Distribution Agreement

In presenting this thesis or dissertation as a partial fulfillment of the requirements for an advanced degree from Emory University, I hereby grant to Emory University and its agents the non-exclusive license to archive, make accessible, and display my thesis or dissertation in whole or in part in all forms of media, now or hereafter known, including display on the world wide web. I understand that I may select some access restrictions as part of the online submission of this thesis or dissertation. I retain all ownership rights to the copyright of the thesis or dissertation. I also retain the right to use in future works (such as articles or books) all or part of this thesis or dissertation.

Signature:

_____ _____

Hari K. Somineni Date

Biological Insights from Integrative Genetic, Epigenetic and Microbial analysis of Inflammatory Bowel Disease

By

Hari K. Somineni

Doctor of Philosophy

Graduate Division of Biological and Biomedical Sciences

Genetics and Molecular Biology Program

_____
Subra Kugathasan, M.D.
Co-Advisor

_____
Greg Gibson, Ph.D.
Co-Advisor

_____
Alicia K. Smith, Ph.D.
Committee Member

_____
David J. Cutler, Ph.D.
Committee Member

_____
Carlos Moreno, Ph.D.
Committee Member

_____
Roger Deal, Ph.D.
Committee Member

Accepted:

_____
Lisa A. Tedesco, Ph.D.
Dean of the James T. Laney School of Graduate Studies

_____
Date

Biological Insights from Integrative Genetic, Epigenetic and Microbial analysis of Inflammatory Bowel Disease

By

Hari K. Somineni
M.S, Wright State University

Advisors: Subra Kugathasan, M.D. and Greg Gibson, Ph.D.

An abstract of
A dissertation submitted to the Faculty of the
James T. Laney School of Graduate Studies of Emory University
in partial fulfillment of the requirements for the degree of

Doctor of Philosophy

Graduate Division of Biological and Biomedical Sciences

Genetics and Molecular Biology Program

2019

ABSTRACT

Biological Insights from Integrative Genetic, Epigenetic and Microbial analysis of Inflammatory Bowel Disease

By Hari K. Somineni

Inflammatory bowel diseases, Crohn's disease and ulcerative colitis, are chronic inflammatory disorders of the gastrointestinal tract, that are therapeutically or surgically manageable but not curable. The pathogenesis of inflammatory bowel disease is hypothesized to involve complex interactions between genetic, immunologic and environmental factors, including the microbiota, that remain largely undescribed. Although some of these pathological components, including common variants and the microbiome, were extensively studied in isolation, the lack of translation of these associations into biological insights has been a roadblock to understanding disease biology and for subsequent targeted prevention and therapy. During the course of this study, we first aimed: i) to facilitate new locus discovery of common and rare variants in a population that remains understudied; ii) to define DNA methylation signatures that might play a causal role in the development of Crohn's disease; iii) to provide a state-of-the art review on the current understanding of the role of the gut microbiota in disease pathogenesis, diagnosis, and therapeutic management; and iv) to gain preliminary insights into the spatial and temporal dynamics of the oral microbiota in the pathogenesis and diagnosis of inflammatory bowel disease, and its relation to inflammation. Second, whenever possible, we performed integrative analyses of some of these pathogenetic datasets to facilitate biological insights into the underpinnings of inflammatory bowel disease, and propose that future studies could use this conceptual framework for integrating genetic, epigenetic and transcriptomic or microbial data. Lastly, based on the knowledge gained over the course of this study, and acknowledging the current gaps in our understanding, we provide a futuristic perspective on how to gain deeper biological insights in order to systematically tackle some of the over-arching objectives that have crystallized in the past decade.

Biological Insights from Integrative Genetic, Epigenetic and Microbial analysis of Inflammatory Bowel Disease

By

Hari K. Somineni
M.S, Wright State University

Advisors: Subra Kugathasan, M.D. and Greg Gibson, Ph.D.

A dissertation submitted to the Faculty of the
James T. Laney School of Graduate Studies of Emory University
in partial fulfillment of the requirements for the degree of

Doctor of Philosophy

Graduate Division of Biological and Biomedical Sciences

Genetics and Molecular Biology Program

2019

ACKNOWLEDGEMENTS

I am extremely grateful to both my advisors, Subra Kugathasan and Greg Gibson, who have provided immense support, invaluable opportunities, and lasting friendships over the past three years. This dissertation would not have been the same without your continued guidance, enthusiasm and ceaseless faith in me throughout these years. You both will forever be my role models and heroes, and I truly cherish all the interactions that we shared.

I would like to express my deepest appreciation to the members of my committee: Alicia K. Smith, David J. Cutler, Carlos Moreno, and Roger Deal for their guidance and mentorship and sharing their expertise, without which this dissertation would not have been possible.

As our longstanding close-collaborators, I truly appreciate the mentorship of Dave Cutler, Karen Conneely, and Alicia Smith. They have all provided extensive knowledge and valuable advice.

I am very appreciative to the past and present members of the Kugathasan Lab – David Okou, Anne Dodd, Chenthan Krishnakumar, Sylvia Doan, Barbara Joanna Niklinska-Schirtz, Livia Lindoso Lima, Khuong Le, Ranjit Singh Pelia, Jason Matthews, Suresh Venkateswaran, Kalifa Shabazz, Mahadev Prasad, Raguraj Chandradevan, Jarod Prince, Chantrice Rogers, Hannah Emetulu, Bernadette Martineau, Kajari Mondal and Zijun Liu – I could not have asked for nicer folks to share an office with. It's been an absolute pleasure working with you all.

To members of the Gibson Group, especially – Urko Marigorta, Biao Zeng, Dalia Arafat, Angela Mo, Ruoyu Tian and Sini Nagpal – for making my time in the Gibson Lab cheerful and scientifically stimulating.

I would like to give a heartfelt, special thanks to colleagues I've gotten to know during my PhD – Urko Marigorta, Biao Zeng, Anna Knight, Sarah Curtis, and Varun Kilaru – from whom I've learned a great deal; thank you for your positive demeanor and willingness to help with any task.

I am also thankful to Marko Bajic, Paja Sijacic, Bob Haines, Cristina Trevino, and Chris Scharer for taking significant amount of time out of their very busy schedules to help me practice my seminar talks and conference presentations.

Thank you Ben Rambo-Martin for suggesting me to check out the Kugathasan Lab for my PhD research, Binta Jalloh for being my buddy since the recruitment weekend in 2015.

I would also like to thank the many friends and colleagues I've gotten to know during my time at Emory – Edwin Corgiat, Sarah Curtis, Trenell Mosley, and everyone in GMB-2015 cohort.

I am especially indebted to Livia Lindoso Lima, Chenthan Krishnakumar, Sylvia Doan, Khuong Le, Raguraj Chandradevan, and Anne Dodd for putting up with me; thank you for providing me with all your friendships, kindness and companionships – you became family. Thank you for patiently listening to my daily rants. Cheers for all the lunchtime banter, second shift strolls, and especially the dinnertime banter.

Lastly, I want to thank my family, especially my parents, grandparents and my dear brother, Chaitanya Somineni for the unconditional love, support, understanding and encouragement that they have shown me in everything that I do.

TABLE OF CONTENTS

TABLE OF TABLES

TABLE OF FIGURES

Chapter 1

Overview and Current Understanding of Inflammatory Bowel Disease

INTRODUCTION

Clinical and epidemiological overview of inflammatory bowel disease. Inflammatory bowel disease is a chronic, life-long condition characterized by intestinal ulceration, pain, rectal bleeding, loss of quality of life and a need for bowel surgery. Crohn's disease and ulcerative colitis are the two classical forms of inflammatory bowel disease. Although the peak age-of-onset is between the 20's and 30's, inflammatory bowel disease can present at any stage of life; pediatric, adolescent or any particular stage of adulthood[1-3]. Similarly, its prevalence is becoming increasingly evident across all populations; Europeans, African Americans, Asians, and Latinos *etc*. In the U.S. itself, there are currently about 1.6 million people (~0.5% of the total population) suffering with this debilitating disease, and in general, inflammatory bowel disease affects about 300 of every 100,000 individuals. Although its prevalence has plateaued in western countries[4], current trends point toward the emergence of inflammatory bowel disease as an epidemic in developing nations[1,5-7].

Inflammatory bowel diseases are thought to arise in the context of complex interactions between genetic, environmental, microbial and immunological factors – most of which are yet to be identified[8]. These interactions result in an overwhelming complexity arguing against studying each of these pathogenic components in isolation. In addition, the intricate, bi-directional, dynamic interactions between disease and inflammation further add to this complexity – necessitating the need to sift through signatures of inflammation to analyze causes of inflammatory bowel disease[9].

Genetics of inflammatory bowel disease. Familial and twin studies lead to estimates that the heritability of inflammatory bowel disease is ~30%[10], making genetic liability as the single strongest known risk factor for developing this complex disorder. Despite both forms of inflammatory bowel disease being transmitted genetically, the heritable component is relatively stronger in the case of probands with Crohn's disease when compared to ulcerative colitis[10]. Attempts to improve our understanding of the genetic basis of inflammatory bowel disease have increased exponentially both in frequency and sample size over the past decade[8,11,12]. In particular, genome-wide association studies (GWAS) have been successful in identifying

about 240 loci that are associated with an increased inflammatory bowel disease risk[8,11,12]. While two-thirds of these risk loci are shared between the two forms of inflammatory bowel disease, genetic effects at the remaining loci were found to be specific to either Crohn's disease or ulcerative colitis[8,11,12].

Inflammatory bowel disease has been at the forefront of common complex diseases in regard to the advancements made through genetic studies[8,11,12]; the number of inflammatory bowel disease risk loci identified by far surpasses the germline common variants mapped for any single polygenic disease, and fine-mapping efforts have already been successful in resolving some of these risk loci (n = 45; 18 associations with 95% certainty; and an additional 27 loci with >50% certainty) to single causal substitutions, while efforts are underway to hone in on causal variants at the remaining loci[13].

Despite these unparalleled successes, there are three critical issues that need to be addressed in order to translate genetics of inflammatory bowel disease into biological insights to facilitate genome-based personalized implementation of targeted prevention and therapy. First, despite the incredible success in identifying several robust and replicable GWAS-associations, all these genetic factors cumulatively explain only a small fraction of phenotypic variance – 13% for Crohn's disease and 8% for ulcerative colitis – indicating that a majority of the genetic contributions to inflammatory bowel disease are yet to be uncovered[8,11,12]. Second, a vast majority of the known risk loci span several kb in length, containing, at times hundred to several thousands of highly correlated variants – presenting a key challenge in prioritizing variants or identifying causal variants within GWAS-associated regions[13]. Third, about 90% of the established risk loci reside in non-coding regions – presenting a key challenge in identifying the relevant genes that the prioritized or causal variants act upon at a disease associated locus[8,11-13].

Environmental component of inflammatory bowel disease. There is growing evidence that suggests that environmental factors play a prominent role in inflammatory bowel disease predisposition, incidence and maintenance. Rapid increases in the incidence of inflammatory bowel disease in developed nations during the second half of the 20th century[4], and its rising prevalence in developing countries parallels

westernization of lifestyle and industrialization[1,5-7]. Several studies have shown an epidemiological evidence of association of environmental factors, including urbanization, westernized diet, air pollution, smoking, and exposure to antibiotics with inflammatory bowel disease[14-19]; however, the underlying mechanism behind these remain unknown. Therefore, translation of this surging epidemiological evidence into mechanisms of inflammatory bowel disease is of great interest, and may aid in both understanding the underlying pathophysiology of this complex disease and in developing novel therapeutic interventions.

Epigenetics of inflammatory bowel disease. It is also becoming increasingly evident that gene products and the by-products of environmental insult often interact at the molecular level, and hence, considering gene-environment interactions may improve our understanding of the causes of complex disease, and can assist in developing targeted therapies. Epigenetic regulation in inflammatory bowel disease patients has recently become an intensely studied area because of its potential in mediating gene-environmental interactions.

DNA methylation is one of several epigenetic modifications that has been shown to play a role in mediating the impact of environmental exposures on the risk of various complex diseases, including inflammatory bowel disease. Covalent addition or removal of a chemical methyl group to the 5th position of Cytosine (C) when followed by a Guanine (G) can regulate gene expression without changing the DNA sequence, thereby influencing the molecular phenotypes of complex diseases. Disruption of methylation patterns is a characteristic feature in many biological processes including inflammation and associated diseases such as asthma, psoriasis, atopic eczema, and inflammatory bowel disease[20,21].

Previous studies have linked site-specific DNA methylation differences in blood and intestinal mucosal biopsies to inflammatory bowel disease[20-29]. However, these associations cannot be assumed to causally underlie disease development, unless proven, as DNA methylation modifications are vulnerable to confounding and reverse causation. Therefore, distinguishing disease-associated methylation signatures that are causal to disease from those that result from disease and disease-related clinical characteristics is

critical in order to establish whether the methylome plays a causal role and hence can be leveraged for therapeutic benefits.

The microbiome of inflammatory bowel disease. Gut microbial dysbiosis, a change in the composition or function of the microbiota, is one of the very well recognized factors in the initiation of inflammatory bowel disease[30-34]. Dysbiotic states of the microbiome may serve as an environmental stimulus in altering the host's mucosal defenses and trigger immune responses. Changes in the structure and function of the microbiota has long been shown to be associated with inflammatory bowel disease using cross-sectional as well as longitudinal investigations[31-35]. Consequently, there has been a surge of interest in microbiome-based drug development as a therapeutic means to achieve and sustain remission of disease; however, the causal role of dysbiotic microbial status in inflammatory bowel disease as well as the beneficial role of its therapeutic modulation, remains controversial[36].

The most convincing evidence for a role for the microbiome in inflammatory bowel disease originates from GWAS associations; genes implicated by GWAS signals of inflammatory bowel disease, *ATG16L1* and *NOD2*, were consistent with the notion that disease-susceptibility variants may contribute to disease pathology via defects in sensing protective signals from the microbiome[8,37]. Further, various microbiome-based case-control studies have shown that disease-specific microbial signature exists, at both global level – represented by changes in the alpha- and/or beta-diversity, and at the individual microbial member level – indicated by the shift in abundance of commensal and pathogenic microbes and their by-products[30-34]. For instance, both forms of inflammatory bowel disease have been linked to reduced overall gut microbial diversity and richness; whereas, depletion of the members of the phylum Firmicutes and expansion of the members of the Proteobacteria phylum were robustly and reproducibly found to be associated with inflammatory bowel disease, along with emerging evidence for several other microbial taxa. Similarly, lower levels of butyrate[31,38-40] and elevated levels of hydrogen sulfide[41-45] has commonly been noted in patients with inflammatory bowel disease relative to healthy individuals. However, despite these associative lines of evidence, systematic investigations into potential underlying causal relationships have not yet been

performed to understand the precise role of microbial dysbiosis in the pathology of inflammatory bowel disease; this, may in part, have contributed to the initial disappointment of the much-hyped exploitation of microbiome-centric interventions for therapeutic benefits of inflammatory bowel disease[36].

On the other hand, non-invasive microbiome-based approaches have proven to be successful in diagnosing, monitoring and stratifying risk for patients with inflammatory bowel disease[31,35,46]. Fecal samples closely mirror the microbial status of the gut; this intricate relationship has provided a strong rationale to leverage fecal microbiota as a non-invasive approach in diagnosing inflammatory bowel disease. Several studies of fecal microbial composition or function have shown that inflammatory bowel disease cases can be distinguished from healthy controls with an area under the curve approaching 0.8 and above – an arbitrary cut-off that is deemed to be clinically meaningful. Surprisingly, at times, it has outperformed the most commonly used non-invasive diagnostic assays of fecal calprotectin. More importantly, data from the latest studies highlight the potential of fecal microbiota in distinguishing inflammatory bowel disease subtypes, therapy responders from non-responders, and even in predicting future clinical outcomes[35,47,48]. While translating these findings into the clinical setting to help with uncertain clinical-decision making is a work in progress, there is mounting interest in exploiting microbiome data obtained from other more easily accessible tissue sources as a noninvasive strategy to enable more accurate diagnoses of inflammatory bowel disease.

Integration of distinct data types (genetics, molecular, microbial, and clinical data) in inflammatory bowel disease. While studying each of the distinct data types, including genetics, epigenetics or microbiome, in isolation provides different and partly independent insights into their potential role in disease pathology, careful retrospective evaluation of the existing gaps in our knowledge stresses the need for more information than provided by each of the individual data sets. For instance, despite the decade long successes in identifying numerous genetic variants associated with inflammatory bowel disease subtypes, their annotation and biological interpretation remains challenging. Similarly, in spite of the overwhelmingly emerging molecular and microbial signatures in inflammatory bowel disease, delineating

the causal versus consequential nature of these associations has been a roadblock in translating such association signals into biological insights.

Unifying data from different sources is, therefore, an important part of understanding the etiopathogenesis of complex diseases. Integrative analyses of a combination of genetic, environmental (epigenetic and/or microbiome), and clinical data may aid in providing a more comprehensive mechanistic view, enabling genome-based personalized implementation of targeted prevention and therapy. However, integrating different types of data with a unifying background to pinpoint causal alterations and their functional consequences is a challenging task.

Overview of current study. Over the course of this study, we attempt to understand the etiopathogenesis of inflammatory bowel disease at genetic, epigenetic and microbial levels; first by studying these data types in isolation, and then using integrative approaches whenever possible. **In chapter 2**, using whole-genome sequence data from a total of 3418 American subjects with African ancestry, we first performed both common and rare variant scans to facilitate new locus discovery of alleles that are either specific to inflammatory bowel disease patients of African descent or shared across divergent populations. By surveying variants that were not previously studied, we implicate two new genes in inflammatory bowel disease that are specific to African Americans. By performing a comparative analysis, we conclude that while the genetic risk of inflammatory bowel disease conferred by common to low-frequency variants is shared across populations, rare variant contributions exhibit population-specific effects.

In chapter 3, surveying methylation profiles in DNA from blood samples obtained from 74 non-inflammatory bowel disease controls (controls) and 164 newly diagnosed, treatment naïve pediatric patients with Crohn's disease, we identified 1189 5'-cytosine-phosphate-guanosine-3' (CpG) sites that associate with Crohn's disease at diagnosis. We provide convincing evidence that these blood-based DNA methylation signatures of Crohn's disease capture inflammatory status of the patient. Then, using longitudinal samples obtained from the same subjects, we demonstrate that the disrupted DNA methylation

patterns at diagnosis of Crohn's disease revert back to normal during the course of treatment of inflammation. Finally, we mapped the temporal dynamics of DNA methylation and relapsing-remitting disease behaviors of Crohn's disease, and then supplemented this longitudinal framework with genetic association and the concept of Mendelian randomization (MR), to define methylation changes that causally contribute to Crohn's disease development. Our proposed conceptual framework for integrating genetic, epigenetic and clinical data, can be a useful approach to identify methylation changes that causally underlie various complex diseases.

In chapter 4, we provide a state-of-the-art review of some of the current human inflammatory bowel disease microbiome findings, describe the cause-effect relationships between the gut microbiome and inflammatory bowel disease, and discuss the possibility of using microbiome-based approaches in the diagnosis, therapy, and management of disease. In addition, the potential role of microbiome-based interventions in the treatment of human inflammatory bowel disease is also discussed.

In chapter 5, using a prospectively recruited cohort of pediatric patients with inflammatory bowel disease (n = 47) and unrelated healthy controls (n = 18), we examine the spatial and temporal distribution of microbiota within the various oral microenvironments, represented by saliva, tongue, buccal mucosa and plaque, and compared them with stool, to test if oral samples are indicative of inflammatory bowel disease, and if so, which type of oral sample is the most informative. We further assessed to what extent gut and oral microbial disease markers converge in terms of their composition in inflammatory bowel disease.

In chapter 6, based on our findings over the course of this study, we provide recommendations for future studies.

REFERENCES

1. Benchimol, E.I. *et al.* Increasing incidence of paediatric inflammatory bowel disease in Ontario, Canada: evidence from health administrative data. *Gut* **58**, 1490-7 (2009).
2. Van Limbergen, J. *et al.* Definition of phenotypic characteristics of childhood-onset inflammatory bowel disease. *Gastroenterology* **135**, 1114-22 (2008).
3. Ruel, J., Ruane, D., Mehandru, S., Gower-Rousseau, C. & Colombel, J.F. IBD across the age spectrum: is it the same disease? *Nat Rev Gastroenterol Hepatol* **11**, 88-98 (2014).
4. Manichanh, C., Borruel, N., Casellas, F. & Guarner, F. The gut microbiota in IBD. *Nat Rev Gastroenterol Hepatol* **9**, 599-608 (2012).
5. Benchimol, E.I. *et al.* Changing age demographics of inflammatory bowel disease in Ontario, Canada: a population-based cohort study of epidemiology trends. *Inflamm Bowel Dis* **20**, 1761-9 (2014).
6. Kaplan, G.G. The global burden of IBD: from 2015 to 2025. *Nat Rev Gastroenterol Hepatol* **12**, 720-7 (2015).
7. Ng, S.C. *et al.* Worldwide incidence and prevalence of inflammatory bowel disease in the 21st century: a systematic review of population-based studies. *Lancet* **390**, 2769-2778 (2018).
8. Jostins, L. *et al.* Host-microbe interactions have shaped the genetic architecture of inflammatory bowel disease. *Nature* **491**, 119-24 (2012).
9. Somineni, H.K. *et al.* Blood-derived DNA Methylation Signatures of Crohn's Disease and Severity of Intestinal Inflammation. *Gastroenterology* **In press**(2019).
10. Halme, L. *et al.* Family and twin studies in inflammatory bowel disease. *World J Gastroenterol* **12**, 3668-72 (2006).
11. Liu, J.Z. *et al.* Association analyses identify 38 susceptibility loci for inflammatory bowel disease and highlight shared genetic risk across populations. *Nat Genet* **47**, 979-986 (2015).
12. de Lange, K.M. *et al.* Genome-wide association study implicates immune activation of multiple integrin genes in inflammatory bowel disease. *Nat Genet* **49**, 256-261 (2017).
13. Huang, H. *et al.* Fine-mapping inflammatory bowel disease loci to single-variant resolution. *Nature* **547**, 173-178 (2017).
14. De Filippo, C. *et al.* Impact of diet in shaping gut microbiota revealed by a comparative study in children from Europe and rural Africa. *Proc Natl Acad Sci U S A* **107**, 14691-6 (2010).
15. Schnorr, S.L. *et al.* Gut microbiome of the Hadza hunter-gatherers. *Nat Commun* **5**, 3654 (2014).
16. Yatsunenko, T. *et al.* Human gut microbiome viewed across age and geography. *Nature* **486**, 222-7 (2012).
17. Martinez, I. *et al.* The gut microbiota of rural papua new guineans: composition, diversity patterns, and ecological processes. *Cell Rep* **11**, 527-38 (2015).
18. Bernstein, C.N. & Shanahan, F. Disorders of a modern lifestyle: reconciling the epidemiology of inflammatory bowel diseases. *Gut* **57**, 1185-91 (2008).
19. Hviid, A., Svanstrom, H. & Frisch, M. Antibiotic use and inflammatory bowel diseases in childhood. *Gut* **60**, 49-54 (2011).
20. Ventham, N.T. *et al.* Integrative epigenome-wide analysis demonstrates that DNA methylation may mediate genetic risk in inflammatory bowel disease. *Nat Commun* **7**, 13507 (2016).
21. Howell, K.J. *et al.* DNA Methylation and Transcription Patterns in Intestinal Epithelial Cells From Pediatric Patients With Inflammatory Bowel Diseases Differentiate Disease Subtypes and Associate With Outcome. *Gastroenterology* **154**, 585-598 (2018).
22. Karatzas, P.S., Gazouli, M., Safioleas, M. & Mantzaris, G.J. DNA methylation changes in inflammatory bowel disease. *Ann Gastroenterol* **27**, 125-132 (2014).
23. McDermott, E. *et al.* DNA Methylation Profiling in Inflammatory Bowel Disease Provides New Insights into Disease Pathogenesis. *J Crohns Colitis* **10**, 77-86 (2016).
24. Li Yim, A.Y.F. *et al.* Peripheral blood methylation profiling of female Crohn's disease patients. *Clin Epigenetics* **8**, 65 (2016).

25. Nimmo, E.R. *et al.* Genome-wide methylation profiling in Crohn's disease identifies altered epigenetic regulation of key host defense mechanisms including the Th17 pathway. *Inflamm Bowel Dis* **18**, 889-99 (2012).
26. Adams, A.T. *et al.* Two-stage genome-wide methylation profiling in childhood-onset Crohn's Disease implicates epigenetic alterations at the VMP1/MIR21 and HLA loci. *Inflamm Bowel Dis* **20**, 1784-93 (2014).
27. Harris, R.A. *et al.* DNA methylation-associated colonic mucosal immune and defense responses in treatment-naive pediatric ulcerative colitis. *Epigenetics* **9**, 1131-7 (2014).
28. Harris, R.A. *et al.* Genome-wide peripheral blood leukocyte DNA methylation microarrays identified a single association with inflammatory bowel diseases. *Inflamm Bowel Dis* **18**, 2334-41 (2012).
29. Taman, H. *et al.* Genome-wide DNA Methylation in Treatment-naive Ulcerative Colitis. *J Crohns Colitis* **12**, 1338-1347 (2018).
30. Frank, D.N. *et al.* Molecular-phylogenetic characterization of microbial community imbalances in human inflammatory bowel diseases. *Proc Natl Acad Sci U S A* **104**, 13780-5 (2007).
31. Gevers, D. *et al.* The treatment-naive microbiome in new-onset Crohn's disease. *Cell Host Microbe* **15**, 382-392 (2014).
32. Sokol, H. *et al.* Faecalibacterium prausnitzii is an anti-inflammatory commensal bacterium identified by gut microbiota analysis of Crohn disease patients. *Proc Natl Acad Sci U S A* **105**, 16731-6 (2008).
33. Darfeuille-Michaud, A. *et al.* Presence of adherent Escherichia coli strains in ileal mucosa of patients with Crohn's disease. *Gastroenterology* **115**, 1405-13 (1998).
34. Petersen, A.M. *et al.* A phylogenetic group of Escherichia coli associated with active left-sided inflammatory bowel disease. *BMC Microbiol* **9**, 171 (2009).
35. Shaw, K.A. *et al.* Dysbiosis, inflammation, and response to treatment: a longitudinal study of pediatric subjects with newly diagnosed inflammatory bowel disease. *Genome Med* **8**, 75 (2016).
36. Somineni, H.K. & Kugathasan, S. The Microbiome in Patients With Inflammatory Diseases. *Clin Gastroenterol Hepatol* **17**, 243-255 (2019).
37. Chu, H. *et al.* Gene-microbiota interactions contribute to the pathogenesis of inflammatory bowel disease. *Science* **352**, 1116-20 (2016).
38. Takaishi, H. *et al.* Imbalance in intestinal microflora constitution could be involved in the pathogenesis of inflammatory bowel disease. *Int J Med Microbiol* **298**, 463-72 (2008).
39. Sokol, H. & Seksik, P. The intestinal microbiota in inflammatory bowel diseases: time to connect with the host. *Curr Opin Gastroenterol* **26**, 327-31 (2010).
40. Sokol, H. *et al.* Low counts of Faecalibacterium prausnitzii in colitis microbiota. *Inflamm Bowel Dis* **15**, 1183-9 (2009).
41. Loubinoux, J., Bronowicki, J.P., Pereira, I.A., Mougenel, J.L. & Faou, A.E. Sulfate-reducing bacteria in human feces and their association with inflammatory bowel diseases. *FEMS Microbiol Ecol* **40**, 107-12 (2002).
42. Zinkevich, V.V. & Beech, I.B. Screening of sulfate-reducing bacteria in colonoscopy samples from healthy and colitic human gut mucosa. *FEMS Microbiol Ecol* **34**, 147-155 (2000).
43. Verma, R., Verma, A.K., Ahuja, V. & Paul, J. Real-time analysis of mucosal flora in patients with inflammatory bowel disease in India. *J Clin Microbiol* **48**, 4279-82 (2010).
44. Mills, D.J. *et al.* Dietary glycated protein modulates the colonic microbiota towards a more detrimental composition in ulcerative colitis patients and non-ulcerative colitis subjects. *J Appl Microbiol* **105**, 706-14 (2008).
45. Pitcher, M.C., Beatty, E.R. & Cummings, J.H. The contribution of sulphate reducing bacteria and 5-aminosalicylic acid to faecal sulphide in patients with ulcerative colitis. *Gut* **46**, 64-72 (2000).
46. Halfvarson, J. *et al.* Dynamics of the human gut microbiome in inflammatory bowel disease. *Nat Microbiol* **2**, 17004 (2017).

47. Haberman, Y. *et al.* Pediatric Crohn disease patients exhibit specific ileal transcriptome and microbiome signature. *J Clin Invest* **124**, 3617-33 (2014).
48. Ananthakrishnan, A.N. *et al.* Gut Microbiome Function Predicts Response to Anti-integrin Biologic Therapy in Inflammatory Bowel Diseases. *Cell Host Microbe* **21**, 603-610 e3 (2017).

Chapter 2

Whole-Genome Sequencing of African Americans Identifies Novel Rare Variants Associated with Inflammatory Bowel Disease

This chapter has been adapted from a manuscript in preparation.

Hari K. Somineni, Talin Haritunians, Claire L. Simpson, David J. Cutler, David T. Okou, Yuval Itan, Suresh Venkateswaran, Christine Stevens, Lisa W. Datta, Tanvi A. Dhere, Mark G. Lazarev, Multicenter African American Inflammatory Bowel Disease Study, Emory African American Inflammatory Bowel Disease Consortium, Michael E. Zwick, Greg Gibson, Judy H. Cho, Mark J. Daly, Dermot P. B. McGovern, Steven R. Brant, and Subra Kugathasan

ABSTRACT

Here we performed the first and the most comprehensive whole-genome sequencing study of 1774 patients with inflammatory bowel disease and 1644 matched controls from Americans with African ancestry, to facilitate new locus discovery and interrogate the contribution of both common and rare variation in this understudied population. We detected aggregate associations of rare, likely deleterious variants in genes not previously associated with inflammatory bowel disease. We implicate for the first time an ATPase, *ATP1A4*, involved in the maintenance of Na^+ and K^+ electrochemical gradients in Crohn's disease and inflammatory bowel disease; and a Ca^{2+} binding neuro-immunomodulator, *CALB2*, in ulcerative colitis. While our results support an overall overlap of common variant risk for inflammatory bowel disease susceptibility between individuals with African and European ancestries, they highlight the possibility of population specificity in rare variant contributions to inflammatory bowel disease risk.

INTRODUCTION

Inflammatory bowel diseases, Crohn's disease and ulcerative colitis, arise in the context of an inappropriate activation of the intestinal immune system in response to an environmental trigger in individuals who are genetically predisposed. Genome-wide association studies (GWAS) of common and low-frequency variants have so far identified 240 loci that confer significant risk for disease susceptibility[1-4]. Despite inflammatory bowel disease being one of the most successfully studied polygenic diseases with respect to identifying many risk alleles, three major challenges remain, including: (i) missing heritability – as only a small fraction of disease liability is explained by the thus far known genetic risk factors (13% for Crohn's disease and 8% for ulcerative colitis)[2]; (ii) lack of molecular insights into established genetic signals – as a vast majority of these loci (~90%) localize to non-coding regions; and (iii) lack of causal insights into underlying GWAS associations – as a majority of the known risk loci span several kb in length, containing, credible sets of hundreds to thousands of highly correlated variants that depict similar evidence of association with disease.

Genetic discoveries of inflammatory bowel disease have been made primarily in populations of European ancestry and utilizing genome-wide genotype data[1-4]. This predominance, combined with a focus primarily on common alleles has left our understanding of the role of rare variants and alleles restricted to non-European populations incomplete. To this end, we have performed the first and the most comprehensive whole-genome sequencing study of a total of 3610 cases and matched controls from Americans with African ancestry. Our goal was two-fold: first, we hypothesized that genetic analysis of this understudied population would facilitate new locus discovery of common variants that are either specific to African-ancestry populations or shared across divergent populations. Second, given the high genetic diversity in African populations, we hypothesized that rare variants within or near protein coding genes contribute to inflammatory bowel disease risk that have yet to be identified.

METHODS

Study samples. This was a multi-center collaborative study involving self-identified African American subjects recruited from five primary sites and their collaborating centers across the US. These sample recruitment centers include: Emory University (recruited as part of the GENESIS study and Emory African American Inflammatory Bowel Disease Consortium) and 12 other collaborating centers; Johns Hopkins/Rutgers (recruited as part of the Multicenter African American Inflammatory Bowel Disease Study) and 17 other collaborating centers; Cedars Sinai Medical Center; Mount Sinai Medical center, and Washington University (recruited as part of the Centers for Common Disease Genomics network). Sample breakdown, along with the proportion of cases vs controls, per center is shown in **Table 2-1**.

Whole-genome sequencing. All DNA samples investigated in this study (a total of 3,610 before quality control) were sequenced at the Broad Institute of Harvard and MIT (Cambridge, MA) following the same protocol. On an average, each sample was sequenced to a depth of 30x. Sequences were aligned to human reference genome build hg38 (GRCh38 assembly). Variants were called jointly using the Genome Analysis Toolkit (GATK) pipeline[5], and were annotated using our in-house Bystro[6] software. After sample quality

control procedures, we excluded a total of 192 samples. This included samples with sex discrepancies (n = 35), missing phenotypes (n = 3), duplicated samples or related individuals (n = 122), missing variant data (n = 42), outlying heterozygotic/homozygotic changes (n = 44), theta (n = 12), exonic theta (n = 13), exonic transition/transversion (n = 1), silent/replacement (n = 1), silent transition/transversion (n = 1), and replacement transition/transversion (n = 1). Following sample quality control, we filtered out variants with missingness > 5%, and those that showed a significant deviation from Hardy-Weinberg equilibrium in controls ($P < 1 \times 10^{-9}$). These procedures resulted in a final dataset of 3,418 samples and 93.4 million variants that include both SNPs and short insertion-deletions (indels).

Principal component analysis of sequence data. After excluding samples and variants with low quality, principal component analysis of whole-genome sequencing dataset was performed using EIGENSTRAT[7]. Principal components were computed based on a pruned version of the dataset consisting of 1.8 million LD-independent ($r^2 < 0.1$), high frequency (minor-allele frequency (MAF) > 1%) variants. The first five principal components were included as covariates to control for population stratification within the whole-genomes dataset for all analyses (**Fig. 2-1**).

Common variation association testing of sequence data. We defined common variants as those that are present in at least 1% of the general African population from the gnomAD database, and have an observed MAF > 1% in this dataset, yielding 14.9 million variants. We used a logistic regression model to test for association at these variants with the first five principal components of the genotype matrix included as covariates. Variants were separately tested for association with Crohn's disease, ulcerative colitis and inflammatory bowel disease. Genomic control (λ_{GC}) values ranged from 1.102 to 1.141, indicating little or no inflation or deflation due to population stratification.

Rare variation association testing of sequence data. We defined rare variants as those that are either absent or present at a MAF of < 0.1% in general African population from the gnomAD database. With these criteria, we observed 64.2 million rare variants in our dataset. *Individual rare variant association testing*.

We tested each individual rare variant for association with Crohn's disease, ulcerative colitis and inflammatory bowel disease, separately, using a logistic regression framework conditioned on the first five genotypic principal components. Genomic control (λ_{GC}) values for these individual analyses of rare variants ranged from 0.612 to 0.841, indicating deflation due to the limited number of rare alleles in the dataset. *Aggregate rare variant association testing.* Using the optimal sequence kernel association test (SKAT-O)[8], we performed both gene-wide and window-wide analyses to detect aggregate association of rare, likely deleterious (Combined Annotation Dependent Depletion (CADD) > 15) variants with the three traits. For aggregate tests, we selected all rare, likely deleterious (CADD > 15) variants (n = 1.5 million) across the genome and assigned them to the nearest gene for gene-wide analysis or aggregated into sets of certain length of 10, 20, 30, 40, 50, 75, and 100 kb, based on their physical location in the genome. We then assessed the association of each set in a SKAT-O model implemented in the R package 'SKAT'. Quantile-quantile plots and genomic control (λ_{GC}) values of aggregate association tests indicating no inflation or deflation are presented in **Figs 2-5 – 2-7**. To interpret statistical significance, we applied experimental-wide, Bonferroni-corrected significance thresholds of $P < 2.2 \times 10^{-6} = 0.05/22{,}521$ for gene-wide analysis, $P < 2.5 \times 10^{-7} = 0.05/201{,}672$ for 10 kb windows, $P < 3.6 \times 10^{-7} = 0.05/139{,}546$ for 20 kb windows, $P < 5.9 \times 10^{-7} = 0.05/84{,}298$ for 30 kb windows, $P < 7.7 \times 10^{-7} = 0.05/64{,}617$ for 40 kb windows, $P < 9.6 \times 10^{-7} = 0.05/52{,}245$ for 50 kb windows, $P < 1.4 \times 10^{-6} = 0.05/35{,}209$ for 75 kb windows, and $P < 2.3 \times 10^{-6} = 0.05/21{,}410$ for 100 kb windows.

GWAS genotype data, quality control, imputation and association testing. Sample information, genotype data, and the application of quality control procedures for the two existing GWAS cohorts considered in the current study were described extensively elsewhere[9]. Briefly, genome-wide genotype data from non-overlapping African American cases and matched controls generated using either the Illumina Omni Array (398 Crohn's disease, 238 ulcerative colitis and 1551 controls) or the Affymetrix Axiom Genome-Wide AFR 1 World Array (451 Crohn's disease, 186 ulcerative colitis and 3038 controls) SNP chips were considered for replicative evidence. Sample and variant quality control, determination of

principal components, removal of outliers was done as described in the original paper[9]. Both datasets were lifted from human reference build hg19 to hg38 using liftOver. *Imputation*. The whole-genome sequences described above (n = 3,418; after quality control) were phased with Eagle v2.4[10] to create a reference panel. These pre-phased whole genome sequences with MAF > 0.5% were imputed into each GWAS dataset, separately, via minimac3 software[11]. By design, all the sequenced individuals are of African descent, and about half of these are inflammatory bowel disease cases, thereby enriching the reference panel for African-specific alleles that increase or decrease inflammatory bowel disease risk. *Common variant association testing for replicative evidence*. After removing samples that were directly sequenced in the discovery phase, genotyped and imputed variants with INFO score > 0.6 were tested for association with Crohn's disease, ulcerative colitis and inflammatory bowel disease, separately, within each GWAS case-control dataset using SNPTEST 2.5.2[12], performing an additive frequentist association test conditioned on the first ten principal components. For sites that were present in both the datasets, and passed our quality control filters, we performed meta-analysis using METAL[13]. For a common variant with genome-wide significance of $P < 5 \times 10^{-08}$ in the discovery cohort, to be inferred to be associated with a trait, it has to have a directionally consistent effect and demonstrate at least a nominal evidence of association ($P < 0.05$) in the meta-analysis of the two GWAS datasets.

Genetic risk score calculation. We used the --*score* function available in plink to compute weighted risk scores for all the individuals in our whole-genome sequencing dataset using a model derived from the observed genotypes and allele dosage at variants of interest, and their corresponding effect sizes from large meta-analyses. For genetic risk score, we considered a model derived from the sentinel SNPs from each of the established inflammatory bowel disease risk loci observed in the recent meta-analyses of GWASs in participants of primarily European ancestry. To compute genome-wide genetic risk scores based on relevance to inflammatory bowel disease in African Americans, we divided our whole-genome sequencing cohort, at random, into a training set with 70% of the samples and the remaining 30% samples retained as a test set. We tested for association with inflammatory bowel disease at all 14.9 million common variants

in the training dataset using a logistic regression model conditioned on the first five principal components, and selected all the variants (n = 1.2 million) that exhibited at least nominal evidence ($P < 0.05$) for association with inflammatory bowel disease along with their corresponding effect sizes to score the genome-wide genetic risk of inflammatory bowel disease in remaining individuals in the test dataset.

RESULTS

After quality control and principal component analysis (**Fig. 2-1**) of our deeply sequenced whole-genomes (median coverage of 30x), we present analyses of a total of 3418 subjects; 1774 cases (1335 Crohn's disease; 407 ulcerative colitis; and 32 inflammatory bowel disease-unknown) and 1644 matched controls (**Table 2-1**), at 93 million variants that comprise both SNPs and short insertion-deletion variants (indels). These data include 14.9 million common variants with minor-allele frequency (MAF) > 1% that were individually tested for association with Crohn's disease, ulcerative colitis and inflammatory bowel disease (Crohn's disease and ulcerative colitis together with inflammatory bowel disease-unknown) in a logistic regression framework conditioned on the first five principal components (see Methods). Following these discovery analyses of common variation, we sought replication of the obtained results in an independent cohort of African Americans that were previously genotyped using Axiom or Omni genome-wide SNP arrays[9] (Methods; **Table 2-2**). Briefly, following quality control, we imputed our whole-genome sequences into these two existing GWAS datasets, thereby enriching the panel for inflammatory bowel disease risk alleles; and performed case-control association testing using a logistic regression model, separately, within each dataset. Results from the meta-analysis of these two GWAS datasets served as our replicative evidence.

We identified 2 independent loci associated with decreased risk of Crohn's disease in African Americans, including 22 intergenic variants near *PTGER4* (~260 kb) and a novel intronic variant in *KIF1B* (**Table 2-3**). Following our previous report of suggestive evidence of association at *PTGER4* locus for Crohn's disease in African Americans[9], here we present the first evidence of genome-wide significance. All 22 variants in *PTGER4* locus were consistent with exerting a protective effect, and are in strong linkage

disequilibrium (LD) with each other (r^2 > 0.8; **Fig. 2-2**). While the newly discovered intronic variant in *KIF1B* also associated with inflammatory bowel disease, variants in *PTGER4* locus did not reach genome-wide significance in the combined discovery cohort.

A protective role in Crohn's disease of *PTGER4* locus variants has previously been implicated in populations of European ancestry[14], with a total of 2,819 common variants at the locus depicting genome-wide significant association[15]. Using fine-mapping, this region was further refined to a subset of 189 credible variants representing four independent signals that are more likely to be causal to Crohn's disease[15] (**Fig. 2-3**). The 22 variants that we detected in this locus in African Americans are in high LD with the strongest signal (signal 1) comprising 2 variants – rs7711427 and rs397897680 – from the fine-mapping analysis in European populations[15] (**Fig. 2-4**). We note that, while rs397897680 was not called in our dataset, rs7711427 had been excluded during our initial quality control procedure (see Methods).

With the sequencing data, we next assessed the contribution of rare variants (MAF < 0.1%) to inflammatory bowel diseases, both individually and in aggregate. Our data was comprised of 64.2 million rare variants that include many alleles that were not genotyped or imputed in previous GWAS of inflammatory bowel disease. For aggregate gene-wide investigations, first, we selected all rare, likely deleterious (CADD > 15) variants across the genome and assigned them to the nearest gene. In total, 1.5 million such variants were assigned to 22,521 genes with an average of 68 variants per gene (range = 1 to 3,593). Using the SKAT-O approach[8], we then tested whether any of these gene sets with a collection of rare, likely deleterious variants have an aggregate association with inflammatory bowel diseases. To interpret statistical significance, we applied a Bonferroni-corrected significance threshold of $P_{SKAT} < 2.2 \times 10^{-6}$ (0.05 corrected for 22,521 tests).

We implicate *ATP1A4* and *CALB2* in inflammatory bowel diseases for the first time. We detected an aggregate association of 66 rare, likely deleterious, heterozygous variants within or near *ATP1A4* with inflammatory bowel disease ($P_{SKAT} = 3.20 \times 10^{-8}$; **Fig. 2-5**) and Crohn's disease ($P_{SKAT} = 1.38 \times 10^{-6}$; **Fig. 2-6**) after Bonferroni correction. Of these, in particular, we identified a missense variant, 1:160155117,

within *ATP1A4* that was seen 9 times in cases as opposed to 48 times in controls. When tested individually using a logistic regression framework conditioned on the first five principal components, 1:160155117 demonstrated a suggestive evidence of association (P = 1.64 x 10^{-6}; OR = 0.17 for inflammatory bowel disease, and P = 2.34 x 10^{-5}; OR = 0.18 for Crohn's disease). On the other hand, we did not observe any individual evidence of association at the remaining variants within or near *ATP1A4* (P > 0.1), suggesting that the aggregate rare variant association signal at *ATP1A4* is likely driven by 1:160155117.

This rare variant signal was independent of the nearby common allele, rs4656958 (~700 kb away) reported in prior GWAS[2,3], indicating that these associations represent unique effects. To confirm whether these effects are specific to populations of African descent, we assessed the presence of these alleles, and evidence of their association, both individually and gene-wide, in whole-genome sequences of 8,000 inflammatory bowel disease cases and 15,000 matched controls with European ancestry (unpublished data from Carl Anderson's lab, Sanger Institute, UK). We observed no aggregate association signal at this locus in European population samples. Further, the result from individual association analysis of 1:160155117 was most compatible with no important effect in European individuals, implicating the possibility of population specificity in rare variant contributions to inflammatory bowel disease.

ATP1A4 encodes an ATPase involved in establishing and maintaining the electrochemical gradients of Na^+ and K^+ ions across the plasma membrane, which is crucial for cell ion homeostasis, cell membrane resting potential, and the transport of a variety of nutrients across the cell surface. *ATP1A4* forms the catalytic component of the ATPase that catalyzes the hydrolysis of ATP, which is coupled by the active exchange of intracellular Na^+ for extracellular K^+. Strikingly, electrolyte imbalances have previously been implicated in the pathology of inflammatory bowel diseases[16,17]. Intestinal inflammatory processes reduce the absorption of Na^+ while they increase K^+ secretion; inflammatory bowel disease-associated mucosal inflammation and the consequent impaired secretion and absorption of electrolytes often result in electrolyte and acid-base imbalance in inflammatory bowel disease patients[16,17]. Here, we present the first direct human

genetic evidence for the involvement of electrolyte imbalance in the pathogenesis of inflammatory bowel disease.

The second new association was seen at *CALB2*. A collection of 35 rare, likely deleterious, heterozygous variants within or near *CALB2* showed an aggregate association with ulcerative colitis (P_{SKAT} = 1.61 x 10^{-6}; **Fig. 2-7**). Half of these variants were observed more frequently in patients with ulcerative colitis compared to controls, while the other half were seen less frequently, representing a typical SKAT type of signal. When tested individually, we discovered an African-specific intronic variant, rs200083611, with a nominal evidence of association for increased risk of ulcerative colitis, showing a MAF of 0.009 in cases and 0.0003 in controls (P = 0.001; OR = 30.5). However, given the high-risk, but weak evidence of association at rs200083611, it appears that *CALB2* gene-wide signal was driven by additional rare variants that had yet to be identified. This *CALB2* signal was approximately 3 Mb away from, and independent of, the nearby common variant, rs1728785, with an established association for ulcerative colitis[2,3,18]. In order to examine the population specific role of *CALB2*, we tested for the aggregate and individual associations within or near this gene in unpublished data from the Anderson lab, confirming the African-specific role of *CALB2* in ulcerative colitis risk.

CALB2 encodes an intracellular calcium-binding protein, calbindin 2 (also known as calretinin) that plays an important role in neuronal physiology, and the maintenance of Ca^{2+} intracellular homeostasis. *CALB2* has a common expression pattern in central and peripheral nervous system, with high expression in brain, and intermediate expression in sigmoid and transverse colon. The absence of *CALB2* in nerve fibers in colon is a widely used marker for Hirschsprung's disease[19,20], whereas, elevated expression of *CALB2* has been reported as a hallmark of rapidly proliferating cancerous cell lines, including in colorectal cancer cell lines[21-25]. Hirschsprung's disease shares many of the clinical features with inflammatory bowel disease, where the latter is more commonly reported in patients who had surgical treatment for Hirschsprung's disease. On the other hand, longstanding inflammatory bowel disease is an established risk factor for various types of cancers, including colorectal cancer[26-29]. Given the intricate relationship of inflammatory

bowel disease with these companion diseases, our implication of Hirschsprung's disease- and colorectal cancer-associated *CALB2* in ulcerative colitis makes this signal noteworthy.

Further support for these two new associations emerges from a slightly different approach that we used to aggregate variants into small number of sets. When we collapsed rare, likely deleterious variants into sets based on a defined region of certain length ranging from 10 kb to 100 kb windows, we noted significant association at regions harboring *ATP1A4* or *CALB2* with the respective aforementioned phenotypes, regardless of the size of the defined regions. Additionally, specifically with 50 kb windows, we observed an aggregate association signal at a region harboring *SOX5*, a transcription factor involved in the regulation of embryonic development and in the determination of cell fate, with inflammatory bowel disease. A total of 74 rare, likely deleterious variants in this 50 kb region demonstrated a significant aggregate association ($P_{SKAT} = 2.24 \times 10^{-7}$; **Fig. 2-8**). Consistent with a SKAT type of signal, no individual variants in this region depicted associative evidence ($P > 0.1$). This signal is ~12 Mb away from, and independent of, the known common variant, rs11612508 implicated in ulcerative colitis, and about 20 Mb away from the *NOD2* locus with multiple large-effect risk alleles for Crohn's disease[2,30-32]. Notably, Sox5 by interacting with c-Maf, has been shown to induce T helper type 17 (Th17) cell differentiation[33]; multiple studies have highlighted a pathogenic role for Th17 cells in various autoimmune diseases, including inflammatory bowel disease.

With our whole-genomes data, we next set to assess whether the genetic landscape at the previously known inflammatory bowel disease risk loci is shared between populations of European and African descent, and whether trans-ethnic comparative analysis can be leveraged to further refine the established GWAS signals. Of the 236 lead variants from the thus far established loci that we found data for in the recent meta-analyses of cohorts of European descent[2,3], we had data for 227 of them in our whole-genome sequence dataset. Of these, 73% showed directional consistency for inflammatory bowel disease risk between the two populations (**Fig. 2-9**). Similarly, we noted a significant directional consistency between European and African American populations for Crohn's disease and ulcerative colitis (**Fig. 2-9**). On the other hand, we did observe genetic heterogeneity at some of the established risk loci driven either by differences in the

direction of effect, effect size and/or minor allele frequency (**Fig. 2-9 and Table 2-4**). For instance, of the two risk variants with large effect sizes for Crohn's disease in European populations, while the missense variant in *NOD2* showed differences pertaining to both effect size and allele frequency, heterogeneity at the intronic variant in *IL23R* was exclusively driven by differences in allele frequency between the two populations (**Table 2-4**). Nevertheless, our data confirm the previous notion that the genetic risk of inflammatory bowel diseases conferred by common variants is, to the most extent, shared across divergent populations.

This motivated us to specifically test whether genetic risk scores derived from all the known inflammatory bowel disease risk loci that were originally identified primarily in European populations, would distinguish our African American cases from controls. On average, genetic risk derived from the previously known GWAS signals (236 lead SNPs; see Methods) was significantly higher in cases compared to controls (**Fig. 2-10**). However, the discriminatory power of the genetic risk score derived from this model was barely any better than what is expected by random chance (AUC = 0.54; **Fig. 2-10**). On the other hand, polygenic risk score derived from a genome-wide feature set of common to low-frequency variants has recently been proven successful in risk stratifying and predicting individuals at high risk for various polygenic diseases, including inflammatory bowel disease, in European individuals[34]. In line with this, the risk model derived from a genome-wide set of 1.2 million common variants with effect sizes estimated in African American population samples (see Methods) outperformed the risk model derived from just the top 236 known signals established by previous GWAS scans in distinguishing our African American cases from controls (AUC of 0.64 vs 0.56; **Fig. 2-11**). Similarly, the African American-genome-wide genetic risk model demonstrated a 7-fold risk gradient between the bottom decile and the top decile, outperforming the 2-fold gradient achieved with the Eurocentric genetic risk model (**Fig. 2-12**).

DISCUSSION

To further define and resolve the genetic architecture of inflammatory bowel diseases, we have performed the first and the most comprehensive whole-genome sequencing analysis that include many alleles that

were not previously examined, in a population that remains vastly understudied. We implicate two new genes, *ATP1A4* and *CALB2* in Crohn's disease and ulcerative colitis, respectively. Our study highlights that multiple rare variants with small to moderate effects exist, and, at least, when clustered in a small number of sets (genes, windows *etc*.), are likely to account for some of the missing heritability; however, large-scale deep sequencing studies will be needed to precisely estimate the variance in disease liability explained by such variants. Besides providing further evidence for the emerging notion that the genetic risk of inflammatory bowel disease conferred by common alleles is shared across populations, our data highlight the possibility that rare variant contributions may exert population-specific effects. While this calls for the expansion of samples within each individual ethnic background, and further methodological development to facilitate a direct comparison of trans-ethnic rare variant discoveries, it remains to be seen whether such population-specific rare variant contributions may provide insights into widening ethnic disparities in health care.

Table 2-1: Whole-genome sequencing (discovery) cohort – sample break down and proportion of cases per site

Site	Total samples (% IBD cases)
Wash U	1274 (0%)
Johns Hopkins	1114 (74%)
Emory University	935 (94%)
Cedars Sinai	195 (59%)
Mount Sinai	92 (87%)

Table 2-2: GWAS (validation) cohorts – sample break down per genotype array and proportion of cases per array

SNP array	Total samples	% IBD	% CD	% UC
Omni	2187	29%	18%	11%
Axiom	3675	17%	12%	5%

Table 2-3: Common variants associated with Crohn's disease in African Americans

CHR	SNP	Position	A1	A2	OR	P	MAF in cases	MAF in controls	Location	Nearest Gene
5	rs11742570	40410482	T	C	0.74	3.24E-08	0.32	0.39	intergenic	*PTGER4*
5	rs6451493	40410833	G	T	0.74	4.55E-08	0.32	0.39	intergenic	*PTGER4*
5	rs6451494	40411189	T	C	0.74	4.09E-08	0.32	0.39	intergenic	*PTGER4*
5	rs12655810	40412093	T	C	0.74	4.09E-08	0.32	0.39	intergenic	*PTGER4*
5	rs1992661	40414887	G	A	0.74	3.65E-08	0.32	0.39	intergenic	*PTGER4*
5	rs1992660	40414965	C	T	0.74	4.55E-08	0.32	0.39	intergenic	*PTGER4*
5	rs2371720	40417739	C	T	0.74	3.86E-08	0.32	0.39	intergenic	*PTGER4*
5	rs12654092	40418033	G	C	0.74	3.86E-08	0.32	0.39	intergenic	*PTGER4*
5	rs6873829	40418162	T	G	0.74	4.30E-08	0.32	0.39	intergenic	*PTGER4*
5	rs6874571	40418639	T	C	0.74	3.21E-08	0.32	0.39	intergenic	*PTGER4*
5	rs114400949	40419041	G	C	0.74	2.46E-08	0.32	0.39	intergenic	*PTGER4*
5	rs149200362	40419327	T	C	0.74	4.04E-08	0.32	0.39	intergenic	*PTGER4*
5	rs10473192	40419847	C	G	0.74	3.06E-08	0.32	0.39	intergenic	*PTGER4*
5	rs7705462	40420564	C	G	0.73	1.71E-08	0.32	0.39	intergenic	*PTGER4*
5	rs56344733	40420780	GT	G	0.74	2.34E-08	0.32	0.39	intergenic	*PTGER4*
5	rs6897767	40422760	T	G	0.74	4.45E-08	0.32	0.39	intergenic	*PTGER4*
5	rs6876242	40422836	A	C	0.74	3.98E-08	0.32	0.39	intergenic	*PTGER4*
5	rs7716887	40423132	C	T	0.74	2.68E-08	0.32	0.39	intergenic	*PTGER4*
5	rs7730591	40423178	G	A	0.73	1.88E-08	0.32	0.39	intergenic	*PTGER4*
5	rs6896969	40424324	A	C	0.74	3.13E-08	0.32	0.39	intergenic	*PTGER4*
5	rs6879489	40425088	T	G	0.74	3.12E-08	0.32	0.39	intergenic	*PTGER4*
5	rs957100	40426318	G	T	0.73	9.37E-09	0.32	0.40	intergenic	*PTGER4*
1	1:10350106	10350106	C	T	0.26	1.47E-09	0.007	0.029	intronic	*KIF1B*

A1 = minor (effective) allele; A2 = major (reference) allele; OR = odds ratio; MAF = minor allele frequency

Table 2-4: Genetic heterogeneity pertaining to effect sizes and/or allele frequency at known inflammatory bowel disease risk loci between populations

	CHR	SNP	BP (hg38)	Gene	Minor Allele	Major Allele	OR	*P* value	MAF
European population	16	rs2066844	50712015	*NOD2*	A	G	2.13	9×10^{-214}	0.04
African Americans	16	rs2066844	50712015	*NOD2*	A	G	1.47	0.072	0.01
European population	1	rs41313262	67240217	*IL23R*	A	G	0.36	8×10^{-114}	0.01
African Americans	1	rs41313262	67240217	*IL23R*	A	G	0.38	0.09	0.003

Figure 2-1: Principal component analysis. Principal component plots of genetic data for the 3418 African American subjects included in the current study. Individuals were color coded based on either the site they came from (left) or case-control status (right). These plots were drawn based on 1.8 million, LD-pruned, high frequency variants.

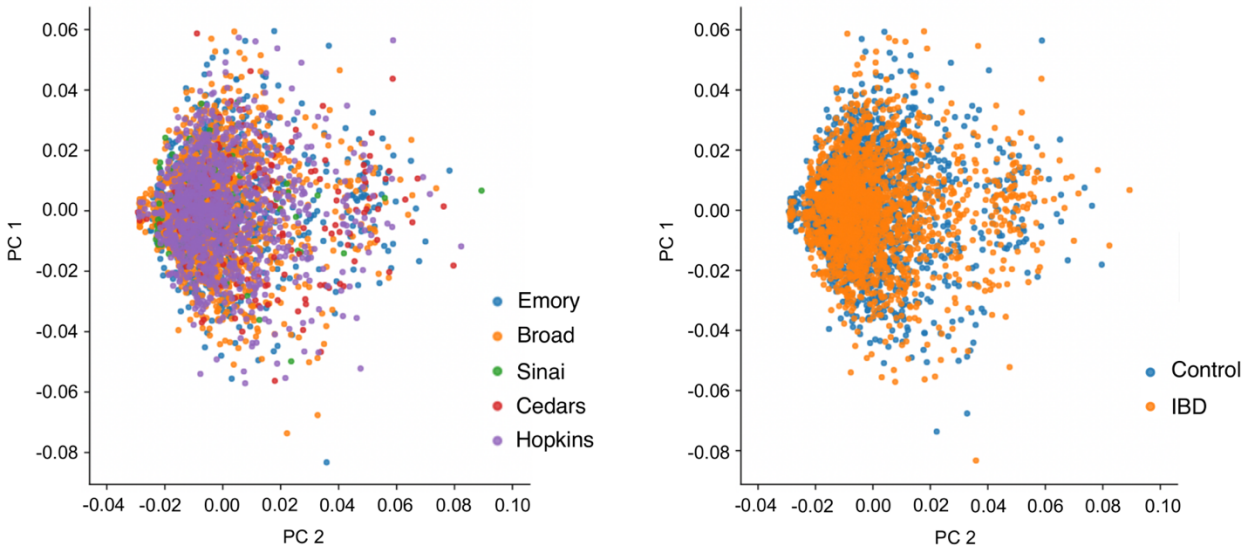


Figure 2-2: LocusZoom plot of variants in *PTGER4* locus with genome-side significant association for Crohn's disease in African Americans. All 22 variants in the *PTGER4* locus attaining $P < 5 \times 10^{-8}$ (*y* axis) in discovery whole-genome sequence data set are shown. The sentinel SNP that achieved suggestive evidence for Crohn's disease in our previous African American GWAS study is highlighted in purple. Red color indicates pair-wise LD with the SNP shown in purple. SNPs with missing LD information are in gray. Genomic location is shown on *x* axis.

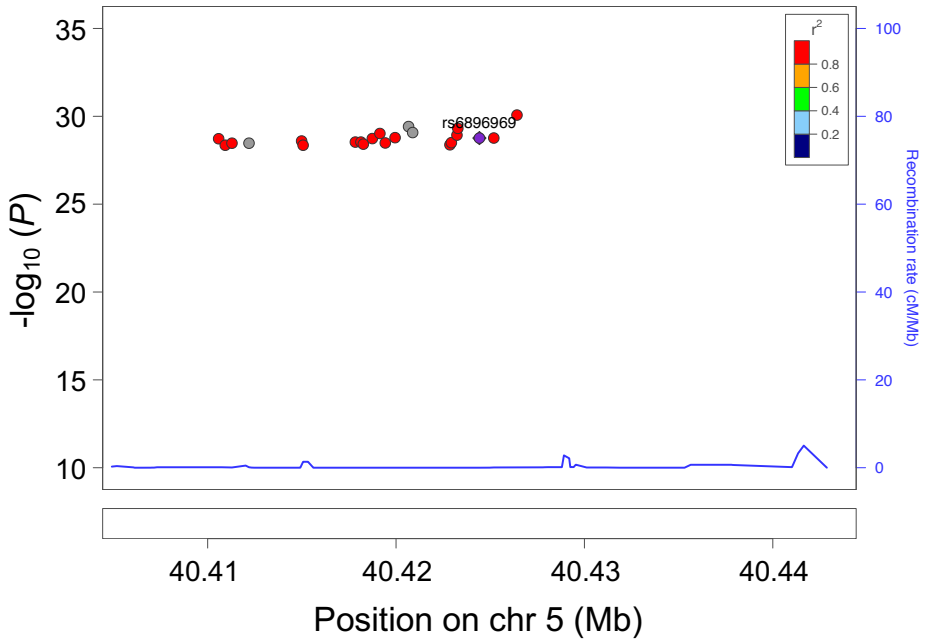


Figure 2-3: LocusZoom plot of credible variants in *PTGER4* locus fine-mapped recently in European population samples. 189 credible variants from Huang *et al*., representing four independent signals within the *PTGER4* locus are shown. Genomic location is shown on *x* axis. Association evidence with Crohn's disease in large meta-analysis of European population samples is shown on *y* axis. The sentinel SNP from signal 3 is shown in purple. The remainder of the credible SNPs are color coded based on their pair-wise LD with the SNP shown in purple. SNPs with missing LD information are in gray.

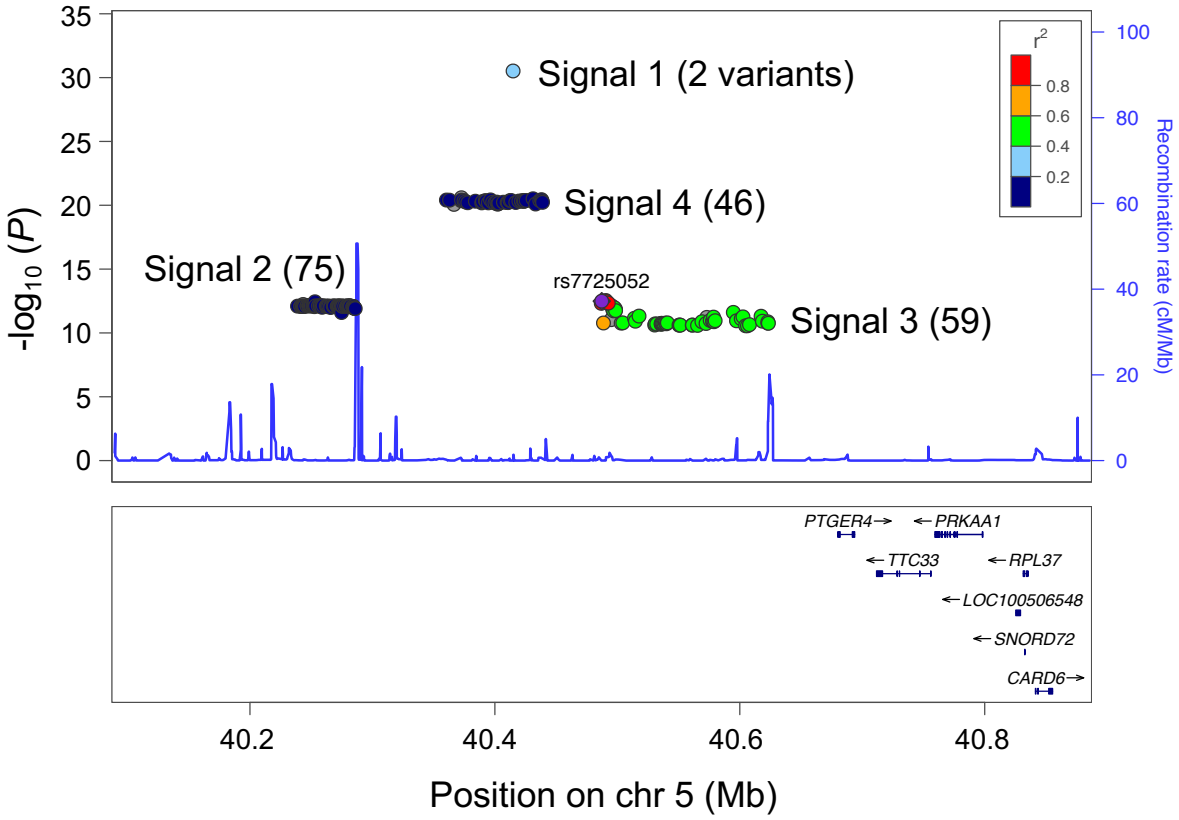


Figure 2-4: African American variants in *PTGER4* locus are in strong LD with signal 1 fine mapped in populations of European ancestry. Purple dot represents the tag variant from signal 1 from Huang *et al*. Variants identified in this dataset are color coded based on their pair-wide LD with the tag variant from Huang *et al*. Variants with missing LD information are in gray.

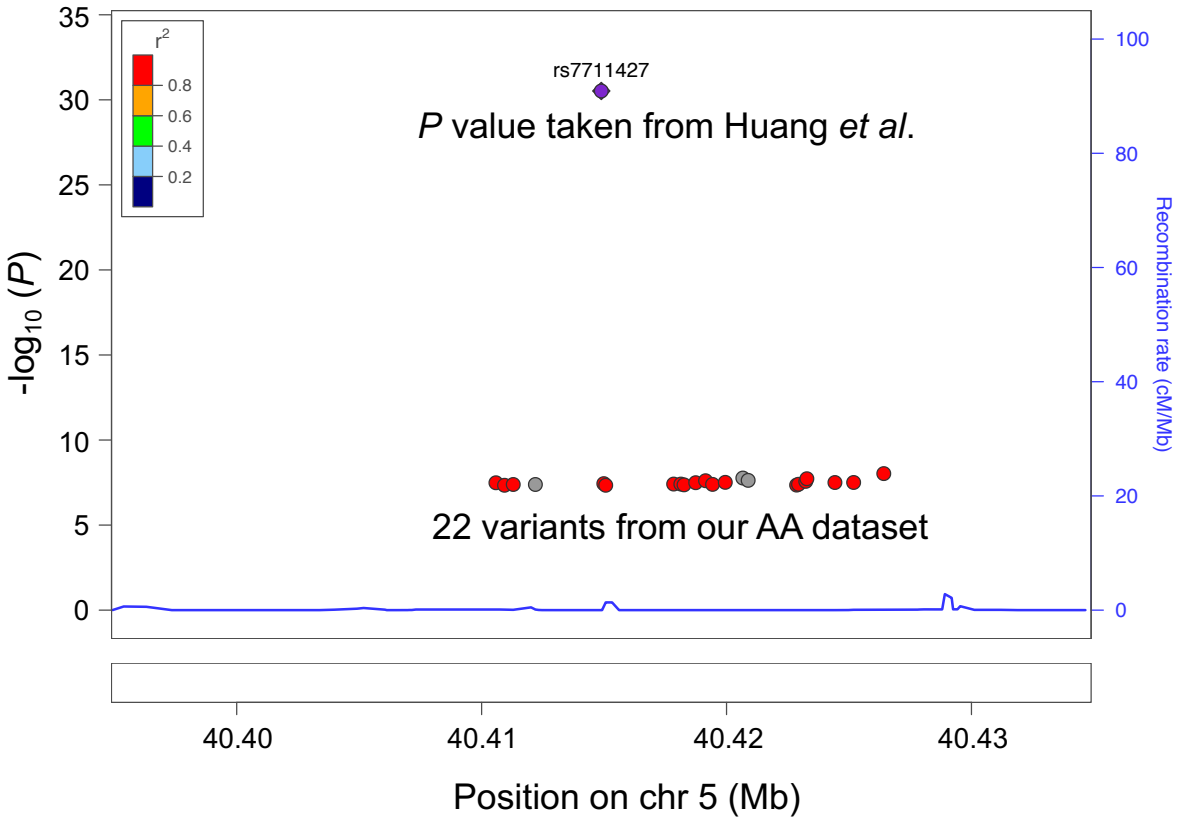


Figure 2-5: Rare, likely deleterious variants within or near *ATP1A4* have an aggregate association with inflammatory bowel disease. Each dot represents a gene with a collection of rare, likely deleterious variants. The observed *P* value of each gene is plotted as a function of the expected *P* value.

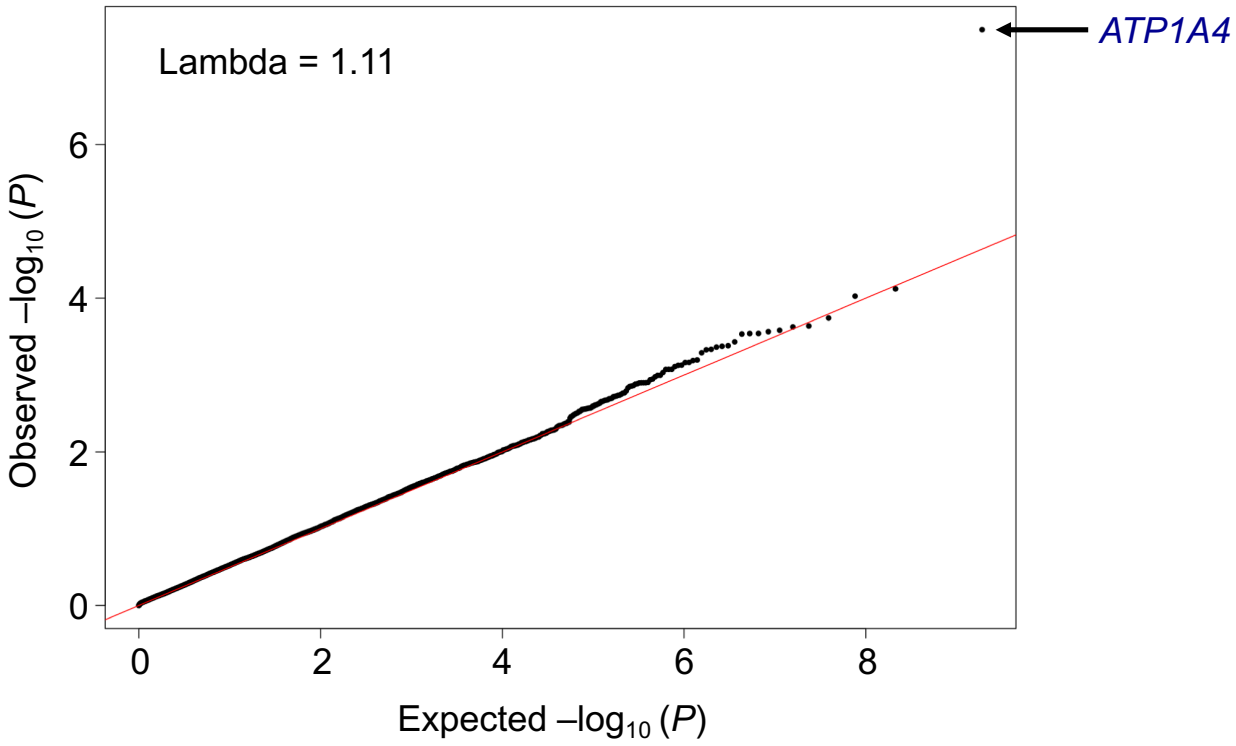


Figure 2-6: Rare, likely deleterious variants within or near *ATP1A4* have an aggregate association with Crohn's disease. Each dot represents a gene with a collection of rare, likely deleterious variants. The observed *P* value of each gene is plotted as a function of the expected *P* value.

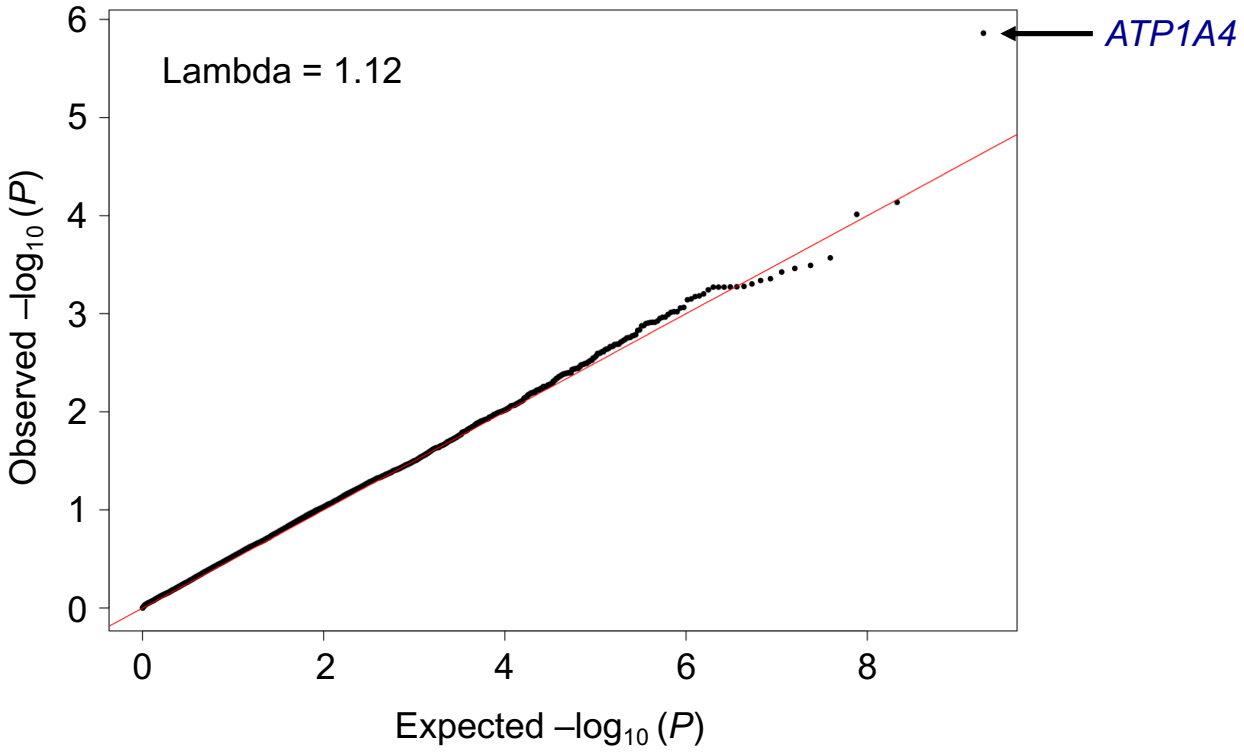


Figure 2-7: Rare, likely deleterious variants within or near *CALB2* have an aggregate association with ulcerative colitis. Each dot represents a gene with a collection of rare, likely deleterious variants. The observed *P* value of each gene is plotted as a function of the expected *P* value.

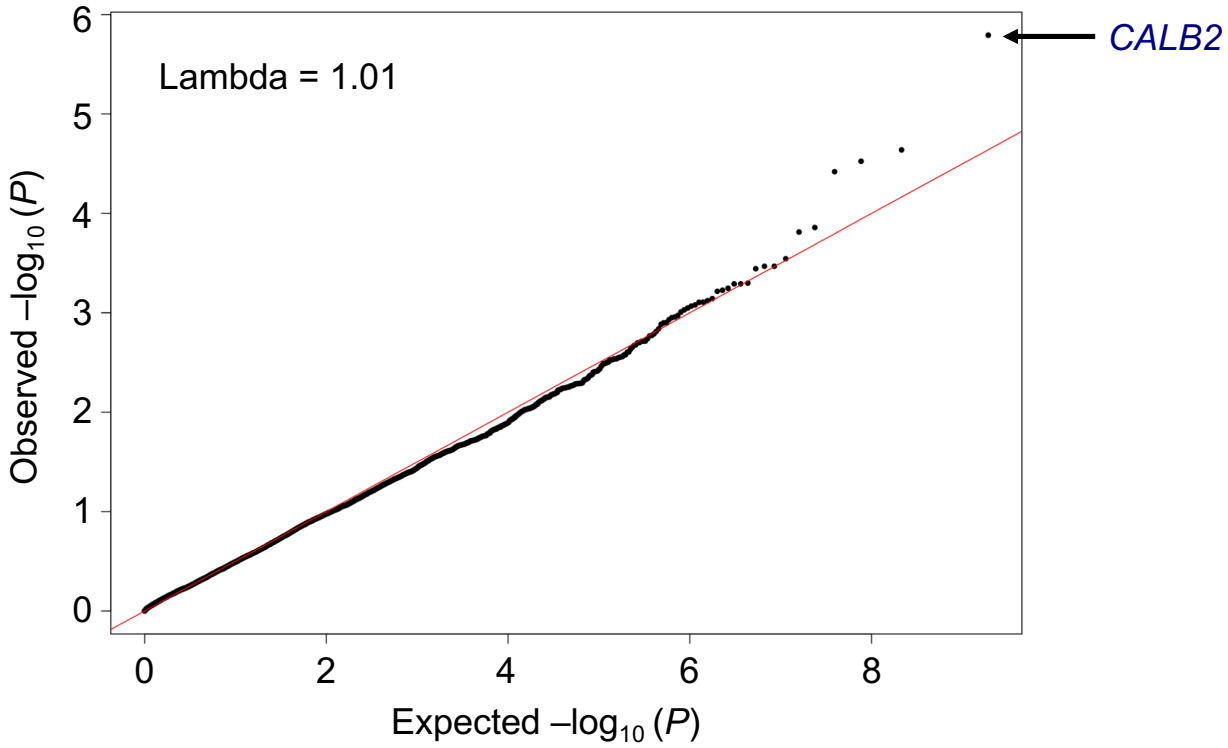


Figure 2-8: Rare, likely deleterious variants in a 50 kb window containing *SOX5* demonstrate an aggregate association with inflammatory bowel disease. Each dot represents a 50 kb region with a collection of rare, likely deleterious variants. The observed *P* value of each region is plotted as a function of the expected *P* value. Windows of 50 kb length on Chr 12 harboring *SOX5* and on Chr 1 harboring *ATP1A4* that reached genome-wide significance are indicated.

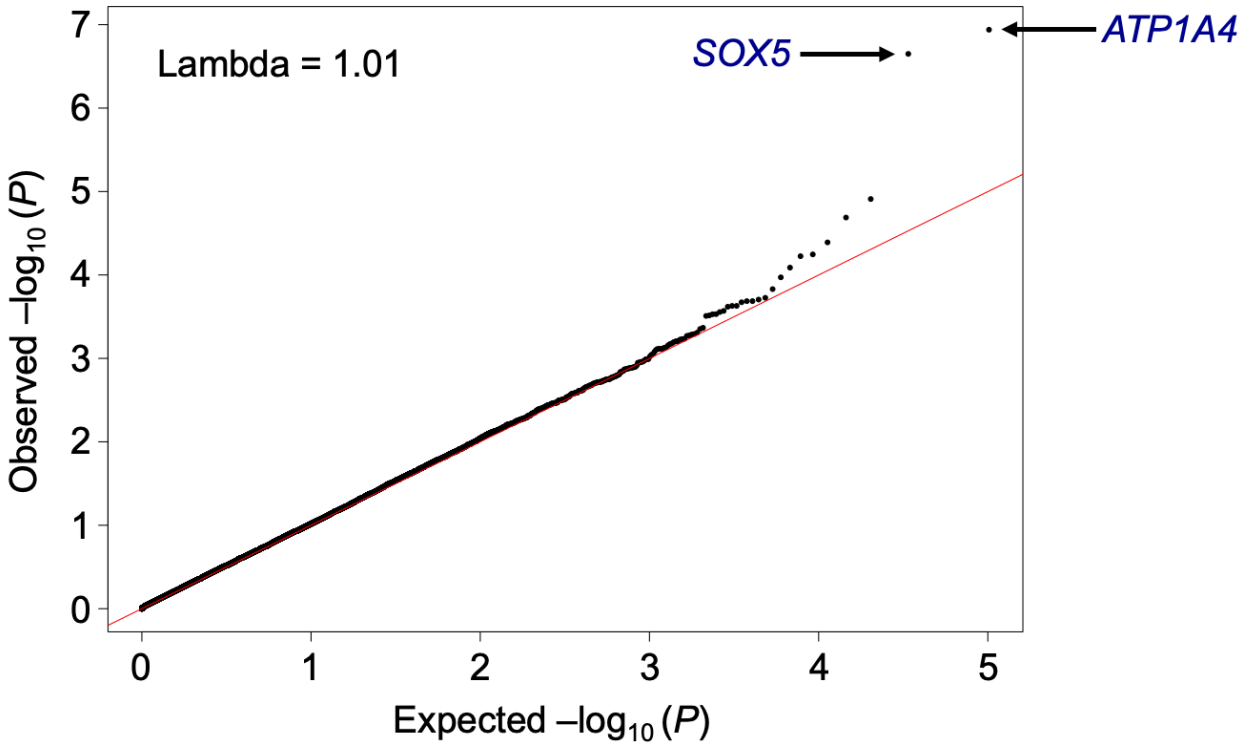
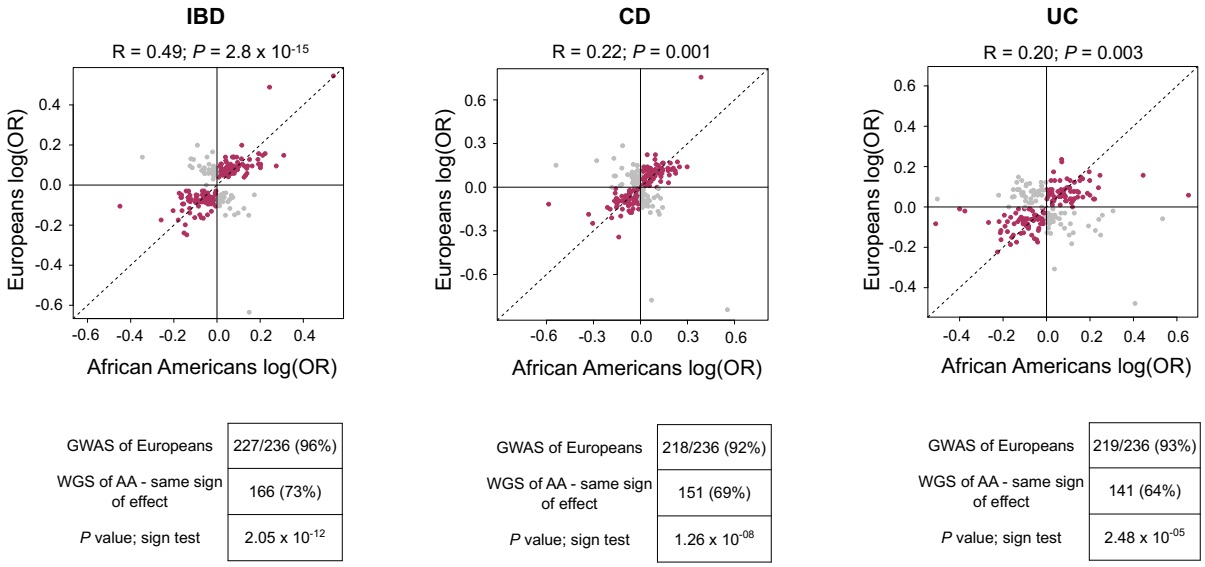


Figure 2-9: Directionally consistent effects at many of the known loci in Europeans vs African Americans. Comparison of estimated effect sizes for lead variants from each of the established risk loci on inflammatory bowel disease phenotypes between European vs African individuals are shown. The estimated effects in European population samples were obtained from large meta-analysis. Effect sizes for African Americans were obtained from the current whole-genome sequence dataset. Dots shown in maroon indicate variants with directionally consistent effects between the two populations. Directionally inconsistent ones are shown in gray.



	IBD	CD	UC
GWAS of Europeans	227/236 (96%)	218/236 (92%)	219/236 (93%)
WGS of AA - same sign of effect	166 (73%)	151 (69%)	141 (64%)
P value; sign test	2.05 x 10^{-12}	1.26 x 10^{-08}	2.48 x 10^{-05}

Figure 2-10: Risk for inflammatory bowel diseases in African Americans according to genetic effects at known disease risk loci estimated in European population samples. Genetic risk of African American cases vs controls in the current dataset. Risk scores were computed based on the 236 lead variants from previously established inflammatory bowel disease risk loci. Receiving operating characteristic curve of individuals genetic risk score was plotted to distinguish cases from controls. The area under the curve (AUC) is indicated. A perfect classifier would have an AUC of 1, and a random classifier would score 0.5.

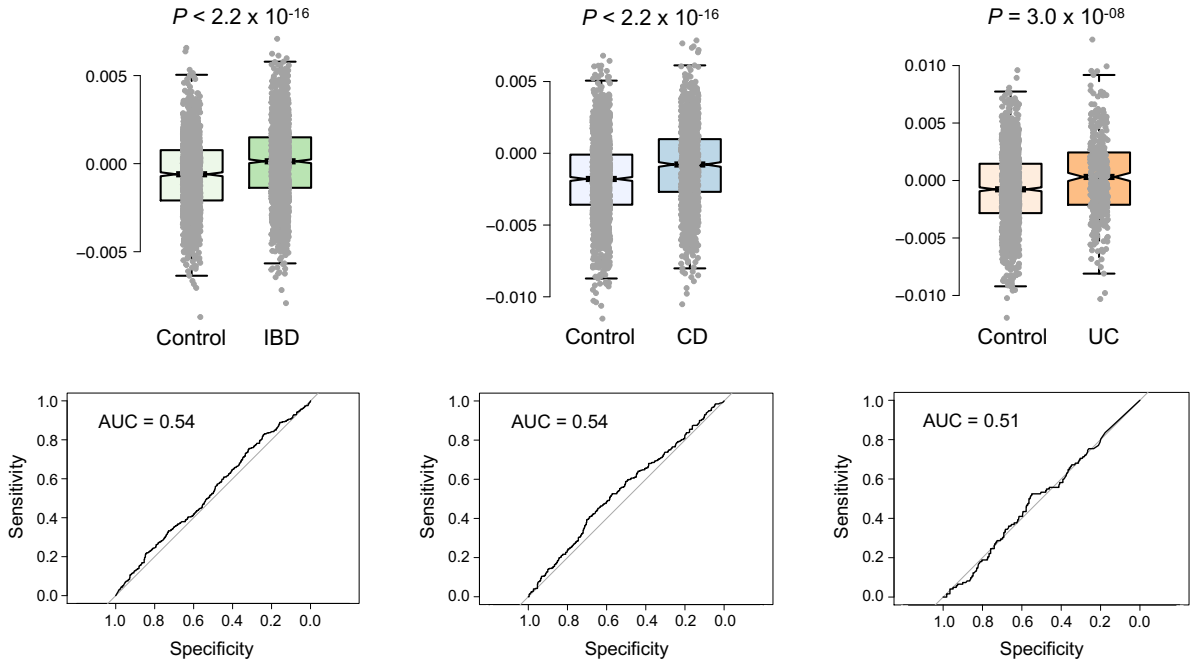


Figure 2-11: Genetic risk vs genome-wide genetic risk for inflammatory bowel disease in African Americans. Genetic risk or genome-wide genetic risk for inflammatory bowel disease in a subset of African American cases vs controls in the current dataset. Risk scores were computed based on a genome-wide feature set of 1.2 million variants estimated in African American samples (left) or from the 236 lead variants from previously established inflammatory bowel disease risk loci (right). Receiving operating characteristic curve of individuals genetic or genome-wide genetic risk score was plotted to distinguish cases from controls. The area under the curve (AUC) is indicated.

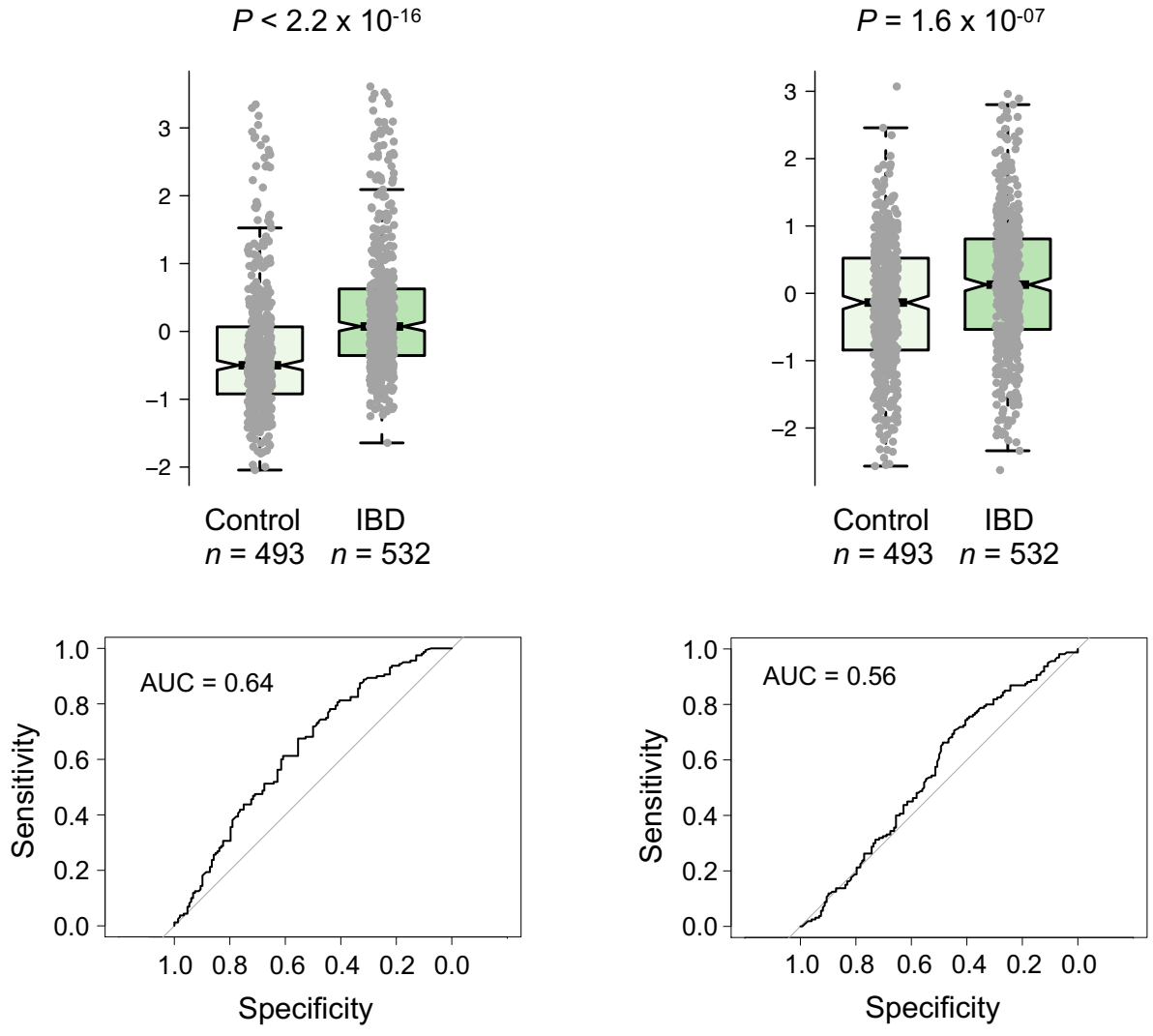


Figure 2-12: Risk stratification potential of genome-wide genetic risk score (left) vs genetic risk score (right). The prevalence of inflammatory bowel disease per group binned according to the decile of the genetic or genome-wide genetic risk score.

Total number of subjects = 1025; Number of subjects in each bin = ~102

REFERENCES

1. Jostins, L. *et al.* Host-microbe interactions have shaped the genetic architecture of inflammatory bowel disease. *Nature* **491**, 119-24 (2012).
2. Liu, J.Z. *et al.* Association analyses identify 38 susceptibility loci for inflammatory bowel disease and highlight shared genetic risk across populations. *Nat Genet* **47**, 979-986 (2015).
3. de Lange, K.M. *et al.* Genome-wide association study implicates immune activation of multiple integrin genes in inflammatory bowel disease. *Nat Genet* **49**, 256-261 (2017).
4. Luo, Y. *et al.* Exploring the genetic architecture of inflammatory bowel disease by whole-genome sequencing identifies association at ADCY7. *Nat Genet* **49**, 186-192 (2017).
5. Van der Auwera, G.A. *et al.* From FastQ data to high confidence variant calls: the Genome Analysis Toolkit best practices pipeline. *Curr Protoc Bioinformatics* **43**, 11 10 1-33 (2013).
6. Kotlar, A.V., Trevino, C.E., Zwick, M.E., Cutler, D.J. & Wingo, T.S. Bystro: rapid online variant annotation and natural-language filtering at whole-genome scale. *Genome Biol* **19**, 14 (2018).
7. Price, A.L. *et al.* Principal components analysis corrects for stratification in genome-wide association studies. *Nat Genet* **38**, 904-9 (2006).
8. Lee, S. *et al.* Optimal unified approach for rare-variant association testing with application to small-sample case-control whole-exome sequencing studies. *Am J Hum Genet* **91**, 224-37 (2012).
9. Brant, S.R. *et al.* Genome-Wide Association Study Identifies African-Specific Susceptibility Loci in African Americans With Inflammatory Bowel Disease. *Gastroenterology* **152**, 206-217 e2 (2017).
10. Loh, P.R. *et al.* Reference-based phasing using the Haplotype Reference Consortium panel. *Nat Genet* **48**, 1443-1448 (2016).
11. Das, S. *et al.* Next-generation genotype imputation service and methods. *Nat Genet* **48**, 1284-1287 (2016).
12. Marchini, J., Howie, B., Myers, S., McVean, G. & Donnelly, P. A new multipoint method for genome-wide association studies by imputation of genotypes. *Nat Genet* **39**, 906-13 (2007).
13. Willer, C.J., Li, Y. & Abecasis, G.R. METAL: fast and efficient meta-analysis of genomewide association scans. *Bioinformatics* **26**, 2190-1 (2010).
14. Libioulle, C. *et al.* Novel Crohn disease locus identified by genome-wide association maps to a gene desert on 5p13.1 and modulates expression of PTGER4. *PLoS Genet* **3**, e58 (2007).
15. Huang, H. *et al.* Fine-mapping inflammatory bowel disease loci to single-variant resolution. *Nature* **547**, 173-178 (2017).
16. Barkas, F., Liberopoulos, E., Kei, A. & Elisaf, M. Electrolyte and acid-base disorders in inflammatory bowel disease. *Ann Gastroenterol* **26**, 23-28 (2013).
17. Priyamvada, S. *et al.* Mechanisms Underlying Dysregulation of Electrolyte Absorption in Inflammatory Bowel Disease-Associated Diarrhea. *Inflamm Bowel Dis* **21**, 2926-35 (2015).
18. Consortium, U.I.G. *et al.* Genome-wide association study of ulcerative colitis identifies three new susceptibility loci, including the HNF4A region. *Nat Genet* **41**, 1330-4 (2009).
19. Rakhshani, N. *et al.* Hirschsprung Disease Diagnosis: Calretinin Marker Role in Determining the Presence or Absence of Ganglion Cells. *Iran J Pathol* **11**, 409-415 (2016).
20. Anbardar, M.H., Geramizadeh, B. & Foroutan, H.R. Evaluation of Calretinin as a New Marker in the Diagnosis of Hirschsprung Disease. *Iran J Pediatr* **25**, e367 (2015).
21. Blum, W., Pecze, L., Felley-Bosco, E. & Schwaller, B. Overexpression or absence of calretinin in mouse primary mesothelial cells inversely affects proliferation and cell migration. *Respir Res* **16**, 153 (2015).
22. Marilley, D. & Schwaller, B. Association between the calcium-binding protein calretinin and cytoskeletal components in the human colon adenocarcinoma cell line WiDr. *Exp Cell Res* **259**, 12-22 (2000).

23. Gotzos, V., Wintergerst, E.S., Musy, J.P., Spichtin, H.P. & Genton, C.Y. Selective distribution of calretinin in adenocarcinomas of the human colon and adjacent tissues. *Am J Surg Pathol* **23**, 701-11 (1999).
24. Gotzos, V., Schwaller, B., Gander, J.C., Bustos-Castillo, M. & Celio, M.R. Heterogeneity of expression of the calcium-binding protein calretinin in human colonic cancer cell lines. *Anticancer Res* **16**, 3491-8 (1996).
25. Doglioni, C. *et al.* Calretinin: a novel immunocytochemical marker for mesothelioma. *Am J Surg Pathol* **20**, 1037-46 (1996).
26. Scharl, S. *et al.* Malignancies in Inflammatory Bowel Disease: Frequency, Incidence and Risk Factors-Results from the Swiss IBD Cohort Study. *Am J Gastroenterol* (2018).
27. Choi, C.R., Bakir, I.A., Hart, A.L. & Graham, T.A. Clonal evolution of colorectal cancer in IBD. *Nat Rev Gastroenterol Hepatol* **14**, 218-229 (2017).
28. Peneau, A. *et al.* Mortality and cancer in pediatric-onset inflammatory bowel disease: a population-based study. *Am J Gastroenterol* **108**, 1647-53 (2013).
29. Aardoom, M.A., Linda Joosse, M.E., de Vries, A.C.H., Levine, A. & de Ridder, L. Malignancy and Mortality in Pediatric-onset Inflammatory Bowel Disease: A Systematic Review. *Inflamm Bowel Dis* **24**, 732-741 (2018).
30. Hugot, J.P. *et al.* Association of NOD2 leucine-rich repeat variants with susceptibility to Crohn's disease. *Nature* **411**, 599-603 (2001).
31. Rivas, M.A. *et al.* Deep resequencing of GWAS loci identifies independent rare variants associated with inflammatory bowel disease. *Nat Genet* **43**, 1066-73 (2011).
32. Ogura, Y. *et al.* A frameshift mutation in NOD2 associated with susceptibility to Crohn's disease. *Nature* **411**, 603-6 (2001).
33. Tanaka, S. *et al.* Sox5 and c-Maf cooperatively induce Th17 cell differentiation via RORgammat induction as downstream targets of Stat3. *J Exp Med* **211**, 1857-74 (2014).
34. Khera, A.V. *et al.* Genome-wide polygenic scores for common diseases identify individuals with risk equivalent to monogenic mutations. *Nat Genet* **50**, 1219-1224 (2018).

Chapter 3

Blood-Derived DNA Methylation Signatures of Crohn's Disease and Severity of Intestinal Inflammation

This chapter has been adapted and was originally published in *Gastroenterology;* Volume 156, Issue 8, June 2019, Pages 2254-2265.e3

Hari K Somineni, Suresh Venkateswaran, Varun Kilaru, Urko M Marigorta, Angela Mo, David T Okou, Richard Kellermayer, Kajari Mondal, Dawayland Cobb, Thomas D Walters, Anne Griffiths, Joshua D Noe, Wallace V Crandall, Joel R Rosh, David R Mack, Melvin B Heyman, Susan S Baker, Michael C Stephens, Robert N Baldassano, James F Markowitz, Marla C Dubinsky, Judy Cho, Jeffrey S Hyams, Lee A Denson, Greg Gibson, David J Cutler, Karen N Conneely, Alicia K Smith and Subra Kugathasan

ABSTRACT

Background & Aims: Crohn's disease is a relapsing and remitting inflammatory disorder with a variable clinical course. Although most patients present with an inflammatory phenotype (B1), approximately 20% of patients rapidly progress to complicated disease, which includes stricturing (B2), within 5 years. We analyzed DNA methylation patterns in blood samples of pediatric patients with Crohn's disease at diagnosis and later time points to identify changes that associate with and might contribute to disease development and progression.

Methods: We obtained blood samples from 164 pediatric patients (1–17 years old) with Crohn's disease (B1 or B2) who participated in a North American study and were followed for 5 years. Participants without intestinal inflammation or symptoms served as controls (n = 74). DNA methylation patterns were analyzed in samples collected at time of diagnosis and 1–3 years later at approximately 850,000 sites. We used genetic association and the concept of Mendelian randomization to identify changes in DNA methylation patterns that might contribute to the development of or result from Crohn's disease.

Results: We identified 1189 5'-cytosine–phosphate–guanosine-3' (CpG) sites that were differentially methylated between patients with Crohn's disease (at diagnosis) and controls. Methylation changes at these sites correlated with plasma levels of C-reactive protein. A comparison of methylation profiles of DNA collected at diagnosis of Crohn's disease vs during the follow-up period showed that, during treatment, alterations identified in methylation profiles at the time of diagnosis of Crohn's disease more closely resembled patterns observed in controls, irrespective of disease progression to B2. We identified methylation changes at 3 CpG sites that might contribute to the development of Crohn's disease. Most CpG methylation changes associated with Crohn's disease disappeared with treatment of inflammation and might be a result of Crohn's disease.

Conclusions: Methylation patterns observed in blood samples from patients with Crohn's disease accompany acute inflammation; with treatment, these change to resemble methylation patterns observed in

patients without intestinal inflammation. These findings indicate that Crohn's disease–associated patterns of DNA methylation observed in blood samples are a result of the inflammatory features of the disease and are less likely to contribute to disease development or progression.

INTRODUCTION

Inflammatory Bowel Diseases encompassing Crohn's disease and ulcerative colitis arise in the context of complex interactions between genetic and environmental factors. While these diseases can manifest at any age, pediatric-onset Crohn's disease has a higher incidence than ulcerative colitis[1,2], and patients diagnosed with Crohn's disease in childhood are more likely to suffer from an aggressive and severe disease course[2].

DNA methylation, occurring predominantly in the cytosine-guanine (CpG) dinucleotide context, is a key epigenetic mechanism that can regulate gene expression and thereby influence the development and progression of complex diseases. Cross-sectional studies of DNA methylation have begun to reveal epigenetic associations with inflammatory bowel disease in both pediatric and adult populations; across a range of cell and tissue types[3-11]. For instance, site-specific DNA methylation differences in peripheral blood[3] and blood-derived mononuclear cells[5] of adult patients with inflammatory bowel disease have been reported. Similarly, studies of mixed or purified cells from blood and intestinal mucosa of pediatric populations revealed distinct methylation profiles in relevance to inflammatory bowel disease[8,9]. Howell *et al*., recently reported a gut segment-specific methylation signature in pediatric patients with inflammatory bowel disease in the purified intestinal epithelial cells, and its persistence during the course of the disease[12]. However, due to the relapsing-remitting behavior of Crohn's disease, and the dynamic nature of DNA methylation and its resulting vulnerability to confounding and reverse causation, delineating the causal role of methylation in Crohn's disease requires longitudinal studies along with the application of integrative analytical approaches. Understanding how the methylome changes during the course of the disease, as a result of varying clinical characteristics, and how disease complications evolve may aid in the identification of potentially causal epigenetic targets, which could subsequently be leveraged for therapeutic benefits.

Here, we performed an epigenome-wide association analysis of DNA methylation in peripheral blood at ~850,000 sites and Crohn's disease 1) at diagnosis and 2) at later stages (1 to 3 years after diagnosis) during which time ~33% of the patients progressed from an initial stage of B1 inflammatory behavior to B2 stricturing behavior. Study participants (summarized in **Table 3-1**) were sampled from the RISK cohort[13], a pediatric prospective inception Crohn's disease cohort. Since the current Crohn's disease therapeutics systematically targets the peripheral immune system, and considerable genetic and cell biological evidence including previous epigenetic studies implicates the immune system in the etiology of Crohn's disease[3,14], we investigated methylation changes in peripheral blood with respect to their potential causal versus consequential roles in disease.

METHODS

Study Population. We utilized a subset of pediatric subjects recruited under the Risk Stratification and Identification of Immunogenetic and Microbial Markers of Rapid Disease Progression in Children with Crohn's Disease (RISK) study[13]. The RISK inception cohort study is thus far, the largest pediatric Crohn's disease cohort recruited at 28 sites in the USA and Canada to identify genetic, clinical, microbial and immunologic factors that predispose Crohn's disease patients (B1) to a complicated disease course (B2 or B3). Briefly, the RISK study recruited children with ages 1-17 who presented to gastroenterology clinics with suspected inflammatory bowel disease and followed them for a period of 5 years at regular intervals to determine the incidence of inflammatory bowel disorders or complications of an established disorder. The RISK study design, recruitment details, inclusion-exclusion criteria, disease behaviors, and data collection have been described in detail elsewhere[13].

Study design. The initial recruitment and follow-up have been previously described[13]. A subset of age-, sex-, and ethnicity-matched non-inflammatory bowel disease control subjects (controls) and Crohn's disease patients with B1 inflammatory behavior and B2 stricturing behavior were drawn from the RISK cohort[13], based on the availability of patient samples at two time points – at diagnosis and at a follow-up visit 1 to 3 years after diagnosis (**Table 3-1**). Subjects who were negative for gut inflammation and depicted

no bowel pathology on endoscopy, and remained inflammatory bowel disease symptom-free during the course of the follow-up period served as controls. Peripheral blood DNA samples from 164 newly diagnosed, treatment-naïve pediatric patients with Crohn's disease (cases) and 74 controls were considered for baseline analysis to identify Crohn's disease associated CpGs. Of these, 150 cases presented purely with an inflammatory phenotype (non-complicated Crohn's disease; B1) while the remaining 14 presented with stricturing phenotype (B2) at diagnosis. However, sensitivity analysis comparing 150 B1 cases or 14 B2 cases to 74 controls versus 164 cases (150 B1, 14 B2) to 74 controls showed that our findings are robust to disease behavior states (B1 or B2), allowing grouping of all cases (both B1 and B2) at diagnosis into a single large cohort.

The longitudinal analysis relied on follow-up samples taken from established cases (n = 164) as part of a longitudinal follow-up in the RISK study, which was 1 to 3 years from diagnosis. Exact details are provided in **Table 3-1**. Of the 150 B1 cases at diagnosis, 55 of them progressed to B2 (progressors) during the course of the follow-up period, while the rest (n = 95) remained as B1 at the time of the follow-up sampling (non-progressors). We note that in order to increase statistical power to define (if any) methylomic changes involved in disease progression, we purposefully inflated the number of progressors by selecting more pediatric cases who experienced B2 complication during the course of their prospective follow-up in the original RISK study[13]. With the 14 at-diagnosis-B2 patients who also remained as B2 at the time of follow-up sampling, we had a total of 95 B1 and 69 B2 at the follow-up. More details about phenotype classification are available in Kugathasan *et al*.[13].

Quantification of genome-wide DNA methylation and data processing. Peripheral blood genomic DNA was extracted using the AllPrep DNA/RNA Mini Kit (Qiagen, Valencia, CA). 500ng of extracted DNA from each sample was subjected to bisulfite treatment using EZ DNA Methylation-GoldTM Kit (Zymo Research, Irvine, CA). Genome-wide DNA methylation was quantified in bisulfite-converted genomic DNA at single-base resolution using the MethylationEPIC BeadChip (Illumina, San Diego, CA). The initial quality control for the data set was performed with the R package CpGassoc[15]. CpG sites called with low

signal or low confidence (detection $P > 0.05$) or with data missing for greater than 10% of samples were removed, and samples with data missing or called with low confidence for greater than 10% of CpG sites were removed. In addition, probes mapping to multiple locations were removed[16]. After the above steps, a total of 807,511 probes and 402 samples (74 controls and 2 samples from each of the 164 cases) remained. Beta values (β) were calculated for each CpG site as the ratio of methylated (M) to methylated and unmethylated (U) signal: β=M/(M+U). Signal intensities were then normalized using the module beta-mixture quantile dilation (BMIQ)[17] to account for the probe design bias in the EPIC array data. These normalized signal intensities were used to perform principal component analysis to further identify sample outliers (**Fig. 3-1**). Differential cell counts of the constituent cell types, $CD4^+$ T cells, $CD8^+$ T cells, NK cells, B cells, monocytes, and granulocytes were estimated for each individual from the methylation data using the Houseman algorithm[18].

Methylation association with Crohn's disease at diagnosis. Crohn's disease-associated methylation changes were profiled using the R package, CATE[19], which implements a state-of-the-art batch correction method to remove inflation and test-statistic bias in association tests. We tested for association between Crohn's disease and methylation at the ~807,000 sites which yielded results with an inflation in the association test of 1.16. Briefly, DNA methylation was regressed on disease status (0 for control, 1 for case) with age, gender, estimated cell proportions and the first 3 genotype-based principal components as covariates in the model. A case-control epigenome-wide association was performed to identify CpGs associated with Crohn's disease at diagnosis in a set of 238 study participants comprised of 164 cases and 74 controls.

Methylation association with Crohn's disease at diagnosis versus follow-up. To analyze longitudinal changes in DNA methylation of Crohn's disease patients, BMIQ-normalized signal intensities were further adjusted using ComBAT[20] to account for chip and position effects. This was done as an alternative to adjustment within CATE, since CATE could not be used to perform the longitudinal analysis. This within-cases longitudinal epigenome-wide association was performed to identify methylation changes during the

course of the disease in 328 samples comprised of 164 at diagnosis samples and 164 follow-up samples from the same subjects. To account for the two time points from each patient, representing methylation levels at diagnosis and follow-up, we used a linear mixed effects model to model the repeated measures where adjusted β values were regressed on the time point (0 for at diagnosis sample, 1 for follow-up sample) with age, gender, estimated cell proportions, 3 genotype-based principal components, and disease behavior states (0 for B1 sample, 1 for B2 sample) as fixed effect covariates and the subject ID as a random effect. Similarly, within-progressors and within-non-progressors epigenome-wide association analyses were performed in 55 progressors and 95 non-progressors, respectively, by comparing their methylation profiles at diagnosis with the follow-up data.

Methylation association with plasma CRP. A subset of subjects (45 controls, 132 cases at diagnosis and 95 cases at follow-up) who underwent methylation profiling had data on plasma CRP levels. To investigate the relationship between DNA methylation in blood and CRP, we performed an epigenome-wide association of CRP in 272 samples using a linear mixed effects model. Methylation β values were regressed on the $\log_2$ transformed plasma CRP (mg/L) levels with age, gender, estimated cell proportions, 3 genotype-based principal components, and disease status (0 for control, 1 for B1, 2 for B2) as fixed effect covariates and the subject ID as a random effect.

Methylation association with PCDAI. PCDAI scores were available for almost all the patients at the time of diagnosis (n = 159) and follow-up (n = 149). Details on how PCDAI was computed can be found elsewhere[13]. We performed epigenome-wide association of PCDAI in 308 samples by regressing methylation beta values on PCDAI with age, gender, estimated cell proportions, 3 genotype-based principal components, and disease behavior states (0 for B1, 1 for B2) as fixed effect covariates and the subject ID as a random effect.

KEGG pathway enrichment analysis. We used missMethyl[21], an R/Bioconductor package, to identify pathways that are more likely to occur in the Crohn's disease associated CpGs than would be expected by

chance, by referencing to KEGG database. Genes with more probes (more CpGs probed) on the MethylationEPIC array are more likely to have differentially methylated CpGs which could introduce potential bias when performing pathway enrichment analysis. The *gometh* function implemented in missMethyl takes into account the varying number of differentially methylated CpGs by computing prior probability for each gene based on the gene length and the number of CpGs probed per gene on the array.

Genotyping and data processing. Peripheral blood DNA samples from the 238 subjects with methylation data were genotyped using the Infinium Multi-Ethnic Global-8 Kit (Illumina, San Diego, CA) and genotypes were called using the GenomeStudio software. All these subjects had call rates >95% and inferred gender consistent with the clinical records. We tested for relatedness among the subjects by calculating pairwise identity by descent based on 59,889 LD-independent SNPs ($r^2 < 0.1$), which confirmed no relatedness among the subjects. The Multi-Ethnic array contained 1,762,905 variants before quality control. Removal of (i) SNPs with low call rate (< 95%), (ii) SNPs not in Hardy-Weinberg equilibrium ($P < 1.0\times10^{-3}$), and (iii) SNPs with minor allele frequency (MAF) < 5%, resulted in the retention of 1,751,369 SNPs, 1,736,281 SNPs and 651,370 SNPs, respectively. We further removed non-autosomal SNPs and SNPs mapping to multiple locations. This resulted in a data set consisting of 636,006 high quality SNPs. All quality control procedures were performed in PLINK[22].

Genotype-based principal components. Principal components were computed based on a pruned version of the data set consisting of 59,889 LD-independent SNPs ($r^2 < 0.1$) and MAF > 0.05. Unless stated otherwise, the first 3 genotype-based principal components were used to control for population stratification in all analyses (**Fig. 3-1**).

Genetic risk scores. We used the *score* function available in PLINK to compute weighted genetic risk scores. These scores were calculated based on the observed genotypes at 93 of the genotyped Crohn's disease risk SNPs and their corresponding effect sizes reported for Caucasian population in Liu *et al*.[23].

Methylation quantitative trait loci (mQTL) analysis. After exclusion of CpGs with a SNP(s) in their probe sequence, methylation proportions at each of the 625,464 CpG sites from baseline peripheral blood methylation data in 238 subjects was tested for associations with local genetic variants (±500 kb; *cis*-mQTLs) using a linear mixed model implemented in GEMMA[24]. This model allows for the adjustment of the population structure and relatedness among individuals as a random effect by providing a genetic relationship matrix (GRM) using LD-pruned SNP data set which could then be used as a covariate in the mQTL analysis. In addition to GRM, we included covariates for age, gender, disease status, estimated cell proportions and 3 genotype-based principal components, and modeled methylation (CpG) as the outcome and SNP as an explanatory variable. For each CpG site, all SNPs residing within ±500 kb were individually tested for association for a total of 144,916,995 tests, genome-wide. To adjust for multiple tests, statistically significant SNP-CpG pairs were inferred at FDR < 5%.

Genetic association and the concept of Mendelian randomization. To clarify the role of methylation changes that are associated with Crohn's disease, we used genetic association and the concept of Mendelian randomization as described in Wahl *et al*[25]. The fundamental idea of Mendelian randomization is shown in **Fig. 3-17**. Briefly, Mendelian randomization makes the following assumptions: (i) an instrumental variable (individual SNP or a combination of SNPs, such as a genetic risk score) has an association with the intermediate phenotype, (ii) the instrumental variable has no association with the outcome except through the intermediate phenotype, and (iii) the instrumental variable is not influenced by any of the measured or unmeasured confounding factors. If the intermediate phenotype is causally associated with the outcome, in an adequately powered study, the instrumental variable (associated with the intermediate phenotype) should also be associated with the outcome. Hence, assignment of directionality to the intermediate phenotype-outcome relationship via Mendelian randomization relies on the observed association between the instrumental variable and the outcome, which would typically require tens of thousands of subjects to achieve adequate power. For studies with limited sample size, as described in Wahl *et al*.[25], if there exists a potential causal relationship between the intermediate phenotype and the outcome, we would expect the

estimated effect (β coefficient) of the instrumental variable on the outcome (β_3 in **Fig. 3-18**) to be consistent (directional consistency) if not equivalent to its predicted effect mediated through the intermediate phenotype (β_1 x β_2).

DNA methylation cause of Crohn's disease. To identify Crohn's disease associated CpGs that are potentially causal, we used the most significantly associated mQTL (sentinel mQTL, defined as the *cis*-mQTL with the smallest *P* value) as the instrumental variable, methylation as the intermediate phenotype and Crohn's disease as the outcome (Model 1; **Fig. 3-17**). The effect size between SNPs and corresponding CpGs (sentinel mQTL-CpG pair from our *cis* mQTL analysis; β_1) was estimated via simple linear regression models with methylation as response and SNP as explanatory variable. The effect size between CpGs and disease status (β_2) was estimated via simple linear regression models with disease status (0 for control, 1 for Crohn's disease) as response and CpG as explanatory variable. The effect size between sentinel mQTL SNP and disease status (β_3) was obtained from large, meta-analysis of Crohn's disease GWAS[23]. For these, the odds ratios estimated via logistic regression models with disease status (0 for control, 1 for Crohn's disease) as response and SNP as explanatory variable in GWAS meta-analysis[23] were log transformed to make the effects linear. The reason behind fitting simple linear regression models despite the response variable being binary is that the relationship between effect sizes denoted in equation in **Fig. 3-18** holds true when linear regression models are fit, but no analogous relationship exists for logistic regression models. Because this relationship between effect sizes was important for our assessment of consequence versus causality depicted in **Fig. 3-18**, we chose to use linear rather than logistic regression. We note that, while the normality assumption of linear regression is clearly violated by the use of a binary dependent variable, leading to incorrect estimates of the standard errors, the estimated effect sizes will be unbiased estimates of the expected change in outcome due to a 1-unit change in the predictor.

From our mQTL analysis we identified mQTL associations (FDR < 0.05) for 194 of the 1189 Crohn's disease associated CpGs. Of these, 174 CpGs for which the associated mQTL SNP (or proxy SNP (n = 6): LD $r^2 \geq 0.8$; **Supplementary Table 3-14**) had been analyzed in a previously published GWAS[23] were

subsequently evaluated for their causal role. None of the sentinel mQTLs associated with the selected CpGs showed deviance from the assumptions made in order to be a valid instrument. For example, no sentinel mQTL showed significant association with the outcome (Crohn's disease) after conditioning on methylation levels at the corresponding CpGs, allowing us to investigate the potential causal relationships between DNA methylation in blood at all the sentinel mQTL-CpGs and Crohn's disease. The predicted effect sizes and standard errors were estimated as $\beta_{pred} = \beta_1 \times \beta_2$; and $SE_{pred} = (SE^2_1 \times SE^2_2 + SE^2_1 \times \beta^2_2 + SE^2_2 \times \beta^2_1)^{1/2}$, respectively. FDR < 0.05 was considered statistically significant for individual CpGs.

DNA methylation consequence of Crohn's disease. To identify Crohn's disease associated CpGs where changes in methylation are a consequence of the disease, we used weighted Crohn's disease genetic risk score as the instrumental variable, Crohn's disease as the intermediate phenotype, and methylation as the outcome (Model 2; **Fig. 3-17**). The effect size between *z* scored weighted genetic risk score and Crohn's disease (β_1) was estimated via simple linear regression models with Crohn's disease as response and risk score as explanatory variable. The effect size between methylation and Crohn's disease (β_2) was estimated via simple linear regression with methylation as response and Crohn's disease as explanatory variable. The effect size between methylation and weighted genetic risk score (β_3) was estimated via simple linear model with methylation as response and genetic risk score as explanatory variable. All of the 194 sentinel mQTL-CpGs were tested for methylation consequence of Crohn's disease. The predicted effect sizes and standard errors were computed as described above. FDR < 0.05 was considered statistically significant for individual CpGs.

Diagnostic utility of peripheral blood methylation signatures. To ascertain if peripheral blood methylation could distinguish patients with Crohn's disease from controls, we divided the baseline methylation dataset consisting of 238 subjects (164 cases and 74 controls) at random into equally weighted (cases and controls) training and testing datasets with 70% of the samples going into the training dataset. The training dataset was fit with a logistic regression model using the R package, glmnet[26], and the fitted

model was used to predict the case status for the test dataset. Diagnostic accuracy was assessed via area under the receiver operator characteristic curve.

RESULTS

Differentially methylated CpGs associated with Crohn's disease at diagnosis. Epigenome-wide association analysis of 164 newly diagnosed, treatment naïve pediatric patients with Crohn's disease (150 B1, 14 B2; cases) and 74 controls identified 1189 CpG sites associated with Crohn's disease in blood at diagnosis (FDR < 0.05; **Fig. 3-2** and **Supplementary Table 3-1**). Of these, 976 CpG sites (82%) had increased methylation in cases relative to controls and 213 (18%) had decreased methylation. Because disease-associated inflammation can influence expression within a cell population and create differences in total cell composition (**Fig. 3-3**), our analysis included covariates to adjust for estimated proportions[18] of the 6 dominant cell types ($CD4^+$ T cells, $CD8^+$ T cells, B cells, NK cells, monocytes and granulocytes) in blood (see Methods; **Fig. 3-4**). Sensitivity analyses demonstrated that our findings are robust to disease behavior states (B1 or B2; **Fig. 3-5** and **Supplementary Table 3-1**) and manifestation in the bowel (L1, L2 or L3; **Fig. 3-6**), suggesting that baseline methylomic contributions to Crohn's disease do not vary strongly by disease behavior or location. The strongest association signals were found on chromosomes 16, 17 and 19 (**Fig. 3-2**), with CpGs in a long non-coding RNA, *LOC100996291* (*LINCO1993*), showing the peak association with Crohn's disease at diagnosis (**Supplementary Table 3-1**). Apart from identifying novel CpG sites, we replicated several findings that were previously associated with Crohn's disease (including *TMEM49* (*VMP1*), *SBNO2*, *RPS6KA2*, *ITGB2*, and *TXK*)[3] (**Supplementary Table 3-2**). CpGs annotated to prominent inflammatory bowel disease therapeutic targets such as Tumor Necrosis Factor (*TNF*), Janus Kinase 3 *(JAK3)*, Interleukin 12B *(IL12B)*, Interleukin 23 Subunit Alpha *(IL23A)*, and Interleukin 1 Receptor Type1 (*IL1R1*) were amongst the disease-associated CpGs (**Supplementary Table 3-1**). Notably, enrichment analysis of our Crohn's disease associated CpGs indicated that they are more likely to occur in gene bodies (OR = 1.67, $P < 2.2 \times 10^{-16}$) and CpG shelves (OR = 1.42, $P = 5.0 \times 10^{-4}$), and are less likely to be in gene promoters (OR = 0.35, $P = 8.2 \times 10^{-13}$ for less than 200 base pairs from transcription start site

(TSS); and OR = 0.57, $P = 5.9 \times 10^{-08}$ for less than 1500 base pairs from TSS), and CpG islands (OR = 0.14, $P < 2.2 \times 10^{-16}$) and shores (OR = 0.59, $P = 9.7 \times 10^{-10}$). Exact details of the distribution of Crohn's disease associated CpGs in relation to gene regions and CpG islands are provided in **Supplementary Tables 3-3, 3-4**.

Gene expression profiles of differentially methylated genes in Crohn's disease. The 1189 Crohn's disease associated CpGs mapped to 717 unique genes. To better understand how these CpGs might reflect functional processes that are perturbed during the diagnosis of Crohn's disease, we examined the expression profiles of our differentially methylated genes in blood RNA-Seq data available from an independent data set consisting of 60 newly diagnosed pediatric patients with Crohn's disease and 12 controls[27]. Of the 585 (of 717) genes available for analysis after quality control, 162 (28%) of those were differentially expressed at FDR < 0.05 and 233 (40%) at the less stringent threshold of $P < 0.05$ (**Fig. 3-7** and **Supplementary Table 3-5**). Overlapped with these 233 differentially expressed genes were 295 of the Crohn's disease associated CpGs with an average of 1.3 CpG sites associated per gene (range = 1-4). As shown in **Fig. 3-8**, the direction of effects between DNA methylation and gene expression changes in relation to Crohn's disease appears to be context dependent with some CpG methylation-gene expression probes demonstrating negative association while others showed positive relationship, irrespective of the position of the CpG site in the associated gene (Fisher's test, $P > 0.05$; **Supplementary Table 3-6**). Collectively, these observations suggest the integrative involvement of methylomic and transcriptomic processes underlying Crohn's disease pathogenesis.

Biological processes enriched in Crohn's disease associated CpGs. Next, we evaluated whether the disease-associated CpGs in blood were enriched for biological processes relevant to Crohn's disease. Our pathway enrichment analysis identified 164 KEGG pathways that were more likely to occur in the Crohn's disease associated CpGs than would be expected by chance (FDR < 0.05; **Supplementary Table 3-7**). Among these were pathways relevant to immune function including TNF-alpha, Jak-STAT, Rap1 and

PI3K-Akt signaling; and inflammation such as the IL-17 signaling pathway, cytokine-cytokine receptor interaction and chemokine signaling.

Relationship between DNA methylation signatures of Crohn's disease and inflammation. To further evaluate the relationship between the disease-associated methylation signatures and inflammation, we tested the 1189 CpG sites for association with plasma C-reactive protein (CRP) levels, a marker of inflammation, and compared the estimated effects of methylation changes on CRP versus Crohn's disease at diagnosis. The relationship was extremely strong ($R = 0.91$, $P < 2.2 \times 10^{-16}$) suggesting that the methylation signatures of Crohn's disease either cause the inflammatory status of the patient, or directly result from it (**Fig. 3-9**). 1155 (97%) of the 1189 Crohn's disease CpGs exhibited directional consistency, and 872 (73%) showed statistically significant association with CRP ($P < 0.05$; **Fig. 3-9** and **Supplementary Table 3-8**).

Next to assess the relevance of these methylation signatures to Crohn's disease related inflammation, we compared the effect sizes of Crohn's disease associated CpGs on Crohn's disease and CRP in our dataset to a recently published meta-analysis of epigenome-wide association of CRP in subjects that were not selected for any particular disorder[28]. These meta-analyses comprised of 8863 participants that were sampled from 9 different prospective cohort studies with a wide-range of focus from cardiometabolic phenotypes to physical activity, intelligence, and aging. Surprisingly, we noted an extremely strong correlation between the estimated effects of Crohn's disease associated CpGs on Crohn's disease and chronic low-grade inflammation that is associated with a broad range of complex diseases, including diabetes and cardiovascular disease (**Fig. 3-9**). To validate this inference, we examined the overlap and directional consistency of previously reported Crohn's disease CpGs[3] with the CRP meta-analysis[28], obtaining consistent results (**Supplementary Table 3-2**).

Longitudinal dynamics of DNA methylation in Crohn's disease. In order to establish the direction of causality of this strong association, we next examined the longitudinal dynamics of inflammation and

disease-associated methylation profiles during the course of the disease. Since frontline treatment of inflammatory bowel disease attempts to lower inflammation in the patients, and as expected, CRP levels in patients at follow-up 1 to 3 years after diagnosis were dramatically lower than at diagnosis ($P = 8.4 \times 10^{-9}$; **Fig. 3-10**), the direction of causality seems obvious. The patients received treatment known to lower inflammation, and the primary marker for inflammation was much lower. Next to assess the dynamics of DNA methylation pre- and post-treatment, we compared the methylation profiles at follow-up to the profiles at diagnosis. Here the effects at diagnosis reflect differences between newly diagnosed patients and controls, and the effects at follow-up reflect differences in patients before and after treatment. At 1179 (99.2%) of the 1189 sites associated with Crohn's disease at diagnosis, the sign of the effect had reversed, while the magnitude of the change remained the same, generating a strong negative correlation ($R = -0.93$, $P < 2.2 \times 10^{-16}$; **Fig. 3-11** and **Supplementary Table 3-8**). In fact, after treatment, methylation at these sites is largely indistinguishable in patients versus controls (**Figs. 3-12, 3-13**). We noted similar results even after stratifying patients based on disease progression to B2 ($R = -0.91$, $P < 2.2 \times 10^{-16}$ for progressors; $R = -0.90$, $P < 2.2 \times 10^{-16}$ for non-progressors; **Fig. 3-11b**, **c**). Collectively, our data establish that during the course of the disease, methylation patterns that are disrupted at the diagnosis of Crohn's disease revert back to the levels seen in controls, irrespective of the disease behavior states (B1 or B2). Only 10 (0.8%) CpGs had the same sign of effect during diagnosis versus follow-up; these CpGs corresponded to 8 unique genes (*RORC*, *CXXC5*, *GMNN*, *GPR183*, *DIDO1*, *SMARCD3*, *ESPNL*, and *EPS8L3*; **Supplementary Table 3-8**). Interestingly, genes such as *RORC*, *SMARCD3* and *EPS8L3* have previously been linked with inflammatory bowel disease, including in genome-wide association studies and gene expression studies[29-33]. For instance, *RORC* encodes a key transcription factor for the Th17 pathway involved in transcriptional regulation of the effector cytokines *IL17A*, *IL17F*, *IL21*, *IL22*, *IL26* and *CCL20*[34] and was previously reported to be differentially expressed in peripheral blood and intestinal Crohn's disease samples compared to healthy controls[30].

Relationship between DNA methylation signatures of Crohn's disease and disease activity. Next, to test whether our finding of methylomic and inflammatory reversion extends to other clinical and laboratory measurements, we examined the measures of the pediatric Crohn's disease activity index (PCDAI), a multi-item index which incorporates clinical symptoms, laboratory parameters, and endoscopic findings[35] and noted higher PCDAI scores (median score of 30; $n = 159$) during diagnosis which were significantly lower during the follow-up (median score of 5; $n = 149$; $P < 2.2 \times 10^{-16}$; **Fig. 3-14**). Following association analysis of PCDAI with the 1189 sites ($n = 308$ samples; see Methods), estimated effect sizes demonstrated a strong correlation with their estimated effects on Crohn's disease ($R = 0.91$, $P < 2.2 \times 10^{-16}$), suggesting a potential relationship between disease activity (based on PCDAI) and DNA methylation in blood (**Supplementary Table 3-8** and **Fig. 3-15**).

Role of medication in DNA methylation reversal. To evaluate the potential impact of therapy on the reversal of Crohn's disease associated methylation signatures, we stratified patients based on the class of medications they were taking at the time of the follow-up sampling (**Table 3-1**). Comparative analysis of methylation levels in blood at the time of the follow-up did not reveal any genome-wide significant differences between subsets of patients who received biologics, immunomodulators, biologics plus immunomodulators, or other drugs, except for one CpG, cg24052338 (in the 3'UTR region of *ZNF837*), that showed significant association with other drugs (FDR < 0.05; **Supplementary Tables 3-9 to 3-12**), indicating that the medication is probably not the primary contributor to the methylomic reversion during the course of the disease. Boxplots depicting methylation beta values of follow-up patients' samples stratified based on the class of medications at the top 5 disease-associated CpGs were shown in **Fig. 3-16**. However, it is possible that the medication-induced reductions in inflammation and consequently disease activity may account for the reversal of the disrupted methylomic signatures in blood. Consistent with our interpretation, a study of site-specific methylation differences in peripheral blood mononuclear cells[10] and a different study of 2 colonic mucosa samples[9], both showed methylomic reversion in response to treatment and/or disease remission via modulation of the disease-specific inflammatory characteristics. In contrast,

another study reported stable methylation differences in patients with newly diagnosed (treatment-naïve) versus established inflammatory bowel disease (exposed to inflammatory bowel disease medications)[3]. However, our finding that the reversion of disease-associated methylation patterns associates with clinical characteristics of the disease (CRP, PCDAI) rather than medication underscores the importance of having prospectively followed inception cohorts with well-documented disease measures.

Understanding the causal versus consequential roles of DNA methylation in Crohn's disease. Given the methylomic reversion occurring during the course of the disease, and its strong relationship with plasma CRP levels, it appears that Crohn's disease associated methylation signatures are tightly linked to inflammation rather than the disease development itself. However, if methylation at specific sites plays a role in disease development, their identification would provide valuable therapeutic targets. To distinguish sites that may have causal versus consequential roles in Crohn's disease, we employed the concept of Mendelian randomization as operationalized by Wahl *et al*.[25]. As shown in **Fig. 3-17,** CpGs that emerge on the path between the instrumental variable and the outcome (Crohn's disease; Model 1), where methylation appears to mediate genetic risk of Crohn's disease, are interpreted to be causal rather than being the consequence (Model 2) of the disease. 194 out of the 1189 Crohn's disease CpGs at diagnosis associated with DNA sequence variation in a *cis* methylation quantitative trait loci (mQTL) analysis (FDR < 0.05; **Supplementary Table 3-13**). Of these, 174 CpGs with genetic data available for the associated mQTL SNPs from a large meta-analysis of genome-wide association studies (GWAS) of Crohn's disease[23] were evaluated for potential causal relationships between methylation in blood and Crohn's disease. For each CpG, we identified the most significantly associated SNP (sentinel mQTL) and applied the concept of Mendelian randomization using the sentinel mQTL SNP as the instrumental variable, CpG as a mediator, and Crohn's disease as the outcome for methylation cause of Crohn's disease (Model 1, **Supplementary Fig. 3-17**).

Causal role of DNA methylation in Crohn's disease. Using this set of sentinel SNP-CpG pairs, we first investigated SNP to DNA methylation (β coefficient; β_1) and DNA methylation to Crohn's disease (β_2)

relationships to obtain predicted effects (β_1 x β_2) of the corresponding SNPs on Crohn's disease via DNA methylation (**Fig. 3-18**). Subsequently, genetic effect sizes of SNPs on Crohn's disease were obtained from large GWAS meta-analyses of Crohn's disease[23] to assess the observed effects of genotypes at these SNPs on Crohn's disease (β_3). If methylation contributes causally to Crohn's disease, we would expect the observed effect of SNP on phenotype to be consistent, if not equivalent to its predicted effect mediated through methylation. Notably, methylation changes at 3 CpGs (cg15706657, cg23216724: near *GPR31*; and cg20406979: near *RNASET2*) showed significant causal associations with Crohn's disease at diagnosis (FDR < 0.05; **Fig. 3-18** and **Supplementary Table 3-14**). Consistent with the potentially causal effect, methylation levels at cg23216724 and cg20406979 became even more pronounced or remained about the same without exhibiting signs of reversion during the follow-up (**Fig. 3-19**), supporting our inference regarding causality. Further support for their potentially causal influence is provided by the observation that all 3 CpGs are influenced by the known inflammatory bowel disease-associated SNP, rs1819333, identified through large GWAS[23,36]. The inflammatory bowel disease-risk locus containing the SNP rs1819333 harbors (within 1 Mb flanking rs1819333) key genes *RPS6KA2*, *RNASET2* and *CCR6* that have previously been implicated in inflammatory bowel disease pathology at both genomic and/or molecular levels, including in transcriptomic and epigenomic studies[3,23,36-38]. Although genetic variation at rs1819333 has been associated with significant risk for Crohn's disease susceptibility, underlying causal gene(s) and molecular mechanisms of this strong GWAS association are yet to be elucidated. Remarkably, all 3 identified potentially causal CpGs that are associated with rs1819333 were recently shown to causally regulate transcriptional levels of *RPS6KA2* in peripheral blood using a summary data-based Mendelian randomization (SMR) approach[39]. Taken together, these findings suggest DNA methylation as a potential mediator of genetic effects of rs1819333 on Crohn's disease, possibly through transcriptional regulation of *RPS6KA2*.

Consequential role of DNA methylation in Crohn's disease. Conversely, to identify Crohn's disease associated sites where changes in methylation are more likely to be the consequence of the disease, we used

a weighted Crohn's disease genetic risk score (see Methods) as an instrumental variable, Crohn's disease as the mediator and methylation as the outcome (Model 2, **Fig. 3-17**). An extremely strong correlation ($R = 0.86$; $P < 2.2 \times 10^{-16}$) between the observed effect of the weighted genetic risk score on methylation and its predicted effect through Crohn's disease was seen (**Fig. 3-18**). In particular, we identified 8 CpGs corresponding to 7 genes that showed significant consequential associations with Crohn's disease at diagnosis (FDR < 0.05; **Fig. 3-5b** and **Supplementary Table 3-15**). In keeping with their consequential role, methylation levels at these CpG sites demonstrated drastic changes approaching levels seen in controls during the follow-up (**Fig. 3-20**). Differential methylation at cg18942579: *TMEM49* and cg17501210: *RPS6KA2*, CpGs that have been consistently found to be associated with Crohn's disease[3,8], appears to be a consequence of the disease rather than exerting causal effects. For instance, cg17501210 has previously been reported to be the top-most differentially methylated CpG site in peripheral blood of inflammatory bowel disease patients, whose effects were (i) even more pronounced in purified $CD14^+$ monocytes; (ii) strongly correlated with disease-relevant markers, including CRP, albumin and hemoglobin; and (iii) not influenced by treatment status. Overall, the strong correlation between the observed and predicted effects in **Fig. 3-18** suggests that most disease-associated methylation changes are triggered by the onset of Crohn's disease. This finding is consistent with findings from other complex diseases[25,40,41], suggesting that only a minority of the trait-associated methylation changes are likely to exert causal effects.

Role of DNA methylation in diagnosis and prognosis of Crohn's disease. Biological data that enable accurate diagnosis and/or prognosis of inflammatory bowel disease has always been of considerable interest from the point of view of clinical application. In line with previous studies[3], we noted that peripheral blood DNA methylation data could indeed distinguish patients with Crohn's disease from controls (AUC = 0.91; **Fig. 3-21**). However, given the non-inflammatory nature of the sampled control subjects, supplemented by our finding that the signatures of methylation observed at diagnosis of Crohn's disease capture general inflammation rather than Crohn's disease-specific, we are hesitant to propose peripheral methylation as a diagnostic biomarker for Crohn's disease based on the prevailing evidence. Future studies of side-by-side

evaluation of methylation data from patients with different immune-mediated inflammatory diseases along with the disease-relevant tissue-specific inflammatory characterization are required to definitively establish the diagnostic potential conferred by methylation signatures of complex diseases.

Next, to assess the utility of methylation in prognosis, by stratifying Crohn's disease patients based on subsequent progression to complicated disease (see Methods), we asked if methylation signatures at diagnosis could predict who would in time progress to complicated Crohn's disease. In keeping with our finding from **Fig. 3-5** and **Supplementary Table 3-1**, we did not find any CpGs showing significant differences when the baseline methylation profiles of subsequent progressors were compared to non-progressors. Taken together, our data suggests that peripheral blood methylation profiles do not predict or change in relevance to the evolution or presence of Crohn's disease complications.

DISCUSSION

In conclusion, we characterized temporal relationships connecting methylation changes in blood with varying inflammatory characteristics at diagnosis and during treatment for Crohn's disease in children. Systemic inflammation has long been understood to be a pathogenetic hallmark of Crohn's disease, and medication to relieve the burden of inflammation has been part of all frontline treatment strategies for managing the disease. Our results provide convincing evidence that the signatures of methylation observed at diagnosis accompany acute inflammation that declines with treatment, but revert toward the levels seen in controls despite ongoing bowel disease, arguing that they are primarily a symptom of the disease rather than a cause. If so, treatment of inflammation signatures may fundamentally be treating the symptoms of Crohn's disease rather than the etiology, partially explaining why inflammatory bowel disease often remains a life-long remitting and relapsing disorder, despite effective treatment of the inflammation symptoms.

A caveat to this interpretation is that we measured circulating immune cells whereas the inflammation is manifest in the bowel. Data for gut-resident immune cells will be required to establish whether the

methylomic reversion we describe is also observed in the gut and affected by the diverse treatment regimens independent of the epithelial signature. By contrast, a recent study[12] of methylation in intestinal epithelial cells described a distinct inflammatory bowel disease profile more related to disease than inflammation, which was stable over time in the 23 patients examined, further suggesting the need for new drugs that treat the cells in which, persistent molecular changes underlie disease pathogenesis. Future epigenetic studies of inflammatory bowel disease could profile circulating, gut-resident immune, and gut epithelial cells in parallel, using our framework to definitively identify causal CpGs, and leverage the epigenome for the development of targeted therapeutics. This, in combination with the current armamentarium of inflammatory bowel disease medications that hold promise for successfully managing the disease by targeting the immune system, may put us one step closer to sustained remission and mucosal healing.

One of the long-term complications of inflammatory bowel disease is inflammation-related cancers[42-45], and our finding that aberrant DNA methylation of Crohn's disease is predominantly a consequence of inflammation provides a strong rationale for the molecular link between inflammatory bowel disease and cancer, as both chronic inflammation and aberrant DNA methylation having a known role in malignancy development. It remains to be seen whether these methylation signatures detected in blood as a consequence of inflammation may in part predict new onset, incident cancer, a major clinical consequence associated with inflammatory bowel disease.

Our study has certain limitations. Although, it was well powered to detect CpGs that are different between controls and newly-diagnosed Crohn's disease patients, and examine how they change during the course of the disease, we were limited in terms of power to apply a Mendelian randomization framework to infer causal associations, and hence, we may have missed detecting some CpGs with a potential causal influence. Despite this limitation, we identified differential methylation of several CpG sites associated with an inflammatory bowel disease-associated SNP, rs1819333, to be potentially causal to Crohn's disease. This result, however, should be interpreted with caution given the lack of replication. Nevertheless, Mendelian randomization revealed results consistent with findings from our longitudinal framework that the peripheral

blood methylation changes associated with Crohn's disease in children are predominantly a consequence of disease. Causal versus consequential analyses of adult cohorts should confirm the potential impact of blood-derived DNA methylation or the lack thereof in adult patients diagnosed with Crohn's disease.

Table 3-1: Summary of patient characteristics

	non-IBD controls	Crohn's disease at diagnosis	Crohn's disease at follow-up	*P* value; Crohn's disease Vs non-IBD	*P* value; Crohn's disease at diagnosis Vs follow-up
Number of samples	74	164	164		
Age in yrs [median (IQR)]	12.6 (9.8 to 14.3)	12.6 (10.6 to 15.0)	15.2 (13.2 to 17.3)	0.237	2.5×10^{-14}
Female sex (%)	34 (46%)	68 (41%)	68 (41%)	0.523	
Disease state [B1/B2]*	0/0	150/14	95/69		
Disease location [L1/L2/L3/missing]*		39/39/69/17	31/21/95/17		
Number of samples with CRP data	45	132	95		
CRP_{mgPerL} [median (IQR)]	0.5 (0.1 to 3.4)	4.4 (1.5 to 14.5)	1.0 (0.5 to 3.0)	0.037	0.0001
Number of samples with CRP < 1/CRP > 1	30/15	25/107	52/43		
Number of samples with PCDAI data	NA	159	149		
PCDAI [median (IQR)]	NA	30.0 (20.0 to 42.5)	5.0 (0.0 to 15.0)		2.2×10^{-16}
Number of samples with PCDAI ≤ 10/PCDAI > 10	NA	19/140	99/50		
Treatment Naïve (number of subjects)	74	164	0		
Biologics	NA	NA	76		
Immunomodulators	NA	NA	26		
Biologics plus immunomodulators	NA	NA	43		
Others*	NA	NA	11		
Medication data missing	NA	NA	8		

*In the Montreal classification of Crohn's disease behavior, B1 corresponds to inflammatory behavior with no stricturing or luminal penetrating complications and B2 to stricturing behavior with no luminal penetrating complications. Similarly, for disease location, L1 corresponds to disease located in the ileum, L2 in the colon, and L3 ileocolon. Others include patients who received 5-ASA, Steroids, and/or Antibiotics. NA – Not Applicable.

Figure 3-1: Principal component plots of baseline DNA methylation and genotype data for the 238 subjects included in the current study. Maroon dots represent Crohn's disease patients and grey dots indicate controls.

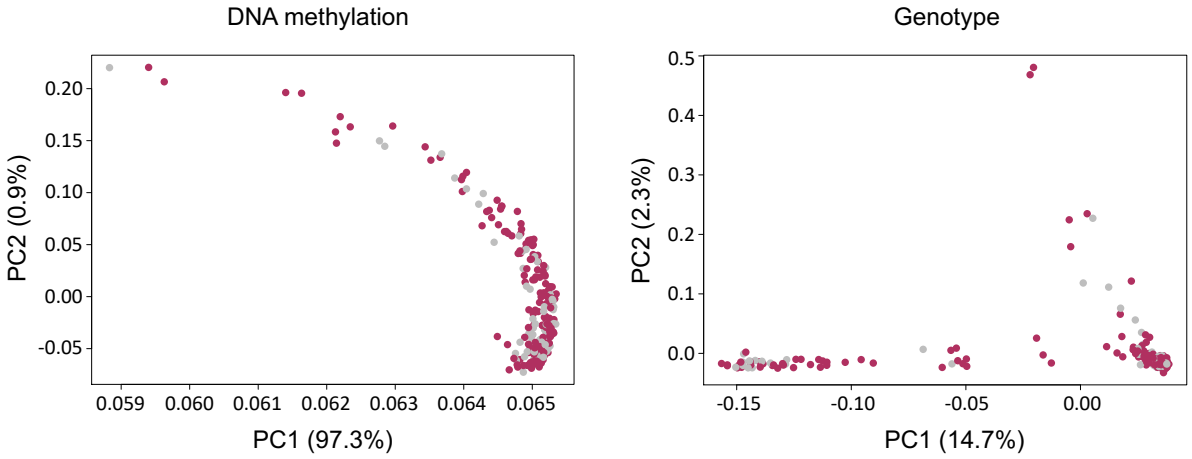


Figure 3-2: Crohn's disease at diagnosis is associated with methylation changes at 1189 CpG sites in blood. All the ~850 K CpG sites represented by dots are ordered by genomic position per chromosome (*x* axis). *P* values ($-\log_{10}$) of site-specific association with Crohn's disease is shown on *y* axis. Dots above the blue line represent CpGs reaching epigenome-wide significance (FDR < 0.05). Dots above the red line represent CpGs reaching epigenome-wide significance after Bonferroni correction (*n* = 114 CpGs).

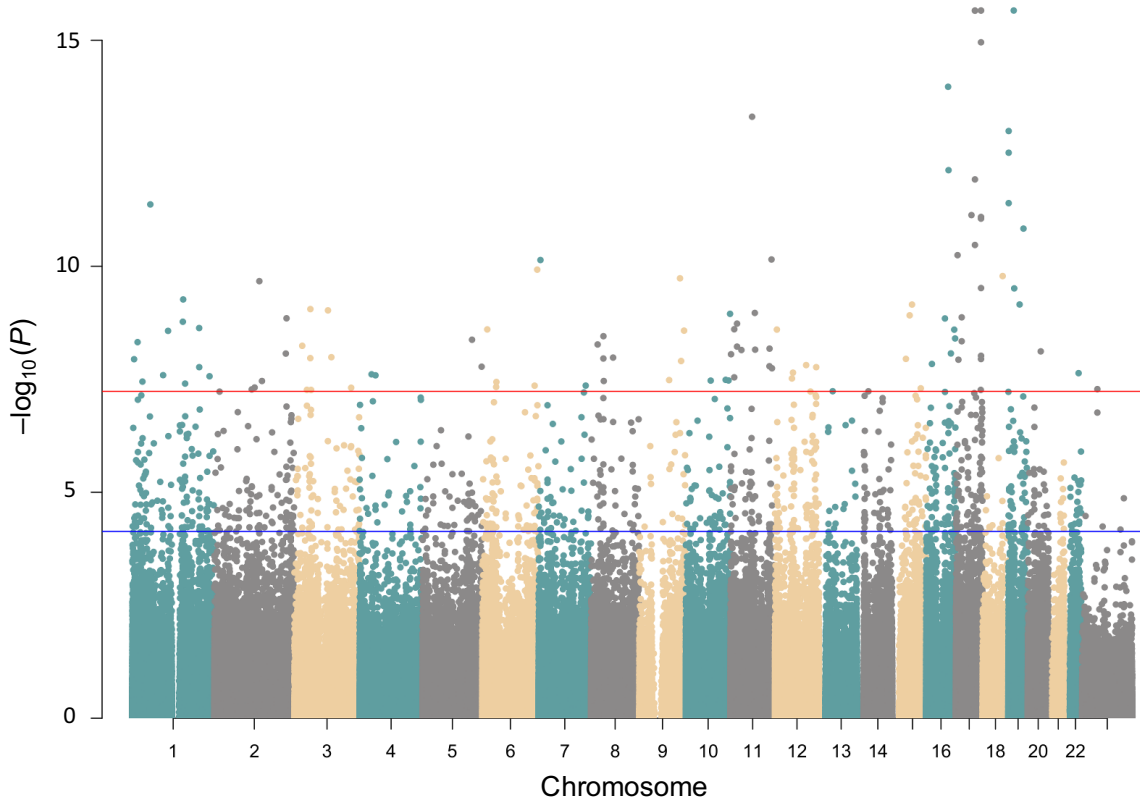


Figure 3-3: Boxplots depicting the estimated cell proportions of the 6 dominant cell types in blood. Cell subset estimates were computed based on DNA methylation data using the Houseman algorithm.

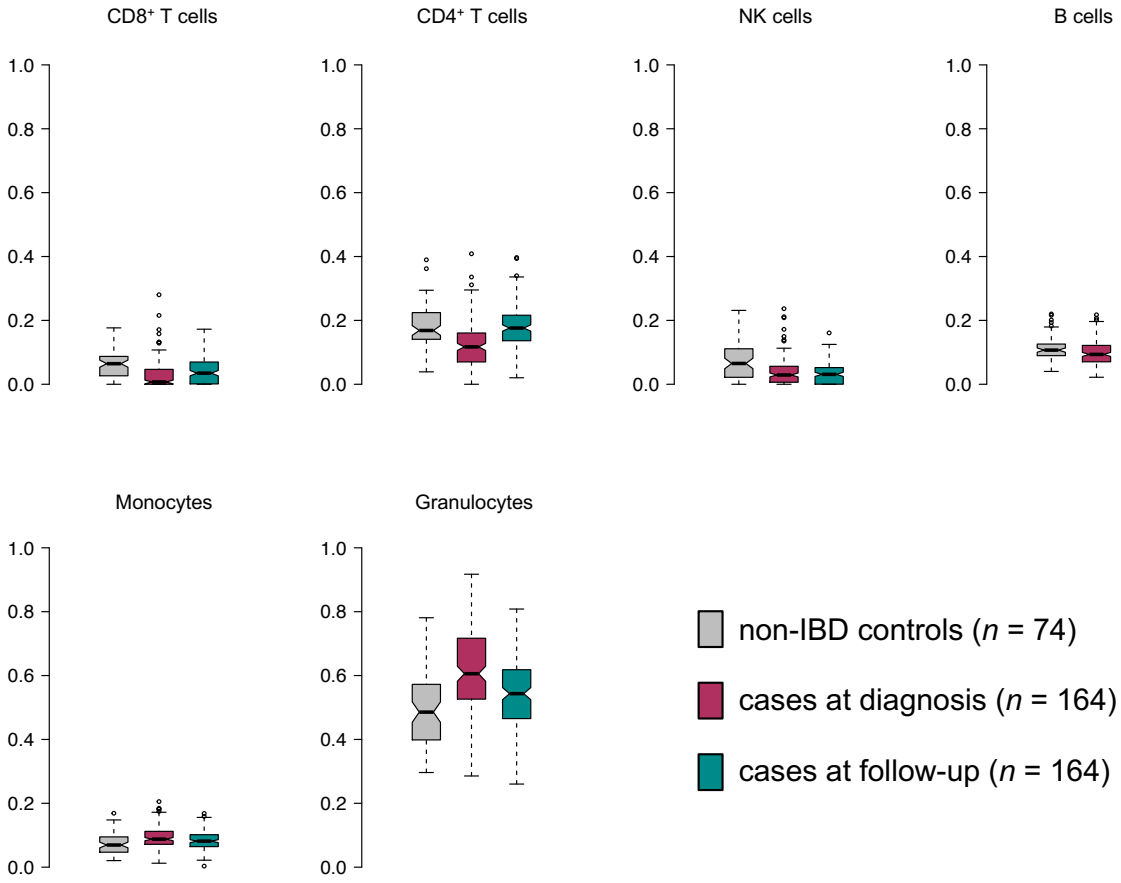


Figure 3-4: CpGs associated with Crohn's disease at diagnosis with or without adjusting for estimated cell subsets. (a, b) Volcano plots depicting the methylation difference (*x* axis) between controls and cases at diagnosis (a) with or (b) without adjusting for the estimated cell subsets, besides controlling for age, gender, and 3 genotype-based principal components. 1189 CpGs in (a) and 3188 in (b) reaching epigenome-wide significance after multiple test correction (FDR < 0.05; *y* axis) are shown in maroon.

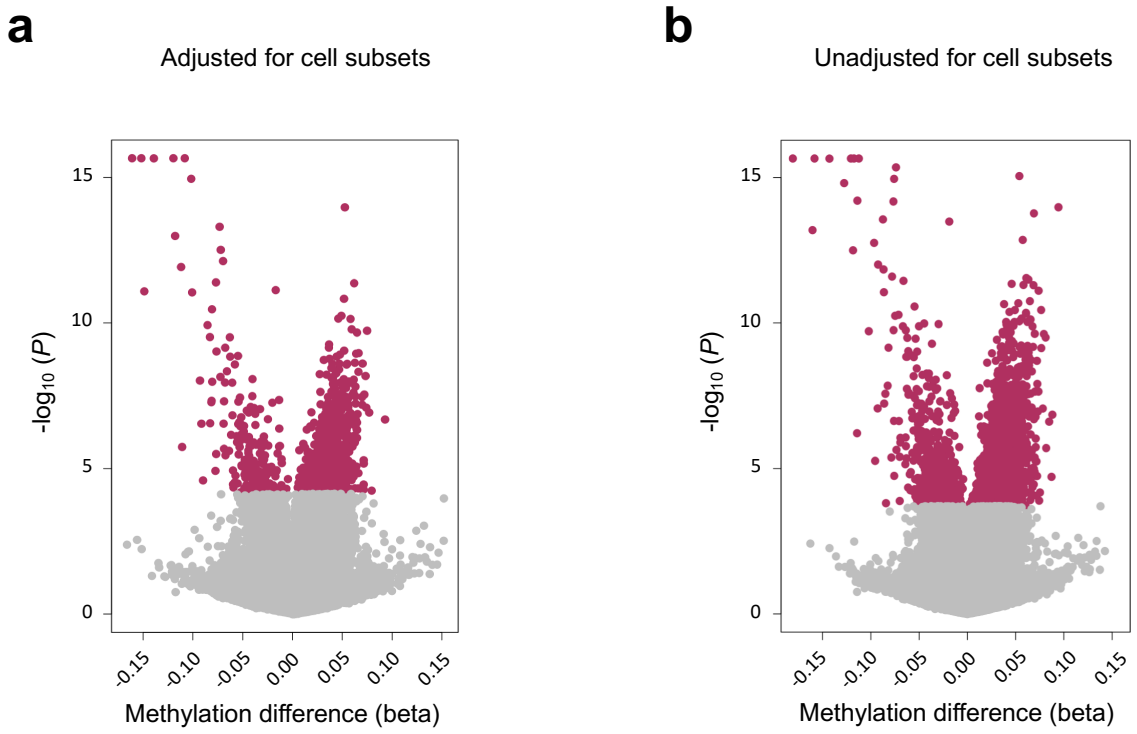


Figure 3-5: Shared methylomic contributions to B1 and B2 at diagnosis. (a-c) Effect of methylation changes at the ~850K sites on B1 versus B2 during diagnosis with (a) 1189 significant CpGs (74 controls Vs 150 B1, 14 B2), (b) 1007 significant CpGs (74 controls Vs 150 B1), and (c) 211 significant CpGs (74 controls Vs 14 B2) shown in maroon. The correlation coefficient and the *P* value of correlation is for the significantly differentially methylated CpGs (FDR < 0.05) that are colored maroon.

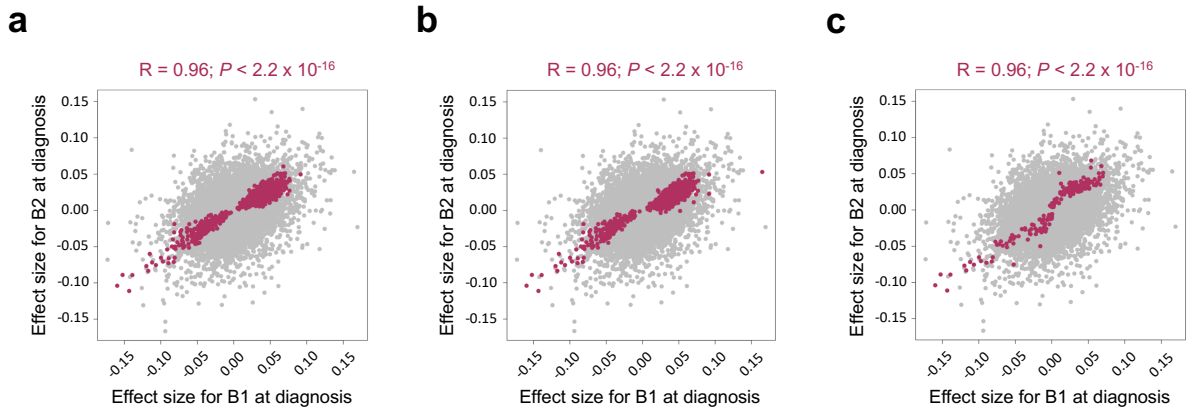


Figure 3-6: CpGs associated with Crohn's disease at diagnosis with or without adjusting for disease location. Scatterplot depicting the correlation between the effects of methylation changes on Crohn's disease at diagnosis with (*y* axis) or without (*x* axis) adjusting for disease location, besides controlling for age, gender, cell type proportions, and 3 genotype-based principal components. 1189 significant CpGs (74 controls Vs 164 Crohn's disease at diagnosis; unadjusted for disease location) are shown in maroon.

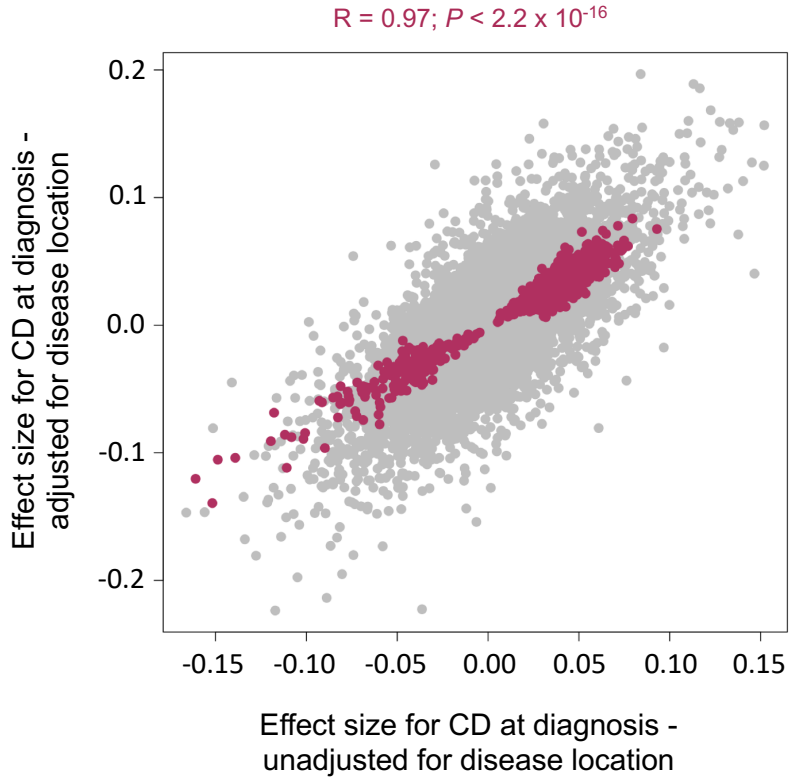


Figure 3-7: Volcano plot of differential gene expression in blood at diagnosis. Of the 585 out of 717 differentially methylated genes available for analysis in RNA-Seq data from an independent data set of 12 controls and 60 newly diagnosed pediatric patients with Crohn's disease, 233 genes highlighted in maroon showed differential expression ($P < 0.05$). Log_2 fold change difference between cases and controls is shown on *x* axis and $-\log_{10} P$ value of association on *y* axis.

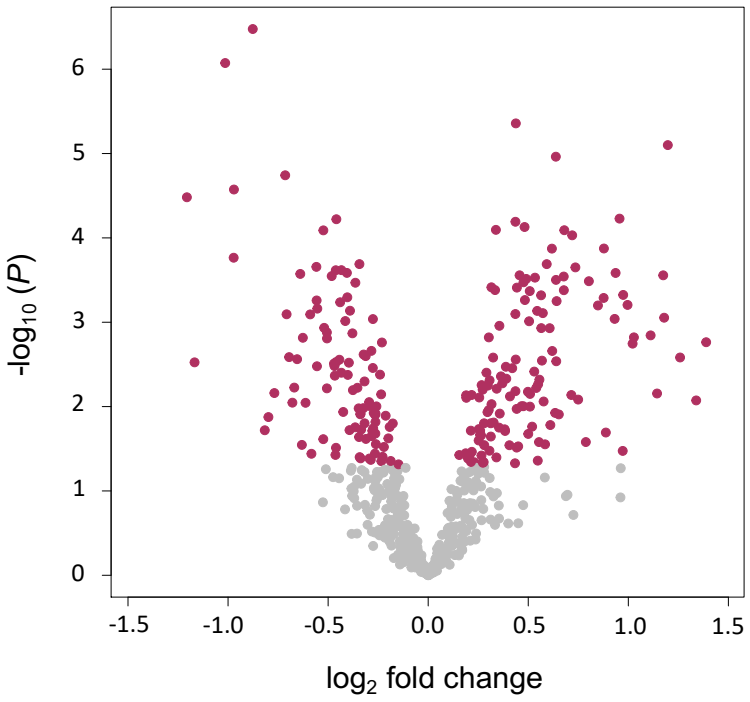


Figure 3-8: Overlap between DNA methylation and transcriptional changes at genes that are both differentially methylated and differentially expressed in Crohn's disease. Scatter plot depicting the relationship between CpG methylation (*x* axis) and their putative gene expression (*y* axis) changes in Crohn's disease at diagnosis. Colors represent position of methylation probes in relation to the gene. Number of CpG-gene expression probes in each quadrant (A-D) are shown.

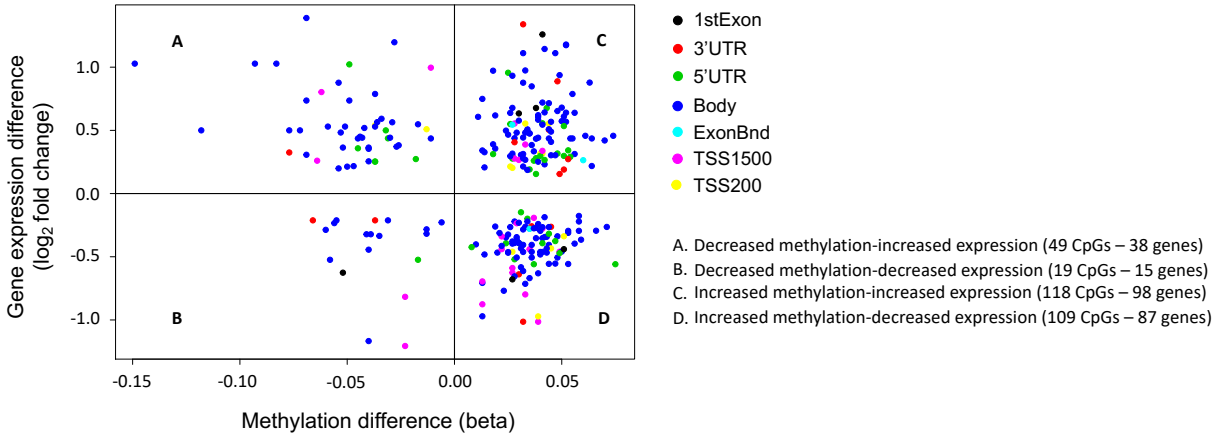


Figure 3-9: Methylation signatures of Crohn's disease reflect inflammatory status of the patient. (a) For the 1189 Crohn's disease associated CpGs, estimated effects (n = 164 cases and 74 controls) on Crohn's disease at diagnosis (x axis) are strongly correlated with the estimated effects (n = 272: 45 controls, 132 at diagnosis and 95 follow-up samples) on plasma CRP levels within the same subjects (y axis). Maroon dots represent Crohn's disease CpGs that showed significance with plasma CRP ($P < 0.05$). (**b, c**) At the 206 (of 218) CRP-associated CpGs in the latest meta-analysis (n = 8863) by the Ligthart *et al.*[28] (y axis), (**b**) 199 had effects on Crohn's disease at diagnosis, and (**c**) 196 had effects on CRP, in the same direction in our data. Maroon dots are CpGs from[28] that showed significance with (**b**) Crohn's disease and (**c**) CRP in our data ($P < 0.05$).

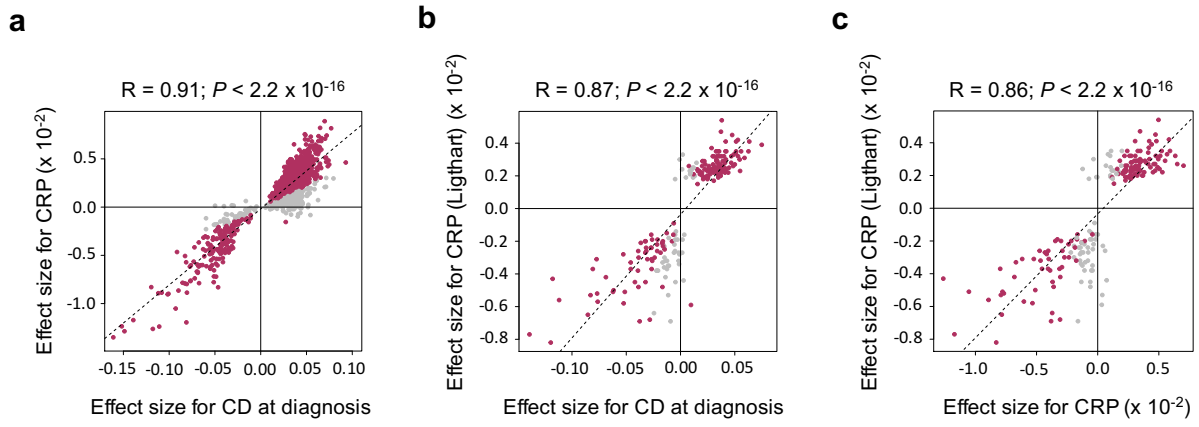


Figure 3-10: Boxplot depicting the log_2 transformed plasma CRP levels (mg/L) between controls, patients at diagnosis and during follow-up. Patients at diagnosis had higher levels of CRP compared to controls (P = 6.9 x 10^{-6}), and were significantly lower at the time of the follow-up (P = 8.4 x 10^{-9} Vs patients at diagnosis).

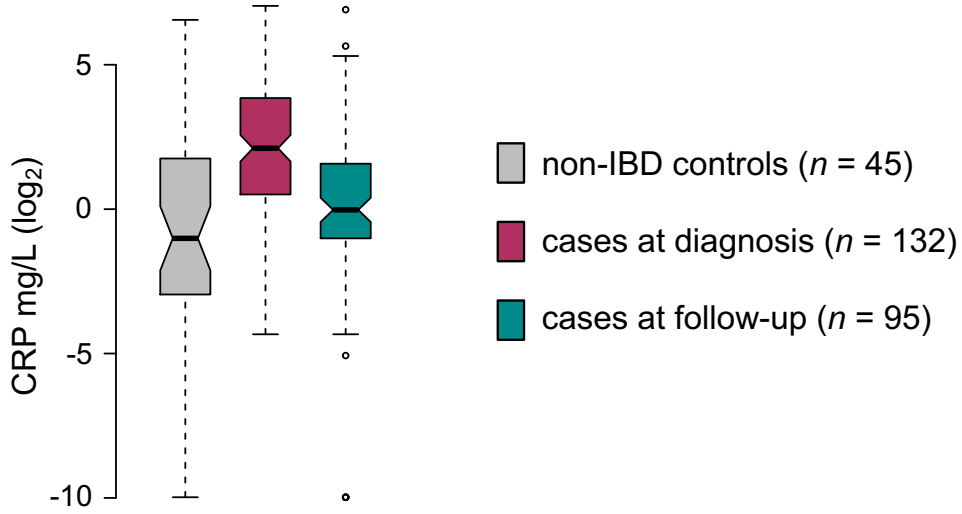


Figure 3-11: Disrupted methylation patterns during the diagnosis of Crohn's disease revert back during the course of the disease. (**a**) Estimated methylation differences between diagnosis (n = 164) and follow-up (n = 164; y axis) were of similar magnitude to baseline differences between cases (n = 164) and controls (n = 164; x axis). All ~850K CpG sites are shown; the correlation coefficient and P value is for the 1189 Crohn's disease CpGs that are colored maroon. (**b**) Same comparison for patients (n = 55) who received an initial diagnosis of B1 at the time of diagnosis and progressed to B2 during the course of the follow-up period. (**c**) Same comparison for patients (n = 95) who started and remained as B1 during diagnosis and follow-up.

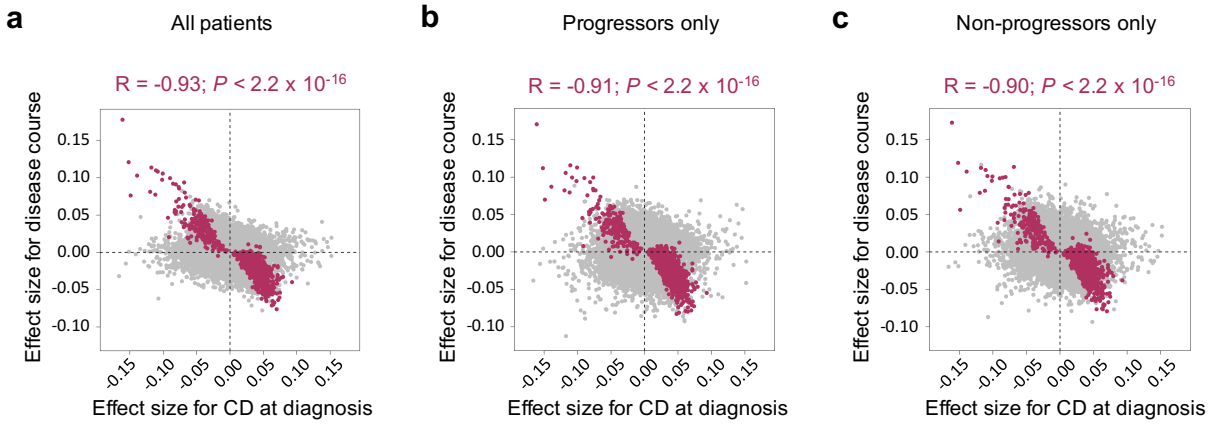


Figure 3-12: Heat map depicting the *z* scored methylation proportions in blood of controls, patients with Crohn's disease at diagnosis and follow-up. Grey, light pink and light green bars immediately above the heat map indicate controls (*n* = 74), patients with Crohn's disease at diagnosis (*n* = 164) and follow-up (*n* = 164), respectively. Each row corresponds to one of the 1189 CpG sites that are associated with Crohn's disease at diagnosis compared to controls after adjusting for age, gender, estimated cell subset proportions and genotype based principal components. The heat map is color-indexed according to the *z* score of each CpG site from low (light blue) to high (red) methylation beta value.

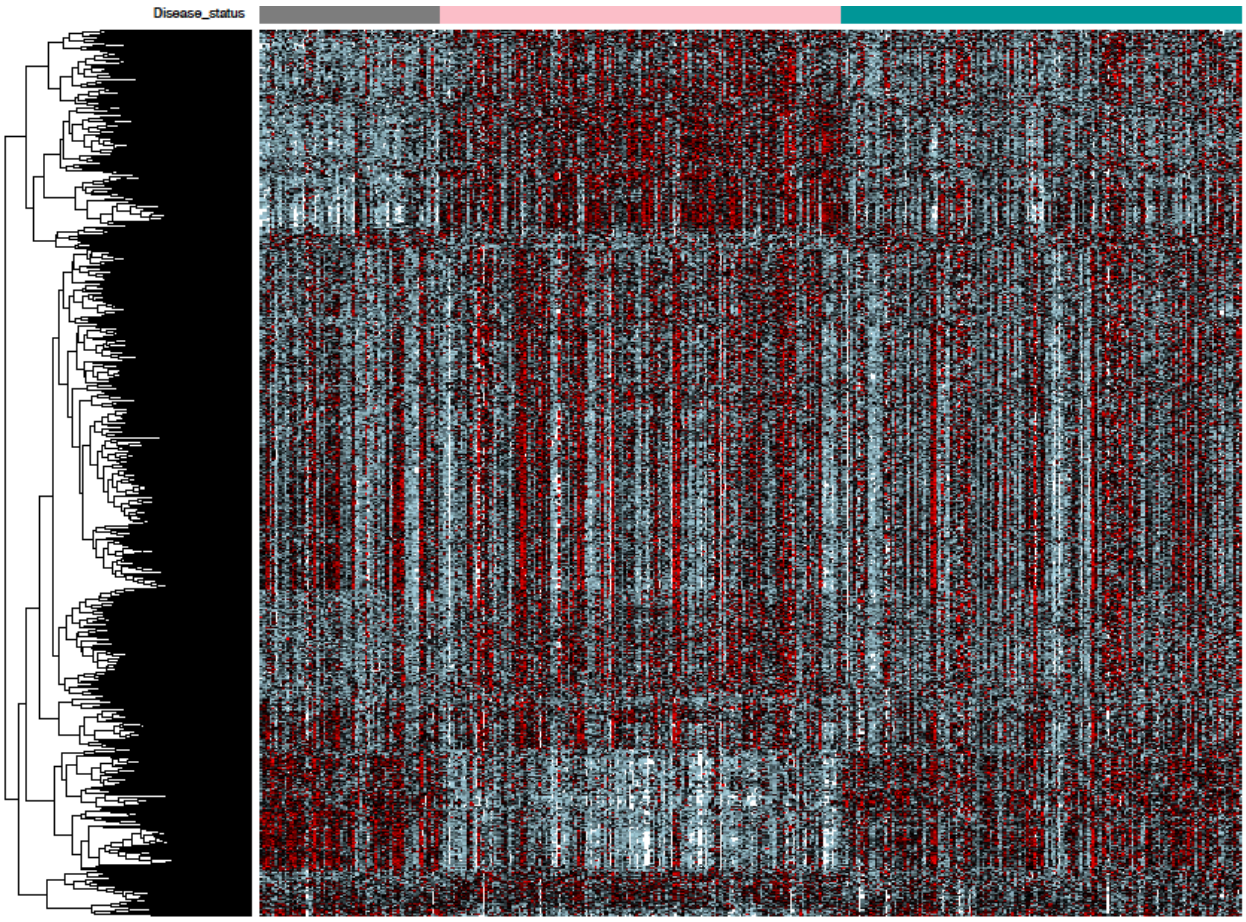


Figure 3-13: Boxplots depicting the methylation proportions of controls, patients at diagnosis and during follow-up at the top 6 CpG sites associated with Crohn's disease at diagnosis.

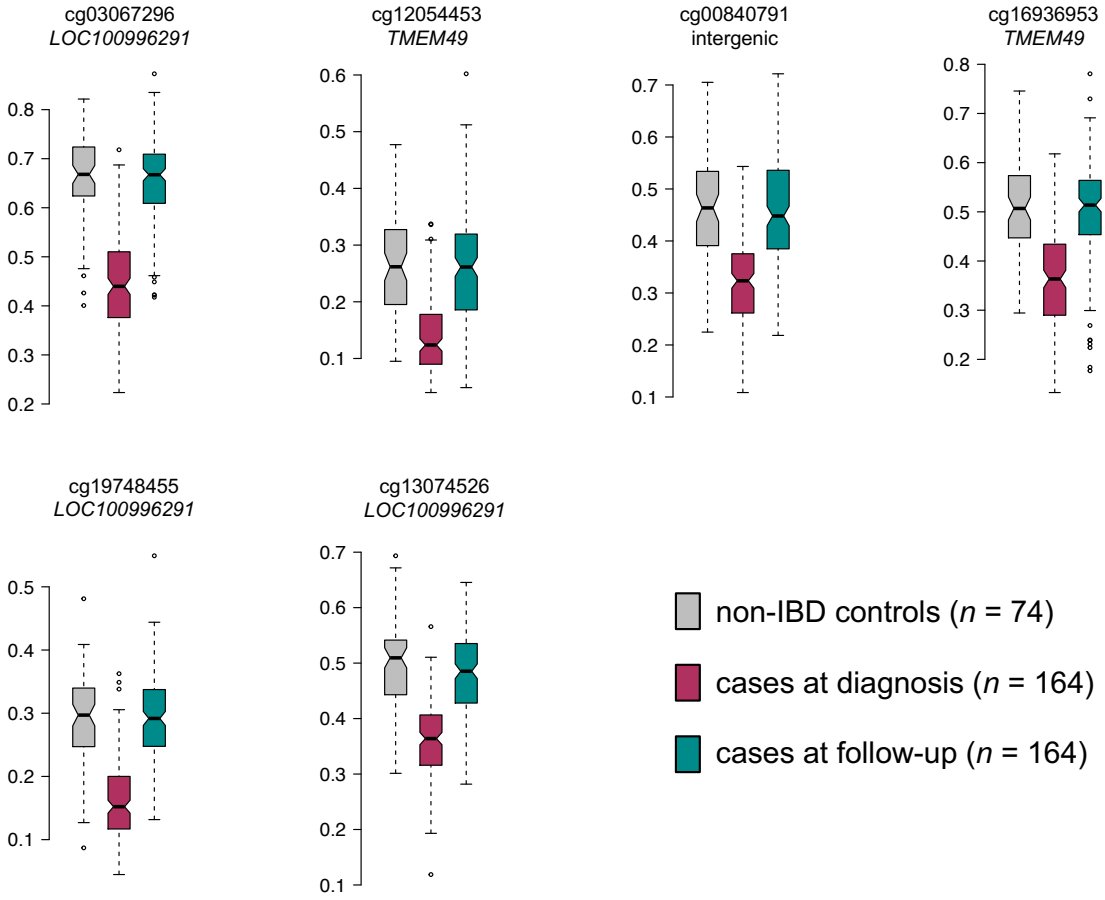


Figure 3-14: Boxplot depicting the PCDAI scores of patients at diagnosis and during follow-up. Patients at diagnosis had a median disease activity score of 30, which was significantly lower at the time of the follow-up (median score of 5; $P < 2.2 \times 10^{-16}$).

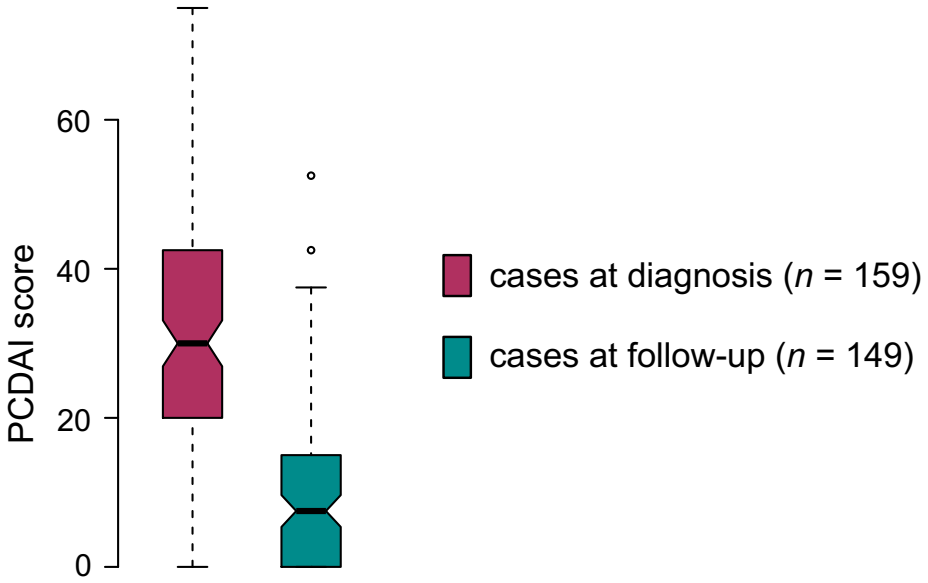


Figure 3-15: Effect of DNA methylation changes at the 1189 CpG sites on Crohn's disease at diagnosis (n = 164 cases and 74 controls; x axis) is strongly correlated with the effect on PCDAI scores (n = 308; y axis). 799 CpGs that showed significant association with PCDAI ($P < 0.05$) are shown in maroon.

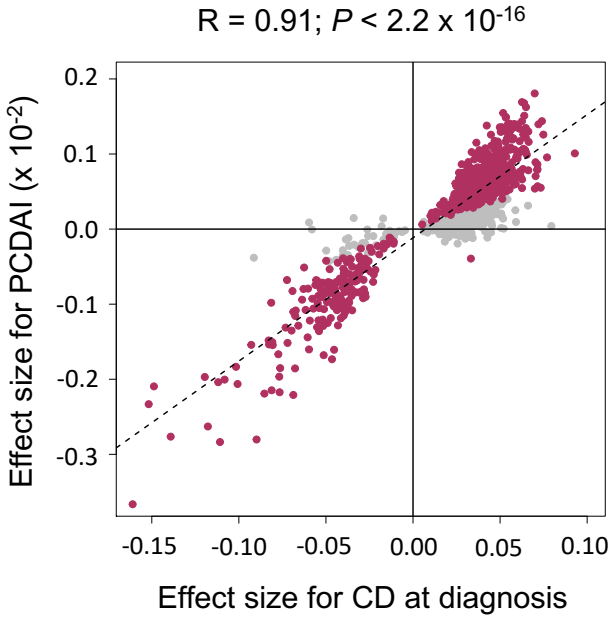


Figure 3-16: Boxplots demonstrating the impact of the class of medications on methylation beta values at the top-five Crohn's disease associated CpGs during follow-up. Methylation levels for the same CpGs in controls and Crohn's disease patients at diagnosis are also shown.

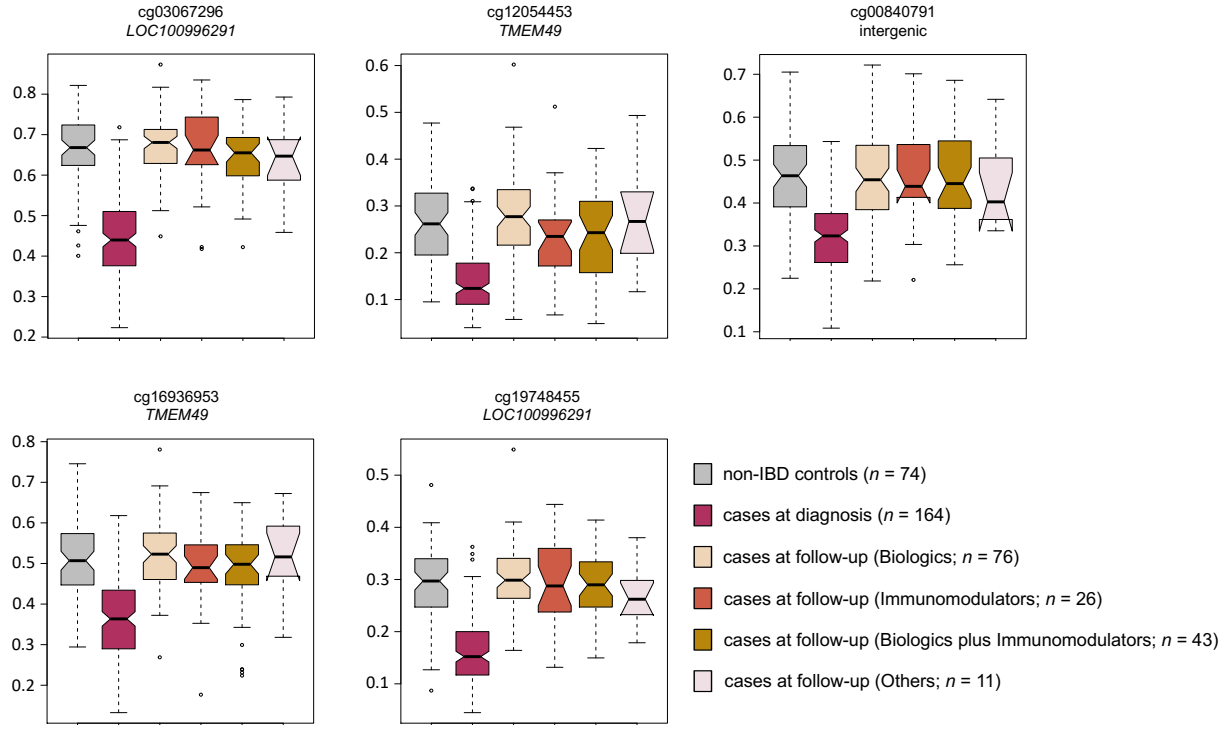


Figure 3-17: Possible models when applying genetic association and the concept of Mendelian randomization to epigenome-wide association studies. In Model 1, methylation changes at the CpG site (intermediate phenotype) of interest falls on the causal path between the instrumental variable (SNP or a cumulative effect of a combination of SNPs) and the outcome (Crohn's disease), where methylation appears to mediate genetic risk of the outcome. Such methylation changes are more consistent with a causal role on the outcome rather than consequential, and are therefore inferred to be causal to the outcome. In Model 2, methylation changes at the site of interest falls outside the causal path and such methylation changes are inferred to be a consequence of Crohn's disease. In this model, Crohn's disease is considered as an intermediate phenotype and DNA methylation as the outcome. Another possible model is the model of pleiotropy (Model 3), where DNA methylation and Crohn's disease are independent, but found associated because of the shared pleiotropic effect of a genetic variation.

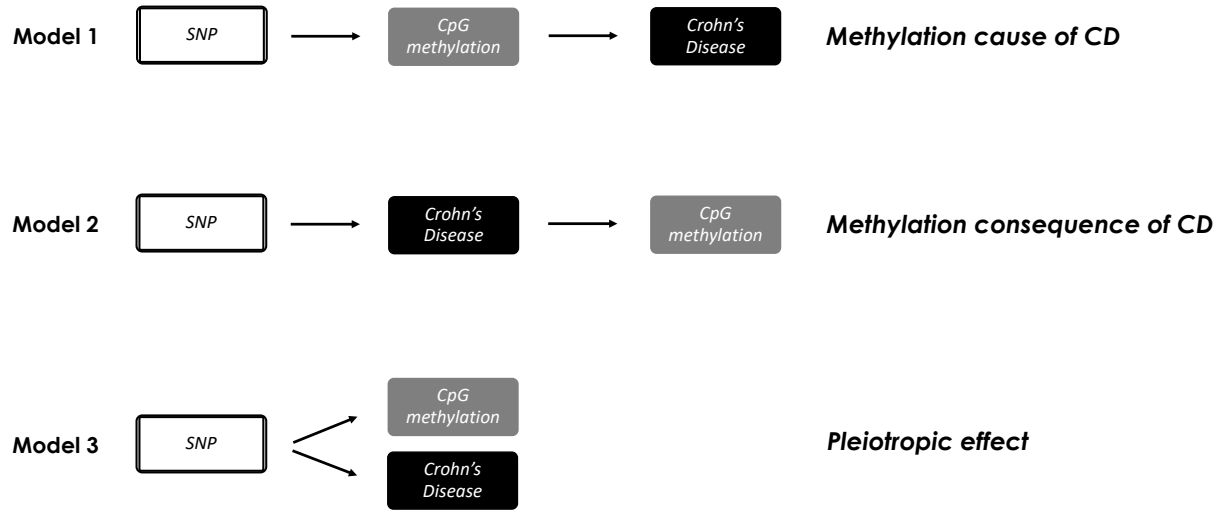


Figure 3-18: Evaluation of directionality among Crohn's disease associated CpG sites. Schematic diagram and results of genetic association and the concept of Mendelian randomization framework implemented to clarify the causal versus consequential role of Crohn's disease associated methylation changes in blood. (**a**) Overall strength of causality of the 174 CpGs tested for methylation cause of Crohn's disease is inferred from the correlation between the observed (*y* axis) and predicted effects (*x* axis) of SNP on Crohn's disease. To infer causality of individual CpG sites, the association of the sentinel mQTL with Crohn's disease should be significant (FDR < 0.05). 95% CI error bars are shown for the 3 CpG sites with an associated mQTL that also associated with Crohn's disease. 82 of the 174 CpGs that demonstrated directional consistency are shown in maroon; CpGs that are directionally inconsistent with the observed versus predicted effects are shown in grey. (**b**) Observed effect of Crohn's disease genetic risk score on methylation (*y* axis) is highly correlated with its predicted effect (*x* axis) through Crohn's disease, suggesting a strong consequential signal at the 194 CpG sites investigated. 95% CI error bars are shown for 8 CpGs demonstrating statistically significant consequential association (FDR < 0.05). 142 of the 194 CpGs with directionally consistent effects are shown in maroon; CpGs that are directionally inconsistent are shown in grey.

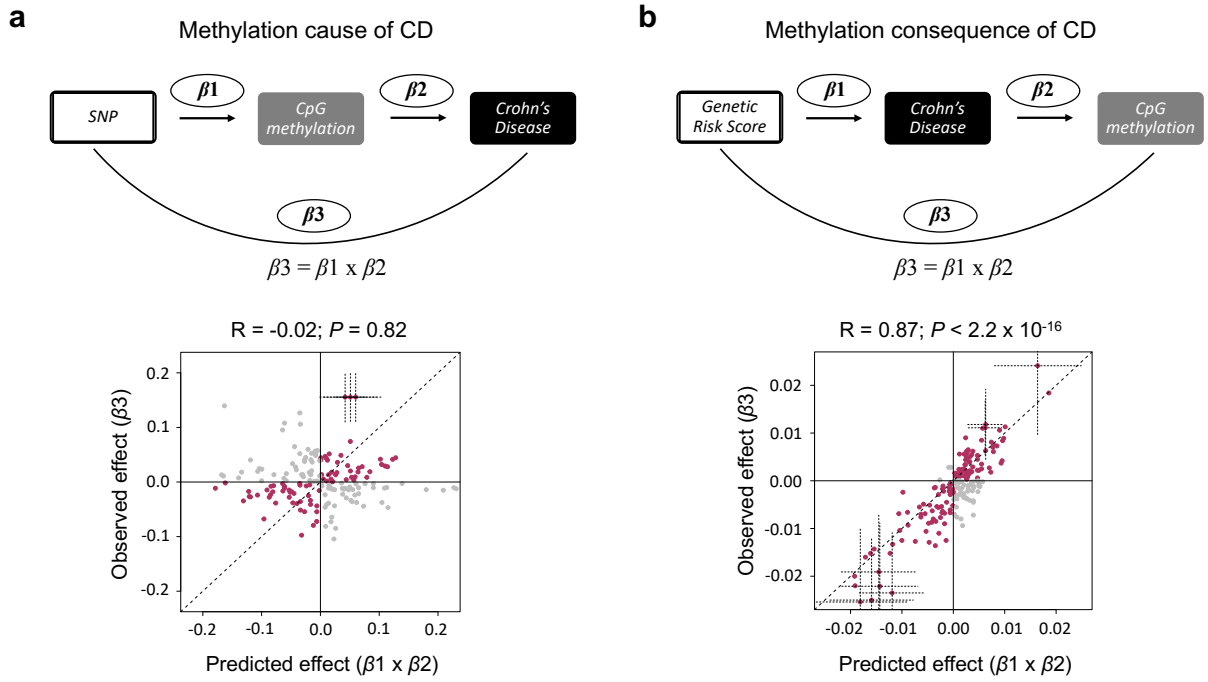


Figure 3-19: Boxplots depicting the methylation proportions of controls, patients at diagnosis and during follow-up at the 3 CpG sites that showed significant causal association with Crohn's disease at diagnosis.

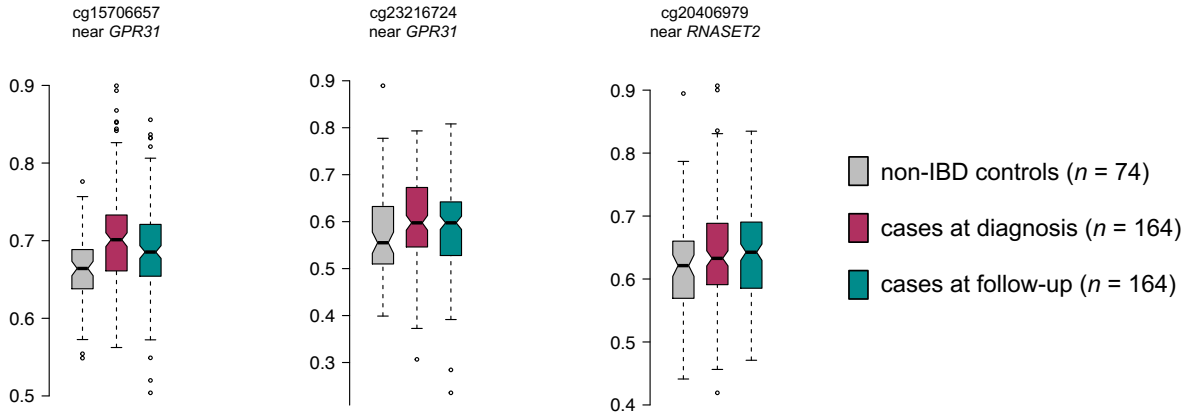


Figure 3-20: Boxplots depicting the methylation proportions of controls, patients at diagnosis and during follow-up at the top 6 CpG sites that showed significant consequential association with Crohn's disease at diagnosis.

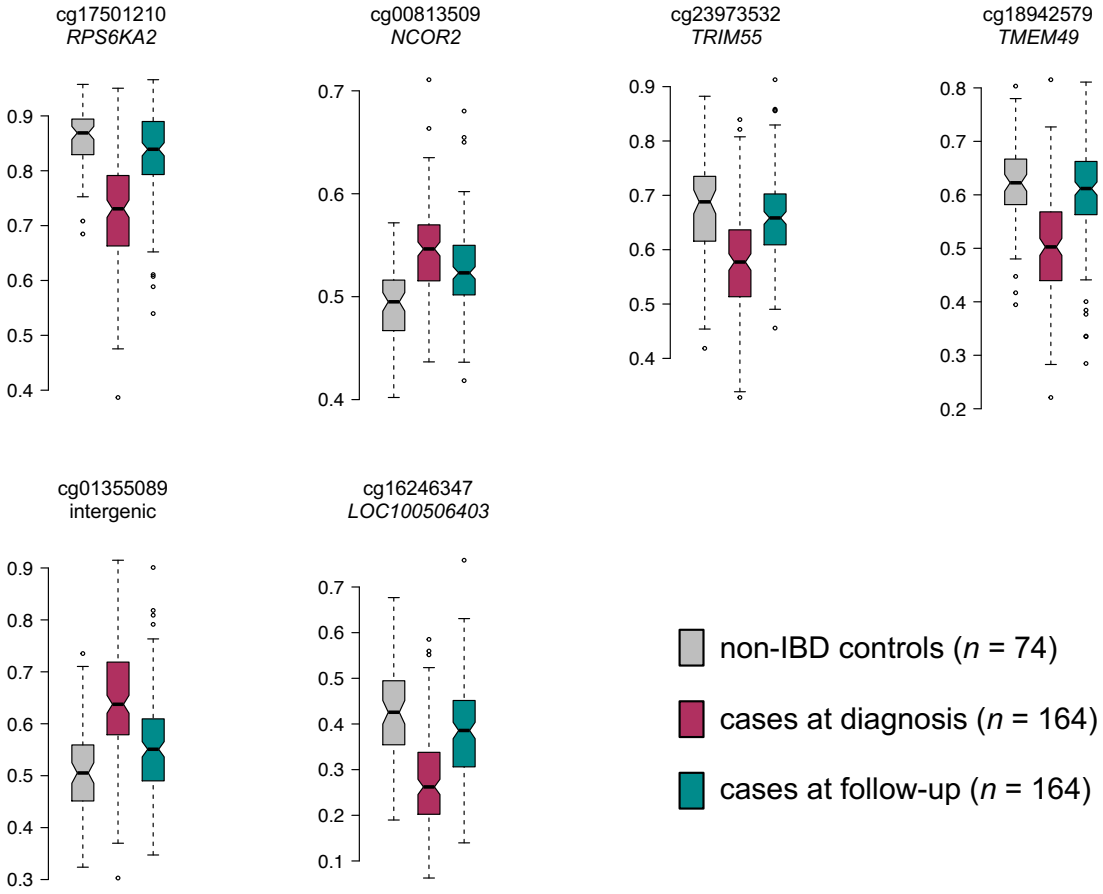
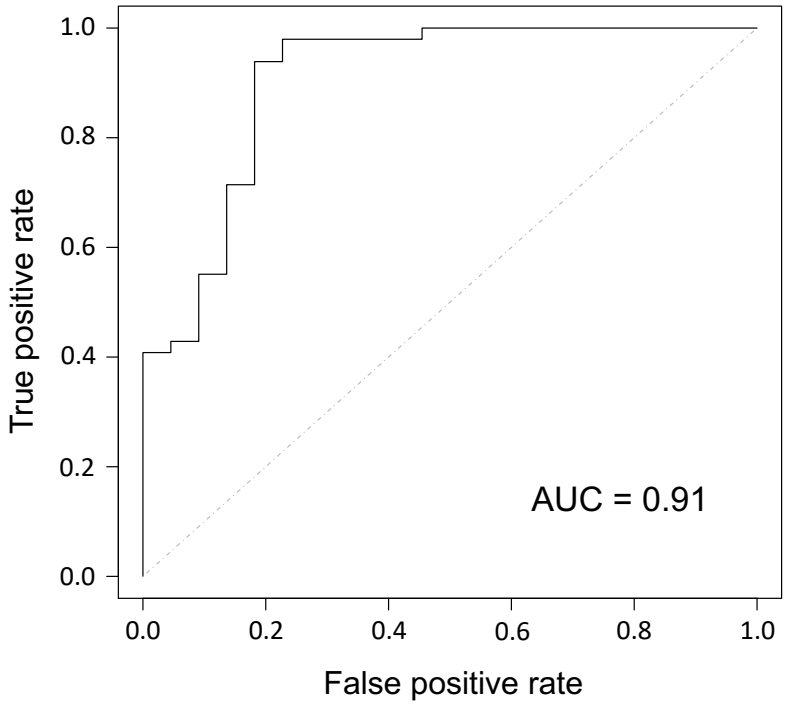


Figure 3-21: Receiver operating characteristic (ROC) curve of baseline peripheral blood methylation data was plotted to differentiate Crohn's disease patients from controls. The area under the curve (AUC) is indicated. A perfect classifier would have an AUC of 1, and a random classifier would score 0.5.



Supplementary Tables

Supplementary Tables 3-1 to 3-15 can be accessed at the following link
https://www.sciencedirect.com/science/article/pii/S001650851930397X

REFERENCES

1. Van Limbergen, J. *et al.* Definition of phenotypic characteristics of childhood-onset inflammatory bowel disease. *Gastroenterology* **135**, 1114-22 (2008).
2. Ruel, J., Ruane, D., Mehandru, S., Gower-Rousseau, C. & Colombel, J.F. IBD across the age spectrum: is it the same disease? *Nat Rev Gastroenterol Hepatol* **11**, 88-98 (2014).
3. Ventham, N.T. *et al.* Integrative epigenome-wide analysis demonstrates that DNA methylation may mediate genetic risk in inflammatory bowel disease. *Nat Commun* **7**, 13507 (2016).
4. Karatzas, P.S., Gazouli, M., Safioleas, M. & Mantzaris, G.J. DNA methylation changes in inflammatory bowel disease. *Ann Gastroenterol* **27**, 125-132 (2014).
5. McDermott, E. *et al.* DNA Methylation Profiling in Inflammatory Bowel Disease Provides New Insights into Disease Pathogenesis. *J Crohns Colitis* **10**, 77-86 (2016).
6. Li Yim, A.Y.F. *et al.* Peripheral blood methylation profiling of female Crohn's disease patients. *Clin Epigenetics* **8**, 65 (2016).
7. Nimmo, E.R. *et al.* Genome-wide methylation profiling in Crohn's disease identifies altered epigenetic regulation of key host defense mechanisms including the Th17 pathway. *Inflamm Bowel Dis* **18**, 889-99 (2012).
8. Adams, A.T. *et al.* Two-stage genome-wide methylation profiling in childhood-onset Crohn's Disease implicates epigenetic alterations at the VMP1/MIR21 and HLA loci. *Inflamm Bowel Dis* **20**, 1784-93 (2014).
9. Harris, R.A. *et al.* DNA methylation-associated colonic mucosal immune and defense responses in treatment-naive pediatric ulcerative colitis. *Epigenetics* **9**, 1131-7 (2014).
10. Harris, R.A. *et al.* Genome-wide peripheral blood leukocyte DNA methylation microarrays identified a single association with inflammatory bowel diseases. *Inflamm Bowel Dis* **18**, 2334-41 (2012).
11. Taman, H. *et al.* Genome-wide DNA Methylation in Treatment-naive Ulcerative Colitis. *J Crohns Colitis* **12**, 1338-1347 (2018).
12. Howell, K.J. *et al.* DNA Methylation and Transcription Patterns in Intestinal Epithelial Cells From Pediatric Patients With Inflammatory Bowel Diseases Differentiate Disease Subtypes and Associate With Outcome. *Gastroenterology* **154**, 585-598 (2018).
13. Kugathasan, S. *et al.* Prediction of complicated disease course for children newly diagnosed with Crohn's disease: a multicentre inception cohort study. *Lancet* **389**, 1710-1718 (2017).
14. Farh, K.K. *et al.* Genetic and epigenetic fine mapping of causal autoimmune disease variants. *Nature* **518**, 337-43 (2015).
15. Barfield, R.T., Kilaru, V., Smith, A.K. & Conneely, K.N. CpGassoc: an R function for analysis of DNA methylation microarray data. *Bioinformatics* **28**, 1280-1 (2012).
16. McCartney, D.L. *et al.* Identification of polymorphic and off-target probe binding sites on the Illumina Infinium MethylationEPIC BeadChip. *Genom Data* **9**, 22-4 (2016).
17. Teschendorff, A.E. *et al.* A beta-mixture quantile normalization method for correcting probe design bias in Illumina Infinium 450 k DNA methylation data. *Bioinformatics* **29**, 189-96 (2013).
18. Houseman, E.A. *et al.* DNA methylation arrays as surrogate measures of cell mixture distribution. *BMC Bioinformatics* **13**, 86 (2012).
19. Wang, J., Zhao, Q., Hastie, T. & Owen, A.B. Confounder adjustment in multiple hypothesis testing. *arXiv* (2015).

20. Johnson, W.E., Li, C. & Rabinovic, A. Adjusting batch effects in microarray expression data using empirical Bayes methods. *Biostatistics* **8**, 118-27 (2007).
21. Phipson, B., Maksimovic, J. & Oshlack, A. missMethyl: an R package for analyzing data from Illumina's HumanMethylation450 platform. *Bioinformatics* **32**, 286-8 (2016).
22. Purcell, S. *et al.* PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am J Hum Genet* **81**, 559-75 (2007).
23. Liu, J.Z. *et al.* Association analyses identify 38 susceptibility loci for inflammatory bowel disease and highlight shared genetic risk across populations. *Nat Genet* **47**, 979-986 (2015).
24. Zhou, X. & Stephens, M. Genome-wide efficient mixed-model analysis for association studies. *Nat Genet* **44**, 821-4 (2012).
25. Wahl, S. *et al.* Epigenome-wide association study of body mass index, and the adverse outcomes of adiposity. *Nature* **541**, 81-86 (2017).
26. Friedman, J., Hastie, T. & Tibshirani, R. Regularization Paths for Generalized Linear Models via Coordinate Descent. *J Stat Softw* **33**, 1-22 (2010).
27. Mo, A. *et al.* Disease-specific regulation of gene expression in a comparative analysis of juvenile idiopathic arthritis and inflammatory bowel disease. *Genome Med* **10**, 48 (2018).
28. Ligthart, S. *et al.* DNA methylation signatures of chronic low-grade inflammation are associated with complex diseases. *Genome Biol* **17**, 255 (2016).
29. Bogaert, S. *et al.* Differential mucosal expression of Th17-related genes between the inflamed colon and ileum of patients with inflammatory bowel disease. *BMC Immunol* **11**, 61 (2010).
30. Granlund, A. *et al.* Whole genome gene expression meta-analysis of inflammatory bowel disease colon mucosa demonstrates lack of major differences between Crohn's disease and ulcerative colitis. *PLoS One* **8**, e56818 (2013).
31. Sipos, F. *et al.* Peripheral blood based discrimination of ulcerative colitis and Crohn's disease from non-IBD colitis by genome-wide gene expression profiling. *Dis Markers* **30**, 1-17 (2011).
32. Kang, J. *et al.* Improved risk prediction for Crohn's disease with a multi-locus approach. *Hum Mol Genet* **20**, 2435-42 (2011).
33. de Lange, K.M. *et al.* Genome-wide association study implicates immune activation of multiple integrin genes in inflammatory bowel disease. *Nat Genet* **49**, 256-261 (2017).
34. Maddur, M.S., Miossec, P., Kaveri, S.V. & Bayry, J. Th17 cells: biology, pathogenesis of autoimmune and inflammatory diseases, and therapeutic strategies. *Am J Pathol* **181**, 8-18 (2012).
35. Hyams, J. *et al.* Evaluation of the pediatric crohn disease activity index: a prospective multicenter experience. *J Pediatr Gastroenterol Nutr* **41**, 416-21 (2005).
36. Jostins, L. *et al.* Host-microbe interactions have shaped the genetic architecture of inflammatory bowel disease. *Nature* **491**, 119-24 (2012).
37. Marigorta, U.M. *et al.* Transcriptional risk scores link GWAS to eQTLs and predict complications in Crohn's disease. *Nat Genet* **49**, 1517-1521 (2017).
38. Hannon, E., Weedon, M., Bray, N., O'Donovan, M. & Mill, J. Pleiotropic Effects of Trait-Associated Genetic Variation on DNA Methylation: Utility for Refining GWAS Loci. *Am J Hum Genet* **100**, 954-959 (2017).
39. Wu, Y. *et al.* Integrative analysis of omics summary data reveals putative mechanisms underlying complex traits. *Nat Commun* **9**, 918 (2018).
40. Mendelson, M.M. *et al.* Association of Body Mass Index with DNA Methylation and Gene Expression in Blood Cells and Relations to Cardiometabolic Disease: A Mendelian Randomization Approach. *PLoS Med* **14**, e1002215 (2017).
41. Elliott, H.R. *et al.* Role of DNA Methylation in Type 2 Diabetes Etiology: Using Genotype as a Causal Anchor. *Diabetes* **66**, 1713-1722 (2017).
42. Scharl, S. *et al.* Malignancies in Inflammatory Bowel Disease: Frequency, Incidence and Risk Factors-Results from the Swiss IBD Cohort Study. *Am J Gastroenterol* (2018).
43. Choi, C.R., Bakir, I.A., Hart, A.L. & Graham, T.A. Clonal evolution of colorectal cancer in IBD. *Nat Rev Gastroenterol Hepatol* **14**, 218-229 (2017).

44. Peneau, A. *et al.* Mortality and cancer in pediatric-onset inflammatory bowel disease: a population-based study. *Am J Gastroenterol* **108**, 1647-53 (2013).
45. Aardoom, M.A., Linda Joosse, M.E., de Vries, A.C.H., Levine, A. & de Ridder, L. Malignancy and Mortality in Pediatric-onset Inflammatory Bowel Disease: A Systematic Review. *Inflamm Bowel Dis* **24**, 732-741 (2018).

Chapter 4

The Microbiome in Patients with Inflammatory Diseases

This chapter has been adapted and was originally published in *Clinical Gastroenterology and Hepatology*; January 2019 Volume 17, Issue 2, Pages 243–255

Hari K. Somineni and Subra Kugathasan

ABSTRACT

Microbial dysbiosis characterized by alterations in the structure and function of the gut microbiome has long been implicated in the pathogenesis of inflammatory bowel diseases. To date, most human inflammatory bowel disease microbiome studies are focused on microbial composition rather than function; however, with the latest technical advancements complemented by the rapidly dropping costs, studies focusing on the functional aspects of microbial dysbiosis are on the rise. Several compelling and complimentary pieces of evidence support the notion that the gut microbiome and their metabolites play an important role in the development of inflammatory bowel disease. Data from preclinical studies overwhelmingly support the notion that changes in the gut microbiome causally underlie inflammatory bowel disease pathogenesis. Hence there is considerable interest in modulating the state and function of the gut microbiome to achieve therapeutic benefits. While the causal potential of the gut microbiome remains an active area of current research in the clinical setting, accumulating correlative evidence support the view that microbial dysbiosis parallels increased incidence of inflammatory bowel disease. In this review, we intend to provide a brief overview of the current human inflammatory bowel disease microbiome findings, describe the cause-effect relationships between the gut microbiome and inflammatory bowel disease, and discuss the possibility of using microbiome-based approaches in the diagnosis, therapy, and management of disease. In addition, the potential role of microbiome-based interventions in the treatment of human inflammatory bowel disease is also discussed.

INTRODUCTION

The interplay between the immune system and the gut microbiome is well established in the pathogenesis of immune-mediated inflammatory diseases such as Crohn's disease and ulcerative colitis, the two classical forms of inflammatory bowel disease. Inflammatory bowel disease is a group of complex, multifactorial disorders characterized by chronic relapsing inflammation in the gut. Approximately 5 million people across the globe are affected by these diseases, hinting at the potential emergence of inflammatory bowel disease as a worldwide epidemic[1-3]. Industrialized nations in North America and Europe experienced rapid

increases in the incidence of inflammatory bowel disease during the second half of the 20th century[4]. Although inflammatory bowel disease incidence rates seem plateaued in some developed countries, it is on the rise in developing nations as well as in some parts of the developed nations[5-8].

While the exact causal mechanisms underlying inflammatory bowel disease are yet to be understood, it is thought to arise in the context of dysregulated immune response to commensal gut bacteria in subjects who are genetically predisposed[9]. There is considerable genetic evidence indicating that the impaired recognition and processing of bacteria contributes to inflammatory bowel disease pathology. Many of the so far identified inflammatory bowel disease susceptibility variants and their putative genes regulate host-microbial interactions[9, 10]. For instance, *NOD2* is an intracellular bacterial sensor and its loss of function mutations are associated with the development of Crohn's disease[11-13]. Similarly, missense mutations in Crohn's disease-associated *ATG16L1*, a critical autophagic effector, impairs autophagic function associated with defects in bacterial killing[14-17]. Mutations in another such autophagy-related gene, *IRGM,* were also found to be associated with inflammatory bowel disease[18]. Further, cross-talk between these autophagy proteins and NOD2 impairs antimicrobial peptide-secretory function of the intestinal Paneth cells in response to bacteria, thereby leading to the accumulation of ileal bacteria[19]. In addition to these genetic data, there is increasing epidemiological and environmental evidence implicating gut microbiome in the etiology of inflammatory bowel disease. Association of increasing incidence of inflammatory bowel disease with urbanization/industrialization in the developing world which is accompanied by changes in diet and environmental exposures has tempted to postulate an inciting role for the microbiota in inflammatory bowel disease pathogenesis. Rapid urbanization in the developing world has been shown to attenuate gut microbial diversity[20-23].

Similarly, several different surrogates of urbanization, including the westernized diet which is generally characterized by low intake of fiber and vegetables and high intake of saturated fats, red meat and carbohydrates, overuse of antibiotics, pollution, and improved hygiene status, have been shown to be associated with inflammatory bowel disease incidence and altered intestinal microbial composition and

function[20, 24, 25] (**Fig. 4-1**). Taken together, all these evidence has led to the hypothesis that intestinal microbiota may at least in part modulate the pathological effects of urbanization and genetic variation in inflammatory bowel disease pathogenesis. However, the mechanistic insights linking microbial alterations to inflammatory bowel disease pathology remains elusive.

What is currently known from inflammatory bowel disease microbiome research in humans.

To date, many human inflammatory bowel disease microbiome studies have focused on characterizing the microbial compositional differences associated with inflammatory bowel disease, primarily by sequencing the 16S ribosomal RNA (rRNA) gene, and has demonstrated all inflammatory bowel disease subtypes to be associated with overall microbial diversity, a shift in balance between commensal and potentially pathogenic microorganisms, and the relative abundance of specific bacterial taxa (**Fig. 4-1**). Reduced microbial diversity and richness has long been recognized as a hallmark in inflammatory bowel disease pathogenesis[26, 27]. Inflammatory bowel disease, in general, has been found to be associated with a shift in balance between the protective and aggressive resident bacteria: patients typically exhibit a depletion of bacteria with anti-inflammatory effects, including *Faecalibacterium prausnitzii* and an expansion of pro-inflammatory bacterial species such as *Escherichia coli* and *Clostridium difficile*[28-30]. Strikingly, some of these bacterial alterations that are characteristic of inflammatory bowel disease, were also noted in the unaffected family members of patients with inflammatory bowel disease who are likely to share genetic and environmental features, and consequently have a higher risk of developing inflammatory bowel disease compared to general population[31, 32]. For instance, depletion of *Faecalibacterium prausnitzii* with anti-inflammatory properties was seen in both affected and unaffected relatives with inflammatory bowel disease compared to healthy controls with no affected family members[31]. Attempts to identify single bacterium associations of inflammatory bowel disease have implicated numerous taxa that are differentially abundant in patients with inflammatory bowel disease and healthy controls, but, inconsistency among these observations and lack of repeated measures of microbiome data from prospective longitudinal cohorts thwarted the efforts to investigate the definitive causal roles of specific bacterial groups in the

etiopathogenesis of inflammatory bowel disease. Nevertheless, members such as adherent-invasive *Escherichia coli*, *Clostridium difficile* and *Fusobacterium nucleatum* are understood to be potentially pathogenic to inflammatory bowel disease[29, 30, 33, 34].

Similarly, nonbacterial microbial dysbiosis has also been found to be associated with inflammatory bowel disease. For instance, certain viruses, including bacteriophages are more abundant in fecal and mucosal samples of inflammatory bowel disease patients compared to healthy controls[35]. Increased virome richness, especially the expansion of Caudovirales bacteriophages has been noted in multiple cohorts of inflammatory bowel disease[35, 36]. Other nonbacterial microbiome, including fungi and archaea were also found to exhibit compositional differences in inflammatory bowel disease and healthy controls[37-40]. For example, a study of mucosa-associated fungal composition demonstrated an overall increase in mycobial diversity in inflammatory bowel disease, while the bacterial diversity was significantly reduced in the same samples[38]. Similar data from stool was reported by Sokol *et al.*, suggesting that inflammatory bowel disease microbiome may favor expansion of fungi at the expense of bacterial community structure[37]. At the individual mycobial member level, expansion of *Candida albicans* and depletion of *Saccharomyces cerevisiae* was noted in fecal samples from patients with inflammatory bowel disease compared to healthy controls[37]. However, investigations aimed towards identifying fungal species that play a direct role in inflammatory bowel disease pathogenesis have yielded mixed results making the findings less rigorous to pursue further. Overall, data from inflammatory bowel disease-related nonbacterial microbiome remain scarce when compared to bacterial dysbiosis, and hence their definitive role in disease pathology and clinical utility remains to be determined.

On the other hand, functional investigations of microbial dysbiosis aimed at understanding how disturbances in microbial composition might reflect functional processes (metabolites) that are perturbed during the pathogenesis of inflammatory bowel disease, have provided valuable insights into the pathophysiology of inflammatory bowel disease. For instance, patients with inflammatory bowel disease exhibit low abundance of butyrate-producing bacteria and high abundance of sulfate-reducing bacteria,

resulting in lower butyrate production and higher levels of hydrogen sulfide (H_2S), respectively (**Fig. 4-1**). Butyrate production has been proposed to exert beneficial effects in the gastrointestinal health of inflammatory bowel disease, while H_2S has been thought to have adverse effects[41] (discussed below in detail). Additionally, elevated intestinal oxygen levels resulting in the breakage of homeostasis between obligate anaerobes and facultative anaerobes has also been proposed as a potential mechanism connecting inflammatory bowel disease-associated microbial dysbiosis to disease pathogenesis[42]. Collectively, despite the compositional evidence demonstrating an association between microbial dysbiosis and inflammatory bowel disease, and functional analyses suggesting potential mechanisms through which dysbiosis may contribute to inflammatory bowel disease pathogenesis, the precise causative role of microbial dysbiosis in the development of inflammatory bowel disease is yet to be defined.

Causal potential of the gut microbiome in human inflammatory bowel disease

Unlike the overwhelming evidence from animal models of inflammatory bowel disease supporting the causal potential of the gut microbiome, evidence from human studies remains scarce. Nonetheless, there were reports of clinical observations that suggest an inciting role for the gut microbiota in disease pathogenesis. For instance, the clinical observation that at least a proportion of patients with inflammatory bowel disease respond to antibiotics, attests to a potentially causal role of the microbiome in inflammatory bowel disease[32, 43-46]. Localization of inflammation to anatomical regions (terminal ileum and rectum) involved in the storage of feces at a standstill, supported by the success of fecal diversion therapy in managing Crohn's disease, is consistent with the notion that gut microbiota may contribute to inflammatory effects seen in the gut[47-49]. Further evidence supporting an essential role for the gut microbiome in inflammatory bowel disease comes from an observation suggesting that the postoperative recurrence of Crohn's disease is triggered by fecal contents[50]. Inflammatory remission and mucosal healing is achieved in the excluded intestinal segment of patients with Crohn's disease following a fecal stream diversion, which showed relapse after reinfusion with the intestinal contents[50]. Collectively, despite these evidence being consistent with a potentially causal influence, the precise role of microbial dysbiosis in inflammatory

bowel disease pathogenesis, including the identification of a candidate set of specific bacterial groups that cause inflammatory bowel disease or contribute to the causal underpinnings of inflammatory bowel disease remain elusive. Efforts towards delineating the causal versus consequential roles of the microbiome in inflammatory bowel disease pathogenesis are thwarted due mostly to reasons such as: (i) the cross-sectional nature of most inflammatory bowel disease microbiome studies; (ii) usage of 16S rRNA gene sequencing rather than shotgun metagenomics with deep sequencing to infer species- and strain-level taxonomic classifications; (iii) difficulties in culturing candidate microbial organisms to investigate the effects of commensal organisms and study the pathogenesis of human infectious diseases; and to an extent (iv) to the usage of fecal samples as a source for profiling disease-specific microbial changes as fecal bacterial community structures do not fully reflect mucosally associated bacterial profiles; yet the majority of the inflammatory bowel disease microbiome studies to date has relied on fecal samples. Nonetheless, pleiotropic associations of inflammatory bowel disease and microbiome, with a number of factors, including the host-genetics, diet, industrialization, antibiotic use, and social status, suggests that the relationship between the microbiome and inflammatory bowel disease is much more complex and dynamic rather than a simple cause-effect. Despite the underlying complexity in delineating the directionality of these interactions, there is considerable evidence to support the notion that microbial dysbiosis parallels increased incidence of inflammatory bowel disease, providing a strong rationale to exploit the gut microbiome for diagnostic as well as therapeutic benefits of inflammatory bowel disease.

Dysbiosis in diagnosing inflammatory bowel disease

Harnessing the associations between the state and function of the gut microbiota and inflammatory bowel disease phenotypes, great strides have been made in the past decade to leverage gut microbial dysbiosis for diagnostic purposes. In our previous investigations of the RISK cohort[51], by studying the treatment-naïve microbiome in the largest pediatric Crohn's disease inception cohort to date, we introduced the concept of microbial dysbiosis index derived from both the intestinal mucosal biopsies and stool[26]. Strongly correlated with disease severity, this microbiome-based index distinguished patients from healthy controls with great

precision; the best performance was obtained by the intestinal mucosal samples, which was closely followed by stool samples[26]. We have further shown that these disease-associated microbial shifts occur during the early stages of Crohn's disease which may provide the clinical benefits of early diagnosis[26]. In a different study, we showed that microbial dysbiosis index computed from pre-treatment stool samples, besides distinguishing inflammatory bowel disease patients from controls, can differentiate Crohn's disease from ulcerative colitis, as well as therapy responders from non-responders[52]. Interestingly, by defining a 'healthy' gut microbial plane, Halfvarson *et al*. demonstrated that patients of all inflammatory bowel disease subtypes have a distinct disease profile which deviates substantially from the healthy plane, with the chronic inflammation or bowel resection cases showing the greatest departure[53]. Notably, departure distance from the healthy plane diagnosed disease better than fecal calprotectin, the most commonly used non-invasive diagnostic approach.

Further, disease-specific microbial signatures confer the potential to inform treatment choices and subsequent clinical outcomes. For example, disease-associated gut microbial structures at the time of diagnosis has been found to predict 6 month steroid-free remission rates[54]. Similarly, Ananthakrishnan *et al*. has reported that the baseline functional analyses of Crohn's disease-associated gut microbiome which showed higher abundance of butyrate-producers and enrichment of 13 microbial pathways can predict response to anti-integrin therapy[55]. In both these studies, incorporation of the microbiome data into respective predictive models based on clinical measures, has resulted in improved predictive ability[54, 55]. In fact, the diagnostic potential of microbiome in inflammatory bowel disease has reached far beyond the gut. In our recent study (Somineni *et al*., unpublished data - in submission) we investigated if oral samples are indicative of inflammatory bowel disease, and if so, which type of the oral sample, including saliva, plaque, tongue or buccal mucosa is the most informative, by comparing their diagnostic potential to stool samples obtained from the same individuals. Strikingly we noticed for the first time that saliva samples are more informative to use in diagnosing inflammatory bowel disease, closely matched by stool and other oral sites. This appears to be because inflammatory bowel disease-associated dysbiosis exist in a site- and taxa-

specific manner. Inspired from these rapidly expanding diagnostic evidence of the microbiome, we are currently in the process of developing the notion of "Microbiome Risk Score" to refine the potential of inflammatory bowel disease-associated microbial profiles to distinguish inflammatory bowel disease patients from healthy controls, differentiate Crohn's disease from ulcerative colitis, and to predict subsequent disease severity, response to therapy, and risk of complications which may constitute a potential strategy for personalized medicine.

Targeting of dysbiosis for therapy

Despite limited clinical evidence demonstrating the causal role of microbial dysbiosis in the development of inflammatory bowel disease, there has been considerable therapeutic interest to support the expansion of a healthy microbiota by various means – microbiome-based interventions, including probiotics, prebiotics and synbiotics.

Probiotics

Probiotics are living microorganisms that confer health benefit when consumed in adequate amounts. Administering live bacterial strains, often the good bacteria, has long been viewed as a safer and more sustainable therapeutic approach for inflammatory bowel disease. Probiotics are thought to exert beneficial effects in inflammatory bowel disease via several modes of action, including (i) the modification of the disease-associated intestinal microbial composition thereby relieving intestinal dysbiosis; (ii) regulation of the metabolic activity of the intestinal microbiota; (iii) suppression of pro-inflammatory processes; and (iv) immunomodulation.

Several bacterial strains were thus far examined for their therapeutic efficacy in inducing or maintaining remission in patients with inflammatory bowel disease, ulcerative colitis in particular. Single strain probiotic preparations such as *E coli* Nissle 1917[56, 57] and bifidobacterium-fermented milk[58-60] were shown to be as effective as conventional medications such as Mesalazine in the maintenance of ulcerative colitis remission. Unlike single strain probiotic preparations, a blend of multiple beneficial bacterial strains was

also tested for their clinical effect in the management of ulcerative colitis. For instance, VSL#3 is a cocktail of eight different strains of live bacteria consisting of 4 strains of *Lactobacilli* (*L paracasei*, *L plantarum*, *L acidophilus*, and *L delbrueckii*), 3 strains of *Bifidobacteria* (*B longum*, *B breve*, and *B infantis*) and a strain of *Streptococcus* (*Streptococcus thermophilus*)[61]. To date, the most consistent and reproducible evidence supporting the favorable action of probiotics in inflammatory bowel disease and gut inflammation comes from the usage of VSL#3 preparations. Administration of VSL#3 twice daily for 12 weeks showed a significant decrease in the Ulcerative Colitis Disease Activity Index and improvement in disease symptoms (rectal bleeding and stool frequency) at weeks 6 and 12 compared with the placebo group in a multicenter, randomized, double-blind study[62]. Another randomized, placebo-controlled trial in children with ulcerative colitis reported effectiveness of VSL#3 (dosed based on patient's body-weight) in the maintenance of ulcerative colitis remission[63]. In addition, evidence from an uncontrolled pilot study and two other open-label studies, also support the notion that VSL#3 preparations are effective in maintaining remission in ulcerative colitis[64-66]. Notably, many studies have challenged the notion of using probiotics in the management of ulcerative colitis. For example, daily supplements of a blend of the bacterial strains *Lactobacillus acidophilus* La-5 and *Bifidobacterium animalis* subsp. *lactis* BB-12 (Probio-Tec AB-25) for 52 weeks demonstrated no significant difference in maintaining remission of ulcerative colitis in a randomized (2:1) double-blind placebo-controlled trial consisting of 32 ulcerative colitis patients[67]. Similarly, a Cochrane review of meta-analyzing some of the prevailing inflammatory bowel disease probiotic studies reported that, in comparison to placebo preparations, probiotics offer no therapeutic value in induction or maintenance of remission in ulcerative colitis, and hence their use cannot be recommended based on the existing evidence[68 69].

On the other hand, the efficacy of probiotics in Crohn's disease has not very well been documented. Probiotic Crohn's disease studies are relatively limited, smaller in sample sizes, and the clinical data that are currently available are not as favorable as compared to ulcerative colitis probiotic studies. Nevertheless, although weak, there is evidence from either randomized or open label pilot studies suggesting the efficacy

of *E coli* Nissle[70], VSL#3[71], and Saccharomyces boulardii[72] in maintaining Crohn's disease remission. In a small open-label pilot study, Gupta *et al*.,[73] reported that administration of *Lactobacillus rhamnosus* GG for 6 months could ameliorate clinical activity in children with mild to moderate Crohn's disease, however, favorable actions of this strain was subsequently challenged when Schultz *et al*.,[74] found no difference in clinical remission rates of Crohn's disease in adult patients after taking the probiotic preparations for 6 months. In agreement, *Lactobacillus rhamnosus* GG given for an year was found ineffective in preventing the rate of Crohn's disease recurrence after surgery in a randomized placebo-controlled study[75]. Similarly, *Lactobacillus johnsonii* LA1 has been reported to be ineffective[76]. Results from this 6-month randomized, double blind trial showed that 49% of Crohn's disease patients experienced recurrence compared to 64% in the placebo group[76]. Further, a recent, fairly large multicenter randomized, placebo-controlled trial evaluating VSL#3 for the prevention of Crohn's disease recurrence did not show much difference when compared to the placebo group[71]. Overall, not so clinically favorable outcomes from the thus far evaluated set of probiotic strains for the prevention of Crohn's disease recurrence, suggest the notion that the right probiotic strain(s) that may exert beneficial effects in Crohn's disease are yet to be identified. Collectively, despite their success in preclinical studies, probiotics do not seem to be effective in achieving clinical benefits in patients with inflammatory bowel disease, leading to a perception that the claims regarding the clinical benefits of probiotics in inflammatory bowel disease are highly overestimated.

These mixed results between the preclinical and clinical settings could at least in part be attributed to the host-related factors, including age, sex, diet, disease location, severity, familial history of inflammatory bowel disease. In addition, characteristics of the probiotic preparations such as the strain type, concentration, mode of delivery, and colonization potential and survival rates of strains. Factors such as the dose and duration of probiotic administration are also hypothesized to play a prominent role in the success of this, otherwise attractive therapeutic approach with minimal to no adverse effects. Hence, pilot studies of comparative analyses to better understand the strain-specificity, optimize the ideal dose, duration, and mode of delivery of probiotic preparations are of an immediate requirement. Findings from such studies

should subsequently be followed by validation in a larger, well designed prospective trials to definitively clarify the therapeutic efficacy of probiotics in the management of inflammatory bowel disease.

Prebiotics

Prebiotics are usually non-digestible food ingredients that get fermented only when they reach the colon, where they can be used as a feed to selectively stimulate the growth and activity of beneficial microbes in the gut. Prebiotics are even more attractive therapeutics when compared to probiotics, as they are capable of inducing clinically favorable effects by stimulating the growth of indigenous organisms, without having to deal with the administration of live probiotic strains which sometimes confer pathogenic effects. Fructo-oligosaccharides, germinated barley foodstuff, galacto-oligosaccharides, lactulose and resveratrol are some of the prebiotic preparations that have been tested for their clinical efficacy in inflammatory bowel disease. A small open-label study with 10 patients demonstrated that supplementation of 15 grams per day of fructo-oligosaccharides for 3 weeks induce significant reduction in disease activity in moderately active ileocolonic Crohn's disease patients[77]. Fructo-oligosaccharides enrich the growth of intestinal bifidobacterial species which confer immunoregulatory benefits by inducing dendritic cell-mediated IL-10 release. This study has reported a significant expansion of fecal *Bifidobacteria* and mucosal IL-10 positive dendritic cells[77]. On the other hand, a well powered study with 103 Crohn's disease patients receiving 15 grams per day of fructo-oligosaccharides for 4 weeks[78], and another small study with 17 Crohn's disease patients receiving 10 grams of lactulose daily for a period of 4 months[79], demonstrated no clinical improvements when compared to their respective placebo groups. Similarly, there is yet inadequate evidence to support the favorable actions of prebiotic supplements in the induction and maintenance of remission in patients with ulcerative colitis. Casellas *et al*. conducted a randomized, placebo-controlled trial involving 19 patients with active ulcerative colitis where patients were randomly assigned to mesalazine plus Synergy 1 (insulin/oligofructose growth substrate; 12 grams per day) supplement group or mesalazine plus placebo (12 grams per day of maltodextrin) group to investigate the potential of prebiotics in ulcerative colitis maintenance[80]. Synergy 1 group did not show any difference in the patient's clinical scores, including

Rachmilewitz score, dyspepsia-related score and inflammatory bowel disease-related quality of life, despite a significant reduction in fecal calprotectin. However, this was a 2 week study, which is probably too short for the prebiotic to exert clinical benefits. Another study consisting of 121 ulcerative colitis patients evaluated the beneficial effects of oral supplement enriched with fructo-oligosaccharides, fish oil, gum arabic, and antioxidants (vitamin E, vitamin C, selenium) for an extended period of 6 months with a particular focus on disease activity and prescribed medication use[81]. Both supplement and placebo groups demonstrated clinically relevant improvements with no significant differences between the groups; however, oral supplement was associated with a significant reduction in the required dose of prednisone compared to the placebo[81]. Similarly, a range of different prebiotic agents have been tested for their therapeutic potential in ulcerative colitis patients, yet the evidence is inadequate to support their use in the clinical management of ulcerative colitis[82].

Synbiotics

Another attractive therapeutic option pertaining to dysbiosis theory that surfaced is the use of synbiotics. A synbiotic is a synergistic combination of a probiotic and prebiotic. To date, few studies have been conducted using prebiotics in combination with probiotics in inflammatory bowel disease patients, with varying degrees of success. A synbiotic preparation, combination of the probiotic strain *Bifidobacterium longum* with the prebiotic component, Synergy 1 (insulin/oligofructose growth substrate), improved clinical symptoms in patients with active Crohn's disease but not ulcerative colitis, in double blind, randomized controlled trials[59, 83]. In patients with ulcerative colitis (n = 18), consumption of synbiotic twice daily for 4 weeks did not demonstrate a significant reduction in colitis at the macroscopic and microscopic level[59]. However, they found significant increase in mucosal *bifidobacterium* concentration, and a decrease in mucosal pro-inflammatory cytokine levels, including *IL-1a* and *TNF-a*[59]. Treating patients with active Crohn's disease (n = 35) with the same synbiotic for an extended period of 6 months, showed significant improvements in clinical outcomes demonstrated by reduced Crohn's disease activity index and histological scores[83]. Synbiotic patients had increased mucosal *bifidobacteria*; however, no significant differences were

noted in the mucosal inflammatory profiles between the synbiotic and the placebo groups[83]. On the other hand, another multi-center, randomized, placebo-controlled synbiotic study with 30 Crohn's disease patients reported no beneficial effects of Synbiotic 2000, a cocktail containing 4 probiotic bacteria and 4 prebiotics, on clinical symptoms and endoscopic scores[84]. This study found no difference in the postoperative recurrence of Crohn's disease between the Synbiotic 2000 and the placebo groups. Notably, by enrolling 120 patients with ulcerative colitis, Fujimori *et al*. conducted a randomized controlled trial to prove the efficacy of synbiotic treatment (*Bifidobacterium longum* plus *psyllium*) compared to probiotic (*Bifidobacterium longum*) or prebiotic (*psyllium*) alone[85]. This study reported that patients who received synbiotic treatment for 4 weeks experienced greater quality-of-life changes evaluated through inflammatory bowel disease-relevant questionnaire compared to the probiotic or prebiotic groups. Fujimori *et al*. unfortunately did not perform endoscopic or histological investigations on this cohort to support the role of these microbiome-based interventions in the clinical management of inflammatory bowel disease.

Overall, the number of clinical studies investigating the efficacy of probiotics, prebiotics and synbiotics in the induction and maintenance of inflammatory bowel disease are limited, and the majority of the existing studies had certain limitations, including, small sample sizes, high dropout rates, poor patient compliance and short duration of treatment regimens. Despite these microbiome-based interventions demonstrating considerable success in animal models, observations from human inflammatory bowel disease studies yielded mixed results which are more biased towards the notion that their use cannot be prescribed based on the prevailing evidence. One major limitation that may have contributed to this disparity could be the lack of mechanistic evidence in clinical studies. While such investigations are adequately performed in animal studies, data regarding human studies are scarce and remains a challenge for the future. Another limitation of these microbiome-based therapeutic preparations is that they are currently considered to be food products, and are thus subject to food regulations, which are less stringent than FDA approved medications. To gain FDA approval as a food product, rigorous clinical evidence attesting for the safety and efficacy of these products are not mandated. This could at least in part be contributing to the

disappointment of microbiome-based interventions in the clinical setting. Another major deficit of current microbiome-based therapeutic studies is the smaller number of patient enrollment, which is mostly due to the lack of research funding. Further large, well-designed clinical trials are required to definitively establish the efficacy of microbiome-based products.

Other microbiome-based therapeutic interventions for the management of inflammatory bowel disease

Do inflammatory bowel disease patients benefit from butyrate replacement?

A prominent short-chain fatty acid metabolite, butyrate, synthesized primarily by the colonic microbiota, is hypothesized to exert therapeutic benefits in inflammatory bowel disease. At the individual microbial member level, inflammatory bowel disease has consistently been found to be associated with a depletion of Firmicutes and an expansion of Proteobacteria. The butyrate producing bacteria in the human gut are members of the phylum Firmicutes, belonging predominantly to clostridial clusters IV (families *Ruminococcaceae*) and XIVa (family *Lachnospiraceae*), the two families that are recognized as strong contributors to microbial dysbiotic signatures in inflammatory bowel disease, and Crohn's disease in particular[28, 86-88]. Although functional analyses characterizing the role of the depletion of these taxa in inflammatory bowel disease pathogenesis are limited, there is enough evidence in the literature documenting lower butyrate content in inflammatory bowel disease cases compared to healthy controls, which suggests a potential link between the compositional disturbances and functional processes relevant to lower butyrate production of inflammatory bowel disease microbiome. Butyrate plays a prominent role in intestinal homeostasis and energy metabolism by possessing immunomodulatory anti-inflammatory properties via inhibition of NF-kB activation and by serving as a rich source of energy for colonocytes (epithelial cells of the colon), respectively. In addition, butyrate has been proposed to promote the integrity of epithelial barrier function by increasing the expression of tight junction proteins, thereby decreasing intestinal epithelial permeability. Taken together, butyrate replenishment may provide a valuable therapeutic option for patients with inflammatory bowel disease.

Several human inflammatory bowel disease microbiome studies have tested the hypothesis that replenishment of butyrate may improve the ability of the host to repair the damaged intestinal epithelium and to regulate inflammation. However, these studies have yielded mixed results, which are due at least in part to the mode of delivery of butyrate. For instance, butyrate administration in the form of enema ameliorated colonic inflammation in patients with ulcerative colitis[89]. Administering oral butyrate in the form of enteric-coated tablets for 8 weeks to patients with mild to moderate ileocolonic Crohn's disease demonstrated effectiveness in inducing clinical remission/improvement by downregulating ileocecal inflammation[90]. Surprisingly, administration of individual or a cocktail of butyrate-producing bacteria in the form of probiotics, enabling *in situ* production of butyrate, yielded results consistent with the idea that their use cannot be recommended based on the available evidence. Taken together, establishing whether butyrate replenishment confer clinical benefits in the management of inflammatory bowel disease remains a challenge for the future.

Do inflammatory bowel disease patients benefit from sulfate-reduction?

Sulfate-reducing bacteria are a group of phylogenetically diverse anaerobic microbes that may represent a keystone member of the microbiome active in inflammatory bowel disease. Several lines of evidence indicate that patients with inflammatory bowel disease exhibit higher counts of sulfate-reducing bacteria in the gut and stool microbiomes, suggesting a potential link between these bacteria and the etiopathogenesis of inflammatory bowel disease[91-94]. Depletion of sulfate-reducing bacteria in patients treated with 5-aminosalicylic acid-based therapy further strengthens this relationship[95].

The mucus layer lining the colonic epithelium plays an essential role in protecting the epithelium by limiting the exposure of epithelial cells to toxins, luminal insults, and microbes. The mucus layer is constituted by highly glycosylated networks of mucins (glycoproteins), which are interconnected by disulfide bonds. Alterations to the structure or composition of the mucus layer results in impaired mucus barrier function, which has long been recognized as a pathogenic hallmark of inflammatory bowel disease[96, 97]. It has been shown that hydrogen sulfide (H_2S), produced mostly by the sulfate-reducing bacteria, has a role in reducing

the disulfide bridges of the mucus layer. Further, H_2S has been found to be associated with DNA damage and alterations in inflammatory cell populations[98]. Thereby, an expansion of sulfate-reducing bacteria, by increasing the concentrations of H_2S, may contribute to the etiopathogenesis of inflammatory bowel disease. Therefore, manipulating the gut microbiome of inflammatory bowel disease to lower sulfate-reducing bacteria may offer an exciting therapeutic option for inflammatory bowel disease. Surprisingly, therapeutic manipulation of sulfate-reducing bacteria, as well as sulfide-reduction in inflammatory bowel disease has not yet been well characterized in human studies.

Do inflammatory bowel disease patients benefit from fecal microbiota transplantation?

Inspired from the compelling finding that the fecal microbiota transplantation (FMT) procedure is effective in treating refractory *Clostridium difficile* infection[99], a gastrointestinal disease thought to arise in the context of intestinal microbial dysbiosis, there has been a great excitement to explore the therapeutic potential of FMT in other dysbiosis-associated diseases that include inflammatory bowel disease. FMT procedure involves infusing the intestinal microbial contents from a healthy donor (in a suspension of stool) to repopulate the diseased gut habitat with a healthy microbiome rather than aiming to restore the diseases-specific gut microbial imbalance. Current results from human inflammatory bowel disease FMT studies remain varied with some studies suggesting its favorable outcomes, while others demonstrate no such benefits. For instance, Paramsothy *et al*. conducted a multicenter, double-blind trial by assigning 85 active ulcerative colitis patients randomly (1:1) to FMT or placebo-groups and reported that 27% of the FMT group achieved the primary endpoint of steroid-free clinical remission with endoscopic improvement compared to 8% in the placebo group[100]. These FMT preparations were derived from multiple donors, and were administered 5 days a week for 8 weeks. Another double-blind, randomized, placebo-controlled trial consisting of 75 active ulcerative colitis patients showed the effectiveness of FMT in inducing remission[101]. This study used a less-intensive regimen of administering 50 mL once weekly for 6 weeks via enema. In contrast, another double-blind, randomized trial consisting of 50 patients with mild to moderately active ulcerative colitis showed that the FMT procedure is ineffective in inducing clinical remission[102]. FMT

infusions were administered twice, once at the beginning of the study and 3 weeks later through nasoduodenal tube. Notably, fecal microbial assessment found that the responders (from both FMT and control groups) were associated with the expansion of Clostridium clusters IV and XIVa[102]. Taken together, discrepancies in the effectiveness of FMT in human inflammatory bowel disease could be attributed at least in part to the heterogeneity of the administered regimens with regards to frequency, duration and mode of administration. On the other hand, FMT data from patients with Crohn's disease is limited and the prevailing findings remain inconclusive[103, 104].

DISCUSSION

Great strides have been made in recent years to understand the complex network of events that underlie the etiopathogenesis of inflammatory bowel disease. While host genetics have been studied extensively, less clearly understood are the contributions of specific environmental determinants, including the gut microbial dysbiosis, to inflammatory bowel disease risk and severity and their interplay with the host genetics. Several compelling and complimentary lines of evidence from animal studies, as well as correlative data from human inflammatory bowel disease microbiome studies collectively point to an essential role of the gut microbiome and their metabolites in the development of inflammatory bowel disease. However, prevailing results from the early therapeutic trials exploring the possibility of microbiome-based interventions in the management of inflammatory bowel disease have been largely disappointing. Probiotic preparations that were tested so far in combination with or without prebiotic supplements had a weak effect, which is mostly transient, in sustaining the remission of both Crohn's disease and ulcerative colitis. No such evidence has so far been documented to induce remission. This setback highlights the need to re-visit the conclusions drawn thus far from the existing microbiome-based studies keeping in mind the limitations of respective studies, and to re-evaluate the therapeutic potential the gut microbiome has to offer.

To effectively translate the current microbiome-based findings into viable therapeutic approaches, it is critical to definitively determine how intestinal dysbiosis, and in particular, which specific bacterial groups play a causal role in conferring inflammatory bowel disease risk. Surprisingly, mechanistic studies

investigating the cause-effect relationships of gut microbiota and their metabolites with inflammatory bowel disease in the clinical setting are lacking. From our recent experience from DNA methylation (another environmental determinant which is increasingly being identified to be associated with inflammatory bowel disease) investigations of Crohn's disease (Somineni *et al.*, *Gastroenterology* 2019), it would not be a surprise to learn that the majority of the disease-associated signatures in the state and function of the gut microbiome are likely a result of the disease rather than exerting causal effects. However, if alterations in state and/or function of a specific microbial taxa plays a causal role in disease development, their identification would provide valuable therapeutic targets.

FUTURE DIRECTIONS

Need for large, well-designed prospective trials. Despite overwhelmingly favorable results from preclinical studies, prevailing efforts directed towards the development of microbiome-focused interventions to achieve therapeutic benefits in humans has been largely disappointing. It is important to note, however, that there are only a limited number of large, well-designed prospective trials that evaluated the therapeutic potential of microbial manipulation in the induction or maintenance of remission in inflammatory bowel disease. Although the interim results have brought about promising implications for future therapeutic strategies, there are numerous challenges that need to be overcome in order to effectively translate the microbiome-based findings into clinic. For instance, intersubject variability is often one of the biggest challenges when dealing with the microbiome-focused studies. The genetic makeup of the enrolled subjects, environment they live in, diet they consume *etc*. should be accounted for in order to minimize the cross-over between the host-specific and disease-relevant effects in evaluating the beneficial actions of microbiome-based interventions. Similarly, issues with the optimal composition, dose, duration, and mode of delivery of microbiome-based preparations that are possibly contributing to the poor outcomes of these microbial therapies should be sorted in order to amplify their favorable clinical response. Lastly, despite numerous studies demonstrating compositional differences in the microbiomes of patients with inflammatory bowel disease and healthy controls, mechanisms connecting dysbiosis to inflammatory bowel

disease remain largely unknown. Future studies focusing more on the functional aspects of microbial dysbiosis using metagenomic characterization, and integrating these data further with metatranscriptomics and metaproteomics from the same subjects may aid in closing these gaps.

Mendelian randomization to identify causal associations. Next, it should be of utmost importance to establish whether the microbiome plays a causal role in humans and can be leveraged for therapeutic benefits before we continue with a surge of microbiome-based drug developmental processes. Delineating disease-associated microbial changes that are capable of exerting causal effects from signatures that are merely a symptom of disease, is essential to leverage the modulation of such signatures for therapeutic benefits. It is being increasingly recognized that the gut microbiome related to inflammatory bowel disease is tightly regulated by the host genetics among many other things[9, 13, 105-113]. This intimate relationship between the host-genetic variation and the microbiome could be harnessed with the latest analytical advancements such as Mendelian randomization for example, to identify strains of bacteria that causally underlie inflammatory bowel disease pathogenesis. Genetic variants associated with the abundance of bacterial species or their metabolites could be used as instrumental variables to delineate the causal versus consequential roles of such microbiome-centric findings.

Role of the gut microbiome in disease course. The role of the gut microbiome in disease progression of human inflammatory bowel disease remains an untapped horizon. Evidence from studies of the largest pediatric inflammatory bowel disease cohorts, RISK and PROTECT, implicate a potential role for the gut microbiota in disease progression. The RISK study findings demonstrate that inflammatory B1 Crohn's disease patients who subsequently progress to B2 stricturing behavior exhibit a distinct mucosal microbial signature at baseline, which is different from patients that remain complication-free during the course of the 3 year follow-up period[51]. Similarly, baseline microbial profiles showed promise in predicting eventual disease severity or remission within the first 6 months after diagnosis[26]. In particular, depletion of Enterobacteriaceae and expansion of Fusobacterium and Haemophilus were found to have substantial effects in predicting subsequent severe disease[26]. On the other hand, latest findings from the PROTECT

ulcerative colitis inception study (Schirmer *et al.*, Cell Host & Microbe - In press) suggest a promising link between microbial dysbiosis and differential disease severity and subsequent need for colectomy. For instance, depletion of 43 bacteria taxa and expansion of 7 taxa in the baseline gut microbiota of new-onset, treatment naïve ulcerative colitis patients was associated with initial disease severity, and showed a continuous depletion or expansion with worsening disease during the course of the 1 year follow-up period (**Figs. 4-2, 4-3**). Notably, all 7 taxa with increased abundance in more severe disease represent microbes that are typically found to reside in the oral cavity, including *Veillonella parvula*, *Veillonella dispar*, *Aggregatibacter segnis* and *Haemophilus parainfluenzae* (**Fig. 4-2**). In keeping with a potential role for the gut microbiota in disease course, 15 of the 50 microbes that showed associations with initial disease severity, were also found to be associated with eventual colectomy within the first year after diagnosis. Several of these microbes that are indicative of patients who are at risk for medically refractory disease and consequently need a colectomy demonstrated a pronounced decrease in microbial stability over time (**Fig. 4-3**). Despite these evidence being correlative, it raises the possibility that the gut microbiome may play a role in differential clinical course of inflammatory bowel disease. However, additional studies are needed to establish a causal link between the gut microbial dysbiosis and disease course, and subsequently explore the possibility of manipulating the microbiome to achieve therapeutic benefits in modulating the disease course.

Figure 4-1: Major factors underlying the inflammatory bowel disease-gut microbiome associations. Genetic and environmental factors that were hypothesized to contribute to inflammatory bowel disease and key microbial perturbations that ultimately lead to dysbiosis characteristic of inflammatory bowel disease. Many factors known to be associated with inflammatory bowel disease and inflammatory bowel disease-subphenotypes were also found to shift the gut microbiota from the state of normobiosis to dysbiosis, thereby challenging the interpretation of causal versus consequential roles of the microbiome in inflammatory bowel disease. Dotted lines represent known associations, arrows indicate the direction of associations. Continuous lines with arrows correspond to depletion or expansion of microbial features in the state of dysbiosis.

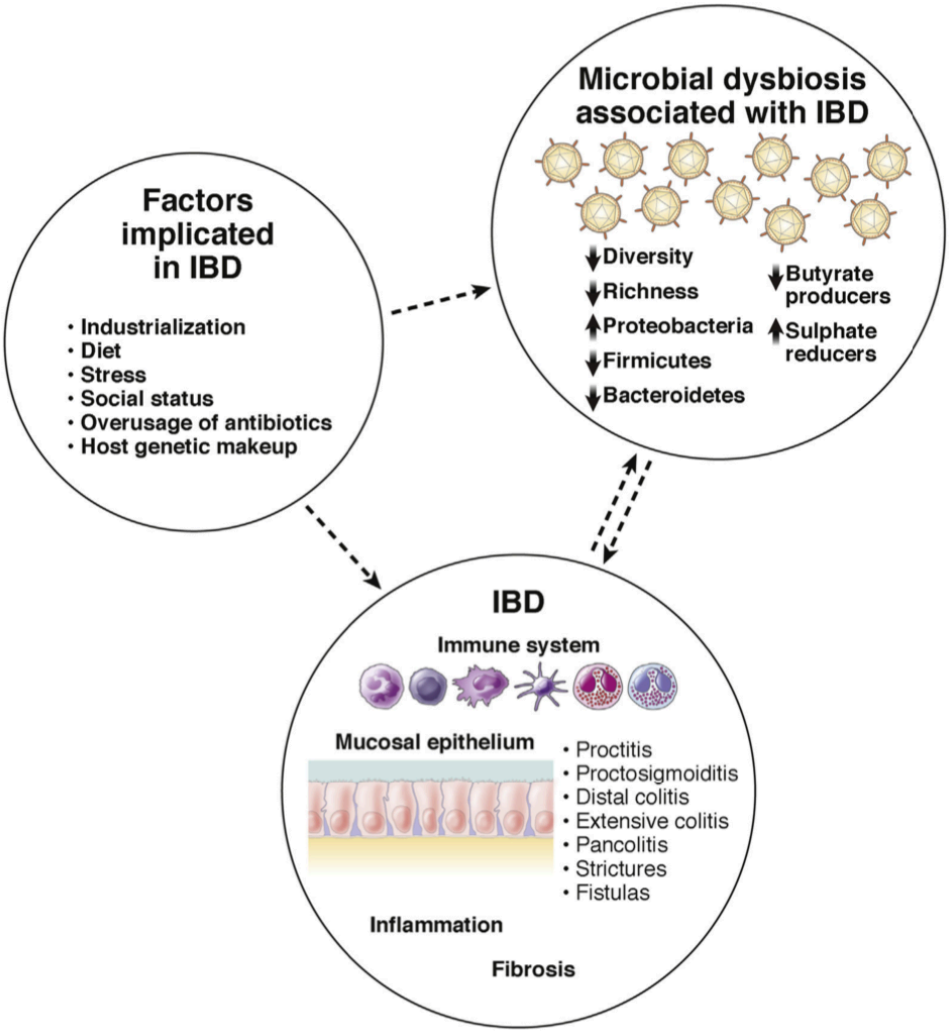


Figure 4-2: Role of the gut microbiota in differential disease severity. Depletion or expansion of a candidate set of microbes in the baseline gut microbiota of new-onset, treatment-naïve ulcerative colitis patients was associated with initial disease severity. Microbes with increased abundance in more sever disease patients at diagnosis are indicated in orange. Microbes that were depleted in patients with severe disease are shown in violet.

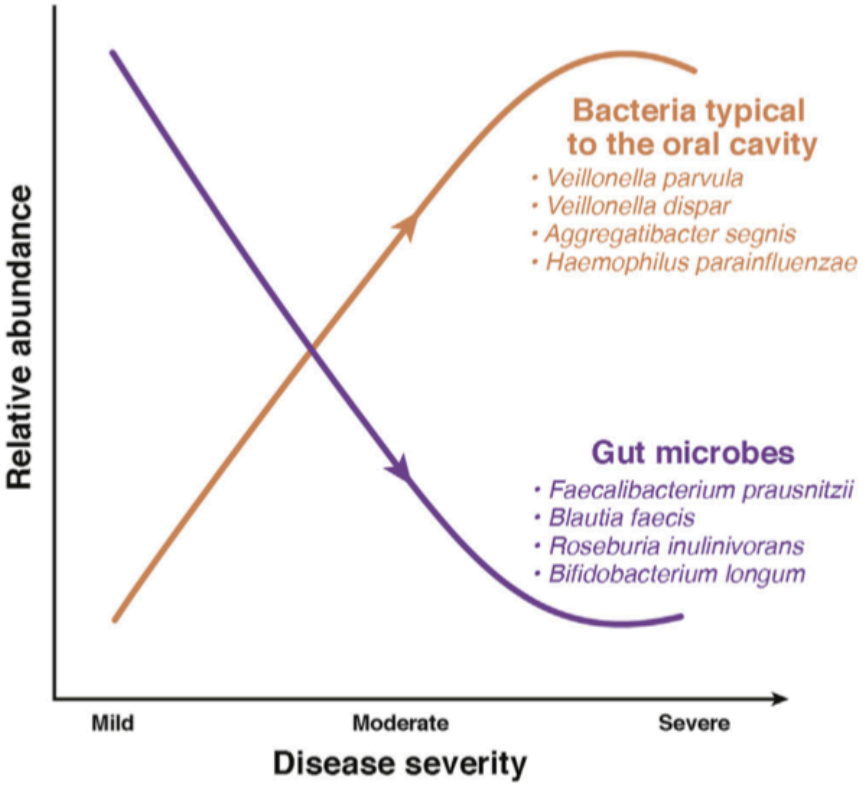
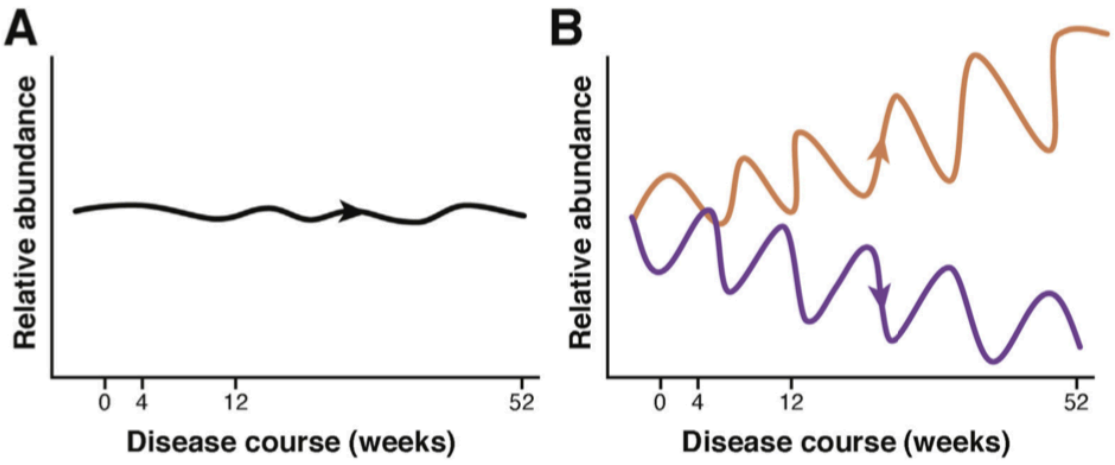


Figure 4-3: Role of the gut microbiota in disease progression. Longitudinal trajectory of microbes whose expansion (orange) or depletion (violet) in the baseline gut microbiota was indicative of patients who are at risk for eventual colectomy. Temporal changes in abundance of specific species associated with poor-prognosis defined as the need for colectomy within 1 year after diagnosis showed consistent increase or decrease with worsening disease, with reduced microbial stability over time.



REFERENCES

1. Molodecky NA, Soon IS, Rabi DM, et al. Increasing incidence and prevalence of the inflammatory bowel diseases with time, based on systematic review. Gastroenterology 2012;142:46-54 e42; quiz e30.
2. Rocchi A, Benchimol EI, Bernstein CN, et al. Inflammatory bowel disease: a Canadian burden of illness review. Can J Gastroenterol 2012;26:811-7.
3. Hammer T, Nielsen KR, Munkholm P, et al. The Faroese IBD Study: Incidence of Inflammatory Bowel Diseases Across 54 Years of Population-based Data. J Crohns Colitis 2016;10:934-42.
4. Manichanh C, Borruel N, Casellas F, et al. The gut microbiota in IBD. Nat Rev Gastroenterol Hepatol 2012;9:599-608.
5. Kaplan GG. The global burden of IBD: from 2015 to 2025. Nat Rev Gastroenterol Hepatol 2015;12:720-7.
6. Ng SC, Shi HY, Hamidi N, et al. Worldwide incidence and prevalence of inflammatory bowel disease in the 21st century: a systematic review of population-based studies. Lancet 2018;390:2769-2778.
7. Benchimol EI, Manuel DG, Guttmann A, et al. Changing age demographics of inflammatory bowel disease in Ontario, Canada: a population-based cohort study of epidemiology trends. Inflamm Bowel Dis 2014;20:1761-9.
8. Benchimol EI, Guttmann A, Griffiths AM, et al. Increasing incidence of paediatric inflammatory bowel disease in Ontario, Canada: evidence from health administrative data. Gut 2009;58:1490-7.
9. Jostins L, Ripke S, Weersma RK, et al. Host-microbe interactions have shaped the genetic architecture of inflammatory bowel disease. Nature 2012;491:119-24.
10. Cho JH. The genetics and immunopathogenesis of inflammatory bowel disease. Nat Rev Immunol 2008;8:458-66.
11. 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.
12. 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.
13. Inohara N, Ogura Y, Fontalba A, et al. Host recognition of bacterial muramyl dipeptide mediated through NOD2. Implications for Crohn's disease. J Biol Chem 2003;278:5509-12.
14. Hampe J, Franke A, Rosenstiel P, et al. A genome-wide association scan of nonsynonymous SNPs identifies a susceptibility variant for Crohn disease in ATG16L1. Nat Genet 2007;39:207-11.
15. 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;139:1630-41, 1641 e1-2.
16. Travassos LH, Carneiro LA, 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.
17. Boada-Romero E, Serramito-Gomez I, Sacristan MP, et al. The T300A Crohn's disease risk polymorphism impairs function of the WD40 domain of ATG16L1. Nat Commun 2016;7:11821.
18. Parkes M, Barrett JC, Prescott NJ, et al. Sequence variants in the autophagy gene IRGM and multiple other replicating loci contribute to Crohn's disease susceptibility. Nat Genet 2007;39:830-2.
19. Bevins CL, Salzman NH. Paneth cells, antimicrobial peptides and maintenance of intestinal homeostasis. Nat Rev Microbiol 2011;9:356-68.
20. De Filippo C, Cavalieri D, Di Paola M, et al. Impact of diet in shaping gut microbiota revealed by a comparative study in children from Europe and rural Africa. Proc Natl Acad Sci U S A 2010;107:14691-6.
21. Schnorr SL, Candela M, Rampelli S, et al. Gut microbiome of the Hadza hunter-gatherers. Nat Commun 2014;5:3654.

22. Yatsunenko T, Rey FE, Manary MJ, et al. Human gut microbiome viewed across age and geography. Nature 2012;486:222-7.
23. Martinez I, Stegen JC, Maldonado-Gomez MX, et al. The gut microbiota of rural papua new guineans: composition, diversity patterns, and ecological processes. Cell Rep 2015;11:527-38.
24. Bernstein CN, Shanahan F. Disorders of a modern lifestyle: reconciling the epidemiology of inflammatory bowel diseases. Gut 2008;57:1185-91.
25. Hviid A, Svanstrom H, Frisch M. Antibiotic use and inflammatory bowel diseases in childhood. Gut 2011;60:49-54.
26. Gevers D, Kugathasan S, Denson LA, et al. The treatment-naive microbiome in new-onset Crohn's disease. Cell Host Microbe 2014;15:382-392.
27. Frank DN, St Amand AL, Feldman RA, 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.
28. Sokol H, Pigneur B, Watterlot L, et al. Faecalibacterium prausnitzii is an anti-inflammatory commensal bacterium identified by gut microbiota analysis of Crohn disease patients. Proc Natl Acad Sci U S A 2008;105:16731-6.
29. Darfeuille-Michaud A, Neut C, Barnich N, et al. Presence of adherent Escherichia coli strains in ileal mucosa of patients with Crohn's disease. Gastroenterology 1998;115:1405-13.
30. Petersen AM, Nielsen EM, Litrup E, et al. A phylogenetic group of Escherichia coli associated with active left-sided inflammatory bowel disease. BMC Microbiol 2009;9:171.
31. Varela E, Manichanh C, Gallart M, et al. Colonisation by Faecalibacterium prausnitzii and maintenance of clinical remission in patients with ulcerative colitis. Aliment Pharmacol Ther 2013;38:151-61.
32. Joossens M, Huys G, Cnockaert M, et al. Dysbiosis of the faecal microbiota in patients with Crohn's disease and their unaffected relatives. Gut 2011;60:631-7.
33. Ohkusa T, Okayasu I, Ogihara T, et al. Induction of experimental ulcerative colitis by Fusobacterium varium isolated from colonic mucosa of patients with ulcerative colitis. Gut 2003;52:79-83.
34. Strauss J, Kaplan GG, Beck PL, et al. Invasive potential of gut mucosa-derived Fusobacterium nucleatum positively correlates with IBD status of the host. Inflamm Bowel Dis 2011;17:1971-8.
35. Wagner J, Maksimovic J, Farries G, et al. Bacteriophages in gut samples from pediatric Crohn's disease patients: metagenomic analysis using 454 pyrosequencing. Inflamm Bowel Dis 2013;19:1598-608.
36. Norman JM, Handley SA, Baldridge MT, et al. Disease-specific alterations in the enteric virome in inflammatory bowel disease. Cell 2015;160:447-60.
37. Sokol H, Leducq V, Aschard H, et al. Fungal microbiota dysbiosis in IBD. Gut 2017;66:1039-1048.
38. Liguori G, Lamas B, Richard ML, et al. Fungal Dysbiosis in Mucosa-associated Microbiota of Crohn's Disease Patients. J Crohns Colitis 2016;10:296-305.
39. Ott SJ, Kuhbacher T, Musfeldt M, et al. Fungi and inflammatory bowel diseases: Alterations of composition and diversity. Scand J Gastroenterol 2008;43:831-41.
40. Chehoud C, Albenberg LG, Judge C, et al. Fungal Signature in the Gut Microbiota of Pediatric Patients With Inflammatory Bowel Disease. Inflamm Bowel Dis 2015;21:1948-56.
41. Li J, Butcher J, Mack D, et al. Functional impacts of the intestinal microbiome in the pathogenesis of inflammatory bowel disease. Inflamm Bowel Dis 2015;21:139-53.
42. Rigottier-Gois L. Dysbiosis in inflammatory bowel diseases: the oxygen hypothesis. ISME J 2013;7:1256-61.
43. Khan KJ, Ullman TA, Ford AC, et al. Antibiotic therapy in inflammatory bowel disease: a systematic review and meta-analysis. Am J Gastroenterol 2011;106:661-73.
44. Wang SL, Wang ZR, Yang CQ. Meta-analysis of broad-spectrum antibiotic therapy in patients with active inflammatory bowel disease. Exp Ther Med 2012;4:1051-1056.

45. Casellas F, Borruel N, Papo M, et al. Antiinflammatory effects of enterically coated amoxicillin-clavulanic acid in active ulcerative colitis. Inflamm Bowel Dis 1998;4:1-5.
46. Lepage P, Hasler R, Spehlmann ME, et al. Twin study indicates loss of interaction between microbiota and mucosa of patients with ulcerative colitis. Gastroenterology 2011;141:227-36.
47. Harper PH, Lee EC, Kettlewell MG, et al. Role of the faecal stream in the maintenance of Crohn's colitis. Gut 1985;26:279-84.
48. Rutgeerts P, Goboes K, Peeters M, et al. Effect of faecal stream diversion on recurrence of Crohn's disease in the neoterminal ileum. Lancet 1991;338:771-4.
49. Janowitz HD, Croen EC, Sachar DB. The role of the fecal stream in Crohn's disease: an historical and analytic review. Inflamm Bowel Dis 1998;4:29-39.
50. D'Haens GR, Geboes K, Peeters M, et al. Early lesions of recurrent Crohn's disease caused by infusion of intestinal contents in excluded ileum. Gastroenterology 1998;114:262-7.
51. Kugathasan S, Denson LA, Walters TD, et al. Prediction of complicated disease course for children newly diagnosed with Crohn's disease: a multicentre inception cohort study. Lancet 2017;389:1710-1718.
52. Shaw KA, Bertha M, Hofmekler T, et al. Dysbiosis, inflammation, and response to treatment: a longitudinal study of pediatric subjects with newly diagnosed inflammatory bowel disease. Genome Med 2016;8:75.
53. Halfvarson J, Brislawn CJ, Lamendella R, et al. Dynamics of the human gut microbiome in inflammatory bowel disease. Nat Microbiol 2017;2:17004.
54. Haberman Y, Tickle TL, Dexheimer PJ, et al. Pediatric Crohn disease patients exhibit specific ileal transcriptome and microbiome signature. J Clin Invest 2014;124:3617-33.
55. Ananthakrishnan AN, Luo C, Yajnik V, et al. Gut Microbiome Function Predicts Response to Anti-integrin Biologic Therapy in Inflammatory Bowel Diseases. Cell Host Microbe 2017;21:603-610 e3.
56. Scaldaferri F, Gerardi V, Mangiola F, et al. Role and mechanisms of action of Escherichia coli Nissle 1917 in the maintenance of remission in ulcerative colitis patients: An update. World J Gastroenterol 2016;22:5505-11.
57. Rembacken BJ, Snelling AM, Hawkey PM, et al. Non-pathogenic Escherichia coli versus mesalazine for the treatment of ulcerative colitis: a randomised trial. Lancet 1999;354:635-9.
58. Ishikawa H, Akedo I, Umesaki Y, et al. Randomized controlled trial of the effect of bifidobacteria-fermented milk on ulcerative colitis. J Am Coll Nutr 2003;22:56-63.
59. Furrie E, Macfarlane S, Kennedy A, et al. Synbiotic therapy (Bifidobacterium longum/Synergy 1) initiates resolution of inflammation in patients with active ulcerative colitis: a randomised controlled pilot trial. Gut 2005;54:242-9.
60. Kato K, Mizuno S, Umesaki Y, et al. Randomized placebo-controlled trial assessing the effect of bifidobacteria-fermented milk on active ulcerative colitis. Aliment Pharmacol Ther 2004;20:1133-41.
61. Chapman TM, Plosker GL, Figgitt DP. VSL#3 probiotic mixture: a review of its use in chronic inflammatory bowel diseases. Drugs 2006;66:1371-87.
62. Sood A, Midha V, Makharia GK, et al. The probiotic preparation, VSL#3 induces remission in patients with mild-to-moderately active ulcerative colitis. Clin Gastroenterol Hepatol 2009;7:1202-9, 1209 e1.
63. Miele E, Pascarella F, Giannetti E, et al. Effect of a probiotic preparation (VSL#3) on induction and maintenance of remission in children with ulcerative colitis. Am J Gastroenterol 2009;104:437-43.
64. Venturi A, Gionchetti P, Rizzello F, et al. Impact on the composition of the faecal flora by a new probiotic preparation: preliminary data on maintenance treatment of patients with ulcerative colitis. Aliment Pharmacol Ther 1999;13:1103-8.
65. Bibiloni R, Fedorak RN, Tannock GW, et al. VSL#3 probiotic-mixture induces remission in patients with active ulcerative colitis. Am J Gastroenterol 2005;100:1539-46.

66. Huynh HQ, deBruyn J, Guan L, et al. Probiotic preparation VSL#3 induces remission in children with mild to moderate acute ulcerative colitis: a pilot study. Inflamm Bowel Dis 2009;15:760-8.
67. Wildt S, Nordgaard I, Hansen U, et al. A randomised double-blind placebo-controlled trial with Lactobacillus acidophilus La-5 and Bifidobacterium animalis subsp. lactis BB-12 for maintenance of remission in ulcerative colitis. J Crohns Colitis 2011;5:115-21.
68. Mallon P, McKay D, Kirk S, et al. Probiotics for induction of remission in ulcerative colitis. Cochrane Database Syst Rev 2007:CD005573.
69. Naidoo K, Gordon M, Fagbemi AO, et al. Probiotics for maintenance of remission in ulcerative colitis. Cochrane Database Syst Rev 2011:CD007443.
70. Malchow HA. Crohn's disease and Escherichia coli. A new approach in therapy to maintain remission of colonic Crohn's disease? J Clin Gastroenterol 1997;25:653-8.
71. Fedorak RN, Feagan BG, Hotte N, et al. The probiotic VSL#3 has anti-inflammatory effects and could reduce endoscopic recurrence after surgery for Crohn's disease. Clin Gastroenterol Hepatol 2015;13:928-35 e2.
72. Guslandi M, Mezzi G, Sorghi M, et al. Saccharomyces boulardii in maintenance treatment of Crohn's disease. Dig Dis Sci 2000;45:1462-4.
73. Gupta P, Andrew H, Kirschner BS, et al. Is lactobacillus GG helpful in children with Crohn's disease? Results of a preliminary, open-label study. J Pediatr Gastroenterol Nutr 2000;31:453-7.
74. Schultz M, Timmer A, Herfarth HH, et al. Lactobacillus GG in inducing and maintaining remission of Crohn's disease. BMC Gastroenterol 2004;4:5.
75. Prantera C, Scribano ML, Falasco G, et al. Ineffectiveness of probiotics in preventing recurrence after curative resection for Crohn's disease: a randomised controlled trial with Lactobacillus GG. Gut 2002;51:405-9.
76. Marteau P, Lemann M, Seksik P, et al. Ineffectiveness of Lactobacillus johnsonii LA1 for prophylaxis of postoperative recurrence in Crohn's disease: a randomised, double blind, placebo controlled GETAID trial. Gut 2006;55:842-7.
77. Lindsay JO, Whelan K, Stagg AJ, et al. Clinical, microbiological, and immunological effects of fructo-oligosaccharide in patients with Crohn's disease. Gut 2006;55:348-55.
78. Benjamin JL, Hedin CR, Koutsoumpas A, et al. Randomised, double-blind, placebo-controlled trial of fructo-oligosaccharides in active Crohn's disease. Gut 2011;60:923-9.
79. Hafer A, Kramer S, Duncker S, et al. Effect of oral lactulose on clinical and immunohistochemical parameters in patients with inflammatory bowel disease: a pilot study. BMC Gastroenterol 2007;7:36.
80. Casellas F, Borruel N, Torrejon A, et al. Oral oligofructose-enriched inulin supplementation in acute ulcerative colitis is well tolerated and associated with lowered faecal calprotectin. Aliment Pharmacol Ther 2007;25:1061-7.
81. Seidner DL, Lashner BA, Brzezinski A, et al. An oral supplement enriched with fish oil, soluble fiber, and antioxidants for corticosteroid sparing in ulcerative colitis: a randomized, controlled trial. Clin Gastroenterol Hepatol 2005;3:358-69.
82. Laurell A, Sjoberg K. Prebiotics and synbiotics in ulcerative colitis. Scand J Gastroenterol 2017;52:477-485.
83. Steed H, Macfarlane GT, Blackett KL, et al. Clinical trial: the microbiological and immunological effects of synbiotic consumption - a randomized double-blind placebo-controlled study in active Crohn's disease. Aliment Pharmacol Ther 2010;32:872-83.
84. Chermesh I, Tamir A, Reshef R, et al. Failure of Synbiotic 2000 to prevent postoperative recurrence of Crohn's disease. Dig Dis Sci 2007;52:385-9.
85. Fujimori S, Gudis K, Mitsui K, et al. A randomized controlled trial on the efficacy of synbiotic versus probiotic or prebiotic treatment to improve the quality of life in patients with ulcerative colitis. Nutrition 2009;25:520-5.

86. Takaishi H, Matsuki T, Nakazawa A, et al. Imbalance in intestinal microflora constitution could be involved in the pathogenesis of inflammatory bowel disease. Int J Med Microbiol 2008;298:463-72.
87. Sokol H, Seksik P. The intestinal microbiota in inflammatory bowel diseases: time to connect with the host. Curr Opin Gastroenterol 2010;26:327-31.
88. Sokol H, Seksik P, Furet JP, et al. Low counts of Faecalibacterium prausnitzii in colitis microbiota. Inflamm Bowel Dis 2009;15:1183-9.
89. Scheppach W, Sommer H, Kirchner T, et al. Effect of butyrate enemas on the colonic mucosa in distal ulcerative colitis. Gastroenterology 1992;103:51-6.
90. Di Sabatino A, Morera R, Ciccocioppo R, et al. Oral butyrate for mildly to moderately active Crohn's disease. Aliment Pharmacol Ther 2005;22:789-94.
91. Loubinoux J, Bronowicki JP, Pereira IA, et al. Sulfate-reducing bacteria in human feces and their association with inflammatory bowel diseases. FEMS Microbiol Ecol 2002;40:107-12.
92. Zinkevich VV, Beech IB. Screening of sulfate-reducing bacteria in colonoscopy samples from healthy and colitic human gut mucosa. FEMS Microbiol Ecol 2000;34:147-155.
93. Verma R, Verma AK, Ahuja V, et al. Real-time analysis of mucosal flora in patients with inflammatory bowel disease in India. J Clin Microbiol 2010;48:4279-82.
94. Mills DJ, Tuohy KM, Booth J, et al. Dietary glycated protein modulates the colonic microbiota towards a more detrimental composition in ulcerative colitis patients and non-ulcerative colitis subjects. J Appl Microbiol 2008;105:706-14.
95. Pitcher MC, Beatty ER, Cummings JH. The contribution of sulphate reducing bacteria and 5-aminosalicylic acid to faecal sulphide in patients with ulcerative colitis. Gut 2000;46:64-72.
96. Moehle C, Ackermann N, Langmann T, et al. Aberrant intestinal expression and allelic variants of mucin genes associated with inflammatory bowel disease. J Mol Med (Berl) 2006;84:1055-66.
97. Sheng YH, Hasnain SZ, Florin TH, et al. Mucins in inflammatory bowel diseases and colorectal cancer. J Gastroenterol Hepatol 2012;27:28-38.
98. Ijssennagger N, Belzer C, Hooiveld GJ, et al. Gut microbiota facilitates dietary heme-induced epithelial hyperproliferation by opening the mucus barrier in colon. Proc Natl Acad Sci U S A 2015;112:10038-43.
99. Gough E, Shaikh H, Manges AR. Systematic review of intestinal microbiota transplantation (fecal bacteriotherapy) for recurrent Clostridium difficile infection. Clin Infect Dis 2011;53:994-1002.
100. Paramsothy S, Kamm MA, Kaakoush NO, et al. Multidonor intensive faecal microbiota transplantation for active ulcerative colitis: a randomised placebo-controlled trial. Lancet 2017;389:1218-1228.
101. Moayyedi P, Surette MG, Kim PT, et al. Fecal Microbiota Transplantation Induces Remission in Patients With Active Ulcerative Colitis in a Randomized Controlled Trial. Gastroenterology 2015;149:102-109 e6.
102. Rossen NG, Fuentes S, van der Spek MJ, et al. Findings From a Randomized Controlled Trial of Fecal Transplantation for Patients With Ulcerative Colitis. Gastroenterology 2015;149:110-118 e4.
103. Suskind DL, Brittnacher MJ, Wahbeh G, et al. Fecal microbial transplant effect on clinical outcomes and fecal microbiome in active Crohn's disease. Inflamm Bowel Dis 2015;21:556-63.
104. Cui B, Feng Q, Wang H, et al. Fecal microbiota transplantation through mid-gut for refractory Crohn's disease: safety, feasibility, and efficacy trial results. J Gastroenterol Hepatol 2015;30:51-8.
105. Hall AB, Tolonen AC, Xavier RJ. Human genetic variation and the gut microbiome in disease. Nat Rev Genet 2017;18:690-699.
106. Kostic AD, Xavier RJ, Gevers D. The microbiome in inflammatory bowel disease: current status and the future ahead. Gastroenterology 2014;146:1489-99.
107. Khor B, Gardet A, Xavier RJ. Genetics and pathogenesis of inflammatory bowel disease. Nature 2011;474:307-17.

108. Philpott DJ, Sorbara MT, Robertson SJ, et al. NOD proteins: regulators of inflammation in health and disease. Nat Rev Immunol 2014;14:9-23.
109. Knights D, Silverberg MS, Weersma RK, et al. Complex host genetics influence the microbiome in inflammatory bowel disease. Genome Med 2014;6:107.
110. Chassaing B, Rolhion N, de Vallee A, et al. Crohn disease--associated adherent-invasive E. coli bacteria target mouse and human Peyer's patches via long polar fimbriae. J Clin Invest 2011;121:966-75.
111. Lapaquette P, Bringer MA, Darfeuille-Michaud A. Defects in autophagy favour adherent-invasive Escherichia coli persistence within macrophages leading to increased pro-inflammatory response. Cell Microbiol 2012;14:791-807.
112. Li D, Achkar JP, Haritunians T, et al. A Pleiotropic Missense Variant in SLC39A8 Is Associated With Crohn's Disease and Human Gut Microbiome Composition. Gastroenterology 2016;151:724-32.
113. Iliev ID, Funari VA, Taylor KD, et al. Interactions between commensal fungi and the C-type lectin receptor Dectin-1 influence colitis. Science 2012;336:1314-7.

Chapter 5

Site- and Taxa-Specific Disease-Associated Oral Microbial Structures Distinguish Patients with Inflammatory Bowel Disease

This chapter has been adapted and was originally submitted to *Clinical Gastroenterology and Hepatology.*

Hari K. Somineni, MS; Jordan H. Weitzner, MD; Suresh Venkateswaran, PhD; Anne Dodd, BS; Khuong U. Le, MS; Jarod Prince, BS; Arjuna Karikaran; Natalie Hoeting, MD; Cary G. Sauer, MD; Shelly Abramowicz, MD, DMD; Michael E. Zwick, PhD; David J. Cutler, PhD; David T. Okou, PhD; Pankaj Chopra, PhD and Subra Kugathasan, MD

ABSTRACT

Background: The gut and oral microbiome have independently been shown to be associated with inflammatory bowel disease. However, it is not known to what extent gut and oral microbial disease markers converge in terms of their composition in inflammatory bowel disease. Further, the spatial and temporal variation within the oral microenvironments of inflammatory bowel disease remain to be elucidated.

Methods: We used a prospectively recruited cohort of pediatric patients with inflammatory bowel disease (n = 47) and unrelated healthy controls (n = 18) to examine the spatial and temporal distribution of microbiota within the various oral microenvironments, represented by saliva, tongue, buccal mucosa and plaque, and compared them with stool. Microbiome characterization was performed using 16S rRNA gene sequencing.

Results: The oral microbiome displays inflammatory bowel disease-associated dysbiosis, in a site- and taxa-specific manner. Plaque samples depicted a relatively severe degree of dysbiosis, while the disease-associated dysbiotic bacterial groups were predominantly the members of the phylum Firmicutes. Our 16S rRNA gene analyses demonstrate that oral microbiota can discriminate inflammatory bowel disease patients from healthy controls, with salivary microbiota performing the best, closely matched by the stool and other oral sites. Longitudinal profiles of microbial composition suggest that some taxa are more consistently perturbed than others, preferentially in a site-dependent fashion.

Conclusions: Collectively, these data indicate the potential of using oral microbial profiles in screening and monitoring patients with inflammatory bowel disease. Furthermore, these results support the importance of spatial and longitudinal microbiome sampling to interpret disease-associated dysbiotic states and eventually to gain insights into disease pathogenesis.

Keywords: Inflammatory bowel disease, 16S rRNA, Microbiota, Microbiome, Spatial, Temporal, and Longitudinal

INTRODUCTION

Inflammatory bowel disease is a life-long condition characterized by intestinal ulceration, pain, rectal bleeding, loss of quality of life, and a need for bowel surgery. Its increasing prevalence has been documented within the developed world[1-5]. Crohn's disease and ulcerative colitis are the two classical forms of inflammatory bowel disease. Both Crohn's disease and ulcerative colitis share many clinical and extraintestinal manifestations and hence it is often difficult to make an accurate diagnosis, particularly at the earliest stages of disease. While most inflammatory bowel disease patients respond to standard-of-care clinical treatment, some patients rapidly progress to complicated disease behaviors such as perforated bowel, stricturing due to fibrosis and/or penetrating fistulas[6]. Endoscopic evaluation combined with histopathological examination of the mucosal-biopsy is the gold-standard for diagnosis or disease monitoring in inflammatory bowel disease. However, these invasive procedures are associated with high cost and relatively low patient acceptance rate and are not ideal for disease monitoring and assessing response to therapy. Serologic studies have been proposed to help diagnose and monitor inflammatory bowel disease, but suffer from a low sensitivity and specificity. Therefore, there is a compelling need for the identification of novel noninvasive, cost-effective, robust, and reproducible biomarkers for accurate diagnosis, treatment selection and disease monitoring.

Although the exact mechanism is not known, the pathogenesis of inflammatory bowel disease has been attributed to a dysregulated immune response to alterations in the gut microbial composition in genetically susceptible individuals[7, 8]. Genes and susceptibility genetic loci implicated in inflammatory bowel disease have been shown to be enriched for pathways involving bacterial recognition or host response to microbial infections suggesting a microbial contribution to disease pathogenesis[9]. Screening for perturbations in fecal microbial composition, referred to as microbial dysbiosis, has emerged as a promising noninvasive approach for inflammatory bowel disease screening[10, 11]. We and others have previously shown that pre-treatment stool microbial dysbiosis is present in inflammatory bowel disease and that a fecal microbial dysbiosis index could be used as a screening tool to diagnose inflammatory bowel disease, differentiate

Crohn's disease from ulcerative colitis as well as therapy responders from non-responders[10, 11]. Furthermore, a recent study by Ananthakrishnan *et al.*, has demonstrated the potential of gut microbiome in predicting response to anti-integrin therapy in inflammatory bowel disease[12]. Although the mechanistic framework underlying the therapeutic potential of gut microbiota has not yet been fully elucidated, there is ample experimental evidence suggesting the causal role for gut microbial dysbiosis in inflammatory bowel disease susceptibility and progression[13]. However, it is not clear whether this dysbiosis is specific to gut microbial community or is a systemic phenomenon in inflammatory bowel disease.

Studies have begun to reveal oral microbial alterations in inflammatory bowel disease[14, 15]. A pediatric study that included 40 Crohn's disease patients and 43 non-inflammatory bowel disease controls reported a significant decrease in overall diversity of tongue microbiota in Crohn's disease[15]. Phylum level analysis of salivary microbiota revealed increased abundance of Bacteroidetes and reduced abundance of Proteobacteria in inflammatory bowel disease patients[16]. Furthermore, perturbed salivary microbial communities in inflammatory bowel disease showed statistically significant associations with inflammatory cytokines such as *IL-1β* and *IL-8* and lysozyme levels[16]. Collectively, these evidence provide a rationale for using oral microbial markers for the detection of the presence or severity of inflammatory bowel disease. The oral microbiota includes a large repertoire of about 700 bacterial species or phylotypes; however, less is known about which anatomical location within the oral cavity is more indicative of inflammatory bowel disease. Further, it is not known to what extent oral and gut microbial disease markers converge in terms of their composition in subjects with inflammatory bowel disease. Here we used a prospectively recruited cohort of pediatric inflammatory bowel disease subjects, both treatment naïve inflammatory bowel disease and established inflammatory bowel disease, along with unrelated healthy controls, to examine the (1) spatial and temporal dynamics of the oral microbiota in inflammatory bowel disease, (2) concordance and divergence between oral and gut microbiota in inflammatory bowel disease and (3) predictive potential of the oral and gut microbiota in assessing the presence of the disease.

METHODS

Study population

Treatment naïve-newly diagnosed or established inflammatory bowel disease patients were recruited from the Children's Healthcare of Atlanta inpatient wards and outpatient pediatric inflammatory bowel disease clinics. Criteria to participate in the study included Crohn's disease or ulcerative colitis diagnosis confirmed by colonoscopy and/or magnetic resonance enterography, willingness to participate, and ability to maintain close follow-up. The control population composed of unrelated, age- and gender-matched healthy individuals who volunteered to participate upon request. Exclusion criteria included subjects who were on or had a recent history (within the preceding month) of antibiotic treatment at the time of enrollment, and subjects who demonstrated oral infections or manifestations relevant to inflammatory bowel disease or any oral diseases.

A total of 65 subjects (31 Crohn's disease, 16 ulcerative colitis and 18 healthy controls) ranging in age from 5 to 20 years (median age of 14 years) were enrolled in the study between January 2015 and February 2017 (**Fig. 5-1**). Of the 65, 44 of them (25 Crohn's disease, 9 ulcerative colitis and 10 healthy controls) were followed longitudinally at regular intervals for up to a maximum period of 88 weeks, which yielded up to 6 follow-up samples over time (**Fig. 5-1**). Of the 47 inflammatory bowel disease patients, 26 (55%) were treatment naïve at the time of enrollment (17 Crohn's disease, 9 ulcerative colitis). Subjects with suspected diagnosis of inflammatory bowel disease based on the symptoms and lab work were approached for participation in the new-onset portion of the study. These patients did not have a prior inflammatory bowel disease diagnosis, prior history of immunomodulator therapy or biologic therapy. Whereas, rest of the inflammatory bowel disease patients (n = 21; 45%), were *a priori* established for one of the two forms of inflammatory bowel disease (13 Crohn's disease, 8 ulcerative colitis) and were on concomitant therapy or had a prior history of immunomodulator and/or biologic therapy at the time of enrollment. Inflammatory bowel disease diagnoses were conducted according to the Paris Classification[17] at the time of enrollment. Demographic and phenotypic data were collected on each subject enrolled via patient interview and chart review at the time of sample collection. Abbreviated pediatric Crohn's disease activity index (PCDAI)[18, 19]

or pediatric ulcerative colitis activity index (PUCAI)[20] was obtained at all clinical visits. Medical treatment was not affected by participating in this study. Although, no subject was on antibiotic therapy during the enrollment, a small number of patients (n = 6; **Supplementary Table 5-1A**) reported short courses of antibiotic usage during the course of the follow-up period. All participants and families provided informed consent and assent for specimen collection and analysis under the study protocol approved by the Institutional Review Board of Children's Healthcare of Atlanta.

Specimen collection and processing

Oral microbiota samples were collected using DNA swabs (Isohelix, United Kingdom) from four anatomically different regions within the oral cavity – saliva, tongue, plaque, and buccal mucosa. Locations sampled include dorsum of the tongue (for tongue samples), buccaneers surface of central and lateral incisors at the gum line (plaque samples), and inside of cheek (buccal mucosa). All samples were immediately stored at -80 °C until further processing. Bacterial genomic DNA was extracted from all the oral samples using BiOstic Bacteremia DNA Isolation Kit (MO BIO Laboratories Inc., Carlsbad, CA) according to manufacturer's guidelines. All samples from the same subject were processed together to minimize batch effects. Fecal specimens were collected along with oral samples from subjects whenever possible. Each fecal sample was collected in a Para-Pak vial (Meridian Bioscience Inc., Cincinnati, OH) which contained no additional additive. Fecal specimen was stored at -20 °C until it was aliquoted into smaller workable units then stored at -80 °C. DNA from fecal samples was extracted using the MagAttract PowerMicrobiome DNA/RNA Kit (Qiagen, Valencia, CA) according to manufacturer's instructions.

16S rRNA gene sequencing and curation

The V4 region of the 16S rRNA gene was PCR amplified and sequenced on an Illumina MiSeq platform using a 2 x 250-bp paired-end protocol adapted from the Human Microbiome Project[21, 22]. The forward and reverse primer sequences were provided in **Supplementary Table 5-1B**. The obtained sequences were curated using the mothur pipeline (v1.38)[23, 24]. Briefly, paired-end reads were merged into contigs, screened for quality following the mothur MiSeq standard operating procedure, and then aligned to SILVA 16S

rRNA gene sequence database. Aligned sequences were then screened for chimeras using the VSEARCH algorithm[25]. Sequences were classified using a naive Bayesian classifier trained against a 16S rRNA gene training set provided by the Ribosomal Database Project (RDP)[26]. Sequences were then clustered into operational taxonomic units (OTUs) using a 97% similarity cutoff with the average neighbor clustering algorithm.

16S rRNA gene sequencing data analysis

OTU-based overall microbial diversity was estimated by calculating three alpha-diversity indices, Shannon, Simpson and alpha. OTU-based overall richness was determined by calculating the Chao1 richness estimate. Differences in overall microbial community structure was visualized by calculating Bray-Curtis dissimilarity measures between all pairs of samples. The significant differences in Principal Coordinate Analysis (PCoA) plots were analyzed using PERMANOVA < 0.05. All the available samples from each subject per site were used for estimating the overall diversity, richness and community structural differences. To statistically test for the individual microbial member level differences in the relative abundances of taxa at different taxonomic levels, phyla, class, order or family, between groups, we used the metagenomeSeq[27], a Bioconductor package, which uses a zero-inflated Gaussian distribution mixture model. Cumulative sum scaling, using default settings was used to normalize the data set prior to fitting the model. Differential abundance analysis was carried out with respect to the baseline samples that were collected on the first visit. In case where the first visit sample is unavailable, we used the first available sample from each subject. We included age, gender, ethnicity, and anti-TNF treatment status as covariates in the model. The *P* values presented for the differential abundance analyses were obtained from 10,000 permutations, which were then corrected for multiple hypothesis testing using the false-discovery rate (FDR) method.

Random Forest

To ascertain if the oral microbiome could distinguish inflammatory bowel disease patients from healthy controls, we used a random forest classifier which is a decision tree based algorithm[28]. We used the same

data set as in differential abundance tests in this analysis, *i.e* from each site, we only selected those samples that were collected on the first visit. If the sample at first visit was missing, we selected the sample from the next visit. Details of the number of samples per group and per site used for the random classifier and the AUC statistics are given in **Supplementary Table 5-1C**. Data from each site was divided into training ($2/3^{rd}$ of the data set) and test ($1/3^{rd}$) data sets. The random forest classifier was run on the training set, with 10,000 random trees, and the predicted model was then used on the test data set to get AUC. We created 100 such random splits for the training and test datasets and used the random forest classifier to get the AUC prediction for each site.

RESULTS

16S rRNA gene data processing

We sequenced bacterial 16S ribosomal RNA gene using the Illumina MiSeq platform with primers targeting the V4 variable regions. Using this approach, we generated a data set consisting of a median of at least ~33,000 reads per sample for each site (**Fig. 5-2** and **Supplementary Table 5-1D)**. Of these, sequences that passed the quality control criteria were sorted into operational taxonomic units (OTUs). All the samples across the habitats and over time were then rarefied to 6,575 reads per sample to minimize the effects of uneven sampling, which resulted in a data set with 768 samples and 4,259 OTUs (**Supplementary Table 5-1A**). The rarefaction curves for all the sites, collectively and individually are shown in **Fig. 5-3**. To ensure robustness, we applied three separate filters to this data set which resulted in three independent data sets for downstream analyses. Our first filter retained OTUs that were present in at least 1% of the total samples (n = 768), yielding 753 OTUs. The second filter retained OTUs present in at least 5% of the total samples, which resulted in a data set consisting of 268 OTUs and the third filter retained OTUs that were present in at least 1% of the total samples and had a minimum total read count of 50 for all samples. Using the third approach, we generated a data set consisting of 462 OTUs. All the downstream analyses were performed on all three datasets; however, the results presented below pertain to the data set with 753 OTUs, unless stated otherwise.

Site-specific microbial composition within the oral habitat

In order to examine beta diversity, we assessed differences in overall microbial community structure across all the habitats and over time from both cases and controls using a non-phylogeny based Bray-Curtis dissimilarity metric. A relatively small Bray-Curtis distance implies that the two communities are similar where majority of the species are shared. Consistent with previous notion, Bray-Curtis distance based principal coordinates analysis revealed strong primary clustering by habitat, rather than by disease status or over time reflecting the high heterogeneity of the sampled habitats (**Fig. 5-4a**). Within the oral microenvironments, buccal samples showed clear separation from tongue (**Fig. 5-4a** and **Fig. 5-5**), suggesting that anatomical regions within the oral cavity were akin to microbial "islands", possessing distinct bacterial communities that persisted temporally. This observation is consistent with the previous reports from the Human Microbiome Project (HMP) and other studies demonstrating that most oral bacterial taxa are habitat specialists[29, 30]. On the other hand, microbial diversity between saliva and plaque appears to be nominal and the two sites were almost indistinguishable from one another on the plotted ordination axes (**Fig. 5-4a** and **Fig. 5-5**).

Richness, diversity and relative abundance of oral microbiota

Microbial diversity and richness vary by anatomic site[31], partly in response to the local environment and biology of each body habitat. To this end we evaluated spatial trends in the structure of the bacterial communities by using both global parameters as well as at an individual microbial member level, across the four oral sites, saliva, tongue, plaque and buccal mucosa. When compared across sites using ANOVA followed by Tukey's test (number of samples per site are presented in **Supplementary Table 5-1E**), we identified statistically significant differences in the overall diversity of buccal microbiota as measured by the Shannon diversity index ($P < 1 \times 10^{-7}$, compared to other three oral sites); however, no significant differences were noted amongst the other three oral sites (**Fig. 5-4b**). On the other hand, we did not observe any significant differences between oral sites in overall richness as measured by Chao1 (**Fig. 5-4c**). As

expected, all four oral sites showed significant differences in both overall diversity and richness when compared with stool (**Figs. 5-4b** and **5-4c**).

Further, we noted changes in composition of individual microbial members, between habitats (**Figs. 5-6, 5-7)**. The human microbiota is typically dominated by the four bacterial phyla, Firmicutes, Bacteroidetes, Proteobacteria and Actinobacteria[31-33]. Our analysis of oral microbiota at the phylum level showed Firmicutes (46%), Bacteroidetes (12%), Proteobacteria (25%) and Actinobacteria (10%) in saliva; Firmicutes (42%), Bacteroidetes (18%), Proteobacteria (21%) and Actinobacteria (13%) in tongue; Firmicutes (35%), Bacteroidetes (11%), Proteobacteria (21%) and Actinobacteria (22%) in plaque; Firmicutes (53%), Bacteroidetes (9%), Proteobacteria (25%) and Actinobacteria (7%) in buccal (**Fig. 5-6)**. Although the same four phyla dominated the microbiota in stool, Firmicutes (72%), Bacteroidetes (5%), Proteobacteria (3%) and Actinobacteria (17%), we noted a pronounced reduction in terms of the relative abundance of Firmicutes and an increase in Proteobacteria in oral cavity compared to stool (**Fig. 5-6)**. This shift in Firmicutes and Proteobacteria in oral microbiota is interesting in particular, as is in line with the shift seen in the intestinal microbiota of severe ulcerative colitis which is characterized by a decline in Firmicutes and an increase in Proteobacteria when compared to mild or moderate ulcerative colitis[34]. When looked at the members of these two phyla, most members (6/7) of the phylum Proteobacteria depicted increased abundance in oral cavity compared to stool which is directionally consistent with the shift seen at the phyla level (**Fig. 5-7**). Whereas, members belonging to the class Clostridia (5/6) of the phylum Firmicutes, demonstrated directional consistency while members of the class Bacilli showed polarizing shifts (**Fig. 5-7**).

Site- and taxa-specific oral microbial dysbiosis in inflammatory bowel disease

To assess overall differences in microbial community structure in inflammatory bowel disease patients and controls, we calculated measures of alpha- and beta-diversity in all the four profiled oral sites as well as in the fecal microbiota. As shown in **Fig. 5-8**, beta-diversity entropy measured using Bray-Curtis dissimilarity depicted statistically significant differences between inflammatory bowel disease patients (Crohn's disease

or ulcerative colitis) and healthy controls, in a site-specific manner (PERMANOVA < 0.05). Similarly, we noted site-specific microbial differences in alpha diversity. In agreement with the previous notion, inflammatory bowel disease was associated with a reduced diversity in stool as indicated by the alpha index ($P = 0.003$) and reduced richness as indicated by the Chao1 index ($P = 0.001$) (**Fig. 5-9**). Interestingly, we noted a similar trend in terms of overall richness in saliva as measured by the Chao1 index ($P = 0.082$), whereas, no such trend was found for any other oral sites including tongue, plaque and buccal mucosa ($P > 0.1$; **Fig. 5-9**). However, it should be noted that our findings from both stool and saliva are not robust to other alpha-diversity measures (**Fig. 5-9**).

Next we surveyed oral microbial samples for inflammatory bowel disease associated changes at an individual microbial member level, with respect to the first available sample, adjusting for age, gender, ethnicity, and anti-TNF treatment status. Because anti-TNF therapy may skew microbiota composition, we used anti-TNF status as a covariate, besides age, gender, and ethnicity. Antibiotic usage was not included as a covariate in our differential abundance analysis since no subject was reported to be on antibiotics with respect to the first available sample. At the population level, we noted significant differences or pronounced shifts in relative abundances of several bacterial members at different taxonomic levels between inflammatory bowel disease cases and healthy controls. For instance, at the phyla level, Actinobacteria, Bacteroidetes and Spirochaetes showed a trend for enrichment in inflammatory bowel disease patients across all four oral sites, whereas Fusobacteria, Firmicutes and Proteobacteria were among the phyla that showed a trend for depletion in patients with inflammatory bowel disease (**Fig. 5-10**). Similar but significant associations were previously reported between the salivary Bacteroidetes, Proteobacteria and inflammatory bowel disease[16].

At an increased resolution, interestingly, we observed inflammatory bowel disease-associated oral microbial dysbiosis in a site- and taxa-specific manner. For example, relative abundance of the order Bacillales in the phylum Firmicutes, showed significant reduction in inflammatory bowel disease patients compared to healthy controls in both plaque ($P = 0.015$) and buccal ($P = 0.019$) microbiotas; however, this

difference did not reach significance in the other two oral sites (**Fig. 5-11**). Similarly, inflammatory bowel disease-associated depletion of Carnobacteriaceae, a family in the phylum Firmicutes, was confined to plaque ($P = 0.007$) and saliva ($P = 0.013$) samples (**Fig. 5-11**). Overall, based on the number of bacterial groups that showed significant differences (FDR < 0.05) between inflammatory bowel disease cases and controls, the degree of the disease-associated dysbiosis was relatively severe in plaque compared to other three oral sites (**Fig. 5-11**). On the other hand, microbes that depicted inflammatory bowel disease-associated dysbiosis are predominantly members of the phylum Firmicutes (**Fig. 5-11**). Notably, perturbed Firmicutes and Actinobacteria abundances has long been implicated in inflammatory bowel disease. For instance, inflammatory bowel disease has been shown to be associated with an overall depletion of Firmicutes in disease-relevant intestinal mucosal biopsies as well as in stool samples[11, 35], whereas, Actinobacteria were reported to be substantially more abundant in inflammatory bowel disease compared to healthy controls[36]. In agreement, we found depletion of the members of the phylum Firmicutes and enrichment of the members of the phylum Actinobacteria (family Actinomycetaceae and genus *Actinomyces*) in the oral microbiota of inflammatory bowel disease patients (**Fig. 5-11**).

On the other hand, as expected, several bacterial groups in fecal microbiota showed correlation with disease phenotype. However, to our surprise, most of the inflammatory bowel disease-associated microbial signal from stool was either lost or trended in the opposite direction in oral samples (**Fig. 5-12**). We performed these differential abundance analyses on the other two data sets, with 462 OTUs and 268 OTUs, and obtained similar results (data not shown). Collectively, these data highlight the importance of site- and taxa-specific dysbiosis in inflammatory bowel disease. Statistical significance was evaluated with a permutation test (number of permutations = 10,000) which is then corrected for multiple comparisons using False Discovery Rate (FDR).

Oral microbiota can differentiate inflammatory bowel disease subjects from healthy controls

To ascertain if the oral microbiota could distinguish inflammatory bowel disease patients from healthy controls, we used random forest classifier on the first available sample from each subject and compared its

prediction accuracy to stool. Details of the number of samples per group and per body site are given in **Supplementary Table 5-1C**. Our classifier for inflammatory bowel disease in oral samples attained an average area under the ROC curve (AUC) ranging from 0.652 to 0.726 depending on the location, suggesting that oral microbial composition across all four profiled sites has the potential for distinguishing inflammatory bowel disease subjects from healthy controls (**Table 5-1**). When compared amongst sites, saliva performed best, closely matched by the stool samples. In one hundred random splits of the data between training and test sets, our classifier for inflammatory bowel disease, attained an average AUC of 0.726 for saliva versus an average AUC of 0.669 for stool (**Table 5-1**). One such split of the data was shown in **Fig. 5-13**. Other oral sites examined, buccal mucosa (AUC = 0.703), plaque (AUC = 0.667) and tongue (AUC = 0.652) were also comparable to stool in classification accuracy (**Table 5-1**). Furthermore, despite limited sample size, we were able to make a classifier from oral microbiotas that sorted both Crohn's disease and ulcerative colitis samples from healthy controls (**Table 5-1**). It is worth noting that a significant proportion of our subjects were on concomitant therapy during the enrollment sampling which may have affected bacterial communities when compared to treatment naïve subjects. However, medication should have had systemic effects on microbial composition and hence we assume no significant bias was introduced since we are only interested in comparing the diagnostic utility of oral microbiota to stool. Nevertheless, our results suggest that oral samples, saliva in particular, can differentiate inflammatory bowel disease subjects from healthy controls and it may be used as a surrogate to diagnose or monitor the presence of inflammatory bowel disease. This finding was robust to other QC criteria (**Supplementary Table 5-1F**).

Longitudinal trajectory of oral microbiota

Cross-sectional studies have shown inflammatory bowel disease to be associated with site-specific dysbiosis, reduced diversity and species richness[11, 37-40]; however, the longitudinal trajectory of these disease-associated changes has not been thoroughly investigated. Understanding the dynamic behavior of microbiota is crucial to elucidate the mechanistic basis of microbial contributions to human health and for

the advancement of microbiome-based diagnostic and therapeutic interventions. To this end, we examined the longitudinal trajectories of the relative abundance of individual microbial members across sites from subjects with at least three over time samples (n = 19; 11 Crohn's disease, 5 ulcerative colitis and 3 healthy controls). We observed two general patterns: 1) global stability and 2) global variability. The global stability group included microbial members whose relative abundance remained fairly stable during the course of the follow-up period across individuals, irrespective of disease status. Microbial organisms belonging to the phyla Bacteroidetes, Fusobacteria, and Proteobacteria were among the globally stable group, across sites, including saliva, tongue and plaque; however, this pattern appears to be disrupted in buccal samples (**Fig. 5-14a to c** and **Figs. 5-15 to 17**). On the other hand, the global variability group included members whose relative abundances displayed inter- and intra-individual variability patterns, intermittently disappearing and reappearing over time. Among these were phyla Firmicutes, SR1 and Actinobacteria (**Fig. 5-14d to f** and **Figs. 5-15 to 17**). Interestingly, previous analysis of over time samples has reported Firmicutes as more temporally dynamic within the gut microbiomes of individuals[29]. Collectively, the longitudinal trajectory findings from oral sites support the view that composition of some microbial organisms is more consistent over time while others exhibit relatively frequent transitions across subjects, irrespective of disease status. Identifying individual microbes that remain fairly consistent over time in controls while exhibiting transitions in inflammatory bowel disease would make great candidates for future microbiome-based functional studies; however, our efforts to this end were thwarted by limited control subjects with over time samples.

DISCUSSION

To our knowledge, this is the first investigation to characterize the spatial and temporal dynamics of the oral microbiota as it relates to inflammatory bowel disease. To obtain an integrated view of the spatial and temporal distribution of the oral microbiota, we examined bacteria from four anatomically different sites within the oral cavity. Our findings confirm that, although a candidate set of bacterial members are shared between all oral sites, each site harbored a characteristic microbiota, differing in both composition and

diversity, that persisted longitudinally. We further noticed that oral microbiota in inflammatory bowel disease patients featured site- and taxa-specific dysbiosis. Interestingly, oral microbiota, salivary microbial structure in particular, performed similar, if not better, than fecal microbiota in discriminating inflammatory bowel disease patients from healthy controls.

The primary goal of this study was to define and understand the spatial and temporal dynamics of oral microbial composition in inflammatory bowel disease, as an effort aimed to subsequently test whether site-specific oral microbial dysbiosis can be used as a surrogate marker of inflammatory bowel disease *in lieu* of or in addition to stool. To this end, we analyzed microbiota samples from a wide array of oral microenvironments including saliva, tongue, plaque and buccal mucosa. Despite limited information, studies have demonstrated oral microbial differences in inflammatory bowel disease implicating the potential of the oral microbiome in diagnosing and monitoring patients with inflammatory bowel disease. However, these studies were limited to a specific region in the oral cavity, which is mainly chosen based on the ease of obtaining or availability or mostly at random. For instance, Said *et al*., profiled salivary microbiota and identified microbial dysbiosis in inflammatory bowel disease patients and healthy controls[16], whereas, Docktor *et al*., employed tongue and buccal samples and reported overall reduced tongue microbial diversity in pediatric Crohn's disease subjects compared to healthy controls[15]. On the other hand, Kelsen *et al*., demonstrated Crohn's disease associated microbial dysbiosis using subgingival plaque samples[14]. Surprisingly, when analyzed at the population level, although several changes were reported independently from one or more oral sites, there is no consistent pattern of change between inflammatory bowel disease cases and controls which could, at least partly, be attributed to baseline differences in composition and diversity based on the region of the mouth sampled. As expected from previous work[29, 30], we noted anatomical location as the strongest driver of microbial composition within the oral cavity, which supports the view that it is critical to define and understand baseline spatial differences with respect to region of the oral cavity profiled in order to interpret disease-associated dysbiotic states and eventually to gain insights into disease etiology.

Next we surveyed oral microbial samples for inflammatory bowel disease associated changes with respect to the first available sample, cross-referencing with fecal samples, collected from the same subjects whenever possible. Fecal microbial dysbiosis has long been regarded to proxy intestinal mucosal microbiota and has been shown to be associated with inflammatory bowel disease. For instance, we and others have previously shown that reduced overall diversity, richness and altered abundances of fecal bacterial taxa are associated with inflammatory bowel disease or one of its two main forms, and that it can accurately classify inflammatory bowel disease patients from healthy controls[10, 11]. Alternatively, inflammatory bowel disease, being a systemic disease, may have effects on microbial community structure that is not restricted to the gastrointestinal system alone. Previous studies have shown alterations in the oral microbiome of systemic diseases, where oral microbiome has been suggested to play a role in systemic health through immune regulation, nutrition absorption and metabolism[41]. In line with these evidence, we hypothesized that oral microbiota may depict inflammatory bowel disease-associated changes. When assessed at the global level or at the population level, oral samples reflected dysbiosis in inflammatory bowel disease cases and healthy controls which are, surprisingly, site- and taxa-specific. For instance, at the population level, saliva, plaque and buccal samples depicted inflammatory bowel disease-associated dysbiosis of certain taxa at different taxonomic levels, while no significant inflammatory bowel disease-associated differences at any of the examined taxonomic levels, phyla, class, order, family and genus were noticed in tongue samples. On the other hand, we identified taxa that were perturbed between cases and controls in one or more, but not all four studied oral microenvironments. For example, abundances of the family Carnobacteriaceae was negatively associated with inflammatory bowel disease in saliva and plaque microbiotas, but not in buccal and tongue samples, while depletion of the family Bacillales_Incertae_Sedis_XI in inflammatory bowel disease was limited to plaque samples. Alternatively, there were taxa that depicted differences or trends in relative abundance between inflammatory bowel disease patients and healthy controls which are consistent across the four profiled oral sites. For example, at the phyla level, Actinobacteria, Bacteroidetes and Spirochaetes were enriched in inflammatory bowel disease across all four oral sites, whereas Fusobacteria, Firmicutes and Proteobacteria were among the phyla that are depleted in patients with inflammatory bowel

disease. In contrast, phyla Fusobacteria and Proteobacteria are significantly enriched in inflammatory bowel disease fecal microbiotas, while Actinobacteria showed a trend for depletion. Such striking differences between oral and stool microbiotas in relation to inflammatory bowel disease were also evident at the other taxonomic levels including class, order and family. These polarizing shifts in relative abundances between stool and oral microbiotas could, at least partly, contributing to distinct clustering between habitats as evidenced from principal coordinate analysis of Bray-Curtis dissimilarity. Collectively, our data is compatible with a differential effect of inflammatory bowel disease depending on taxa and sample type, which warrants future investigation. Nevertheless, we demonstrated for the first time that the oral microbiota has discriminatory power for classifying inflammatory bowel disease subjects from healthy controls, regardless of the location and surprisingly, salivary microbiota performed even better than stool, which is widely believed to hold the potential of noninvasive diagnostic approach. Given the fact that oral samples are significantly easier to obtain than fecal samples, and less invasive than intestinal biopsies, this creates an opportunity to use oral microbial sampling approach to diagnose and monitor patients with inflammatory bowel disease. It would be of tremendous importance in the future to investigate whether oral microbiota can discriminate Crohn's disease from ulcerative colitis and its usefulness in monitoring or predicting treatment effects. Identification of oral microbial biomarkers that predict changes in disease flares and the risk of developing disease associated complications may help identify patients that are at high risk, and may facilitate preemptive treatments.

There is tremendous interest for using measurements of the microbiome as a means to diagnose and improve different aspects of human health. To this end, understanding the dynamic behavior of microbiota as it relates to the trait of interest, is critical to be able to translate disease-associated microbiome measurements into the clinical setting. Herein, we contribute to addressing this knowledge gap by analyzing the longitudinal trajectories of the oral microbiome, in the context of inflammatory bowel disease. Our findings indicate that the composition of some taxa are more consistently perturbed than others. For instance, abundances of the phyla Bacteroidetes, Fusobacteria, and Proteobacteria remained relatively stable over

time, whereas Firmicutes, SR1 and Actinobacteria showed drastic changes, disappearing and reappearing intermittently. This finding indicates that making conclusions based on single time point microbiome measurements in case-control studies is problematic, especially when aiming for the identification of disease-specific microbial candidates. We noticed these patterns of global stability or variability across the subjects irrespective of their disease state. Future studies are warranted to identify taxa that are potentially pathogenic by selecting for those that stay stable over time in controls, but exhibit striking shifts in inflammatory bowel disease patients with respect to changes in disease flares, severity and treatment effects. Our efforts to this end, to make biologically meaningful observations were thwarted by limited number of healthy controls with longitudinal follow-up samples.

Our study has several limitations. We did not have any measures pertaining to diet, oral health status, duration since the last oral hygiene, and stages of dentition, all of which are particularly relevant to oral microbial composition. We do acknowledge that our population size and study design (to include established inflammatory bowel disease subjects) is sub-optimal. Although, we were primarily interested in comparing the structural composition and diagnostic potential of microbiota of various sites within the same subjects, it is possible that treatment may have had site-specific effects on microbial composition, which could potentially influence our findings. Nevertheless, our findings highlight the proposition that oral microbial surveillance can serve as a diagnostic marker to discriminate inflammatory bowel disease patients from healthy controls. Given the fact that obtaining oral samples is significantly easier than stool and intestinal biopsies, this creates an opportunity to perform microbiome-based studies in larger cohort sizes, preferentially in a longitudinal fashion. Our findings also highlight the importance of understanding baseline spatial differences within the oral microenvironments in order to interpret disease-associated changes. Given the differences and directional inconsistency between stool and oral microbiotas, fueled by our previous observation that stool is not a perfect reflection of intestinal microbial community structure[11], it would be of value to perform side by side investigation of intestinal biopsy, stool and oral microbiotas in future studies of the role of the microbiome in inflammatory bowel disease.

Table 5-1: ROC analysis of the site-specific microbiotas for the classification of inflammatory bowel disease or Crohn's disease or ulcerative colitis patients from healthy controls

SITE	IBD - AUC (mean)	IBD - AUC (SD)	CD - AUC (mean)	CD - AUC (SD)	UC - AUC (mean)	UC - AUC (SD)
SALIVA	0.726	0.106	0.694	0.131	0.751	0.128
BUCCAL	0.703	0.120	0.660	0.113	0.685	0.119
STOOL	0.669	0.110	0.639	0.113	0.744	0.136
PLAQUE	0.667	0.125	0.696	0.132	0.654	0.137
TONGUE	0.652	0.110	0.647	0.103	0.730	0.139

mean = average of 100 random splits between the respective group and healthy controls; *SD* = *standard deviation*

Figure 5-1: Illustrative time series for each subject are shown per site. Time series (in days) is presented on the *x* axis and subject ID on the *y* axis.

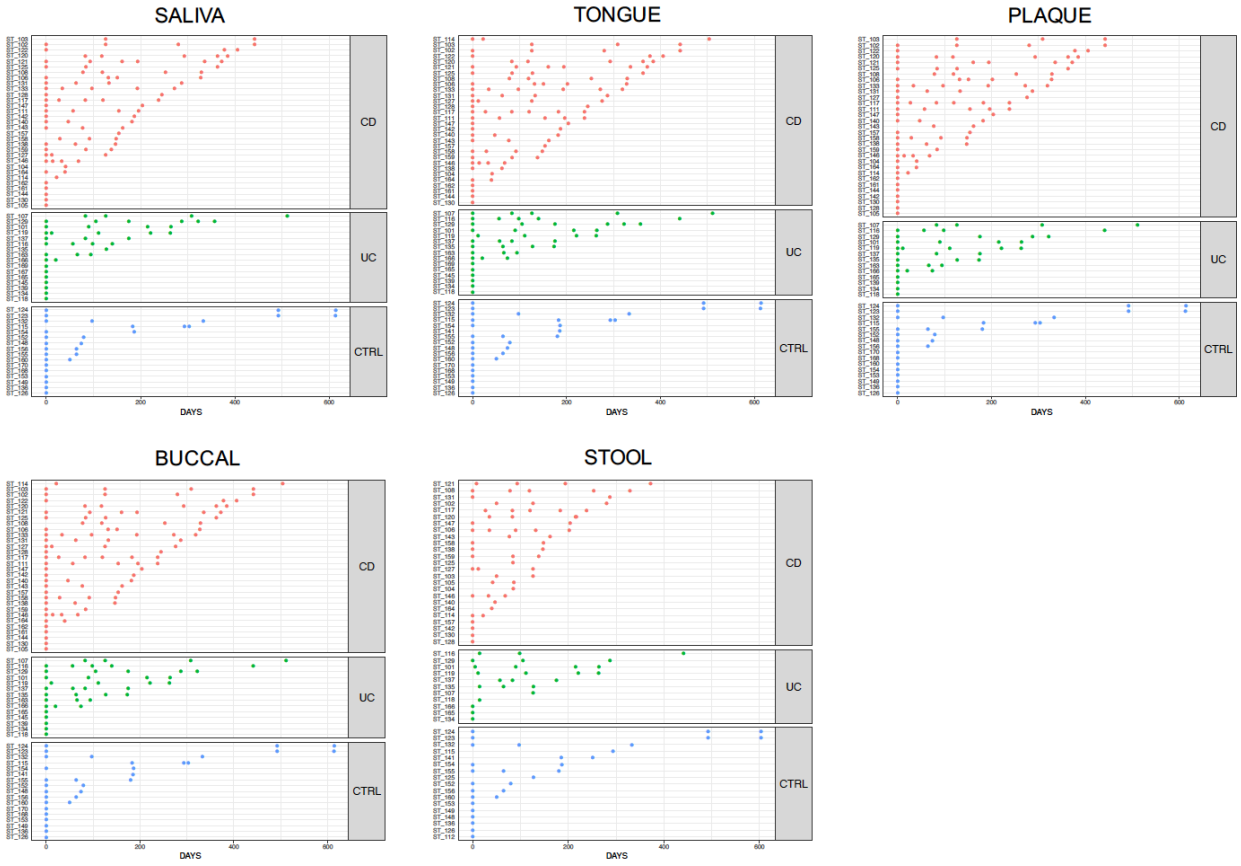


Figure 5-2: Boxplots displaying median and quartiles of total read counts across sites.

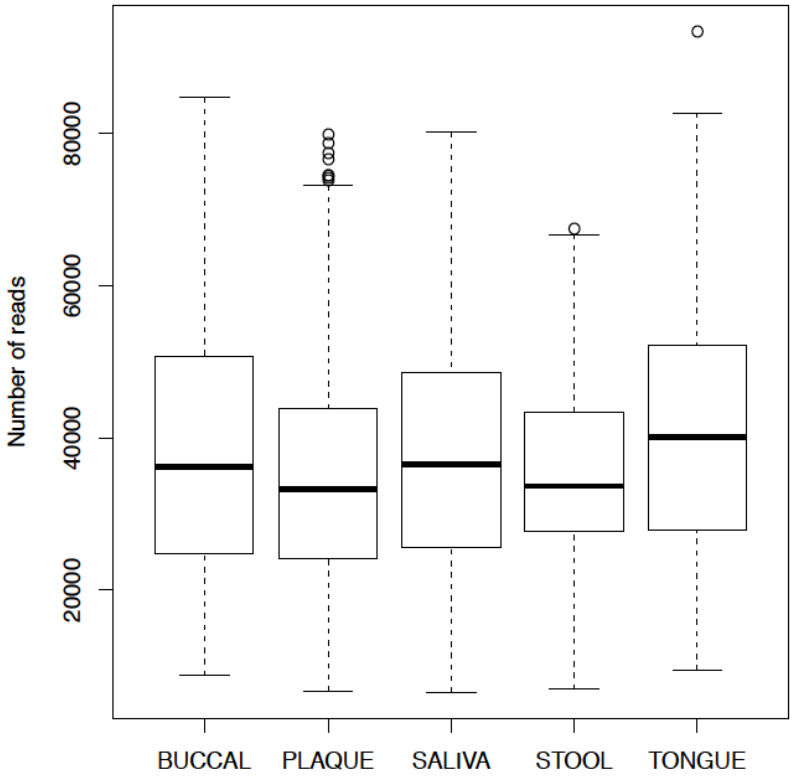


Figure 5-3: Rarefaction curves showing the relationship between the total read counts and microbial species richness across sites, collectively and individually. Each curve represents a different sample. The number of samples for each plot are presented at the top.

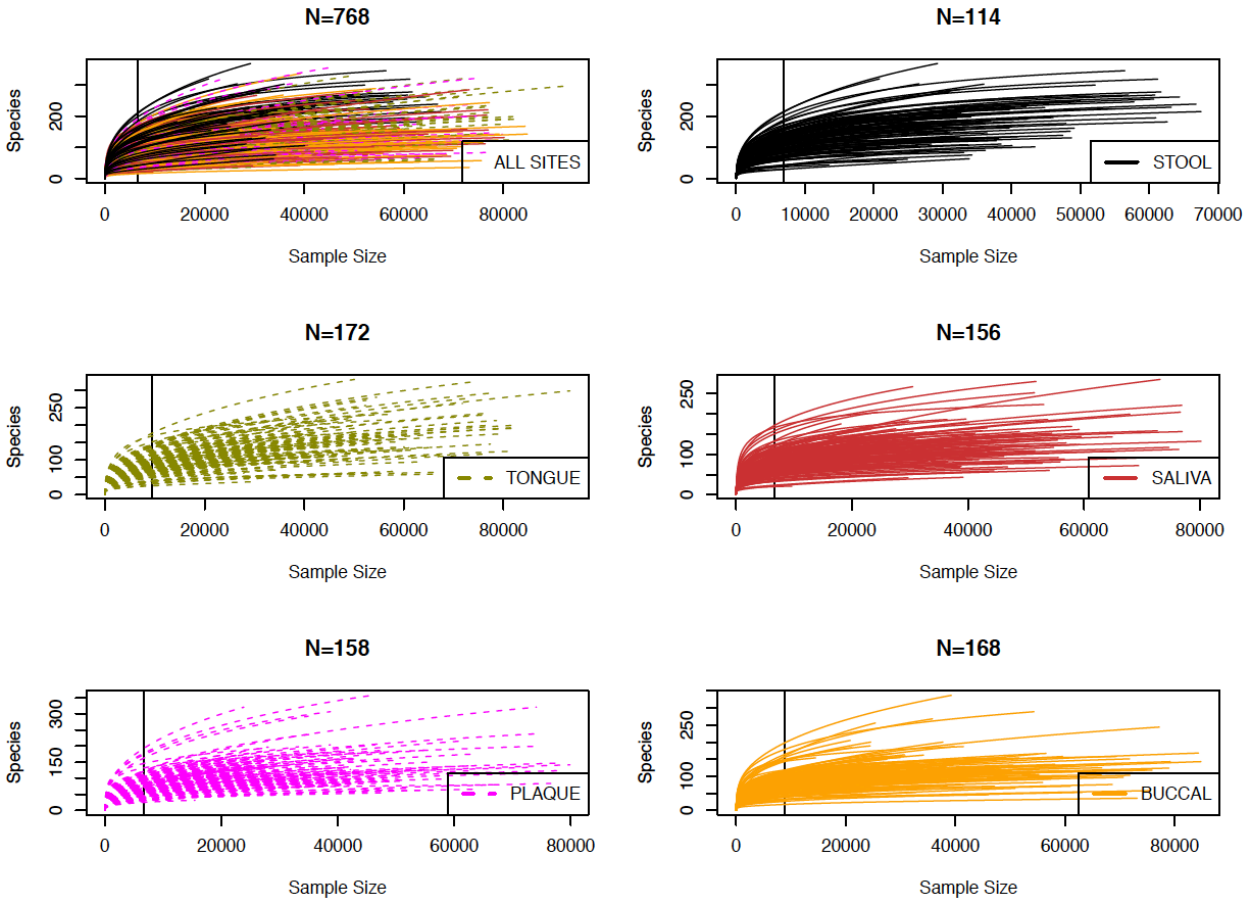


Figure 5-4: Overall microbial community structure, diversity and richness across sites. (a) Principal Coordinate Analysis (PCoA) of microbial community structure using Bray-Curtis distance. Each dot on the PCoA plot corresponds to a sample colored by either anatomical location, disease status or collection time point. The percentage of variation explained by the plotted principal coordinates is indicated on the axes. (**b**) Overall microbial diversity across sites as measured by the Shannon diversity index. (**c**) Overall microbial richness as measured by the Chao1 index. Data shown here was obtained from all the available samples from each site (156 saliva, 172 tongue, 158 plaque, 168 buccal mucosa and 114 stool samples).

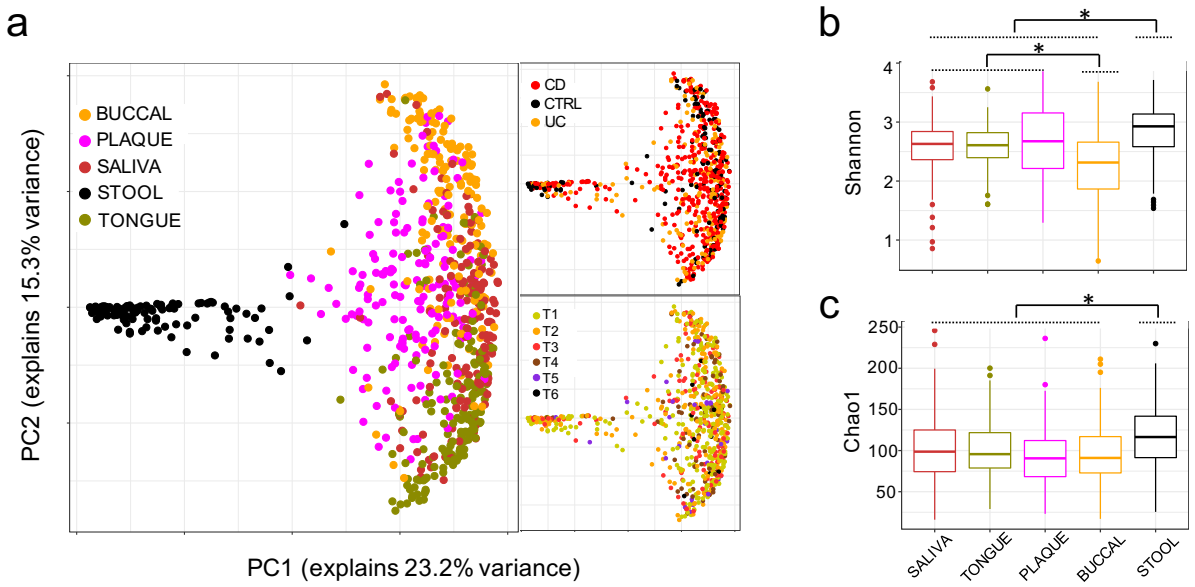


Figure 5-5: Principal Coordinate Analysis (PCoA) of oral microbial community structure using Bray-Curtis distance. Each dot on the PCoA plot corresponds to an oral sample colored by anatomical location. The percentage of variation explained by the plotted principal coordinates is indicated on the axes.

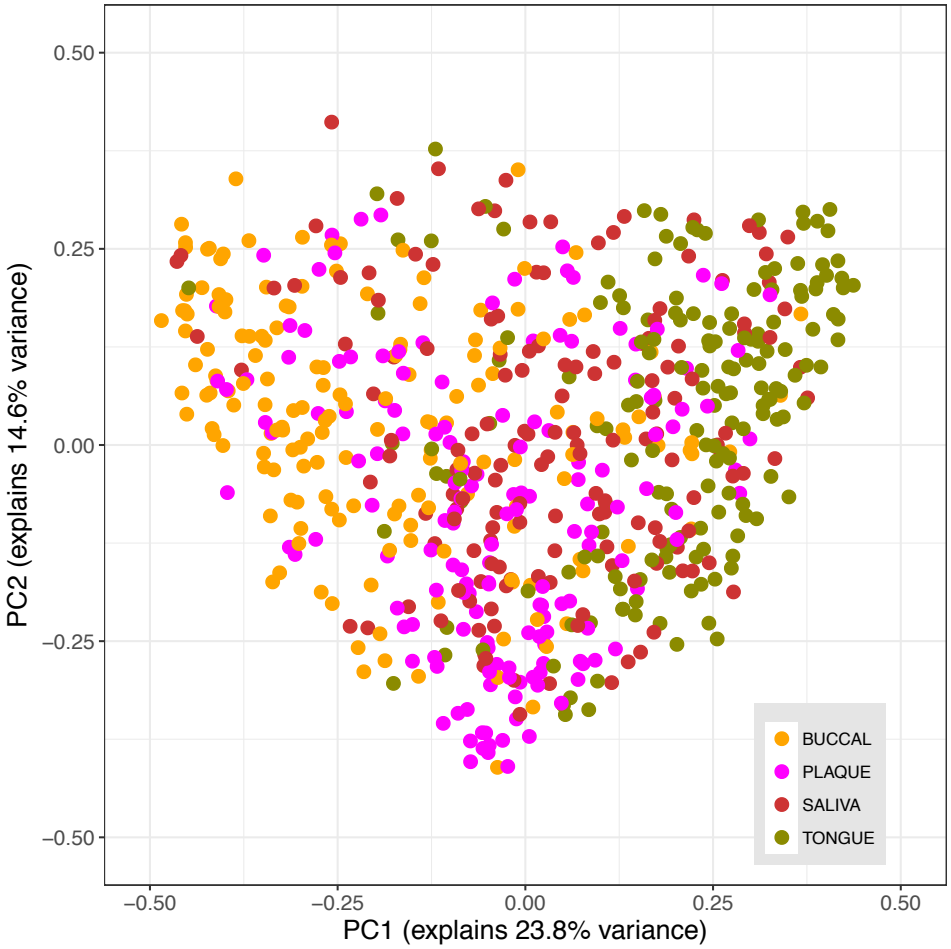


Figure 5-6: Relative abundances of bacterial groups across sites. Packed bubble graphs depicting the relative abundances of bacterial groups at the level of the phylum. Each bubble in the plot represents a single phylum with the size of the bubble corresponding to the relative abundance within the microbiota of each individual site. Phyla that showed profound shifts between stool and oral microbiotas were indicated. Data shown here was obtained from the first available sample per subject from each site.

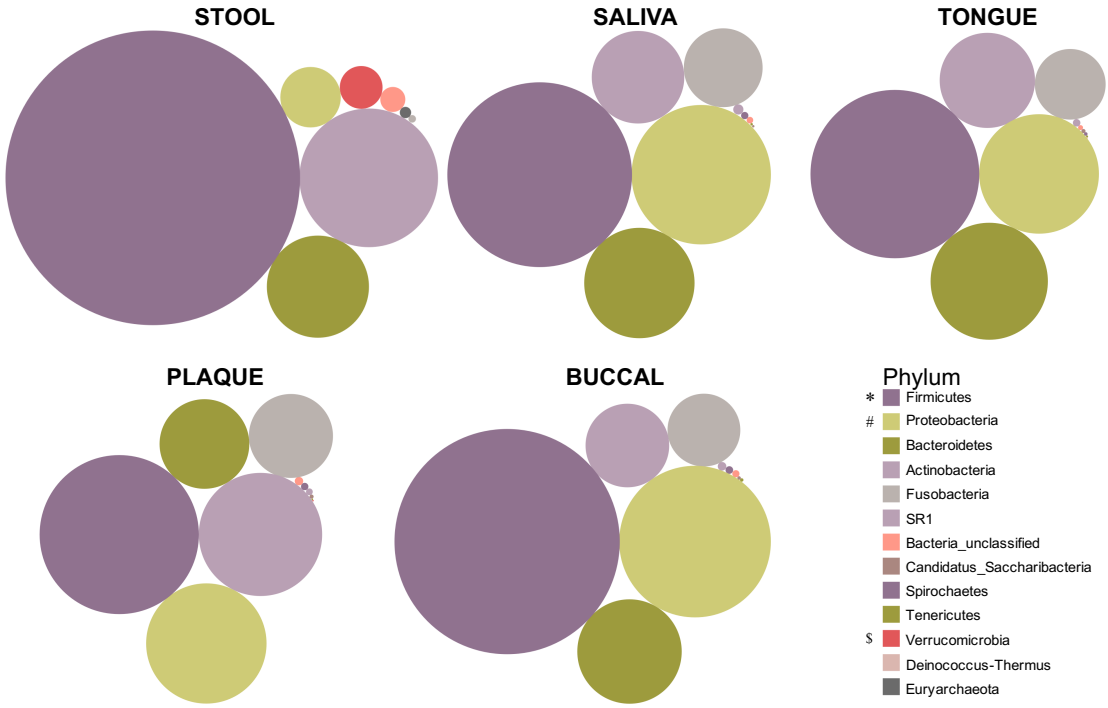


Figure 5-7: Relative abundances of bacterial groups across sites. At the family taxonomic level, relative abundances represented by the size of the bubble were shown for each site. Family names were presented in the order from highest abundance to lowest in saliva samples. Members of the phylum Firmicutes were indicated with the asterisk and the Proteobacteria by the number (#). Data shown here was obtained from the first available sample per subject from each site.

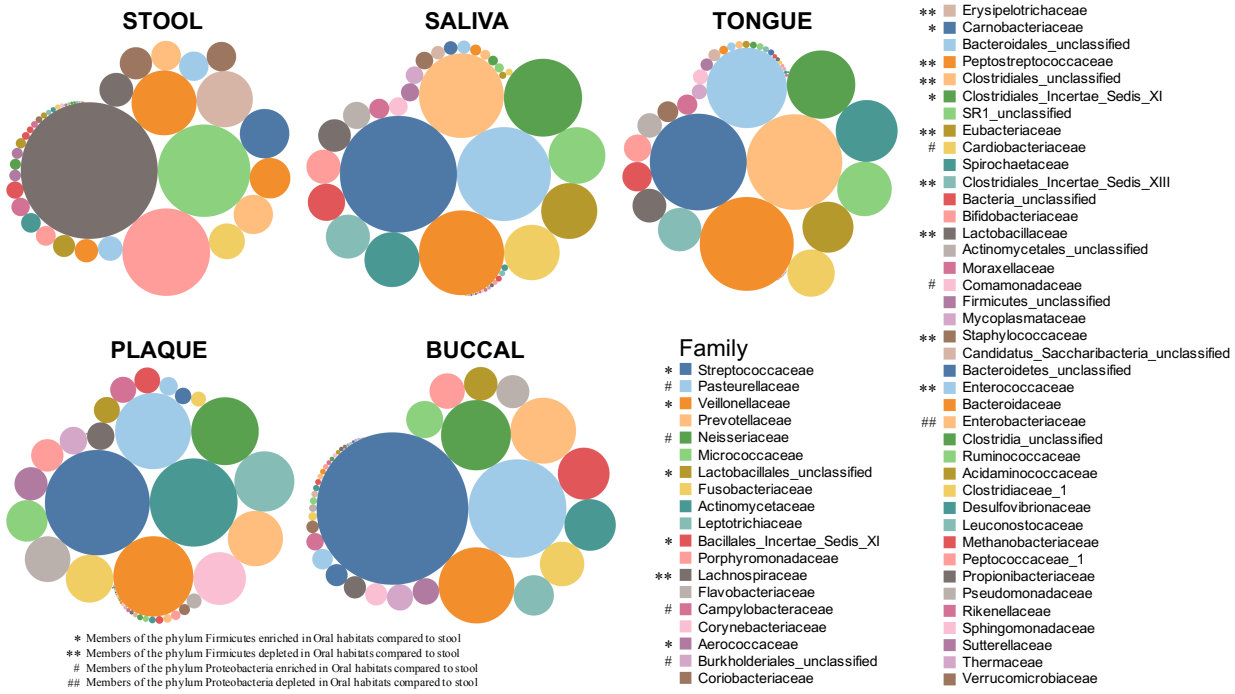


Figure 5-8: Site- and taxa-specific oral microbial dysbiosis in inflammatory bowel disease. PCoA of microbial community structure using Bray-Curtis distance. Each dot on the PCoA plot corresponds to a sample colored by disease status. The percentage of variation explained by the plotted principal coordinates is indicated on the axes.

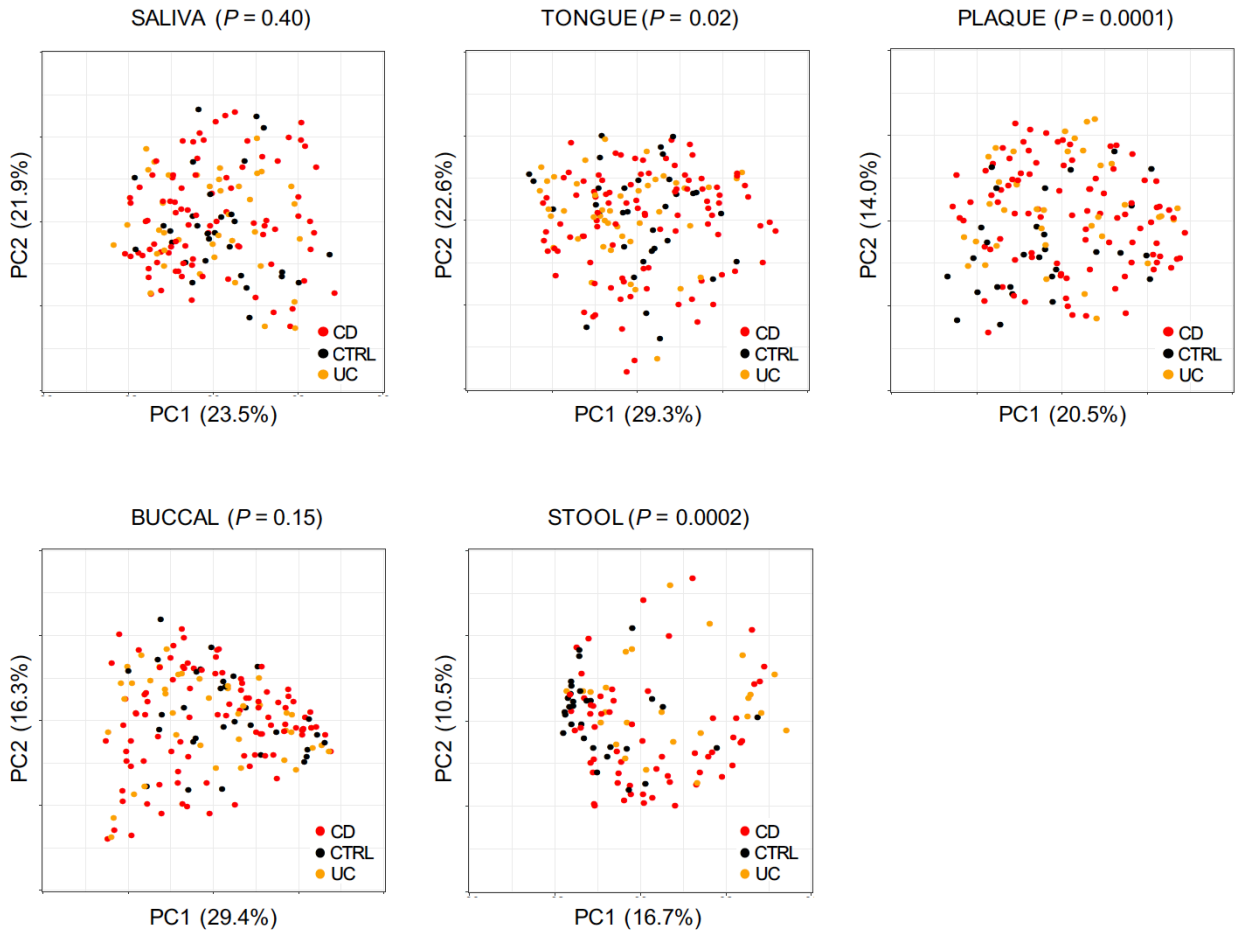


Figure 5-9: Differences in the overall microbial diversity and richness of fecal and salivary microbiotas between cases (Crohn's disease or ulcerative colitis) and controls (CTRL). (**a**) Overall fecal microbial diversity between the groups as measured using different indices, Shannon, Alpha, and Simpson, which exhibit different sensitivities. Overall richness estimates between groups were obtained using the Chao1 method. (**b**) Overall diversity and richness between groups in saliva samples. ANOVA *P* values and the number of samples (all the available samples from each site) per group are presented.

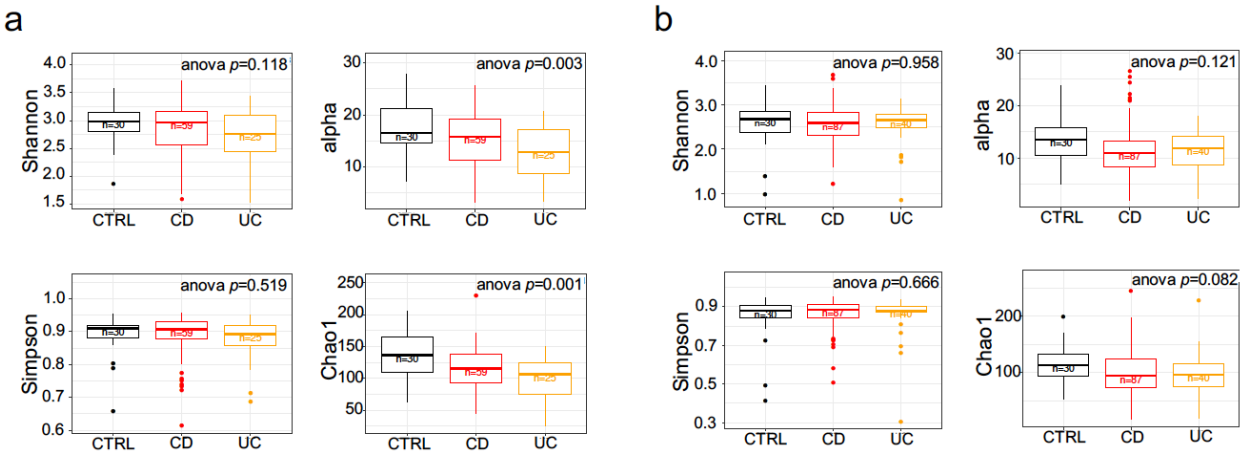


Figure 5-10: Inflammatory bowel disease-associated shifts at the phylum level that are consistent across the four profiled oral sites (**a**) saliva, (**b**) tongue, (**c**) plaque and (**d**) buccal mucosa. Phyla that displayed trends in relative abundance between inflammatory bowel disease patients and healthy controls at baseline (first available sample from each subject). Bars in turquoise represent phyla that showed a trend of enrichment in inflammatory bowel disease compared to healthy controls. Bars in red corresponds to phyla that showed a trend of depletion in inflammatory bowel disease. The $\log_2$ fold changes in the relative abundance of each phylum were shown on the *x* axis.

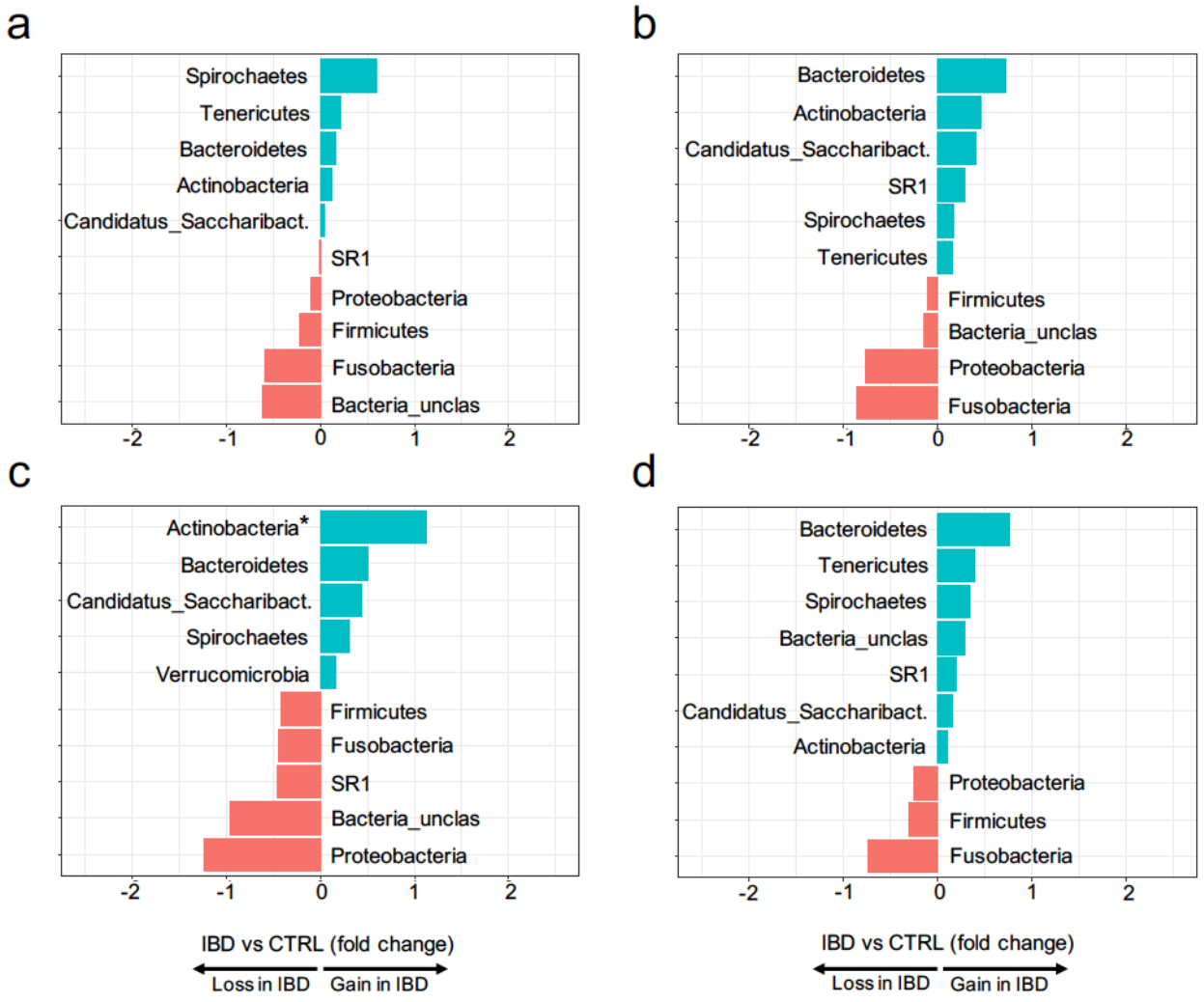


Figure 5-11: Site- and taxa-specific oral microbial dysbiosis in inflammatory bowel disease. Individual microbial members that demonstrated significant differences between inflammatory bowel disease cases and healthy controls at baseline (first available samples from each subject). Relative abundances are shown on the *x* axis (log_{10} scale). FDR adjusted *P* values (P) demonstrating the associations between microbial abundances and disease phenotype were presented. Members of the phylum Firmicutes were represented by the asterisk.

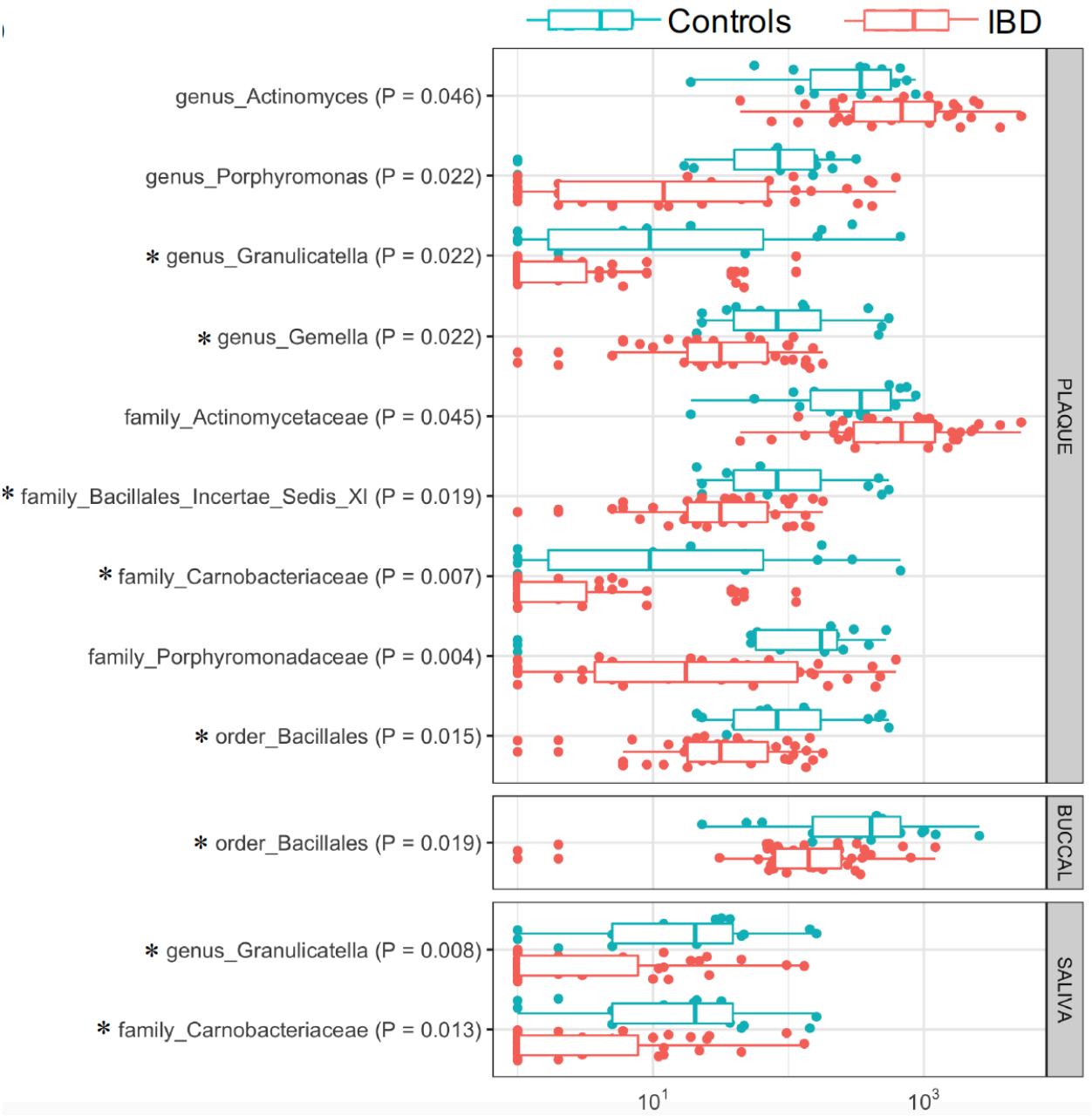


Figure 5-12: Directional inconsistency in inflammatory bowel disease-associated microbial signatures between stool and oral microbiotas. inflammatory bowel disease-associated taxa from baseline stool samples (FDR < 0.2) were selected to compare and contrast the direction of change between stool and oral microbiotas at baseline. Data from all the five profiled sites are presented, at various taxonomic levels, (**a**) phylum, (**b**) class, (**c**) order, and (**d**) family. Bars to the right of the dotted line at the center of each plot represent bacterial groups that are enriched in inflammatory bowel disease cases compared to healthy controls. Bars to the left corresponds to taxa that are depleted in inflammatory bowel disease. The $\log_2$ fold changes in the relative abundance of each taxon between cases and controls were shown on the *x* axis.

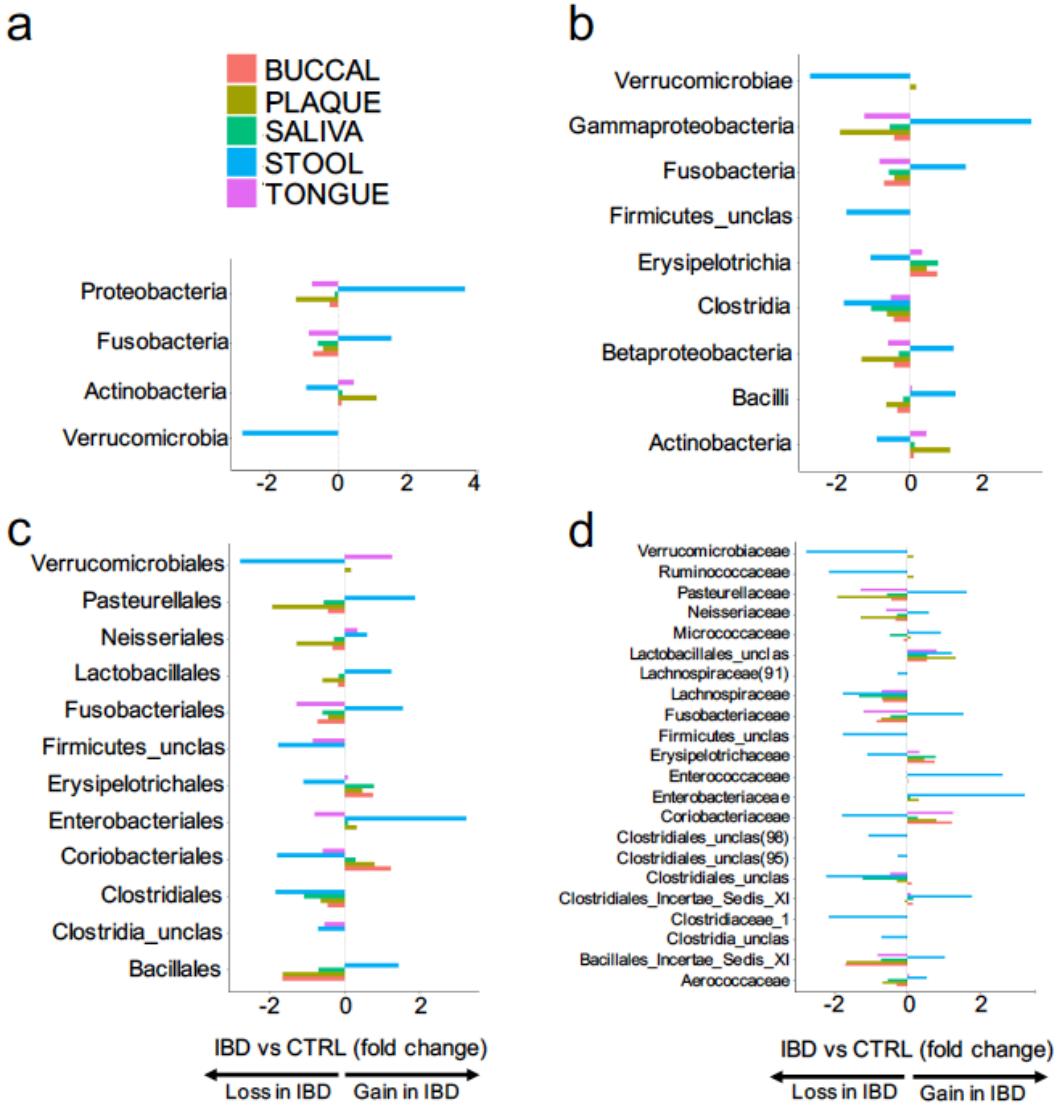


Figure 5-13: Performance of microbiome-based random forest classifiers in differentiating inflammatory bowel disease patients from healthy controls. Receiver operating characteristic (ROC) curves of baseline (first available) salivary (47 inflammatory bowel disease, 16 healthy controls) and fecal (36 inflammatory bowel disease, 16 healthy controls) microbiotas were plotted to differentiate inflammatory bowel disease patients from healthy controls. The area under the curve (AUC) of salivary (red line) and fecal (black line) microbiotas are indicated. A perfect classifier would have an AUC of 1, and a random classifier would score 0.5.

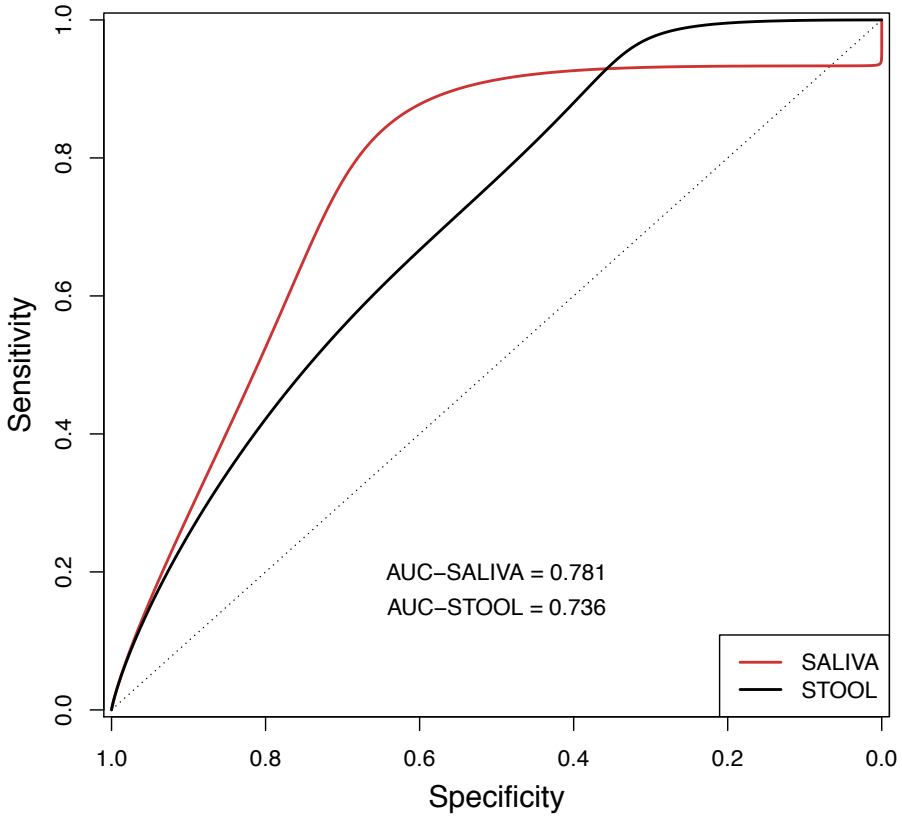


Figure 5-14: Temporal dynamics of the selected taxa in salivary microbiome. **(a-c)** Global stability group consisting of the phyla that remained fairly stable over time across subjects. Heat map of the relative abundances of the selected taxa for 19 subjects with at least three over time samples are shown. Each row represents the relative abundance (*x* axis, $\log_{10}$ scale) of a particular phylum across 3 or more consecutive time points from a particular subject. The subjects consist of 11 Crohn's disease (dark green bar on the left), 5 ulcerative colitis (light green) and 3 healthy controls (CTRL; light pink). Illustrative time series (in days) for each subject are shown in **Fig. 1**. (**d-f**) Global variability group consisting of the phyla that displayed inter- and intra-individual variability patterns, intermittently disappearing and reappearing over time.

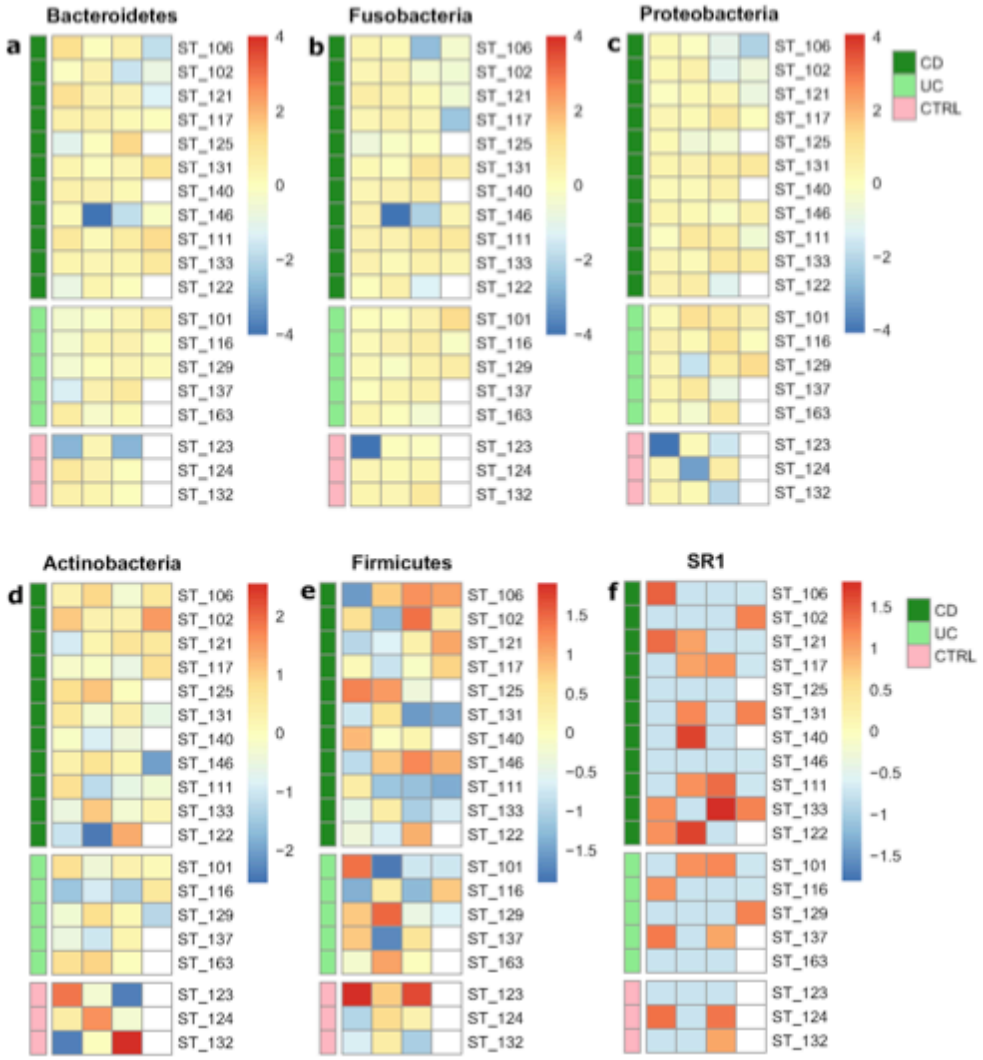


Figure 5-15: Temporal dynamics of selected taxa in tongue microbiota. (**a-c**) Global stability group consisting of the phyla that remained fairly stable over time across subjects. Heat map of the relative abundances of the selected taxa for 19 subjects with at least three over time samples are shown. Each row represents the relative abundance (*x* axis, $\log_{10}$ scale) of a particular phylum across 3 or more consecutive time points from a particular subject. The subjects consist of 11 Crohn's disease (dark green bar on the left), 5 ulcerative colitis (light green) and 3 healthy controls (CTRL; light pink). Illustrative time series (in days) for each subject are shown in **Fig. 5-1**. (**d-f**) Global variability group consisting of the phyla that displayed inter- and intra-individual variability patterns, intermittently disappearing and reappearing over time.

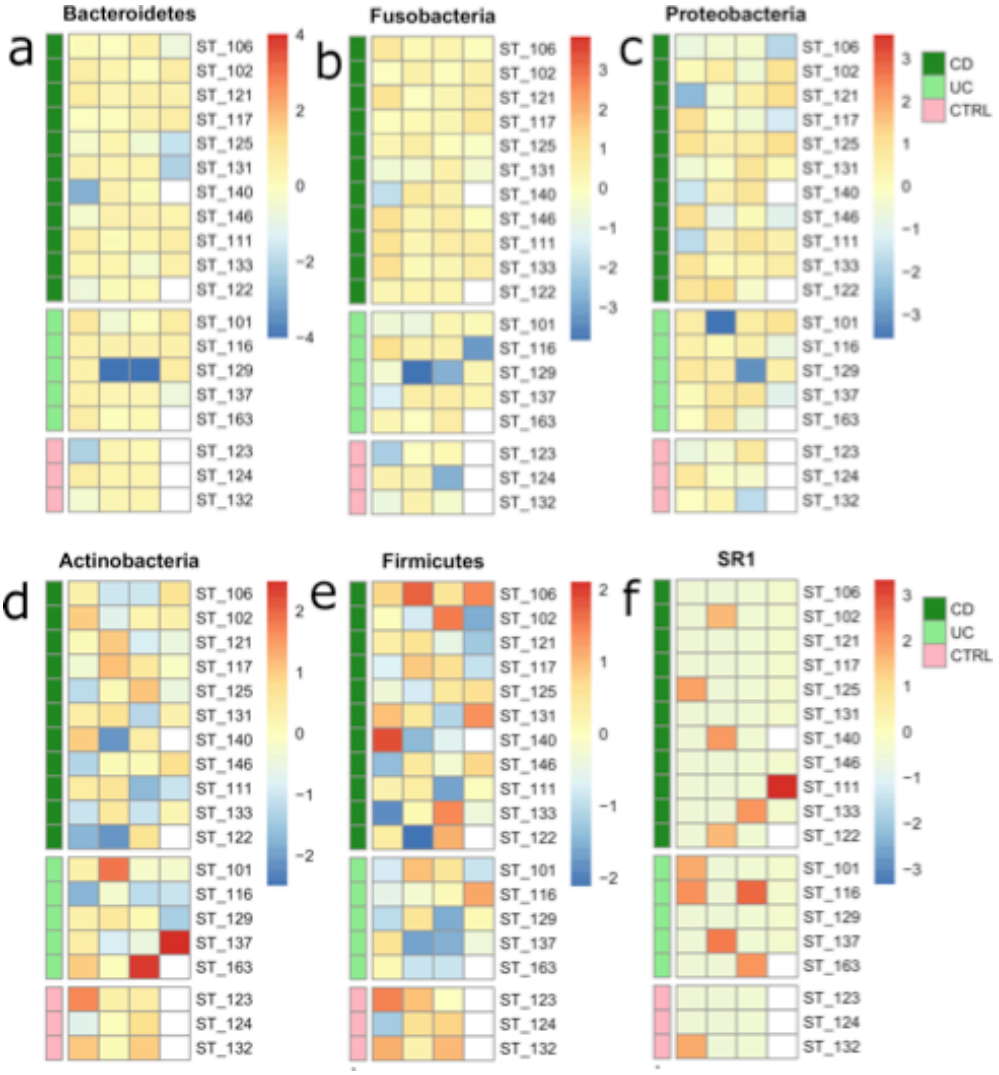


Figure 5-16: Temporal dynamics of selected taxa in plaque microbiota. (**a-c**) Global stability group consisting of the phyla that remained fairly stable over time across subjects. Heat map of the relative abundances of the selected taxa for 19 subjects with at least three over time samples are shown. Each row represents the relative abundance (*x* axis, $\log_{10}$ scale) of a particular phylum across 3 or more consecutive time points from a particular subject. The subjects consist of 11 Crohn's disease (dark green bar on the left), 5 ulcerative colitis (light green) and 3 healthy controls (CTRL; light pink). Illustrative time series (in days) for each subject are shown in **Fig. 5-1**. (**d-f**) Global variability group consisting of the phyla that displayed inter- and intra-individual variability patterns, intermittently disappearing and reappearing over time.

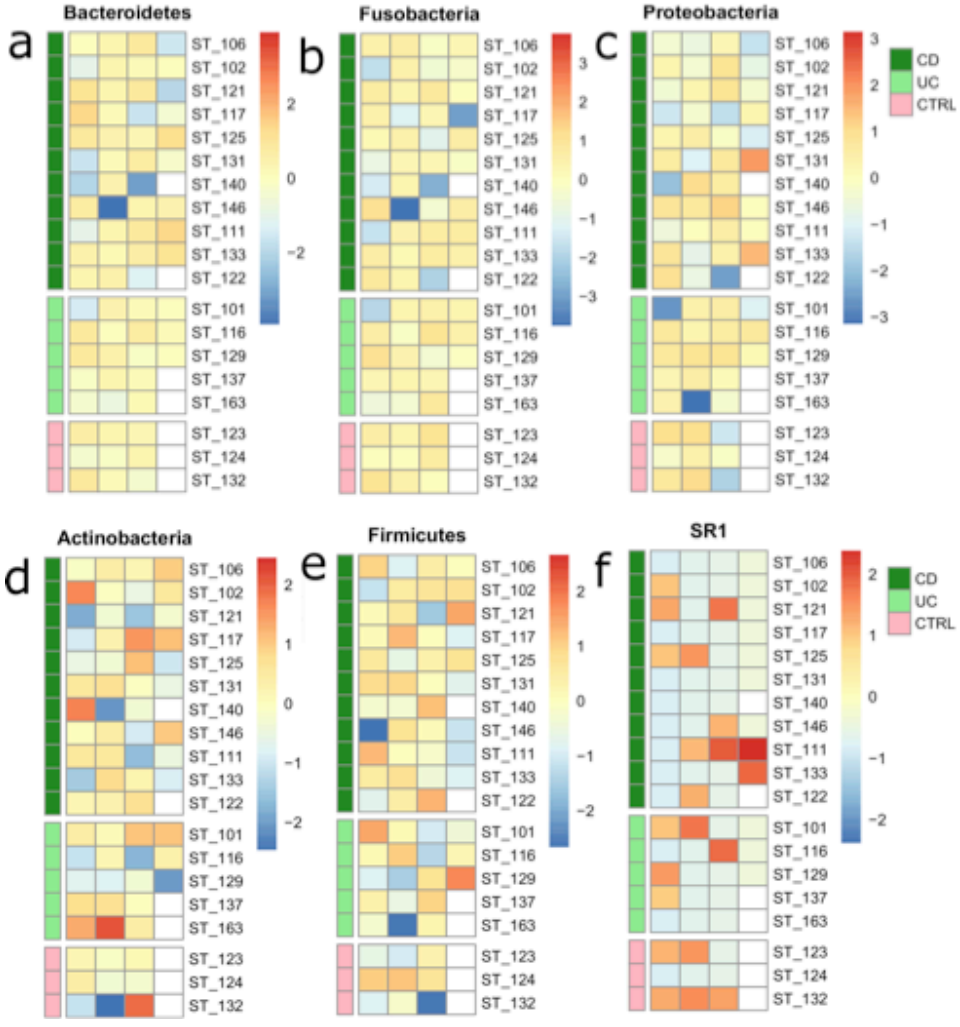
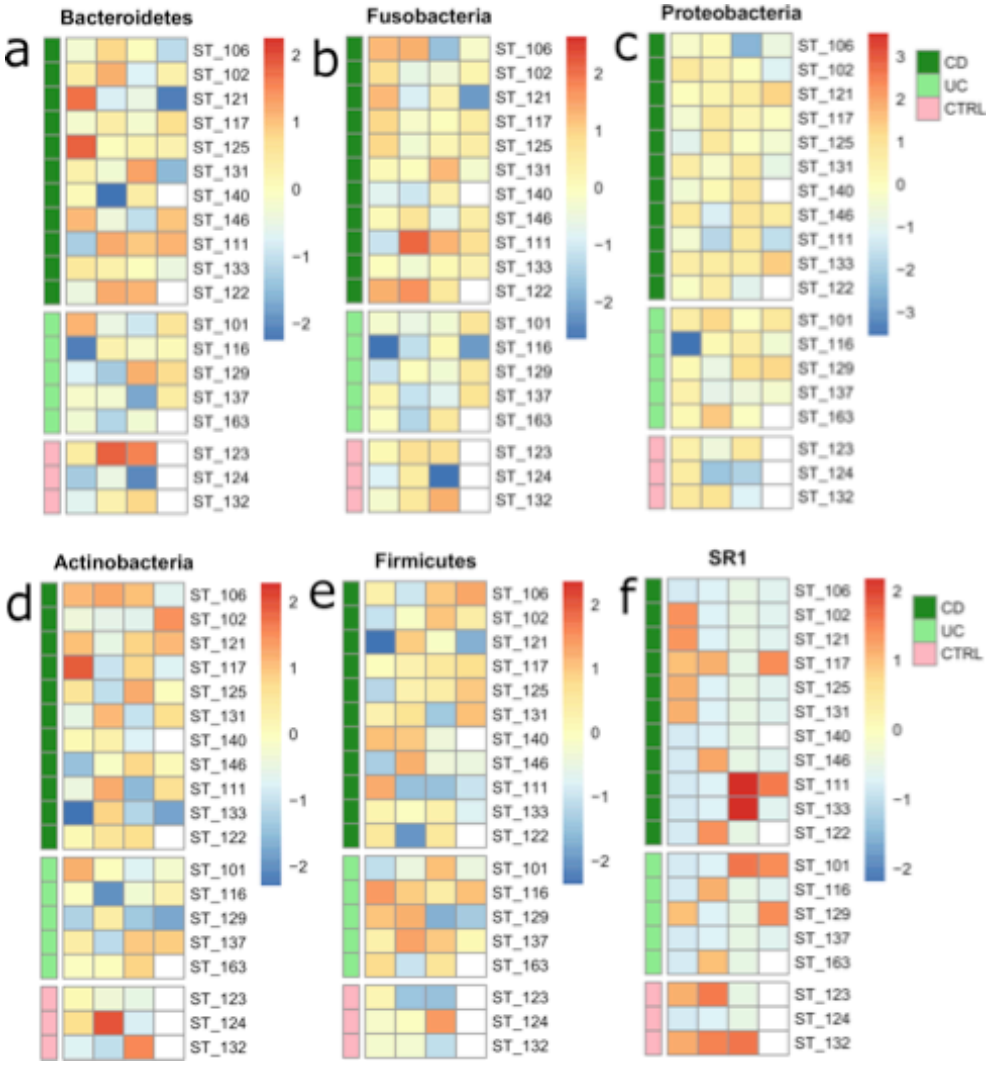


Figure 5-17: Temporal dynamics of the selected taxa in buccal mucosal microbiota. (**a-c**) Global stability group consisting of the phyla that remained fairly stable over time across subjects. Heat map of the relative abundances of the selected taxa for 19 subjects with at least three over time samples are shown. Each row represents the relative abundance (*x* axis, $\log_{10}$ scale) of a particular phylum across 3 or more consecutive time points from a particular subject. The subjects consist of 11 Crohn's disease (dark green bar on the left), 5 ulcerative colitis (light green) and 3 healthy controls (CTRL; light pink). Illustrative time series (in days) for each subject are shown in **Fig. 5-1**. (**d-f**) Global variability group consisting of the phyla that displayed inter- and intra-individual variability patterns, intermittently disappearing and reappearing over time.



Supplementary Tables

Supplementary Tables 5-1A to 5-1F can be accessed at the following dropbox link
https://www.dropbox.com/s/6hb9a7s0hatizn7/Chapter%205%20-%20Supplementary%20Table.xlsx?dl=0

REFERENCES

1. Burisch J, Jess T, Martinato M, et al. The burden of inflammatory bowel disease in Europe. J Crohns Colitis 2013;7:322-37.
2. Benchimol EI, Guttmann A, Griffiths AM, et al. Increasing incidence of paediatric inflammatory bowel disease in Ontario, Canada: evidence from health administrative data. Gut 2009;58:1490-7.
3. Benchimol EI, Mack DR, Guttmann A, et al. Inflammatory bowel disease in immigrants to Canada and their children: a population-based cohort study. Am J Gastroenterol 2015;110:553-63.
4. Molodecky NA, Soon IS, Rabi DM, et al. Increasing incidence and prevalence of the inflammatory bowel diseases with time, based on systematic review. Gastroenterology 2012;142:46-54.e42; quiz e30.
5. Herrinton LJ, Liu L, Lewis JD, et al. Incidence and prevalence of inflammatory bowel disease in a Northern California managed care organization, 1996-2002. Am J Gastroenterol 2008;103:1998-2006.
6. Kugathasan S, Denson LA, Walters TD, et al. Prediction of complicated disease course for children newly diagnosed with Crohn's disease: a multicentre inception cohort study. Lancet 2017;389:1710-1718.
7. de Lange KM, 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. Liu JZ, van Sommeren S, Huang H, et al. Association analyses identify 38 susceptibility loci for inflammatory bowel disease and highlight shared genetic risk across populations. Nat Genet 2015;47:979-86.
9. Knights D, Silverberg MS, Weersma RK, et al. Complex host genetics influence the microbiome in inflammatory bowel disease. Genome Med 2014;6:107.
10. Shaw KA, Bertha M, Hofmekler T, et al. Dysbiosis, inflammation, and response to treatment: a longitudinal study of pediatric subjects with newly diagnosed inflammatory bowel disease. Genome Med 2016;8:75.
11. Gevers D, Kugathasan S, Denson LA, et al. The treatment-naive microbiome in new-onset Crohn's disease. Cell Host Microbe 2014;15:382-392.
12. Ananthakrishnan AN, Luo C, Yajnik V, et al. Gut Microbiome Function Predicts Response to Anti-integrin Biologic Therapy in Inflammatory Bowel Diseases. Cell Host Microbe 2017;21:603-610.e3.
13. Lewis JD, Chen EZ, Baldassano RN, et al. Inflammation, Antibiotics, and Diet as Environmental Stressors of the Gut Microbiome in Pediatric Crohn's Disease. Cell Host Microbe 2015;18:489-500.
14. Kelsen J, Bittinger K, Pauly-Hubbard H, et al. Alterations of the Subgingival Microbiota in Pediatric Crohn's Disease Studied Longitudinally in Discovery and Validation Cohorts. Inflamm Bowel Dis 2015;21:2797-805.
15. Docktor MJ, Paster BJ, Abramowicz S, et al. Alterations in diversity of the oral microbiome in pediatric inflammatory bowel disease. Inflamm Bowel Dis 2012;18:935-42.
16. Said HS, Suda W, Nakagome S, et al. Dysbiosis of salivary microbiota in inflammatory bowel disease and its association with oral immunological biomarkers. DNA Res 2014;21:15-25.
17. Levine A, Griffiths A, Markowitz J, et al. Pediatric modification of the Montreal classification for inflammatory bowel disease: the Paris classification. Inflamm Bowel Dis 2011;17:1314-21.

18. Shepanski MA, Markowitz JE, Mamula P, et al. Is an abbreviated Pediatric Crohn's Disease Activity Index better than the original? J Pediatr Gastroenterol Nutr 2004;39:68-72.
19. Hyams JS, Ferry GD, Mandel FS, et al. Development and validation of a pediatric Crohn's disease activity index. J Pediatr Gastroenterol Nutr 1991;12:439-47.
20. Turner D, Otley AR, Mack D, et al. Development, validation, and evaluation of a pediatric ulcerative colitis activity index: a prospective multicenter study. Gastroenterology 2007;133:423-32.
21. Human Microbiome Project C. Structure, function and diversity of the healthy human microbiome. Nature 2012;486:207-14.
22. Human Microbiome Project C. A framework for human microbiome research. Nature 2012;486:215-21.
23. 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.
24. Kozich JJ, Westcott SL, Baxter NT, et al. Development of a dual-index sequencing strategy and curation pipeline for analyzing amplicon sequence data on the MiSeq Illumina sequencing platform. Appl Environ Microbiol 2013;79:5112-20.
25. Rognes T, Flouri T, Nichols B, et al. VSEARCH: a versatile open source tool for metagenomics. PeerJ 2016;4:e2584.
26. Wang Q, Garrity GM, Tiedje JM, et al. Naive Bayesian classifier for rapid assignment of rRNA sequences into the new bacterial taxonomy. Appl Environ Microbiol 2007;73:5261-7.
27. Paulson JN, Stine OC, Bravo HC, et al. Differential abundance analysis for microbial marker-gene surveys. Nat Methods 2013;10:1200-2.
28. Svetnik V, Liaw A, Tong C, et al. Random forest: a classification and regression tool for compound classification and QSAR modeling. J Chem Inf Comput Sci 2003;43:1947-58.
29. Lloyd-Price J, Mahurkar A, Rahnavard G, et al. Strains, functions and dynamics in the expanded Human Microbiome Project. Nature 2017;550:61-66.
30. Mark Welch JL, Rossetti BJ, Rieken CW, et al. Biogeography of a human oral microbiome at the micron scale. Proc Natl Acad Sci U S A 2016;113:E791-800.
31. Costello EK, Lauber CL, Hamady M, et al. Bacterial community variation in human body habitats across space and time. Science 2009;326:1694-7.
32. Dethlefsen L, McFall-Ngai M, Relman DA. An ecological and evolutionary perspective on human-microbe mutualism and disease. Nature 2007;449:811-8.
33. Reid G, Younes JA, Van der Mei HC, et al. Microbiota restoration: natural and supplemented recovery of human microbial communities. Nat Rev Microbiol 2011;9:27-38.
34. Walujkar SA, Dhotre DP, Marathe NP, et al. Characterization of bacterial community shift in human Ulcerative Colitis patients revealed by Illumina based 16S rRNA gene amplicon sequencing. Gut Pathog 2014;6:22.
35. Morgan XC, Tickle TL, Sokol H, et al. Dysfunction of the intestinal microbiome in inflammatory bowel disease and treatment. Genome Biol 2012;13:R79.
36. Frank DN, St Amand AL, Feldman RA, 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.
37. Hansen R, Russell RK, Reiff C, et al. Microbiota of de-novo pediatric IBD: increased Faecalibacterium prausnitzii and reduced bacterial diversity in Crohn's but not in ulcerative colitis. Am J Gastroenterol 2012;107:1913-22.
38. Sokol H, Seksik P. The intestinal microbiota in inflammatory bowel diseases: time to connect with the host. Curr Opin Gastroenterol 2010;26:327-31.
39. Alipour M, Zaidi D, Valcheva R, et al. Mucosal Barrier Depletion and Loss of Bacterial Diversity are Primary Abnormalities in Paediatric Ulcerative Colitis. J Crohns Colitis 2016;10:462-71.

40. Knights D, Lassen KG, Xavier RJ. Advances in inflammatory bowel disease pathogenesis: linking host genetics and the microbiome. Gut 2013;62:1505-10.
41. Wade WG. The oral microbiome in health and disease. Pharmacol Res 2013;69:137-43.

Chapter 6:

Lessons Learnt and Recommendations for Future Studies

Need for large, case-control cohorts of non-European ancestry. African Americans constitute about 15% of the US population (50 million). Despite having the same disease burden compared to individuals of European ancestry, African Americans are largely understudied in genetic and clinical research. A vast majority of the inflammatory bowel disease-related genetic discoveries were made in cohorts of European descent[1-3], with a handful of loci originally discovered in populations from Japan[4], India[5] and Korea[6], that were subsequently validated in European cohorts. As a result of these Eurocentric biases, prevailing efforts directed toward the translation of genetic discoveries into biological insights to enable precision medicine, are overwhelmingly in favor of certain populations compared to others. For instance, a recent study investigating the utility of polygenic risk scores in predicting a person's risk of various complex diseases, demonstrated that the existing genetic data performs well in European-descent individuals compared to others, with least accuracy reported in African-descent individuals[7], highlighting the genetic heterogeneity across populations that is highly likely driven by differences in genetic architecture, effect sizes and/or allele frequency. This disparity reinforces the need for more diversity in genetic studies. Future studies focusing more on substantially understudied population samples may aid in identifying non-European alleles of inflammatory bowel disease, which may account for, at least, a portion of the missing heritability. Expansion of other population samples, especially of African-descent individuals with shorter linkage disequilibrium intervals can further resolve the genetic architecture of common complex diseases and might put us one step closer toward effectively translating genome-based findings into health care.

Leveraging the genetic heterogeneity across populations to understand widening ethnic disparities in inflammatory bowel disease. Compared to European-descent individuals, the clinical course of disease in African Americans is more aggressive and severe leading to worse outcomes including, perianal, fistulizing and gastroduodenal disease. However, it remains undescribed whether these differences are due to hyper-activation of common molecular mechanisms or involvement of different biological mechanisms. Our finding that rare variant contributions to inflammatory bowel disease risk exist in a population-specific manner (Chapter 2), provides a global context for understanding reasons for differential disease severity

across ethnicities at genetic and/or molecular levels. Comparative analyses of African and European Americans are the first step toward understanding reasons for disparity in disease course and severity of inflammatory bowel disease. However, given the fact that non-European sample sizes are considerably smaller due to the Eurocentric biases in genetic studies, it remains to be seen to what extent such differences in the genetic architecture of inflammatory bowel disease risk across populations provide insights into the reasons behind widening ethnic disparities.

Un-interpreted genetic signals, and trans-ethnic summary statistic fine-mapping analysis of causal variants in inflammatory bowel disease. As is typical of complex diseases, one of the substantial issues that needs to be overcome is to extract maximum useful information from genetic data where >90% of the variants implicated by GWAS or sequencing studies localize to non-coding sequences, and are co-inherited with nearby causal regulatory variants. Because of the widespread correlation structure within the genome, numerous variants that are highly correlated with each other depict similar evidence for association with a trait, presenting a key challenge in distinguishing causal variants from surrogate variants. This phenomenon often known as linkage disequilibrium exists in a population specific manner. Therefore, subtle differences in statistical evidence of association from large trans-ethnic GWAS or sequencing data sets might significantly improve fine-mapping of causal variants[3,8].

African Americans in particular, are a recently admixed population, with a median genome composition of ~80% African and 20% European ancestries, and consequently, have relatively higher genetic diversity and shorter linkage disequilibrium blocks in regions of African ancestry than do European Americans. These properties are consistent with the notion that they might readily facilitate fine-mapping efforts. For instance, our comparison of the estimated effects between European- vs African-descent individuals apparently indicate that some of the lead variants with directionally inconsistent effects are less likely to be causal to inflammatory bowel disease (Chapter 2). In the recent fine-mapping efforts undertaken in large GWAS data sets of European ancestry[8], of the 139 independent inflammatory bowel disease associated regions that were investigated, 18 were resolved to a single causal variant with >95% certainty and an additional 27 with

moderate confidence (>50% certainty). At the remaining 94 associated signals, a total of at least 4,900 SNPs demonstrated similar evidence for association, with many SNPs depicting equal evidence. Because African Americans have different, and relatively shorter linkage disequilibrium patterns, re-doing the fine-mapping analysis jointly with data from both the European- and African-descent individuals might dramatically change the evaluation at many of these remaining 94 signals, as well as at the other known loci where fine-mapping had not yet been attempted.

DNA methylation data as a functional tool to identify critical variants in inflammatory bowel disease risk loci. Besides using summary statistics from large GWAS or sequencing data sets to pinpoint causal variants, functional annotations of genetic variants has also been shown to reveal critical-regulatory variants[9]. Nevertheless, testing all the variants generated by GWAS and burgeoning whole-genome sequencing studies for functional analyses using gold-standard assays is very low-throughput and cumbersome. Therefore, more scalable approaches to sift through the disease-associated genetic intervals and determine if and how variants affect disease risk demands more attention. However, in order to annotate the regulatory consequences of variants in inflammatory bowel disease risk loci, previously undertaken fine-mapping studies intersected genetic data with publicly available gene expression and epigenetic information, generated by various consortia such as the Roadmap Epigenomics Mapping Consortium, as part of establishing reference transcriptomic and epigenomic chromatin marks across different tissue types and cells obtained from healthy individuals[8]. Central to our study, molecular architecture is dynamic, tissue- and context-specific, and its variation has been implicated in many complex diseases; therefore, it is critical that appropriate cell types, likely under appropriate conditions of stimulus including overt pathology, are evaluated for QTL and chromatin effects.

DNA methylation can be influenced by general or disease-related environmental exposures, stochastic perturbations or genetic variation in *cis* by the local sequence context or in *trans* by chromosomal looping; where the latter informs that methylation QTL could be as informative as expression and other epigenetic QTL. A recent study has demonstrated the potential of non-coding genetic variants that enrich as mQTLs

in driving differential methylation in inflammatory bowel disease[10]. In line with this, our finding that some of the Crohn's disease-associated methylation changes correlate with local genetic variation (Chapter 3), extends further support to the notion of annotating genetic signals with methylation information. Furthermore, our unpublished data from the lab (Venkateswaran *et al*; in preparation), indicate that approximately 40% of the Crohn's disease risk loci that localize predominantly to non-coding sequences show association with DNA methylation in blood. Based on these evidence, it is tempting to postulate that DNA methylation data could be used as a functional tool to sift through genetic regions and hone in on critical DNA regulatory variants that underlie inflammatory bowel disease-associated GWAS signals. However, the usefulness of disease- and context-specific DNA methylation in further resolving or improving fine-mapping remains a challenge for the future.

Integrative epigenetic and transcriptomic analysis of genetic associations to gain molecular insights into GWAS signals. GWAS loci established thus far in the pathogenesis of inflammatory bowel disease, collectively implicate a total of about 900 genes (that are present within ±500 kb from the lead variant) in disease susceptibility[1-3]; therefore, identifying the relevant genes and underlying molecular mechanisms that mediate genetic effects of inflammatory bowel disease is critical to gain mechanistic insights. Intersecting genetic data with the context-specific functional genomic information may facilitate molecular insights into how regulatory non-coding variants with robust and replicable association with inflammatory bowel disease can contribute to disease pathology. Expression quantitative trait loci (eQTL) are being used to fine map which genes are most likely to account for the GWAS signals, but in many cases the locations of the causal nucleotide changes and relevant genes often remain unknown[8,11]. This gap in understanding is a roadblock for targeted prevention and therapy. Our unpublished data (Venkateswaran *et al*) indicate that about 40% of the Crohn's disease risk loci that localize predominantly to non-coding sequences show association with DNA methylation in blood, which in turn correlates with gene expression patterns; however, it is not known if this is the mechanism by which genetic variants contribute to disease pathology. Our integrative analysis of genetic association and the concept of Mendelian randomization, in combination

with publicly available summary-data based Mendelian randomization results indicate that an inflammatory bowel disease-associated locus on chromosome 6, tagged by rs1819333, contribute to disease pathology by influencing DNA methylation changes that result in transcriptional silencing of the gene, *RPS6KA2* (Chapter 3). This supports the utility of DNA methylation to hone in on which genes are most likely to account for the GWAS signals, and to facilitate mechanistic insights into how such disease-associated regulatory variants can contribute to disease pathology. Future studies with genetic, epigenetic and transcriptomic data along with the clinical phenotypic information, could use our integrative analytical framework to gain molecular insights into the remaining GWAS loci.

Quantifying the impact of environmental exposures on the epigenome and establishing the causal potential of exposure-associated DNA methylation in inflammatory bowel disease. DNA methylation is the most well studied epigenetic modification in relation to environmental influences. Despite emerging correlative evidence across a wide range of complex diseases, the precise causal relationships or mechanisms underlying environmental exposure-DNA methylation associations remain elusive. Of our 1189 Crohn's disease-associated CpG sites (Chapter 3), while we were able to test the causal versus consequential roles of differential methylation at 194 CpG sites for which genetic proxies were readily available (through mQTL analysis), likely causal potential at the remaining CpGs is unclear. These methylation changes that are not under local genetic influence may be attributed to one or more of the following reasons: i) result of stochastic alterations; ii) result of Crohn's disease; and/or iii) result of individual or a mixture of environmental influences. Our unpublished environmental epigenetics of Crohn's disease data suggest that the effect of environmental exposures on DNA methylation in blood remain consistent at the two-time points examined – at diagnosis and during the 3-yr follow-up period, rooting for the longitudinal stability of environmental exposure-DNA methylation associations in Crohn's disease. This evidence implies that the remaining Crohn's disease-associated DNA methylation changes are presumably under some mode of "environmental influence". While it is highly likely that DNA methylation may serve as a molecular mechanism in mediating the risk of environmental exposures in Crohn's disease,

future work should establish the precise causal relationships between the exposures, DNA methylation and inflammatory bowel disease, and define their functional consequences.

Genetics of microbiome and its integration with the epigenome and transcriptome to gain causal and molecular insights. Gut microbial dysbiosis influenced both by extrinsic factors and by host genetics has been shown to be associated robustly and reproducibly with various complex diseases, including inflammatory bowel disease; however, the potentially causal nature of the microbiome and its functional mechanistic links to disease remain largely unknown. As we proposed for the first time in our review on the present and future of microbiome in complex diseases (see Chapter 4), a recent study has successfully leveraged an integrative Mendelian randomization approach to precisely define functional microbial changes that are potentially causal to type 2 diabetes[12]. Due to the limited sample size and the lack of genome-wide genotype data from our oral microbiome cohort (Chapter 5), we were unable to implement these integrative analyses to identify changes in microbial composition that might exert potentially causal effects in inflammatory bowel disease. However, future large microbiome-wide association studies of inflammatory bowel disease could implement this framework to resolve causal relationships that underlie gut or oral microbial associations with inflammatory bowel disease.

On the other hand, increasing evidence indicates that changes in gut microbiota and their metabolites affect host physiology and susceptibility to disease[12-14]; however, the mechanisms by which microbial changes contribute to pathology remain largely undescribed. Recent evidence implicates that microbiota and their metabolites might contribute to disease by affecting host epigenetic states, including histone methylation and acetylation which in turn result in transcriptional reprogramming[15-17]. We present an analytical framework pertaining to the integration of genetic, epigenetic and transcriptomic data with microbial signatures of inflammatory bowel disease in order to gain systematic mechanistic insights into how changes in microbiota could affect disease. Future studies focusing more on the functional aspects of microbial dysbiosis using metagenomic characterization, and integrating these data further with host genomes, epigenomes, and transcriptomes from the same subjects may aid in closing these gaps.

REFERENCES

1. Jostins, L. *et al.* Host-microbe interactions have shaped the genetic architecture of inflammatory bowel disease. *Nature* **491**, 119-24 (2012).
2. Liu, J.Z. *et al.* Association analyses identify 38 susceptibility loci for inflammatory bowel disease and highlight shared genetic risk across populations. *Nat Genet* **47**, 979-986 (2015).
3. de Lange, K.M. *et al.* Genome-wide association study implicates immune activation of multiple integrin genes in inflammatory bowel disease. *Nat Genet* **49**, 256-261 (2017).
4. Asano, K. *et al.* A genome-wide association study identifies three new susceptibility loci for ulcerative colitis in the Japanese population. *Nat Genet* **41**, 1325-9 (2009).
5. Juyal, G. *et al.* Genome-wide association scan in north Indians reveals three novel HLA-independent risk loci for ulcerative colitis. *Gut* **64**, 571-9 (2015).
6. Yang, S.K. *et al.* Genome-wide association study of Crohn's disease in Koreans revealed three new susceptibility loci and common attributes of genetic susceptibility across ethnic populations. *Gut* **63**, 80-7 (2014).
7. Martin, A.R. *et al.* Clinical use of current polygenic risk scores may exacerbate health disparities. *Nat Genet* **51**, 584-591 (2019).
8. Huang, H. *et al.* Fine-mapping inflammatory bowel disease loci to single-variant resolution. *Nature* **547**, 173-178 (2017).
9. Mahajan, A. *et al.* Fine-mapping type 2 diabetes loci to single-variant resolution using high-density imputation and islet-specific epigenome maps. *Nat Genet* **50**, 1505-1513 (2018).
10. Ventham, N.T. *et al.* Integrative epigenome-wide analysis demonstrates that DNA methylation may mediate genetic risk in inflammatory bowel disease. *Nat Commun* **7**, 13507 (2016).
11. Marigorta, U.M. *et al.* Transcriptional risk scores link GWAS to eQTLs and predict complications in Crohn's disease. *Nat Genet* **49**, 1517-1521 (2017).
12. Sanna, S. *et al.* Causal relationships among the gut microbiome, short-chain fatty acids and metabolic diseases. *Nat Genet* **51**, 600-605 (2019).
13. Martinez, K.B., Leone, V. & Chang, E.B. Microbial metabolites in health and disease: Navigating the unknown in search of function. *J Biol Chem* **292**, 8553-8559 (2017).
14. Roager, H.M. & Licht, T.R. Microbial tryptophan catabolites in health and disease. *Nat Commun* **9**, 3294 (2018).
15. Qin, Y. *et al.* An obesity-associated gut microbiome reprograms the intestinal epigenome and leads to altered colonic gene expression. *Genome Biol* **19**, 7 (2018).
16. Kelly, D. *et al.* Microbiota-sensitive epigenetic signature predicts inflammation in Crohn's disease. *JCI Insight* **3**(2018).
17. Krautkramer, K.A., Rey, F.E. & Denu, J.M. Chemical signaling between gut microbiota and host chromatin: What is your gut really saying? *J Biol Chem* **292**, 8582-8593 (2017).