Manchester Metropolitan University

1 **Crohn's disease siblings exhibit a biologically relevant dysbiosis in mucosal**

2 **microbial metacommunities**

3

4 **Authors**

5 Charlotte Hedin[1,5a,] Christopher J. van der Gast[2a], Geraint B. Rogers[3], Leah Cuthbertson[2], Sara McCartney[4],

6 Andrew J. Stagg[5], James O. Lindsay[6,7], Kevin Whelan[1]

7

8 [1] King's College London, Faculty of Life Sciences & Medicine, Diabetes and Nutritional Sciences Division,

9 London, United Kingdom

10 [2] NERC Centre for Ecology & Hydrology, Wallingford, Oxfordshire, United Kingdom

11 [3] South Australian Health and Medical Research Institute, Infection and Immunity Theme, Flinders

12 University, Adelaide, Australia

13 [4] University College London, Centre for Gastroenterology and Nutrition, London, United Kingdom

14 [5] Blizard Institute, Queen Mary University of London, Centre for Immunology and Infectious Disease,

15 London, United Kingdom

16 [6] Blizard Institute, Queen Mary University of London, Centre for Digestive Diseases, London, United

17 Kingdom

18 [7] Barts Health NHS Trust, Department of Gastroenterology, London, United Kingdom

19 [a] These authors contributed equally to this work

20

1

21 **Corresponding author**

22 Professor Kevin Whelan

23 King's College London, Faculty of Life Sciences & Medicine, Diabetes and Nutritional Sciences Division,

24 Franklin Wilkins Building, 150 Stamford Street, London, SE1 9NH, United Kingdom

25

26 Email: kevin.whelan@kcl.ac.uk

27

28 Word count: 4162

29 Number of figures: 4

30 Number of tables: 3

31 Number of references: 39

32 Number of supplementary files for online only publication: 3

33

37

38 **Author contributions**

39 CRH: study concept and design; obtained funding; recruitment of participants and acquisition of data;

40 analysis and interpretation of data; statistical analysis; drafting of the manuscript; critical revision of the

2

48

49   **ABSTRACT**

50   Objective

51   Siblings of patients with Crohn's disease (CD) have elevated risk of developing CD and display aspects of

52   disease phenotype, including faecal dysbiosis. Whether the mucosal microbiota is disrupted in these at-risk

53   individuals is unknown. Objective: To determine the existence of mucosal dysbiosis in siblings of CD

54   patients using 454 pyrosequencing and to comprehensively characterise, and determine the influence of

55   genotypic and phenotypic factors, on that dysbiosis.

56   Design

57   Rectal biopsy DNA was extracted from 21 patients with quiescent CD, 17 of their healthy siblings and 19

58   unrelated healthy controls. Mucosal microbiota was analysed by 16S rRNA gene pyrosequencing and were

59   classified into core and rare species. Genotypic risk was determined using Illumina Immuno BeadChip,

60   faecal calprotectin by ELISA and blood T-cell phenotype by flow cytometry.

61   Results

62   Core microbiota of both CD patients and healthy siblings were significantly less diverse than controls.

63   Metacommunity profiling (Bray-Curtis (S_{BC}) index) showed the sibling core microbial composition to be

64   more similar to CD ($S_{BC}=0.70$) than to HC, whereas the sibling rare microbiota was more similar to HC

65   ($S_{BC}=0.42$). *Faecalibacterium prausnitzii* contributed most to core metacommunity dissimilarity both

66   between siblings and controls, and between patients and controls. Phenotype/genotype markers of CD-risk

67   significantly influenced microbiota variation between and within groups, of which genotype had the largest

68   effect.

69   Conclusion

4

70 Individuals with elevated CD-risk display mucosal dysbiosis characterised by reduced diversity of core

71 microbiota and lower abundance of *F. prausnitzii*. This dysbiosis in healthy people at-risk of CD implicates

72 microbiological processes in CD pathogenesis.

73 **SUMMARY BOX**

74 What is already known about this subject:

75 • Patients with CD have mucosal dysbiosis, including reduced abundance of *Faecalibacterium*

76 *prausnitzii*

77 • Low mucosal *Faecalibacterium prausnitzii* predicts relapse after surgery in CD patients

78 • Healthy siblings of CD patients have increased risk of developing CD and have altered abundance of

79 key species in the gut lumen

80 What are the new findings:

81 • There is a distinct dysbiosis in the mucosal microbiota of healthy siblings of CD patients

82 • The sibling dysbiosis comprises a fundamental distortion of microbial community composition,

83 most notably reduced diversity of core microbiota and low abundance of mucosal *Faecalibacterium*

84 *prausnitzii*

85 • Mucosal microbiota disruption is not merely a consequence of the inflammation in CD but is

86 present at healthy individuals at risk of CD

87

88 How might it impact on clinical practice in the foreseeable future?

89 • Identification of this at risk dysbiosis signals pathways in CD pathogenesis and raises the possibility

90 of CD risk identification and CD risk intervention

6

91 **INTRODUCTION**

92 Disruption of gut microbiota (dysbiosis) is an established feature of inflammatory bowel disease (IBD). The

93 dysbiosis in Crohn's disease (CD) has been well described and includes reduced microbial diversity, reduced

94 abundance of Firmicutes particularly *Faecalibacterium prausnitzii*, reduced abundance of Bifidobacteria,

95 increased γ-proteobacteria and disturbances in Bacteroides populations.[1] The involvement of several CD

96 susceptibility genes in the recognition and handling of bacteria (e.g., NOD2, ATG16L1, IRGM) reinforces the

97 position of the gut microbiota at the centre of IBD pathogenesis. [2, 3, 4]

98 Whether the CD dysbiosis is involved with pathogenesis is uncertain. The dependence on the presence of

99 gut microbiota for the development of inflammation in animal models[5] as well as CD patients,[6] and the

100 association between reduced mucosal *F. prausnitzii* and post-operative relapse[7] implies a pathogenic

101 role. Conversely, the lack of therapeutic benefit of manipulating the microbiota,[8,9] suggests that

102 dysbiosis in CD may not drive inflammation, but rather is consequent to established disease, reflecting for

103 example, the differential survival of various species in an inflamed environment. Moreover, attempts to

104 identify aspects of the CD dysbiosis that were present at disease initiation, which therefore potentially have

105 a role in pathogenesis may be obfuscated by both the mature disease phenotype of the patients studied

106 and the effect of the medical, surgical and patient-initiated attempts to treat and control symptoms.

107 Siblings of CD patients have a relative risk (RR) of developing CD of up to 35 times that of the general

108 population.[10] This risk is partly genetic, but is also driven by non-genetic factors many of which they

109 share with their CD-affected sibling.[10,11] Several of these non-genetic risk factors, such as mode of

110 delivery, breast feeding, maternal inoculum, home environment and weaning diet,[12] potentially impact

111 gut microbial acquisition and development. It follows that any aspect of the CD dysbiosis which is also

112 present in a healthy sibling cannot be disrupted as a consequence of disease, and rather may be implicated

113 in processes driving CD pathogenesis.[12]

7

114 Attempts have been made to determine whether aspects of the CD phenotype are present in patients'

115 unaffected relatives. These have assessed dysbiosis[13] and other features of the CD phenotype such as

116 raised faecal calprotectin (FC), increased intestinal permeability (IP) and the presence of anti-microbial

117 antibodies.[12] Using PCR probes selected to detect dominant species that comprise the dysbiosis in CD,

118 we have previously indicated that a faecal dysbiosis exists in healthy siblings of CD patients characterised by

119 reduced faecal Firmicutes including *F. prausnitzii*.[14] Moreover, we previously demonstrated in siblings

120 that a combination of luminal dysbiosis, raised FC, reduced abundance of circulating naïve T-cells,

121 disturbances in their expression of gut-homing β7 integrin and at-risk genotype could be combined to

122 create a multidimensional risk phenotype, which significantly distinguished healthy siblings of CD patients

123 from healthy, unrelated controls.[14]

124 It has been speculated that mucosal microbiota are of greater significance in CD pathogenesis than luminal

125 microbiota given their closer spatial relationship to the gut immune system. Yet, studies comparing

126 mucosal microbiota in CD patients, their families and healthy controls are rare due to the invasiveness of

127 procedures required to obtain mucosal samples from otherwise healthy individuals. However, the potential

128 rewards of obtaining such samples have been amplified by recent advances in the analysis of large, diverse

129 and complex microbial communities. Pyrosequencing technology and meta-community profiling enables

130 greater sampling depth permitting detection not only of dominant microbial community members but also

131 low-abundance (rare) taxa.[15, 16] The capacity to characterise core and rare microbial communities

132 separately may reveal microbial features associated with disease not otherwise readily apparent.

133 Furthermore, 16S rRNA gene pyrosequencing and other next-generation technologies have demonstrated

134 that microbial diversity can be orders of magnitude higher than previously appreciated.[16] Measuring

135 diversity may be significant as healthy gut microbiota high diversity compared with microbial populations in

136 other human body habitats.[18] Moreover, gut microbial diversity is consistently described as reduced both

137 in CD,[1] and other human diseases including obesity,[18, 19] colorectal cancer,[20] eczema,[21] and in

138 addition has been linked with smoking.[22]

8

139 Therefore, we sought to use 454 pyrosequencing and metacommunity analysis to comprehensively

140 characterise the structure and composition of the mucosal microbial community in an at-risk group of CD

141 siblings compared with CD patients and healthy controls.

142 **MATERIALS AND METHODS**

143 Patients with inactive CD (Crohn's Disease Activity Index (CDAI) <150 and C-reactive protein (CRP) ≤5mg/L,

144 and their healthy siblings (both 16-35 years) were recruited from clinics at Barts Health NHS Trust and

145 University College Hospitals NHS Foundation Trust (London, UK). Patients required a confirmed diagnosis of

146 CD for >3months. All healthy siblings who volunteered and did not meet exclusion criteria (detailed in

147 supplementary Table S1) were included, to limit bias in the selection of siblings with specific characteristics.

148 Healthy controls were recruited by email sent to staff and students at King's College London (London, UK),

149 during the same period. Participants were informed that involvement in the study did not constitute

150 screening for disease and that detection of clinical disease in any sibling or control would lead to exclusion

151 from the study. Only participants consenting to rectoscopy and providing analysable biopsies were

152 included. All participants provided written, informed consent. Ethical approval was provided by Bromley

153 Local Research Ethics Committee (reference 07/H0805/46).

154 At screening, demographics, medical and drug exposure history, physical examination, CRP, inclusion and

155 exclusion criteria were assessed. Instructions regarding avoidance of prebiotics/probiotics for 4 weeks (to

156 prevent impact on microbiota), non-steroidal anti-inflammatory drugs for 1 week and alcohol for 24h

157 before the study (to prevent impact on IP) were provided. Blood samples were taken for routine

158 haematology/biochemistry, T-cell analysis and genotyping. Participants completed a 5h urine collection for

159 measurement of IP and underwent flexible rectoscopy without bowel cleansing. Biopsies from non-

160 inflamed rectum were snap frozen, and stored at -80°C before processing for histological and

161 microbiological analyses. Stool was obtained and stored at -20°C before processing for FC quantification.

162 **Faecal calprotectin**

9

163 FC extraction and ELISA analysis (Calpro AS, Lysaker, Norway) were carried out according to manufacturer's

164 instructions using duplicate appropriately diluted samples. FC concentration ($\mu g/g$) was determined relative

165 to standard curves.

166 **Peripheral blood T-cell flow cytometry**

167 Whole blood, collected in lithium-heparin Vacutainer tubes (BD Bioscience), was stored at room

168 temperature for ≤4 h before labelling with fluorescently conjugated monoclonal antibodies to detect CD3 T-

169 cells, naïve (CD45RA$^+$) and memory (CD45RA$^-$) subsets of CD4 and CD8 T-cells. Integrin $\alpha4\beta7$ expression was

170 assessed by labelling with anti-$\beta7$ (see supplementary methods for antibodies used). Data were acquired

171 using a LSRII 4-colour flow cytometer (BD Bioscience) and collected using fluorescence-activated cell sorting

172 Diva software V.4.1.2 (BD Bioscience) using Flow-Count fluorospheres (Beckman Coulter) for absolute

173 quantitation. Colour compensation was performed offline using Winlist V.6.0 (Verity Software House).

174 **Genotyping**

175 Human DNA was extracted from whole blood using the phenol chloroform-isoamyl alcohol method.

176 Genotyping was performed using the Illumina Infinium Immunochip.[2, 23] To increase detection of NOD2

177 mutations and capture the enhanced risk of NOD2 compound heterozygosity, three NOD2 mutations

178 (rs2066845/G908R, rs2066844/R702W and rs5743293/3020insC) were individually assessed. Cumulative

179 genotype relative risk (GRR) for each participant was therefore calculated across 72 CD-risk loci. A

180 population distribution model of CD-risk was generated using the REGENT R program[24] and previously

181 published odds ratios.[2] Participants were categorised into reduced, average, elevated or high genotype

182 risk with reference to this model.[25]

183 **Intestinal permeability**

184 IP was measured using lactulose-rhamnose tests as previously described.[14]

185 **Gut mucosal microbiota**

10

186   Biopsy DNA extraction was carried out using a phenol/chloroform based method, as described

187   previously.[26] A detailed extraction protocol is provided in the supplementary methods. DNA extracts

188   were quantified using the Picodrop Microlitre Spectrophotometer (GRI, Braintree, UK). Negative controls

189   (sterile water), were included in the DNA extraction and PCR amplification steps.

190   Bacterial tag-encoded FLX amplicon pyrosequencing (bTEFAP) was performed as described previously using

191   Gray28F 5'-TTTGATCNTGGCTCAG-3' and Gray519r 5'-GTNTTACNGCGGCKGCTG-3'.[27] Detailed protocols

192   for 16S rRNA gene sequencing and sequence data processing are provided in the supplementary methods.

193   To assign bacterial identities to 16S rRNA gene sequences, sequence data were de-noised, assembled into

194   OTU clusters at 97% identity, and queried using a distributed .NEt algorithm that utilises Blastn+

195   (KrakenBLAST, www.krakenblast.com) against a database of high quality 16S rRNA gene bacterial

196   sequences. Using a .NET and C# analysis pipeline the resulting BLASTn+ outputs were compiled, data

197   reduction analysis performed, and sequence identity classification carried out, as described previously.[28]

198   **Statistical analyses**

199   Bacterial species within each metacommunity were partitioned into common and rare groups using a

200   modification of a previously described method.[15] Three complementary measurements of diversity were

201   used to compare microbial diversity between samples, as previously described: species richness (S*, the

202   total number of species), Shannon-Wiener (H', a metric accounting for both number and relative

203   abundance of species), and Simpson's (1-D, a measure of the probability that two species randomly

204   selected from a sample will differ).[15, 26] To avoid potential bias due to varying sequences per sample, all

205   measures were calculated using randomised re-sampling to a uniform number of sequence reads per

206   sample.[26] Mean diversity measures were calculated from the re-sampling of the reads from each

207   specimen to the lowest number of sequence reads among all specimens for 1000 iterations. Diversity

208   analysis was performed in R.[29] Two sample t-tests, regression analysis, coefficients of determination (r2),

209   residuals and significance (P) were calculated using Minitab software (version 16, Minitab, University Park,

11

210 PA, USA). Canonical correspondence analysis (CCA), analysis of similarity (ANOSIM), similarity of

211 percentages (SIMPER) analysis were performed using the PAST (Palaeontological Statistics, version 3.01)

212 program available from the University of Oslo website link (http://folk.uio.no/ohammer/ past) run by

213 Øyvind Hammer. The Bray-Curtis quantitative index of similarity was used as the underpinning community

214 similarity measure for CCA, ANOSIM, and SIMPER tests.

215 **RESULTS**

216 Demographic and disease characteristics of the 21 patients with quiescent CD, 17 of their healthy siblings,

217 and 19 unrelated healthy controls that were included are summarised in Table 1. At the time of the study

218 only one patient was cohabiting with one of the included siblings. GRR, FC, faecal Firmicute abundance and

219 circulating T-cell characteristics were all significantly different in both CD patients and healthy siblings

220 compared with healthy controls as previously published,[14] and as summarised in Table 1.

221

222 **Table 1 Summary of demographic variables in patients, siblings and controls as well as clinical**

223 **characteristics in patients. The features of the at-risk phenotype that have previously been delineated in**

224 **this cohort are also displayed.**

		Patients (n=21)	Siblings (n=17)	Controls (n=19)	*P*-value
Mean age years, (SD)		27.7 (6.6)	25.5 (4.5)	27.7 (5.8)	0.783*
Males, n (%)		13 (62)	11 (65)	9 (47)	0.515†
Body Mass Index, kg/m^2 (SD)		24.5 (5.0)	24.5 (3.6)	23.9 (3.4)	0.870*
Ethnicity n (%)	White British	17 (81)	15 (88)	17 (90)	
	Asian/Asian British	3 (14)	1 (6)	0 (0)	0.469†
	Black British or mixed black/white	1 (5)	1 (6)	2 (11)	
Smoking n (%)	Never	14 (67)	10 (59)	12 (63)	
	Current	4 (19)	5 (29)	3 (16)	0.830†
	Previous	3 (14)	2 (12)	4 (21)	
Age at diagnosis, n (%)	Below 16 years	7 (33)			
	16-40 years	14 (67)			
Disease location, n (%)	Ileal	7 (33)			
	Colonic	5 (24)			
	Ileocolonic	9 (43)			
Concomitant upper GI disease, n (%)		1 (5)			
Disease behaviour, n (%)	Non-stricturing, non-penetrating	11 (52)			
	Stricturing	5 (24)			
	Penetrating	5 (24)			
Perianal disease, n (%)		4 (19)			
Current 5-ASA n (%)		11 (52)			
Current immuno-suppressant, n (%)	Azathioprine	7 (33)			
	Mercaptopurine	2 (10)			
	Methotrexate	1 (5)			
Ileocaecal resection / right hemicolectomy, n (%)		9 (43)			
Isolated small bowel resection, n (%)		1 (5)			
Genotype relative risk,[‖] n (%)	High	3 (14)	1 (6)	0 (0)	
	Elevated	2 (10)	3 (18)	0 (0)	0.175†
	Average	10 (48)	8 (47)	8 (42)	

13

Reduced	6 (29)	5 (29)	11 (58)	
Fecal calprotectin, m/g (IQR)	281 (144-855)	30 (13-83)	13 (7-33)	<0.001‡
Faecal *F. prausnitzii*, % (IQR)	0.1 (0.0-2.9)	3.7 (0.4-7.1)	5.2 (2.3-7.2)	0.001‡
T-cells with memory phenotype, % (IQR)	73 (63-82)	74 (67-83)	65 (54-70)	0.011‡
Naïve $CD4^+$ T-cells, cells/ ml (IQR)	194,132 (71,053-341,156)	198,220 (128,550-296,351)	380,256 (279,118-564,861)	<0.001‡
Naïve $CD4^+$ T-cells expressing β7 integrin, % (IQR)	76 (63-85)	74 (61-83)	52 (32-71)	0.003‡
Intestinal permeability: urinary lactulose-rhamnose ratio, (IQR)§	0.061 (0.033-0.111)	0.034 (0.024-0.056)	0.038 (0.025-0.050)	0.081‡

225 * One-way ANOVA
226 † Chi-squared test
227 ‡ Kruskall-Wallis test
228 § Data from 20 patients, 17 siblings and 16 controls contributed to the intestinal permeability analysis
229 ‖ Cumulative genotype relative risk (GRR) for each participant was calculated across 72 CD-risk loci (detected using the Illumina
230 Infinium Immunochip). Participants were categorised into reduced, average, elevated or high genotype risk with reference to a
231 population distribution model of CD-risk, previously described.[14]
232

233 A total of 180,696 bacterial sequence reads (mean per sample 3235 ± SD 205), identifying 160 genera and

234 351 distinct operational taxonomic units (OTUs) classified to species level (Table S2), were generated from

235 all samples combined. The numbers of bacterial sequence reads per sample were similar among the three

236 cohorts (mean ± SD): CD, 3296 ± 258 (n =21); siblings, 3190 ± 423 (n =17); and healthy, 3210 ± 393 (n =19).

Species abundance was directly correlated with distribution

238 We have previously established that the categorisation of human microbiota into core and rare species

239 revealed important aspects of metacommunity species-abundance distributions that would be neglected

240 without such a distinction.[15] A coherent metacommunity could be expected to exhibit a direct

241 relationship between prevalence and abundance of individual species within the constituent communities.

242 Consistent with this prediction, the abundance of species in each study group significantly correlated with

243 the number of individual sample communities those species occupied (CD (R^2 =0.62, $F_{1,227}$ =366.9, P <

244 0.0001); siblings (R^2 =0.71, $F_{1,259}$ =590.1, P < 0.0001); and healthy controls (R^2 =0.68, $F_{1,258}$ =552.6, P <

245 0.0001)), (Fig. 1).

In CD patients a lower proportion of the mucosal microbiota were core species

247 Individual species in each cohort metacommunity were then classified as core or rare based on their falling

248 within or outside the upper quartile of subject occupancy, respectively (Fig. 1). Of the 229 species that

249 comprised the CD metacommunity, only 7 were core and 222 were rare species. The healthy siblings

250 metacommunity (261 species) comprised 18 core and 243 rare species, and the healthy controls

251 metacommunity (260 species) comprised 25 and 235 species, respectively. In addition, the core species

252 within each cohort metacommunity accounted for 44.7% ± 4.8% (CD), 67.6%,± 5.5% (healthy siblings) and

253 67.4% ± 4.6 (healthy controls) of the mean (± SD) relative abundance. The mean relative abundances in the

254 CD core microbiota were significantly lower than the healthy siblings and healthy controls (P < 0.0001 in

255 both cases), but were not different between the siblings and healthy controls (P =0.907).

Microbial diversity was lower in both siblings and patients compared with controls

15

257   The mean microbial diversity of subject communities for each cohort was compared using three indices of

258   diversity (Fig. 2). Diversity was compared between the three cohorts for the whole microbiota, as well as

259   core and rare species groups (Fig. 2). These analyses revealed the siblings' whole and core microbiota to be

260   significantly more diverse than the CD cohort, but the sibling core microbiota was significantly less diverse

261   than the healthy core microbiota. No significant difference in diversity was observed between the whole

262   microbiota between the siblings and healthy cohorts, emphasising the advantage of analysing core and rare

263   populations separately. In addition, the CD rare microbiota was significantly less diverse than the other two

264   rare species cohorts, which in turn were not significantly different from each other. All of these

265   observations were underpinned by all three measures of diversity in each instance (Fig. 2).

266   Interestingly, within the CD population, diversity of the whole microbiota was lower in the nine patients

267   with an ileocaecal resection / right hemicolectomy compared with the 11 patients without these operations

268   (as shown by Richness $P<0.0001$; Shannon-Wiener $P=0.046$; but not Simpson's $P=0.768$). This was largely

269   driven by lower diversity of rare taxa (as shown by Richness $P<0.0001$; Shannon-Wiener $P=0.019$; but not

270   Simpson's $P=0.159$) rather than core taxa (Richness $P=0.523$; Simpson's $P=0.612$; Shannon-Wiener

271     $P=0.824$).

272   **Significant divergence in whole and core microbial composition between CD patients and healthy**

273   **controls, but not between CD patients and healthy siblings**

274   The distribution of the microbiota within the three cohorts was determined by direct ordination using Bray-

275   Curtis similarity measures. Using Analysis of Similarities (ANOSIM) tests, the CD and healthy whole and core

276   microbiota were significantly divergent from each other. However, the whole and core microbiota of

277   siblings were not significantly divergent from either that of the CD or healthy controls (Fig. 3). In all

278   instances rare microbiota were significantly divergent between cohorts, including between siblings and

279   healthy controls.

Lower *Faecalibacterium prausnitzii* made the greatest contribution to the dissimilarity in microbiota between both healthy siblings and healthy controls and between CD and healthy controls

Given the involvement of core species in differences of relative abundance, diversity and microbiota composition, the contribution of individual taxa to the dissimilarity between core microbiota was assessed by Similarities of Percentages (SIMPER) analyses (Table 2). Both *F. prausnitzii* and *Escherichia fergusonii* contributed the most to the dissimilarity between all cohorts. As a proportion of core species *F. prausnitzii* had a higher relative abundance in the healthy controls (30.9%) than both the CD (22.4%) and siblings (24.2%). Conversely, *E. fergusonii* was more abundant in the CD cohort (21.4%) than in siblings (9.7%) and healthy controls (4.1%).

17

Table 2 Similarity of Percentages (SIMPER) analysis of microbial community dissimilarity (Bray-Curtis)

291 **between core species groups for (A) CD and siblings, (B) healthy and siblings, and (C) CD and healthy**

292 **cohorts.** Given is mean % abundance of sequences for core species only across the samples each was

293 observed to occupy and the average dissimilarity between samples (overall mean (A) =73.4% and (B)

294 =55.0%, (C) =73.0%). Percentage contribution is the mean contribution divided by mean dissimilarity across

295 samples. The list of species is not exhaustive so cumulative % value does not sum to 100%. Species level

296 identities of detected taxa are reported here. However, given the length of the ribosomal sequences

297 analysed, these identities should be considered putative.

A Name	Crohn's mean abundance	Siblings mean abundance	% Contribution	Cumulative %
Faecalibacterium prausnitzii	22.4	24.2	20.7	20.7
Escherichia fergusonii	21.4	9.7	15.9	36.6
Shigella flexneri	13.6	7.2	10.7	47.3
Ruminococcus gnavus	13.1	5.2	8.9	56.2
Bacteroides vulgatus	13.2	7.6	7.8	64.0
Eubacterium rectale	9.8	6.4	6.6	70.6
Oscillospira guilliermondii	0	8.0	5.9	76.5
Escherichia coli	6.5	0	4.5	81.0
Sutterella wadsworthensis	0	6.0	4.5	85.5
Bacteroides dorei	0	5.7	4.2	89.6
Roseburia faecis	0	4.0	2.9	92.6
B Name	Healthy mean abundance	Siblings mean abundance	% Contribution	Cumulative %
Faecalibacterium prausnitzii	30.9	24.2	18.9	18.9
Escherichia fergusonii	4.1	9.7	10.6	29.5
Sutterella wadsworthensis	8.7	6.0	9.4	38.9
Shigella flexneri	3.6	7.2	8.4	47.3
Bacteroides vulgatus	8.0	7.6	8.4	55.7
Eubacterium rectale	9.9	6.4	7.0	62.8
Oscillospira guilliermondii	8.5	8.0	7.0	69.8
Bacteroides dorei	0	5.7	5.4	75.2
Ruminococcus gnavus	4.1	5.2	4.1	79.3
Bacteroides uniformis	2.9	2.0	3.1	82.4
Roseburia faecis	2.4	4.0	3.0	85.4
Coprococcus eutactus	2.3	0	2.2	87.7
Shigella dysenteriae	2.1	0	2.1	89.8
Blautia producta	2.0	1.8	1.8	91.6
C Name	Crohn's mean abundance	Healthy mean abundance	% Contribution	Cumulative %
Faecalibacterium prausnitzii	22.4	30.9	22.4	22.4
Escherichia fergusonii	21.4	4.1	14.4	36.7
Shigella flexneri	13.6	3.6	9.3	46.0
Ruminococcus gnavus	13.1	4.1	8.6	54.6
Bacteroides vulgatus	13.2	8.0	8.0	62.7
Eubacterium rectale	9.8	9.9	7.4	70.1
Sutterella wadsworthensis	0	8.7	6.4	76.5
Oscillospira guilliermondii	0	8.5	6.2	82.8

18

Escherichia coli	6.5	0	4.4	87.2
Bacteroides uniformis	0	2.9	2.1	89.3
Roseburia faecis	0	2.4	1.8	91.1

298

299

Genotype and phenotypic features associated with CD and CD-risk significantly explained microbiota variation

Canonical correspondence analysis (CCA) was used to relate the variability in the distribution of microbiota between cohorts to clinical and demographic variables (Table 3 and Fig. 4). Variables that significantly explained variation in mucosal microbiota were determined with forward selection (999 Monte Carlo permutations; $P < 0.05$) and used in CCA. Based on the direct ordination approach, the microbiota between cohorts was significantly influenced by factors listed in Table 3. The same analytical approach was used to assess the extent to which variance in the microbiota distribution within cohorts could be accounted for by variation in measures of clinical and demographic factors, (Table 3). GRR was the most significant factor in explaining variance between the three cohorts, but also within each cohort. FC was also significant in explaining variance between cohorts, particularly in the core microbiota. However, in the within-group analyses FC was significant in explaining microbial variance in patients and siblings but not in controls. Blood T-cell factors explained a higher proportion of variance in siblings and controls than in patients. Conversely, age significantly associated with variance in controls but not in patients or siblings.

20

314 **Table 3 Canonical correspondence analyses for determination of percent variation in the whole, core, and rare microbiota between and within the three**

315 **subject cohorts by clinical variables significant at the P < 0.05 level**. * Ileal/Colonic involvement in CD patients used as a factor for the corresponding siblings. n/a

316 denotes not applicable for between cohort or within healthy cohort analyses.

Variable	Between Cohorts			Within Crohn's			Within Siblings			Within Healthy		
	Whole	Core	Rare	Whole	Core	Rare	Whole	Core	Rare	Whole	Core	Rare
Age	-	-	-	-	-	-	-	-	-	5.37	6.08	5.08
Blood concentration of naïve CD4$^+$ T-cells (cells /ml)	-	2.8	-	4.15	3.69	4.54	7.34	8.84	8.30	-	-	-
Calprotectin	1.7	3.4	2.7	4.68	5.36	4.63	4.85	5.70	7.10	-	-	-
Gender	1.8	-	2.4	5.03	-	6.20	6.04	9.12	7.03	6.31	4.57	7.19
Genotype relative risk (GRR)	5.1	4.3	4.7	8.56	6.75	9.11	6.53	12.54	5.64	5.57	6.26	5.34
Ileal/Colonic involvement*	n/a	n/a	n/a	5.50	-	5.87	4.94	6.40	5.81	n/a	n/a	n/a
Proportion of blood T-cells with memory phenotype (%)	2.0	-	2.4	-	-	-	8.71	5.77	9.24	5.19	-	6.87
Proportion of CD4$^+$ naïve T-cells expressing β7 integrin (%)	-	3.6	-	4.85	5.84	5.19	5.30	-	7.81	5.60	10.08	4.15
Undetermined	89.4	85.9	87.8	67.2	78.4	64.5	56.3	51.6	49.1	72.0	73.0	71.4

317 * Ileal/Colonic involvement in Crohn's patients used as a factor for the corresponding Sibling subjects. n/a denotes not applicable for between cohort or within
318 Healthy cohort analyses.

21

319 **DISCUSSION**

320 This is the first study to detail the mucosal microbiota of clinically and genetically well-characterised

321 healthy siblings of CD patients, and to compare them with both their CD-affected siblings and healthy

322 controls. Moreover, this study is unique in uncovering interactions of mucosal microbiota with genotype

323 and features of the CD-risk phenotype. This manuscript is a significant advance on the preliminary account

324 of the multidimensional risk phenotype previously described, which centred on qPCR sampling of faecal

325 microbiota.[14] The current study not only focuses on the mucosal microbiota but also employs next-

326 generation sequencing and advanced statistical analysis to reveal the complexity of the metacommunities

327 in healthy siblings of CD patients. The core mucosal microbiota in siblings was characterised by lower

328 diversity compared with controls, and lower abundance of *F. prausnitzii* made the greatest contribution to

329 the dissimilarity between these two groups. Genetic CD-risk explained the highest proportion of microbial

330 variance both between all three groups, and within the patient and sibling groups. These findings are

331 unlikely to be confounded by cohabitation as only one patient cohabited with one sibling.

332 Although related healthy individuals are known to harbour similar gut microbiota,[19] the similarity in the

333 microbiota between CD patients and their healthy siblings is of considerable pathogenic relevance. Previous

334 studies have shown that when one sibling has CD, familial microbial similarity is disrupted, even in disease-

335 discordant monozygotic twins.[30] Thus, microbial features which are similar between affected and

336 unaffected siblings, but which are not present in low CD-risk healthy individuals, may be part of the CD-risk

337 phenotype and therefore pertinent to CD pathogenesis. In order to discern these features associated with

338 familial risk, comparison with healthy, unrelated individuals is essential.

339 The validity of the data presented is supported by the correlation between species-abundance and

340 distribution, which is consonant with a coherent metacommunity structure and is similar to distributions

341 described in other ecological communities.[15] This feature of community structure facilitated delineation

342 of core species which are abundant and persistent, and allowed resolution of features of the mucosal

343 microbiota without obfuscation from rare microbiota which may be highly variable, transient and scarce. A

22

344   significantly higher proportion of the microbiota in CD patients belonged to the rare group compared with

345   healthy siblings and healthy controls. As described below this is at least in part attributable to loss of

346   principal members of the core group, most notably Firmicutes.

347   Reduced microbial diversity is an almost universally reported feature of mucosal CD dysbiosis.[1] The

348   current study reveals that core microbiota diversity is also lost in siblings of CD patients, indicating that this

349   may be a fundamental step in CD pathogenesis. Reduced diversity may be an indicator of the health of

350   human microbial communities, as it is reduced in a variety of disorders.[18-21] Lower diversity may be

351   associated with incomplete occupation of ecological niches resulting in reduced resistance to pathogen

352   colonisation; additionally a more restricted gut metagenome contains a lower array of genes which may

353   result in the loss of key functions.

354   Lower diversity indicates altered mucosal microbial composition, and microbial composition in CD patients

355   and healthy controls were significantly distinct from one another. In contrast, the composition of the whole

356   and core microbiota in healthy siblings was not significantly different from either CD patients or healthy

357   controls, indicating that from a microbial metacommunity perspective, siblings lie somewhere between

358   patients and controls. The greater variability in the composition of the microbiota in at-risk siblings

359   (illustrated by larger 95% concentration ellipse in Figure 3 (panel B)) probably reflects the range of CD-risk

360   contained within this group, with siblings with higher CD-risk lying closer to or within the CD region. In

361   addition, diversity was lower in core and rare microbiota in patients with ileocaecal resection/ right

362   hemicolectomy, potentially explained by differences in disease phenotype, or the absence of the ileocaecal

363   valve that would otherwise constitute a barrier between small and large intestinal microbiota.

364   Consonant with previous work highlighting the importance of *F. prausnitzii* in CD dysbiosis,[7, 12, 14] *F.*

365   *prausnitzii* made the greatest contribution to the dissimilarity between CD patients and healthy control

366   microbiota. The prominence of *F. prausnitzii* has biological significance as it is the only microbial factor

367   shown to be predictive of the natural history of CD,[7] and response to treatment.[31] Strikingly, *F.*

23

368 *prausnitzii* was also the biggest contributor to the dissimilarity of the core mucosal microbiota between

369 healthy siblings and healthy controls, establishing that mucosal *F. prausnitzii* not only correlates to the

370 natural history of CD, but is also a key feature of the at-risk phenotype. Taken together these findings

371 strongly support the hypothesis that depletion of *F. prausnitzii* is part of CD pathogenesis rather than

372 consequent to established CD. Several mechanisms exist whereby *F. prausnitzii* and other Firmicutes may

373 contribute to gut health, including the production of short-chain fatty acids (SCFAs),[32, 33] SCFA-

374 independent, NFκB-mediated effects,[7] and via production of longer-chain fatty acids such as conjugated

375 linoleic acid.[34]

376 The pathogenic role of reduced *F. prausnitzii* in CD has been questioned by a study describing increased

377 mucosal *F. prausnitzii* in newly-diagnosed pediatric IBD.[35] However, whether increased abundance of *F.*

378 *prausnitzii* is a distinctive feature of pediatric-onset IBD, with low *F. prausnitzii* being associated with later-

379 onset CD, or whether the abundance of *F. prausnitzii* may bloom in childhood and then critically decline in

380 those at risk of CD, may only be determined by longitudinal studies.

381 Other species contributing to the dissimilarity in the core mucosal microbiota between CD patients and

382 healthy controls were congruent with species previously identified as characterising the CD dysbiosis,

383 including a greater abundance of most Proteobacteria such as *E. fergusonii* and *Escherichia coli*. Similar

384 species contributed to the dissimilarity between siblings and controls. However, the presence of *E. coli* was

385 specific to CD mucosa, and therefore may be a feature of established CD rather than pathogenic. Features

386 of the inflamed gut such as increased activity of nitric oxide synthases[36], or reduction in faecal butyrate

387 producers which will result in a rise in pH, potentially favour the survival of organisms that are inhibited at

388 acidic pH such as *E. coli*.[37]

389 GRR was the factor associated most strongly with the variation in the microbiota in both the between-

390 group analysis, and analysis within each of the three groups. Although the proportion of variation in

391 mucosal microbiota explained by GRR was small, it is nevertheless significant. The combination of loci used

24

392   to estimate GRR in the current study does not include more recently detected risk loci and can be expected

393   to account for a limited proportion of the genetic risk.[38] Therefore, these data will tend to underestimate

394   the effect of genotype. Furthermore, since other factors known to affect gut microbiota such as diet were

395   not controlled, this signal of the interaction between genotype and the mucosal microbiota is striking.

396   The direction of the vector in figure 3 illustrates that FC contributed to the axis separating patients from the

397   other two groups in the whole, core and rare microbiota, implying that microbial composition in CD is

398   partly associated with the degree of inflammation. This would support the hypothesis that CD-specific

399   elements of the dysbiosis may be consequent to intestinal inflammation, through mechanisms such as the

400   enhanced survival of *E. coli* in an inflamed environment as proposed above.

401   When each group was considered separately, the effect of each factor in different groups could be

402   compared. Several factors were significant in all groups (GRR, gender, proportion of CD4$^+$ naïve T-cells

403   expressing β7 integrin). Other factors were significant in patients and siblings but not controls: FC and

404   blood naïve CD4$^+$ T-cell concentration were significant only in patients and siblings, whereas age was

405   significant only in controls. Disease phenotype was significant in explaining microbial variation within the

406   CD group as would be predicted from previous studies.[30] However; we have also demonstrated that for

407   healthy siblings, disease site in their affected relative was significantly associated with the variation in their

408   own microbiota. This would suggest that specific risk phenotypes are associated with different disease

409   phenotypes.

410   Overall these factors accounted for a higher proportion of the variance in the microbial composition in

411   siblings, compared with controls or patients, indicating that this multidimensional risk phenotype is specific,

412   and that in low CD-risk individuals the microbial composition is associated with other factors, such as age.

413   Furthermore, it would appear that in CD the influence of factors associated with the original risk phenotype

414   is obfuscated by established CD and its surgical and medical management.

415   **CONCLUSION**

25

416   Healthy siblings of CD patients, who themselves have elevated risk of CD, have a dysbiosis of the core

417   mucosal microbiota characterised by reduced diversity and loss of Firmicutes, notably *F. prausnitzii*.

418   Genotype determines a proportion of the at-risk mucosal microbial phenotype. Notwithstanding the

419   limited extent to which known loci account the observed CD-risk,[39] it is also clear that the sibling risk

420   goes beyond genotype and that non-genetic factors within families contribute to the development of an at-

421   risk microbiota. How and why patients and their siblings acquire the microbiota that marks out this risk is

422   not known. However, knowledge of the at-risk microbial phenotype illuminates possible pathways in CD

423   pathogenesis and raises the prospect of intervention to impact human health and influence disease risk.

428   **Competing interest statement:** None declared

429

FIGURE LEGENDS

431 Fig. 1. The distribution and abundance of bacterial species within microbiota samples within the (a) CD, (b)

432 siblings, and (c) healthy control cohort metacommunities. Given is the number of mucosal samples for

433 which each bacterial taxon was observed to occupy, plotted against the mean abundance across all samples

434 ((a) n =21, r2 =0.62, F1, 227 =366.9, P < 0.0001; (b) n =17, r2 =0.71, F1, 259 =590.1, P < 0.0001; and (c) n

435 =19, r2 =0.68, F1, 258 =552.6, P < 0.0001). Core species were defined as those that fell within the upper

436 quartile (dashed lines), and rare species defined as those that did not.

437

438 Fig. 2. Diversity of whole, core and rare microbiota within the CD (black columns), siblings (grey), and

439 healthy (white) control cohorts. Given are three indices of diversity; Species richness (S^*), Simpson's index

440 of diversity (1-D), and Shannon-Wiener index of diversity *(H')*. Error bars represent the standard deviation

441 of the mean (CD *n* =21, siblings *n* =17, and healthy *n* =19). Asterisks denote significant differences in

442 comparisons of diversity at the P < 0.05 level determined by two sample *t*-tests.

443

444 Fig. 3. Analysis of similarities (ANOSIM) of whole, common, and rare microbiota between subject cohorts.

445 Given is the ANOSIM test statistic (*R*) and probability (*P*) that two compared groups are significantly

446 different at the *P* < 0.05 level (* denotes *P* < 0.001 and ** *P* < 0.0001). ANOSIM *R* and *P* values were

447 generated using the Bray-Curtis measure of similarity. *R* scales from +1 to -1. +1 indicates that all the most

448 similar samples are within the same groups. *R* = 0 occurs if the high and low similarities are perfectly mixed

449 and bear no relationship to the group. A value of -1 indicates that the most similar samples are all outside

450 of the groups.

451

452   Fig. 4. Canonical correspondence biplots for (a) whole, (b) core, and (c) rare microbiota. Red crosses

453   represent microbiota samples from the CD cohort, yellow filled triangles for the siblings cohort, and green

454   diamonds for the healthy cohort. In each instance, the 95 % concentration ellipses are given for the CD

455   (red), siblings (yellow), and healthy (green) cohort microbiota. Biplot lines for clinical variables that

456   significantly accounted for variation within the microbiota at the $P < 0.05$ level (see Table 3) show the

457   direction of increase for each variable, and the length of each line indicates the degree of correlation with

458   the ordination axes. CCA field labels: Calprotectin, Gender, "% Memory T-cells" – Proportion of blood T-cell

459   with memory phenotype (%), "CD4$^+$ T-cells" – Blood concentration of naïve CD4$^+$ T-cells (cells /ml), "β7

460   integrin" – Proportion of CD4 naïve T-cells expressing β7 integrin (%),"GRR" – genotype relative risk,

461   (cumulative genotype relative risk (GRR) for each participant was calculated across 72 CD-risk loci (detected

462   using the Illumina Infinium Immunochip), participants were categorised into reduced, average, elevated or

463   high genotype risk with reference to a population distribution model of CD-risk). Percentage of community

464   variation explained by each axis is given in parentheses.

465

466   **REFERENCE LIST**

467   1. Kostic AD, Xavier RJ, Gevers D. The microbiome in inflammatory bowel disease: current status and the

468   future ahead. *Gastroenterology* 2014;146:1489-1499.

469   2. Franke A, McGovern DP, Barrett JC *et al*. Genome-wide meta-analysis increases to 71 the number of

470   confirmed Crohn's disease susceptibility loci. *Nat Genet* 2010;42:1118-25

471   3. Anderson CA, Boucher G, Lees CW *et al.* Meta-analysis identifies 29 additional ulcerative colitis risk loci,

472   increasing the number of confirmed associations to 47. Nat Genet. 2011;43:246-52.

473   4. Jostins L, Ripke S, Weersma RK, *et al.* Host-microbe interactions have shaped the genetic architecture of

474   inflammatory bowel disease. Nature. 2012;491:119-24.

475   5. Sellon RK, Tonkonogy S, Schultz M *et al*. Resident enteric bacteria are necessary for development of

476   spontaneous colitis and immune system activation in interleukin-10-deficient mice. *Infect Immun*

477   1998;66:5224-5231.

478   6. Harper PH, Lee EC, Kettlewell MG *et al*. Role of the faecal stream in the maintenance of Crohn's colitis.

479   *Gut* 1985;26:279-284.

480   7. Sokol H, Pigneur B, Watterlot L *et al*. *Faecalibacterium prausnitzii* is an anti-inflammatory commensal

481   bacterium identified by gut microbiota analysis of Crohn disease patients. *Proc Natl Acad Sci U S A*

482   2008;105:16731-16736.

483   8. Ghouri YA, Richards DM, Rahimi EF *et al.* Systematic review of randomized controlled trials of probiotics,

484   prebiotics, and synbiotics in inflammatory bowel disease. Clin Exp Gastroenterol. 2014;9:473-87.

485   9. Benjamin JL, Hedin CR, Koutsoumpas A *et al*. Randomised, double-blind, placebo-controlled trial of

486   fructo-oligosaccharides in active Crohn's disease. *Gut* 2011;60:923-929.

487   10. Calkins BM, Mendeloff AI. Epidemiology of inflammatory bowel disease. *Epidemiol Rev* 1986;8:60-91.

30

488   11. Halfvarson J. Genetics in twins with Crohn's disease: less pronounced than previously believed?

489   *Inflamm Bowel Dis* 2011;17:6-12.

490   12. Hedin CR, Stagg AJ, Whelan K *et al*. Family studies in Crohn's disease: new horizons in understanding

491   disease pathogenesis, risk and prevention. *Gut* 2012;61:311-318.

492   13. Joossens M, Huys G, Cnockaert M *et al*. Dysbiosis of the faecal microbiota in patients with Crohn's

493   disease and their unaffected relatives. *Gut* 2011;60:631-637.

494   14. Hedin CR, McCarthy NE, Louis P *et al*. Altered intestinal microbiota and blood T cell phenotype are

495   shared by patients with Crohn's disease and their unaffected siblings. *Gut* 2014;63:1578-1586.

496   15. van der Gast CJ, Walker AW, Stressmann FA *et al*. Partitioning core and satellite taxa from within cystic

497   fibrosis lung bacterial communities*. ISME J* 2011;5:780-791.

498   16. Sogin ML, Morrison HG, Huber JA *et al*. Microbial diversity in the deep sea and the underexplored "rare

499   biosphere". *Proc Natl Acad Sci U S A* 2006;103:12115-12120.

500   17. Li K, Bihan M, Yooseph S *et al*. Analyses of the microbial diversity across the human microbiome. *PLoS

501   One* 2012;7:e32118.

502   18. Le CE, Nielsen T, Qin J *et al*. Richness of human gut microbiome correlates with metabolic markers.

503   *Nature* 2013;500:541-546.

504   19. Turnbaugh PJ, Hamady M, Yatsunenko T *et al*. A core gut microbiome in obese and lean twins. *Nature*

505   2009;457:480-484.

506   20. Ahn J, Sinha R, Pei Z *et al*. Human gut microbiome and risk for colorectal cancer. *J Natl Cancer Inst*

507   2013;105:1907-1911.

508   21. Ismail IH, Oppedisano F, Joseph SJ *et al*. Reduced gut microbial diversity in early life is associated with

509   later development of eczema but not atopy in high-risk infants. *Pediatr Allergy Immunol* 2012;23:674-681.

31

510   22. Biedermann L, Zeitz J, Mwinyi J *et al*. Smoking cessation induces profound changes in the composition

511   of the intestinal microbiota in humans. *PLoS One* 2013;8:e59260.

512   23. Cortes A, Brown MA. Promise and pitfalls of the Immunochip. *Arthritis Res Ther* 2011;13:101.

513   24. Crouch DJ, Goddard GH, Lewis CM. REGENT: a risk assessment and classification algorithm for genetic

514   and environmental factors. *Eur J Hum Genet* 2013;21:109-111.

515   25. Goddard GH, Lewis CM. Risk categorization for complex disorders according to genotype relative risk

516   and precision in parameter estimates. *Genet Epidemiol* 2010;34:624-632.

517   26. Rogers GB, Cuthbertson L, Hoffman LR *et al*. Reducing bias in bacterial community analysis of lower

518   respiratory infections. *ISME J* 2013;7:697-706.

519   27. Rogers GB, Kozlowska J, Keeble J *et al*. Functional divergence in gastrointestinal microbiota in physically-

520   separated genetically identical mice. *Sci Rep* 2014;4:5437.

521   28. Dowd SE, Wolcott RD, Sun Y, *et al.* Polymicrobial nature of chronic diabetic foot ulcer biofilm infections

522   determined using bacterial tag encoded FLX amplicon pyrosequencing (bTEFAP). *PLoS One* 2008;3:e3326.

523   29. R Core Team.R Foundation for Statistical Computing. Vienna A. R: A Language and Environment for

524   Statistical Computing. 2014.

525   30. Willing B, Dicksved J, Halfvarson J *et al*. A pyrosequencing study in twins shows that GI microbial profiles

526   vary with inflammatory bowel disease phenotypes. *Gastroenterology* 2010;39:1844-1854.

527   31. Rajca S, Grondin V, Louis E *et al.* Alterations in the intestinal microbiome (dysbiosis) as a predictor of

528   relapse after infliximab withdrawal in Crohn's disease. *Inflamm Bowel Dis*. 2014;20:978-86.

529   32. Macfarlane GT, Macfarlane S. Fermentation in the human large intestine: its physiologic consequences

530   and the potential contribution of prebiotics. *J Clin Gastroenterol* 2011;45 Suppl:S120-S127.

32

531   33. Vinolo MA, Rodrigues HG, Nachbar RT *et al*. Regulation of inflammation by short chain fatty acids.

532   *Nutrients* 2011;3:858-876.

533   34. Bassaganya-Riera J, Hontecillas R, Horne WT *et al*. Conjugated linoleic acid modulates immune

534   responses in patients with mild to moderately active Crohn's disease. *Clin Nutr* 2012.

535   35. Hansen R, Russell RK, Reiff C *et al*. Microbiota of de-novo pediatric IBD: increased Faecalibacterium

536   prausnitzii and reduced bacterial diversity in Crohn's but not in ulcerative colitis. *Am J Gastroenterol*

537   2012;107:1913-1922.

538   36. Rachmilewitz D, Karmeli F, Okon E *et al*. Experimental colitis is ameliorated by inhibition of nitric oxide

539   synthase activity. *Gut* 1995;37:247-255.

540   37. Duncan SH, Louis P, Thomson JM *et al*. The role of pH in determining the species composition of the

541   human colonic microbiota. *Environ Microbiol* 2009;11:2112-2122.

542   38. Park JH, Wacholder S, Gail MH *et al*. Estimation of effect size distribution from genome-wide association

543   studies and implications for future discoveries. *Nat Genet* 2010;42:570-575.

544   39. Manolio TA, Collins FS, Cox NJ *et al*. Finding the missing heritability of complex diseases. *Nature*

545   2009;461:747-753.
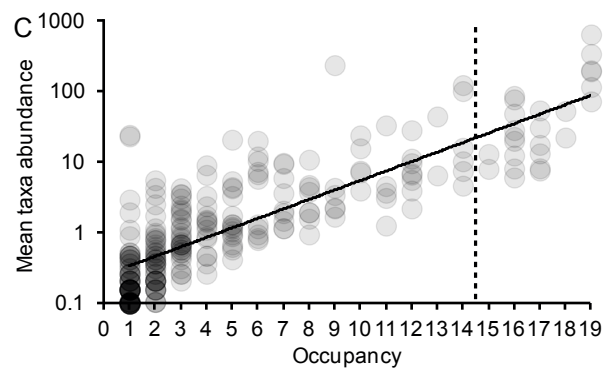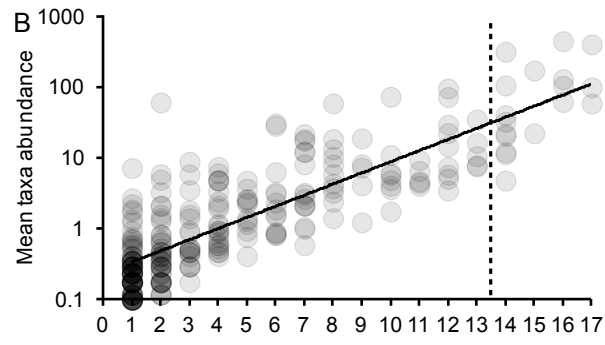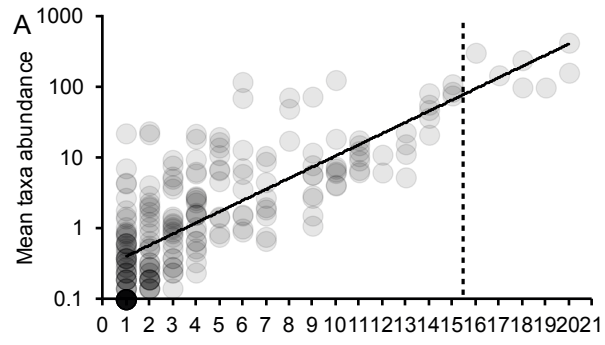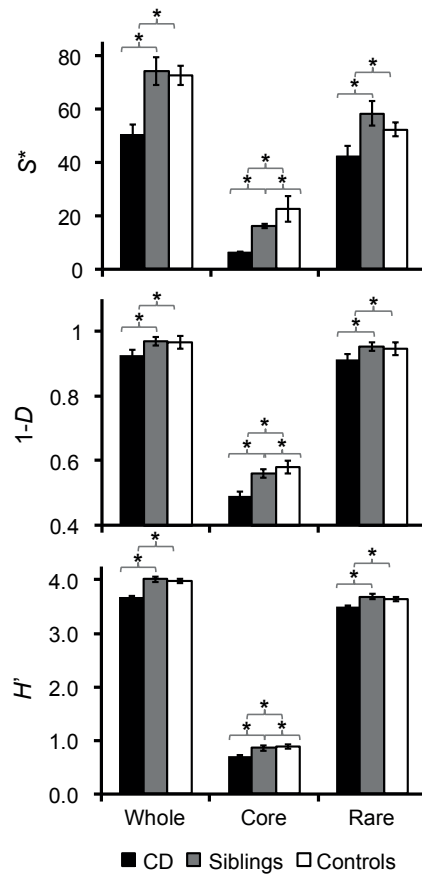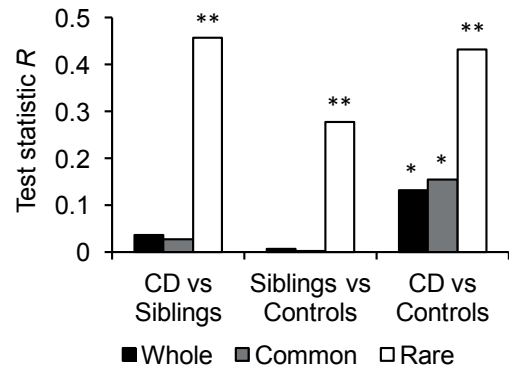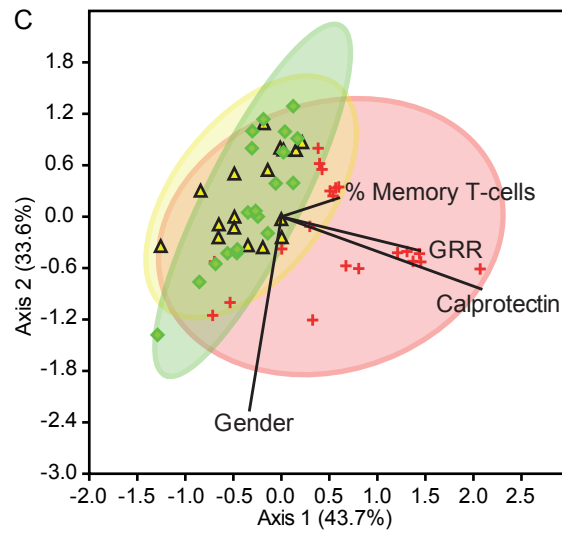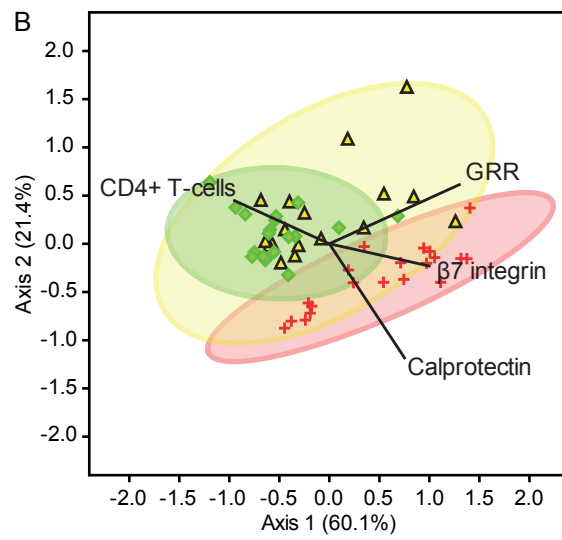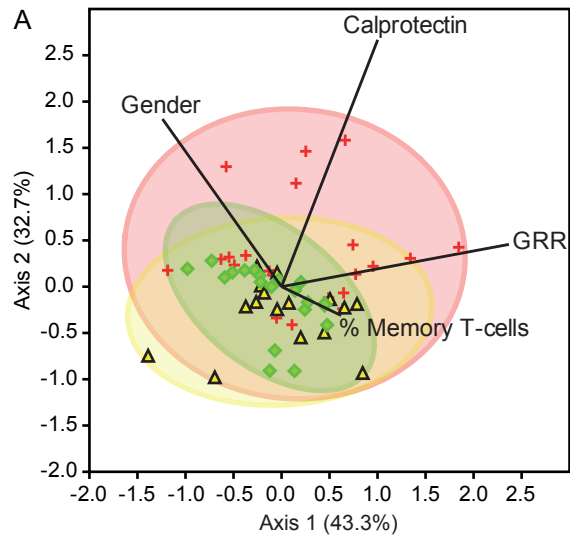
Supplementary table S1. List of exclusion criteria by group

All participants	Patients	Siblings	Controls
Unable to consent due to mental illness/ dementia/ learning disability	Evidence of active CD as defined by a CDAI of greater than 150	Previous diagnosis of IBD	Previous diagnosis of IBD
Current infection with an enteric pathogen	Purely perianal CD	Symptoms fulfilling Rome III criteria for IBS	Symptoms fulfilling Rome III criteria for IBS
Use of antibiotics within the last month	Change in dose of oral steroids within the last 4 weeks		A first or second degree relative with IBD
Consumption of any probiotic or prebiotic within the last month	Dose of steroids exceeding 10mg prednisolone per day or equivalent		
Pregnancy or lactation	Change in dose of oral 5-ASA products within the last 4 weeks		
Participant requiring hospitalization	Commencement of azathioprine or methotrexate within the last 4 months, or change in dose of these drugs within the last 4 weeks		
Significant hepatic, renal, endocrine, respiratory, neurological or cardiovascular disease as determined by the principal investigator	Infusion of biological therapies (e.g. infliximab) within the last 3 months*		
A history of cancer with a disease free state of less than two years	Use of rectal 5-ASA or steroids within the last 2 weeks		
CRP greater than 5mg/L at screening, as measured by the local laboratory	Use of NSAIDs within the last 2 weeks		
	Imminent need for surgery		
	Short bowel syndrome		
	Previous proctocolectomy		

* No patient had been previously exposed to biological therapies

Supplementary methods

Peripheral blood T-cell flow cytometry

The fluorescently labeled antibodies used were: anti-CD3 Pacific Blue (clone OKT3, Biolegend, San Diego, CA,USA), anti-CD45RA PE-Cy7 (clone L48, BD Bioscience, Franklin Lakes, NJ, USA), anti-CD8 PerCP-Cy5.5 (clone SK1, BD Bioscience) and anti-CD4 APC (clone RPA-T4, BD Bioscience), anti-β7 PE (clone FIB504, BD Pharmingen).

Isotype-matched controls for mIgG1κ PE-Cy7 (clone MOPC-21, BD Pharmingen), rat IgG2a PE (clone R35-95, BD Bioscience), mIgG1 PE (clone MOPC-21, BD Bioscience), rIgM FITC (clone R4-22, BD Pharmingen) and mIgG1 FITC (clone MOPC-21, BD Pharmingen) were used to set positive and negative regions for gating during analysis. Anti-CD8 FITC (clone LT8, AbD Serotec, Kidlington, UK), anti-β7 PE (clone FIB504, BD Pharmingen), anti-CD8 PerCP-Cy5.5 (clone SK1, BD Bioscience), anti-CD45RA PE-Cy7 (clone L48, BD Bioscience) anti-CD3 PB (clone OKT3, Biologend) and anti-CD4 APC (clone RPA-T4, BD Bioscience) conjugated antibodies were used for off-line compensation.

Gut mucosal microbiota

DNA extraction protocol

Biopsy DNA extraction was carried out using a phenol/chloroform based method, as follows: Guanidinium thiocyanate–EDTA–sarkosyl (500 μL) and PBS (500 μL), pH 8.0, were added to biopsy samples. Cell disruption was achieved using a Fastprep-24 Instrument (MP Biomedicals Europe, Illkirch, France) 6.5 m/s, 60 s, followed by incubation at 90 °C for 1 min and −20 °C for 5 min. Cell debris was pelleted by centrifugation at 12 000 × g for 2 min at 4 °C. Supernatant was transferred to a fresh microfuge tube. NaCl (to a final concentration of 0.5 mol/L and polyethylene glycol (to a final concentration of 15%) were added and DNA precipitated at 4 °C for 30 min. DNA was pelleted by

centrifugation at 12 000 × g for 2min at 4 °C and resuspended in 300 µL of sterile distilled water. Samples were heated at 90 °C for 30 s and vortexed. Phenol/chloroform (1:1) (300 µL) was added, and samples were vortexed for 20 s before centrifugation at 12 000 × g at 4 °C for 3min. The upper phase was then transferred to a fresh microfuge tube. Total DNA was then precipitated by the addition of an equal volume of isopropanol, a 0.1-volume 10 mol/L ammonium acetate, and 1 µL of GenElute linear polyacrylamide (Sigma-Aldrich, Gillingham, UK) and incubated at −20 °C for 25 min. DNA was pelleted by centrifugation at 12 000 × g at 4 °C for 5 min. Pelleted DNA was then washed 3 times in 70% ethanol, dried, and resuspended in 50 µL of sterile distilled water. DNA extracts were quantified using the Picodrop Microlitre Spectrophotometer (GRI, Braintree, UK). Negative controls, consisting of sterile water, were included in the PMA treatment, DNA extraction, and PCR amplification steps.

16S rRNA gene sequencing

Bacterial tag-encoded FLX amplicon pyrosequencing (bTEFAP) was performed as described previously using Gray28F 5'-TTTGATCNTGGCTCAG-3' and Gray519r 5'-GTNTTACNGCGGCKGCTG-3').[1] A single-step 30 cycle PCR using HotStarTaq Plus Master Mix Kit (Qiagen, Valencia, CA) performed under the following conditions: 94°C for 5minutes, followed by 28 cycles of: 94°C for 30 seconds, 53°C for 40 seconds, and 72°C for 1 minute. Amplification was followed by a final elongation step at 72°C for 5 minutes. Following PCR, all amplicon products from different samples were mixed in equal concentrations and purified using Agencourt Ampure beads (Agencourt Bioscience Corporation, MA, USA). Samples were sequenced utilizing Roche 454 FLX titanium instruments and reagents following manufacturer's guidelines.

Sequence data analysis was carried out. Here, the Q25 sequence data derived from the sequencing process was processed using standard analysis pipeline processes (MR DNA, Shallowater, USA). Sequences were depleted of barcodes and primers then short sequences, 200 bp removed, as were sequences with ambiguous base calls removed, and sequences with homopolymer runs exceeding 6

bp, sequences were denoised and chimeras removed.[2-8] Operational taxonomic units were defined

after removal of singleton sequences, clustering at 3% divergence (97% similarity). Final OTUs were

taxonomically classified using BLASTn against a curated databased derived fromGreenGenes, NCBI

and RDP databases.[9] Normalized and de-noised files were then rarefied and run through QIIME[10] to

generate alpha and beta diversity data. Additional statistical analyses were performed with

NCSS2007 (NCSS, UT) and XLstat 2012 (Addinsoft, NY).

References for supplementary methods

1. Rogers GB, Kozlowska J, Keeble J, *et al.* Functional divergence in gastrointestinal microbiota in physically-separated genetically identical mice. Sci Rep. 2014:4;5437.

2. Dowd SE, Callaway TR, Wolcott RD, *et al.* Evaluation of the bacterial diversity in the feces of cattle using 16S rDNA bacterial tag-encoded FLX amplicon pyrosequencing (bTEFAP). BMC Microbiol. 2008:8;125.

3. Dowd SE, Sun Y, Wolcott RD, *et al.* Bacterial tag-encoded FLX amplicon pyrosequencing (bTEFAP) for microbiome studies: bacterial diversity in the ileum of newly weaned Salmonella-infected pigs. Foodborne Pathog. Dis. 2008:5;459–72.

4. Edgar RC. Search and clustering orders of magnitude faster than BLAST. Bioinformatics 2010:26;2460–1.

5. Capone KA, Dowd SE, Stamatas GN, *et al.* Diversity of the human skin microbiome early in life. J. Invest. Dermatol. 2011:131;2026–32.

6. Dowd SE, Delton Hanson J, Rees E, *et al.* Survey of fungi and yeast in polymicrobial infections in chronic wounds. J. Wound Care 2011:20;40–7.

7. Eren AM, Zozaya M, Taylor CM, *et al.* Exploring the diversity of Gardnerella vaginalis in the genitourinary tract microbiota of monogamous couples through subtle nucleotide variation. PLoS One 2011:6;e26732;

8. Swanson KS, Dowd SE, Suchodolski JS, *et al.* Phylogenetic and gene-centric metagenomics of the canine intestinal microbiome reveals similarities with humans and mice. ISME J. 2011:5;639–49.

9. DeSantis TZ, Hugenholtz P, Larsen N, *et al.* Greengenes, a chimera-checked 16S rRNA gene database and workbench compatible with ARB. Appl. Environ. Microbiol. 2006:72;5069–72.

10. Caporaso JG, Kuczynski J, Stombaugh J, *et al.* QIIME allows analysis of high-throughput community sequencing data. Nat Methods. 2010;7:335-6.