Molecular analysis of the gut microbiota of identical twins with Crohn's disease

Johan Dicksved[1*], Jonas Halfvarson[2*], Magnus Rosenquist[1], Gunnar Järnerot[2], Curt Tysk[2,5], Juha Apajalahti[3], Lars Engstrand[4], and Janet K. Jansson[1,6]

*Authors contributed equally to this work

[1] Swedish University of Agricultural Sciences, Department of Microbiology, Uppsala, Sweden

[2] Division of Gastroenterology, Department of Medicine, Örebro University Hospital, Örebro, Sweden.

[3] Alimetrics Ltd, Höyläämötie 14, 00380 Helsinki, Finland

[4] Swedish Institute for Infectious Disease Control, Department of Bacteriology, Solna, Sweden

[5] School of Health and Medical Sciences, Örebro University, Örebro, Sweden.

[6] Lawrence Berkeley National Laboratory, Division of Earth Sciences, 1 Cyclotron Rd., Berkeley, CA 94720

Keywords, Crohn's disease, terminal-restriction fragment length polymorphism (T-RFLP), discordant twins, concordant twins, *Bacteroides*

Running title. Twin study of Crohn's disease

ABSTRACT

Increasing evidence suggests that a combination of host genetics and the composition of the gut microbiota are important for development of Crohn's disease (CD). Our aim was to study identical twins with CD to determine microbial factors independently of host genetics. Fecal samples were studied from 10 monozygotic twin pairs with CD (discordant n=6, concordant n=4) and 8 healthy twin pairs. DNA was extracted, 16S rRNA genes were PCR amplified and T-RFLP fingerprints generated using general bacterial and *Bacteroides* group specific primers. The microbial communities were also profiled based on their % G+C contents. *Bacteroides* 16S rRNA genes were cloned and sequenced from a subset of the samples. The bacterial diversity in each sample and similarity indices between samples were estimated based on the T-RFLP data using a combination of statistical approaches. Healthy individuals had a significantly higher bacterial diversity compared to individuals with CD. The fecal microbial communities were more similar between healthy twins than between twins with CD, especially when these were discordant for the disease. The microbial community profiles of individuals with ileal CD were significantly different from healthy individuals and those with colonic CD. Also, CD individuals had a lower relative abundance of *B. uniformis* and higher relative abundances of *B. ovatus* and *B. vulgatus*. Our results suggest that genetics and/or environmental exposure during childhood in part determine the gut microbial composition. However, CD is associated with dramatic changes in the gut microbiota and this was particularly evident for individuals with ileal CD.

INTRODUCTION

Crohn's disease (CD) is a chronic relapsing inflammatory disorder of the gastrointestinal tract with an unknown etiology. Available data suggests that inflammation occurs due to an imbalanced mucosal immune response to the commensal bacteria in genetically susceptible individuals (Sartor *et al.*, 2006).

The knowledge about genetic factors that are relevant for CD has increased considerably during recent years and several susceptibility genes have been associated with CD. For example, polymorphisms in pattern recognition receptors, such as CARD15/NOD2 that recognize microbial components, have highlighted the importance of the microbiota in pathogenesis of CD (Sartor *et al.*, 2006). The genetic influence is also supported by higher concordance rates (approximately 50%) for CD occurrence in monozygotic twins (Halfvarson *et al*., 2003; Jess *et al*., 2005; Orholm *et al*., 2000; Tysk *et al*., 1988). Still, approximately 50% of identical twin pairs are discordant for CD (i.e. one is diseased and one is healthy) demonstrating that environmental factors are also important for disease incidence (Halfvarson *et al*., 2006; Loftus *et al*., 2004).

It has been difficult to correlate specific causative bacterial agents to CD. An increased prevalence of mucosal bacteria has been observed in CD patients, with higher levels of *E. coli* and *Bacteroides* species (Keighley *et al*., 1978; Swidsinski *et al*., 2002; Swidsinski *et al*., 2005). Representatives of these bacteria have also been demonstrated to induce colitis when inoculated into germ free animals, however, with conflicting results (Sartor et al., 2003). Recently, increased levels of adherent, invasive *E. coli* (AIEC) were found in ileal tissues of CD patients (Barnich and Darfeuille-Michaud 2007; Darfeuille-Michaud *et al*., 1998). Furthermore, there have been reports of reduced numbers and a lower diversity of *Firmicutes* in the gut microbiota in individuals with CD (Gophna *et al*., 2006; Manichanh *et al*., 2006). However, the total microbiota of a patient with CD has not yet been profiled to a degree

where its composition can be considered predictive for disease development, and specific bacterial species responsible for the bowel inflammation have not yet been identified.

The search for a causative disease agent is complicated by the great individuality of the gut microbiota with little overlap between individuals (Dicksved *et al*., 2007; Eckburg *et al*., 2005; Zoetendal *et al*., 1998). However, previous findings have shown that there is a high similarity in the composition of the fecal microbial communities in monozygotic twins (Stewart *et al*., 2005; Van de Merwe *et al*., 1983; Zoetendal *et al*., 2001). Therefore, one way to unravel the respective contributions of host genetics and commensal bacteria towards CD development and establishment would be to study identical twins.

The aims of this study were 1) to determine if the gut microbiota in healthy twins have a higher degree of similarity than in twins that have CD and 2) to determine whether there are differences in the composition of the gut microbiota in individuals that have CD compared to healthy individuals. In particular, we focused on a set of discordant twin pairs, where one individual is healthy and one has CD, because they provided each other's genetically matched control thus enabling us to focus on changes in the gut microbiota according to disease state.

To test these hypotheses, we used molecular approaches to provide community profiles of the fecal microbiota. By focusing on the nucleic acid composition of the gut microbiome, we were able to overcome the known biases with cultivation-based approaches. Current estimates are that only 20% of the total species residing in the human gut have been cultivated to date (Eckburg *et al*., 2005). Therefore, we are still greatly limited in our knowledge about the physiology and ecology of the majority of the gut microbiota.

In this study, we used the molecular fingerprinting approach, terminal-restriction fragment length polymorphism (T-RFLP), to monitor the bacterial community architecture in concordant and discordant identical twins with CD, and healthy twins. In addition, we used percent guanine + cytosine (%G+C) profiling of the total bacterial microbiome as a

complementary approach. Furthermore, we aimed to identify members of the microbiota that could be linked to CD incidence or development. We specifically targeted members of the *Bacteroides* genera since these are dominant members of the commensal biota with functionally important roles in the gut. Some *Bacteroides* species have previously been shown to be present in increased levels in persons suffering from IBD (Gophna *et al*., 2006; Swidsinski *et al*., 2005), yet animal models have provided conflicting evidence as to which species may cause CD, warranting closer examination.

MATERIAL AND METHODS

Patient cohort

The twins with CD were derived from a Swedish twin population, described previously (Halfvarson *et al*., 2003; Halfvarson *et al*., 2004; Tysk *et al*., 1988). In short, twin pairs where at least one twin in each pair had been hospitalized for IBD, were identified by running the Swedish twin registry against the Swedish Hospital Discharge Register. All twins were sent a questionnaire concerning diagnosis of IBD, general gastrointestinal symptoms and exposure to environmental factors. After consent from each twin, the medical notes of all twins were evaluated, to verify or refute the diagnosis of IBD and to characterize the disease according to the Montreal classification (Satsangi *et al*., 2006). Zygosity was assessed by a questionnaire-based method, applied by the Swedish twin registry (Cederlöf *et al*., 1961). It relies on questions on childhood resemblance and has been shown to be very accurate (Lichtenstein *et al*., 2002). Monozygotic twin pairs with CD born between 1936 and 1986, who had approved further contact and had not undergone extensive CD related surgical resections, i.e. colectomy, were invited to undergo colonoscopy. Ten monozygotic twin pairs were studied, six of these were discordant and four were concordant for CD, for a total of 14 individuals

with CD. Data on age, disease location, disease duration, behavior at diagnosis in the CD twins is presented in Table 1. All diseased twins, except two (labeled; 10b and 15a), were in clinical remission according to the Harvey Bradshaw score (Harvey and Bradshaw 1980). All remaining twins were in endoscopic remission or had only post-inflammatory changes. All twins were asked to send fecal samples 7-10 days prior to the colonoscopy. In addition they submitted responses to a questionnaire regarding, usage of antibiotics or Non-Steroid Anti-inflammatory Drugs (NSAIDs) within the last 12 months, gastroenteritis within the last three months or specific dietary habits and this information is provided in Supplementary Table S1.

Table 1. Clinical characteristics of the twins with Crohn's disease according to the Montreal classification, (n=14)

	CD twins (n=14)
Mean age (y)	49 (20-70)
Age at diagnosis	
< 40 years	9
≥ 40 years	5
Location	
Terminal ileum	5
Colon	6
Ileocolon	2
Ileocolon + Upper GI	1
Behavior	
Non-stricturing non-penetrating	11
Stricturing	2
Penetrating	1
Perianal disease	0
Median (range) Harvey Bradshaw score	1.5 (1-6)

Abbreviations: CD, Crohn's Disease; GI, gastrointestinal

Eight healthy twin pairs, five monozygotic and three dizygotic pairs, not suffering from any gastrointestinal disease, were also invited to participate but did not undergo colonoscopy. The mean (range) age of these twins was 19 (6-56) years. All healthy twins sent

fecal samples and responded to the same questionnaire described above. All collected fecal samples were placed in a freezer at –70°C, immediately after arrival, i.e., at most one day after the samples were collected, and were stored there until analysis. For detailed characteristics of the twins, see Supplementary Table S1. The Örebro County Ethical Committee approved the use of human subjects for this study (Dnr;167/03)

Percent guanine + cytosine profiling of the bacterial community DNA

Bacterial cells were extracted from 0.5 g fecal samples by differential centrifugation as previously described (Apajalahti *et al*., 1998). The isolated bacteria were then lysed and DNA was purified by a protocol comprising enzymatic, chemical and physical steps as described elsewhere in detail (Apajalahti *et al*., 2001; Apajalahti *et al*., 1998). The DNA was fractionated by 72 h CsCl equilibrium density gradient centrifugation, which fractionates chromosomes of the component taxa, based on their characteristic G+C content as described previously (Apajalahti *et al*., 2001; Apajalahti *et al*., 1998; Holben *et al*., 2004). This separation is based on differential density imposed by the AT-dependent DNA-binding dye bis-benzimidazole. Following ultracentrifugation, a Brandel model SYR-94 syringe pump (Brandel, Inc., Gaithersburg, Md.) was used to pass the formed gradients through an ISCO UA-5 UV absorbance detector (ISCO, Inc., Lincoln, Nebr.) set to 280 nm. The %G+C content represented by each gradient fraction was determined by linear regression analysis ($r^2 > 0.99$) of data obtained from control gradients containing standard DNA samples of known %G+C composition as described previously (Apajalahti et al., 1998). This procedure requires a minimum of 30 μg high molecular weight DNA from each sample.

PCR amplification conditions

DNA was extracted from duplicate 250 mg samples from each fecal sample using the MoBio Power Soil DNA Kit (MoBio, Solana Beach, CA), according to the manufacturer's instructions. 16S rRNA genes were PCR amplified from each DNA extract (two technical replicates per extract) using the general bacterial primers Bact- 8F (5'-AGAGTTTGATCCTGGCTCAG-3') (Edwards *et al*., 1989), 5′ end-labeled with 6-carboxyfluorescein (6-FAM), and 926r (5'-CCGTCAATTCCTTTRAGTTT-3') (Muyzer *et al*., 1993) using conditions that have been described in detail elsewhere (Dicksved *et al*., 2007).

In addition, 16S rRNA genes of the *Bacteroides fragilis* subgroup were specifically PCR amplified using a *Bacteroides fragilis* subgroup specific reverse primer, g-Bfra-R (5'-CCAGTATCAACTGCAATTTTA -3') (Matsuki *et al*., 2002) in combination with the same end-labeled Bact-8F general bacterial forward primer mentioned above. PCR amplification was carried out with an initial denaturation step at 95°C for 3 min, followed by 30 cycles consisting of 20 s at 95°C, 20 s at 49°C and 30 s at 72°C. The reaction was completed with a final primer elongation step at 72°C for 5 min. PCR amplified DNA product amounts and sizes were confirmed by agarose gel electrophoresis using GeneRuler 100bp DNA ladder Plus (Fermentas Life Sciences, Burlington, Canada) as a size marker.

Terminal-restriction fragment length polymorphism (T-RFLP)

PCR products were digested with the HaeIII restriction enzyme (GE Healthcare, Uppsala, Sweden) and the digested fragments were separated on an ABI 3700 capillary sequencer (ABI), as previously described (Hjort *et al*., 2007). The sizes of the fluorescently labelled fragments were determined by comparison with the internal GS ROX-500 size standard (ABI). T-RFLP electropherograms were imaged using GeneScan software (ABI). Relative

peak areas of each terminal restriction fragment (TRF) were determined by dividing the area of the peak of interest by the total area of peaks within the following threshold values; lower threshold at 50 bp and upper threshold at 500 bp. Data was normalized by applying a threshold value for relative abundance at 0.5% and only TRFs with higher relative abundances were included in the remaining analyses.

Cloning and sequencing

Cloning and sequencing of 16S rRNA genes from DNA extracted from the fecal samples was performed to confirm the identities of bacterial species corresponding to dominant TRFs from the *Bacteroides* dataset. DNA samples from four twin pairs (one healthy, one concordant and two discordant pairs) were amplified using the *Bacteroides fragilis* subgroup specific primer g-Bfra-R in combination with the general Bact-8F primer. Three replicate PCR products from each individual were pooled and gel purified using the Qiagen gel extraction kit (Qiagen, Hilden, Germany). A total of eight libraries were constructed by inserting PCR products into TOPO TA pCR 4.0 vectors (Invitrogen, Carlsbad, CA), followed by transformation into *Escherichia coli* TOP 10 competent cells (Invitrogen). A total of 24 inserts from each library were PCR amplified using vector primers M13f and M13r (Invitrogen) using the same thermal cycling program as described above for amplification using general bacterial primers for T-RFLP. The PCR products were diluted 50-fold and used in a nested PCR reaction with primers g-Bfra-R and fluorescently tagged primer Bact-8F for T-RFLP analysis of inserted clones, with the same running conditions as described previously for these primers (see above). All clones with unique TRF sizes were selected for sequencing, in addition to several clones from redundant TRFs, for a total of 136 clones. Obtained sequences were examined using MacVector 8.1.1 (Accelrys Software Inc, San Diego, CA), to remove redundant sequences. The remaining sequences were aligned against GenBank database entries using

standard nucleotide BLAST at NCBI (URL: www.ncbi.nlm.nih.gov). Hits defined as unknown or uncultured bacteria were subsequently aligned against sequenced bacterial genomes (genomic BLAST at NCBI), as well as examined with the Ribosomal Database Project II Sequence Match, in an attempt to classify them. Sequences with 99-100% identity, were given the same name as the species hit. Sequences with 97-99% identity were assigned "spp-like". Sequences were aligned using the online MAFFT (standard FFT-NS-i) aligner (Katoh et al., 2002), followed by construction of a circular Neighbor-Joining tree, using BioNJ settings in PAUP4b10 (Swofford, 1993). Unique sequences were deposited in GenBank at NCBI, under the following accession numbers: EU381163-EU381180.

Statistical analysis

The samples were initially statistically assessed as a blind study; i.e. without any prior knowledge of disease status or twin relationship, to avoid potential biases in subsequent data analyses. T-RFLP data from each individual was normalized and entered into a data matrix that consisted of the TRFs as variables and individuals as objects. A consensus T-RFLP profile, from each biological replicate, was constructed by averaging the technical duplicates. Principal component analysis (PCA) plots were generated using the multivariate statistics software Canoco (version 4.5, Microcomputer Power, Ithaca, N.Y.) and statistical significance of ordination was tested using a Monte Carlo permutation test with 999 permutations. Diversity, defined as evenness and richness of the bacterial community members detected as TRFs by T-RFLP analysis, was calculated using Simpson's index of diversity (D) (Begon *et al*., 2006) and Shannon's diversity (H) and equitability index (E) (Begon *et al*., 2006). Differences in diversity between different groups of twins were analyzed by Mann Whitney's U test. Agreement of diversity within twin pairs was analyzed by calculating the intra-class correlation coefficient (ICC) according to Dunn (Dunn *et al*., 1989). Good agreement is

indicated by an ICC value higher than 0.8, fair agreement by values between 0.8-0.4 and a great disagreement by negative values. Differences in bacterial composition (TRF data) within each of the twin pairs were computed with Manhattan distances, and significance between the groups was tested with an ANOVA and Tukey's post hoc test. T-RFLP binary data, i.e. presence or absence of TRFs, was analyzed by cluster analysis using Jaccard's similarity index. P values <0.05 were considered statistically significant.

RESULTS

Percent G+C profiling

Percent G+C profiling of the bacterial chromosomes recovered from the fecal samples was used to detect major differences in the fecal bacterial communities of healthy and diseased individuals in a subset of the twin samples. The power of this method is its robustness; i.e. it examines a large pool of DNA representing the microbial community of interest, and is not susceptible to biases caused by primer mismatches or PCR inhibitors. Six monozygotic twin pairs (one healthy, two discordant and three concordant) were analyzed using this approach. The %G+C profiles that were obtained from the healthy twin pair were very similar which shows that the major bacterial genera present were similar (Figure 1). Microbial communities in the feces of the twin pairs that were concordant for CD were also very similar, but the profiles from discordant twin pairs were dissimilar, illustrating that the microbial composition differed when one of the twins was healthy and the other had CD (Figure 1). It was not possible, however, to distinguish a common pattern for sick or healthy individuals by assessment of the %G+C profiles in this sample set.

T-RFLP profiles using general bacterial primers

T-RFLP was used to obtain bacterial community profiles from fecal samples obtained from 10 monozygotic twin pairs with CD (concordant; n= 4 and discordant; n = 6) and 8 healthy pairs. The reproducibility of the T-RFLP data was very high within technical and biological duplicates. Similarity scores for biological replicates were generally higher than 90%, regardless if abundance data (Manhattan index) or binary data (Jaccard's index) were used. Similar to the %G+C profiling results shown in Figure 1, we found that the T-RFLP patterns were more similar for healthy twins, and for some of the concordant twin pairs, whereas the discordant twins had large differences in their T-RFLP profiles. An illustrative example of the distribution of TRFs for discordant, concordant and healthy twins is shown in Figure 2A.

Community diversity based on TRF diversity

Diversity indices were used to determine the richness (number of TRFs) and evenness of the T-RFLP profiles. The TRF diversity was significantly higher in the healthy group, median (range) 0.91 (0.82-0.93) than in CD patients, median (range) 0.87 (0.71-0.94) when Simpson's index of diversity was used (P = 0.029). However, this significance could not be reproduced for the entire sample group with Shannon's diversity (H) and equitability (E) index (Figure 3). Nevertheless, all healthy individuals in the discordant twin pair sets had a higher TRF diversity than their matched disease twin according to pair wise comparisons using all diversity indices. Using the intra-class coefficient (ICC), a high agreement was observed within healthy pairs (ICC=0.51). In contrast, this high agreement was not observed in discordant (ICC=-0.16) or concordant twin pairs with CD (ICC=-0.05).

Multivariate analyses of T-RFLP profiles

The T-RFLP data representing the gut microbial community profiles were analyzed using multivariate statistics separately for the healthy twin pairs (Figure 4A) and the twin pairs that were concordant or discordant for CD (Figure 4B). Principal component analyses of the T-RFLP profiles obtained from the healthy twin pairs clearly demonstrated that the bacterial community profiles were highly similar to each other for both the first and the second principal component (x and y-axes on the PCA plot, respectively), for individuals of a given pair (Figure 4A). The second principal component differed for only one of the pairs (6a and 6b in Figure 4A). The dizygotic twins [(3a,b, 5a,b, and 8a,b) in Figure 4A] were as similar to each other as the monozygotic pairs in their microbial profiles (Figure 4A).

The bacterial community profiles from fecal samples of twin pairs that were concordant or discordant for CD were less similar to each other compared to those from the healthy twin pairs (Figure 4B). In particular, two of the discordant twin pairs showed large differences in their community profiles (16a,b and 18a,b in Figure 4B). Interestingly, all of the healthy twins in the discordant pairs, grouped to the left of the PCA plot, suggesting that the microbial communities of the healthy individuals share some characteristics that differentiate them from many of the CD individuals. There was, however, no clear gradient that differentiated the whole CD group from the healthy group. However, the bacterial community profiles of twins with ileal involvement were separated from the others and grouped to the right of the PCA plot (Figure 4B). In contrast, the community profiles of twins with colonic disease were similar to those of the healthy individuals, and localized to the left region of the PCA plot (Figure 4B). This separation according to location of the disease was highly significant in ordination space ($P = 0.001$).

To verify the results shown with PCA, Manhattan distances were computed for each twin pair to determine the similarities of the microbial communities within twin pairs. The T-RFLP profile similarities within pairs were significantly different when making between

group comparisons (P = 0.008), with the highest degree of similarity in healthy pairs compared to concordant (P = 0.019) or discordant (P = 0.033) pairs. In addition, by comparing Manhattan differences within discordant pairs, we observed that individuals with ileal CD involvement were less similar to their healthy matching twin compared to discordant twins with colonic CD.

Binary analyses of the T-RFLP data

The T-RFLP binary data, i.e. presence or absence of TRFs, was analyzed by cluster analysis using Jaccard's similarity index. In this analysis all sample data were analyzed together, including healthy, concordant and discordant twin pairs. Four out of eight of the healthy twin pairs, both monozygotic and dizygotic, were more closely clustered to each other than to other individuals, supporting the PCA analyses shown above. Two out of four of the concordant, and one out of six discordant twin pairs, were also similar in their microbial compositions, according to binary similarities (Figure 4C). One of the older healthy pairs (6a,b) had community profiles that were similar according to Jaccard's similarity index of the binary data, but this similarity was not reflected to the same extent in the PCA plots when abundance values were included. Another older twin pair (1a,b) had community profiles that were closely clustered on the PCA plots, but not so when analyzing binary data.

Similar to the PCA analyses, the samples grouped into several clusters (Figure 4C). In particular, CD patients with ileal involvement, except for three individuals (10b, 17a, and 12b), clustered separately from all others. Patients with colonic disease tended to cluster with healthy individuals. There was, however, a large cluster with 16 healthy individuals and only two CD patients, one with ileal involvement (12b) and one with colonic disease (14b).

T-RFLP analysis using primers targeting *Bacteroides*

When using general bacterial primers some trends were observed in the abundances of TRFs within discordant twin pairs with CD, possibly representing *Bacteroides* spp. (Figure 2). For example, CD individuals tended to have a higher relative abundance of TRF 264 (Figure 2A, black areas) and a lower relative abundance of TRF 262 (Figure 2A, dark blue areas). According to *in silico* digestion of 16S rRNA genes deposited in existing databases these TRF sizes could represent *Bacteroides* spp. Although other genera could potentially have similar TRF lengths, it is highly likely that these are representative of *Bacteroides* in our samples since it is known that *Bacteroides* spp. are dominant members of the fecal microbiota (Ott *et al*., 2004; Scanlan *et al*., 2006; Seksik *et al*.; 2003) and our *Bacteroides* clone sequences from the same material had the same TRF sizes (see below). Therefore, a group specific primer set was used during T-RFLP, to focus on the *Bacteroides* group in the same DNA extracts from the fecal samples that were previously analyzed using general bacterial primers. The T-RFLP profiles of the *Bacteroides* community generally had a low complexity, with one or a few dominant peaks shared by most of the individuals, however with large differences in the abundances between individuals (Supplementary Figure S1). In contrast to the results obtained using the general bacterial primers (Figure 4A), the *Bacteroides* profiles within the healthy twin pairs were not more similar compared to discordant or concordant pairs ($P = 0.85$). However, some of the twin pairs had highly similar *Bacteroides* profiles (over 85% similarity based on Manhattan distances), which could not be correlated to disease state (Figure 2B, and Supplementary Figure S1). Interestingly, when the discordant and concordant CD pairs were analyzed by PCA the pattern of the clustering was similar to that observed with general bacterial primers; i.e. the samples from healthy twins in discordant twin pairs grouped together with individuals with colonic disease and the individuals with ileal involvement were significantly separated from the others ($P = 0.030$, Supplementary Figure S2).

Clone libraries of *Bacteroides* spp.

To determine the identities of the different *Bacteroides* spp. detected in the T-RFLP profiles, clone libraries of the amplified 16S rRNA genes were made from four twin pairs [one concordant (15a,b), one healthy (6a,b) and two discordant pairs (12a,b and 18a,b)]. The same *Bacteroides* group specific target regions for PCR amplification were used as those used for T-RFLP of the *Bacteroides* group. A total of 24 clones from each clone library were screened for their TRF fragment sizes (192 total) and 136 of these were sequenced. Most of the clones were identified as *Bacteroides vulgatus, B. uniformis* and *B. ovatus*, (Figure 5). TRFs 262 and 264 matched to sequences corresponding to *B. uniformis* and B. *ovatus,* respectively, and both TRFs 83 and 142 matched to *B. vulgatus* sequences. Some of the CD individuals had a higher relative abundance of the TRF corresponding to *B. ovatus* and a lower relative abundance of the TRF corresponding to *B. uniformis* compared to healthy individuals, but this trend did not hold for the entire sample cohort. However, when looking at disease location the TRF representative of *B. uniformis* was present in significantly lower abundances in twins with ileal involvement ($P = 0.0005$, average abundance; 21± st. dev.11%) compared with both healthy (average abundance; 45±15%, $P = 0.006$) and twins with colonic disease (average abundance; 54±19%, $P = 0.0003$). By contrast, there was a trend that the TRFs corresponding to *B. ovatus* (264; $P = 0.08$) and *B. vulgatus* (83 and 142, $P = 0.12$) were present in higher abundances in patients with ileal involvement (Supplementary Figure S1). Some TRFs had no sequence matches in the clone library and this was generally the case for those TRFs that had a low relative abundance. The relative proportions of specific populations detected by T-RFLP and by cloning and sequencing were highly correlated (Supplementary Table S2).

DISCUSSION

The most widely accepted hypothesis about the pathogenesis of CD is that it is due to a combination of microbial colonization, environmental factors, immune dysfunction, and host genetics. Untangling the possible contribution of microorganisms to CD has been complicated by the large variability in the composition of the gut microbiota in humans. Basically, each human has an individual fecal microbial fingerprint (Dicksved *et al*., 2007; Eckburg *et al*., 2005; Zoetendal *et al*., 1998). However, the study of monozygotic twins basically eliminates the variable of host genetics, except for potential epigenetic factors. In particular the study of a set of discordant monozygotic twin pairs, where one had CD and the other was healthy, was extremely valuable for determination of differences in the gut microbiota, independent of host genetics.

The microbial compositions in fecal samples collected from healthy twin pairs were highly similar, using both T-RFLP and %GC profiling techniques, supporting the hypothesis that genetics has a strong influence on the composition of the gut microbiota. However, six out of eight healthy pairs were young and were still living in the same household. This could contribute to their similar microbial profile. Nevertheless, it was particularly interesting to note that the microbial community profiles of individuals in healthy twin pairs that had lived apart for many years/decades, for example, twin pairs 1a,b and 6a,b, were still highly similar. Zoetendal *et al*. (2001) also observed high similarities among identical twins that had lived separated for more than five years.

In our sample set, three of the healthy twin pairs were dizygotic and the similarities were not higher within the monozygotic compared to the dizygotic twins, although too few pairs were studied to determine the significance of these observations. Even if dizygotic twins share a certain genetic relatedness they are not as closely related as monozygotic twins and

therefore, our results also lend support to the hypothesis that there is environmental programming of the gut microbiota soon after birth (Ley *et al*., 2006). In addition, all of the dizygotic healthy twin pairs were very young, (7-8 years old), and were still living in the same household and this could also contribute to their high similarities in profiles.

Another important finding in this study was that patients with CD ileal involvement, had a significantly different gut microbiota than healthy individuals and those with colonic CD. It is increasingly apparent that Crohn´s disease is not a homogenous disease but a tissue response to various etologic factors (Järnerot, 1996), and our results lend support to this hypothesis. Pairwise comparisons of the microbial profiles from twin pairs also showed that all discordant twins with ileal involvement had community profiles that were less similar to their healthy twin compared to discordant pairs with colonic disease. A possible confounding factor could be surgical impacts prior to sampling, such as ileocecal or ileocolonic resection. However, for the subjects included in this study, their prior surgery was not sufficiently extent for short bowel syndrome to develop. In support of our findings, ileal CD has previously been reported to differ from colonic CD with dysbiosis of the ileal mucosa-associated microbiota correlating to the ileal disease phenotype (Baumgart *et al*., 2007). Also, there are differences in genetic susceptibility (Ahmad *et al*., 2002) and adaptive immune responses (Targan *et al*., 2005) of CD patients with ileal disease compared to those with colonic CD. In this study, dysbiosis of the fecal microbiota correlated with ileal involvement of CD, suggesting that fecal samples could be used as a potential diagnostic marker for the ileal disease phenotype. Some reports suggest that feces are not appropriate for diagnosis of CD, as they may not reflect the composition of mucosa-associated bacteria that are more directly responsible for inflammation (Lepage *et al*., 2005). However, previous findings show that components of feces are relevant for ileal CD recurrence (Rutgeerts *et al*., 1991) and induction of inflammation (Harper *et al*., 1985). Importantly, fecal samples are non-invasive and easier to

obtain than biopsies and our findings provide encouragement in the use of fecal samples for eventual monitoring and/or diagnosis of CD.

Previous studies have found that the microbial diversity in the gut is lower in individuals with CD compared to healthy individuals (Manichanh *et al*., 2006; Scanlan *et al*., 2006; Seksik *et al*., 2003). For example, there have been reports of a reduced diversity of *Firmicutes* (Gophna *et al*., 2006; Manichanh *et al*., 2006) and *Bacteroides* (Frank *et al*., 2007; Ott *et al*., 2004) in CD patients. We also found a significantly higher bacterial diversity in healthy individuals (based on T-RFLP profiles) compared to the CD patients. In addition, all healthy individuals in the discordant twin pairs had higher T-RFLP profile diversities compared to their respective CD twin. Assessment of the ICC coefficient of T-RFLP diversity for the twin pairs also showed a higher agreement within healthy pairs compared to those found within discordant and concordant pairs. These results highlight the power of studying twins as genetically matched controls.

In this study, all of the CD twins, except for two (10b and 15a), were in clinical remission. It has been previously shown that the microbiota of CD patients differ from healthy individuals regardless of disease state (Seksik *et al*., 2003). The two individuals that had active disease in this study had a low degree of inflammation. A possible advantage of studying patients in remission is that during the active stage of the disease, alterations in the microbiota could be an echo, rather than a cause of inflammation (Manichanh *et al*., 2006). In addition, patients in remission have a lower probability of taking medication that could induce changes in the microbiota independent of disease state.

We focused our attention on *Bacteroides* species as these seemed to differentiate between healthy individuals and those with CD when using general bacterial primers for T-RFLP. *Bacteroides vulgatus, B. uniformis*, and *B. ovatus* were the most abundant *Bacteroides* species detected. *B. uniformis* (TRF 262) was present in all samples and *B. ovatus* (TRF 264)

and *B. vulgatus* (TRFs 83 and 142), were present in most of the samples (Supplementary Figure S1). The reason for two representative TRFs for *B. vulgatus* is probably due to different strain variations of the 16S rRNA gene within this species. The *Bacteroides* communities were not significantly similar within any of the twin sets, including the healthy twins (Supplementary Figure S1). However, it has been previously shown that the *Bacteroides* group has a very large inter-individual variation (Eckburg *et al*., 2005; Jernberg *et al*., 2007).

Interestingly, our data suggest that there are differences in the composition of *Bacteroides* species in healthy individuals and CD patients with ileal involvement (Supplementary Figure S2). This difference was largely due to lower relative abundances of *B. uniformis* and higher abundances of *B. ovatus* and *B. vulgatus*, in patients with ileal involvement compared with both healthy twins and twins with colonic disease. Several previous reports have also shown an abnormal *Bacteroides* community in CD patients compared to healthy individuals (Ott *et al*., 2004; Scanlan *et al*., 2006; Seksik *et al*., 2003). For example, Scanlan *et al*. (2006) reported a lower complexity of DGGE profiles within the *B. fragilis* subgroup in CD patients than in healthy individuals and a difficulty in obtaining PCR products from CD patients compared to controls. In this study, it was not problematic to obtain PCR products for the *Bacteroides* group, which could simply reflect differences in *Bacteroides* abundances within the sample groups of the two studies.

One main conclusion of this study was that the healthy twins and some of the concordant twins that were sampled had similar microbial community profiles and these were closely matched within a particular twin set. However, this similarity did not hold for discordant twins suggesting that the diseased individuals had a different microbial community structure than their healthy twins. This finding was made using two independent molecular approaches: %G+C profiling and T-RFLP. By assessing the T-RFLP diversity within twins a higher agreement was found within healthy twin pairs compared to that within discordant and

concordant twin pairs. In addition, the bacterial diversity was higher in healthy twins compared to CD twins. Interestingly, cluster analysis of binary T-RFLP data as well as ordination techniques of T-RFLP abundance data showed that CD twins with ileal involvement differed from healthy twins as well as from twins with colonic disease. This difference could partly be explained by a shift of the dominant *Bacteroides* community members of CD patients with ileal involvement.

ACKNOWLEDGEMENTS

This project was funded by the Örebro University Hospital Research Foundation, Örebro County Research Foundation, Bengt Ihre´s foundation, the Uppsala BioX Micprof project, the Swedish University of Agricultural Sciences and in part by U. S. Department of Energy Contract DE-AC02-05CH11231 with Lawrence Berkeley National Laboratory. Maria Hellman is thanked for her assistance with DNA extraction from the samples and Kerstin Eriksson for her assistance with collection of biological material.

REFERENCES

Ahmad T, Armuzzi A, Bunce M, Mulcahy-Hawes K, Marshall SE, Orchard TR, *et al.* (2002). The molecular classification of the clinical manifestations of Crohn's disease. *Gastroenterology* **122**: 854-866.

Apajalahti JHA, Kettunen A, Bedford MR, Holben WE. (2001). Percent G+C profiling accurately reveals diet-related differences in the gastrointestinal microbial community of broiler chickens. *Appl Environ Microbiol* **67**: 5656-5667.

Apajalahti JHA, Sarkilahti LK, Maki BRE, Heikkinen JP, Nurminen PH, Holben WE. (1998). Effective recovery of bacterial DNA and percent-guanine-plus-cytosine-based analysis of community structure in the gastrointestinal tract of broiler chickens. *Appl Environ Microbiol* **64**: 4084-4088.

Barnich N, Darfeuille-Michaud A.(2007). Adherent-invasive *Escherichia coli* and Crohn's disease. *Curr Opin Gastroenterol* **23**: 16-20.

Baumgart M, Dogan B, Rishniw M, Weitzman G, Bosworth B, Yantiss R, *et al.* (2007). Culture independent analysis of ileal mucosa reveals a selective increase in invasive *Escherichia coli* of novel phylogeny relative to depletion of *Clostridiales* in Crohn's disease involving the ileum. ISME journal **1**: 403-418.

Begon M, Harper JL, Townsend CR. (2006). Ecology: from individuals to ecosystems, 4th ed. Blackwell, Oxford, **pp 471-472**.

Cederlöf R, Friberg L, Jonsson E, Kaij L. (1961). Studies on similarity diagnosis in twins with aid of mailed questionnaires. *Acta Genet Stat Med* **11**: 338-362.

Darfeuille-Michaud A, Neut C, Barnich N, Lederman E, Di Martino P, Desreumaux P, *et al.* (1998). Presence of adherent *Escherichia coli* strains in ileal mucosa of patients with Crohn's disease. *Gastroenterology* **115**: 1405-1413.

Dicksved J, Flöistrup H, Bergstrom A, Rosenquist M, Pershagen G, Scheynius A, *et al.* (2007). Molecular fingerprinting of the fecal microbiota of children raised according to different lifestyles. *Appl Environ Microbiol* **73**: 2284-2289.

Dunn G. (1989). Design and analysis of reliability studies. The statistical evaluation of measurement errors. Oxford University Press, New York, **pp 34-35**.

Eckburg PB, Bik EM, Bernstein CN, Purdom E, Dethlefsen L, Sargent M, *et al.* (2005). Diversity of the human intestinal microbial flora. *Science* **308**: 1635-1638.

Edwards U, Rogall T, Blocker H, Emde M, Bottger EC. (1989). Isolation and direct complete nucleotide determination of entire genes - characterization of a gene coding for 16S-ribosomal RNA. *Nucleic Acids Res* **17**: 7843-7853.

Frank DN, St Amand AL, Feldman RA, Boedeker EC, Harpaz N, Pace NR (2007). Molecular-phylogenetic characterization of microbial community imbalances in human inflammatory bowel diseases. *Proc Natl Acad Sci USA* **104**: 13780-13785.

Gophna U, Sommerfeld K, Gophna S, Doolittle WF, van Zanten SJ (2006). Differences between tissue-associated intestinal microfloras of patients with Crohn's disease and ulcerative colitis. *J Clin Microbiol* **44**: 4136-4141.

Halfvarson J, Bodin L, Tysk C, Lindberg E, Järnerot G (2003). Inflammatory bowel disease in a Swedish twin cohort: A long-term follow-up of concordance and clinical characteristics. *Gastroenterology* **124**: 1767-1773.

Halvarson J, Jess T, Magnuson A, Montgomery SM, Orholm M, Tysk C, *et al.* (2006). Environmental factors in inflammatory bowel disease: A co-twin control study of a Swedish-Danish twin population. *Inflamm Bowel Dis* **12**: 925-933.

Halfvarson J, Tysk C, Järnerot G. (2004). Decreasing pair concordance in monozygotic twins with Crohn's disease. *Gastroenterology* **126**: A45-A45.

Harper PH, Lee ECG, Kettlewell MGW, Bennett MK, Jewell DP. (1985). Role of the fecal stream in the maintenance of Crohn's colitis. *Gut* **26**: 279-284.

Harvey RF, Bradshaw JM. (1980). A simple index of disease activity. *Lancet* **315**: 514.

Hjort K, Lembke A, Speksnijder A, Smalla K, Jansson JK. (2007). Community structure of actively growing bacterial populations in plant pathogen suppressive soil. *Microb Ecol* **53**: 399-413.

Holben WE, Feris KP, Kettunen A, Apajalahti JHA. (2004). GC fractionation enhances microbial community diversity assessment and detection of minority populations of bacteria by denaturing gradient gel electrophoresis. *Appl Environ Microbiol* **70**: 2263-2270.

Jernberg C, Löfmark S, Edlund C, Jansson JK. (2007). Long-term ecological impacts of antibiotic administration on the human intestinal microbiota. ISME journal **1**: 56-66.

Jess T, Gamborg M, Matzen P, Munkholm P, Sorensen TIA. (2005). Increased risk of intestinal cancer in Crohn's disease: A meta-analysis of population-based cohort studies. *Am J Gastroenterol* **100**: 2724-2729.

Järnerot G. (1996). Future aspects on inflammatory bowel disease. *Scand J Gastroenterol* **220**:87-90.

Katoh K, Misawa K, Kuma K, Miyata T. (2002). MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucleic Acids Res*. **30**: 3059-66

Keighley MRB, Arabi Y, Dimock DW, Allan RN, Alexander-Williams J. (1978). Influence of inflammatory bowel disease on intestinal microflora. *Gut* **19**: 1099-1104.

Lepage P, Seksik P, Sutren M, de la Cochetiere MF, Jian R, Marteau P*, et al.* (2005). Biodiversity of the mucosa-associated microbiota is stable along the distal digestive tract in healthy individuals and patients with IBD. *Inflamm Bowel Dis* **11**: 473-480.

Ley RE, Peterson DA, Gordon JI. (2006). Ecological and evolutionary forces shaping microbial diversity in the human intestine. *Cell* **124**: 837-848.

Lichtenstein P, De Faire U, Floderus B, Svartengren M, Svedberg P, Pedersen NL. (2002). The Swedish twin registry: a unique resource for clinical, epidemiological and genetic studies. *J Int Med* **252**: 184-205.

Loftus Jr EV. (2004). Clinical epidemiology of inflammatory bowel disease: Incidence, prevalence, and environmental influences. *Gastroenterology* **126**: 1504-1517.

Manichanh C, Rigottier-Gois L, Bonnaud E, Gloux K, Pelletier E, Frangeul L*, et al.* (2006). Reduced diversity of faecal microbiota in Crohn's disease revealed by a metagenomic approach. *Gut* **55**: 205-211.

Matsuki T, Watanabe K, Fujimoto J, Miyamoto Y, Takada T, Matsumoto K*, et al.* (2002). Development of 16S rRNA-gene-targeted group-specific primers for the detection and identification of predominant bacteria in human feces. *Appl Environ Microbiol* **68**: 5445-5451.

Muyzer G, Dewaal EC, Uitterlinden AG. (1993). Profiling of complex microbial populations by denaturing gradient gel electrophoresis analysis of polymerase chain reaction-amplified genes coding for 16S ribosomal-RNA. *Appl Environ Microbiol* **59**: 695-700.

Orholm M, Binder V, Sorensen TIA, Rasmussen LP, Kyvik KO. (2000). Concordance of inflammatory bowel disease among Danish twins - Results of a nationwide study. *Scand J Gastroenterol* **35**: 1075-1081.

Ott SJ, Musfeldt M, Wenderoth DF, Hampe J, Brant O, Folsch UR*, et al.* (2004). Reduction in diversity of the colonic mucosa associated bacterial microflora in patients with active inflammatory bowel disease. *Inflamm Bowel Dis* **53**: 685-693.

Rutgeerts P, Goboes K, Peeters M, Hiele M, Penninckx F, Aerts R*, et al.* (1991). Effect of faecal stream diversion on recurrence of Crohn's disease in the neoterminal ileum. *Lancet* **338**: 771-774.

Sartor RB. (2003). Targeting enteric bacteria in treatment of inflammatory bowel diseases: why, how and when. *Curr Opin Gastroenterol* **19**: 358-365.

Sartor RB. (2006). Mechanisms of disease: pathogenesis of Crohn's disease and ulcerative colitis. *Nat Clin Pract Gastroenterol Hepatol* **3**: 390-407.

Satsangi J, Silverberg MS, Vermeire S, Colombel JF. (2006). The Montreal classification of inflammatory bowel disease: controversies, consensus, and implications. *Gut* **55**: 749-753.

Scanlan P, Shanahan F, O' Mahony C, Marchesi JR. (2006). Culture-independent analyses of temporal variation of the dominant fecal microbiota and targeted bacterial subgroups in Crohn's disease. *J Clin Microbiol* **44**: 3980-3988.

Seksik P, Rigottier-Gois L, Gramet G, Sutren M, Pochart P, Marteau P*, et al.* (2003). Alterations of the dominant faecal bacterial groups in patients with Crohn's disease of the colon. *Gut* **52**: 237-242.

Stewart JA, Chadwick VS, Murray A. (2005). Investigations into the influence of host genetics on the predominant eubacteria in the faecal microflora of children. *J Med Microbiol* **54**: 1239-1242.

Swidsinski A, Ladhoff A, Pernthaler A, Swidsinski S, Loening-Baucke V, Ortner M*, et al.* (2002). Mucosal flora in inflammatory bowel disease. *Gastroenterology* **122**: 44-54.

Swidsinski A, Weber J, Loening-Baucke V, Hale LP, Lochs H. (2005). Spatial organization and composition of the mucosal flora in patients with inflammatory bowel disease. *J Clin Microbiol* **43**: 3380-3389.

Swofford DL. (1993). Paup - a computer-program for phylogenetic inference using maximum parsimony. *J Gen Physiol* **102**:A9-A9.

Targan SR, Landers CJ, Yang HY, Lodes MJ, Cong YZ, Papadakis KA*, et al.* (2005). Antibodies to CBir1 flagellin define a unique response that is associated independently with complicated Crohn's disease. *Gastroenterology* **128**: 2020-2028.

Tysk C, Lindberg E, Järnerot G, Floderus-myrhed B. (1988) Ulcerative colitis and Crohn's-disease in an unselected population of monozygotic and dizygotic twins - A study of heritability and the influence of smoking. *Gut* **29**: 990-996.

Van de Merwe JP, Stegeman JH, Hazenberg MP. (1983). The resident faecal flora is determined by genetic characteristics of the host. Implications for Crohn's disease? Antonie van Leeuwenhoek **49**: 119-124.

Zoetendal EG, Akkermans ADL, Akkermans-van Vliet WM, de Visser JAGM, De Vos WM. (2001) The host genotype affects the bacterial community in the human gastrointestinal tract. *Microbial Ecol Health Dis* **13**: 129-134.

Zoetendal EG, Akkermans ADL, De Vos WM. (1998) Temperature gradient gel electrophoresis analysis of 16S rRNA from human fecal samples reveals stable and host-specific communities of active bacteria. *Appl Environ Microbiol* **64**: 3854-3859.

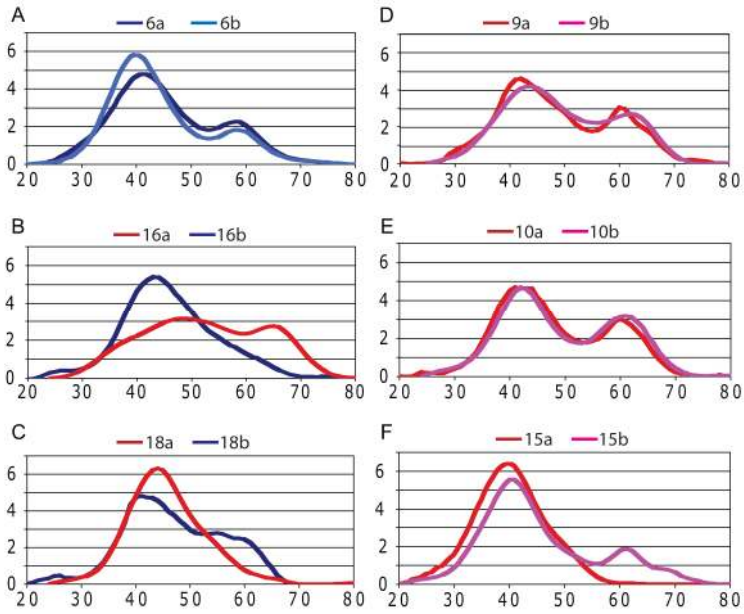

Figure 1. Percent guanine+cytosine profiles for six twin pairs: healthy pair (A), two discordant (B, C), and three concordant (D-F) twin pairs for Crohn's disease (CD). Healthy individuals are labeled dark or light blue and CD individuals are labeled pink or red. Sample identifications are provided at the top of each panel according to assignments given in Supplementary Table S1. Horizontal axis shows GC content in percent and the vertical axis indicates relative absorbance values in percent.

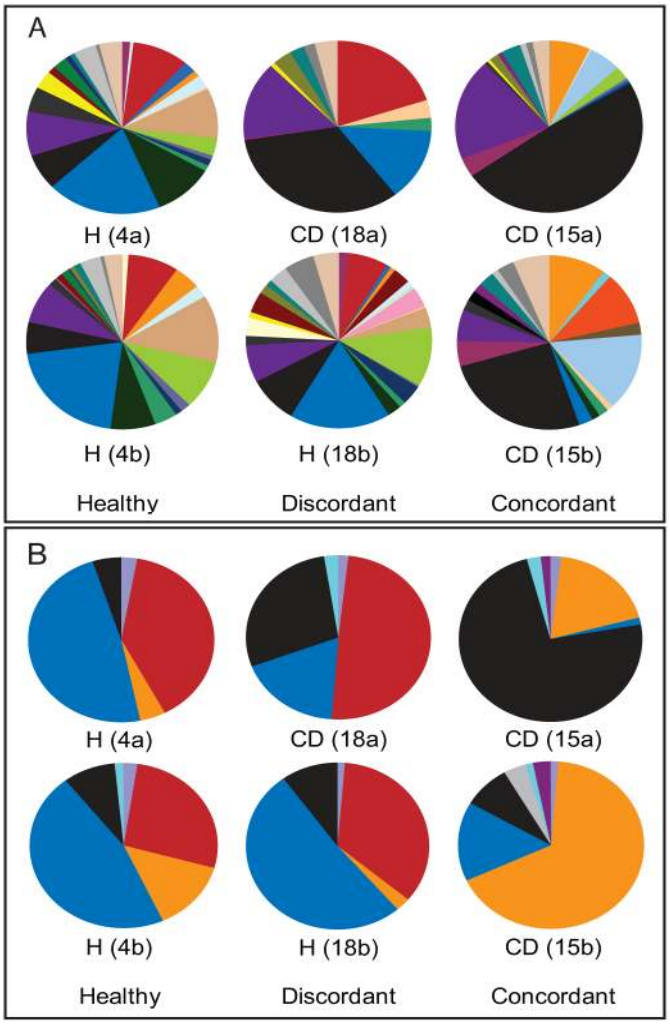

Figure 2. Terminal restriction fragment (TRF) distributions generated using general bacterial primers (Panel A), and a *Bacteroides* group specific reverse primer (Panel B), for three twin pairs (one healthy, one discordant and one concordant for Crohn's disease). Identification of individuals according to assignments given in Supplementary Table S1 is shown below each pie chart. Each area represents the relative abundance of a particular TRF. TRFs of the same size are the same color for all individuals and for both panels A and B considering that the forward primer used for PCR was the same in all cases.

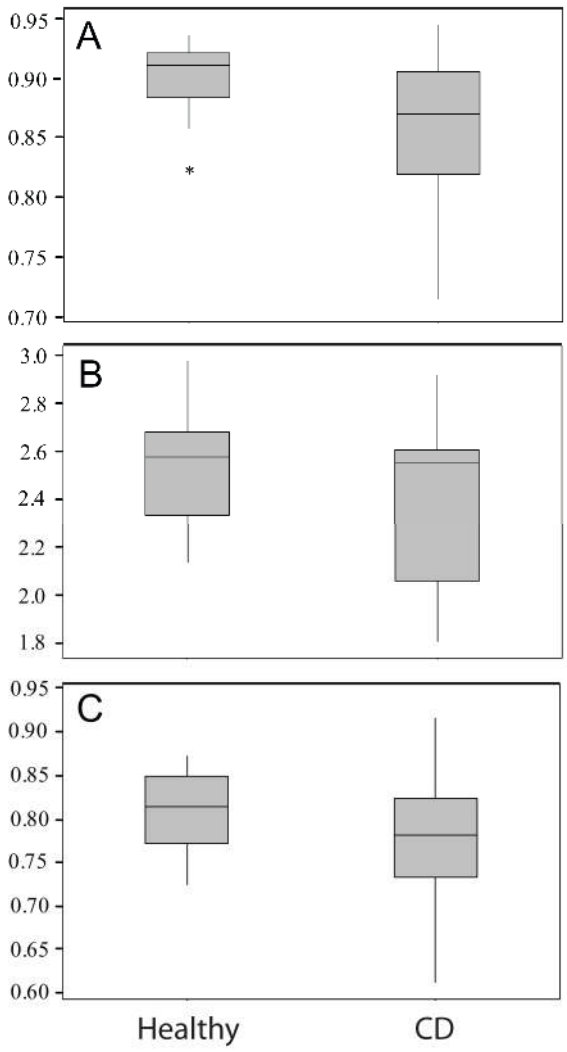

Figure 3. Box plots comparing the diversity of T-RFLP profiles calculated for healthy individuals to those with CD: (A) Simpson's index of diversity, (B) Shannon's richness index and (C) Shannon's evenness index.

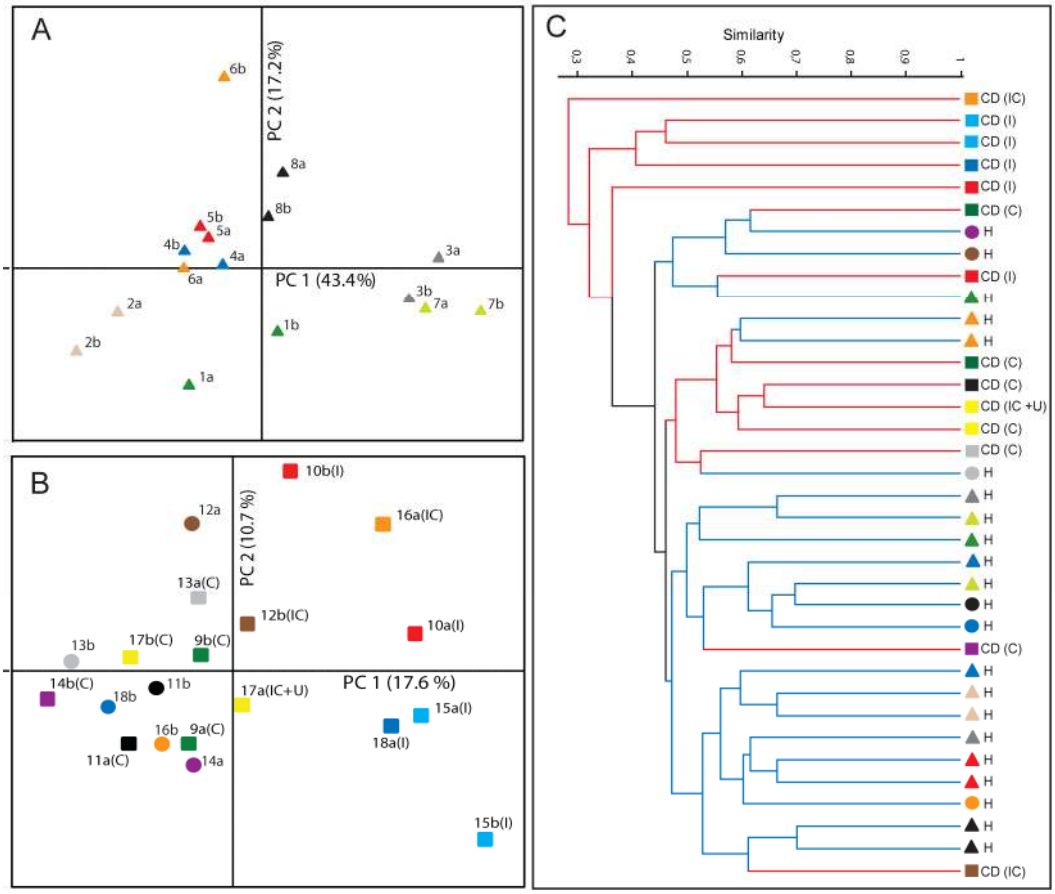

Figure 4. Principal component analysis (PCA) plots for T-RFLP profiles (including TRF size and relative abundance data) for healthy twins (Panel A) and twins discordant or concordant for CD (Panel B); Eigenvalues are shown in parentheses for PC1 and PC2. Panel C shows a similarity plot based on the binary T-RFLP data (i.e. presence or absence of TRFs) for all individuals calculated using Jaccard's index and UPGMA as a cluster method. Twin pairs are coded according to colors and shapes of symbols: healthy individuals in control group, closed triangles; healthy individuals in discordant twin pairs, closed circles; individuals with CD, closed squares. Individuals within a twin pair have the same colored symbol. In Panel C, clusters comprised of healthy individuals are indicated with blue branches whereas those for CD individuals with red. Sample ID is shown next to the symbols on the PCA plots (panels A and B) according to the assignments given in Supplementary Table S1. Abbreviated disease locations: I: Ileum, C: colon, IC: Ileocolon, and U: Upper gastrointestinal tract.

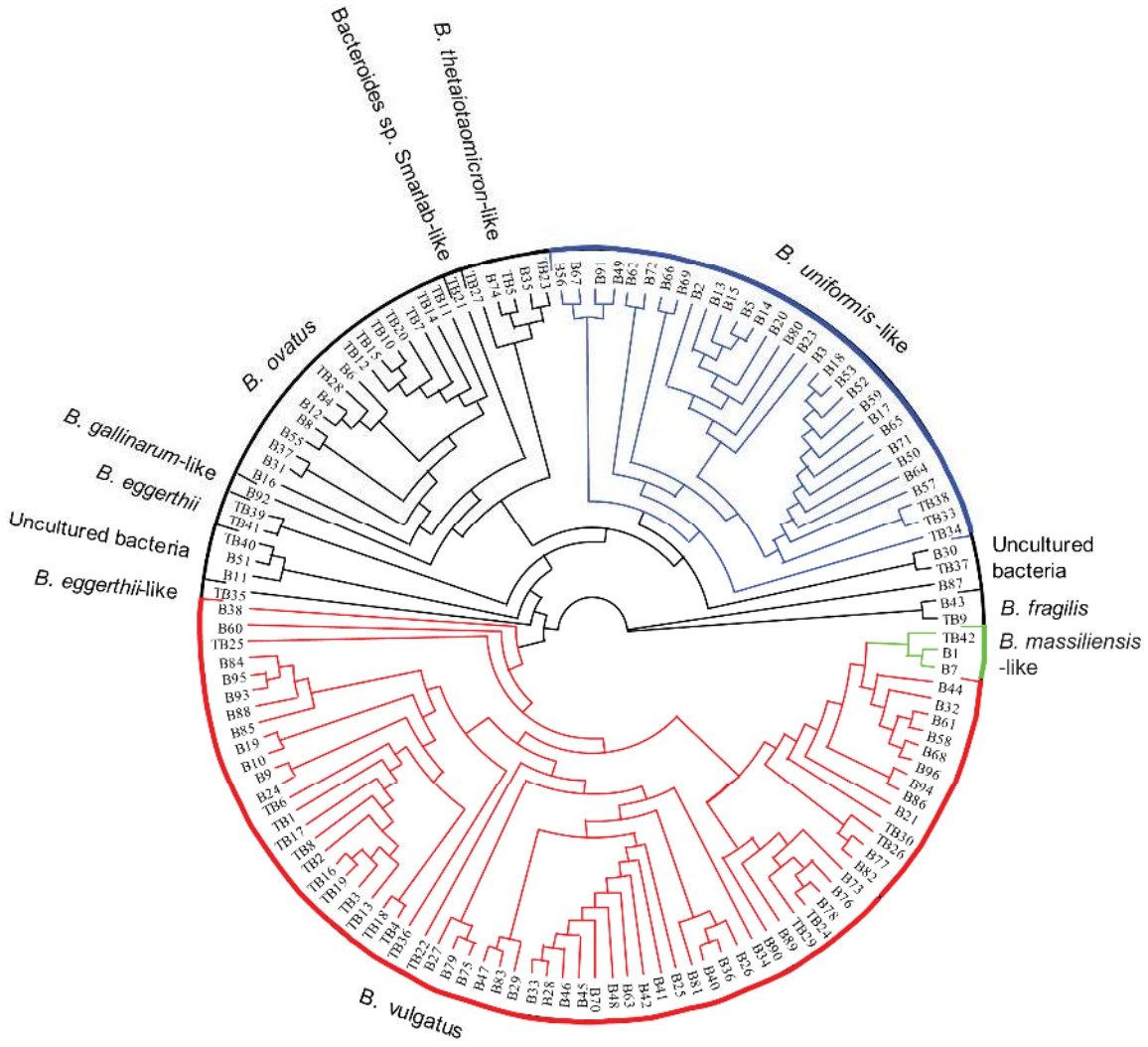

Figure 5. Neighbor joining tree showing clustering of the *Bacteroides* clone sequences based on libraries created from one healthy twin pair (6a,b), one concordant CD pair (15a,b) and two discordant CD pairs (16a,b and 18a,b); the T-RFLP abundance data for these individuals are shown in Supplementary Figure S1. Coloring of branches illustrate the respective *Bacteroides* species that matched to the clone sequences. For sequences where the species names are given, the matches were 99-100%, and species with sequence identities >97%, were called *Bacteroides* spp-like.