Gastroenterology 2023;165:670–681

Gut Microbiome Composition Is Associated With Future Onset of Crohn's Disease in Healthy First-Degree Relatives

Juan Antonio Raygoza Garay,[1,2,*] **Williams Turpin**,[2,*] **Sun-Ho Lee**,[1,2,*] Michelle I. Smith,[2] Ashleigh Goethel,[2] Anne M. Griffiths,[3] Paul Moayyedi,[4] Osvaldo Espin-Garcia,[5,6] Maria Abreu,[7] Guy L. Aumais,[8] Charles N. Bernstein,[9] Irit A. Biron,[10] Maria Cino,[1] Colette Deslandres,[11] Iris Dotan,[10] Wael El-Matary,[12] Brian Feagan,[13] David S. Guttman,[14] Hien Huynh,[15] Levinus A. Dieleman,[16] Jeffrey S. Hyams,[17] Kevan Jacobson,[18] David Mack,[19] John K. Marshall,[4] Anthony Otley,[20] Remo Panaccione,[21] Mark Ropeleski,[22] Mark S. Silverberg,[1] A. Hillary Steinhart,[1] Dan Turner,[23] Baruch Yerushalmi,[24] Andrew D. Paterson,[5,25] Wei Xu,[5,6] the CCC GEM Project Research Consortium, and Kenneth Croitoru[1,2]

[1]Division of Gastroenterology & Hepatology, Temerty Faculty of Medicine, University of Toronto, Toronto, Ontario, Canada; [2]Zane Cohen Center for Digestive Diseases, Mount Sinai Hospital, Toronto, Ontario, Canada; [3]Division of Gastroenterology, The Hospital for Sick Children, Toronto, Ontario, Canada; [4]Department of Medicine, Farncombe Family Digestive Health Research Institute, McMaster University, Hamilton, Ontario, Canada; [5]Division of Biostatistics, Dalla Lana School of Public Health, University of Toronto, Toronto, Ontario, Canada; [6]Biostatistics Department, Princess Margaret Cancer Center, University Health Network, Toronto, Ontario, Canada; [7]Division of Gastroenterology, Department of Medicine, University of Miami, Miller School of Medicine, Miami, Florida; [8]Hopital Maisonneuve-Rosemont, Montreal, Quebec, Canada; [9]Inflammatory Bowel Disease Clinical and Research Center and Department of Internal Medicine, Max Rady College of Medicine, Rady Faculty of Health Sciences, University of Manitoba, Manitoba, Winnipeg, Canada; [10]Division of Gastroenterology, Rabin Medical Center, Petah-Tikva, Israel; [11]Department of Hepatology and Pediatric Nutrition, Centre Hospitalier Universitaire Sainte-Justine, Montreal, Quebec, Canada; [12]Pediatric Gastroenterology, Max Rady College of Medicine, University of Manitoba, Manitoba, Winnipeg, Canada; [13]Departments of Epidemiology and Biostatistics, University of Western Ontario, London, Ontario, Canada; [14]Center for the Analysis of Genome Evolution and Function, University of Toronto, Toronto, Ontario, Canada; [15]Department of Pediatrics, University of Alberta, Edmonton, Alberta, Canada; [16]Division of Gastroenterology, Department of Medicine, University of Alberta, Edmonton, Alberta, Canada; [17]Division of Digestive Diseases, Hepatology, and Nutrition, Connecticut Children's Medical Center, Hartford, Connecticut; [18]Research Institute, British Columbia Children's Hospital, Vancouver, British Columbia, Canada; [19]Division of Gastroenterology, Hepatology & Nutrition, Children's Hospital of Eastern Ontario and University of Ottawa, Ottawa, Ontario, Canada; [20]Division of Gastroenterology, Izaak Walton Killam Hospital, Dalhousie University, Halifax, Nova Scotia, Canada; [21]Inflammatory Bowel Disease Unit, University of Calgary, Calgary, Alberta, Canada; [22]Gastrointestinal Diseases Research Unit, Department of Medicine, Queen's University, Kingston, Ontario, Canada; [23]The Juliet Keidan Institute of Pediatric Gastroenterology and Nutrition, Shaare Zedek Medical Center, The Hebrew University of Jerusalem, Jerusalem, Israel; [24]Pediatric Gastroenterology Unit, Soroka University Medical Center and Faculty of Health Sciences, Ben-Gurion University of the Negev, Beer-Sheva, Israel; and [25]Genetics and Genome Biology, The Hospital for Sick Children Research Institute, The Hospital for Sick Children, Toronto, Ontario, Canada

BACKGROUND & AIMS: The cause of Crohn's disease (CD) is unknown, but the current hypothesis is that microbial or environmental factors induce gut inflammation in genetically susceptible individuals, leading to chronic intestinal inflammation. Case-control studies of patients with CD have cataloged alterations in the gut microbiome composition; however, these studies fail to distinguish whether the altered gut microbiome composition is associated with initiation of CD or is the result of inflammation or drug treatment. **METHODS:** In this prospective cohort study, 3483 healthy first-degree relatives (FDRs) of patients with CD were recruited to identify the gut microbiome composition that precedes the onset of CD and to what extent this composition predicts the risk of developing CD. We applied a machine learning approach to the analysis of the gut microbiome composition (based on 16S ribosomal RNA sequencing) to define a microbial signature that associates with future development of CD. The performance of the model was assessed in an independent validation cohort. **RESULTS:** In the validation cohort, the microbiome risk score (MRS) model yielded a hazard ratio of 2.24 (95% confidence interval, 1.03-4.84; $P = .04$), using the median of the MRS from the discovery cohort as the threshold. The MRS demonstrated a temporal validity by capturing individuals that developed CD up to 5 years before disease onset (area under the curve > 0.65). The 5 most important taxa contributing to the MRS included *Ruminococcus torques*, *Blautia*, *Colidextribacter*, an uncultured genus-level group from *Oscillospiraceae*, and *Roseburia*. **CONCLUSION:** This study is the first to demonstrate that gut microbiome composition is associated with future onset of CD and suggests that gut microbiome is a contributor in the pathogenesis of CD.

Keywords: Microbiome; *Faecalibacterium*; Fecal Calprotectin; Vitamins B; Preclinical Inflammatory Bowel Disease.

Crohn's disease (CD) is one of the inflammatory bowel diseases (IBDs) characterized by chronic, relapsing inflammation of the intestine. The cause of CD is unknown, but the current hypothesis is that microbial or environmental factors induce gut inflammation in genetically susceptible individuals, leading to chronic intestinal inflammation and damage.[1,2] Case-control studies of patients with established CD have cataloged alterations in the gut microbiome composition[3–7]; however, these studies fail to distinguish whether the altered gut microbiome composition is associated with initiation of CD or is the result of inflammation or drug treatment.[3]

To address these issues, the Crohn's and Colitis Canada Genetic Environmental Microbial (GEM) project, a prospective cohort study of healthy first-degree relatives (FDRs) of individuals with CD, was designed to identify the parameters associated with the development of CD. Among these parameters, we were interested in profiling the gut microbiome composition that precedes the onset of CD and to what extent this composition predicts the risk of developing CD. Specifically, we applied a machine learning approach to the analysis of the gut microbiome composition in a large cohort of healthy FDRs of patients with CD (N = 3483) to define a microbial signature that is associated with the risk of developing CD.

WHAT YOU NEED TO KNOW

BACKGROUND AND CONTEXT

Case-control studies of established Crohn's disease fail to distinguish whether alterations of gut microbiome composition are associated with future onset of Crohn's disease or a result of the inflammation.

NEW FINDINGS

This study identifies a preclinical gut microbiome signature that is associated with future development of Crohn's disease and suggests that microbiome communities are implicated in its pathogenesis.

LIMITATIONS

In the validation cohort, the predictive performance of the microbiome risk score to predict Crohn's disease onset was modest, with a concordance index of 0.67. However, the predictive value of the microbiome was replicated using other machine learning algorithms.

CLINICAL RESEARCH RELEVANCE

This study suggests that the microbiome risk score could offer the possibility to stratify healthy at-risk individuals who would benefit from interventions aimed at modifying the microbial imbalance and possibly reducing the risk of developing Crohn's disease.

BASIC RESEARCH RELEVANCE

This study suggests that gut microbiome is a potential contributor in the pathogenesis of Crohn's disease. We found that the microbial community rather than individuals' taxa are associated with risk of Crohn's disease.

Material and Methods

Subject Recruitment

The GEM project is a prospective cohort study of healthy FDRs recruited between 2008 and 2017 (see Supplementary Methods for inclusion and exclusion criteria). The recruiting centers were from Canada, the United States of America, Israel, the United Kingdom, Ireland, and New Zealand (Supplementary Table 1). All subjects were contacted every 6 months via a phone call (Supplementary Notes 1 and 2). If a subject disclosed that they had a diagnosis of CD (as of February 28, 2020, the data freeze date for this study), this was confirmed by their treating physician based on clinical, endoscopic, radiographic, or histologic reports (Supplementary Note 3). All subjects or their guardians gave written informed consent to participate in the study. The study was approved by the Mount Sinai Hospital Research Ethics Board and the local recruitment centers.

Profiling of the Fecal Microbiota

Stool samples at enrollment were collected in a commode specimen collector and frozen before their delivery to the local study site where they were stored at −80°C. The stool DNA extraction was performed using the QIAamp DNA Stool Mini Kit (QIAGEN) and V4 hypervariable region of bacterial 16S ribosomal RNA (16S rRNA) was amplified using the 515F/806R primer pair[8] and sequenced in paired-end mode (2 × 150 base pairs). A median of 75,904 reads per sample were imported into a QIIME artifact and were denoised using the dada2 plugin[9] (see Supplementary Methods). Imputed bacterial function was generated with the use of the picrust2_pipeline.py script of the PICRUSt2[10] 2.4.1 package.

Assessment of Gut Inflammation Using Fecal Calprotectin

Fecal calprotectin concentration was measured by the BÜHLMANN fCAL ELISA test (Schöonrenbuch, Switzerland) following the manufacturer's protocol, and the average of the duplicate values was used to define the calprotectin concentration (see Supplementary Methods).

Assessment of Stool Metabolomics

For stool metabolomics, we used samples obtained at enrollment of healthy FDRs from the nested case-control subset of the cohort with an available microbiome risk score (MRS) and metabolomic measurements from the Crohn's and Colitis

*** Authors share co-first authorship.**

Abbreviations used in this paper: **abs, absolute value; CCC, Crohn's and Colitis Canada; CD, Crohn's disease; CI, confidence interval; FDR, first-degree relative; GEM, Genetic Environmental Microbial; HR, hazard ratio; IBD, inflammatory bowel disease; MRS, microbiome risk score; rRNA, ribosomal RNA; RSF, random survival forests; SD, standard deviation; TMAP, N,N,N-trimethyl-L-alanyl-L-proline betaine.**

Most current article

0016-5085/$36.00
https://doi.org/10.1053/j.gastro.2023.05.032

Canada (CCC) GEM Project (see Supplementary Methods). Stool metabolomic measurements were performed with Metabolon using the Metabolon's DiscoveryHD4 Platform, following manufacturer instructions, from the same stool sample as the 16S profile.

Construction of the Microbiome Risk Score

The gut microbiota is increasingly recognized as an ecologic niche comprising many different taxa that act as a community.[11,12] In this context, changes in the abundance of a given bacterial taxon can impact the capacity of other taxa to thrive in a given environment. To address this complexity in the microbiome, we applied a random survival forests (RSF) methodology to assess the relationship between the baseline gut microbiota community and the future risk of developing CD.[13]

Using this approach, we developed a risk score that combines the effects of relative abundances of bacterial genera, as defined by stool 16S rRNA sequencing in a nonlinear, nonparametric manner, to quantify an individual's risk of developing CD (see Supplementary Methods). The RSF model allows for combining the high-dimensional microbiome composition data and generating a score that quantifies the risk of developing CD.[13] The RSF model can efficiently handle the relatively small number of events compared with the large number of nonevents using resampling techniques.[13,14] Finally, the RSF model detects and incorporates high-order interactions of the variables, which contribute to the risk score.[13,14] Besides the relative abundances of bacterial taxa, the RSF model included the covariates age, sex, Shannon alpha diversity, and the number of sequencing reads (see Supplementary Methods).

To develop the risk score model and to ensure its predictive ability, we first randomly assigned two-thirds of the original cohort into a discovery cohort and one-third into an independent validation cohort. More precisely, the subjects in the validation cohort were never used in the model construction. The discovery cohort was used to develop the risk score using the RSF methodology (see Supplementary Methods). The generated MRS ranks the individuals according to their risk of developing CD. We then applied the MRS to the independent validation cohort to evaluate its performance. Here, we assessed the predictive ability of the MRS as a continuous variable using Cox's proportional hazards model.

Results

Microbial Composition Risk Score Is Associated With Future Risk of Developing Crohn's Disease

We analyzed 3483 healthy FDRs with baseline microbiome data (Supplementary Figures 1–5 and Supplementary Tables 1–3). The subjects were monitored for a median of 5.4 years. The median age at recruitment was 17.0 years (range, 6–35 years), and 48% were <18 years old. In this cohort, 73 individuals developed CD (pre-CD) (Supplementary Table 2), with a median time from enrollment to CD diagnosis of 3.1 years and a median age of new CD onset of 17.7 years (Table 1).

To develop the MRS and ensure its predictive ability, we first randomly assigned two-thirds of the original cohort into a discovery cohort (n = 2321) and one-third into an independent validation cohort (n = 1162) (Table 1, see

Table 1. Discovery and Validation Cohort Demographics

Variable	Discovery set (n = 2321)	Validation set (n = 1162)	*P* value
Pre-CD[a,b,c]	43	30	.16
Female sex[a]	1252 (53.9)	587 (50.5)	.06
Age at recruitment, y[d]			
Median	17.0	17.0	.88
Mean	18.1	18.1	
SD	7.8	7.6	
Country[a]			
Canada	1395 (60.1)	705 (60.7)	.76
United States	289 (12.5)	148 (12.7)	.82
Israel	354 (15.3)	175 (15.1)	.92
United Kingdom	245 (10.6)	118 (10.2)	.76
New Zealand	30 (1.3)	10 (0.9)	.31
Ireland	8 (0.3)	6 (0.5)	.57
Time in the study, y[b]			
Median	5.5	5.3	.42
Mean	3.1	5.8	
SD	2.1	3.0	

NOTE. Data are presented as number (%), unless indicated otherwise. All percentages are a function of the total number of individuals in the respective cohort. Country indicates the country of recruitment of the individuals.
[a]The *P* values were calculated using Fisher's exact test.
[b]Time in the study is defined as the difference between stool sample collection date and last follow-up date for healthy individuals and sample collection date to date of diagnosis for individuals who later developed CD.
[c]Individuals healthy at the time of recruitment who later developed CD.
[d]The *P* values were calculated using the 2-sided Kruskal-Wallis rank test.

Supplementary Methods). The MRS that was developed using RSF in the discovery cohort was subsequently applied to the validation cohort. The MRS yielded a hazard ratio (HR) of 1.58 per standard deviation (SD) of the MRS based on discovery cohort (95% confidence interval [CI], 1.14–2.18), a concordance index (C index) of 0.67, and $P = .0057$ (Supplementary Table 4).

Using the value corresponding to the median of the MRS from the discovery cohort as the threshold (see Supplementary Methods), we obtained a HR of 2.24 (95% CI, 1.03–4.84; $P = .04$) (Figure 1 and Supplementary Figure 6) in the validation cohort. The model was also compared against other machine learning methodologies, including Neural Networks and eXtreme Gradient Boosting applied to survival data (Supplementary Table 5), wherein the RSF model had the highest concordance index (C index = 0.67) of all models.

Microbial Composition Risk Score Predicts Crohn's Disease Up to 5 Years Before Disease Onset

Because the MRS was measured in healthy asymptomatic at-risk individuals who later developed CD at different times, we evaluated the relationship of the MRS with the time

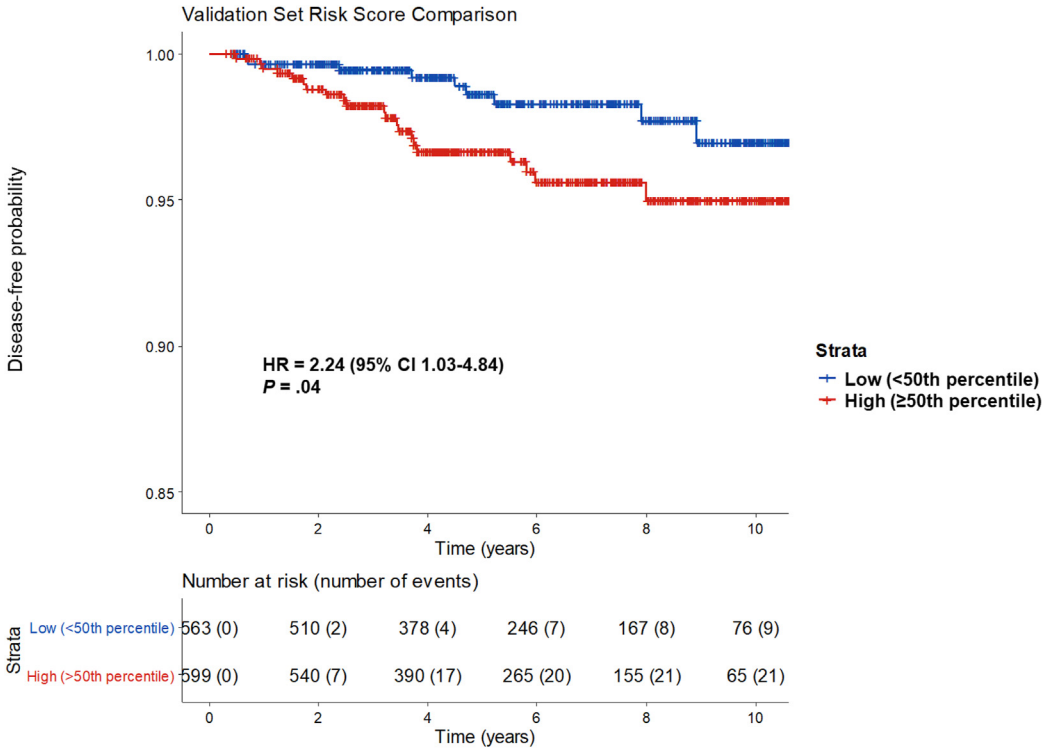


Figure 1. Kaplan-Meier plot shows the risk score performance in the validation cohort. As described, we assigned the validation cohort individuals (n = 1162) into 1 of 2 groups defined by the median of the MRS based on the discovery cohort and compared the relative survival between these 2 groups. See the Supplementary Methods for term definitions.

to CD onset. To do this, we examined the performance of the MRS in the validation cohort to predict the risk of developing CD when measured within 1.5, 3, and 5 years after the baseline stool sample was collected, compared with those who remained healthy and were monitored for the same period. We found that the predictive accuracy of the model for those subjects who developed CD within 1.5 years of having their MRS measured had an area under the curve of 0.70 (Supplementary Figure 7; see Supplementary Methods). For those subjects who developed CD within 3 and 5 years of having their MRS measured, we obtained an area under the curve of 0.71 and 0.67, respectively (Supplementary Figure 7).

Microbial Taxa Contributing to the Microbial Composition Risk Score

To assess the specific contribution of the individual taxa (in the gut microbial community) as part of the MRS to predict future onset of CD, we calculated the importance of each taxon as defined by the RSF model by using both model- and permutation-based approaches on the discovery cohort (see Supplementary Methods and Supplementary Tables 6–9). We found that the 5 most important taxa include the genera *Ruminococcus torques*, *Blautia*, *Colidextribacter*, an uncultured genus-level group from *Oscillospiraceae*, and *Roseburia* (Figure 2*A*).

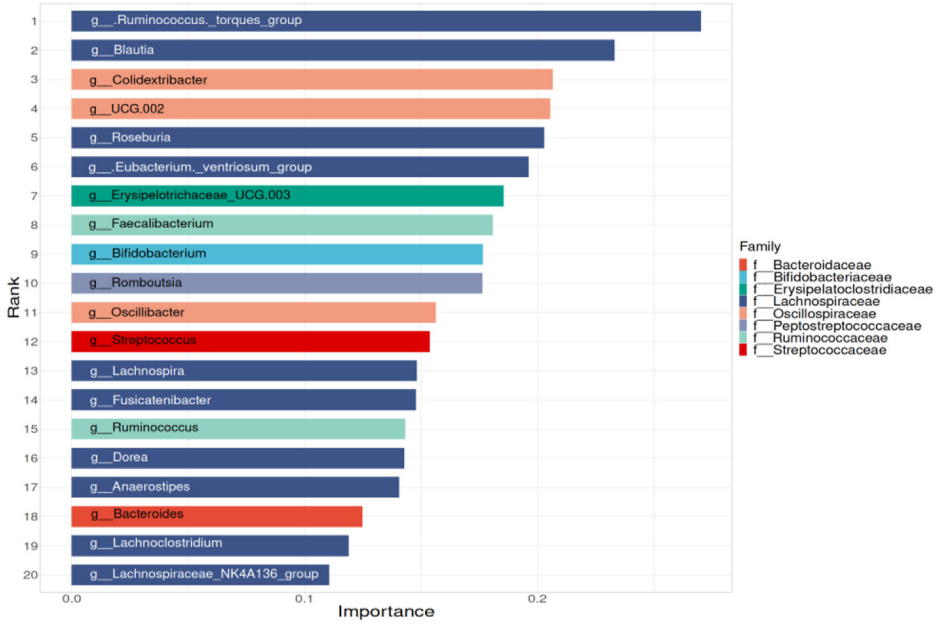
The most important taxon according to the RSF model was the genus *R torques* group. This taxon was positively associated with an increased MRS (discovery: Spearman's $\rho = 0.96$; $P < 2e^{-16}$; validation: Spearman's $\rho = 0.95$; $P < 2e^{-16}$) (see Supplementary Methods-Association between increasing MRS decile and taxa relative abundance; and Figure 2*B*). The second most important taxon was the genus *Blautia*. This taxon was also positively associated with the MRS (discovery: Spearman's $\rho = 0.98$; $P < 2e^{-16}$; validation: Spearman's $\rho = 0.98$; $P < 2e^{-16}$). Another important taxon was the *Roseburia* genus, which was negatively associated with the risk score (discovery: Spearman's $\rho = 0.91$; $P = 5e^{-4}$; validation: Spearman's $\rho = -0.61$; $P = .066$) (Figure 2*B*). Finally, we found that an increase in the abundance of the *Faecalibacterium* genus (eighth most important taxon) was inversely associated with an increase of MRS (discovery: Spearman's $\rho = -1$, $P < 2e^{-16}$; validation: Spearman's $\rho = -0.79$, $P = .009$) (see Supplementary Methods, Supplementary Table 10, and Supplementary Figure 8).

Finally, the relative abundance distribution for the top 10 taxa comparing those who remained healthy vs those who developed CD followed a similar pattern as dictated by the RSF (Figure 3, Supplementary Table 10, and Supplementary Figure 9). We found that the top 20 taxa have a large number of interactions that define the MRS via the RSF model compared with features of lower importance (Figure 2*C*; Supplementary Figure 10, and Supplementary Table 11).

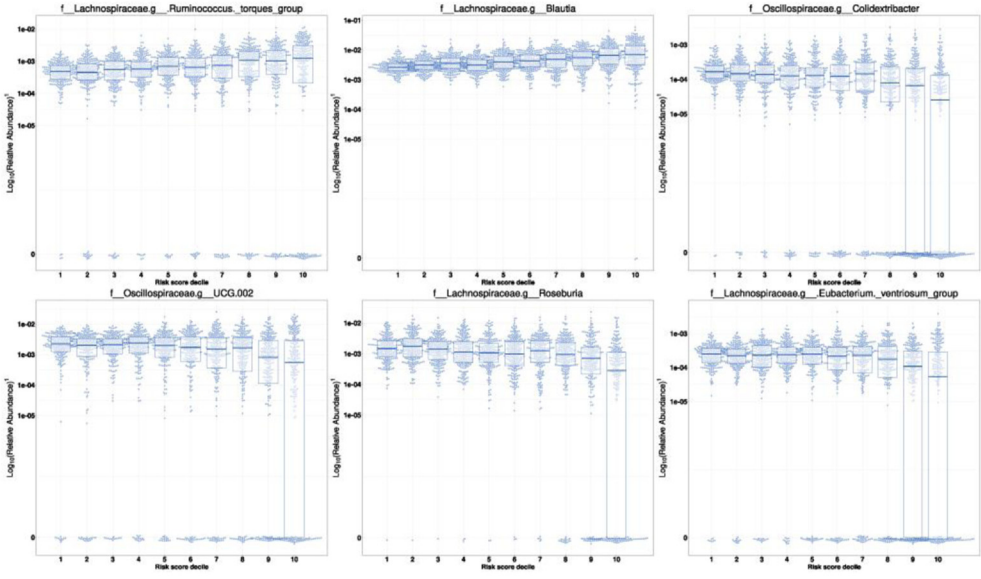
The Microbial Risk Score Comparison With the Ecologic Partitioning of the Microbiome to Predict Crohn's Disease Onset

We next assessed whether the MRS for the entire cohort was associated with a specific ecologic partitioning of the

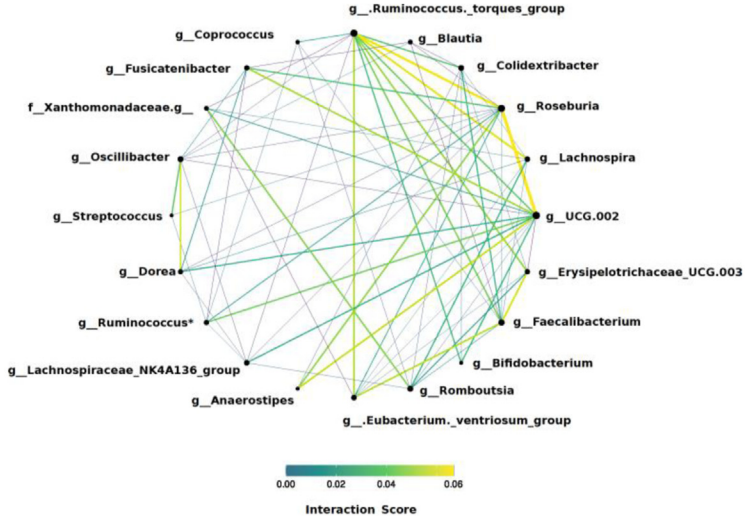
A.


g__.Ruminococcus._torques_group
g__Blautia
g__Colidextribacter
g__UCG.002
g__Roseburia
g__.Eubacterium._ventriosum_group
g__Erysipelotrichaceae_UCG.003
g__Faecalibacterium
g__Bifidobacterium
g__Romboutsia
g__Oscillibacter
g__Streptococcus
g__Lachnospira
g__Fusicatenibacter
g__Ruminococcus
g__Dorea
g__Anaerostipes
g__Bacteroides
g__Lachnoclostridium
g__Lachnospiraceae_NK4A136_group
Rank
Importance
0.0
0.1
0.2
Family
f__Bacteroidaceae
f__Bifidobacteriaceae
f__Erysipelatoclostridiaceae
f__Lachnospiraceae
f__Oscillospiraceae
f__Peptostreptococcaceae
f__Ruminococcaceae
f__Streptococcaceae

B.


f__Lachnospiraceae.g__.Ruminococcus._torques_group
f__Lachnospiraceae.g__Blautia
f__Oscillospiraceae.g__Colidextribacter
f__Oscillospiraceae.g__UCG.002
f__Lachnospiraceae.g__Roseburia
f__Lachnospiraceae.g__.Eubacterium._ventriosum_group
Log10(Relative Abundance)
Risk score decile

C.


g__.Ruminococcus._torques_group
g__Coprococcus
g__Blautia
g__Fusicatenibacter
g__Colidextribacter
f__Xanthomonadaceae.g__
g__Roseburia
g__Oscillibacter
g__Lachnospira
g__Streptococcus
g__UCG.002
g__Dorea
g__Erysipelotrichaceae_UCG.003
g__Ruminococcus*
g__Faecalibacterium
g__Lachnospiraceae_NK4A136_group
g__Bifidobacterium
g__Anaerostipes
g__Romboutsia
g__.Eubacterium._ventriosum_group
0.00
0.02
0.04
0.06
Interaction Score

microbiome, previously described as enterotype[15] (see Supplementary Methods). We found that the MRS was significantly and positively associated with the Firmicutes enterotype (using a generalized estimating equations linear model, $P = 2.0 \times 10^{-16}$) (see Supplementary Methods and Supplementary Figure 11). This finding is not surprising, because many of the important taxa in the RSF model belong to the Firmicutes phylum (Figure 3). However, the Firmicutes ecologic partition itself or any other enterotype groups were not enriched in pre-CD individuals (χ^2 $P = .1$), indicating the limitation of such classification to identify individuals at risk of developing CD.

Microbiome Risk Score and Imputed Microbiome Function

Although the risk score was built using the gut microbiome composition and provides an initial understanding of the taxa that may be involved with the development of CD, it lacks information about the specific functional processes that may potentially be associated with CD pathogenesis. In an attempt to understand which microbial functions are related to the microbial signature of developing CD, we performed microbial functional imputation of our microbiome data using PICRUSt2 (see Supplementary Methods).

After imputation, we performed a correlation analysis at the pathway level against the predicted MRS for our entire cohort (N = 3483) (Figure 4 and Supplementary Table 12). This analysis found 46 of 409 microbial functions with a correlation coefficient abs(ρ) > 0.25 and a false discovery rate-adjusted $P < .05$. The top 10 most significant correlated pathways were all negatively correlated with the MRS and included the reductive acetyl coenzyme A pathway I, mycolate biosynthesis, palmitoleate biosynthesis I, oleate biosynthesis IV, stearate biosynthesis II, superpathway of fatty acid biosynthesis initiation, (5z)-dodecenoate biosynthesis I, palmitate biosynthesis, 8-amino-7-oxononanoate biosynthesis I, and biotin biosynthesis I.

Stool Metabolomic Assessment of the Microbiome Risk Score

Although a previous study showed a high degree of concordance between PICRUSt2 imputed metagenomic data to their corresponding shotgun sequencing data,[10] PICRUSt2 may suffer from imputation bias. Moreover, PICRUSt2 and shotgun sequencing data only provide information about the genomic potential present in the community and may not necessarily correlate with proteomic or metabolomic profiles of the community.[16] For this reason and to further understand how our pre-CD gut microbiome signature may play a functional role, we explored the baseline stool metabolomics data set available from a subset of the cohort (n = 122), comprising 56 pre-CD and 66 individuals who remained healthy (see Supplementary Methods). Those who later developed CD (n = 56) were closely matched 1:1 by age, sex, follow-up duration, and geographic location with control FDRs remaining healthy (n = 66) (Supplementary Table 13). This exploratory analysis identified 24 metabolites of 1029 stool metabolites with a correlation coefficient abs(ρ) > 0.25 with a nominal significance ($P < .05$) with the MRS (Supplementary Tables 14 and 15 and Supplementary Figures 12–14). The top 10 most significant correlated metabolites were all negatively correlated with the MRS and include cytosine, N,N,N-trimethyl-L-alanyl-L-proline betaine (TMAP), cytidine, 2,3-dihydroxyisovalerate, gentisate, nicotinate, guanine, xylose, 8-hydroxyguanine, and β-alanine.

The Microbiome Risk Score Is Associated With Crohn's Disease Independent of Fecal Calprotectin

Because subclinical gut inflammation may affect the gut microbial composition, we assessed the possible relationship between a marker of gut inflammation and MRS performance. We used fecal calprotectin levels measured from the same stool sample from which we measured the 16S profile as a proxy for subclinical gut inflammation. We performed a Cox proportional hazards model using the MRS in the subset of the validation cohort consisting of individuals with microbiome composition and fecal calprotectin measured (n = 1109). We categorized individuals by the presence or absence of gut inflammation indicated by their fecal calprotectin levels, using a threshold of >120 μg/g fecal calprotectin.[17–19] In the validation set with available

Figure 2. Description of taxa associations with our risk score for developing CD. (*A*) Taxa relative importance defined by the RSF using the discovery cohort. Importance values were calculated by the mean position relative to the root of every regression tree built using the bootstrap samples from the discovery cohort. The colors of the individual *bars* correspond to the family the taxon belongs to (see Supplementary Methods and Supplementary Table 6). The *x-axis* represents the importance value of a given taxa. The *y-axis* represents the taxa contributing to the MRS. Family-level taxonomies of the given figure are color coded as indicated on the *right* of the figure. (*B*) Log_{10} relative abundance distribution of the 6 most important taxa (*x-axis*) according to the RFS model and their association with each decile of increasing risk score. (^{1}*y-axis*, log transformation is for visualization purposes). This information shows the direction of the contribution of each taxon to the risk score model generated. Spearman correlation between log(median relative abundance) and decile group: *Ruminococcus torques* group ($\rho = 0.96$; $P < 2e^{-16}$), *Blautia* ($\rho = 0.98$; $P < 2e^{-16}$), *Colidextriacter* ($\rho = -0.86$; $P = .003$), *UCG.002* ($\rho = -0.90$; $P = 9e^{-4}$), *Roseburia* ($\rho = 0.91$; $P = 5e^{-4}$), and *Eubacterium ventriosum* group ($\rho = -0.80$; $P = .008$). We can observe the nonlinear relationship between the relative abundance of each taxon and the resulting risk score (Supplementary Table 10). (*C*) Network representation of the interactions according to the RSF model of the top 20 important taxa. Interactions are defined as outlined in the Supplementary Methods. Each *node* corresponds to one of the top 20 most important taxa, and their *labels* correspond to the rank column in *A*. The size of the node is proportional to the degree of the node (ie, number of edges in each node). The presence of an edge between any 2 taxa indicates an interaction score >0. The *edge color* and *line weight* represent the value of the interaction score for a given pair of taxa. (See Supplementary Methods and Supplementary Table 11.) Range of values for interaction score of all taxa pairs is 0 to 0.067.

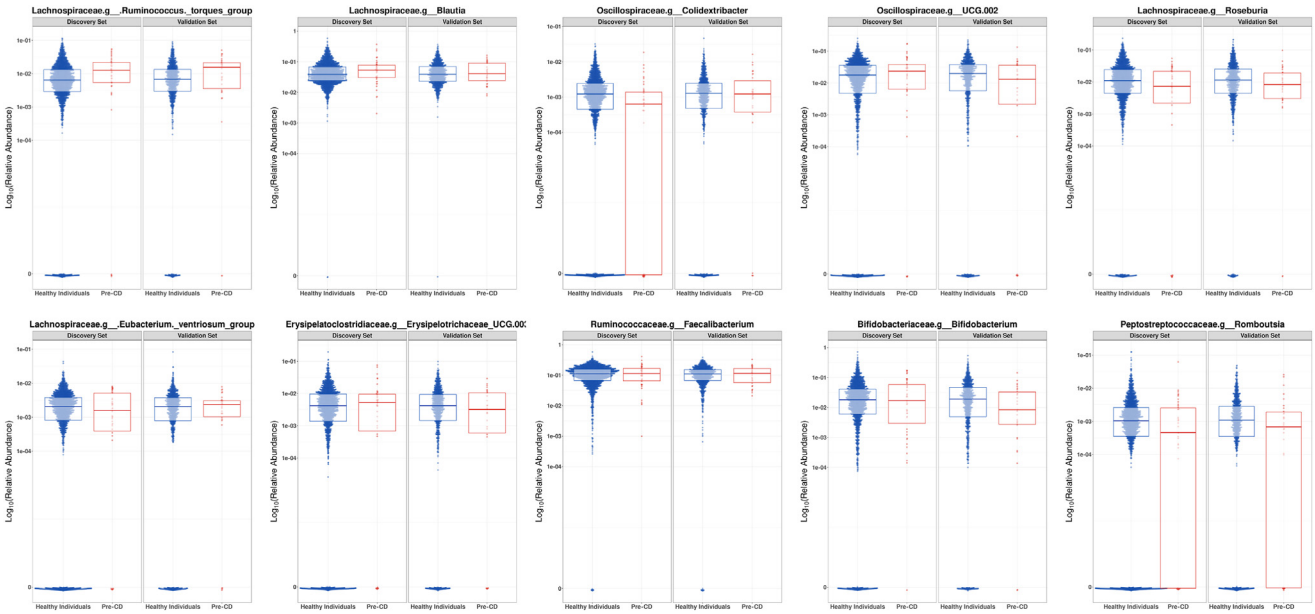


Figure 3. Relative abundance distribution of the top 10 important taxa from the RSF model in the individuals from the discovery and validation cohorts. The *dark blue dots* and *dark blue box plots* represent the relative abundance for a specific taxon for every healthy individual in the corresponding cohort. The *red dots* and *red box plots* represent the relative abundance for a specific taxon for every individual that later developed CD in the corresponding cohort. The *y-axis* represents the $\log_{10}$ transformation of the relative abundances for each individual and taxon (for visualization purposes). The distribution of the relative abundances for the most important taxa in the RSF follow a similar pattern (with exceptions) in the entire cohort divided by health status at the time of analysis. Demographic information is presented in Table 1. The figure simply indicates the consistency of direction of differences in the discovery and validation sets with the direction in which individual taxa are associated with the MRS. When assessed individually, the taxa are not statistically significantly associated with CD development (q-value > 0.05) when evaluated with a Cox proportional method. This result potentially suggests the nonlinearity and multidimensional nature of the association of the microbial composition with future development of CD.

fecal calprotectin data, a Cox's proportional model having only the continuous MRS yielded a HR of 1.51 per SD (95% CI, 1.10–2.08 per SD; $P = .011$). When we included the fecal calprotectin indicator in the model using the same data set, the continuous MRS yielded an HR of 1.42 per SD (95% CI, 1.02–1.98 per SD; $P = .041$) (Supplementary Table 16 and Supplementary Figure 15).

Notably, 93% (515 of 555) of the unaffected FDRs who had elevated baseline fecal calprotectin >120 μg/g remained asymptomatic during a mean follow-up duration of 6.0 years and up to a maximum follow-up of 11.4 years (Supplementary Figure 16). As a sensitivity analysis, we applied the MRS to the subgroup of FDRs with low baseline fecal calprotectin (<120 μg/g) in the validation cohort. The HR of the MRS for this subgroup (1.29 per SD; 95% CI, 0.72–2.31 per SD) showed consistent direction of effect compared with that of the entire validation cohort (1.51 per SD; 95% CI, 1.10–2.08 per SD), although not statistically significant (Supplementary Figure 17). Finally, an additional sensitivity analysis adjusting for dichotomized fecal calprotectin based on cutoffs of 50 or 100 μg/g showed consistent results (Supplementary Figures 18 and 19).

Discussion

Until now, most studies suggesting involvement of the microbiome in CD have been cross-sectional case-control studies of patients with established CD. However, CD activity of inflammation and related treatment in those with established disease may introduce confounding effects on the microbial composition. The use of prospective preclinical cohort studies minimizes such confounders and is a more powerful approach to characterizing the contribution of the microbiome to CD onset.[20,21] However, due to the rare incidence rate of CD, prospective studies require significant time and resources to observe a meaningful number of incident cases. To address these issues, the CCC GEM project, a prospective cohort study of healthy FDRs of individuals with CD, was designed to identify the parameters associated with the development of CD.

Using data collected from the GEM cohort, we developed an MRS model from a subset of the original cohort capable of classifying individuals that later go on to develop CD. We hypothesized that differences in the baseline gut microbiota composition between healthy individuals who later develop CD compared with those who remain healthy provide insight in the microbial determinants of CD pathogenesis.

To date, most microbiome studies have focused on defining the association of individual taxa.[22–24] The gut microbiota is increasingly recognized as an ecologic niche comprising many different taxa that act as a community.[11,12] In this context, changes in the abundance of a given bacterial taxon can impact the capacity of other taxa to thrive in a given environment. To address this complexity in the microbiome, we applied an RSF methodology to assess the

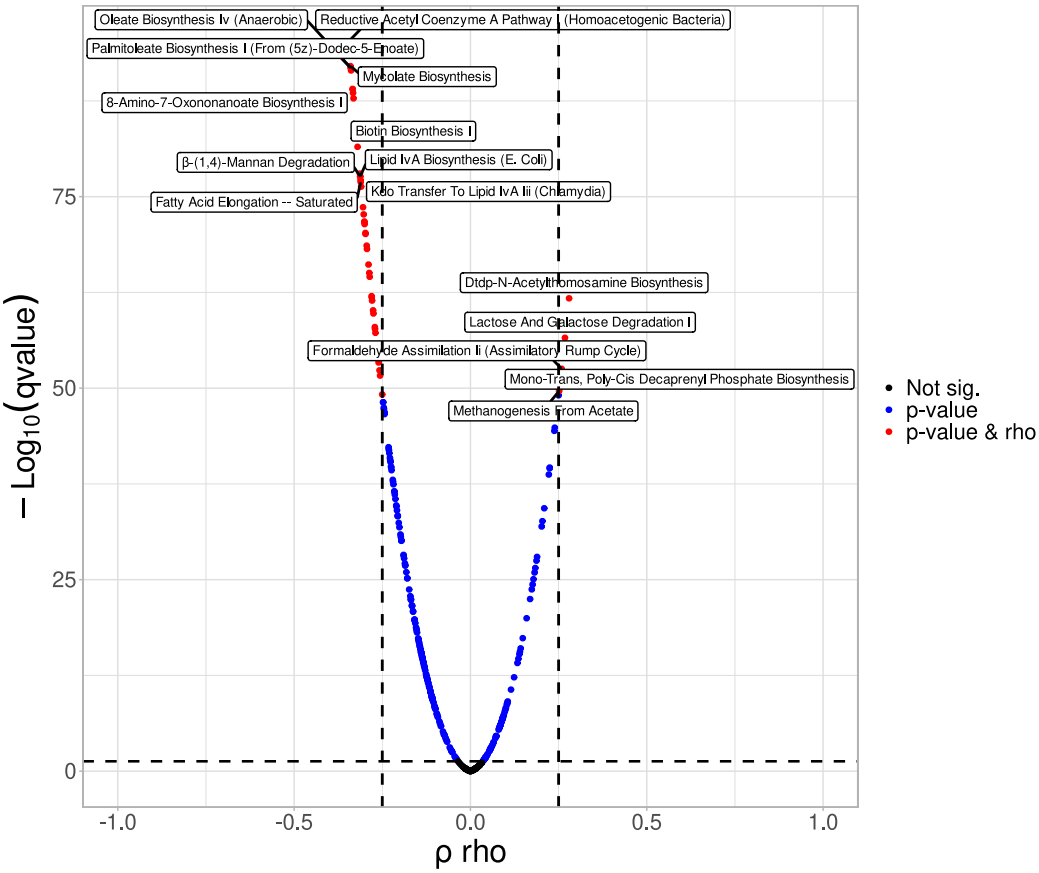


Figure 4. Results of the analysis of imputed microbial function in the entire cohort. *Volcano plot* of Spearman correlation values (ρ) vs P values of imputed microbial function using PICRUSt2 against the microbial risk score. *q*-value, false discovery rate-corrected *P* value. The *red dots* represent functions whose correlations are statistically significant (*q*-value < 0.05) and abs(ρ) > 0.25. The *blue dots* are those microbial functions whose correlation coefficients are statistically significant (*q*-value < 0.05), but whose abs(ρ) < 0.25. The *black dots* represent functions whose correlations are neither statistically significant (*q*-value > 0.05) and abs(ρ) < 0.25. We have labeled the top 10 functions with both significant *q*-value and abs(ρ) > 0.25. (See Supplementary Table 12.) The *x-axis* represents the abs(ρ) and *y-axis* the $\log_{10}$(*q*-value).

relationship between the baseline gut microbiota community and the future risk of developing CD.[13] The RSF model allows for the high dimensionality of the microbiome composition data in generating a score that quantifies the risk of developing CD.[13] After developing and validating the MRS we then identified the most important microbial features from the machine learning model in an attempt to obtain insights. These microbial features may represent nonlinear, nonparametric functions that cannot be easily interpreted with univariable analysis or traditional statistical models. It is noteworthy that when assessed individually, none of the taxa reached statistical significance (after multiple testing correction; *q* value > 0.05 with Cox's proportional model) in the discovery cohort and in the entire cohort, suggesting that a nonlinear and multidimensional nature of the microbial composition contributes to the risk of CD.

Of the taxa contributing the most to the MRS, we found that the increased abundance of *R torques* and *Blautia* was positively correlated with the MRS, suggesting that these taxa might be important contributors to CD onset. Indeed, *R torques* species are mucin degraders that have been shown to be increased in patients with established CD.[25] *R torques* appears to induce an increase in other mucin-using bacteria, perhaps aiding in compromising this protective barrier in the gut epithelium,[26] whereas *Blautia* is another mucin degrader observed to be increased in IBD and primary sclerosing cholangitis compared with healthy controls.[27] Contrarily, we found that the abundance of the *Roseburia* genus correlates negatively with the MRS, suggesting that this genus might harbor protective function against CD onset. Species from the *Roseburia* genus have been identified as one of the taxa decreased in studies of newly diagnosed patients with CD[6] and established IBD.[28] In a mouse model of colitis, this taxon was shown to increase the ratio of regulatory T cells and decrease interleukin 17.[28]

Finally, we found that an increase in the abundance of the *Faecalibacterium* genus (eighth most important taxon) was inversely associated with an increased MRS, which is consistent with previous studies reporting that *F prausnitzii* is depleted in patients with CD compared with healthy controls.[29–31] The protective effect of *F prausnitzii* on the development of inflammation is hypothesized to be due to anti-inflammatory factors found in the supernatant of

cultures of *F prausnitzii*.[30,32] Thus, this is the first study to show that the decreased abundance of *Faecalibacterium* may be a preclinical signature of CD that can be observed many years before CD onset. Whether restoring the abundance of *Faecalibacterium* during the preclinical phase could delay or prevent the development of CD remains to be seen.

Also, although this study was designed to capture CD, some individuals later developed ulcerative colitis.[33] Our results indicate that none of the 10 taxa contributing the most to the MRS were associated with ulcerative colitis onset. At this time, it is tempting to speculate that there are differences in organisms involved in the onset of either CD or ulcerative colitis onset.

We further assessed the imputed function of the microbiome by PICRUSt2 and its relation to the MRS. One of the microbial pathways negatively associated with the MRS was that of biotin biosynthesis ($\rho = -0.32$, false discovery rate–adjusted $P = 3.1 \times 10^{-84}$). Biotin (vitamin B_7 or vitamin H) is a micronutrient obtained by humans through dietary intake and is also produced by gut bacteria. Micronutrient deficiencies, including vitamin B_7 deficiency, have been documented in patients with IBD.[34] There is also experimental evidence of biotin supplementation alleviating a colitis-like phenotype in a mouse model.[34]

The second interesting function identified, which was also negatively correlated with the MRS, was β-(1,4)-mannan degradation ($\rho = -0.31$; false discovery rate–adjusted $P = 2.3 \times 10^{-78}$). β-(1,4)-Mannans are widely present in the human diet as part of hemicellulose and thickening additives. β-(1,4)-Mannans are resistant to human enzyme degradation and are processed almost entirely by the gut microbiome.[35] β-(1,4)-Mannans have emerged as popular prebiotics to combat the action of mucin degraders that may disrupt the protective mucus layers in the gut epithelium.[35] This points to an essential function whose reduction might have a strong impact despite an increased consumption of β-(1,4)-mannans through dietary changes.

We were also able to detect stool metabolomics associated with the MRS in a subset of the cohort. Notably, cytosine, which was previously shown to be decreased in CD compared with healthy controls,[36] together with its derivate cytidine, had the strongest negative correlation with the MRS ($\rho = -0.39$, $P = 1.2 \times 10^{-5}$, and $\rho = -0.35$, $P = 8.2 \times 10^{-5}$). Additionally, the MRS pre-CD signature was associated with a reduction of metabolites that have anti-inflammatory or antioxidant activity. Specifically, gentisate and nicotinate were negatively correlated with MRS ($\rho = -0.26$ to -0.31, $P = 3.2 \times 10^{-4}$–2.3×10^{-3}). Previously, nicotinate, also known as niacin or vitamin B_3, was reported to suppress inflammation and exhibit antioxidant properties, whereas gentisate, showed a broad spectrum of anti-inflammatory, antioxidant,[37] and antimicrobial properties[38,39]; both metabolites were depleted in patients with IBD compared with healthy controls.[40–43]

These findings suggest that a reduction in gut microbiome-derived anti-inflammatory metabolites may precede the development of CD. Interestingly, these protective metabolites were also positively correlated with the abundance of *Faecalibaterium* and *Lachnospira* (some of the top contributors of our pre-CD microbiome signature), which indicates a potential biological interaction between the abundance of these metabolites and the microbial composition. (Supplementary Figure 10).

In contrast, the metabolomics data showed a higher abundance of (12 or 13)-methylmyristate (A15:0 or I15:0) and (14 or 15)-methylpalmitate A17:0 or I17:0) with the MRS ($\rho = 0.27$, $P = 3.0 \times 10^{-3}$ and $\rho = 0.26$, $P = 4.3 \times 10^{-3}$). These metabolites belong to the sphingolipid class of organic compounds that can act as signaling metabolites once synthesized by bacteria to communicate with the host.[44–46] Sphingolipids were previously found to be increased in IBD patients.[47,48] Our findings suggest that the dysregulation of sphingolipid abundance may already be present before the development of CD. These risk metabolites were positively correlated with *R torques* (our top important taxon in the MRS), suggesting a microbial effect on both the metabolite abundance and disease risk (Supplementary Figure 14).

Finally, the untargeted metabolomics analysis also identified TMAP, which was negatively correlated with MRS ($\rho = -0.35$, $P = 7.9 \times 10^{-5}$). Although, the structure of TMAP was recently identified, little is known regarding its biological activity.[49] Further studies will need to explore how TMAP may be involved in CD pathogenesis. Of note, 79 metabolites were only nominally associated with CD onset ($P < .05$), likely due to the small sample size of this exploratory analysis (see Supplementary Methods and Supplementary Figure 13). Studies examining larger samples will be required to validate these results.

Notably, a proportion of healthy FDRs may already have subclinical gut inflammation at the time of recruitment. However, when we included the fecal calprotectin indicator in the model, the MRS yielded an HR of 1.42 per SD (95% CI, 1.02–1.98 per SD; $P = .041$) (Supplementary Table 16 and Supplementary Figure 15). The slight decrease in the HR suggests that the MRS is not greatly confounded by the absence or presence of subclinical gut inflammation, as reflected by elevated fecal calprotectin. Notably, when considering adjusting for continuous fecal calprotectin levels, we found evidence of a nonlinear effect of fecal calprotectin on CD risk (Supplementary Figure 20). Adjusting for both nonlinear and linear components of continuous fecal calprotectin should be further explored in future studies that are appropriately powered to account for this non-linear effect.

Another potential limitation of our study is that our findings are representative of the healthy at-risk FDR population and thus may not pertain to the general population. Moreover, although the MRS was statistically associated with future onset of CD, the predictive performance of the MRS remains relatively modest, indicating that the microbiome in combination with other factors, such as genetic risk and dietary pattern, may improve the risk stratification of CD onset among healthy FDRs.

Finally, we only have access to a single time point, which means that dynamic changes in individuals that could lead to disease may not be captured. Despite the considerable resources required, we recommend including longitudinal samples to capture the succession of events that lead to

disease onset. Experimental studies will be needed to assess whether the associations presented in this study represent a cause or effect of CD pathogenesis. Nevertheless, we believe that our study sets a framework for future prospective cohort studies to examine the relationship between microbial composition and the risk of developing chronic immune-mediated diseases such as CD.

Conclusion

This study is the first to demonstrate that gut microbiome composition is associated with future onset of CD, which suggests that the gut microbiome is a potential contributor in the pathogenesis of CD. We further demonstrated that microbiome composition can define an individual's risk to develop CD according to our prediction model. The involvement of the microbiome in disease onset was validated in an independent cohort and remained associated when adjusted for inflammation as measured by fecal calprotectin, and the derived MRS predictability persisted even when the MRS was measured >5 years before disease onset. Integration of the MRS with other biomarkers for improved risk stratification remains to be explored in future studies. Understanding the biomarkers associated with the risk of CD and the biology of these biomarkers as they relate to pathogenesis will also be necessary to develop novel strategies for disease prevention in high-risk populations and for improvement in treatment in patients with established disease.

Supplementary Material

Note: To access the supplementary material accompanying this article, visit the online version of *Gastroenterology* at www.gastrojournal.org, and at https://dx.doi.org/10.1053/j.gastro.2023.05.032.

References

1. Turpin W, Goethel A, Bedrani L, et al. Determinants of IBD heritability: genes, bugs, and more. Inflamm Bowel Dis 2018;24:1133–1148.
2. Torres J, Burisch J, Riddle M, et al. Preclinical disease and preventive strategies in IBD: perspectives, challenges and opportunities. Gut 2016;65:1061–1069.
3. Wright EK, Kamm MA, Teo SM, et al. Recent advances in characterizing the gastrointestinal microbiome in Crohn's disease: a systematic review. Inflamm Bowel Dis 2015;21:1219–1228.
4. Gevers D, Kugathasan S, Denson L, et al. The treatment-naive microbiome in new-onset Crohn's disease. Cell Host Microbe 2014;15:382–392.
5. Olaisen M, Flatberg A, Granlund AvB, et al. Bacterial mucosa-associated microbiome in inflamed and proximal noninflamed ileum of patients with Crohn's disease. Inflamm Bowel Dis 2021;27:12–24.
6. Kowalska-Duplaga K, Gosiewski T, Kapusta P, et al. Differences in the intestinal microbiome of healthy children and patients with newly diagnosed Crohn's disease. Sci Rep 2019;9:18880.
7. Pittayanon R, Lau JT, Leontiadis GI, et al. Differences in gut microbiota in patients with vs without inflammatory bowel diseases: a systematic review. Gastroenterology 2020;158:930–946.e1.
8. Caporaso JG, Lauber CL, Walters WA, et al. Ultra-high-throughput microbial community analysis on the Illumina HiSeq and MiSeq platforms. ISME J 2012;6:1621–1624.
9. Callahan BJ, McMurdie PJ, Rosen MJ, et al. DADA2: High-resolution sample inference from Illumina amplicon data. Nat Methods 2016;13:581–583.
10. Douglas GM, Maffei VJ, Zaneveld JR, et al. PICRUSt2 for prediction of metagenome functions. Nat Biotechnol 2020;38:685–688.
11. Boon E, Meehan CJ, Whidden C, et al. Interactions in the microbiome: communities of organisms and communities of genes. FEMS Microbiol Rev 2014;38:90–118.
12. Wright RJ, Gibson MI, Christie-Oleza JA. Understanding microbial community dynamics to improve optimal microbiome selection. Microbiome 2019;7:85.
13. Ishwaran H, Kogalur UB, Chen X, et al. Random survival forests for high-dimensional data. Stat Anal Data Min 2011;4:115–132.
14. Ishwaran H, Kogalur UB, Blackstone EH, et al. Random survival forests. Ann Appl Stat 2008;2:841–860.
15. **Costea PI, Hildebrand F**, Arumugam M, et al. Enterotypes in the landscape of gut microbial community composition. Nat Microbiol 2018;3:8–16.
16. Mallick H, Franzosa EA, McLver LJ, et al. Predictive metabolomic profiling of microbial communities using amplicon or metagenomic sequences. Nat Commun 2019;10:3136.
17. von Roon AC, Karamountzos L, Purkayastha S, et al. Diagnostic precision of fecal calprotectin for inflammatory bowel disease and colorectal malignancy. Am J Gastroenterol 2007;102:803–813.
18. Reenaers C, Bossuyt P, Hindryckx P, et al. Expert opinion for use of faecal calprotectin in diagnosis and monitoring of inflammatory bowel disease in daily clinical practice. United European Gastroenterol J 2018;6:1117–1125.
19. Kittanakom S, Shajib MS, Garvie K, et al. Comparison of fecal calprotectin methods for predicting relapse of pediatric inflammatory bowel disease. Can J Gastroenterol Hepatol 2017;2017:1450970.
20. Kim ES, Tarassishin L, Eisele C, et al. Longitudinal changes in fecal calprotectin levels among pregnant women with and without inflammatory bowel disease and their babies. Gastroenterology 2021;160:1118–1130.e3.
21. **Lee SH, Turpin W, Espin-Garcia O**, et al. Anti-microbial antibody response is associated with future onset of Crohn's disease independent of biomarkers of altered gut barrier function, subclinical inflammation, and genetic risk. Gastroenterology 2021;161:1540–1551.
22. Jackson MA, Verdi S, Maxan M-E, et al. Gut microbiota associations with common diseases and prescription medications in a population-based cohort. Nat Commun 2018;9:2655.

23. Lloyd-Price J, Abu-Ali G, Huttenhower C. The healthy human microbiome. Genome Med 2016;8:51.
24. Durack J, Lynch SV. The gut microbiome: Relationships with disease and opportunities for therapy. J Exp Med 2019;216:20–40.
25. Schirmer M, Garner A, Vlamakis H, et al. Microbial genes and pathways in inflammatory bowel disease. Nat Rev Microbiol 2019;17:497–511.
26. Png CW, Linden SK, Gilshenan KS, et al. Mucolytic bacteria with increased prevalence in IBD mucosa augment in vitro utilization of mucin by other bacteria. Am J Gastroenterol 2010;105:2420–2428.
27. Torres J, Bao X, Goel A, et al. The features of mucosa-associated microbiota in primary sclerosing cholangitis. Aliment Pharmacol Ther 2016;43:790–801.
28. Zhu C, Song K, Shen Z, et al. *Roseburia intestinalis* inhibits interleukin17 excretion and promotes regulatory T cells differentiation in colitis. Mol Med Rep 2018;17:7567–7574.
29. Dörffel Y, Swidsinski A, Loening-Baucke V, et al. Common biostructure of the colonic microbiota in neuroendocrine tumors and Crohn's disease and the effect of therapy. Inflamm Bowel Dis 2012;18:1663–1671.
30. Packey CD, Sartor RB. Commensal bacteria, traditional and opportunistic pathogens, dysbiosis and bacterial killing in inflammatory bowel diseases. Curr Opin Infect Dis 2009;22:292–301.
31. Cao Y, Shen J, Ran ZH. Association between *Faecalibacterium prausnitzii* reduction and inflammatory bowel disease: a meta-analysis and systematic review of the literature. Gastroenterol Res Pract 2014;2014:872725–872725.
32. Quévrain E, Maubert MA, Michon C, et al. Identification of an anti-inflammatory protein from *Faecalibacterium prausnitzii*, a commensal bacterium deficient in Crohn's disease. Gut 2016;65:415–425.
33. **Galipeau HJ, Caminero A, Turpin W, Bermudez-Brito M**, et al. Novel fecal biomarkers that precede clinical diagnosis of ulcerative colitis. Gastroenterology 2021;160:1532–1545.
34. Jayawardena D, Dudeja PK. Micronutrient deficiency in inflammatory bowel diseases: cause or effect? Cell Mol Gastroenterol Hepatol 2020;9:707–708.
35. La Rosa SL, Leth ML, Michalak L, et al. The human gut Firmicute *Roseburia intestinalis* is a primary degrader of dietary beta-mannans. Nat Commun 2019;10:905.
36. Kolho KL, Pessia A, Jaakkola T, et al. Faecal and serum metabolomics in paediatric inflammatory bowel disease. J Crohns Colitis 2017;11:321–334.
37. Mardani-Ghahfarokhi A, Farhoosh R. Antioxidant activity and mechanism of inhibitory action of gentisic and alpha-resorcylic acids. Sci Rep 2020;10:19487.
38. Vandal J, Abou-Zaid MM, Ferroni G, et al. Antimicrobial activity of natural products from the flora of Northern Ontario, Canada. Pharm Biol 2015;53:800–806.
39. **Han X, Guo J**, Gao Y, et al. Gentisic acid prevents diet-induced obesity in mice by accelerating the thermogenesis of brown adipose tissue. Food Funct 2021;12:1262–1270.
40. Lappas M, Permezel M. The anti-inflammatory and antioxidative effects of nicotinamide, a vitamin B(3) derivative, are elicited by Fox03 in human gestational tissues: implications for preterm birth. J Nutr Biochem 2011;22:1195–1201.
41. Brown BG, Zhao XQ, Chait A, et al. Simvastatin and niacin, antioxidant vitamins, or the combination for the prevention of coronary disease. N Engl J Med 2001;345:1583–1592.
42. Lloyd-Price J, Arze C, Ananthakrishnan AN, et al. Multi-omics of the gut microbial ecosystem in inflammatory bowel diseases. Nature 2019;569:655–662.
43. **Li J, Kong D, Wang Q**, et al. Niacin ameliorates ulcerative colitis via prostaglandin D2-mediated D prostanoid receptor 1 activation. EMBO Mol Med 2017;9:571–588.
44. An D, Oh SF, Olszak T, et al. Sphingolipids from a symbiotic microbe regulate homeostasis of host intestinal natural killer T cells. Cell 2014;156:123–133.
45. Johnson EL, Heaver SL, Waters JL, et al. Sphingolipids produced by gut bacteria enter host metabolic pathways impacting ceramide levels. Nat Commun 2020;11:2471.
46. Kniazeva M, Crawford QT, Seiber M, et al. Monomethyl branched-chain fatty acids play an essential role in *Caenorhabditis elegans* development. PLoS Biol 2004;2:E257.
47. **Franzosa EA, Sirota-Madi A, Avila-Pacheco J**, et al. Gut microbiome structure and metabolic activity in inflammatory bowel disease (published correction appears in Nat Microbiol 2019;4:898). Nat Microbiol 2019;4:293–305.
48. Ebrahimpour S, Shahbazi M, Khalili A, et al. Elevated levels of IL-2 and IL-21 produced by CD4+ T cells in inflammatory bowel disease. J Biol Regul Homeost Agents 2017;31:279–287.
49. Zhang Q, Ford LA, Evans AM, et al. Structure elucidation of metabolite x17299 by interpretation of mass spectrometric data. Metabolomics 2017;13:92.

Author names in bold designate shared co-first authorship.

Received March 16, 2023. Accepted May 8, 2023.

Correspondence
Address correspondence to: Kenneth Croitoru, MDCM, Zane Cohen Center for Digestive Diseases, Division of Gastroenterology & Hepatology, Department of Medicine, University of Toronto, Mount Sinai Hospital, 600 University Avenue, Room 437, Toronto, Ontario, M5G 1X5, Canada e-mail: ken.croitoru@sinaihealth.ca; and Wei Xu, PhD, Biostatistics Department, Princess Margaret Cancer Center, Dalla Lana School of Public Health, University of Toronto, 10-511, 610 University Avenue, Toronto, M5G 2M9, Canada. e-mail: wei.xu@uhnres.utoronto.ca.

Acknowledgments
The authors thank the members of the CCC GEM Global Project Office, including Kevin Ow, Heather MacAulay, and others for their contributions.

The CCC GEM Project Research Consortium includes Maria Abreu,[1] Paul Beck,[2] Charles Bernstein,[3] Kenneth Croitoru,[4,5] Levinus A. Dieleman,[6] Brian Feagan,[7] Anne Griffiths,[8] David Guttman,[9] Kevan Jacobson,[10] Gilaad Kaplan,[2] Denis O. Krause,[11†] Karen Madsen,[12] John Marshall,[13] Paul Moayyedi,[13] Mark Ropeleski,[14] Ernest Seidman,[15†] Mark Silverberg,[4] Scott Snapper,[16] Andy Stadnyk,[17] Hillary Steinhart,[4] Michael Surette,[18] Dan Turner,[19] Thomas Walters,[20] Bruce Vallance,[21] Guy Aumais,[22] Alain Bitton,[15] Maria Cino,[4] Jeff Critch,[23] Lee Denson,[24] Colette Deslandres,[25] Wael El-Matary,[26] Hans Herfarth,[27] Peter Higgins,[28] Hien Huynh,[29] Jeffrey S. Hyams,[30] David Mack,[31] Jerry McGrath,[32] Anthony Otley,[33] and Remo Panaccione[34]; from the [1]Division of Gastroenterology, Department of Medicine, University of Miami, Miller School of Medicine, Miami, Florida; [2]Department of Medicine, University of Calgary, Calgary, Alberta, Canada; [3]Inflammatory Bowel Disease Clinical and Research Center and Department of Internal Medicine, Max Rady College of Medicine, Rady Faculty of Health Sciences, University of Manitoba, Manitoba, Winnipeg, Canada; [4]Division of

Gastroenterology & Hepatology, Temerty Faculty of Medicine, University of Toronto, Toronto, Ontario, Canada; [5]Zane Cohen Center for Digestive Diseases, Mount Sinai Hospital, Toronto, Ontario, Canada; [6]Division of Gastroenterology, Department of Medicine, University of Alberta, Edmonton, Alberta, Canada; [7]Departments of Epidemiology and Biostatistics, University of Western Ontario, London, Ontario, Canada; [8]Division of Gastroenterology, The Hospital for Sick Children, Toronto, Ontario, Canada; [9]Center for the Analysis of Genome Evolution and Function, University of Toronto, Toronto, Ontario, Canada; [10]Research Institute, British Columbia Children's Hospital, Vancouver, British Columbia, Canada; [11]Department of Medical Microbiology and Infectious Diseases, University of Manitoba, Winnipeg, Manitoba, Canada; [12]Division of Gastroenterology, Department of Medicine, Faculty of Medicine and Dentistry, University of Alberta, Edmonton, Alberta, Canada; [13]Department of Medicine, Farncombe Family Digestive Health Research Institute, McMaster University, Hamilton, Ontario, Canada; [14]Gastrointestinal Diseases Research Unit, Department of Medicine, Queen's University, Kingston, Ontario, Canada; [15]Division of Gastroenterology and Hepatology, McGill University Health Center, Montreal, Quebec, Canada; [16]Division of Gastroenterology, Hepatology and Nutrition, Boston Children's Hospital and Harvard Medical School, Boston, Massachusetts; [17]Department of Microbiology and Immunology, Dalhousie University, Halifax, Canada; [18]Department of Biochemistry and Biomedical Sciences, McMaster University, Hamilton, Ontario, Canada; [19]The Juliet Keidan Institute of Pediatric Gastroenterology and Nutrition, Shaare Zedek Medical Center, The Hebrew University of Jerusalem, Jerusalem, Israel; [20]Division of Gastroenterology, Hepatology and Nutrition, The Hospital for Sick Children, Toronto, Ontario, Canada; [21]BC Children's Hospital Research Institute, University of British Columbia, Vancouver, British Columbia, Canada; [22]Biostatistics Department, Princess Margaret Cancer Center, University Health Network, Toronto, Ontario, Canada; [23]Janeway Children's Health and Rehabilitation Center and Department of Pediatrics, Memorial University, St. John's, Newfoundland, Canada; [24]Department of Pediatrics, Cincinnati Children's Hospital Medical Center, Cincinnati, Ohio; [25]Centre Hospitalier Universitaire Sainte-Justine, Montreal, Quebec, Canada; [26]Pediatric Gastroenterology, Max Rady College of Medicine, University of Manitoba, Manitoba, Winnipeg, Canada; [27]Division of Gastroenterology and Hepatology, University of North Carolina at Chapel Hill, Chapel Hill, North Carolina; [28]Division of Gastroenterology & Hepatology, University of Michigan, Ann Arbor, Michigan; [29]Department of Pediatrics, University of Alberta, Edmonton, Alberta, Canada; [30]Division of Digestive Diseases, Hepatology, and Nutrition, Connecticut Children's Medical Center, Hartford, Connecticut; [31]Division of Gastroenterology, Hepatology & Nutrition, Children's Hospital of Eastern Ontario and University of Ottawa, Ottawa, Ontario, Canada; [32]Provincial Medical Genetics Program, Eastern Health, St. John's, Newfoundland, Canada; [33]Division of Gastroenterology, Izaak Walton Killam Hospital, Dalhousie University, Halifax, Nova Scotia, Canada; and [34]Inflammatory Bowel Disease Unit, University of Calgary, Calgary, Alberta, Canada.

The CCC GEM Project recruitment site directors include Maria Abreu, Guy Aumais, Robert Baldassano, Charles Bernstein, Maria Cino, Lee Denson, Colette Deslandres, Wael El-Matary, Anne M. Griffiths, Charlotte Hedin, Hans Herfarth, Peter Higgins, Seamus Hussey, Hien Huynh, Kevan Jacobson, David Keljo, David Kevans, Charlie Lees, David Mack, John Marshall, Jerry McGrath, Sanjay Murthy, Anthony Otley, Remo Panaccione, Nimisha Parekh, Sophie Plamondon, Graham Radford-Smith, Mark Ropeleski, Joel Rosh, David Rubin, Michael Schultz, Ernest Seidman,[†] Corey Siegel, Scott Snapper, Hillary Steinhart, and Dan Turner.

[†] Deceased.

CRediT Authorship Contributions

Juan Antonio Raygoza Garay, PhD (Conceptualization: Lead; Formal analysis: Lead; Methodology: Lead; Visualization: Lead; Writing – original draft: Lead; Writing – review & editing: Lead).

Williams Turpin, PhD (Conceptualization: Lead; Formal analysis: Equal; Investigation: Equal; Methodology: Equal; Supervision: Equal; Visualization: Supporting; Writing – original draft: Equal; Writing – review & editing: Equal).

Sun-Ho Lee, MD, PhD (Conceptualization: Equal; Formal analysis: Equal; Investigation: Equal; Methodology: Lead; Validation: Equal; Visualization: Equal; Writing – original draft: Equal; Writing – review & editing: Equal).

Michelle I. Smith, PhD (Project administration: Supporting; Resources: Equal).

Ashleigh Goethel, PhD (Project administration: Supporting; Writing – original draft: Supporting).

Anne M. Griffiths, MD (Conceptualization: Lead; Methodology: Equal; Resources: Lead; Writing – review & editing: Supporting).

Paul Moayyedi, MB, PhD (Conceptualization: Lead; Investigation: Equal; Resources: Equal; Writing – review & editing: Lead).

Osvaldo Espin-Garcia, PhD (Formal analysis: Supporting; Methodology: Supporting; Writing – review & editing: Equal).

Maria Abreu, MD (Resources: Equal; Writing – review & editing: Supporting).

Guy L. Aumais, MD (Conceptualization: Supporting; Resources: Supporting).

Charles N. Bernstein, MD (Conceptualization: Supporting; Resources: Supporting).

Irit A. Biron, MD (Resources: Supporting).

Maria Cino, MD (Resources: Supporting).

Colette Deslandres, MD (Resources: Supporting).

Iris Dotan, MD (Resources: Supporting).

Wael El-Matary, MD (Conceptualization: Supporting; Resources: Supporting).

Brian Feagan, MD (Conceptualization: Equal).

David S. Guttman, PhD (Conceptualization: Equal; Methodology: Supporting).

Hien Huynh, MD (Conceptualization: Equal).

Levinus A. Dieleman, MD, PhD (Conceptualization: Equal).

Jeffrey S. Hyams, MD (Conceptualization: Equal).

Kevan Jacobson, MD (Conceptualization: Equal; Resources: Equal).

David R. Mack, MD (Conceptualization: Equal; Resources: Equal; Writing – review & editing: Supporting).

John K. Marshall, MD (Conceptualization: Equal; Resources: Equal; Writing – review & editing: Supporting).

Anthony Otley, MD (Conceptualization: Equal; Resources: Equal; Writing – review & editing: Supporting).

Remo Panaccione, MD (Conceptualization: Equal; Resources: Equal; Writing – review & editing: Supporting).

Mark Ropeleski, MD (Conceptualization: Equal; Resources: Equal).

Mark S. Silverberg, MD, PhD (Conceptualization: Equal).

A. Hillary Steinhart, MD (Conceptualization: Equal; Resources: Equal).

Dan Turner, MD (Conceptualization: Equal; Resources: Equal; Writing – review & editing: Supporting).

Baruch Yerushalmi, MD (Resources: Equal).

Andrew D. Paterson, MD (Conceptualization: Supporting; Methodology: Lead; Supervision: Equal; Writing – original draft: Supporting).

Wei Xu, PhD (Conceptualization: Lead; Methodology: Lead; Supervision: Lead; Writing – review & editing: Supporting).

Ken Croitoru, MDCM (Funding acquisition: Lead; Project administration: Lead; Resources: Lead; Supervision: Lead; Writing – original draft: Lead; Writing – review & editing: Lead).

Conflicts of interest

The authors disclose no conflicts.

Funding

This study was supported by grants from Crohn's and Colitis Canada Grant #CCC-GEMIII, Canadian Institutes of Health Research (CIHR) Grant #CMF108031, and the Leona M. and Harry B. Helmsley Charitable Trust. Williams Turpin is a former recipient of a Postdoctoral Fellowship Research Award from the CIHR Fellowship/Canadian Association of Gastroenterology (CAG)/Ferring Pharmaceuticals Inc. Sun Ho Lee is a former recipient of the Imagine/CIHR/CAG Fellowship Award. Williams Turpin, Sun-Ho Lee, and Juan Antonio Raygoza Garay are recipients of fellowships from the Department of Medicine, Mount Sinai Hospital, Toronto. Kenneth Croitoru is recipient of a Canada Research Chair in Inflammatory Bowel Diseases.

Data Availability

Requests for raw and analyzed data should follow the instructions given at http://www.gemproject.ca/data-access/. The raw microbiome data are publicly accessible under Accession: PRJNA685746. All submissions will be reviewed by the Crohn's and Colitis Canada Genetic Environmental Microbial (GEM) Project Operating Committee to ensure that the requested samples/data will not interfere in any way with the intended GEM Project analysis of the nested cohort as per the original GEM Project Study Design and is not a duplication of analysis already ongoing. Those proposals meeting this evaluation will be distributed to all members of the GEM Project Steering Committee (GPSC) for review and open discussion. This review will focus on the global scientific merit of the proposal. This review will assess the basic scientific merit and the availability of requested samples and data, ensuring there is no compromise of the original intent of the GEM project. It would be of value to contact a member of the GPSC who could help sponsor your application. Those projects achieving majority vote of approval at the GPSC will be informed that the GEM Project will provide a letter of support stating that the requested samples or data will be made available to the applicants once the applicant receives funding from a granting agency that applies an independent peer review process to the proposal. The criteria to be used for review of all submissions will include the "scientific relevance" of the proposal and the judged availability of biological material requested. The budget to be requested from a funding agency must allow for any expenses in processing samples or in setting up the appropriate queries of the database. The intent is to allow sufficient time for applicants to consider submission for funding opportunities. Notably, the source code files for the microbiome based risk scores can be found in the following link: https://github.com/raygozag/GEM-microbiome-gastroenterology-paper.