1 # DNA from fecal immunochemical test can replace

2 # stool for microbiota-based colorectal cancer

3 # screening

4 **Authors:** Nielson T. Baxter[1], Charles C. Koumpouras[1], Mary A.M. Rogers[2], Mack T.

5 Ruffin IV[3], and Patrick D. Schloss[1*]

6 **Affiliations:**

7 [1]Department of Microbiology and Immunology, University of Michigan, Ann Arbor,

8 Michigan.

9 [2]Department of Internal Medicine, University of Michigan, Ann Arbor, Michigan.

10 [3]Department of Family Medicine, University of Michigan, Ann Arbor, Michigan.

11 [*]Corresponding author: *pschloss@umich.edu*

12 **Email addresses:**

13 NTB: *ntbaxter@umich.edu*

14 CCK: *chkoumpo@umich.edu*

15 MAMR: *maryroge@med.umich.edu*

16 MTR: *mruffin@med.umich.edu*

17 PDS: *pschloss@umich.edu*

Abstract

Background: There is a significant demand for colorectal cancer (CRC) screening methods that are noninvasive, inexpensive, and capable of accurately detecting early stage tumors. It has been shown that models based on the gut microbiota can complement the fecal occult blood test and fecal immunochemical test (FIT). However, a barrier to microbiota-based screening is the need to collect and store a patient's stool sample.

Methods: Using stool samples collected from 404 patients we tested whether the residual buffer containing resuspended feces in FIT cartridges could be used in place of intact stool samples.

Results: We found that the bacterial DNA isolated from FIT cartridges largely recapitulated the community structure and membership of patients' stool microbiota and that the abundance of bacteria associated with CRC were conserved. We also found that models for detecting CRC that were generated using bacterial abundances from FIT cartridges were equally predictive as models generated using bacterial abundances from stool.

Conclusions: These findings demonstrate the potential for using residual buffer from FIT cartridges in place of stool for microbiota-based screening for CRC. This may reduce the need to collect and process separate stool samples and may facilitate combining FIT and microbiota-based biomarkers into a single test. Additionally, FIT cartridges could constitute a novel data source for studying the role of the microbiome in cancer and other diseases.

40 **Keywords:** colorectal cancer, gut microbiome, microbiota, fecal immunochemical test,

41 random forest

Background

43 Although colorectal cancer (CRC) mortality has declined in recent decades, it remains

44 the second leading cause of death among cancers in the United States [1]. Early

45 detection of CRC is critical since patients whose tumors are detected at an early stage

46 have a greater than 90% chance of survival [1]. However more than a third of

47 individuals for whom screening is recommended do not adhere to screening guidelines

48 [2]. The high cost and invasive nature of procedures, such as colonoscopy and

49 sigmoidoscopy are barriers for many people [3, 4]. Unfortunately non-invasive tests,

50 such as the guaiac fecal occult blood test (gFOBT), fecal immunochemical test (FIT),

51 and the multitarget DNA test fail to reliably detect adenomas [5, 6] (e.g., sensitivity for

52 nonadvanced adenomas is 7.6% for FIT and 17.2% for the DNA test). Thus, there is a

53 need for novel non-invasive screening methods with improved sensitivity for early stage

54 colonic lesions.

55 Several studies have demonstrated the potential for the gut microbiota to be used to

56 detect CRC [7–10]. Moreover, we and others have shown that combining microbiota-

57 analysis with conventional diagnostics, like gFOBT and FIT, can significantly improve

58 the detection of colonic lesions over either method by itself [7, 8, 10]. One limitation of

59 microbiota-based CRC screening is the need to collect and process separate stool

60 samples for microbiota characterization. Given the widespread use of FIT to collect

61 specimens for screening, the ability to use the same sample for microbiota

62 characterization could make processing more efficient and less expensive. We

3

63   hypothesized that the small amount of fecal material contained in FIT sampling

64   cartridges was sufficient to perform both hemoglobin quantification and microbiota

65   characterization. To test this hypothesis, we isolated bacterial DNA from the residual

66   buffer of OC-Auto® FIT cartridges (Polymedco Inc.) that had already been used for

67   quantifying fecal hemoglobin concentrations. We then compared the bacterial

68   composition of the FIT cartridge to that of DNA isolated directly from a patient's stool

69   sample and assessed the ability of FIT cartridge-derived DNA to be used for microbiota-

70   based CRC screening.

71   **Materials and Methods**

72   **Study Design / Diagnoses / Stool Collection.** Stool samples were obtained through

73   the Great Lakes-New England Early Detection Research Network. Patients were

74   asymptomatic, at least 18 years old, willing to sign informed consent, able to tolerate

75   removal of 58 mL of blood, and willing to collect a stool sample. Patient age at the time

76   of enrollment ranged from 29 to 89 with a median of 60 years. Patients were excluded if

77   they had undergone surgery, radiation, or chemotherapy for current CRC prior to

78   baseline samples or had inflammatory bowel disease, known hereditary non-polyposis

79   CRC, or familial adenomatous polyposis. Patient diagnoses were determined by

80   colonoscopic examination and histopathological review of any biopsies taken.

81   Colonoscopies were performed and fecal samples were collected in four locations:

82   Toronto (Ontario, Canada), Boston (Massachusetts, USA), Houston (Texas, USA), and

83   Ann Arbor (Michigan, USA). Stool samples were packed in ice, shipped to a processing

84   center via next day delivery and stored at -80˚C. Fecal material for FIT was collected

85   from frozen stool aliquots using OC-Auto® FIT sampling bottles (Polymedco Inc.),

4

86   processed using an OC-Auto Micro 80 automated system (Polymedco Inc.), and stored

87   at -20C. The University of Michigan Institutional Review Board approved this study, and

88   all subjects provided informed consent.

89   **16S rRNA gene sequencing.** Processed FIT samples were thawed, and 100 μl of

90   buffer were withdrawn by pipette for DNA extraction. DNA was isolated from FIT

91   samples or matching stool samples using the PowerSoil-htp 96 Well Soil DNA isolation

92   kit (MO BIO Laboratories) and an epMotion 5075 automated pipetting system

93   (Eppendorf). The V4 region of the bacterial 16S rRNA gene was amplified using custom

94   barcoded primers and sequenced as described previously using an Illumina MiSeq

95   sequencer [11]. The 16S rRNA gene sequences were curated using the mothur

96   software package, as described previously [11, 12]. Curated sequences were clustered

97   into operational taxonomic units (OTUs) using a 97% similarity cutoff with the average

98   neighbor clustering algorithm. Sequences were classified using a naive Bayesian

99   classifier trained against a 16S rRNA gene training set provided by the Ribosomal

100   Database Project (RDP) [13]. Species-level classifications for OTUs of interest were

101   determined by using blastn to compare the predominant sequence within each OTU to

102   the NCBI 16S rRNA database. The putative species was only reported for OTUs with

103   greater than 99% sequence identity to a single species in the database; otherwise the

104   consensus RDP classification was used.

105   **Statistical Methods.** All statistical analyses were performed using R (v.3.2.0). Random

106   forest models were generated using the AUC-RF algorithm for feature reduction and

107   maximizing model performance [14]. The most predictive OTUs were determined based

108   on mean decrease in accuracy when removed from the model. The area under the

5

109   curve (AUC) of receiver operator characteristic (ROC) curves were compared using the

110   method described by DeLong et al. [15] as implemented in the pROC R package [16].

Results

112   DNA was isolated and 16S rRNA gene sequencing was performed on stool aliquots and

113   the residual buffer of paired OC-Auto® FIT sampling cartridges from 404 patients.

114   Among these patients, 101 had CRC, 162 had adenomas, and 141 had no colonic

115   lesions. First, we tested whether the bacterial community profiles from FIT cartridges

116   recapitulated their stool counterparts. First, we compared the number of OTUs shared

117   within FIT/stool pairs from the same patient to the number of OTUs shared between

118   patients (Fig. 1A). FIT cartridges and stool from the same patient (red line) had

119   significantly more bacterial populations in common than those taken from different

120   patients ($p < 0.001$, two-sample Kolmogorov-Smirnov test), indicating that community

121   membership was conserved within patients across stool and FIT cartridges. Second, we

122   calculated the similarity in community structure between samples using 1-thetaYC index

123   [17]. This metric compares the presence or absence of bacterial populations and their

124   relative abundance. The bacterial community structure of stool and FIT samples from

125   the same patient (red line) were significantly more similar to each other than to stool or

126   FIT from other patients (Fig. 1B, $p < 0.001$). Finally, we used a Mantel test to determine

127   whether the patient-to-patient thetaYC distances among stool samples were correlated

128   with the patient-to-patient thetaYC distances among FIT cartridges. We found that there

129   was a significant correlation (Mantel test $r = 0.525$, $p < 0.001$), suggesting that the inter-

130   patient variation in community structure between the stool samples of patients was

131   conserved in samples from FIT cartridges.

6

132 Next, we observed a significant correlation between the abundance of each genus in

133 the paired FIT cartridge and stool samples (Fig. 2A, Spearman rho: 0.699, p<0.001).

134 This suggested that the abundance of bacterial genera was conserved. This correlation

135 was especially strong when comparing only the 100 most abundant genera from stool

136 (Spearman rho: 0.886, p<0.001). Several bacterial species have been repeatedly

137 associated with CRC, including *Fusobacterium nucleatum*, *Porphyromonas*

138 *asaccharolytica*, *Peptostreptococcus stomatis*, and *Parvimonas micra* [8–10, 18]. As

139 expected, the abundance of these species in stool was significantly correlated with their

140 abundance in matched FIT cartridges (all p<0.001, Spearman rho ≥0.352)(Fig. 2B). We

141 observed some biases in the abundance of certain taxa. In particular, the genus

142 *Pantoea* was detected in 399 of the 404 FIT cartridges with an average abundance of

143 2.4%, but was only detected in 1 stool sample. The genus *Helicobacter* was detected in

144 172 FIT cartridges, but only 10 stool samples. Likewise several genera of *Actinobacteria*

145 were more abundant in stool samples compared to FIT. Notwithstanding these few

146 exceptions, the abundance of the vast majority of genera were well conserved between

147 stool and FIT cartridges. Overall, these findings suggested that that the overall bacterial

148 community structure and the abundance of specific taxa in FIT cartridges and stool

149 were similar.

150 We tested whether the bacterial relative abundances we observed from FIT cartridges

151 could be used to differentiate healthy patients from those with carcinomas using random

152 forest models as we did previously using intact stool samples [10]. Using DNA from the

153 FIT cartridge, the optimal model utilized 28 OTUs and had an AUC of 0.831 (Fig. 3A).

154 There was not a significant difference in the AUC for this model and the model based on

155   DNA isolated directly from stool, which used 32 OTUs and had an AUC of 0.853

156   (p=0.41). Furthermore, the probabilities of individuals having lesions were correlated

157   between the models generated using DNA isolated from the FIT cartridges and stool

158   samples (Spearman rho: 0.633, p<0.001, Fig. 3B). We also generated random forest

159   models for differentiating healthy patients from those with any type of lesions (i.e.

160   adenoma or carcinoma). There was not a significant difference in AUC between the

161   stool-based model with 41 OTUs (AUC=0.700) and the FIT cartridge-based model with

162   41 OTUs (AUC=0.686, p=0.65, Fig. 3C). Again, the probabilities of individuals having

163   lesions according to the two models were significantly correlated (Spearman rho: 0.389,

164   p<0.001 Fig. 3D). These findings demonstrated that models based on bacterial DNA

165   from FIT cartridges were as predictive as models based on DNA isolated directly from

166   stool.

167   **Discussion**

168   Bacterial DNA isolated from the residual buffer of FIT cartridges recapitulated the

169   community structure and membership of patients' stool microbiota. FIT/stool pairs

170   collected from the same patient were significantly more similar to each other than

171   samples from different patients and the inter-patient differences in stool microbiota

172   structure were conserved in FIT cartridge-derived microbiota. More importantly, random

173   forest models generated using bacterial abundances from FIT cartridge-derived and

174   stool-derived DNA were equally predictive for differentiating healthy patients from those

175   with adenomas and carcinomas.

176   Sinha et al. compared a variety of sampling and storage methods for fecal samples to

177   be used for microbiome analyses [19]. They found reproducible biases according to

178 sampling method and time at ambient temperature. Likewise, we observed biases in the

179 abundance certain bacterial populations in FIT cartridges compared stool. For example,

180 an OTU associated with *Pantoea* was found in 98.8% of FIT cartridge samples and only

181 0.2% of stool samples. There are several possible explanations for this result. It is

182 possible that because the biomass contained in the FIT cartridges is considerably lower

183 than that in stool, the analysis was more sensitive to contaminants in our reagents or

184 the FIT cartridge [20]. Alternatively, storage conditions could have played a role in

185 biasing the relative abundances of certain genera. The feces in the FIT cartridges spent

186 more time exposed to ambient temperatures in order to be analyzed for hemoglobin

187 concentration. Therefore it is possible that certain bacterial populations, especially

188 aerobes, were able to grow. Considering *Pantoea* is rarely found in human feces and is

189 more commonly found in soil, plant surfaces, and air we suspect that it was a

190 contaminant. Regardless of the source of this and the other suspicious populations, any

191 biases were limited since the random forest feature selection process did not select

192 these populations and did not affect the ability to detect CRC from FIT cartridge-derived

193 DNA.

Conclusions

195 This could reduce the need to collect and process separate stool samples, decreasing

196 the cost of screening. It may be possible to use FIT cartridges rather than separate stool

197 samples for future studies on the role of the gut microbiota and cancer. Samples

198 collected from patients who undergo annual FIT screening could be used to monitor

199 temporal changes in a patient's microbiota, making it possible to detect shifts toward a

200 disease-associated microbiota. Since FIT cartridges are currently used for CRC

201   screening, our findings may facilitate large-scale validations of microbiota-based

202   screening methods.

Declarations

Abbreviations:

FIT: fecal immunochemical test

gFOBT: guaic fecal occult blood test

OTU: operational taxonomic unit

AUC: area under the curve

ROC curve: reciever operating characteristic curve

Ethics approval and consent to participate: The University of Michigan Institutional Review Board approved this study, and all subjects provided informed consent. This study conformed to the guidelines of the Helsinki Declaration.

Availability of data and materials: Raw fastq files and a MIMARKS file are available through the NCBI Sequence Read Archive [SRP062005]. The data processing steps for going from the raw sequence data to the final manuscript is available at http://www.github.com/SchlossLab/Baxter_FITs_BMCCancer_2016.

Competing interests: The authors declare no competing financial interests.

Funding: This study was supported by funding from the National Institutes of Health to P. Schloss (R01GM099514, P30DK034933) and to the Early Detection Research Network (U01CA86400).

Author Contributions: PDS, MTR, MAMR, and NTB were involved in the conception and design of the study. NTB and CCK performed DNA extractions and 16S rRNA gene

10

223   sequencing. NTB analyzed the data. All authors interpreted the data. NTB and PDS

224   wrote the manuscript. All authors reviewed and revised the manuscript.

11

Figures

Figure 1. Bacterial community structure from FIT cartridge recapitulates stool.

Density plots showing distribution of the number of shared OTUs (A) and community

similarity (B) between groups of samples (* p<0.001 two-sample Kolmogorov-Smirnov

Test).

Figure 2. Bacterial populations conserved between stool and FIT cartridge. (A)

Scatterplot of the average relative abundance of each bacterial genus in stool and FIT

cartridges colored by phylum. (B) Scatterplots of the relative abundances of 4 species

frequently associated with CRC. All correlations were greater than 0.35 (all p<0.001).

Figure 3. Microbiota-based models from FIT cartridge DNA are as predictive as

models from stool. (A) ROC curves for distinguishing healthy patients from those with

cancer using microbiota-based random forest models using DNA from FIT cartridges or

stool. (B) Probability of having cancer for each patient according to microbiota-based

models from A. (C) ROC curves for distinguishing patients with adenomas or

carcinomas from healthy patients using microbiota-based random forest models using

DNA from FIT cartridges or stool. (D) Probability of having a lesion for each patient

based on the models from C.

References

244   **References**

245   1. Siegel R, DeSantis C, Jemal A: **Colorectal cancer statistics, 2014**. *CA: a cancer*

246   *journal for clinicians*. 2014, **64**:104–117.

247   2. Centers for Disease Control and Prevention: **Vital signs: Colorectal cancer**

248   **screening test use–United states, 2012.** *MMWR Morbidity and mortality weekly*

249   *report*. 2013, **62**:881.

250   3. Hsia J, Kemper E, Kiefe C, Zapka J, Sofaer S, Pettinger M, Bowen D, Limacher M,

251   Lillington L, Mason E, others: **The importance of health insurance as a determinant**

252   **of cancer screening: evidence from the Women's Health Initiative**. *Preventive*

253   *medicine*. 2000, **31**:261–270.

254   4. Jones RM, Devers KJ, Kuzel AJ, Woolf SH: **Patient-reported barriers to colorectal**

255   **cancer screening: a mixed-methods analysis**. *American journal of preventive*

256   *medicine*. 2010, **38**:508–516.

257   5. Hundt S, Haug U, Brenner H: **Comparative evaluation of immunochemical fecal**

258   **occult blood tests for colorectal adenoma detection**. *Annals of Internal Medicine*.

259   2009, **150**:162–169.

260   6. Imperiale TF, Ransohoff DF, Itzkowitz SH, Levin TR, Lavin P, Lidgard GP, Ahlquist

261   DA, Berger BM: **Multitarget stool DNA testing for colorectal-cancer screening**. *New*

262   *England Journal of Medicine*. 2014, **370**:1287–1297.

263   7. Zackular JP, Rogers MA, Ruffin MT, Schloss PD: **The human gut microbiome as a**

264   **screening tool for colorectal cancer**. *Cancer Prevention Research*. 2014, **7**:1112–

265   1121.

266   8. Zeller G, Tap J, Voigt AY, Sunagawa S, Kultima JR, Costea PI, Amiot A, Böhm J,

267   Brunetti F, Habermann N, others: **Potential of fecal microbiota for early-stage**

268   **detection of colorectal cancer**. *Molecular systems biology*. 2014, **10**:766.

269   9. Yu J, Feng Q, Wong SH, Zhang D, Liang Q yi, Qin Y, Tang L, Zhao H, Stenvang J, Li

270   Y, others: **Metagenomic analysis of faecal microbiome as a tool towards targeted**

271   **non-invasive biomarkers for colorectal cancer**. *Gut*. 2015:gutjnl–2015.

272   10. Baxter NT, Ruffin MT, Rogers MA, Schloss PD: **Microbiota-based model**

273   **improves the sensitivity for detecting colonic lesions**. *Genome Medicine*. 2016,

274   **8**:1–10.

275   11. Kozich JJ, Westcott SL, Baxter NT, Highlander SK, Schloss PD: **Development of a**

276   **dual-index sequencing strategy and curation pipeline for analyzing amplicon**

277   **sequence data on the MiSeq Illumina sequencing platform**. *Applied and*

278   *environmental microbiology*. 2013, **79**:5112–5120.

279   12. Schloss PD, Westcott SL, Ryabin T, Hall JR, Hartmann M, Hollister EB, Lesniewski

280   RA, Oakley BB, Parks DH, Robinson CJ, others: **Introducing mothur: open-source,**

281   **platform-independent, community-supported software for describing and**

282   **comparing microbial communities**. *Applied and environmental microbiology*. 2009,

283   **75**:7537–7541.

284   13. Wang Q, Garrity GM, Tiedje JM, Cole JR: **Naive bayesian classifier for rapid**

285   **assignment of rRNA sequences into the new bacterial taxonomy**. *Applied and*

286   *environmental microbiology*. 2007, **73**:5261–5267.

287   14. Calle ML, Urrea V, Boulesteix A-L, Malats N: **AUC-RF: A new strategy for**

288   **genomic profiling with random forest**. *Human heredity*. 2011, **72**:121–132.

289   15. DeLong ER, DeLong DM, Clarke-Pearson DL: **Comparing the areas under two or**

290   **more correlated receiver operating characteristic curves: a nonparametric**

291   **approach**. *Biometrics*. 1988:837–845.

292   16. Robin X, Turck N, Hainard A, Tiberti N, Lisacek F, Sanchez J-C, Müller M: **pROC:**

293   **an open-source package for R and S+ to analyze and compare ROC curves**. *BMC*

294   *bioinformatics*. 2011, **12**:1.

295   17. Yue JC, Clayton MK: **A similarity measure based on species proportions**.

296   *Communications in Statistics-Theory and Methods*. 2005, **34**:2123–2131.

297   18. Warren RL, Freeman DJ, Pleasance S, Watson P, Moore RA, Cochrane K, Allen-

298   Vercoe E, Holt RA: **Co-occurrence of anaerobic bacteria in colorectal carcinomas**.

299   *Microbiome*. 2013, **1**:16.

300   19. Sinha R, Chen J, Amir A, Vogtmann E, Inman KS, Flores-Munguia R, Sampson JN,

301   Knight R, Chia N, others: **Collecting fecal samples for microbiome analyses in**

302   **epidemiology studies**. *Cancer Epidemiology Biomarkers & Prevention*. 2015:cebp–

303   0951.

15

304   20. Salter SJ, Cox MJ, Turek EM, Calus ST, Cookson WO, Moffatt MF, Turner P,

305   Parkhill J, Loman NJ, Walker AW: **Reagent and laboratory contamination can**

306   **critically impact sequence-based microbiome analyses**. *BMC biology*. 2014, **12**:87.