The impact of NOD2 variants on fecal microbiota in Crohn’s disease and controls without gastrointestinal disease

**Nicholas A Kennedy1-2,** Christopher A Lamb3, Susan H Berry4, Alan W Walker5-6, John Mansfield7, Miles Parkes8, Rachel Simpkins9, Mark Tremelling10, Sarah Nutland9, UK IBD Genetics Consortium, Julian Parkhill5, Chris Probert11,Georgina L Hold4§, Charlie W Lees1§.

1GI Unit, Institute of Genetics and Molecular Medicine, University of Edinburgh, Edinburgh, UK, 2IBD pharmacogenetics group, University of Exeter, UK, 4 Institute of Cellular Medicine, Newcastle University, UK 4Gastrointestinal Research Group, University of Aberdeen, Aberdeen, UK, 5Pathogen Genomics Group, Wellcome Trust Sanger Institute, Hinxton, Cambridgeshire, UK, 6Microbiology Group, The Rowett Institute, University of Aberdeen, Aberdeen, UK, 7Dept of Gastroenterology, Royal Victoria Infirmary, Newcastle, UK,  8Dept of Gastroenterology, Addenbrookes Hospital, Cambridge, UK, 9Cambridge BioResource, Cambridge, 10Department of Gastroenterology, Norfolk and Norwich University Hospital, Norwich, UK, 11Institute of Translational Medicine, University of Liverpool, UK.

§ Denotes Joint senior authorship

# Short Title

NOD2 genotype and gut microbiota

# Address for correspondence

Dr Nicholas Kennedy

IBD pharmacogenetics group, RILD South, Royal Devon and Exeter Hospital, Barrack Road, Exeter EX2 5DW

Email: n.kennedy@exeter.ac.uk

Telephone: +44 (0)1392 408457

# Conference presentations

Earlier versions of this work have been presented at BSG 2014 (Manchester, UK), BSG 2015 (London, UK) and DDW 2015 (Washington DC, USA).

# Funding

This work was supported by CORE, the Digestive Diseases Foundation and the Wellcome Trust [grant number 097943 to NAK, 093885 to CAL and 098051 to AWW [and JP]]. AWW receives core funding support from the Scottish Government Rural and Environmental Science and Analysis Service (RESAS).

# Abstract

## Background/Aims

Current models of Crohn’s disease (CD) describe an inappropriate immune response to gut microbiota in genetically susceptible individuals. *NOD2* variants are strongly associated with development of CD, and NOD2 is part of the innate immune response to bacteria. This study aimed to identify differences in fecal microbiota in CD patients and non-IBD controls stratified by *NOD2* genotype.

## Methods

Patients with CD and non-IBD controls of known *NOD2* genotype were identified from patients in previous UK IBD genetics studies and the Cambridge bioresource (genotyped/phenotyped volunteers). Individuals with known CD-associated *NOD2* mutations were matched to those with wild-type genotype. We obtained fecal samples from patients in clinical remission with low fecal calprotectin (<250µg/g) and controls without gastrointestinal disease. After extracting DNA, the V1-2 region of 16S rRNA genes were PCR-amplified and sequenced. Analysis was undertaken using the mothur package. Volatile organic compounds (VOC) were also measured.

## Results

91 individuals were in the primary analysis (37 CD, 30 bioresource controls and 24 household controls). Comparing CD with non-IBD controls, there were reductions in bacterial diversity, *Ruminococcaceae*, *Rikenellaceae* and *Christensenellaceae* and an increase in *Enterobacteriaceae*. No significant differences could be identified in microbiota by *NOD2* genotype, but fecal butanoic acid was higher in Crohn’s patients carrying *NOD2* mutations.

## Conclusions

In this well-controlled study of *NOD2* genotype and fecal microbiota, we identified no significant genotype-microbiota associations. This suggests that the changes associated with *NOD2* genotype might only be seen at the mucosal level, or that environmental factors and prior inflammation are the predominant determinant of the observed dysbiosis in gut microbiota.

# Keywords

Crohn’s Disease, NOD2, Microbiota, genotype

# Authorship Statement

NAK, GLH, CWL, AWW, JP and CP designed the study. NAK and CWL co-ordinated the study. NAK, CAL, JM,MP, RS, MT, SN and CWL recruited patients from the study sites. SHB, GLH and NAK prepared the collected samples for sequencing and measurement of calprotectin. CP generated and analysed volatile organic compound data. AWW and JP generated 16S rRNA sequencing data. NAK analysed the clinical and scientific data and wrote the initial draft of the paper with input from GLH and CWL. All authors reviewed and contributed to subsequent drafts of the paper and approved the final manuscript.

# Conflicts of Interest

None of the authors have any relevant conflicts of interest to report.

# Introduction

The precise aetiology of Crohn’s disease remains unknown. However, the key pathogenic process involves an inappropriate immune response that results in bowel inflammation and damage. The targets of this response are thought to be antigens derived from constituents of the microbiota, a view supported by the benefits of altering the microbiota1 or physically diverting the fecal stream.2 Further, 16S rRNA gene sequencing has shown that the microbiota in IBD is abnormal and characterized by reduced diversity with fewer *Firmicutes* species present.3 The direction of causality between IBD and alterations in microbiota remains incompletely understood, as does the question of whether overall dysbiosis or specific taxa are most important.Recent research has also emphasized the functional aspect of the gut microbiota through measurement of microbial metabolites such as the volatile organic compounds present in feces.4

The last few years has seen rapid advances in the genetics of IBD as a result of large cohort genome-wide association studies (GWAS) of cases and controls. Over 200 IBD susceptibility loci have now been reported.5–7 For some loci, the disease gene and associated point mutations are known (e.g. *NOD2* and *ATG16L1)*. *NOD2* has the largest effect, and a large recent subphenotype-genotype analysis has confirmed that *NOD2* is strongly associated in particular with ileal CD.8 Viewed alongside other functionally inter-related genes that have been associated with CD (e.g. *ATG16L1*, *IRGM*, *XBP1*), an impaired capability of the host to regulate microbial constituents consistently emerges as a major common theme.

NOD2 is a cytosolic pattern recognition receptor (PRR) that is a key player in immunity to intracellular bacteria and inflammatory responses. NOD2 recognises muramyldipeptide (MDP), a ubiquitous component of bacterial cells walls, and its stimulation leads to induction of autophagy in human cells.9 Variants of *NOD2* associated with CD are mutated in the ligand recognition domain and fail to induce autophagy on MDP triggering, which results in aberrant bacterial handling and antigen presentation in these cells.10–12 NOD2 possesses other anti-bacterial effects, including the ability to prime human dendritic cells (DCs) to promote T-helper 17 (Th17) responses (via NOD2-induced-expression of IL-23 and IL-1)13 and the ability to induce antimicrobial peptide defensins in the intestine.14 If expression of CD-variant *NOD2* leads to dysregulated bacterial destruction within the cells in which it is expressed, bacteria may persist abnormally in the mucosa and activate tissue inflammation in these sites.

However, approximately 11-14% of white Europeans are heterozygous and 0.4-0.9% homozygous or compound-heterozygous for CD-risk-variant *NOD2* but remain healthy, which reinforces the role for co-existent genetic or environmental factors in initiation of CD.15,16 The association of defective antibacterial mechanisms with CD-associated polymorphisms in *NOD2* suggest that the presence of these variants may influence the nature of the microbiota over time. This in turn might lead to either a critical dysbiotic state being reached, or the presence of specific microbes emerging to initiate the cycle of inflammation observed in disease. For example, altered release of antibacterial peptides from variant-*NOD2*-expressing Paneth cells, defective Th17 responses or defective autophagic bacterial processing in the gut mucosa could change gut bacterial burden or species diversity.

Little is known of the nature of the microbiota in the presence of NOD2 mutations. Human studies to date have been limited in scope due to small numbers of individuals homozygous for *NOD2* mutations without accurate matching of controls. Frank and colleagues revisited the dataset from their index 2007 study on the microbiota in IBD, stratifying patients retrospectively for *NOD2* and *ATG16L1* genotype. Due to its retrospective design this study is severely constrained by limited power; despite this, they observed clear shifts in microbial composition as a result of genotype.17 The aim of the current study was to prospectively define the role of *NOD2* genotypein influencing the nature of the host microbiota in health and in Crohn’s disease

# Methods

Individuals with CD of known *NOD2* genotype were identified from the UK IBD genetics consortium (Figure 1). Patients were selected if they carried two copies of the CD-associated *NOD2* variants (i.e. homozygotes or compound heterozygotes for R702W [rs2066844], G908R [rs2066845] or L1007fs [rs2066847])18,19 as measured using genotyping arrays for the original genetics studies in which they had been involved (Affymetrix GeneChip 50020 and ImmunoChip6). Patients were recruited if they were deemed by their treating physician to be in clinical remission. Each *NOD2*-mutant patient was matched to a homozygous *NOD2*-wild-type patient. Exclusion criteria for Crohn’s-disease patients included antibiotics within the three months prior to recruitment, active Crohn’s disease (by physician global assessment), and presence of an ileostomy. For all CD patients a household control was approached (usually an unrelated spouse). Healthy volunteers stratified by the same *NOD2* variants were recruited from the Cambridge BioResource.21 The Cambridge BioResource is a panel of around 16,000 volunteers, both with and without health conditions, who have previously submitted DNA for genotyping. Participants can be approached for studies on the basis of genotype and phenotypic characteristics. Volunteers from the BioResource had no known gastrointestinal diagnosis and had had no antibiotics in the preceeding three months. All study participants had fecal calprotectin (FC) measured by a standard ELISA (Calpro AS, Norway). CD patients with FC >250 µg/g and controls (household or BioResource) with FC >50 µg/g were excluded from further analysis.

Clinical data including medical and surgical history, smoking status, medication history, antibiotic use, probiotic use, weight, height, Montreal disease location and behavior22 were collected using a patient questionnaire, interrogation of the medical record and use of previous phenotype information recorded on the IBD cohorts as part of a rephenotyping exercise.8 Probiotic and medication use were documented as being used at the time of sampling. Participants with missing data were excluded from the analysis of that specific datapoint.

Fecal samples were collected from each study participant using the Fisher Fecal Commode Collection Kit. The collection container was held in the toilet bowl using the supplied trivet, and whole fecal samples collected without contamination by urine. Samples were kept cold using phase-change refrigerant gel packs and processed within 24 hours of collection. This short period of storage is not expected to significantly influence molecular estimation of microbial community composition,23 nor the VOC profile (unpublished data). Each sample was thoroughly mixed and aliquots were transferred into lysing matrix E tubes (MP Biomedicals, Santa Ana, CA, USA) for subsequent DNA extraction, head space vials (Supelco, Bellefonte, PA, USA) for volatile organic compound analysis and universal containers for fecal calprotectin analysis. Samples were stored at -80 °C before shipping to a central processing laboratory in Aberdeen, UK, where DNA was extracted within 1 month of collection.

Household controls had not been previously genotyped. Saliva samples were acquired using Oragene kits (DNA Genotek, Ottawa, Canada). DNA was extracted following the manufacturer’s protocol and was genotyped for the three CD-associated *NOD2* variants listed above using TaqMan assays (Applied Biosystems, Carlsbad, CA, USA). Where a genotype was not determined, the allelic discrimination plots were examined manually to ensure a homozygotes for the minor allele had not been missed. For the purposes of analysis of the household controls, missing genotypes were inferred to be wild type

Ethical approval was granted by the North of Scotland Research Ethics Committee (reference 12/NS/0050). All participants provided written consent.

## DNA extraction

For each fecal sample an approximately 400 mg aliquot was placed in a lysing matrix E tube. 978 µl of sodium phosphate buffer and 122 µl MT buffer were added to each tube and vortex mixed. This was then processed using the FastDNA SPIN kit for Soil following the manufacturer's instructions (MP Biomedicals) as described previously.24

## PCR amplification and sequencing

The V1-V2 region of the 16S rRNA gene was amplified using 27F and 338R primers.25 The primers were designed with the Illumina adapter sequences already included and with one of 200 barcode sequences included in the 338R reverse primer, thus avoiding the need for a separate step to add the adapter sequences and barcode. Twenty cycles of PCR amplification were performed using the Q5 polymerase kit following the manufacturer’s instructions (New England Bio, Ipswich, MA, USA). Post-amplification, samples were quantified using a Qubit fluorometer (Thermo Fisher, Waltham, MA, USA) and then pooled to obtain equimolar concentrations.

Sequencing was performed using an Illumina MiSeq sequencer using Illumina V2 chemistry and paired-end 2×250 base pair reads. Initial sequence data processing was performed in the Illumina MiSeq Reporter to de-multiplex samples and strip adapters and primers and sequence data were exported in the FASTQ format.

## Bioinformatics analysis

The 16S rRNA gene sequence data were further processed using mothur26 following the MiSeq SOP.27

Alignment and classification were done against the SILVA v119 reference set.28 Community structures were compared using trees generated using Jaccard and Yue Clayton distance metrics after subsampling to 3,943 reads per sample. The trees were then plotted graphically using the Interactive Tree of Life.29,30 Trees were compared using the parsimony command within Mothur. Subsequent statistical analysis was done in R 3.2.2 (R Foundation for Statistical Computing, Vienna, Austria). Microbial diversity was assessed using inverse Simpson and compared using a Mann Whitney U test.31 Comparisons at the family, genus and OTU level were done using Mann Whitney U tests for binary comparisons and corrected using Holm’s method.32 The twelve most abundant families were selected for plotting graphically.

## Volatile organic compound analysis

VOC data were generated using previously methodology described methodology.33 Briefly, gas chromatography-mass spectroscopy (GCMS) was used to quantify the metabolites in the headspace gas taken from vials containing an aliquot of participants’ faeces. The raw GCMS data were processed using AMDIS (National Institute of Standards and Technology, Gaithersburg, MD, USA). Compounds detected in fewer than 20% of the study population were filtered out. The resultant ion intensity data were log transformed and the limma package used to facilitate running multiple linear models including disease status and *NOD2* genotype as covariates.34 P values were corrected using Holm’s method.

# Results

Out of the 110 individuals recruited, 91 were used in the primary analysis (Table 1). Reasons for exclusion are shown in Supplementary Table 1. There were 37 CD patients [57% *NOD2* mutant], 30 bioresource volunteers [58% *NOD2* mutant] and 24 household controls. All were of white European ethnicity. There were no differences in phenotype within the CD patients by *NOD2* status (Table 1). Five out of 21 genotyped household controls with genotype information had single CD-associated *NOD2­*-associated mutations. Three of these had a first degree relative with Crohn’s.

The total number of raw reads was 3,410,868, with a median number of reads per sample of 34,302. After quality control and removal of samples with very low read numbers, the remaining samples had a minimum of 3,943 reads and median of 20,338. The sequence data are available from the European Nucleotide Archive under Study Accession Number PRJEB21593.

There was a significant reduction in diversity (as assessed by calculating the inverse Simpson index) between CD cases and both bioresource and household controls (p < 0.001 and 0.003 respectively, Figure 2). No difference was observed in diversity by *NOD2* genotype either within the CD cases or the bioresource controls (p=0.32 and 0.65). Hierarchical clustering using the Jaccard metric demonstrated clustering by CD versus controls in either cohort (p<0.001), but not by *NOD2* genotype (p=0.16 within cases, (Figure 3)). The CD cases also clustered with each other rather than their household controls; indeed the Jaccard distance between cases and their household control was no different from the distance between cases and unmatched household controls (p=0.81, Mann Whitney U).

At a family level, there were significant decreases in *Ruminococcaceae*, *Rikenellaceae* and *Christensenellaceae* (p all <0.001 uncorrected and <0.01 corrected) and an increase in *Enterobacteriaceae* (p<0.001 corrected) in samples from CD patients vs. controls (Figure 4A). There were no differences in relative abundance of any bacterial families when stratified by *NOD2* status, either within the CD patients or bioresource controls (Figure 4B). There were also no differences by genotype at the genus or OTU taxonomic levels; in each case, comparisons were made using a Mann Whitney U test with correction for multiple testing using Holm’s method, and no corrected p value was less than 0.05.

## Volatile organic compound analysis

For the VOC analysis, there were 314 compounds identified in at least one sample, and 198 of those were present in at least five CD patients and five bioresource controls.

Linear models were constructed for log2-transformed data of each compound, with the presence of Crohn’s disease and *NOD2* genotype as the independent variables. These analyses revealed significant reductions in CD patients versus controls in pentanoic acid (log2 fold change [logFC] -2.11), piperidinone (logFC -2.43), butanone (logFC -2.19) and acetone (logFC -3.90) (Table 2).When looking at the effect of carrying two of the previously defined *NOD2* mutations, there was a single significant association after correction for multiple testing using the Holm method with butanoic acid (logFC 1.25, corrected p=0.024). On further examination, this VOC was noted to be less abundant specifically in patients with Crohn’s disease with wild type *NOD2* (figure 5).

# Discussion

This prospective study examines the relationship between *NOD2* genotype and the fecal microbiota in human participants stratified by *NOD2* genotype. It further confirms previously identified shifts in gut microbiota in CD patients when compared to non-IBDcontrols, notably a reduction in obligate anaerobic lineages in tandem with an increase in the facultatively anaerobic *Enterobacteriaceae* family. These changes have previously been described in both inflamed and uninflamed tissue, and in both faecal and mucosal samples.35–38 However, no significant differences in fecal microbiota were seen when analysed by *NOD2-*status, at any of the taxonomic levels assessed. The present study also includes VOC data and demonstrates the value of having a means to assess the functional aspects of the gut microbiota, and we were able to demonstrate higher butanoic acid concentrations in Crohn’s patients with *NOD2* mutations than those without.

Earlier animal studies have shown an association between *NOD2* genotype and gut microbiota. Both Rehman *et al.* and Mondot *et al.* showed reductions in diversity and changes in specific taxa when comparing wild type and *NOD2* knockout mice.39,40 However, more recently, Shanahan *et al.* conducted experiments where the knockout and wild type mice were co-housed and failed to demonstrate a *NOD2* genotype-specific effect on gut microbiota. They concluded that the cage environment was more important than genotype. Carmody *et al.* went further and looked at the relative impact of genotype and diet on the gut microbiota in mice; they demonstrated dominant effects of diet, regardless of the underlying host genetics.41 Nonetheless, Nabhani *et al.* found *NOD2*-related differences in mucosal microbiota even when *NOD2-*knockout and wild type mouse embryos were mixed and transferred to wild type surrogates and were subsequently co-housed.42

In humans, others have previously reported an effect of *NOD2* on intestinal microbiota. Knights *et al.* reported results from cohorts comprising a total of 474 individuals with IBD, though not stratified by *NOD2* status.43 They identified an association between six causal *NOD2* variants and increased *Enterobacteriaceae* measured in intestinal biopsies. Of note, they were able to identify similar patterns in UC patients with *NOD2* mutations, suggesting that the observed effect is not just one of disease phenotype. However, in a network of associations between bacterial taxa, host and environmental factors, the effect of *NOD2* genotype was only modest compared to antibiotic usage, immunosuppressants, biopsy location and cohort of origin. Li *et al*. reported differences in intestinal biopsy microbial profile related to *NOD2* genotype alongside disease phenotype, with an increase in the C. coccoides-E. rectales group in patients with ileal Crohn’s carrying a risk *NOD2* allele.44 More recently, Imhann *et al.* reported an interaction between an IBD genetics risk score which included *NOD2* variants and the fecal microbiota, although the impact of *NOD2* on its own was not described.45Strengths of this present study include the use of patients and non-IBD controls of known *NOD2* genotype, with close matching of the phenotypic characteristics across genotypes. Establishing the causal relationship between the gut microbiota and IBD remains challenging; intestinal inflammation is well-established as a cause of dysbiosis.46 The study excluded participants with either clinical or biomarker evidence of active disease, reducing the possibility of confounding by disease activity. While the use of patients in remission will have removed one source of variability, it is also possible that the effects of *NOD2* are manifest during active disease. With regards to limitations, this study explores only the changes in gut microbiota in the fecal contents, which are unlikely to fully reflect changes at the mucosal level, particularly in the terminal ileum where one might expect *NOD2* to exert its strongest effect. This reflects the difficulty in accessing colonoscopic biopsy samples in a cohort of non-IBD controls and patients in remission. Although only a single VOC was significantly different by NOD2 status, this suggests a possibile difference in metabolically active bacteria not well-represented in faecal samples. The patient cohort also had well-established disease, with a history of surgical resection in most participants. This may reflect a higher risk of surgical resection in patients with *NOD2* mutations, noting that the wild type controls were matched using this phenotype among others. Shotgun metagenomic analysis might have facilitated detection of differences at the species or gene level between cohorts that could be missed with 16s rRNA taxonomic analysis.Future studies should explore the relationship between *NOD2* genotype and ileal-associated bacteria, ideally using either cohorts again stratified by genotype or very large cohorts to generate adequate numbers of individuals carrying two disease-associated mutations. Large cohort studies also offer the opportunity to perform more extensive genotype-microbiota-phenotype analyses, which should lead to a better understanding of these complex interactions.

# References

1. Kajiura T, Takeda T, Sakata S, et al. Change of intestinal microbiota with elemental diet and its impact on therapeutic effects in a murine model of chronic colitis. *Dig. Dis. Sci.* 2009;54:1892–1900.

2. Winslet MC, Andrews H, Allan RN, et al. Fecal diversion in the management of Crohn’s disease of the colon. *Dis. Colon Rectum*. 1993;36:757–762.

3. Frank DN, St Amand AL, Feldman R a, et al. Molecular-phylogenetic characterization of microbial community imbalances in human inflammatory bowel diseases. *Proc. Natl. Acad. Sci. U. S. A.* 2007;104:13780–5.

4. Ahmed I, Greenwood R, Costello B, et al. Investigation of faecal volatile organic metabolites as novel diagnostic biomarkers in inflammatory bowel disease. *Aliment. Pharmacol. Ther.* 2016;43:596–611.

5. Liu JZ, Sommeren S van, Huang H, et al. Association analyses identify 38 susceptibility loci for inflammatory bowel disease and highlight shared genetic risk across populations. *Nat. Genet.* 2015;47.

6. Jostins L, Ripke S, Weersma RK, et al. Host–microbe interactions have shaped the genetic architecture of inflammatory bowel disease. *Nature*. 2012;491:119–124.

7. Lange KM de, Moutsianas L, Lee JC, et al. Genome-wide association study implicates immune activation of multiple integrin genes in inflammatory bowel disease. *Nat. Genet.* 2017;49:256–261.

8. Cleynen I, Boucher G, Jostins L, et al. Inherited determinants of Crohn’s disease and ulcerative colitis phenotypes: A genetic association study. *Lancet*. 2016;387:156–167.

9. Philpott DJ, Sorbara MT, Robertson SJ, et al. NOD proteins: regulators of inflammation in health and disease. *Nat. Rev. Immunol.* 2014;14:9–23.

10. Cooney R, Baker J, Brain O, et al. NOD2 stimulation induces autophagy in dendritic cells influencing bacterial handling and antigen presentation. *Nat. Med.* 2010;16:90–7.

11. Travassos LH, Carneiro LAM, Ramjeet M, et al. Nod1 and Nod2 direct autophagy by recruiting ATG16L1 to the plasma membrane at the site of bacterial entry. *Nat. Immunol.* 2010;11:55–62.

12. Homer CR, Richmond AL, Rebert NA, et al. ATG16L1 and NOD2 interact in an autophagy-dependent antibacterial pathway implicated in crohn’s disease pathogenesis. *Gastroenterology*. 2010;139.

13. Beelen AJ van, Zelinkova Z, Taanman-Kueter EW, et al. Stimulation of the Intracellular Bacterial Sensor NOD2 Programs Dendritic Cells to Promote Interleukin-17 Production in Human Memory T Cells. *Immunity*. 2007;27:660–669.

14. Kobayashi KS, Chamaillard M, Ogura Y, et al. Nod2-dependent regulation of innate and adaptive immunity in the intestinal tract. *Science*. 2005;307:731–734.

15. MacArthur D, Palotie A, Metspalu A, et al. gnomAD browser Gene: NOD2. Available at: http://gnomad.broadinstitute.org/gene/ENSG00000167207 [Accessed June 29, 2017].

16. Yazdanyar S, Kamstrup PR, Tybjaerg-Hansen A, et al. Penetrance of NOD2/CARD15 genetic variants in the general population. *CMAJ*. 2010;182:661–5.

17. Frank DN, Robertson CE, Hamm CM, et al. Disease phenotype and genotype are associated with shifts in intestinal-associated microbiota in inflammatory bowel diseases. *Inflamm. Bowel Dis.* 2011;17:179–84.

18. Ogura Y, Bonen DK, Inohara N, et al. A frameshift mutation in NOD2 associated with susceptibility to Crohn’s disease. *Nature*. 2001;411:603–6.

19. Hugot JP, Chamaillard M, Zouali H, et al. Association of NOD2 leucine-rich repeat variants with susceptibility to Crohn’s disease. *Nature*. 2001;411:599–603.

20. Burton PR, Clayton DG, Cardon LR, et al. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature*. 2007;447:661–78.

21. Cambridge Bioresource. Introducing the Cambridge BioResource — The Cambridge NIHR BioResource. 2017. Available at: http://www.cambridgebioresource.group.cam.ac.uk/.

22. Silverberg MS, Satsangi J, Ahmad T, et al. Toward an integrated clinical, molecular and serological classification of inflammatory bowel disease: Report of a Working Party of the 2005 Montreal World Congress of Gastroenterology. *Can. J. Gastroenterol.* 2005;19 Suppl A:5–36.

23. Cardona S, Eck A, Cassellas M, et al. Storage conditions of intestinal microbiota matter in metagenomic analysis. *BMC Microbiol.* 2012;12:158.

24. Kennedy NA, Walker AW, Berry SH, et al. The impact of different DNA extraction kits and laboratories upon the assessment of human gut microbiota composition by 16S rRNA gene sequencing Sanz Y, ed. *PLoS One*. 2014;9:e88982.

25. Bernardo D, Durant L, Mann ER, et al. Chemokine (C-C Motif) Receptor 2 Mediates Dendritic Cell Recruitment to the Human Colon but Is Not Responsible for Differences Observed in Dendritic Cell Subsets, Phenotype, and Function Between the Proximal and Distal Colon. *C. Cell. Mol. Gastroenterol. Hepatol.* 2016;2:22–39.e5.

26. Schloss PD, Westcott SL, Ryabin T, et al. Introducing mothur: open-source, platform-independent, community-supported software for describing and comparing microbial communities. *Appl. Environ. Microbiol.* 2009;75:7537–41.

27. Schloss PD. MiSeq SOP - mothur. 2015. Available at: http://www.mothur.org/wiki/MiSeq\_SOP [Accessed March 20, 2015].

28. Pruesse E, Quast C, Knittel K, et al. SILVA: a comprehensive online resource for quality checked and aligned ribosomal RNA sequence data compatible with ARB. *Nucleic Acids Res.* 2007;35:7188–96.

29. Letunic I, Bork P. Interactive Tree Of Life (iTOL): an online tool for phylogenetic tree display and annotation. *Bioinformatics*. 2007;23:127–8.

30. Letunic I, Bork P. Interactive Tree Of Life v2: online annotation and display of phylogenetic trees made easy. *Nucleic Acids Res.* 2011;39:W475-8.

31. Simpson EH. Measurement of Diversity. *Nature*. 1949;163:688–688.

32. Holm S. A Simple Sequentially Rejective Multiple Test Procedure. *Scand. J. Stat.* 1979;6:65–70.

33. Aggio RB, Mayor A, Reade S, et al. Identifying and quantifying metabolites by scoring peaks of GC-MS data. *BMC Bioinformatics*. 2014;15:374.

34. Smyth GK. Limma: linear models for microarray data. In: Gentleman R, Carey V, Dudoit S, et al., eds. *Bioinformatics and Computational Biology Solutions Using {R} and Bioconductor*. New York: Springer; 2005:397–420.

35. Hold GL, Smith M, Grange C, et al. Role of the gut microbiota in inflammatory bowel disease pathogenesis: what have we learnt in the past 10 years? *World J. Gastroenterol.* 2014;20:1192–210.

36. Walker AW, Sanderson JD, Churcher C, et al. High-throughput clone library analysis of the mucosa-associated microbiota reveals dysbiosis and differences between inflamed and non-inflamed regions of the intestine in inflammatory bowel disease. *BMC Microbiol.* 2011;11:7.

37. Gevers D, Kugathasan S, Denson LA, et al. The treatment-naive microbiome in new-onset Crohn’s disease. *Cell Host Microbe*. 2014;15:382–92.

38. Papa E, Docktor M, Smillie C, et al. Non-invasive mapping of the gastrointestinal microbiota identifies children with inflammatory bowel disease. *PLoS One*. 2012;7:e39242.

39. Mondot S, Barreau F, Nabhani Z Al, et al. Altered gut microbiota composition in immune-impaired Nod2(-/-) mice. *Gut*. 2012;61:634–5.

40. Rehman A, Sina C, Gavrilova O, et al. Nod2 is essential for temporal development of intestinal microbial communities. *Gut*. 2011;60:1354–62.

41. Carmody RN, Gerber GK, Luevano JM, et al. Diet Dominates Host Genotype in Shaping the Murine Gut Microbiota. *Cell Host Microbe*. 2015;17:72–84.

42. Nabhani Z Al, Lepage P, Mauny P, et al. Nod2 Deficiency Leads to a Specific and Transmissible Mucosa-associated Microbial Dysbiosis Which Is Independent of the Mucosal Barrier Defect. *J. Crohn’s Colitis*. 2016:jjw095.

43. Knights D, Silverberg MS, Weersma RK, et al. Complex host genetics influence the microbiome in inflammatory bowel disease. *Genome Med.* 2014:1–11.

44. Li E, Hamm CM, Gulati AS, et al. Inflammatory bowel diseases phenotype, C. difficile and NOD2 genotype are associated with shifts in human ileum associated microbial composition. *PLoS One*. 2012;7:e26284.

45. Imhann F, Vich Vila A, Bonder MJ, et al. Interplay of host genetics and gut microbiota underlying the onset and clinical presentation of inflammatory bowel disease. *Gut*. 2016.

46. Buttó LF, Haller D. Dysbiosis in intestinal inflammation: Cause or consequence. *Int. J. Med. Microbiol.* 2016;306:302–9.

# Figure and table legends

## Table 1: Study demographics

### A: Whole cohort

### B: Crohn’s disease subcohort

## Table 2: Top volatile organic compounds by presence of Crohn’s disease

## Figure 1: Flow diagram of recruitment

## Figure 2: Inverse Simpson index of microbial diversity by *NOD2* status and case typeFigure 3: Hierarchical clustering by Jaccard distance metric of the 16S rRNA gene data showing differences by study group and NOD2 status. The panel on the right shows the relative proportions of the ten most prevalent bacterial families, as well as the cumulative relative proportion of all other bacteria (shown in black).

## Figure 4A: Relative abundance of the 12 most prevalent bacterial families in both Crohn’s patients and non-IBD controls. P values are corrected for multiple testing using Holm’s method across all 59 families seen in the sequencing data. Corrected p values < 0.05 are highlighted in bold.

## Figure 4B: Relative abundance of the 12 most prevalent bacterial families where samples have been grouped by diagnosis and by *NOD2* genotype. P values are corrected for multiple testing using Holm’s method across all 59 families seen in the sequencing data. Mutant *NOD2* is defined here as the presence of two Crohn’s disease associated mutations (rs2066844, rs2066845, rs2066847); wild type *NOD2* is defined as the absence of any of these mutations.

Figure 5: Concentration of butanoic acid stratified by cohort and by *NOD2* status. P values shown are uncorrected and are for Mann Whitney U tests by *NOD2* status within each cohort. Mutant *NOD2* is defined here as the presence of two Crohn’s disease associated mutations (rs2066844, rs2066845, rs2066847); wild type *NOD2* is defined as the absence of any of these mutations.