Gut microbiota signatures in inflammatory bowel disease

Vestergaard, Marie Vibeke; Allin, Kristine H.; Eriksen, Carsten; Zakerska-Banaszak, Oliwia; Arasaradnam, Ramesh P.; Alam, Mohammad T.; Kristiansen, Karsten; Brix, Susanne; Jess, Tine

Published in:
United European Gastroenterology Journal

DOI:
10.1002/ueg2.12485

Publication date:
2024

Document version
Publisher's PDF, also known as Version of record

Document license:
CC BY-NC-ND

Citation for published version (APA):
Gut microbiota signatures in inflammatory bowel disease

Marie Vibeke Vestergaard1 | Kristine H. Allin1,2 | Carsten Eriksen1,3 | Oliwia Zakerska-Banaszak4 | Ramesh P. Arasaradnam5 | Mohammad T. Alam5,6 | Karsten Kristiansen1,7 | Susanne Brix1,3 | Tine Jess1,2

1Center for Molecular Prediction of Inflammatory Bowel Disease, PREDICT, Department of Clinical Medicine, Aalborg University, Copenhagen, Denmark
2Department of Gastroenterology & Hepatology, Aalborg University Hospital, Aalborg, Denmark
3Department of Biotechnology and Biomedicine, Technical University of Denmark, Lyngby, Denmark
4Institute of Human Genetics, Polish Academy of Sciences, Poznan, Poland
5Warwick Medical School & Cancer Research Centre, University of Leicester, Leicester, UK
6Department of Biology, United Arab Emirates University, Al Ain, Abu Dhabi, United Arab Emirates
7Laboratory of Genomics and Molecular Medicine, Department of Biology, University of Copenhagen, Copenhagen, Denmark

Abstract

Background: Inflammatory bowel diseases (IBD), including Crohn's disease (CD) and ulcerative colitis (UC), affect millions of people worldwide with increasing incidence.

Objectives: Several studies have shown a link between gut microbiota composition and IBD, but results are often limited by small sample sizes. We aimed to re-analyze publicly available fecal microbiota data from IBD patients.

Methods: We extracted original fecal 16S rRNA amplicon sequencing data from 45 cohorts of IBD patients and healthy individuals using the BioProject database at the National Center for Biotechnology Information. Unlike previous meta-analyses, we merged all study cohorts into a single dataset, including sex, age, geography, and disease information, based on which microbiota signatures were analyzed, while accounting for varying technical platforms.

Results: Among 2518 individuals in the combined dataset, we discovered a hitherto unseen number of genera associated with IBD. A total of 77 genera associated with CD, of which 38 were novel associations, and a total of 64 genera associated with UC, of which 28 represented novel associations. Signatures were robust across different technical platforms and geographic locations. Reduced alpha diversity in IBD compared to healthy individuals, in CD compared to UC, and altered microbiota composition (beta diversity) in UC and especially in CD as compared to healthy individuals were found.

Conclusions: Combining original microbiota data from 45 cohorts, we identified a hitherto unseen large number of genera associated with IBD. Identification of microbiota features robustly associated with CD and UC may pave the way for the identification of new treatment targets.

KEYWORDS
16S rRNA amplicon sequencing, Crohn’s disease, Faecalibacterium, IBD, Lachnospiraceae NK4A136 group, microbiome, microbiota, Roseburia, Turicibacter, ulcerative colitis
INTRODUCTION

The inflammatory bowel diseases (IBDs), ulcerative colitis (UC) and Crohn's disease (CD), are chronic inflammatory diseases of the gastrointestinal tract affecting an increasing number of individuals worldwide. The pathophysiology of IBD is only partially understood. In addition to genetic and environmental risk factors for IBD, IBD is associated with decreased diversity and altered relative abundance of certain bacterial genera and species of the gut microbiota.

Most microbiota studies are based on amplicon sequencing of the bacterial and archeal 16S rRNA gene. However, the Microbiome Quality Control project has shown that variation in each step of the laboratory and bioinformatics pipeline introduces different biases in the resulting microbiota datasets. Pipeline challenges include sample storage, DNA extraction method, primer choice, sequencing depth, and bioinformatics quality control procedures. Also, the choice of statistical model may influence results, and today no gold standard for the conduct of microbiota studies exists.

Previous meta-analyses of gut microbiota in IBD have primarily summarized published results without combining original 16S rRNA amplicon sequencing data from existing datasets to ensure that bioinformatic processing of sequencing reads and the following statistical analysis are performed consistently across studies. A large, combined sample size increases the chance of identifying robust and consistent IBD signatures across geographic locations, enables adjustment for individual technical choices, and increases the statistical power to detect new genera associated with IBD.

Therefore, the aim of the present study was to identify, combine, and re-analyze original 16S rRNA amplicon sequencing data from internationally available cohorts of patients with IBD and healthy individuals, while taking technical variation between studies into account, to provide well-powered and robust microbiota signatures for IBD.

METHODS

Collection of publicly available original data

To identify available original fecal 16S rRNA amplicon sequencing datasets from IBD patients and healthy individuals worldwide, we searched the BioProject database of the National Center for Biotechnology Information (NCBI) using the search string "(16S OR microbiome OR microbiota OR bacterial AND (colitis OR ibd OR Crohn's OR uc OR cd OR inflammatory))". Of the resulting 2270 matches, we included relevant datasets that were based on at least five stool samples and only contained samples from humans, excluding samples from infants, reducing the number of eligible studies to 61. We were unable to match 12 with a publication (Table S1), and they were excluded. For the remaining 49 datasets, if no cohort information was linked to the sequencing reads data, we sent a request for access to the additional information to the corresponding authors. If more than one sample were available per subject, the baseline sample was selected. Of the 49 original datasets, four were excluded due to no information on sample type (fecal or biopsy, accession number PRJNA436359), non-overlapping paired-end sequencing reads (accession number PRJNA7761255), or forward and reverse reads being inconsistent with each other (accession numbers: PRJNA606913 and PRJNA313074). Non-overlapping reads were removed as a quality filter of the data, while the inconsistency was not possible to surpass due to the replacement of the fastq IDs with SRA IDs in the files. Finally, 45 datasets were eligible for this study, of which we were able to obtain individual-level disease status information on 38 and further information on sex and age on 17 (Table 1, Table S2). All 45 cohorts were included when evaluating technical covariates, but analyses of IBD status were limited to the 17 cohorts, including sex and age information.

Microbiota analysis

All analyses were conducted in R (version 4.1.1). We used the DADA2 pipeline (package version 1.22.0) to infer amplicon sequencing variants (ASVs) from the sequencing reads. Each project was analyzed separately to allow for different error profile estimations and different primers. Taxonomy was assigned to ASVs using the SILVA database version 138.1.

The American Gut project (PRJEB11419) included data from individuals with many different diseases. We selected IBD patients and healthy individuals with no reported diseases and no current pregnancy. Other projects were limited to covering IBD and healthy controls and perhaps one other disease. In this case, we excluded individuals with additional diseases from the dataset. We performed quality control based on the depth of the reads that passed the DADA2 pipeline individually for each project (Figure S1), with the lowest accepted read depth of 2000. To combine all datasets into a single dataset, the microbiota data were pooled at the genus level (Figure S2).
**TABLE 1** Publicly available data from microbiota studies of IBD patients.

<table>
<thead>
<tr>
<th>Study</th>
<th>Country</th>
<th>UC/CD/Hi</th>
<th>% Males</th>
<th>Age span</th>
<th>Status</th>
</tr>
</thead>
<tbody>
<tr>
<td>Papa et al. (2012)</td>
<td>USA</td>
<td>43/23/24</td>
<td>48.9</td>
<td>3–24</td>
<td>1</td>
</tr>
<tr>
<td>Pérez-Brocal et al. (2013)</td>
<td>Spain</td>
<td>0/11/8</td>
<td>-</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td>Gevers et al. (2014)</td>
<td>USA</td>
<td>0/23/71</td>
<td>58.7</td>
<td>Mean: 12.3</td>
<td></td>
</tr>
<tr>
<td>Norman et al. (2015)</td>
<td>USA, United Kingdom</td>
<td>59/39/55</td>
<td>43.1</td>
<td>Mean: 41.5</td>
<td>4</td>
</tr>
<tr>
<td>Pérez-Brocal et al. (2015)</td>
<td>Spain</td>
<td>0/12/12</td>
<td>37.5</td>
<td>18–71</td>
<td></td>
</tr>
<tr>
<td>Duranti et al. (2016)</td>
<td>Italy</td>
<td>14/0/14</td>
<td>39.3</td>
<td>25–45</td>
<td>2</td>
</tr>
<tr>
<td>Jacobs et al. (2016)</td>
<td>USA</td>
<td>10/26/54</td>
<td>55.6</td>
<td>Mean: 14.0</td>
<td>3</td>
</tr>
<tr>
<td>Tedjo et al. (2016)</td>
<td>The Netherlands</td>
<td>0/71/0</td>
<td>46.5</td>
<td>18–70</td>
<td>2</td>
</tr>
<tr>
<td>Bajer et al. (2017)</td>
<td>Czech Republic</td>
<td>32/0/31</td>
<td>47.6</td>
<td>20–72</td>
<td>2</td>
</tr>
<tr>
<td>Ijaz et al. (2017)</td>
<td>Scotland</td>
<td>0/19/31</td>
<td>60.0</td>
<td>IQR: 11–50</td>
<td></td>
</tr>
<tr>
<td>Jacob et al. (2017)</td>
<td>USA</td>
<td>0/20/2</td>
<td>60.0</td>
<td>23–71</td>
<td>2</td>
</tr>
<tr>
<td>Iwasawa et al. (2017)</td>
<td>Japan</td>
<td>16/0/23</td>
<td>43.6</td>
<td>3–23</td>
<td></td>
</tr>
<tr>
<td>Santoru et al. (2017)</td>
<td>Italy</td>
<td>82/50/51</td>
<td>54.1</td>
<td>Mean: 45.9</td>
<td>1</td>
</tr>
<tr>
<td>Doherty et al. (2018)</td>
<td>USA</td>
<td>0/306/0</td>
<td>38.2</td>
<td>Mean: 39.0</td>
<td></td>
</tr>
<tr>
<td>Forbes et al. (2018)</td>
<td>Canada</td>
<td>19/20/23</td>
<td>38.7</td>
<td>Mean: 43.8</td>
<td></td>
</tr>
<tr>
<td>Goyal et al. (2018)</td>
<td>USA</td>
<td>12/7/0</td>
<td>57.1</td>
<td>8–21</td>
<td></td>
</tr>
<tr>
<td>Kennedy et al. (2018)</td>
<td>United Kingdom</td>
<td>0/37/54</td>
<td>47.3</td>
<td>43–65</td>
<td>1</td>
</tr>
<tr>
<td>Kump et al. (2018)</td>
<td>Austria</td>
<td>27/0/0</td>
<td>63.0</td>
<td>Mean: 41.0</td>
<td>2</td>
</tr>
<tr>
<td>Mcdonald et al. (2018)</td>
<td>Many</td>
<td>112/82/1770</td>
<td>53.3</td>
<td>0–89</td>
<td></td>
</tr>
<tr>
<td>Schäffler et al. (2018)</td>
<td>Germany</td>
<td>0/7/10</td>
<td>64.7</td>
<td>Mean: 33.1</td>
<td>2</td>
</tr>
<tr>
<td>Shutkever et al. (2018)</td>
<td>United Kingdom</td>
<td>120/150/0</td>
<td>44.1</td>
<td>Mean: 50.1</td>
<td>1</td>
</tr>
<tr>
<td>Smolinska et al. (2018)</td>
<td>The Netherlands</td>
<td>0/68/0</td>
<td>44.0</td>
<td>18–70</td>
<td>2</td>
</tr>
<tr>
<td>Zhou et al. (2018)</td>
<td>China</td>
<td>51/72/71</td>
<td>56.1</td>
<td>Mean: 33.7</td>
<td>2</td>
</tr>
<tr>
<td>Braun et al. (2019)</td>
<td>Israel</td>
<td>0/45/22</td>
<td>59.7</td>
<td>Mean: 35.6</td>
<td>3</td>
</tr>
<tr>
<td>Cold et al. (2019)</td>
<td>Denmark</td>
<td>7/0/0</td>
<td>71.4</td>
<td>27–50</td>
<td>2, 5</td>
</tr>
<tr>
<td>Li et al. (2019)</td>
<td>China</td>
<td>0/9/9</td>
<td>55.1</td>
<td>Mean: 35.8</td>
<td>2</td>
</tr>
<tr>
<td>Presti et al. (2019)</td>
<td>Italy</td>
<td>31/7/47</td>
<td>56.5</td>
<td>21–74</td>
<td></td>
</tr>
<tr>
<td>Weng et al. (2019)</td>
<td>China</td>
<td>107/173/42</td>
<td>65.8</td>
<td>IQR: 21–53</td>
<td>1</td>
</tr>
<tr>
<td>Alam et al. (2020)</td>
<td>United Kingdom</td>
<td>11/10/9</td>
<td>-</td>
<td>-</td>
<td>2, 5</td>
</tr>
<tr>
<td>Ambroziewicz et al. (2020)</td>
<td>Poland</td>
<td>0/15/9</td>
<td>66.7&lt;sup&gt;a&lt;/sup&gt;</td>
<td>20–65&lt;sup&gt;a&lt;/sup&gt;</td>
<td>2</td>
</tr>
<tr>
<td>Ayelén et al. (2020)</td>
<td>Argentina</td>
<td>20/0/0</td>
<td>-</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td>Diederen et al. (2020)</td>
<td>The Netherlands</td>
<td>0/27/15</td>
<td>52.0&lt;sup&gt;a&lt;/sup&gt;</td>
<td>IQR: 12–15&lt;sup&gt;a&lt;/sup&gt;</td>
<td>2, 5</td>
</tr>
<tr>
<td>MacHiels et al. (2020)</td>
<td>Belgium</td>
<td>0/54/0</td>
<td>47.5</td>
<td>Mean: 51.4</td>
<td>2</td>
</tr>
<tr>
<td>Metwaly et al. (2020)</td>
<td>Spain</td>
<td>0/15/0</td>
<td>27.6</td>
<td>-</td>
<td>2</td>
</tr>
<tr>
<td>Schierová et al. (2020)</td>
<td>Czech Republic</td>
<td>16/0/0</td>
<td>50.0</td>
<td>28–66</td>
<td>3, 5</td>
</tr>
<tr>
<td>Sokol et al. (2020)</td>
<td>France</td>
<td>0/17/0</td>
<td>52.9</td>
<td>IQR: 29–37</td>
<td></td>
</tr>
<tr>
<td>Wang et al. (2020)</td>
<td>China</td>
<td>58/0/72</td>
<td>51.5</td>
<td>Mean: 43.9</td>
<td>2</td>
</tr>
<tr>
<td>Clooney et al. (2021)</td>
<td>Canada, Ireland</td>
<td>228/303/161</td>
<td>54.3</td>
<td>Mean: 50.3</td>
<td>1, 5</td>
</tr>
<tr>
<td>Cortez et al. (2021)</td>
<td>Brazil</td>
<td>12/0/23</td>
<td>68.6</td>
<td>2–21</td>
<td></td>
</tr>
<tr>
<td>Dai et al. (2021)</td>
<td>China</td>
<td>24/0/14</td>
<td>55.3</td>
<td>Mean: 34.9</td>
<td>2</td>
</tr>
</tbody>
</table>
TABLE 1 (Continued)

<table>
<thead>
<tr>
<th>Study</th>
<th>Country</th>
<th>UC/CD/II</th>
<th>% Males</th>
<th>Age span</th>
<th>Status</th>
</tr>
</thead>
<tbody>
<tr>
<td>Frau et al. (2021)35</td>
<td>United Kingdom</td>
<td>20/23/20</td>
<td>50.8</td>
<td>Mean: 43.0</td>
<td>1</td>
</tr>
<tr>
<td>Imai et al. (2021)36</td>
<td>Japan</td>
<td>42/16/45</td>
<td>59.0</td>
<td>Mean: 29.5</td>
<td>4</td>
</tr>
<tr>
<td>Maldonado-Arriaga et al. (2021)37</td>
<td>Mexico</td>
<td>18/0/15</td>
<td>57.6</td>
<td>Mean: 36.9</td>
<td>1</td>
</tr>
<tr>
<td>Schierova et al. (2021)38</td>
<td>Czech Republic</td>
<td>10/17/37</td>
<td>34.4</td>
<td>IQR: 26–44</td>
<td></td>
</tr>
<tr>
<td>Zakerska-Banaszak et al. (2021)39</td>
<td>Poland</td>
<td>10/0/10</td>
<td>45.0</td>
<td>Mean: 46.3</td>
<td>2, 5</td>
</tr>
</tbody>
</table>

Note: A total of 45 studies were eligible for this project. Information on sex and age was collected from each publication for both ulcerative colitis (UC) and Crohn’s disease (CD) patients as well as healthy individuals (II). Status 1 = publicly available data missed individual level information on disease status, age, and sex. Status 2 = publicly available data missed individual level information on disease status, age, and sex. Status 3 = publicly available data missed individual level information on either age or sex. Status 4 = subject names in publication do not match subject names of sequencing data. Status 5 = Additional individual level data was received from authors. -- Not reported in publication.

Abbreviations: CD, Crohn’s disease; HI, healthy individuals; IBD, inflammatory bowel diseases; IQR, inter-quartile range; UC, ulcerative colitis.

*Only provided for patients.

Statistical analysis of diversity measures

We calculated two alpha diversity measures, Chao1 richness and Shannon diversity, based on ASV count data from each project ("estimate richness", phyloseq-package version 1.36.0). Chao1 richness and read depth were transformed using the log-transformation to approach normality (Figure S3). Alpha diversity measures more than three standard deviations (SD) away from the mean were removed as outliers. This removed 42 samples from the Chao1 dataset and 34 from the Shannon dataset out of a total of 5381 samples. The association between alpha diversity and different variables was analyzed using linear models ("lm", stats-package version 4.1.1), and type II ANOVA models ("Anova", car-package version 3.0.13) to generate results for numerical and categorical variables, respectively.

We calculated Bray-Curtis and Jaccard dissimilarities as beta diversity measures based on relative abundance data at the genus level ("distance", phyloseq-package version 1.36.0). We conducted analyses of beta diversity using PERMANOVA models ("adonis2", vegan-package version 2.5.7). They included a permutation design with permutation within projects, but not between, when analyzing age, sex, body mass index (BMI), smoking status, and the disease status (CD, UC, and healthy individuals). Technical features and geographical location are constant within a project and were analyzed without restrictions to the permutation design. We were not included as covariates in models including the project-based permutation design. 999 permutations were used, and the marginal effects were reported.

We evaluated all available technical and biological variables for association with alpha and beta diversity to select important confounding variables. The effects of technical features (V-region, sequencing instrument, sequencing depth, and paired-end/single-end layout) were analyzed by including all in the same model. Important technical features (p-value < 0.05) were included as possible confounders in the following models. Biological features (sex, age, BMI, and smoking status) and geographical location were individually tested for association with alpha and beta diversity due to differences in metadata sparsity in some cohorts. This resulted in 5206, 3202, 2641, 1788 and 1835 samples in the analysis of geographical location, sex, age, BMI, and smoking status, respectively. Potential biological confounders were likewise selected based on p-values < 0.05. In the final model investigating the association between microbiota features and disease status, technical confounders and the geographical continent of the subjects were made into a single variable to avoid collinearity by merging the variables.

Statistical analysis of individual genera versus disease status

We analyzed associations between disease status and microbiota composition at the genus level. The number of genera to be tested was reduced from 1345 to 144 by excluding all genera present in less than 10% of the samples as advised by Nearing et al.10 As the choice of statistical method influences the final results, Nearing et al. recommended the use of multiple models when analyzing microbiota data to reduce this bias.10 We included three different and commonly used models61 that all enable adjustment for confounders, and we only report results consistent across all three methods in regard to statistical significance (adjusted p-value < 0.05) and direction of association. We employed a zero-inflated negative binomial (ZINB) model62 ("zeroinfl", pscl-package version: 1.5.5) to model raw microbiota data without any normalization and transformation. ANCOM-BC63 ("ancombc", ANCOMBC-package version: 1.4.0) and MaAsLin264 ("Maaslin2", Maaslin2-package version: 1.8.0) were selected based on a recommandation from Nearing et al. Further information on the three models is available in Table S3. The ZINB model is a mixture model and thus provides both a result for the negative binomial and zero-inflated part of the model. Thus, only one of these results needed to agree with ANCOM-BC and MaAsLin2 when evaluating the results. p-values were adjusted using the Benjamini-Hochberg method.

The combined variable of technical effects and geographical continent, as well as sex and age of the participants were included as covariates. Further, the log of the sequencing depth was included as an offset in the ZINB model. Eleven of the 144 tested genera failed to
fit the ZINB model and were thus only tested in the two other models.

Finally, we constructed a permutation test to evaluate whether the microbiota profile of CD was more similar to UC than expected if overlaps were random. This was validated by permuting the genera 999 times and re-calculating the number of overlaps between CD and UC findings.

RESULTS

This study included original data from 16S rRNA gene amplicon sequencing of stool samples from 45 IBD cohorts across the world (Table 1, Table S2). The final combined dataset consisted of 5381 stool samples from 36 different countries on five different continents, harboring 1345 different genera. Each sample consisted of at least 2000 summarized genus counts. Information on disease status, sex, and age was available for 2518 samples, of which 934 were patients with IBD (CD, N = 691 and UC, N = 243), and 1584 were healthy individuals (Figure 1). IBD patients consisted of a heterogeneous population of IBD patients including both patients in remission, with active disease, in treatment, naïve to treatment, etc.

Potential technical and biological confounders

Technical (sequencing instrument, sequencing layout, analyzed V-region, and read depth) and biological (sex, age, BMI, and smoking status) variables as well as geographical location were evaluated for association with alpha and beta diversity to assess possible confounding. Most of the technical and biological variables were associated with alpha and beta diversity (Table S4). The combined technical variables obtained an R² value of 0.063 when analyzing Bray-Curtis dissimilarity and 0.074 when analyzing Jaccard dissimilarity. Hence, all the following analyses were corrected for all technical and biological variables, except for BMI and smoking status, which were adjusted only for in a sensitivity analysis to maximize sample size in the main analysis.

Microbiota diversity in IBD versus healthy individuals

Alpha diversity varied across studies (Figure 2a) and was consistently reduced for both UC and CD as compared to healthy individuals (Figure 2b). For CD and UC, respectively, the estimated difference in log-transformed Chao1 richness to healthy individuals was −0.41 (standard error [SE] = 0.025, p-value < 2e−16) and −0.27 (SE = 0.032, p-value < 2e−16). Likewise, the estimated difference in Shannon diversity was −0.34 (SE = 0.047, p-value = 4e−13) for CD and −0.28 for UC (SE = 0.061, p-value = 5e−6) as compared to healthy individuals. Alpha diversity was lower in CD patients compared to UC patients using the log-transformed Chao1 richness for comparison (estimate = −0.14, SE = 0.036, p-value = 1e−4), but not for the Shannon diversity index (estimate = 0.062, SE = 0.069, p-value = 0.36).

The beta diversity measures, Bray-Curtis and Jaccard dissimilarity (Figure 2c), revealed that CD explained 2.41%−3.75% of the microbiota variation between individuals, whereas UC only explained

FIGURE 1  Publicly available microbiota studies on IBD patients and healthy individuals. Forty-five studies had publicly available data usable for this project. (a) Bar plot summarizing the number of studies available for each 16S V-region. The studies are further colored according to the sequencing instrument used. (b) Bar plot summarizing the number of samples available for each disease status. The samples are colored by the sex of the subject. CD, Crohn’s disease; HI, healthy individuals; IBD, inflammatory bowel diseases; NA, not available; UC, ulcerative colitis.
0.70%–1.01% of this variation. Combined, disease status explained 2.41%–3.65% of the variation in the dataset. Healthy individuals were more homogeneous in microbiota composition than IBD patients across cohorts (Figure 2c).

Results were similar upon inclusion of BMI and smoking status in the models (N = 1,602, Table S5).

**IBD-associated genera**

A total of 94 genera were identified to be differentially abundant between patients with IBD and healthy individuals (Table 2, Table S6A, and Figure S4). Of these, 55 showed lower abundance in IBD. A ranking of the 94 identified genera based on absolute effect sizes is presented
<table>
<thead>
<tr>
<th>Higher abundance in both CD and UC</th>
<th>Higher abundance only in CD</th>
<th>Higher abundance only in UC</th>
</tr>
</thead>
<tbody>
<tr>
<td>(Clostridium) innocuum group</td>
<td>(Eubacterium) fissicatena group</td>
<td><strong>Bifidobacterium</strong></td>
</tr>
<tr>
<td>(Eubacterium) brachy group</td>
<td>Anarotrunccus</td>
<td><strong>Gordonibacter</strong></td>
</tr>
<tr>
<td>(Ruminococcus) gnarus group</td>
<td>Atopobium</td>
<td><strong>Lactcasebacillus</strong></td>
</tr>
<tr>
<td>Actinomyces</td>
<td>Candidatus Soleaeferrea</td>
<td><strong>Ruminococcaceae DTU089</strong></td>
</tr>
<tr>
<td>Eggerthella</td>
<td>Dialister</td>
<td><strong>Turicibacter</strong></td>
</tr>
<tr>
<td>Enterococcus</td>
<td>Dielma</td>
<td></td>
</tr>
<tr>
<td>Erysipelatoclostridium</td>
<td>Eisenbergiella</td>
<td></td>
</tr>
<tr>
<td>Faecalibacter</td>
<td>Fusobacterium</td>
<td></td>
</tr>
<tr>
<td>Flavonifractor</td>
<td>Hungatella</td>
<td></td>
</tr>
<tr>
<td>Gemella</td>
<td>Lachnocolstridum</td>
<td></td>
</tr>
<tr>
<td>Haemophilus</td>
<td>Lachnospiraceae UCG-010</td>
<td></td>
</tr>
<tr>
<td>Intestinibacter</td>
<td>Lactobacillus</td>
<td></td>
</tr>
<tr>
<td>Lactococcus</td>
<td>Megasphaera</td>
<td></td>
</tr>
<tr>
<td>Sellimonas</td>
<td>Mogibacterium</td>
<td></td>
</tr>
<tr>
<td>Streptococcus</td>
<td>Ruminococcaceae UBA1819</td>
<td></td>
</tr>
<tr>
<td>Veillonella</td>
<td>Solobacterium</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Tuzzerella</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Tyzcibacter</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Lower abundance in both CD and UC</th>
<th>Lower abundance only in CD</th>
<th>Lower abundance only in UC</th>
</tr>
</thead>
<tbody>
<tr>
<td>(Eubacterium) ruminantium group</td>
<td>(Eubacterium) eligens group</td>
<td><strong>Alloprevotella</strong></td>
</tr>
<tr>
<td>(Eubacterium) sireum group</td>
<td>(Eubacterium) hallii group</td>
<td><strong>Angelakisella</strong></td>
</tr>
<tr>
<td>(Eubacterium) ventriusom group</td>
<td>Colidextrabacter</td>
<td><strong>Bacteroides</strong></td>
</tr>
<tr>
<td>(Eubacterium) xylanophilum group</td>
<td>Collinsella</td>
<td><strong>Bilohila</strong></td>
</tr>
<tr>
<td>Agathobacter</td>
<td>Corynebacterium</td>
<td><strong>Butyricimonas</strong></td>
</tr>
<tr>
<td>Akkermansia</td>
<td>Faecalibacterium</td>
<td><strong>Coprobacter</strong></td>
</tr>
<tr>
<td>Alistipes</td>
<td>Holdemanella</td>
<td><strong>Desulfovibrio</strong></td>
</tr>
<tr>
<td>Barnesiella</td>
<td>Lachnospiraceae FC5020 group</td>
<td><strong>Family XIII UCG-001</strong></td>
</tr>
<tr>
<td>Christensenellaceae R-7 group</td>
<td>Marvinbyantia</td>
<td><strong>Lachnospiraceae AC2044 group</strong></td>
</tr>
<tr>
<td>Coprococcus</td>
<td>Monogobus</td>
<td><strong>Odoribacter</strong></td>
</tr>
<tr>
<td>Funicatenibacter</td>
<td>Pseudomas</td>
<td><strong>Oxalobacter</strong></td>
</tr>
<tr>
<td>Intestimonas</td>
<td>Romboutsia</td>
<td><strong>Prevotella, 9</strong></td>
</tr>
<tr>
<td>Lachnospira</td>
<td>Turicibacter</td>
<td><strong>Vicivallis</strong></td>
</tr>
<tr>
<td>Lachnospiraceae CAG-56</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Lachnospiraceae ND3007 group</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Lachnospiraceae NK4A136 group</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Lachnospiraceae UCG-001</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Oscilospiraceae NK4A214 group</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Oscilospiraceae UCG-002</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Oscilospiraceae UCG-003</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Oscilospiraceae UCG-005</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Parabacteroide</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Paraprevotella</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Phascolarctobacterium</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Prevotella</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Roseburia</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ruminococcaceae CAG-352</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ruminococcus</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Subdoligranulum</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Note: All listed genera had an adjusted p-value < 0.05 in all three statistical models and same direction of association. The genera are divided according to association with CD, UC or both CD and UC, and according to whether their abundance in IBD was higher or lower than in healthy individuals. They are named by genus name and sorted alphabetically. Further information is available in Table S6A and Figure S5A. The names are colored based on whether the genus previously has been associated with IBD. Green: The same direction of association has been reported in literature. Orange: An opposite direction of association has been reported in literature. Black: The genus has not been associated with IBD previously.

Abbreviations: CD, Crohn’s disease; IBD, inflammatory bowel diseases; UC, ulcerative colitis.

In Table S7 and S8 for CD and UC, respectively. We found that 77 genera were statistically significantly associated with CD, and 64 genera were associated with UC upon adjustment for multiple testing. A total of 46 genera had the same direction of association in CD and UC. An overlap of 46 similar associations for CD and UC was significantly more than expected from random association...
(permutation test: p-value < 0.001). Turicibacter was the only genus showing opposite direction of association between CD and UC.

When adjusting for smoking status and BMI, including only datasets where these variables were available, 22 genera were associated with CD and 32 with UC. Of these, 21 and 28, were similar to the genera detected in the main analysis (Figure S5 and Table S6B). An additional sensitivity analysis limited to samples with information on smoking status and BMI, but not including these as covariates in the model, produced similar results (Figure S6).

Distinguishing the gut microbiota in CD and UC

When comparing the microbiota composition in CD versus UC, we found 31 discriminative genera (Table 3, Figure S7, and Table S6C), of which 17 were increased and 14 were decreased in relative abundance in CD patients compared to UC patients.

DISCUSSION

In this study analyzing original 16S fecal microbiota datasets from 45 cohorts across the world, encompassing 934 IBD patients and 1584 healthy individuals, we confirmed reduced alpha diversity in patients with IBD and a microbiota composition that not only differs between IBD and healthy individuals, but also between CD and UC. Through analyses of 144 different genera present in at least 10% of study participants, we discovered the hitherto largest number of genera associated with CD (n = 77) and UC (n = 64). Of these, 46 overlapped across CD and UC. Our analysis was uniquely powered to correct for technical differences that normally introduce bias, was based on coinciding findings from three statistical models, and therefore provides a comprehensive, robust, and generalizable microbiota signature of IBD.

The reduced alpha diversity in patients with IBD compared to healthy individuals and in patients with CD compared to UC were in accordance with former studies of smaller sample size. Also, in analyses of beta diversity, we found altered microbiota composition in UC and especially in CD as compared to healthy individuals.

For the first time, we identified as many as 77 genera associated with CD and 64 associated with UC. When comparing these genera with results from (1) the 45 individual study cohorts eligible for the current study, (2) the four excluded studies, and (3) four previous meta-analyses of IBD microbiota signatures, we found that 38 genera in CD and 28 genera in UC represented novel associations. These novel findings appear to be robust, since (1) the beta diversity plot did not show clear batch effects, (2) we adjusted for technical differences across studies, and (3) we detected a large overlap with previous studies at the genus level, which was consistent using three different analytical methods.

When examining the ranking of absolute effect sizes of genera found to be associated with IBD in this project, we identified a clear pattern for UC, where genera detected previously in other studies for association with IBD on average showed larger effect sizes than the new discoveries. This supports the notion that sample size is the main explanation for the ability of the present study to detect novel associations. This is further supported by the fact that genera reported in smaller studies with associations opposite to what we observe were ranking lowest in the present study. Thus, a small absolute effect size is more likely to result in an inverse association. We also observed a large variance in ranking across the three different statistical methods (NBZI, ANCOM-BC and MaAsLin2), which is likely a contributing factor to differences in results across former microbiota studies. It is a strength of this study, that we used three different methods and only reported overlapping findings since we then reduced bias related to statistical methods applied in individual studies and related to individual approaches of normalization and transformation of data. This approach was chosen to overcome the general limitation of microbiota studies, where the choice of statistical model influences the outcome.

Of the four previously mentioned meta-analyses, the one by Mancabelli et al. may appear partly similar to ours, as the authors re-analyzed data from 2177 IBD patients and healthy individuals using a common bioinformatic pipeline. However, for statistical analysis at the genus level, each project was analyzed separately,
without correcting for technical and biological covariates. Therefore, the authors did not obtain the power obtained in the present study, and the study only detected 18 genera associated with CD and 13 genera associated with UC.

The primary strength and novelty of this study was the combination and re-analysis of publicly available data from all cohort studies on gut microbiota signatures in IBD, resulting in a hitherto unseen large sample size of IBD patients and healthy individuals. Thus, the bioinformatical pipeline and the statistical method were unified across study data, which eliminated bias and enabled controlling for different techniques. Further, our large sample size enabled detection of a wide range of new genera with smaller effect sizes.

Combining different studies will be influenced by a batch effect caused by technical heterogeneity, which must be accounted for in subsequent analyses. Reducing the complexity of the dataset from the ASV level to the genus level reduced the influence of batch effect. We estimated the remaining batch effect using the available technical information and adjusted for these in all analyses in this paper. We further ran a sensitivity analysis testing whether the results of alpha-diversity analyses would be influenced by also including a random intercept with the study ID to further correct for batch effect. The sensitivity analysis produced similar results (results not shown). A similar sensitivity analysis was already included in the permutation design of beta-diversity analyses and was unfortunately not possible in the single genus analysis since all three analysis tools did not allow inclusion of random effects. However, since it did not affect the analyses of diversity, we concluded that the batch effect and technical heterogeneity were sufficiently accounted for in the analysis by including technical and geographical covariates.

The present study also has limitations that need consideration. First, despite access to and combination of original microbiota data from all identified studies, we were unable to obtain complete information on disease status, sex, age, BMI, and smoking status across all study cohorts. This limited the sample size used in analyses. We conducted a sensitivity analysis adjusting for BMI and smoking status. However, the results of the sensitivity analysis were similar to results from analyses just reducing the sample size based on the availability of BMI and smoking status, hence suggesting that differences between main and sensitivity analyses were mainly driven by differences in sample size and not by adjusting for BMI and smoking status in addition to the other covariates.

We corrected for technical covariates available to us (V-region, sequencing instrument, library layout, and sequencing depth); however, other technical choices might also affect the resulting dataset, including sample storage, DNA extraction method, and primer choice. Furthermore, variation in gut microbiota between IBD subtypes may reflect differences in medication use, disease course, disease location, and disease severity, which was not captured consistently by available data resources either. Not considering this cohort heterogeneity in analyses is a potential limitation to the study. However, the detection of numerous significant associations after correction for multiple testing, consistent across three different statistical models, suggests that results are robust across the diverse IBD populations. Quality control of the sequencing depth of samples was performed differently for each project, and the average sequencing depth varied markedly across projects. We used a sequencing depth of >2000 as the lowest threshold for inclusion. We could have also chosen to exclude projects that ended up being outliers in terms of sequencing depth. However, as an example, samples with more extreme sequencing depth than mean 3SD of the log-transformed sequencing depth only accounted for 8 out of 2518 samples and would thus not change the outcomes of analyses. Lastly, a world-wide examination of gut microbiota signatures in IBD, based on publicly available studies, will always be prone to selection bias. 16S IBD projects are more likely to be conducted in countries with higher research investments, whereas national data protection rules, on the other hand, may decrease the likelihood of data being publicly available.

**CONCLUSION**

This comprehensive re-analysis of data on 2518 IBD patients and healthy individuals from 45 cohorts identified 38 novel genera associated with CD and 28 with UC, in addition to confirming an altered alpha and beta diversity in IBD. Genera detected by the single genus analysis may potentially represent novel targets for IBD prediction and therapy. Especially, genera found to be reduced in the microbiota of IBD patients, such as *Lachnospiraceae NK4A136 group*, *Faecalibacterium*, *Roseburia*, *Alistipes* and *Parabacteroides*, may be interesting candidates for future invention strategies. The number of novel findings is surprising considering the large heterogeneity within IBD patients. Future studies would benefit from selecting a more homogenous subgroup or collecting comprehensive characterization of the IBD patients to account for the heterogeneity. Re-analyzing original gut microbiota data while taking technical issues into account may also pave the way for the identification of robust signals of value for the implementation of microbial-based treatments in other disease areas.

**AUTHOR CONTRIBUTIONS**

**Conceptualization:** Marie Vibeke Vestergaard, Kristine H. Allin, Tine Jess. **Data curation:** Marie Vibeke Vestergaard. **Formal Analysis:** Marie Vibeke Vestergaard. **Funding acquisition:** Tine Jess. **Investigation:** Olivia Zakerska-Banaszak, Ramesh P. Arasaratnam, Mohammad T. Alam. **Methodology:** Marie Vibeke Vestergaard, Carsten Eriksen, Kristine H. Allin, Susanne Brix, Karsten Kristiansen. **Project administration:** Marie Vibeke Vestergaard, Tine Jess. **Resources:** Tine Jess. **Software:** Marie Vibeke Vestergaard. **Supervision:** Carsten Eriksen, Kristine H. Allin, Susanne Brix, Karsten Kristiansen, Tine Jess.
ACKNOWLEDGMENTS
This study was supported by a grant from the Danish National Research Foundation (Grant no. DNRF148).

CONFLICT OF INTEREST STATEMENT
The authors have nothing to disclose.

DATA AVAILABILITY STATEMENT
All data used in this study are already publicly available at the BioProject database of the National Center for Biotechnology Information. All source codes used in this study are available at GitHub: https://github.com/marievibeke/16S_IBD_project.

ORCID
Marie Vibeke Vestergaard https://orcid.org/0009-0006-1409-1945
Kristine H. Allin https://orcid.org/0000-0002-6880-5759
Carsten Eriksen https://orcid.org/0000-0003-4400-3855
Ramesh P. Arasaradnam https://orcid.org/0000-0002-2231-3062

REFERENCES


SUPPORTING INFORMATION
Additional supporting information can be found online in the Supporting Information section at the end of this article.