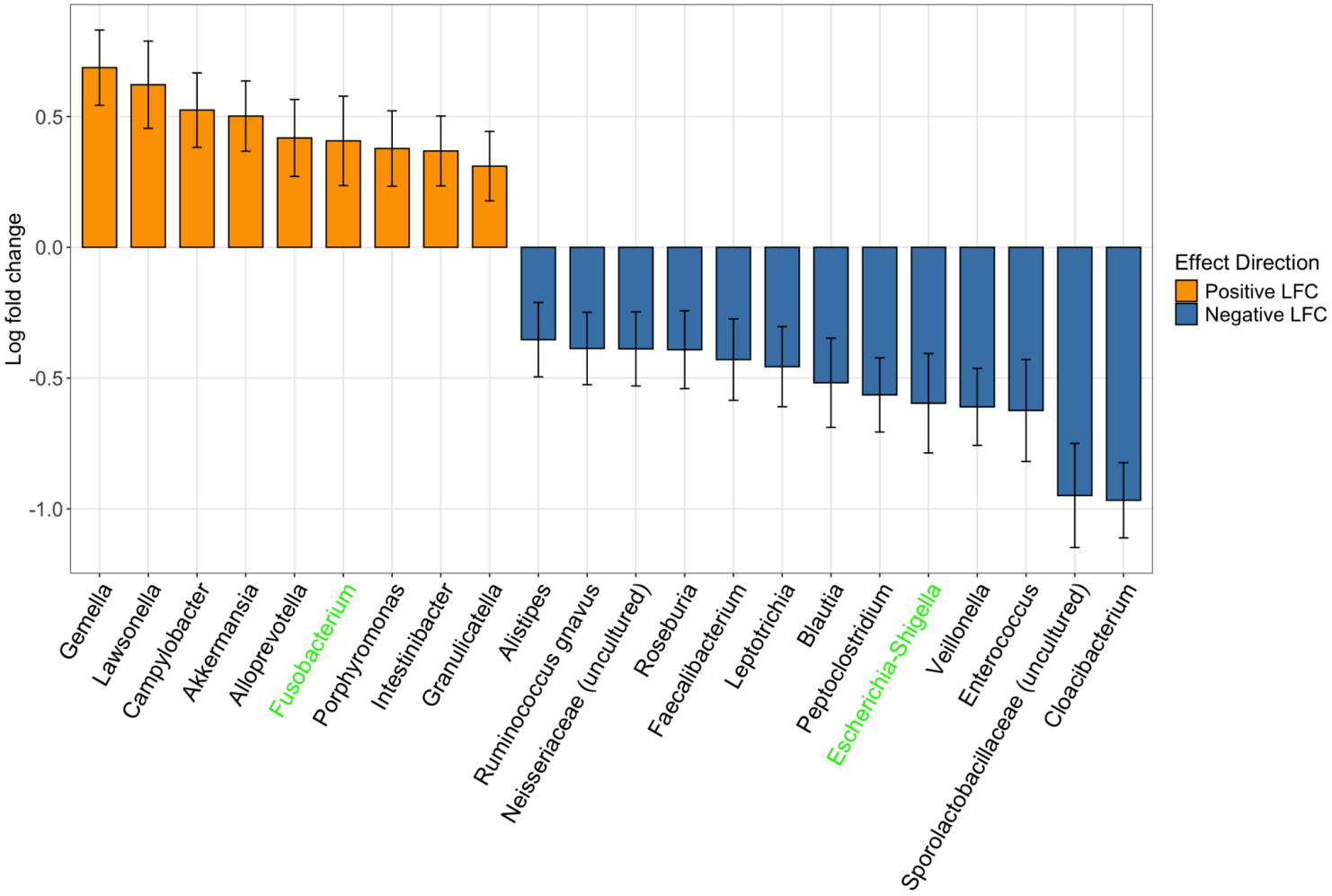
Supplemental Figure 2. Individual effect score estimates (MiDivES: light blue) from aMiAD analysis of the 6 alpha diversity metrics transformed to ENS values and the effect estimate of the adaptive measure aMiAD (aMiDvES: dark blue). Effect score estimates below 0 represent negative associations of alpha diversity with MSI status, with the magnitude of the result indicating the relative strengths of the associations.



Supplemental Figure 3. Logistic regression results of the estimated association of Shannon diversity on MSI status as odds ratios modelled on participant subsets with and without F. nucleatum positivity, adjusted for age at diagnosis, sex, smoking history, and tumor location. P-values of respective odds ratios shown.



Supplemental Figure 4. Differential abundance of taxa at the family level using ANCOM-BC2, adjusted for age at diagnosis, sex, smoking history, and tumor location. Log-fold change in relative abundance plotted by the different families that were statistically significant after FDR correction using the Benjamini-Hochberg method. Families that were enriched in MSI-high tumors are in orange and families that were enriched in MSS/MSI-low tumors are blue. Families labels highlighted in green were robust to sensitivity analysis of variation in pseudocounts.



Supplemental Figure 5. Differential abundance of taxa at the genus level using ANCOM-BC2, adjusted for age at diagnosis, sex, smoking history, and tumor location. Fusobacterium ASVs were subset to only ASVs that were identified as one of the F. nucleatum subspecies (*F. animalis*, *F. nucleatum*, *F. polymorphum*, and *F. vincentii*) by the top hit using BLASTn. Log-fold change in relative abundance plotted by the different genera that were statistically significant after FDR correction with Benjamini-Hochberg method. Genera that were enriched in MSI-high tumors are in orange and genera that were enriched in MSS/MSI-low tumors are blue. Genera labels highlighted in green are robust to sensitivity analysis of variation in pseudocounts.