RESEARCH ARTICLE

REVISED Mucosal microbiome is predictive of pediatric Crohn's disease across geographic regions in North America [version 2; peer review: 2 approved]

Rajesh Shah[1], Kristi Hoffman[2], Lee Denson[3], Subramaniam Kugathasan[4], Richard Kellermayer[2]

[1]Suite 200, Baylor Health Care System, Austin, Texas, 78735, USA
[2]Baylor College of Medicine, Houston, USA
[3]Cincinnati Children's Hospital Medical Center, Cincinnati, USA
[4]Emory University, Atlanta, USA

v2 **First published:** 08 Feb 2022, **11**:156
https://doi.org/10.12688/f1000research.108810.1

Latest published: 04 Jan 2023, **11**:156
https://doi.org/10.12688/f1000research.108810.2

Abstract

Background: Patients with Crohn's disease (CD) have an altered intestinal microbiome, which may facilitate novel diagnostic testing. However, accuracy of microbiome classification models across geographic regions may be limited. Therefore, we sought to examine geographic variation in the microbiome of patients with CD from North America and test the performance of a machine learning classification model across geographic regions.
Methods: The RISK cohort included 447 pediatric patients with CD and 221 non-inflammatory bowel disease controls from across North America. Terminal ileum, rectal and fecal samples were obtained prior to treatment for microbiome analysis. We divided study sites into 3 geographic regions to examine regional microbiome differences. We trained and tested the performance of a machine learning classification model across these regions.
Results: No differences were seen in the mucosal microbiome of patients with CD across regions or in either the fecal or mucosal microbiomes of controls. Machine learning classification algorithms for patients with CD performed well across regions (area under the receiver operating characteristic curve [AUROC] range of 0.85-0.91) with the best results from terminal ileum.
Conclusions: This study demonstrated the feasibility of microbiome based diagnostic testing in pediatric patients with CD within North America, independently from regional influences.

Keywords
Crohn's disease, microbiome, inflammatory bowel disease, machine learning

Open Peer Review

Approval Status ✓ ✓

	1	2
version 2 (revision) 04 Jan 2023	✓ view	✓ view
version 1 08 Feb 2022	✓ view	? view

1. **Jonathan Braun**, Cedars Sinai Medical Center, Los Angeles, USA
2. **Ranko Gacesa**, University Medical Centre Groningen, University of Groningen, Groningen, The Netherlands

Any reports and responses or comments on the article can be found at the end of the article.

Corresponding author: Rajesh Shah (rajesh.shah@bswhealth.org)

Author roles: Shah R: Conceptualization, Data Curation, Formal Analysis, Methodology, Project Administration, Supervision, Validation, Visualization, Writing – Original Draft Preparation, Writing – Review & Editing; **Hoffman K**: Conceptualization, Formal Analysis, Methodology, Project Administration, Writing – Original Draft Preparation, Writing – Review & Editing; **Denson L**: Writing – Original Draft Preparation, Writing – Review & Editing; **Kugathasan S**: Writing – Original Draft Preparation, Writing – Review & Editing; **Kellermayer R**: Conceptualization, Data Curation, Formal Analysis, Methodology, Project Administration, Supervision, Validation, Visualization, Writing – Original Draft Preparation, Writing – Review & Editing

Competing interests: No competing interests were disclosed.

Grant information: RK was supported by philanthropic donations through the Gutsy Kids Fund led by the Brock Wagner and Klaasmeyer families. The RISK consortium was supported by the Crohn's and Colitis Foundation.

How to cite this article: Shah R, Hoffman K, Denson L *et al.* **Mucosal microbiome is predictive of pediatric Crohn's disease across geographic regions in North America [version 2; peer review: 2 approved]** F1000Research 2023, **11**:156 https://doi.org/10.12688/f1000research.108810.2

First published: 08 Feb 2022, **11**:156 https://doi.org/10.12688/f1000research.108810.1

REVISED **Amendments from Version 1**

A supplemental file has been added showing the underlying code used to generate the random forest models.

Any further responses from the reviewers can be found at the end of the article

Introduction

Currently, our understanding is the intestinal microbiome plays a role in the pathogenesis of inflammatory bowel disease (IBD),[1] and specifically Crohn's disease (CD).[2] The RISK consortium found significant differences in the taxonomy of the mucosal and fecal microbiomes of pediatric, treatment naïve patients with CD compared to non-IBD controls.[3] Similar results were demonstrated in a longitudinal study of adult patients with IBD with an emphasis on disruption of the microbiome during periods of disease activity.[4]

Based on an altered microbiome composition in patients with CD, microbiome signatures may be utilized as a diagnostic biomarker. From the original RISK publication,[3] the addition of microbiome data to clinical information improved the performance of their classification models for CD. Similarly, Pascal *et al*. showed microbiome classification models for CD were accurate and performed well across 4 countries in Europe (Spain, Belgium, the UK and Germany).[2]

Recent data suggested that geographic bias, however, may limit the validity of microbiome based diagnostic models. He *et al*. studied 7,009 individuals from 1 Chinese province with 14 districts to determine regional differences in the microbiome.[5] They found strong associations between microbiome composition and host district, which translated into decreased model performance when classifying metabolic diseases across districts. However, they acknowledged that other diseases, such as CD, could not be studied due to a limited sample size. Therefore, we sought to examine differences in the intestinal microbiome of pediatric patients with CD by region and to determine if geographic bias hinders the performance of a machine learning classification model across regions in North America.

Methods

Participants

A *post hoc* analysis of the RISK cohort was performed. The RISK cohort was a multicenter study that enrolled treatment naïve pediatric patients aged 3 to 17 years with CD and non-IBD controls from 28 sites with the United States and Canada from 2008 to 2012.[3] All patients had symptoms suggestive of CD, including abdominal pain or diarrhea that prompted evaluation with a colonoscopy with biopsies from the terminal ileum and rectum. A subset of patients also provided fecal samples. Patients were either diagnosed with CD, based on endoscopic appearance and histology, or a non-inflammatory etiology for their symptoms, which served as the non-IBD controls. Full inclusion and exclusion criteria for the RISK cohort have been described in the original publication.[3] In total, 447 patients with CD and 221 non-IBD controls were included in the original publication and they provided a total of 1,321 samples, including 630 ileal, 387 rectal and 304 fecal samples.

Ethical considerations

IRB approval was not required for this study, as deidentified data was used and consent was previously obtained from participants when they enrolled in the RISK cohort study.

Statistical analysis

Age at diagnosis, sex, race, disease phenotype, and treatment center were examined. To evaluate the influence of geography on microbiome composition, we grouped the treatment centers into 3 subjective regions based on overall geography (North-East, South-East and West, Figure 1A).

16Sv4 rRNA gene analysis was performed in the original cohort study using the Illumina MiSeq platform. For our analysis, the original biom table was obtained and rarefied to 3,441 sequences per sample. This rarefaction depth was chosen to retain the maximum number of samples and preserve the most amount of sequencing data per sample. The alpha and beta diversity and taxonomic composition of the terminal ileum, rectum, and fecal microbiomes were evaluated using the ATIMA interface version 1.0 available through the Baylor College of Medicine Alkek Center for Metagenomics and Microbiome Research. ATIMA is a graphic user interface that allows users to provide a biom table and mapping file for microbiome analysis. To adjust for potential confounding, MaasLin was used to control for variations in age at diagnosis, sex, race, sample type and geographic region.[6]

Finally, we sought to develop a machine learning model to evaluate the accuracy of a microbiome model to identify patients with CD across different regions. A random forest machine learning model was trained on patients from the North-East and tested in the South-East and West using the R package healthcare.ai version 2.5.0 with the default settings.

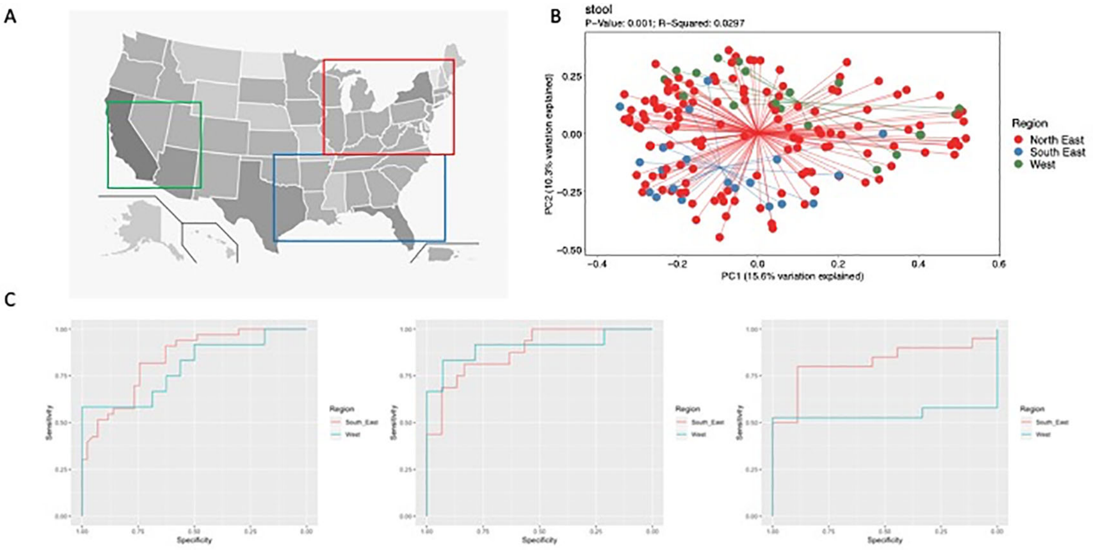

Figure 1. **A) Map of the United States with squares indicating the arbitrarily defined North East (Red), South East (Blue) and West (Green) regions, B) Principal Component Analysis of weighted Bray Curtis metric for feces from patients with CD by region (PERMANOVA), C) Receiver operating characteristic (ROC) curves for a CD prediction model trained using North East data and tested on South East and West data for ileum, rectum and feces, from left to right.**

The healthcare.ai package is an open-source R package that allows for data cleaning, manipulation, imputation, tuning of models and evaluation of model performance. Visualization of model performance with AUROC metrics was done using the R package pROC version 1.18.0.

Results

Based on the terminal ileum biopsies retained after rarefaction, we included 227 patients with CD and 165 controls with a mean age of 12.2 and 12.1 years, respectively. Approximately half of patients with CD and controls were male (58.6% and 53%, respectively) and a larger proportion of patients with CD were Caucasian compared to controls (78.9% and 68.7%, respectively). Since microbiome composition can be influenced by the presence of stricturing/fistulizing disease[7] and these patients present less of a diagnostic challenge, they were excluded from our analysis to create a consistent population with an inflammatory phenotype. After separating into regions, 182 patients were in the North-East, 33 in the South-East and 12 in the West with CD, and 106 patients in the North-East, 43 in the South-East and 16 in the West without IBD.

For patients with CD, no significant differences were found in alpha and beta diversity of the ileal and rectal mucosal microbiome by geography. However, PCoA plots of unweighted and weighted beta diversity (Figure 1B) determined through the Bray Curtis metric revealed significant differences in fecal samples. In controls, no significant differences were found in alpha and beta diversity of the ileum, rectum or fecal samples. In the South-East, patients with CD had a relative increase in Fusobacteria and Bacteroidetes with a decrease in Actinobacteria and Firmicutes in fecal samples compared to the other 2 regions. This corresponded to an increase in the genera *Bacteroides* and *Fusobacterium* with a decrease in *Bifidobacterium* and *Lactobacillus.* However, after adjustment with MaasLin, *Erwinia* was the only genus associated with geographical variation in patients with CD. Specifically, fecal samples from CD patients in the South-East had increased abundance of *Erwinia* compared to other geographic regions in North America (q=0.04).

Random forest models across sample types performed well (Figure 1C, Supplement 1[13]). The best performance occurred with ileal samples (North-East AUROC 0.89, South-East AUROC 0.85 and West AUROC 0.91). The rectal (North-East AUROC 0.87, South-East AUROC 0.83, West AUROC 0.76) and fecal (North-East AUROC 0.82, South-East AUROC 0.85, West AUROC 0.74) samples performed well, but experienced decreased performance in the West. Comparing the models, those for ileum and rectum shared OTUs discriminating CD, which included members of the Lachnospiraceae and Clostridiaceae families and the genus *Blautia.* Intriguingly, ileal biopsies and fecal samples shared top CD-discriminating OTUs from the Erysipelotrichaceae family and *Haemophilus* genus, which were not present between rectal biopsies and fecal samples.

Discussion

Our results indicate that CD influences mucosal microbiome composition to a greater extent than geography in pediatric patients from North America. Machine learning classification models performed well across the regions, despite minor differences in the fecal microbiome of CD patients. Differences in microbiome composition are known to vary across populations in healthy cohorts[8,9] and in patients with metabolic syndrome.[5] Yatsunenko *et al*. showed Westernization may influence fecal microbiome composition by comparing samples from subjects in the US, Venezuela and Malawi.[8] Similar patterns were seen by Pasolli *et al*. when they examined metagenomes from 9,428 samples from 32 countries and noted significant differences in the metagenomes of Western populations.[9] Together, these studies demonstrated microbiome composition varies across populations, however, they did not address microbiome differences within countries. To that end, He *et al*. studied a single province in China and noted differences in microbiome composition between its districts.[5] This suggested, as has been previously reviewed, that a vast number of environmental factors may play a role in shaping the microbiome and may limit the accuracy of microbiome classification models.[10]

Overall, our classification models performed well across regions and is consistent with prior reports. Using 2,045 fecal samples taken from patients with IBD and non-IBD controls across 4 European countries, Pascal *et al*. showed that a microbial signature could be used to discriminate patients with CD from non-IBD controls with an overall sensitivity of 80% and specificity of 94%.[2] In a separate cohort, Franzosa *et al*. used metagenomics and metabolomics to distinguish IBD patients from non-IBD patients also with high accuracy.[11] Our findings in pediatric CD are consistent with these results and demonstrate the feasibility of using microbiome classification models to accurately diagnose CD without geographic bias within North America.

Despite the limitations of our study, our classification models performed well. We were unable to adjust for additional confounders of microbiome composition, such as diet and supplement intake.[12] However, even without this information, our models based on ileal biopsies performed well. Additionally, we noted a decrease in model performance for fecal samples and in the West, but this may be linked to a smaller sample size, which is known to hinder the performance of machine learning models. Further work with larger cohorts and different control groups will be needed to fully determine whether microbiome machine learning models can support the diagnosis of CD in children without geographical bias, and if non-invasive testing with fecal samples is feasible.

In summary, machine learning models can distinguish patients with CD from non-IBD controls without geographic bias in North America. Further development of microbiome machine learning models to diagnose CD may be warranted.

Data availability

Underlying data

NCBI BioProject: human gut metagenome. Accession number PRJNA237362; https://identifiers.org/NCBI/bioproject:PRJNA237362.

The underlying clinical data used for this study is available through the RISK consortium. Consortium approval was required to access de-identified patient data and requests can be placed through the Crohn's and Colitis Foundation IBD Plexus Initiative (www.crohnscolitisfoundation.org/research/granst-fellowships/ibd-plexus).

Extended data

Figshare: Supplement 1. Random Forest Models. https://doi.org/10.6084/m9.figshare.21727862.v1.[13]

This project contains the following extended data:

Supplement 1 includes the relevant code, including used packages, inputs and outputs used to generate the random forest models. It also includes subsequent testing of the models and their outputs.

Data are available under the terms of the Creative Commons Attribution 4.0 International license (CC-BY 4.0)

References

1. Huttenhower C, Kostic AD, Xavier RJ: **Inflammatory Bowel Disease as a Model for Translating the Microbiome.** *Immunity.* 2014; **40**(6): 843–854.
PubMed Abstract | Publisher Full Text

2. Pascal V, Pozuelo M, Borruel N, *et al.*: **A microbial signature for Crohn's disease.** *Gut.* 2017; **66**(5): 813–822.
PubMed Abstract | Publisher Full Text

3. Gevers D, Kugathasan S, Denson LA, *et al.*: **The Treatment-Naive Microbiome in New-Onset Crohn's Disease.** *Cell Host Microbe.* 2014; **15**(3): 382–392.
PubMed Abstract | Publisher Full Text

4. Lloyd-Price J, Arze C, Ananthakrishnan AN, *et al.*: **Multi-omics of the gut microbial ecosystem in inflammatory bowel diseases.** *Nature.* 2019; **569**(7758): 655–662.
PubMed Abstract | Publisher Full Text

5. He Y, Wu W, Zheng H-M, *et al.*: **Regional variation limits applications of healthy gut microbiome reference ranges and disease models.** *Nat. Med.* 2018; **24**(10): 1532–1535.
PubMed Abstract | Publisher Full Text

6. Morgan XC, Tickle TL, Sokol H, *et al.*: **Dysfunction of the intestinal microbiome in inflammatory bowel disease and treatment.** *Genome Biol.* 2012; **13**(9): R79.
PubMed Abstract | Publisher Full Text

7. Kugathasan S, Denson LA, Walters TD, *et al.*: **Prediction of complicated disease course for children newly diagnosed with Crohn's disease: a multicentre inception cohort study.** *Lancet.* 2017; **389**: 1710–1718.
PubMed Abstract | Publisher Full Text

8. Yatsunenko T, Rey FE, Manary MJ, *et al.*: **Human gut microbiome viewed across age and geography.** *Nature.* 2012; **486**(7402): 222–227.
PubMed Abstract | Publisher Full Text

9. Pasolli E, Asnicar F, Manara S, *et al.*: **Extensive Unexplored Human Microbiome Diversity Revealed by Over 150,000 Genomes from Metagenomes Spanning Age, Geography, and Lifestyle.** *Cell.* 2019; **176**: 649–662.e20.
PubMed Abstract | Publisher Full Text

10. Gupta VK, Paul S, Dutta C: **Geography, Ethnicity or Subsistence-Specific Variations in Human Microbiome Composition and Diversity.** *Front. Microbiol.* 2017; **8**: 1162.
PubMed Abstract | Publisher Full Text

11. Franzosa EA, Sirota-Madi A, Avila-Pacheco J, *et al.*: **Gut microbiome structure and metabolic activity in inflammatory bowel disease.** *Nat. Microbiol.* 2019; **4**(2): 293–305.
PubMed Abstract | Publisher Full Text

12. Lewis JD, Chen EZ, Baldassano RN, *et al.*: **Inflammation, Antibiotics, and Diet as Environmental Stressors of the Gut Microbiome in Pediatric Crohn's Disease.** *Cell Host Microbe.* 2015; **18**(4): 489–500.
PubMed Abstract | Publisher Full Text

13. Rajesh S: Supplement 1. Random Forest Models. figshare. Dataset. 2022.
Publisher Full Text

Open Peer Review

Current Peer Review Status: ✓ ✓

Version 2

Reviewer Report 20 January 2023

https://doi.org/10.5256/f1000research.142172.r158956

✓

Ranko Gacesa
Department of Gastroenterology and Hepatology, University Medical Centre Groningen, University of Groningen, Groningen, The Netherlands

Authors have included supplementary material with codes used to train machine learning classifiers. I have no further comments.

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: Bioinformatics, Microbiome, Artificial intelligence, Inflammatory Bowel Disease

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.

Reviewer Report 05 January 2023

https://doi.org/10.5256/f1000research.142172.r158957

✓

Jonathan Braun
F. Widjaja IBD Institute, Cedars Sinai Medical Center, Los Angeles, CA, 90095-1732, USA

The revision (RF model) is helpful. Good addition.

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: Immunology and microbiome in IBD

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.

Version 1

Reviewer Report 14 October 2022

https://doi.org/10.5256/f1000research.120239.r150775

?

Ranko Gacesa
Department of Gastroenterology and Hepatology, University Medical Centre Groningen, University of Groningen, Groningen, The Netherlands

Authors examined consistency of microbiome-trained machine learning models for prediction of Crohn's disease across US geography. They demonstrated that models trained on one of geographic regions perform well on ileal samples, but perform less consistently for fecal and rectal samples.

Study is relevant (as the loss of predictive power of microbiome models due to geographic variation was demonstrated before in the Chinese population) and methodologically sound. The study would, however, benefit from sharing of used codes, produced models and other technical details to facilitate the reproducibility and comparison with other, related, studies.

Is the work clearly and accurately presented and does it cite the current literature?
Yes

Is the study design appropriate and is the work technically sound?
Yes

Are sufficient details of methods and analysis provided to allow replication by others?
Partly

If applicable, is the statistical analysis and its interpretation appropriate?
Yes

Are all the source data underlying the results available to ensure full reproducibility?
Partly

Are the conclusions drawn adequately supported by the results?

Yes

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: Bioinformatics, Microbiome, Artificial intelligence, Inflammatory Bowel Disease

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard, however I have significant reservations, as outlined above.

Author Response 08 Nov 2022

Rajesh Shah, Baylor Health Care System, Austin, USA

We are happy to provide the underlying code that was used to generate the data for this study in a supplement to facilitate reproducibility.

Competing Interests: None

Reviewer Report 05 October 2022

https://doi.org/10.5256/f1000research.120239.r150774

✓ **Jonathan Braun**
F. Widjaja IBD Institute, Cedars Sinai Medical Center, Los Angeles, CA, 90095-1732, USA

Regional contribution to microbiome composition is an important and less-assessed factor in refining conclusions about disease association and microbiome. This study is useful in addressing this question in the context of US regions and the landmark RISK cohort study of Crohn's disease.

My issues with the paper are minor. It references several assessments with negative associations, for which data is not shown. It is fine to do so, but the methods should be briefly noted: what was the alpha diversity metric(s); and, is the unweighted beta diversity (noted in context of Figure 1, which is weighted) a data-not-shown finding, or a clean-up needed in the text?

Figure 1B has "lines" connecting dots in the PCoA plot. What do these lines represent?

Is the work clearly and accurately presented and does it cite the current literature?
Yes

Is the study design appropriate and is the work technically sound?

Yes

Are sufficient details of methods and analysis provided to allow replication by others?
No

If applicable, is the statistical analysis and its interpretation appropriate?
Yes

Are all the source data underlying the results available to ensure full reproducibility?
Yes

Are the conclusions drawn adequately supported by the results?
Yes

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: Immunology and microbiome in IBD

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.

Author Response 11 Oct 2022

Rajesh Shah, Baylor Health Care System, Austin, USA

We greatly appreciate the comments and favorable review. To address the raised points, we limited the presentation of negative findings to allow us to focus on positive findings. In terms of the metrics used, we used the Shannon metric to measure alpha diversity and the Bray-Curtis metric for beta diversity. The lines presented in Figure 1B refer to the centroid of each Region.

Competing Interests: None