Microbiota-based model improves the sensitivity for detecting colonic lesions

Nielson T. Baxter[1], Mack T. Ruffin IV[2], Mary A.M. Rogers[3], and Patrick D. Schloss[1*]

[1]Department of Microbiology and Immunology, University of Michigan, Ann Arbor, Michigan.

[2]Department of Family Medicine, University of Michigan, Ann Arbor, Michigan.

[3]Department of Internal Medicine, University of Michigan, Ann Arbor, Michigan.

[*]Correspondence to: pschloss@umich.edu

Colorectal cancer is the second leading cause of death among cancers in the United States[1]. Although individuals diagnosed early have a greater than 90% chance of survival, more than one-third of individuals do not adhere to screening recommendations partly because the standard diagnostics, colonoscopy and sigmoidoscopy, are expensive and invasive[1–4]. Thus, there is a great need to improve the sensitivity of non-invasive tests to detect early stage cancers and adenomas. Numerous studies have demonstrated a causal link between the formation of colonic lesions and the activity of the gut microbiota in tissue culture and animal models[5–8]. These findings have been complemented by studies in human populations identifying shifts in the composition of the gut microbiota associated with the progression of colorectal cancer[9–13]. These results suggest that the gut microbiota may represent a reservoir of biomarkers that would complement existing non-invasive methods such as the widely used fecal immunochemical test (FIT). Using stool samples from 490 patients we developed a cross-validated random forest classification model that detects colonic lesions using the relative abundance of gut microbiota and the concentration of hemoglobin in stool. The microbiota-based random forest model detected 95.0% of cancers and 57.1% of adenomas while FIT alone detected 75.0% and 15.7%, respectively. Of the colonic lesions missed by FIT, the model detected 80.0% of cancers and 49.1% of adenomas. These findings demonstrate the potential for microbiota analysis to complement existing screening methods to improve detection of colonic lesions. With a high sensitivity and low rate of false negatives, our model could be used to accurately identify asymptomatic individuals with preclinical disease and, thus, save lives.

Colorectal cancer mortality has steadily declined in recent decades, due in large part to increased screening[1]. Yet current screening tests, the fecal immunochemical test (FIT) and the multitarget DNA test, have a sensitivity of 7.6% and 17.2%, respectively, for detecting non-advanced adenoma –

just the type of early lesion that screening is meant to identify[14]. Although structural exams including colonoscopy and sigmoidoscopy are able to detect both adenomas and carcinomas, the high cost and invasive nature are barriers for many people. Fear, discomfort, and embarrassment are among the most cited reasons patients choose to forego CRC screening[4]. Likewise the large disparity in screening rates between those with and without health insurance highlights the need for inexpensive screening methods[1–3]. Unfortunately cheaper, less invasive stool-based tests like guaic fecal occult blood test and FIT are unable to reliably detect adenomas[15]. The newly introduced stool DNA panel has improved accuracy compared to FIT, but is still limited in its ability to accurately detect adenomas[14]. Thus there is need for novel screening methods that are inexpensive and capable of detecting both cancer and adenomas.

The gut microbiota, the collection of microorganisms that inhabit the gastrointestinal tract, are one potential source of biomarkers for detecting colonic lesions. Numerous studies have observed alterations in the gut bacterial communities of patients with CRC[9–13]. Experiments in animal models have demonstrated that such alterations have the potential to accelerate tumorigenesis[5]. Furthermore, several members of the gut microbiota have been shown to potentiate both the development and progression of CRC by a variety of mechanisms[6–8]. Although each of these organisms may play a role in certain cases of CRC, none of them is present in every case. Therefore we postulate that no one organism is an effective biomarker on its own and that focusing on a single bacterial population excludes the potential that the microbial etiology of the disease is actually polymicrobial.

We and others have shown that statistical models that take into account the abundances of multiple bacterial species can be used to distinguish healthy individuals from those with CRC[16,17]. In the present study we expanded upon those findings by demonstrating the potential for microbiota

analysis to complement FIT for improved detection of colonic lesions, including adenomas. We utilized the random forest algorithm, which is a decision tree-based machine learning algorithm for classification that accounts for non-linear data and interactions among features and includes an internal cross-validation to prevent overfitting[18]. By incorporating data on hemoglobin and bacterial abundances into a single model (labeled the Multitarget Microbiota Test or MMT), we were able to improve the sensitivity for adenomas and cancer compared to FIT alone.

We characterized the bacterial communities of stool samples from 490 patients using 16S rRNA gene sequencing. Among these patients, 120 had CRC, 109 had advanced adenomas, 89 had non-advanced adenomas, and 172 had no colonic lesions. We also tested each sample for the concentration of hemoglobin using FIT. With these data we developed a random forest model that incorporated the microbiota and FIT data and would differentiate normal individuals from those with any type of colonic lesion (i.e. adenoma or carcinoma). We determined the optimal model using the AUC-RF algorithm for maximizing the area under the curve (AUC) of the receiver operating characteristic (ROC) curve for the MMT[19]. The optimal model combining hemoglobin results and the microbiota used 23 bacterial populations, or operational taxonomic units (OTUs) (Extended Data Fig. 1). Of those OTUs, 16 were members of the Firmicutes phylum, including 3 from the Ruminococcaceae family and 10 from the Lachnospiraceae family, the predominant producers of butyrate in the gut[20] (Extended Data Fig. 2). Three OTUs were associated with the genus *Bacteroides*. The remaining OTUs were associated with *Porphyromonas*, *Parabacteroides*, *Collinsella*, and Enterobacteriaceae. The OTU associated with *Porphyromonas* was most closely related to *Porphyromonas asaccharolytica*, which has been previously shown to be predictive of CRC[16,21]. Like other studies[13,16] we also observed an OTU associated with *Fusobacterium nucleatum* that was enriched in cancer samples, however its relative abundance did not add sufficient information to be included in the model. Interestingly the majority of OTUs used in the model, especially the

Lachnospiraceae, were enriched in normal patients, suggesting that a loss of beneficial organisms in addition to the emergence of pathogens may be indicative of CRC development.

To determine whether microbiota sequence data could be used to complement FIT, we compared the performance of the MMT to FIT. The AUC for the MMT (AUC=0.755) was significantly higher than FIT (AUC=0.639) for distinguishing adenoma from normal ($p<0.001$) or all lesions from normal (FIT AUC=0.749, MMT AUC=0.829, $p<0.001$), but not cancer from normal (FIT AUC=0.929, MMT AUC=0.952, $p=0.091$) (Fig. 1A). To generate a categorical prediction from the MMT, we determined that the optimal threshold for the models's probability was 0.622 using Youden's J statisitc[22]. Samples scoring above this cutoff were classified as lesions, and those below the cutoff were classified as normal. We then compared the sensitivity and specificity of the MMT to those of FIT using the manufacturer recommended threshold of 100 ng/ml of hemoglobin. At these cutoffs the MMT detected 95.0% of cancers and 57.1% of adenomas compared to 75.0% and 15.7% for FIT (Table 1, Fig. 1B). When adenomas and cancers were pooled together, the MMT detected 71.4% of lesions, while FIT only detected 38.1%. The MMT significantly improved sensitivity for both advanced and non-advanced adenomas as well as multiple stages of cancer (Fig. 2). The increased sensitivity of the MMT was accompanied by a decrease in specificity (83.7%) compared to FIT (97.1%).

To better understand the relationship between the MMT and FIT, we compared the results of the two tests for each sample (Fig. 3). All samples that tested positive by FIT also tested positive by the MMT, indicating that the MMT did not miss any of the lesions that FIT was able to detect. However the MMT was able to detect 80% of cancers and 49.1% of adenomas that FIT had failed to detect, while maintaining a specificity of 86.2% (Extended Data Fig. 3). This result demonstrated that incorporation of data from a subject's microbiota complemented FIT to improve its sensitivity.

The purpose of screening is to identify asymptomatic individuals with early stage disease (i.e., true positives). Therefore, we estimated the number of true positives captured through FIT and MMT in the recommended screening population in the United States (adults ages 50-74 years). The prevalence of lesions in an average-risk population was obtained through a previously published meta-analysis[23]. Tests were utilized in series so that FIT, with a higher specificity (fewer false positives), was applied first to minimize unnecessary diagnostic testing. MMT, with a higher sensitivity (fewer false negatives), was then used to capture additional true positives in those with negative FIT results (Extended Data Table 1). MMT was able to identify a large proportion of true positives among individuals with a negative FIT result (55.1% for cancer, 72.0% for advanced adenoma, 82.5% for non-advanced adenoma).

Previous studies have identified differences in diagnostic test performance for certain demographic groups or for people taking certain medications[24–26]. Therefore we tested whether the MMT performance differed between patient populations. We found no difference in model performance according to age, BMI, NSAID usage, diabetes, smoking, or previous history of polyps (all $p>0.05$). However the model was significantly better at differentiating normal from lesion for females than for males ($p=0.016$; Extended Data Fig. 4). For females the model detected 73.5% of lesions with a specificity of 89.2%. For males the model detected 69.9% of lesions with a specificity of 73.8%. This difference was more pronounced for adenomas. The MMT detected 62.5% of adenomas in females and 53.4% in males. Despite performing more poorly overall for males, the MMT did have a higher sensitivity for cancer among males (98.5%) than females (90.4%). The difference in performance between males and females seems to be due to differences in FIT results rather than differences in the microbiome. After correcting for diagnosis, there was a significant effect of sex on FIT result ($p=0.0057$, two-way ANOVA), but not on the overall structure of the microbiome($p=0.063$, PERMANOVA).

It was recently shown that when FIT was combined with host-associated DNA biomarkers the ability to detect adenomas and carcinomas was significantly improved over FIT alone[14]. The sensitivity of the host-associated DNA screen was 92.3% for CRC and 42.4% for adenomas, which are both slightly lower than what we observed with our MMT. Regardless of the relative performance, such results support the assertion that because of the large interpersonal variation in markers for adenomas and carcinomas, it is necessary to employ a panel of biomarkers and to use a model that integrates the biomarkers. The accuracy of our model may be further improved by incorporating additional biomarkers such as the host-associated biomarkers or those targeting specific genes involved in the underlying mechanism of tumorigenesis such as toxins[7,8,17]. More generally, predictive and diagnostic models for other diseases with a microbial etiology may benefit from a similar approach. For example, we recently demonstrated the ability to detect *Clostridium difficile* infection based on the composition of the microbiota[27]. Such models are likely to be useful as microbiota sequencing gains traction as a tool for characterizing health.

Our findings demonstrate the potential for combining the analysis of a patient's microbiota with conventional stool-based tests to improve CRC detection. Using the random forest algorithm it was possible to interpret FIT results in the context of the microbiota. The MMT had significantly higher sensitivity for lesions at almost all stages of tumorigenesis. Moreover the model detected the majority of lesions that FIT was unable to detect. The shortcoming of the MMT is its lower specificity but, by conducting the FIT and MMT in series, it is possible to maximize the number of correctly identified individuals with preclinical lesions. The potential value of the MMT is its higher sensitivity which, at its core, is the purpose of preventive screening – finding lesions earlier so that cancer would be avoided.

Methods Summary. Fecal samples were collected from 490 subjects in 4 locations: Toronto (Ontario, Canada), Boston (Massachusetts, USA), Houston (Texas, USA), and Ann Arbor (Michigan, USA). Patient diagnoses were determined by colonoscopy and subsequent histopathological examination of any biopsies taken. FIT was performed using OC FIT-CHEK sampling bottles and processed using an OC-Auto Micro 80 automated system (Polymedco Inc.). The V4 region of the bacterial 16S rRNA gene was amplified using custom barcoded primers, sequenced using an Illumina MiSeq sequencer, and analyzed as described previously[28]. A data analysis pipeline and all necessary scripts to generate this paper are available at github.com/SchlossLab/Baxter_glne007Modeling_2015. The sequence data are available in the Sequence Read Archive under accession number SRP062005.

Methods (online only)

Study Design/Patient sampling. Eligible patients for this study were at least 18 years old, willing to sign informed consent, able to tolerate removal of 58 ml of blood, and willing to collect a stool sample. Patient age at the time of enrollment ranged from 29 to 89 with a median of 60. All patients were asymptomatic and were excluded if they had undergone surgery, radiation, or chemotherapy for current CRC prior to baseline samples or had inflammatory bowel disease, known hereditary non-polyposis CRC, or familial adenomatous polyposis. Colonoscopies were performed and fecal samples were collected from subjects in 4 locations: Toronto (Ontario, Canada), Boston (Massachusetts, USA), Houston (Texas, USA), and Ann Arbor (Michigan, USA). Patient diagnoses were determined by colonoscopic examination and histopathological review of any biopsies taken. Patients with an adenoma greater than 1cm, more than three adenomas of any size, or an adenoma with villous histology were classified as advanced adenoma. Whole evacuated stool was collected from each patient either prior to colonoscopy preparation or 1-2 weeks after colonoscopy. This has been shown to be sufficient time for the microbiota to recover from colonoscopy preparation[29].

Stool samples were packed in ice, shipped to a processing center via next day delivery and stored at -80˚C. This study was approved by the University of Michigan Institutional Review Board and all subjects provided informed consent.

Fecal Immunochemical Tests. Fecal material for FIT was collected from frozen stool aliquots using OC FIT-CHEK sampling bottles (Polymedco Inc.) and processed using an OC-Auto Micro 80 automated system (Polymedco Inc.). Hemoglobin concentrations were used for generating ROC curves for FIT and for building the MMT.

16S rRNA Sequencing. DNA was extracted from roughly 50 mg of fecal material from each subject using the PowerSoil-htp 96 Well Soil DNA isolation kit (MO BIO Laboratories) and an epMotion 5075 automated pipetting system (Eppendorf). The V4 region of the bacterial 16S rRNA gene was amplified using custom barcoded primers and sequenced as described previously using an Illumina MiSeq sequencer[28]. The 490 samples were divided into three sequencing runs to increase the per sample sequencing depth. Although the same percentage of samples from the three groups were represented on each sequencing run, samples were randomly assigned to the sequencing runs to avoid confounding our analysis based on diagnosis or demographics.

Sequence Curation. The 16S rRNA gene sequences were curated using the mothur software package, as described previously[28]. Briefly, paired-end reads were merged into contigs, screened for quality, aligned to SILVA 16S rRNA sequence database, and screened for chimeras. Curated sequences were clustered in to operational taxonomic units (OTUs) using a 97% similarity cutoff with the average neighbor clustering algorithm. The number of sequences in each sample was rarefied to 10,000 per sample to minimize the effects of uneven sampling.

Statistical Methods. All statistical analyses were performed using R. Random Forest models were generated using the AUCRF package[19]. The AUC of ROC curves was compared using the method described by DeLong et al.[30]. The optimal cutoff for the MMT was determined using Youden's *J* statistic as implemented in the pROC package in R[22]. The sensitivities of FIT and the MMT were compared using McNemar's chi-squared test. To control for diagnosis while testing the effects of sex on the microbiome we used PERMANOVA as implemented in the adonis funciton in the vegan package.

Data Availability. Raw fastq files and a MIMARKS file are available through the NCBI Sequence Read Archive [SRP062005]. A data analysis pipeline and all necessary scripts are available at github.com/SchlossLab/Baxter_glne007Modeling_2015.

Literature cited

1. Siegel, R., DeSantis, C. & Jemal, A. Colorectal cancer statistics, 2014. *CA: A Cancer Journal for Clinicians* **64,** 104–117 (2014).

2. Disease Control, C. for, (CDC, P. & others. Vital signs: Colorectal cancer screening test use–United states, 2012. *MMWR. Morbidity and Mortality Weekly Report* **62,** 881 (2013).

3. Hsia, J. *et al.* The importance of health insurance as a determinant of cancer screening: evidence from the Women's Health Initiative. *Preventive Medicine* **31,** 261–270 (2000).

4. Jones, R. M., Devers, K. J., Kuzel, A. J. & Woolf, S. H. Patient-reported barriers to colorectal cancer screening: a mixed-methods analysis. *American Journal of Preventive Medicine* **38,** 508–516 (2010).

5. Zackular, J. P. *et al.* The gut microbiome modulates colon tumorigenesis. *MBio* **4,** e00692–13 (2013).

6. Kostic, A. D. *et al.* Fusobacterium nucleatum potentiates intestinal tumorigenesis and modulates the tumor-immune microenvironment. *Cell Host & Microbe* **14,** 207–215 (2013).

7. Wu, S. *et al.* A human colonic commensal promotes colon tumorigenesis via activation of T helper type 17 T cell responses. *Nature Medicine* **15,** 1016–1022 (2009).

8. Arthur, J. C. *et al.* Intestinal inflammation targets cancer-inducing activity of the microbiota. *Science* **338,** 120–123 (2012).

9. Wang, T. *et al.* Structural segregation of gut microbiota between colorectal cancer patients and healthy volunteers. *The ISME Journal* **6,** 320–329 (2012).

10. Chen, H.-M. *et al.* Decreased dietary fiber intake and structural alteration of gut microbiota in patients with advanced colorectal adenoma. *The American Journal of Clinical Nutrition* **97,** 1044–1052 (2013).

11. Chen, W., Liu, F., Ling, Z., Tong, X. & Xiang, C. Human intestinal lumen and mucosa-associated microbiota in patients with colorectal cancer. *PloS One* **7,** e39743 (2012).

12. Shen, X. J. *et al.* Molecular characterization of mucosal adherent bacteria and associations with colorectal adenomas. *Gut Microbes* **1,** 138–147 (2010).

13. Kostic, A. D. *et al.* Genomic analysis identifies association of Fusobacterium with colorectal carcinoma. *Genome Research* **22,** 292–298 (2012).

14. Imperiale, T. F. *et al.* Multitarget stool DNA testing for colorectal-cancer screening. *New England Journal of Medicine* **370,** 1287–1297 (2014).

15. Hundt, S., Haug, U. & Brenner, H. Comparative evaluation of immunochemical fecal occult blood tests for colorectal adenoma detection. *Annals of Internal Medicine* **150,** 162–169 (2009).

16. Zackular, J. P., Rogers, M. A., Ruffin, M. T. & Schloss, P. D. The human gut microbiome as a screening tool for colorectal cancer. *Cancer Prevention Research* **7,** 1112–1121 (2014).

17. Zeller, G. *et al.* Potential of fecal microbiota for early-stage detection of colorectal cancer. *Molecular Systems Biology* **10,** 766 (2014).

18. Liaw, A. & Wiener, M. Classification and regression by randomForest. *R News* **2,** 18–22 (2002).

19. Calle, M. L., Urrea, V., Boulesteix, A.-L. & Malats, N. AUC-RF: A new strategy for genomic profiling with random forest. *Human Heredity* **72,** 121–132 (2011).

20. 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 Microbiology Letters* **217,** 133–139 (2002).

21. Rex, D. K. *et al.* American College of Gastroenterology guidelines for colorectal cancer screening 2008. *The American Journal of Gastroenterology* **104,** 739–750 (2009).

22. Youden, W. J. Index for rating diagnostic tests. *Cancer* **3,** 32–35 (1950).

23. Heitman, S. J. *et al.* Prevalence of adenomas and colorectal cancer in average risk individuals: a systematic review and meta-analysis. *Clinical Gastroenterology and Hepatology* **7,** 1272–1278 (2009).

24. Symonds, E. L. *et al.* Factors affecting faecal immunochemical test positive rates: demographic, pathological, behavioural and environmental variables. *Journal of Medical Screening* 0969141315584783 (2015).

25. Kapidzic, A. *et al.* Gender differences in fecal immunochemical test performance for early detection of colorectal neoplasia. *Clinical Gastroenterology and Hepatology* (2015).

26. Levi, Z. *et al.* Sensitivity, but not specificity, of a quantitative immunochemical fecal occult blood test for neoplasia is slightly increased by the use of low-dose aspirin, NSAIDs, and anticoagulants. *The American Journal of Gastroenterology* **104,** 933–938 (2009).

27. Schubert, A. M., Sinani, H. & Schloss, P. D. Antibiotic-Induced Alterations of the Murine Gut Microbiota and Subsequent Effects on Colonization Resistance against Clostridium difficile. *MBio* **6,** (2015).

28. Kozich, J. J., Westcott, S. L., Baxter, N. T., Highlander, S. K. & Schloss, P. D. Development of a dual-index sequencing strategy and curation pipeline for analyzing amplicon sequence data on the MiSeq Illumina sequencing platform. *Applied and Environmental Microbiology* **79,** 5112–5120 (2013).

29. O'Brien, C. L., Allison, G. E., Grimpen, F. & Pavli, P. Impact of Colonoscopy Bowel Preparation on Intestinal Microbiota. *PLoS ONE* **8,** e62815 (2013).

30. DeLong, E. R., DeLong, D. M. & Clarke-Pearson, D. L. Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. *Biometrics* 837–845 (1988).

Author Contributions

All authors were involved in the conception and design of the study. NTB processed samples and analyzed the data. All authors interpreted the data. NTB and PDS wrote the manuscript. All authors reviewed and revised the manuscript.

Author Information

The authors declare no competing financial interests.

279

Figures

Diagnosis		FIT		Multitarget Microbiota Test	
		True Positives	Sensitivity (95% CI)	True Positives	Sensitivity (95% CI)
Cancer	n=120	90	**75.0** (67.5-82.5)	114	**95.0** (90.8-98.3)
Advanced Adenoma	n=109	21	**19.3** (11.9-27.5)	64	**58.7** (49.5-67.9)
Non Advanced Adenoma	n=89	10	**11.2** (5.62-18)	49	**55.1** (43.8-65.2)
Any Lesions	n=318	121	**38.1** (33-43.4)	227	**71.4** (66.4-76.4)
		True Negatives	Specificity (95% CI)	True Negatives	Specificity (95% CI)
Normal	n=172	167	**97.1** (94.2-99.4)	144	**83.7** (77.9-89)

Table 1. Table of sensitivities and specificities for FIT and MMT. The 95% confidence intervals were computed with 2000 stratified bootstrap replicates.

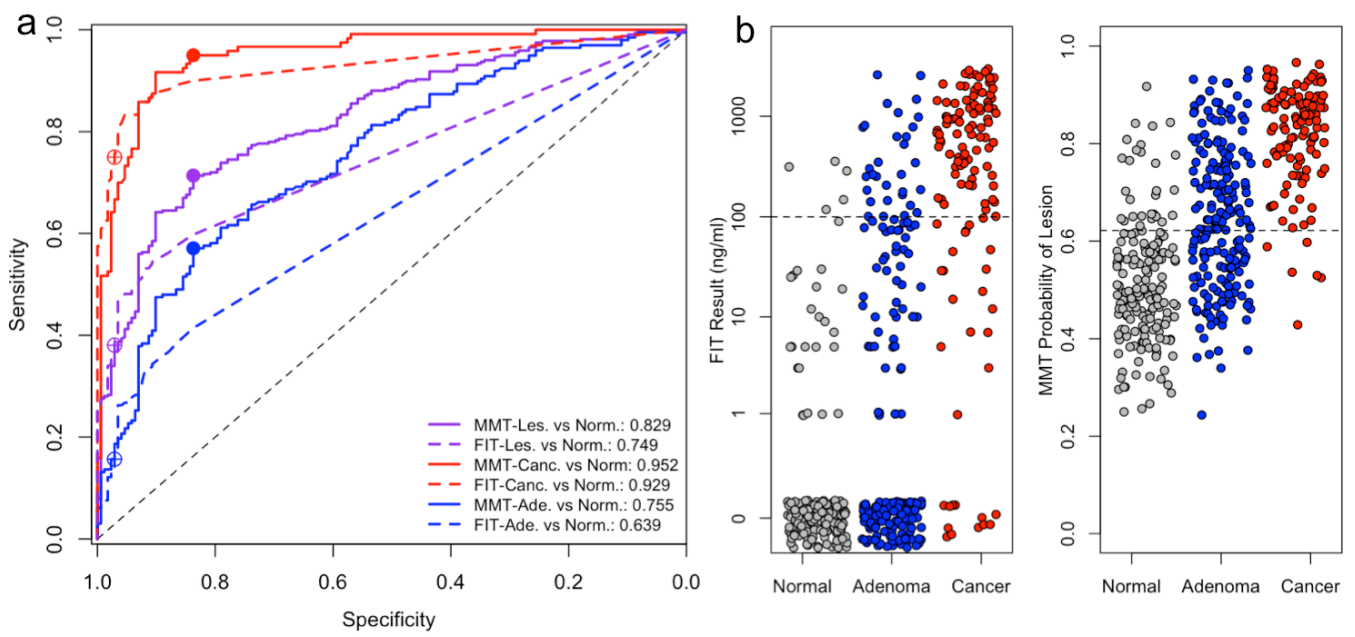
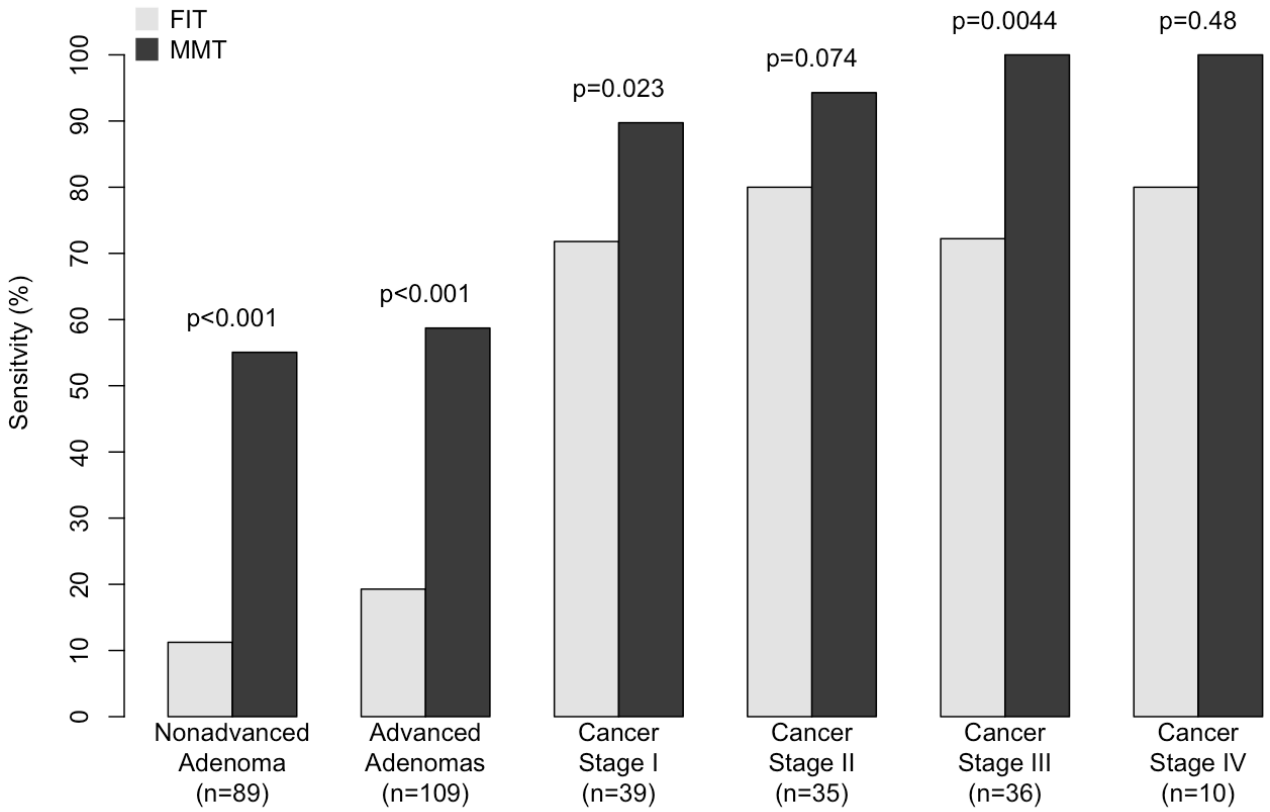


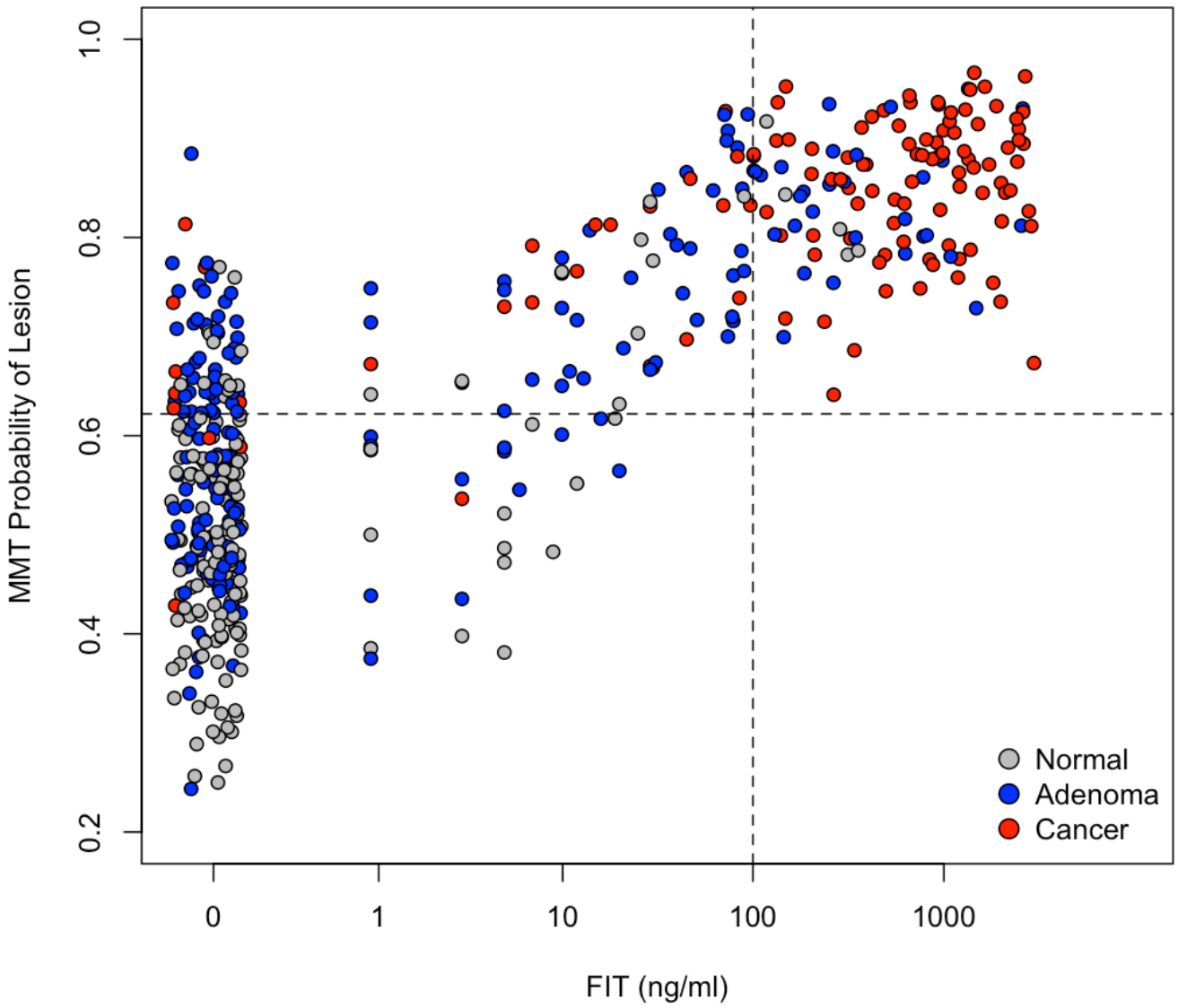
Figure 1. (a) ROC Curves for the MMT (solid lines) and FIT (dashed lines) for distinguishing normal from any lesion (purple), normal from cancer (red) and normal from adenoma (blue). Filled dots show the sensitivity and specificity of the MMT at the optimal cutoff (0.622). Open dots show the sensitivity and specificity of FIT at the 100ng/ml cutoff. (b) Strip charts showing the results for FIT and the MMT. Dashed lines show the cutoff for each test. Points with a FIT result of 0 are jittered to improve visibility.



293

294 **Figure 2.** Barplot of sensitivities for FIT and MMT for each stage of tumor development. P-values

295 based on McNemar's chi-squared test.



296

297 **Figure 3.** Scatter plot of the results of FIT and MMT for each sample. Dashed lines show the cutoff

298 for each test. Points with a FIT result of 0 are jittered to improve visibility.

299

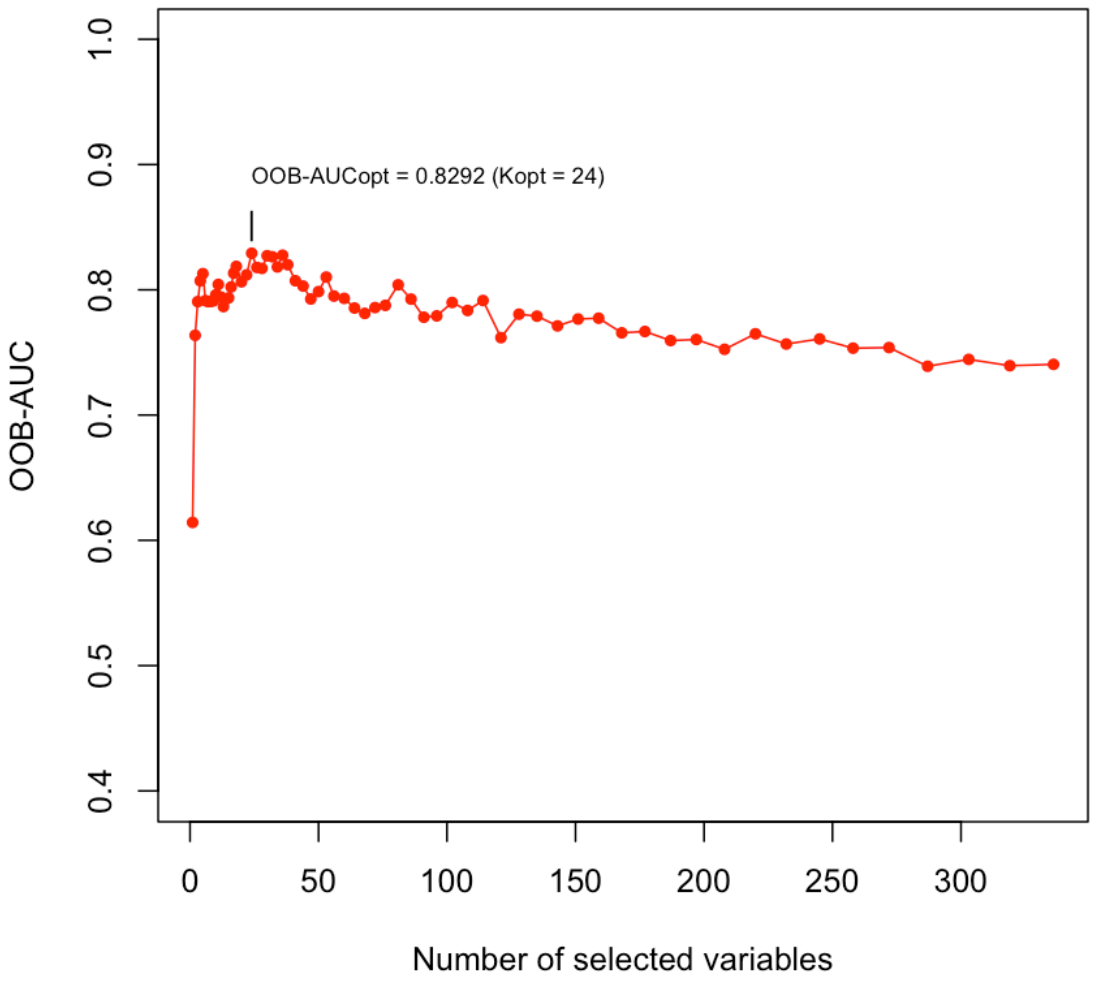
Extended Data Figures

Number and Percentage of True Positives Identified through FIT and MMT

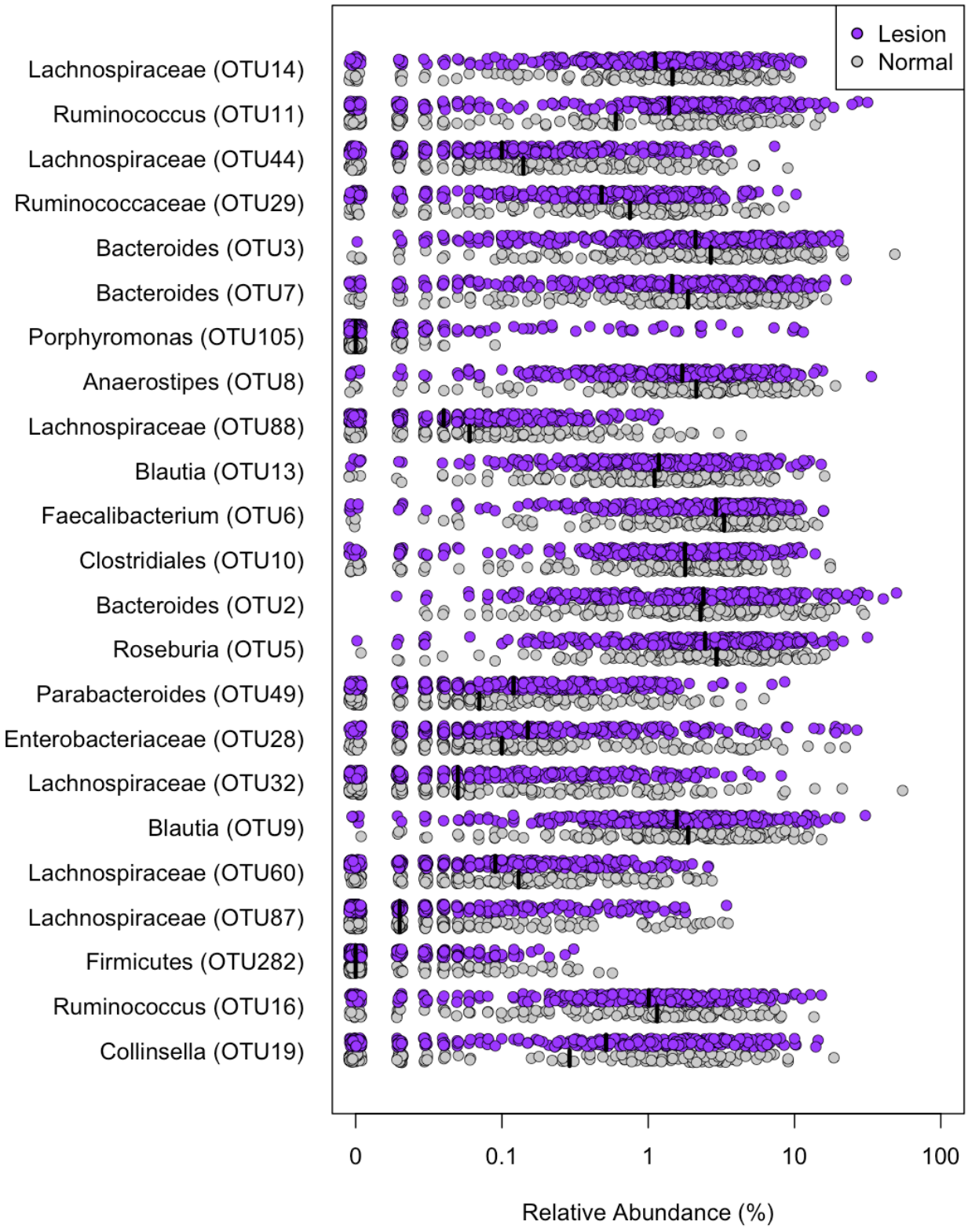
Condition	Prevalence	Number of Persons, ages 50-75 years, with Condition	True Positives identified through FIT	True Positives identified through additional MMT	Percentage of True Positives identified through additional MMT
Cancer	0.3%	241,483	181,112	222,260	55.1%
Advanced Adenoma	5.7%	4,588,174	885,518	2,276,159	72.0%
Non-advanced Adenoma	17.7%	14,247,488	1,595,719	7,507,376	82.5%

*Number of persons in the United States in 2010, 50-75 years of age, was 80,494,283.

Extended Data Table 1. Number and proportion of true positives identified through FIT and MMT in the United States in adults 50-75 years of age, based on published estimates of CRC prevalence. Far right column shows percentage of true positives identified among individuals with a negative FIT result.



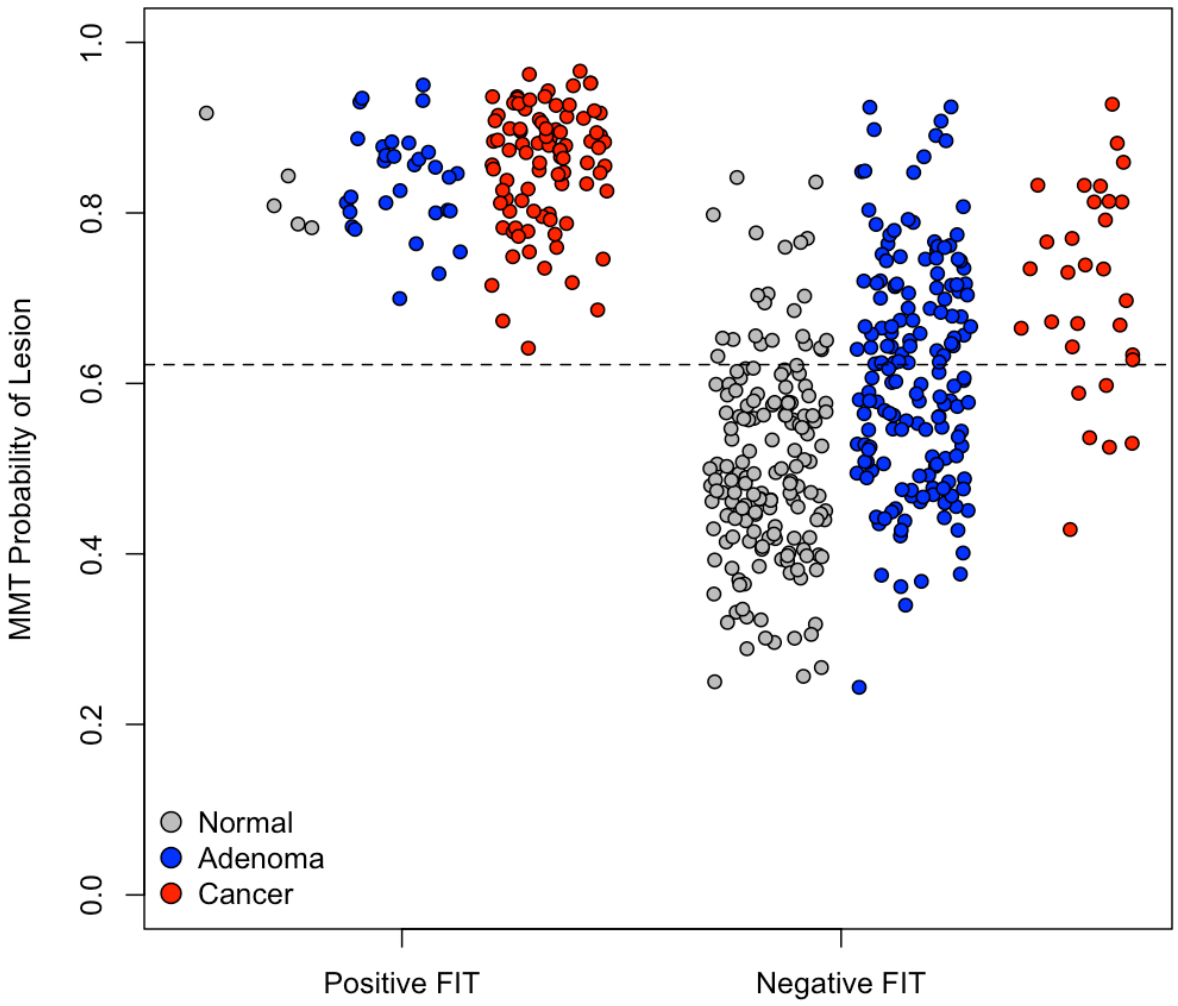
Extended Data Figure 1. Change in out-of-bag AUC with number of features in the MMT. The optimal model contains 24 features and has an AUC of 0.829.



311

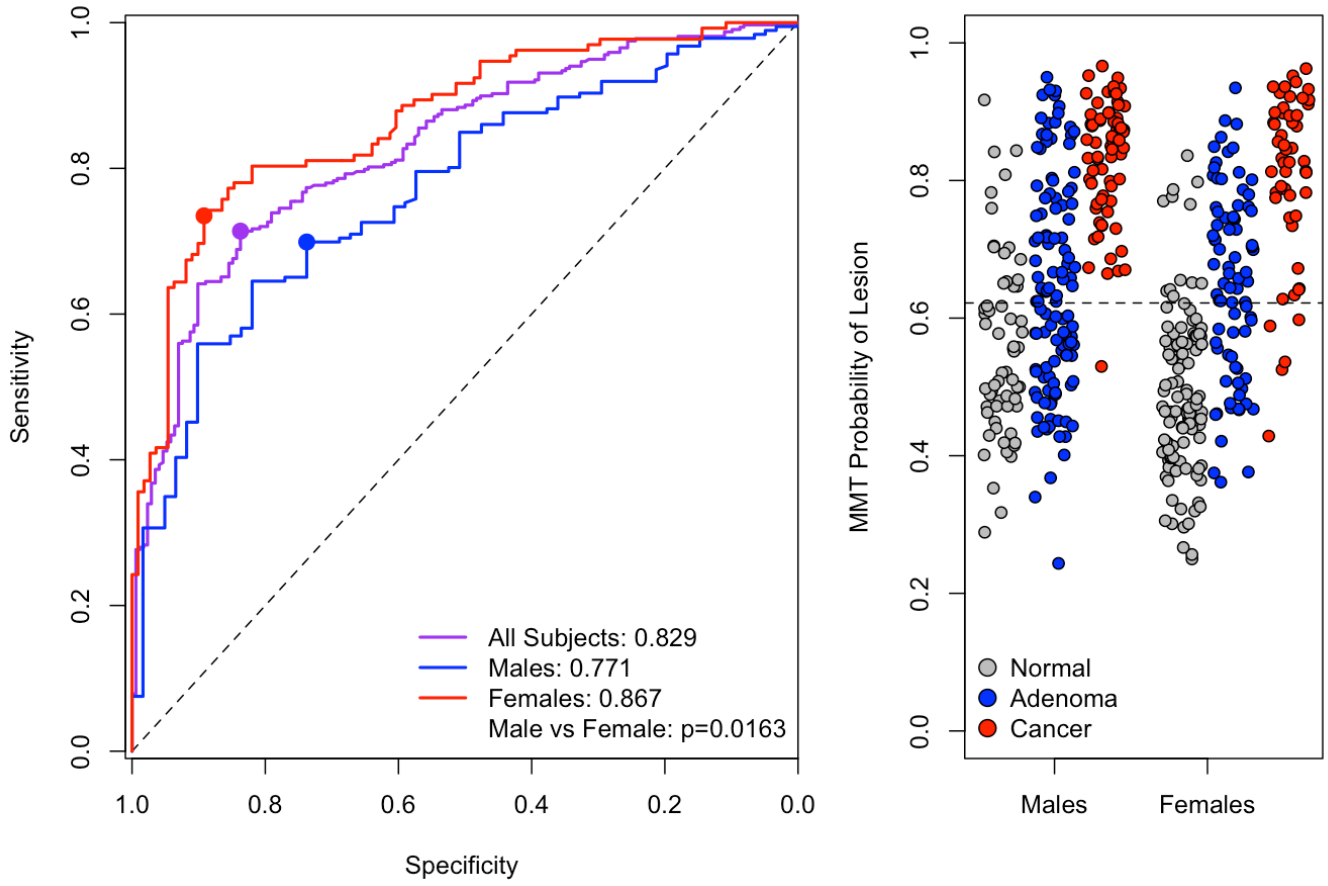
312 **Extended Data Figure 2.** Stripchart of the relative abundances of each OTU in the MMT with black

313 lines at the medians.



314

315 **Extended Data Figure 3.** Stripchart of MMT results for each sample based on FIT result.



316

317 **Extended Data Figure 4.** ROC curves (left) and stripchart (right) of MMT separated by sex.