Distribution Agreement

Signature:

______________________________ _______________
Kelly Ann Shaw Date

Genetic and environmental contributions to gastrointestinal health

By

Kelly Ann Shaw
Doctor of Philosophy

Graduate Division of Biological and Biomedical Sciences
Genetics and Molecular Biology

Jennifer G. Mulle, M.H.S., Ph.D.
Advisor

Michael E. Zwick, Ph.D.
Advisor

David J. Cutler, Ph.D.
Committee Member

Michael P. Epstein, Ph.D.
Committee Member

Timothy D. Read, Ph.D.
Committee Member

Accepted:

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

Date

Genetic and environmental contributions to gastrointestinal health

By

Kelly Ann Shaw
B.S., Michigan State University, 2011

Advisor: Jennifer G. Mulle, M.H.S., Ph.D.
Advisor: Michael E. Zwick, 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
2017

Abstract

Genetic and environmental contributions to gastrointestinal health
By Kelly Ann Shaw

More than 60 million people in the United States (US) are affected by gastrointestinal (GI) diseases. An estimated $100 billion is spent on direct costs for medical care and indirect costs from morbidity and mortality. Genetics, diet, and microbes all play interconnected roles in the development and normal functioning of the GI tract. Through my dissertation work I sought to address the relationship between these factors and GI symptoms and disease. First, I tested a hypothesis generated from parents of individuals with a rare single-gene metabolic disease, classic galactosemia (CG). These parents anecdotally reported their children suffered from GI symptoms. Using an online survey, I found that individuals with CG were 4.5 times more likely to report constipation and 4.2 times more likely to report nausea compared to controls. There were no significant effects of predicted residual GALT activity or dietary galactose restriction, two known modifiers of other long-term outcomes in CG. Secondly, I sought to identify rare genetic variants that may contribute to increased susceptibility to pediatric inflammatory bowel disease (IBD). We found overlap with well-established IBD genes and evidence supporting the contribution of neutrophil function to disease. We also found variants in several extracellular matrix proteins, which have been of recent interest in the field. Finally, I studied gut bacteria in IBD, because host immune response to microbes likely plays a role in disease etiology. Previous work found increases and decreases in specific bacterial families in patients compared to controls. I expanded on their work by studying these bacteria longitudinally. I found that this imbalance in bacteria decreased over time but remained higher than in controls. While abundance of these IBD-associated bacteria was associated with a marker of gut inflammation, it did not differ between patients with and without mucosal healing, a marker of response to treatment. I discovered other bacterial groups that better separated responders to treatment from non-responders; a larger study is needed to follow up on these findings. My dissertation work focused on these two diseases to advance our knowledge of GI health and potentially lead to better prevention, prognosis, and treatment of disease.

Genetic and environmental contributions to gastrointestinal health

By

Kelly Ann Shaw
B.S., Michigan State University, 2011

Advisor: Jennifer G. Mulle, M.H.S., Ph.D.
Advisor: Michael E. Zwick, 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
2017

Acknowledgements

To my wonderful family and friends, who have taught me the most important things in life.

I hope you don’t actually try to read this.

Table of Contents

CHAPTER I. Introduction

Poor gastrointestinal (GI) health is a significant problem for individual and public health.

More than 60 million people in the United States (US) are affected by digestive tract diseases[1] such as constipation, diarrhea, inflammatory bowel disease, irritable bowel syndrome, gastrointestinal infections, hemorrhoids, diverticular disease, abdominal hernia, gallstones, ulcers, hepatitis, and pancreatitis[2]. In 2012, around 32.1 million ambulatory care visits (3.5 percent of all such visits) were associated with diseases of the digestive system[3]. Almost 1 in 4 of those visits were to emergency departments; these 7.5 million visits accounted for 5.8 percent of total emergency visits[4]. There were 21.7 million hospitalizations as a result of digestive diseases in 2010[5] and 245,921 deaths attributable to digestive disease in the US in 2009, representing 10% of all deaths that year[6].

Not only is the impact of GI disease wide in scope, it is also very costly. One estimate from 2004 found not only $97.8 billion in direct medical costs for care of GI diseases, but also an estimated $44 billion in indirect costs due to lost work from disease-associated morbidity and mortality[2,7]. In the 2010 National Health and Wellness Survey of 75,000 people in the US, individuals with GI diseases or experiencing GI symptoms reported worse mental and physical health and higher levels of impairment in work and general activities than individuals without disease[6].

Gastrointestinal health is a multi-faceted problem since the GI tract is one of the primary interfaces of the human body and its environment. Genetics, diet, and microbes all play important and interconnected roles in development and normal functioning of the GI tract. Through my dissertation work I sought to improve our understanding of GI health through studying some of these factors—not only in a disease specific to the digestive tract, inflammatory bowel disease, but also classic galactosemia, a disease not traditionally thought of as having GI involvement.

Single-gene inherited metabolic disorders can provide novel insight into GI health.

GI involvement has been understudied in classic single-gene Mendelian diseases involving inborn errors of metabolism. Though the causative gene is known in these disorders, often the actual pathophysiology that results in clinical manifestations is unknown. Another contributing factor in these disorders is the necessity of avoiding intake of specific nutrients; this can mean lifelong adherence to a diet which is fundamentally different from most of the population.

Classic galactosemia (CG) is one example of a rare, single-gene inherited inborn error of metabolism that could offer insight into issues of GI health. The incidence of CG is approximately 1/50,000 in the US, occurring mainly in populations of European descent[8,9]. The primary metabolic defect of CG, an inability to metabolize galactose, arises from null or low activity of both alleles encoding the galactose-1-phosphate-uridylyltransferase (GALT) enzyme. All states screen for CG as part of newborn screening[8,9], because infants must be identified and stop breastfeeding immediately to prevent severe acute complications including vomiting, diarrhea, failure to thrive, and hepatomegaly from their inability to process the galactose in breastmilk.

Even though simple dietary intervention to remove sources of galactose prevents severe acute illness, many children with the disease still experience long-term complications[10,11]. One negative long-term outcome experienced by more than half of patients is developmental delay including motor, behavioral, speech, and emotional abnormalities, along with cognitive disability[11–13]. Women with CG also experience very high prevalence (80-90%) of primary or premature ovarian insufficiency[11,14] (and reviewed in [15]). Additionally, anecdotal reports from parents in the CG community led to the hypothesis that children with CG might also experience GI problems. However, no study had formally investigated the problem. In the course of this dissertation I helped lay the foundation for GI health research in CG by testing whether these individuals experienced higher prevalence of GI symptoms and whether any known modifiers of long-term outcomes—predicted residual activity or dietary galactose restriction—also showed effects on GI health[16].

While GI health in CG is a new area of research, inflammatory bowel disease is a disorder of the digestive tract which has been studied for decades.

Inflammatory bowel disease (IBD) is characterized by chronic remitting and relapsing inflammation of some portion of the GI tract. The two most common forms of IBD are Crohn's disease (CD) and ulcerative colitis (UC). CD and UC are primarily differentiated by the location and characteristics of inflammation. In CD, inflammation is transmural (spanning all layers of the epidermis) and can occur in a discontinuous pattern anywhere along the GI tract. Abscesses, strictures (narrowing of the GI tract), or fistulas are possible complications. In contrast, inflammation in UC is not transmural and is limited to the colon. In UC the innermost part of the epidermis sloughs off, or ulcerates, leading to a distinctive "cobblestone" appearance in the large intestine (clinical aspects of disease reviewed in [17]).

Both diseases result in substantial quality of life issues. Symptoms vary based on disease location or severity of inflammation, but abdominal discomfort or pain, diarrhea, and passage of blood and/or mucus are common among patients. Additionally, up to 25% of patients present with extraintestinal manifestations which often include inflammation of non-GI tissues such as uveitis, pleuritis, myocarditis, pancreatitis, ankylosing spondylitis, arthritis, and tendonitis[18]. Since IBD is a chronic disease, patients often require medication long-term, and colonoscopies are needed from a younger age for surveillance to counter the possible increased risk of colon cancer. Surgery is also a frequent outcome for individuals with IBD: 70-80% of patients with CD have intestinal surgery within 20 years of diagnosis, and 25-30% of UC patients require colectomy within 25 years[19].

According to a 2016 study of the population of Olmstead county, Minnesota, estimated US prevalence per 100,000 people is 246.7 cases for CD and 286.3 cases of UC, with an annual incidence per 100,000 people of 10.7 and 12.2 new cases of CD and UC, respectively[20]. A more broad study that utilized data from 12 million commercially-insured individuals from 2008-2009 estimated very similar prevalence in adults—241 cases of CD and 263 cases of UC per 100,000 people[21]. These numbers mean that an estimated 1.2-1.6 million people currently have IBD in the US. While most diagnoses of IBD are received in the age range from late 20s to mid-30s[20–22], an estimated 5% of prevalent cases, or 62,000 patients, are younger than 20 years of age[21].

Worryingly, most studies of IBD in the US have found evidence of increasing incidence in both adult and pediatric populations[20,21]. Hospitalizations and associated healthcare costs from IBD follow the same trend of significant increase over time, from approximately 1.2 billion 2012-inflation-adjusted dollars in 1993 to $3.5 billion in 2012[23].

IBD has been a paragon of discovery in genome-wide association studies.

Because family history of IBD is the biggest risk factor for disease (a study of the entire Danish population estimated that up to 12% of all IBD cases in that country were familial[24]), genetic studies have been pursued as one way of understanding disease etiology. In studies of twins, concordance between monozygotic twins—who share 100% of their DNA—was 37.3% and 10% for CD and UC, respectively. For dizygotic twins, who share the same environment *in utero* but are no more genetically similar than other siblings, concordance was 7% for CD and 3% for UC [25]. The higher concordance for MZ twins demonstrates that genetic factors play a role in getting IBD.

Of course with this evidence for heritability, more detailed genetic studies soon followed to identify specific genetic loci that associate with disease. Before high-resolution association studies were possible, linkage studies were originally used to find general areas of the genome that could contribute to IBD risk. Through these approaches, signals associated with either CD or UC were found and replicated across the genome on chromosomes 3, 5, 6, 12, 14, 16, and 19 (reviewed in [25]). The advent of array technologies allowed hundreds of thousands of loci to be genotyped at a reasonable cost. This allowed for larger sample sizes and increased resolution of the genome; as a result the number of IBD-associated loci skyrocketed.

To date, genome-wide association studies (GWAS) have identified well over 200 loci associated with risk for IBD. In 2012, Jostins et al. published the largest meta-analysis of IBD, which included genetic data for 32,628 IBD cases and 29,704 controls[26]. This study identified 163 loci significantly associated with IBD. Thirty loci showed an effect only in Crohn's disease (including *NOD2* with OR >3), 23 loci were specific to ulcerative colitis (the most distinctive being *HLA*), and 110 loci were associated with both diseases (e.g. *IL23R, MUC19*), suggesting genetic architecture of the two is mostly shared. Liu et al. expanded this research in 2015 to more diverse cohorts including 9,846 individuals of Iranian, Indian, or East Asian descent; they replicated the Jostins findings in addition to discovering 38 additional loci. They found that for most associated loci the direction and magnitude of effects were the same across populations, but there were important differences in allele frequency (e.g. *NOD2*), effect size, or both (e.g. *IL23R*) for several loci[27]. Another 25 loci were recently added to the list, 3 of which encode integrin proteins[28]. Though genetic findings in pediatric IBD largely echo findings in adults[29,30], one study of greater than 1,000 pediatric-onset IBD cases and 1,600 controls found slightly increased odds ratios for risk alleles also found in adult populations (including the well-established *NOD2*), and greater burden of these common variants was weakly correlated with earlier age of onset in Crohn's disease[31].

Overall heritability calculated using data from genotyping studies is estimated to be 37% for CD and 27% for UC[32], approximately half of the heritability estimated from twin studies (75% for CD and 67% for UC), reflecting the recurring theme of heritability that is "missing" after GWAS is performed[33,34]. Though IBD is regarded as a GWAS "success" because so many SNPs have been identified, effect sizes for these variants are generally small (with an average OR 1.1), and only account for 13.1% and 8.2% of variance in disease for CD and UC, respectively[27], leaving room for contributions from other genetic features such as rare genetic variants, copy number variation, and epigenetic differences. Most variants in protein-coding sequence are at low frequency[35–37], and the explosive growth of the human population in recent history has led to a corresponding explosion of rare variants[38]. I therefore set out to explore pathway enrichment and rare genetic variation in a cohort of pediatric IBD cases.

There is evidence in IBD that the gut microbiome is an important environmental influence contributing to disease etiology.

As previously mentioned, diagnoses of IBD as well as associated costs have been increasing in the US, but this phenomenon is noted more broadly in the majority of adult[19,22,39] and pediatric[40,41] cohorts worldwide. While the highest rates of IBD are in North America, the UK, and northern Europe, countries experiencing the greatest increase in rates are nations undergoing recent booms in industrialization such as those in East Asia[19,22,42]. Immigrants who move from low-incidence to high-incidence areas are at increased risk for IBD, and this increased risk is also experienced by their descendants[43]. These observations provide evidence that genetics is not everything in IBD—the environment also plays a large role[39,42,44–46].

Many factors change as a country develops. There are changes in occupational exposures as industry grows and more people move to urban locales. Diet may also be impacted as the economy grows and international restaurant chains seek new markets. Of additional importance, developing countries undergo an epidemiologic transition—where society's morbidity and mortality burden shifts from infectious to chronic disease through improvements in public health interventions and medical care.

These diet, lifestyle behavior, sanitation, and environmental exposure changes that accompany industrialization have been linked to development of IBD[44]. Not only does this shift population exposure to microbes in the external environment, but also in the environment they carry around every day. The human microbiome is the collection of microbes, including bacteria, viruses, fungi, and single-celled eukaryotes on and within the human body. There are multiple body sites where microbes have carved out niches to live: the skin, respiratory tract, genitourinary tract, and all along the digestive tract.

The gut microbiome contains the most diverse population of microbes[47], and much research has focused on this site. In addition to its involvement in digestion and response to environmental chemicals, the gut microbiome in humans is important for healthy gut and immune system development, as well as ongoing regulation of the immune system, and prevention of invasion and growth of pathogens (reviewed in [48]). With these important roles in human health, it seems likely that the gut microbiome could also play an important role in disease. Since immune activation and host response to microbes emerge as an important theme in genetic studies of IBD[26–28], defining the gut microbiome in IBD was a high priority. Another compelling reason to pursue the role of the gut microbiome in disease is that we can target the gut microbiome for intervention quite easily—through probiotic supplementation or fecal microbiome transplants.

Preliminary studies of IBD patients' gut microbiomes have found significant differences in their microbiomes compared to controls, including an overall reduction in bacterial diversity as well as altered abundance of specific bacterial groups and gene families found within bacterial genomes[49–55]. Studies have shown that the gut microbiome plays a large role in driving inflammation in IBD[56] and treatment involving antibiotics has been shown to reduce intestinal inflammation in patients[57].

One large study of treatment-naïve pediatric Crohn's patients helped set the stage for microbiome research in IBD.

In 2014, Gevers et al. published a study where they compared intestinal biopsy and fecal samples in 447 children with newly diagnosed, treatment-naïve Crohn's Disease to 221 controls[52]. They discovered that bacterial families Enterobacteriaceae, Pasteurellaceae, Fusobacteriaceae, Neisseriaceae, Veillonellaceae, Gemellaceae were increased in patients.

Bacterial orders Bacteroidales and Clostridiales (excluding Veillonellaceae) and families Erysipelotrichaceae and Bifidobacteriaceae were significantly decreased in patients.

In this study, disease severity was measured by the Pediatric Crohn's Disease Activity Index (or PCDAI), which is the most common measure of disease activity. It involves collection of data including patient recall of their symptoms over the last week, various blood markers and basic clinical exam, weight gain/loss, height trajectory, but it's worth noting it is fairly subjective, not a direct measure of inflammation or treatment response. They showed that when you sum the abundances of these CD-associated taxa in samples, higher abundance associated with worse severity of disease. With the decreased-in-CD taxa, higher abundance associated with less severe disease.

"Dysbiosis" is the general term often used to refer to gut microbiome characteristics that are different in a group compared to controls. Gevers et al. created an IBD-specific quantification called the microbial dysbiosis index. They took the $\log_{10}$ of the total abundance of bacteria associated with Crohn's divided by the abundance of bacteria decreased in Crohn's. The dysbiosis index was therefore a summary measure that maximized the differences between cases and controls, and in separate replication samples of 425 ileal biopsies, 300 rectal biopsies, and 199 stool samples, it successfully separated cases and controls with area under the receiver operating characteristic curve (AUC) of 0.85, 0.78, and 0.66, respectively.

While the Gevers paper did a lot to set the stage for microbiome research in IBD, there were unanswered questions that I sought to address as part of my dissertation work[58]: What happens to the microbiome over time with treatment? Gevers et al. showed dysbiosis associated with a somewhat subjective measure of disease activity, but what about a more objective measure of inflammation? Lastly, does the dysbiosis index associate with treatment outcome?

Because of the significant impact GI disease has on individuals and the healthcare system, it is important to learn more about GI health, whether from diseases of the digestive tract like IBD or disorders like CG in which GI symptoms are secondary.

My dissertation work focused on these two seemingly disparate diseases to advance our knowledge of GI health. In CG, examining whether GI problems are associated with disease could bring clinicians' attention to an as-yet unrecognized issue for their patients, improving quality of life. Potential GI involvement could also contribute to new hypotheses regarding disease pathophysiology, as well as emphasize the need to rigorously study possible contributions of diet to long-term outcomes in CG. For inflammatory bowel disease, finding rare genetic variants in known or novel genes associated with disease could both illuminate etiology or new pathways for therapeutic targeting. Likewise, greater knowledge of the gut microbiome in IBD could be used not only to identify disease, but examination of longitudinal patterns of change could be used to monitor disease status and inform treatment options, including those targeted to the microbiome itself.

REFERENCES

1. National Institutes of Health, U.S. Department of Health and Human Services. Opportunities and Challenges in Digestive Disease Research: Recommendations of the National Commission on Digestive Diseases [Internet]. 2009 [cited 2016 Dec 4]. Available from: https://www.niddk.nih.gov/about-niddk/strategic-plans-reports/Documents/NCDD%20Research%20Plan/NCDD_04272009_ResearchPlan_CompleteResearchPlan.pdf

2. Digestive Diseases Statistics for the United States [Internet]. National Institute of Diabetes and Digestive and Kidney Diseases, National Institutes of Health; 2014 [cited 2016 Dec 4]. Available from: https://www.niddk.nih.gov/health-information/health-statistics/Pages/digestive-diseases-statistics-for-the-united-states.aspx

3. National Center for Health Statistics, Centers for Disease Control and Prevention. National Ambulatory Medical Care Survey: 2012 State and National Summary Tables [Internet]. 2016 [cited 2016 Dec 5]. Available from: http://www.cdc.gov/nchs/data/ahcd/namcs_summary/2013_namcs_web_tables.pdf

4. National Center for Health Statistics, Centers for Disease Control and Prevention. National Hospital Ambulatory Medical Care Survey: 2012 Emergency Department Summary Tables [Internet]. 2016 [cited 2016 Dec 4]. Available from: http://www.cdc.gov/nchs/data/ahcd/nhamcs_emergency/2012_ed_web_tables.pdf

5. National Center for Health Statistics, Centers for Disease Control and Prevention. Number of all-listed diagnoses for discharges from short-stay hospitals, by ICD-9-CM code, sex, age, and geographic region: United States, 2010 [Internet]. [cited 2016 Dec 5].

Available from:
http://www.cdc.gov/nchs/data/nhds/10Detaileddiagnosesprocedures/2010det10_numberalldiagnoses.pdf

6. Peery AF, Dellon ES, Lund J, Crockett SD, McGowan CE, Bulsiewicz WJ, Gangarosa LM, Thiny MT, Stizenberg K, Morgan DR, Ringel Y, Kim HP, DiBonaventura M daCosta, Carroll CF, Allen JK, Cook SF, Sandler RS, Kappelman MD, Shaheen NJ. Burden of Gastrointestinal Disease in the United States: 2012 Update. Gastroenterology. 2012 Nov;143(5):1179–1187.e3. PMCID: PMC3480553

7. Everhart J. The Burden of Digestive Diseases in the United States. National Institute of Diabetes and Digestive and Kidney Diseases, US Department of Health and Human Services; 2008.

8. Pyhtila BM, Shaw KA, Neumann SE, Fridovich-Keil JL. A brief overview of galactosemia newborn screening in the United States. J Inherit Metab Dis. 2014 Jul;37(4):649–650. PMID: 24658844

9. Pyhtila BM, Shaw KA, Neumann SE, Fridovich-Keil JL. Newborn Screening for Galactosemia in the United States: Looking Back, Looking Around, and Looking Ahead. JIMD Rep. 2014 Apr 10; PMID: 24718839

10. Jumbo-Lucioni PP, Garber K, Kiel J, Baric I, Berry GT, Bosch A, Burlina A, Chiesa A, Pico MLC, Estrada SC, Henderson H, Leslie N, Longo N, Morris AAM, Ramirez-Farias C, Schweitzer-Krantz S, Silao CLT, Vela-Amieva M, Waisbren S, Fridovich-Keil JL. Diversity of approaches to classic galactosemia around the world: a comparison of

diagnosis, intervention, and outcomes. J Inherit Metab Dis. 2012 Nov;35(6):1037–1049. PMCID: PMC3774053

11. Waggoner DD, Buist NR, Donnell GN. Long-term prognosis in galactosaemia: results of a survey of 350 cases. J Inherit Metab Dis. 1990;13(6):802–818. PMID: 1706789

12. Ryan EL, Lynch ME, Taddeo E, Gleason TJ, Epstein MP, Fridovich-Keil JL. Cryptic residual GALT activity is a potential modifier of scholastic outcome in school age children with classic galactosemia. J Inherit Metab Dis. 2013 Nov;36(6):1049–1061. PMCID: PMC3657299

13. Shriberg LD, Potter NL, Strand EA. Prevalence and Phenotype of Childhood Apraxia of Speech In Youth with Galactosemia. J Speech Lang Hear Res JSLHR. 2011 Apr;54(2):487–519. PMCID: PMC3070858

14. Kaufman F, Kogut MD, Donnell GN, Koch H, Goebelsmann U. Ovarian failure in galactosaemia. Lancet. 1979 Oct 6;2(8145):737–738. PMID: 90818

15. Fridovich-Keil JL, Gubbels CS, Spencer JB, Sanders RD, Land JA, Rubio-Gozalbo E. Ovarian function in girls and women with GALT-deficiency galactosemia. J Inherit Metab Dis. 2011 Apr;34(2):357–366. PMCID: PMC3063539

16. Shaw KA, Mulle JG, Epstein MP, Fridovich-Keil JL. Gastrointestinal Health in Classic Galactosemia. JIMD Rep. 2016 Jul 1; PMID: 27363831

17. Baumgart DC, Sandborn WJ. Inflammatory bowel disease: clinical aspects and established and evolving therapies. Lancet Lond Engl. 2007 May 12;369(9573):1641–1657. PMID: 17499606

18. Fakhoury M, Negrulj R, Mooranian A, Al-Salami H. Inflammatory bowel disease: clinical aspects and treatments. J Inflamm Res. 2014;7:113–120. PMCID: PMC4106026

19. Cosnes J, Gower-Rousseau C, Seksik P, Cortot A. Epidemiology and natural history of inflammatory bowel diseases. Gastroenterology. 2011 May;140(6):1785–1794. PMID: 21530745

20. Shivashankar R, Tremaine WJ, Harmsen WS, Loftus EV. Incidence and Prevalence of Crohn's Disease and Ulcerative Colitis in Olmsted County, Minnesota From 1970 Through 2010. Clin Gastroenterol Hepatol Off Clin Pract J Am Gastroenterol Assoc. 2016 Nov 14; PMID: 27856364

21. Kappelman MD, Moore KR, Allen JK, Cook SF. Recent trends in the prevalence of Crohn's disease and ulcerative colitis in a commercially insured US population. Dig Dis Sci. 2013 Feb;58(2):519–525. PMCID: PMC3576554

22. Molodecky NA, Soon IS, Rabi DM, Ghali WA, Ferris M, Chernoff G, Benchimol EI, Panaccione R, Ghosh S, Barkema HW, Kaplan GG. Increasing incidence and prevalence of the inflammatory bowel diseases with time, based on systematic review. Gastroenterology. 2012 Jan;142(1):46–54.e42; quiz e30. PMID: 22001864

23. Peery AF, Crockett SD, Barritt AS, Dellon ES, Eluri S, Gangarosa LM, Jensen ET, Lund JL, Pasricha S, Runge T, Schmidt M, Shaheen NJ, Sandler RS. Burden of Gastrointestinal, Liver, and Pancreatic Diseases in the United States. Gastroenterology. 2015 Dec;149(7):1731–1741.e3. PMCID: PMC4663148

24. Moller FT, Andersen V, Wohlfahrt J, Jess T. Familial risk of inflammatory bowel disease: a population-based cohort study 1977-2011. Am J Gastroenterol. 2015 Apr;110(4):564–571. PMID: 25803400

25. Baumgart DC, Carding SR. Inflammatory bowel disease: cause and immunobiology. Lancet. 2007 May 12;369(9573):1627–1640. PMID: 17499605

26. Jostins L, Ripke S, Weersma RK, Duerr RH, McGovern DP, Hui KY, Lee JC, Schumm LP, Sharma Y, Anderson CA, Essers J, Mitrovic M, Ning K, Cleynen I, Theatre E, Spain SL, Raychaudhuri S, Goyette P, Wei Z, Abraham C, Achkar J-P, Ahmad T, Amininejad L, Ananthakrishnan AN, Andersen V, Andrews JM, Baidoo L, Balschun T, Bampton PA, Bitton A, Boucher G, Brand S, Büning C, Cohain A, Cichon S, D'Amato M, De Jong D, Devaney KL, Dubinsky M, Edwards C, Ellinghaus D, Ferguson LR, Franchimont D, Fransen K, Gearry R, Georges M, Gieger C, Glas J, Haritunians T, Hart A, Hawkey C, Hedl M, Hu X, Karlsen TH, Kupcinskas L, Kugathasan S, Latiano A, Laukens D, Lawrance IC, Lees CW, Louis E, Mahy G, Mansfield J, Morgan AR, Mowat C, Newman W, Palmieri O, Ponsioen CY, Potocnik U, Prescott NJ, Regueiro M, Rotter JI, Russell RK, Sanderson JD, Sans M, Satsangi J, Schreiber S, Simms LA, Sventoraityte J, Targan SR, Taylor KD, Tremelling M, Verspaget HW, De Vos M, Wijmenga C, Wilson DC, Winkelmann J, Xavier RJ, Zeissig S, Zhang B, Zhang CK, Zhao H, International IBD Genetics Consortium (IIBDGC), Silverberg MS, Annese V, Hakonarson H, Brant SR, Radford-Smith G, Mathew CG, Rioux JD, Schadt EE, Daly MJ, Franke A, Parkes M, Vermeire S, Barrett JC, Cho JH. Host-microbe interactions have shaped the genetic architecture of inflammatory bowel disease. Nature. 2012 Nov 1;491(7422):119–124. PMCID: PMC3491803

27. Liu JZ, van Sommeren S, Huang H, Ng SC, Alberts R, Takahashi A, Ripke S, Lee JC, Jostins L, Shah T, Abedian S, Cheon JH, Cho J, Daryani NE, Franke L, Fuyuno Y, Hart A, Juyal RC, Juyal G, Kim WH, Morris AP, Poustchi H, Newman WG, Midha V, Orchard TR, Vahedi H, Sood A, Sung JJY, Malekzadeh R, Westra H-J, Yamazaki K, Yang S-K, International Multiple Sclerosis Genetics Consortium, International IBD Genetics Consortium, Barrett JC, Franke A, Alizadeh BZ, Parkes M, B K T, Daly MJ, Kubo M, Anderson CA, Weersma RK. Association analyses identify 38 susceptibility loci for inflammatory bowel disease and highlight shared genetic risk across populations. Nat Genet. 2015 Sep;47(9):979–986. PMCID: PMC4881818

28. de Lange KM, Moutsianas L, Lee JC, Lamb CA, Luo Y, Kennedy NA, Jostins L, Rice DL, Gutierrez-Achury J, Ji S-G, Heap G, Nimmo ER, Edwards C, Henderson P, Mowat C, Sanderson J, Satsangi J, Simmons A, Wilson DC, Tremelling M, Hart A, Mathew CG, Newman WG, Parkes M, Lees CW, Uhlig H, Hawkey C, Prescott NJ, Ahmad T, Mansfield JC, Anderson CA, Barrett JC. Genome-wide association study implicates immune activation of multiple integrin genes in inflammatory bowel disease. Nat Genet. 2017 Feb;49(2):256–261. PMCID: PMC5289481

29. Okou DT, Kugathasan S. Role of genetics in pediatric inflammatory bowel disease. Inflamm Bowel Dis. 2014 Oct;20(10):1878–1884. PMCID: PMC4201539

30. McGovern DPB, Kugathasan S, Cho JH. Genetics of Inflammatory Bowel Diseases. Gastroenterology. 2015 Oct;149(5):1163–1176.e2. PMCID: PMC4915781

31. Cutler DJ, Zwick ME, Okou DT, Prahalad S, Walters T, Guthery SL, Dubinsky M, Baldassano R, Crandall WV, Rosh J, Markowitz J, Stephens M, Kellermayer R,

Pfefferkorn M, Heyman MB, LeLeiko N, Mack D, Moulton D, Kappelman MD, Kumar A, Prince J, Bose P, Mondal K, Ramachandran D, Bohnsack JF, Griffiths AM, Haberman Y, Essers J, Thompson SD, Aronow B, Keljo DJ, Hyams JS, Denson LA, PRO-KIIDS Research Group, Kugathasan S. Dissecting Allele Architecture of Early Onset IBD Using High-Density Genotyping. PloS One. 2015;10(6):e0128074. PMCID: PMC4476779

32. Chen G-B, Lee SH, Brion M-JA, Montgomery GW, Wray NR, Radford-Smith GL, Visscher PM, International IBD Genetics Consortium. Estimation and partitioning of (co)heritability of inflammatory bowel disease from GWAS and immunochip data. Hum Mol Genet. 2014 Sep 1;23(17):4710–4720. PMCID: PMC4119411

33. Sadee W, Hartmann K, Seweryn M, Pietrzak M, Handelman SK, Rempala GA. Missing heritability of common diseases and treatments outside the protein-coding exome. Hum Genet. 2014 Oct;133(10):1199–1215. PMCID: PMC4169001

34. Zaitlen N, Kraft P. Heritability in the genome-wide association era. Hum Genet. 2012 Oct;131(10):1655–1664. PMCID: PMC3432754

35. Tennessen JA, Bigham AW, O'Connor TD, Fu W, Kenny EE, Gravel S, McGee S, Do R, Liu X, Jun G, Kang HM, Jordan D, Leal SM, Gabriel S, Rieder MJ, Abecasis G, Altshuler D, Nickerson DA, Boerwinkle E, Sunyaev S, Bustamante CD, Bamshad MJ, Akey JM, Broad GO, Seattle GO, NHLBI Exome Sequencing Project. Evolution and functional impact of rare coding variation from deep sequencing of human exomes. Science. 2012 Jul 6;337(6090):64–69. PMCID: PMC3708544

36. Marth GT, Yu F, Indap AR, Garimella K, Gravel S, Leong WF, Tyler-Smith C, Bainbridge M, Blackwell T, Zheng-Bradley X, Chen Y, Challis D, Clarke L, Ball EV, Cibulskis K, Cooper DN, Fulton B, Hartl C, Koboldt D, Muzny D, Smith R, Sougnez C, Stewart C, Ward A, Yu J, Xue Y, Altshuler D, Bustamante CD, Clark AG, Daly M, DePristo M, Flicek P, Gabriel S, Mardis E, Palotie A, Gibbs R, 1000 Genomes Project. The functional spectrum of low-frequency coding variation. Genome Biol. 2011 Sep 14;12(9):R84. PMCID: PMC3308047

37. Lek M, Karczewski KJ, Minikel EV, Samocha KE, Banks E, Fennell T, O'Donnell-Luria AH, Ware JS, Hill AJ, Cummings BB, Tukiainen T, Birnbaum DP, Kosmicki JA, Duncan LE, Estrada K, Zhao F, Zou J, Pierce-Hoffman E, Berghout J, Cooper DN, Deflaux N, DePristo M, Do R, Flannick J, Fromer M, Gauthier L, Goldstein J, Gupta N, Howrigan D, Kiezun A, Kurki MI, Moonshine AL, Natarajan P, Orozco L, Peloso GM, Poplin R, Rivas MA, Ruano-Rubio V, Rose SA, Ruderfer DM, Shakir K, Stenson PD, Stevens C, Thomas BP, Tiao G, Tusie-Luna MT, Weisburd B, Won H-H, Yu D, Altshuler DM, Ardissino D, Boehnke M, Danesh J, Donnelly S, Elosua R, Florez JC, Gabriel SB, Getz G, Glatt SJ, Hultman CM, Kathiresan S, Laakso M, McCarroll S, McCarthy MI, McGovern D, McPherson R, Neale BM, Palotie A, Purcell SM, Saleheen D, Scharf JM, Sklar P, Sullivan PF, Tuomilehto J, Tsuang MT, Watkins HC, Wilson JG, Daly MJ, MacArthur DG, Exome Aggregation Consortium. Analysis of protein-coding genetic variation in 60,706 humans. Nature. 2016 Aug 18;536(7616):285–291. PMCID: PMC5018207

38. Keinan A, Clark AG. Recent explosive human population growth has resulted in an excess of rare genetic variants. Science. 2012 May 11;336(6082):740–743. PMCID: PMC3586590

39. Loftus EV. Clinical epidemiology of inflammatory bowel disease: Incidence, prevalence, and environmental influences. Gastroenterology. 2004 May;126(6):1504–1517. PMID: 15168363

40. Benchimol EI, Fortinsky KJ, Gozdyra P, Van den Heuvel M, Van Limbergen J, Griffiths AM. Epidemiology of pediatric inflammatory bowel disease: a systematic review of international trends. Inflamm Bowel Dis. 2011 Jan;17(1):423–439. PMID: 20564651

41. Virta LJ, Saarinen MM, Kolho K-L. Inflammatory Bowel Disease Incidence is on the Continuous Rise Among All Paediatric Patients Except for the Very Young: A Nationwide Registry-based Study on 28-Year Follow-up. J Crohns Colitis. 2017 Feb;11(2):150–156. PMID: 27555642

42. Ananthakrishnan AN. Epidemiology and risk factors for IBD. Nat Rev Gastroenterol Hepatol. 2015 Apr;12(4):205–217. PMID: 25732745

43. Bernstein CN, Shanahan F. Disorders of a modern lifestyle: reconciling the epidemiology of inflammatory bowel diseases. Gut. 2008 Sep;57(9):1185–1191. PMID: 18515412

44. Molodecky NA, Kaplan GG. Environmental risk factors for inflammatory bowel disease. Gastroenterol Hepatol. 2010 May;6(5):339–346. PMCID: PMC2886488

45. Abegunde AT, Muhammad BH, Bhatti O, Ali T. Environmental risk factors for inflammatory bowel diseases: Evidence based literature review. World J Gastroenterol. 2016 Jul 21;22(27):6296–6317. PMCID: PMC4945988

46. Halfvarson J, Jess T, Magnuson A, Montgomery SM, Orholm M, Tysk C, Binder V, Järnerot G. Environmental factors in inflammatory bowel disease: A co-twin control study of a Swedish-Danish twin population. Inflamm Bowel Dis. 2006 Oct 1;12(10):925–933.

47. Hughes JB, Hellmann JJ, Ricketts TH, Bohannan BJ. Counting the uncountable: statistical approaches to estimating microbial diversity. Appl Environ Microbiol. 2001 Oct;67(10):4399–4406. PMCID: PMC93182

48. Shreiner AB, Kao JY, Young VB. The gut microbiome in health and in disease. Curr Opin Gastroenterol. 2015 Jan;31(1):69–75. PMCID: PMC4290017

49. Erickson AR, Cantarel BL, Lamendella R, Darzi Y, Mongodin EF, Pan C, Shah M, Halfvarson J, Tysk C, Henrissat B, Raes J, Verberkmoes NC, Fraser CM, Hettich RL, Jansson JK. Integrated metagenomics/metaproteomics reveals human host-microbiota signatures of Crohn's disease. PloS One. 2012;7(11):e49138. PMCID: PMC3509130

50. Frank DN, St Amand AL, Feldman RA, Boedeker EC, Harpaz N, Pace NR. Molecular-phylogenetic characterization of microbial community imbalances in human inflammatory bowel diseases. Proc Natl Acad Sci U S A. 2007 Aug 21;104(34):13780–13785. PMCID: PMC1959459

51. Frank DN, Robertson CE, Hamm CM, Kpadeh Z, Zhang T, Chen H, Zhu W, Sartor RB, Boedeker EC, Harpaz N, Pace NR, Li E. Disease phenotype and genotype are associated with shifts in intestinal-associated microbiota in inflammatory bowel diseases. Inflamm Bowel Dis. 2011 Jan;17(1):179–184. PMCID: PMC3834564

52. Gevers D, Kugathasan S, Denson LA, Vázquez-Baeza Y, Van Treuren W, Ren B, Schwager E, Knights D, Song SJ, Yassour M, Morgan XC, Kostic AD, Luo C, González A, McDonald D, Haberman Y, Walters T, Baker S, Rosh J, Stephens M, Heyman M, Markowitz J, Baldassano R, Griffiths A, Sylvester F, Mack D, Kim S, Crandall W, Hyams J, Huttenhower C, Knight R, Xavier RJ. The treatment-naive microbiome in new-onset Crohn's disease. Cell Host Microbe. 2014 Mar 12;15(3):382–392. PMCID: PMC4059512

53. Manichanh C, Rigottier-Gois L, Bonnaud E, Gloux K, Pelletier E, Frangeul L, Nalin R, Jarrin C, Chardon P, Marteau P, Roca J, Dore J. Reduced diversity of faecal microbiota in Crohn's disease revealed by a metagenomic approach. Gut. 2006 Feb;55(2):205–211. PMCID: PMC1856500

54. Morgan XC, Tickle TL, Sokol H, Gevers D, Devaney KL, Ward DV, Reyes JA, Shah SA, LeLeiko N, Snapper SB, Bousvaros A, Korzenik J, Sands BE, Xavier RJ, Huttenhower C. Dysfunction of the intestinal microbiome in inflammatory bowel disease and treatment. Genome Biol. 2012;13(9):R79. PMCID: PMC3506950

55. Sepehri S, Kotlowski R, Bernstein CN, Krause DO. Microbial diversity of inflamed and noninflamed gut biopsy tissues in inflammatory bowel disease. Inflamm Bowel Dis. 2007 Jun;13(6):675–683. PMID: 17262808

56. Sartor RB. Microbial influences in inflammatory bowel diseases. Gastroenterology. 2008 Feb;134(2):577–594. PMID: 18242222

57. Casellas F, Borruel N, Papo M, Guarner F, Antolín M, Videla S, Malagelada JR. Antiinflammatory effects of enterically coated amoxicillin-clavulanic acid in active ulcerative colitis. Inflamm Bowel Dis. 1998 Feb;4(1):1–5. PMID: 9552221

58. Shaw KA, Bertha M, Hofmekler T, Chopra P, Vatanen T, Srivatsa A, Prince J, Kumar A, Sauer C, Zwick ME, Satten GA, Kostic AD, Mulle JG, Xavier RJ, Kugathasan S. Dysbiosis, inflammation, and response to treatment: a longitudinal study of pediatric subjects with newly diagnosed inflammatory bowel disease. Genome Med. 2016;8(1):75. PMID: 27412252

CHAPTER II. Gastrointestinal health in classic galactosemia

Coauthors: Jennifer G. Mulle, Michael P. Epstein, and Judith L. Fridovich-Keil

SUMMARY

Classic galactosemia (CG) is an autosomal recessive disorder of galactose metabolism that affects approximately 1/50,000 live births in the United States. Following exposure to milk, which contains large quantities of galactose, affected infants may become seriously ill. Early identification by newborn screening with immediate dietary galactose restriction minimizes or prevents the potentially lethal acute symptoms of CG. However, more than half of individuals with CG still experience long-term complications including cognitive disability, behavioral problems, and speech impairment. Anecdotal reports have also suggested frequent gastrointestinal (GI) problems, but this outcome has not been systematically addressed. In this study we explored the prevalence of GI symptoms among 183 children and adults with CG (cases) and 190 controls. Cases reported 4.5 times more frequent constipation (95% CI 1.8-11.5) and 4.2 times more frequent nausea (95% CI 1.2-15.5) than controls. Cases with genotypes predicting residual GALT activity reported less frequent constipation than cases without predicted GALT activity but this difference was not statistically significant. Because the rigor of dietary galactose restriction varies among individuals with galactosemia, we further tested whether GI symptoms associated with diet in infancy. Though constipation was almost four times as common among cases reporting a more restrictive diet in infancy, this difference was not statistically significant. These data confirm that certain GI symptoms are more common in classic galactosemia compared to controls and suggest future studies should investigate associations with residual GALT activity and dietary galactose restriction in early life.

INTRODUCTION

Classic galactosemia (CG) results from profound deficiency of galactose-1-phosphate uridylyltransferase (GALT) activity and affects approximately 1/50,000 live births in the United States (Pyhtila et al 2015). Following exposure to milk, which contains large quantities of galactose, affected infants can become seriously ill and die if not immediately switched to a low-galactose formula (Berry 2014). Early identification by newborn screening and rapid dietary intervention generally prevents or resolves the potentially lethal acute symptoms of CG (Berry 2014).

Despite early diagnosis and intervention, most individuals with CG experience long-term complications that can include multiple developmental disabilities (Kaufman et al 1995, Waggoner et al 1990). The majority of girls and women with CG also experience primary or premature ovarian insufficiency (Fridovich-Keil et al 2011, Kaufman et al 1979, Spencer et al 2013, Waggoner et al 1990). For years, anecdotal reports of increased gastrointestinal (GI) health problems in CG have been shared by families but not investigated formally. To determine whether children and adults with CG indeed experience increased prevalence of GI symptoms, we performed a systematic survey of GI health among 183 individuals with CG (cases) and 190 controls. To address possible genetic and environmental modifiers of GI outcome in CG we also gathered *GALT* genotype and retrospective diet information for each case.

More than 300 different *GALT* variants have been reported (http://arup.utah.edu/database/GALT/GALT_display.php; (Calderon et al 2007)) and this allelic heterogeneity has been a suspected modifier of outcomes (e.g. (Tyfield et al 1999)). Recently, trace residual GALT activity predicted from a yeast model system for specific genotypes was associated with both improved scholastic (Ryan et al 2013) and ovarian outcomes (Spencer et al 2013), suggesting that residual GALT activity might also modify GI outcomes in CG.

Another potential modifier of GI outcomes in CG is diet. While the majority of healthcare providers recommend lifelong dietary restriction of milk and other dairy products for their patients with CG, some also recommend restriction of non-dairy foods that contain low levels of galactose (Gleason et al 2010, van Calcar et al 2014). As a result, rigor of dietary galactose restriction varies among individuals with CG.

Using GI health outcomes, *GALT* genotype, and retrospective diet information collected for volunteers in our study we sought to address (1) whether cases reported more frequent GI problems than controls, (2) whether presence of predicted residual GALT activity associated with frequency of GI symptoms among cases, and (3) whether rigor of dietary galactose restriction in infancy associated with frequency of GI symptoms among cases.

MATERIALS AND METHODS

Study volunteers

Children and adults with classic galactosemia were ascertained by referral from healthcare professionals or self-referral, often following interactions facilitated by the Galactosemia Foundation (www.galactosemia.org). Controls were recruited in two ways. First, unaffected siblings of CG volunteers participating in the study were recruited as “related controls.” Second, “unrelated controls” were recruited by posting an IRB-approved flyer to the Centers for Disease Control (CDC) parents’ email listserv (a widely subscribed electronic mailing list). All study volunteers completed informed consent prior to joining this IRB-approved study (Emory IRB00024933, PI: JL Fridovich-Keil).

Gastrointestinal health parent- or self-report survey

We developed the gastrointestinal (GI) health survey used in this study to assess how frequently each study volunteer experienced different GI symptoms including abdominal pain, constipation, diarrhea, heartburn, nausea, and vomiting (see Supplemental data). The survey was administered online via Emory's HIPAA-compliant Feedback Server in 2013 and 2014. Surveys were completed by parent/guardians for their children, or by adults for themselves. Symptoms of each GI outcome were rated by frequency: "never", "less than once a month", "at least once a month", "weekly", or "daily". We classified problems that were experienced more than once a month as "frequent" and problems experienced less than or equal to once a month as "infrequent."

In addition to measures of GI health, we also gathered data on potential covariates including probiotic/antibiotic usage within the prior 6 months, date of birth, gender, race, and ethnicity. Our study design did not allow calculation of an overall response rate because the survey distribution routes used prevented us from knowing how many eligible people received the invitation to participate.

Dietary restriction parent-report survey

Our diet survey was developed to assess historical dietary information retrospectively. For individuals with classic galactosemia, this included which food groups were restricted in infancy to avoid galactose exposure. Like the GI health survey, our diet survey was administered online via Emory's Feedback Server. One hundred fourteen cases who responded to the diet survey also completed the GI health survey. We scored dietary restriction of milk/dairy only or milk/dairy plus legumes as "moderate" and restriction of milk/dairy, legumes, plus other food groups (e.g. some fruits or vegetables) as "strict."

Predicted residual GALT activity

We collected all available *GALT* genotype information for cases and calculated predicted GALT enzyme activity using results from a previously described yeast expression system (Fridovich-Keil and Jinks-Robertson 1993, Riehman et al 2001). Cases were classified as having either ≥0.4% predicted residual GALT activity (approximately the limit of detection of the enzyme assay) or <0.4% predicted residual GALT activity based on the average of activities predicted for their two *GALT* alleles.

Statistical analyses

We performed all statistical analyses in R (https://www.r-project.org/). Because there are no good estimates for the relevant population prevalence of the GI symptoms we report, we used the reported symptoms in our controls as a guide for calculating the statistical power of our study. Reported symptoms ranged from a prevalence of 1.6% (nausea) to 6.3% (heartburn) in our controls. With our sample size, we had 80% power to detect an increase in cases of 5.2-7.8%.

To determine if there were significant differences in population structure or outcomes between related and unrelated control groups, we used chi-square tests, t-tests, and Fisher's exact tests, as appropriate. For case/control comparisons we performed logistic regression using generalized estimating equations (GEE) (Liang and Zeger 1986) to account for within-family correlations. With "frequent" (symptom experienced more than once a month) or "infrequent" (symptom experienced once a month or less) GI symptom as the outcome, our full models included “case” or “control” diagnosis as the predictor of interest and age, gender, probiotic use, and antibiotic use as covariates. Covariates were tested individually for association with outcome and retained in our reduced model if their p-value was ≤0.1. To adjust for multiple testing of various GI symptoms, we used permutation procedures that randomly shuffled each subject’s set of GI symptoms within the study. To perform permutations while maintaining the existing familial structure in the dataset, we performed separate shuffling of unrelated subjects (unrelated cases and unrelated controls) and related subjects (related cases and controls). For related subjects, we assigned each individual’s set of GI symptoms randomly among subjects from the same family. Symptoms significantly associated or close to associated with diagnosis (p≤0.1) were subjected to 10,000 such permutations of outcome to account for multiple testing.

For case-only diet and residual activity analyses we used Fisher’s exact tests because all cases were unrelated (independent observations) and at least one cell in each comparison included fewer than five individuals.

RESULTS

Study population characteristics

In total, 499 people responded to our GI health survey. However, we restricted analyses to respondents between ages 1 to 55 because of differences in distribution of cases and controls outside this range. Additionally, because >90% of our cases self-reported as white and non-Hispanic, we restricted our analyses to this demographic. We ultimately analyzed GI health survey results from 183 children and adults with classic galactosemia (cases) and 190 children and adults without classic galactosemia (controls). These 190 controls included 75 volunteers who were related to cases in the study, and 115 unrelated volunteers. There were only 4 reports of frequent vomiting in our entire cohort (evenly split between cases and controls), so we excluded this outcome from our analysis.

Notably, GI outcomes were not significantly different between the related and unrelated control groups for abdominal pain, constipation, heartburn, or nausea (Fisher's exact test p=1, p=1, p=1, and p=0.3, respectively). This is important because it suggests there were not strong "household" effects impacting the outcomes studied here. However, 13 unrelated controls reported severe diarrhea, compared to 0 reports in the related control group (Fisher's exact test p=0.002). Privacy issues prevented us from re-contacting these 13 individuals for clarification, and because they did not clearly differ from other controls with regard to other parameters assessed, we did not exclude them from the study but instead did not test diarrhea as an outcome in subsequent analyses.

Unrelated controls were significantly older than related controls (24 ± 14 years old compared to 18 ± 14 years old; t-test p=0.003), and overall this combined control group was significantly older than the case group (22 ± 14 years old compared to 16 ± 12 years old; p=1E-05 based on t-test). We therefore tested age as a potential covariate in all case/control analyses. We likewise tested gender as a potential covariate because of differences in gender distribution between related and unrelated controls (63% and 42% female, respectively; chi-square p=0.008). However, the gender distributions of cases and combined controls were not significantly different (55% and 50% female, respectively). **Table S1** shows a summary of the numbers of cases and controls used in all comparisons.

Individuals with classic galactosemia experience some GI symptoms more frequently than controls

Our final GEE models comparing frequency of GI symptoms between cases and controls included: probiotic usage for abdominal pain and constipation, age for heartburn, and antibiotic usage for nausea (**Table S2** provides a summary of full and reduced models). Of note, both antibiotic and probiotic usage were similar between cases and controls, so this was not a confounding variable (**Table S3**). Gender did not approach significant association with any outcome (p>0.1 for all analyses) and therefore was not included in any of our reduced models.

Using case/control status as a binary predictor in our GEE framework, we were able to calculate adjusted odds ratios for experience of frequent symptoms controlled for relevant covariates (**Table 1**). Comparing unadjusted prevalence numbers we found that a diagnosis of classic galactosemia was significantly associated with a 4.5-fold increase in frequent constipation (95% CI 1.8-11.5, permuted p=0.0008) and a 4.2-fold increase in frequent nausea (95% CI 1.6-18.7, permuted p=0.03) (Figure S1). Differences in abdominal pain (2.1-fold, 95% CI 0.8-5.4) and heartburn (1.2-fold, 95% CI 0.5-2.9) were not significant.

Residual GALT activity and GI health

For the case-only residual GALT activity question, we had *GALT* genotype information for 153 of the 183 cases who completed our GI health survey, 29 of whom had *GALT* alleles either not yet tested or not appropriate for study in our yeast system (Fridovich-Keil and Jinks-Robertson 1993, Riehman et al 2001). Of the 124 cases for whom we could predict GALT activity, 27 had ≥0.4% predicted residual GALT activity and 97 had <0.4%. *GALT* genotypes and predicted activities for this study are summarized in **Table S4**.

While cases with predicted residual GALT activity ≥0.4% reported one-fifth the frequent constipation reported by cases with lower predicted activity (**Table 2**, odds ratios, upper rows and Figure S2A, unadjusted prevalence), this difference was not statistically significant (95% CI 0.005-1.6, p=0.2). There was no evidence of a difference in frequency of nausea between the two groups (p=1).

Dietary restriction in infancy and GI health

We received completed parent-response diet surveys with historical galactose restriction data for 114 of the 183 cases who also completed our GI health survey. The diet survey asked respondents to indicate categories of food restricted in infancy to avoid galactose. Options included: (1) milk and other high galactose dairy products, (2) legumes, (3) some fruits, (4) some vegetables, and (5) other. Milk and other dairy products were universally restricted among cases in infancy, and most families also restricted legumes which have long been considered a significant source of galactose (Acosta and Gross 1995). A smaller proportion of families also restricted some fruits/vegetables, or other foods believed to contain potentially concerning levels of galactose. Because of this distribution, we defined diets restricting only milk/dairy or milk/dairy plus legumes as “moderate” and diets restricting these plus any additional food groups (e.g. some fruits and vegetables) as “strict.”

Nausea was not significantly different between “moderate” and “strict” dietary groups (**Table 2**, odds ratios, lower rows and **Figure S2B**, unadjusted prevalence). We noted a 3.9-fold increase in odds for frequent constipation in the “strict” group but this result was not significant (95% CI 0.8-38.3, p=0.1). Importantly, our findings were not confounded by the effect of residual GALT activity, because similar proportions of cases in the “moderate” and "strict" dietary groups had ≥0.4% predicted residual activity (19% and 22%, respectively, **Table S5**).

DISCUSSION

The main goal of this study was to test whether there was a link between classic galactosemia and specific GI symptoms among a relatively large cohort of volunteers. Our results demonstrated that cases indeed reported significantly more frequent constipation and nausea than controls. Specifically, we found that individuals with classic galactosemia in our study were 4.5 times more likely to report frequent constipation and 4.2 times more likely to report frequent nausea compared to controls. It is important to note that while these increases were significant, the absolute prevalence of each symptom in our CG study group was fairly low at 11% and 5%, respectively. Therefore, while individuals with classic galactosemia do experience these GI problems more frequently than controls, these symptoms are not universal.

As a first step toward identifying possible genetic and environmental modifiers of GI health outcomes in classic galactosemia we addressed two obvious possibilities: predicted residual GALT activity and diet in infancy. We found suggestive trends for residual GALT activity: cases with ≥0.4% predicted residual GALT activity reported less frequent constipation than individuals with <0.4% predicted residual GALT activity (**Table 2**). We saw no evidence of a difference in frequency of nausea. However, a larger study is needed to confirm or refute the significance of these results.

Considering dietary galactose restriction in infancy, we noted a nearly four-fold increase in reported frequent constipation among cases on strict compared to moderate galactose restriction in infancy. This difference was not statistically significant, but our sample size may not have been adequately powered to detect a difference. We saw no notable difference in the frequency of nausea between the two diet groups.

We did not have concurrent GI health and general nutritional information for our study cohort. It is therefore possible that cases on more restrictive diets in infancy also followed more restrictive diets later in life, potentially leading to lower fiber intake due to a reduction in fruit and/or vegetable consumption. A larger study, with data gathered concurrently for diet and GI symptoms, will be needed to test this possibility. We also did not have information concerning a number of other factors that might have potentially influenced the GI outcomes we measured here, including type of milk substitute consumed, if any, presence or absence of calcium supplementation, psychosocial distress or psychiatric comorbidity, alcohol ingestion, obesity, or use of medications not covered by our survey.

Because classic galactosemia is a rare disorder with limited treatment options, individuals experiencing complications may be more likely to participate in research than those not experiencing complications, resulting in ascertainment bias. However, our observation that less than 12% of cases reported “frequent” experience for each GI symptom helps counter the concern that only those with frequent GI problems were motivated to participate in this study.

One other study limitation is the retrospective nature of our diet survey. Because classic galactosemia is a rare condition (1/50,000 live births), it took many years to assemble our study cohort, at all times welcoming cases of any age to join. While recall bias is therefore potentially a concern, there was no practical way to conduct this study otherwise. Of note, we have anecdotally found that parents of children with classic galactosemia tend to remember incredible detail of their child’s early diet, perhaps because they worried about it so much.

Another potential limitation is our control group. We originally wanted to use siblings of cases to control for shared environment and genetics. However, we worried that parents raising a child with classic galactosemia might be so focused on the considerable health needs of their affected child they might under-report possible health concerns for their non-CG child. A comparison of related and unrelated control groups demonstrated no significant differences in reported frequency of GI symptoms between the two groups (with the exception of frequent vomiting in a small number of unrelated controls as a clear outlier). Additionally, performing GEE analysis of binary outcome data allowed us to account for within-family correlations that could have biased our results.

Importantly, our findings open up new avenues of investigation into pathophysiology of CG and possibilities for therapeutic intervention. One potential explanation for increased GI problems in CG is that defective glycosylation due to perturbation in UDP sugar substrate pools might impact the mucosal layer of the gut, compromising gut barrier function and potentially commensal bacterial population structure (reviewed in (Bergstrom and Xia 2013)). A "leaky" gut, microbiome dysbiosis, or both, could help explain increased GI problems as well as some of the other complications commonly seen in CG.

Importantly, diet also has a significant impact on establishment of the gut microbiome (Albenberg and Wu 2014, David et al 2014), and the diet of infants and children with CG is fundamentally altered because of restriction of galactose-containing foods. Deficiency of the probiotic effect of milk and other dairy products alone could result in differences in the gut microbiome between cases and controls. Perhaps the most appealing aspect of testing this hypothesis is that it could offer opportunities for therapeutic intervention such as dietary supplementation with appropriate probiotics.

ACKNOWLEDGMENTS

First, we thank the Galactosemia Foundation and the many individuals and families who participated in this study; without them none of this work would have been possible. We are also grateful to Drs. Tanvi Dhere and Sandy van Calcar for invaluable guidance in the early stages of this project, and to Dr. K. Alaine Broadaway for early assistance with R. This study was funded in part by NIH grant R01 DK059904 (to JLFK). KAS was supported at different times by funds from training grant ID#1008188 (BWF), training grant T32GM008490 (NIH) and NRSA F31DK107229 (NIH).

REFERENCES

Acosta PB, Gross KC (1995) Hidden sources of galactose in the environment. *Eur J Pediatr* 154: S87-92.

Albenberg L, Wu G (2014) Diet and the intestinal microbiome: associations, functions, and implications for health and disease. *Gastroenterology* 146: 1564-1572.

Bergstrom K, Xia L (2013) Mucin-type O-glycans and their roles in intestinal homeostasis. *Glycobiology* 23: 1026-1037.

Berry G (2014) Classic Galactosemia and Clinical Variant Galactosemia. In R. Pagon, M. Adam, H. Ardinger, T. Bird, C. Dolan, C. Fong, R. Smith and K. Stephens, eds. *GeneReviews*®. Seattle (WA): University of Washington, Seattle,

Calderon F, Phansalkar A, Crockett D, Miller M, Mao R (2007) Mutation database for the galactose-1-phosphate uridyltransferase (<I>GALT</I>) gene. *Human Mutation* 28: 939-943.

David LA, Maurice CF, Carmody RN et al (2014) Diet rapidly and reproducibly alters the human gut microbiome. *Nature* 505: 559-63.

Fridovich-Keil JL, Gubbels CS, Spencer JB, Sanders RD, Land JA, Rubio-Gozalbo E (2011) Ovarian function in girls and women with GALT-deficiency galactosemia. *J Inherit Metab Dis* 34: 357-66.

Fridovich-Keil JL, Jinks-Robertson S (1993) A yeast expression system for human galactose-1-phosphate uridylyltransferase. *Proc Natl Acad Sci U S A* 90: 398-402.

Kaufman F, Kogut MD, Donnell GN, Koch H, Goebelsmann U (1979) Ovarian failure in galactosaemia. *Lancet* 2: 737-8.

Kaufman FR, McBride-Chang C, Manis FR, Wolff JA, Nelson MD (1995) Cognitive functioning, neurologic status and brain imaging in classical galactosemia. *Eur J Pediatr* 154: S2-5.

Liang KY, Zeger SL (1986) Longitudinal Data-Analysis Using Generalized Linear-Models. *Biometrika* 73: 13-22.

Pyhtila BM, Shaw KA, Neumann SE, Fridovich-Keil JL (2015) Newborn screening for galactosemia in the United States: looking back, looking around, and looking ahead. *JIMD Rep* 15: 79-93.

Riehman K, Crews C, Fridovich-Keil JL (2001) Relationship between genotype, activity, and galactose sensitivity in yeast expressing patient alleles of human galactose-1-phosphate uridylyltransferase. *Journal of Biological Chemistry* 276: 10634-10640.

Ryan EL, Lynch ME, Taddeo E, Gleason TJ, Epstein MP, Fridovich-Keil JL (2013) Cryptic residual GALT activity is a potential modifier of scholastic outcome in school age children with classic galactosemia. *J Inherit Metab Dis* 36: 1049-61.

Spencer JB, Badik JR, Ryan EL et al (2013) Modifiers of ovarian function in girls and women with classic galactosemia. *J Clin Endocrinol Metab* 98: E1257-65.

Tyfield L, Reichardt J, Fridovich-Keil J et al (1999) Classical galactosemia and mutations at the galactose-1-phosphate uridyl transferase (GALT) gene. *Human Mutation* 13: 417-430.

van Calcar SC, Bernstein LE, Rohr FJ, Scaman CH, Yannicelli S, Berry GT (2014) A re-evaluation of life-long severe galactose restriction for the nutrition management of classic galactosemia. *Mol Genet Metab* 112: 191-7.

Waggoner DD, Buist NR, Donnell GN (1990) Long-term prognosis in galactosaemia: results of a survey of 350 cases. *J Inherit Metab Dis* 13: 802-18.

TABLES

Table 1: Odds ratios from logistic regression using generalized estimating equations (GEE) to calculate odds of cases experiencing frequent symptoms compared to controls, with 95% confidence intervals, p-values (^ indicates after 10,000 permutations), and number of observations included in model. Asterisk (*) indicates outcome was significantly higher among cases.

Symptom	Odds ratio for cases	95% CI	p-value	N
Abdominal pain	2.1	0.8, 5.4	0.1^	360
Constipation*	4.5	1.8, 11.5	0.0008^	362
Heartburn	1.2	0.5, 2.9	0.8	356
Nausea*	4.2	1.2, 15.5	0.03^	356

Table 2: Results of Fisher's exact tests for association of <0.4% predicted residual GALT activity (upper set of rows) or strict diet (lower set of rows) with frequent experience of GI symptoms.

Association of ≥0.4% predicted residual GALT activity with frequent experience of GI symptoms				
Symptom	**Odds ratio for ≥0.4%**	**95% CI**	**p-value**	**N**
Constipation	0.2	0.005, 1.6	0.2	122
Nausea	0.8	0.02, 7.4	1	118
Association of strict dietary galactose restriction in infancy with frequent experience of GI symptoms				
Symptom	**Odds ratio for strict diet**	**95% CI**	**p-value**	**N**
Constipation	3.9	0.8, 38.3	0.1	112
Nausea	1.2	0.2, 8.4	1	109

SUPPLEMENTAL TABLES

Table S1: Summary of numbers of volunteers included in analyses

Analysis	Groups	Total N
Case/control volunteers for whom we had GI health outcome data	Case	183
	Control	190
Predicted residual GALT activity (of the 183 cases)	≥ 0.4%	27
	< 0.4%	97
	Unknown	59
Dietary galactose restriction in infancy (of the 183 cases)	Moderate	47
	Strict	67
	Unknown	69

Table S2: Odds ratios, 95% confidence intervals (CI), and p-values (* = significant; ^ = p-value after 10,000 permutations) for full and reduced logistic regression models using case/control data in a generalized estimating equations framework. Variables included in the reduced model are reported: 1=diagnosis, 2=age, 3=gender, 4=probiotic usage, 5=antibiotic usage.

Symptom	Model	OR	95% CI	p-value	N
abdominal pain	full: 1 − 5	2.7	0.9, 7.6	0.1	360
	reduced: 1,4	2.1	0.8, 5.4	0.1^	
constipation*	full: 1 – 5	5.0	1.8, 13.6	0.002	362
	reduced: 1,4	4.5	1.8, 11.5	0.0008^	
heartburn	full: 1 – 5	1.1	0.5, 2.9	0.8	356
	reduced: 1,2	1.2	0.5, 2.9	0.8	
nausea*	full: 1 – 5	5.4	1.4, 20.0	0.03	356
	reduced: 1,5	4.2	1.2, 15.5	0.03^	

Table S3: Summary of probiotic or antibiotic usage in the prior 6 months among cases and controls

Drug / supplement	Diagnosis	Answer	N	%
Antibiotic	**Case**	yes	37	20.2
		no	146	79.8
	Control	yes	48	25.3
		no	142	74.7
Probiotic	**Case**	yes	17	9.3
		no	166	90.7
	Control	yes	30	15.8
		no	160	84.2

Table S4: Summary of *GALT* genotypes and predicted residual GALT activities for cases.

First allele	Second allele	Predicted GALT activity for genotype (% of wild-type)	N	%
Q188R	5kb deletion	0	5	4.0
	A320T	0.45	3	2.4
	D197G	16.6	1	0.8
	E308K	62.1	1	0.8
	E363F	35.1	1	0.8
	K285N	0	11	8.9
	L195P	0.4	11	8.9
	M142K	0	1	0.8
	Q188R	0	66	53.2
	Q344K	2.75	6	4.8
	R148Q	0	1	0.8
	R201H	31.4	1	0.8
	R204X	0	2	1.6
	R259W	0	1	0.8
	R333G	0.3	1	0.8
	R333W	0	1	0.8
	Y209C	6.8	1	0.8
K285N	5kb deletion	0	1	0.8
	D98N	9.55	1	0.8
	R148Q	0	2	1.6
	R204X	0	1	0.8
	Y209C	6.8	1	0.8
5kb deletion	5kb deletion	0	1	0.8
	R231C	0	1	0.8
M142K	R259W	0	2	1.6
		TOTAL	124	100

Table S5: Distribution of predicted residual GALT activity levels among cases categorized by rigor of dietary galactose restriction in infancy.

Rigor of dietary galactose restriction when <12 months old	Predicted GALT activity	N	%
Moderate	≥0.4%	6	19.4
	<0.4%	25	80.6
Strict	≥0.4%	11	22
	<0.4%	39	78
Unknown	≥0.4%	10	23.3
	<0.4%	33	76.7

SUPPLEMENTAL FIGURES

Figure S1: The unadjusted overall percentage of cases (shaded bars) and controls (open bars) reporting frequent experience of the indicated GI symptom is shown. Percentage and cohort size are indicated above each bar. Of note, the exact cohort sizes vary slightly between symptoms because of missing data in some survey responses. After adjusting for relevant covariates, only constipation and nausea were significantly more likely to be experienced frequently among cases.

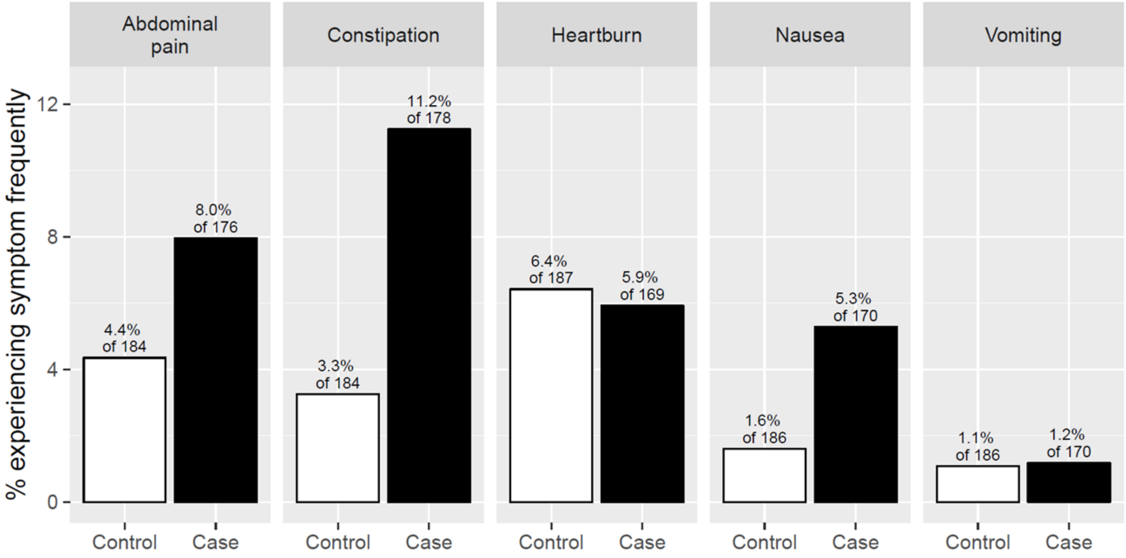
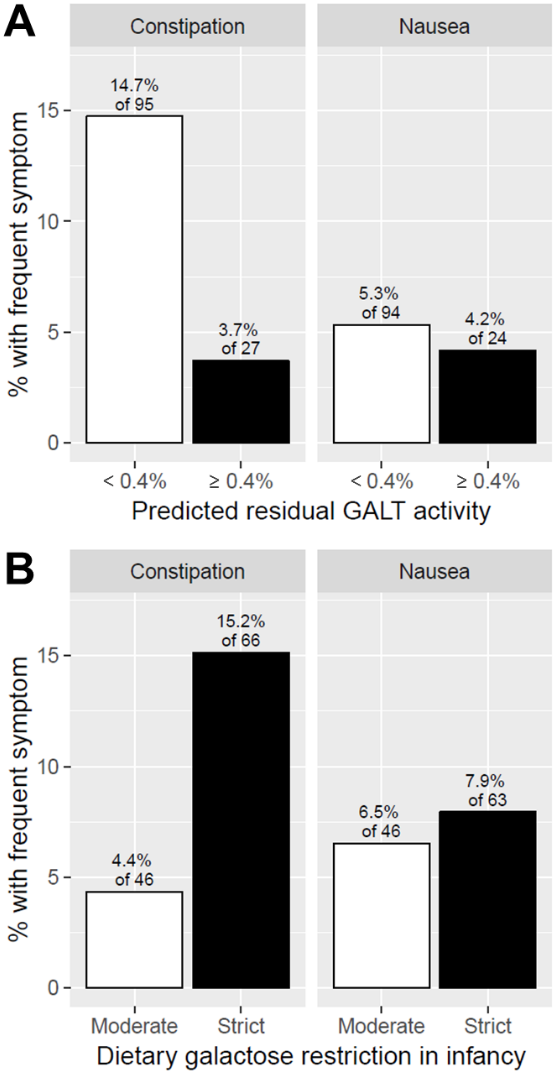
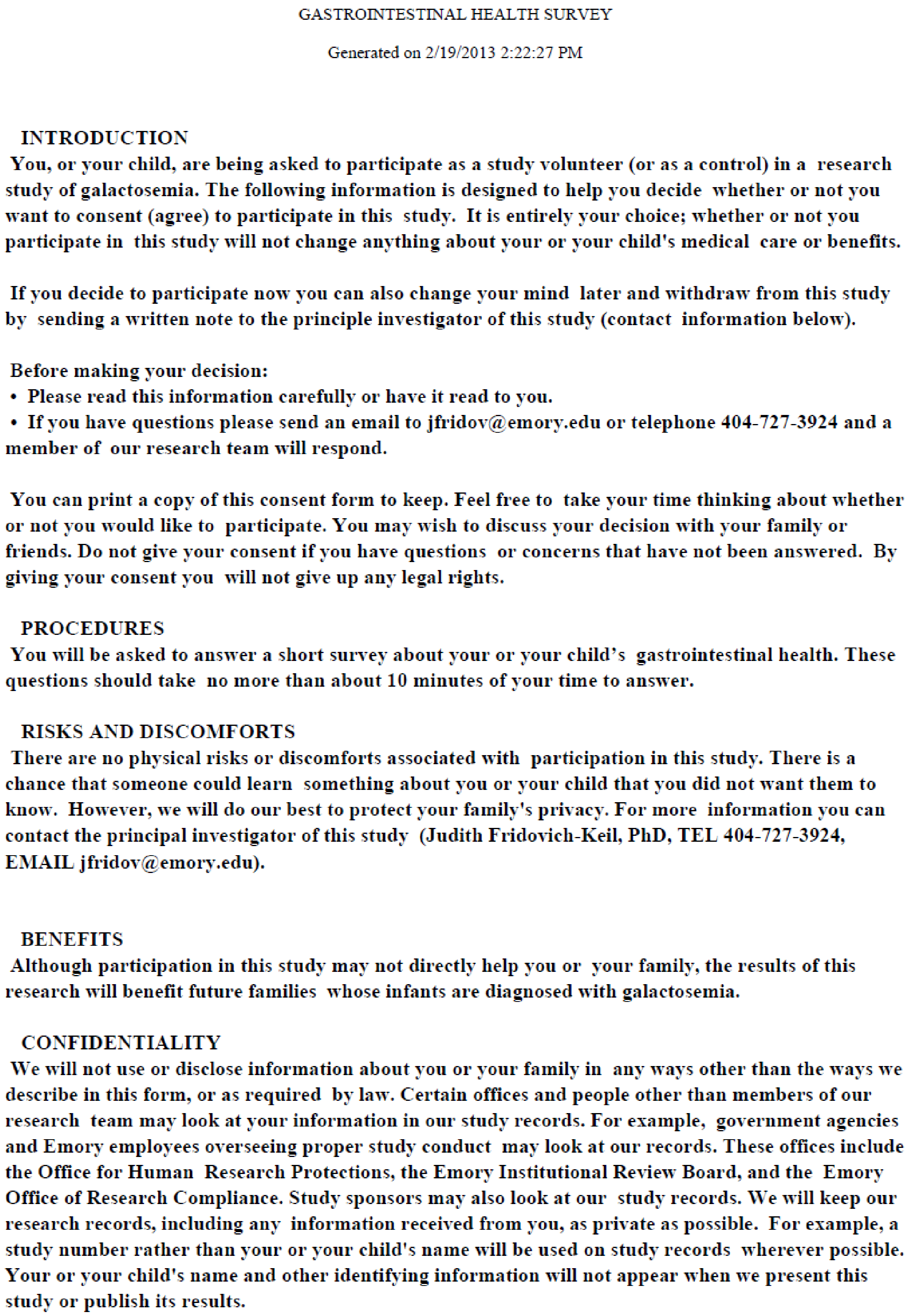


Figure S2: The unadjusted overall percentage of cases reporting frequent GI symptoms broken down by (A) level of predicted residual GALT activity (<0.4%, open bars; or ≥0.4%, shaded bars) and (B) rigor of dietary galactose restriction in infancy, categorized as moderate (restricting only milk/dairy and legumes, open bars) or strict (restricting milk/dairy, legumes, and some fruits, vegetables, or other foods, shaded bars), is shown. Percentage and cohort size are indicated above each bar. Of note, the exact cohort sizes vary slightly between symptoms because of missing data in some survey responses.



SUPPLEMENTAL FILE:



GASTROINTESTINAL HEALTH SURVEY

Generated on 2/19/2013 2:22:27 PM

INTRODUCTION

You, or your child, are being asked to participate as a study volunteer (or as a control) in a research study of galactosemia. The following information is designed to help you decide whether or not you want to consent (agree) to participate in this study. It is entirely your choice; whether or not you participate in this study will not change anything about your or your child's medical care or benefits.

If you decide to participate now you can also change your mind later and withdraw from this study by sending a written note to the principle investigator of this study (contact information below).

Before making your decision:

- Please read this information carefully or have it read to you.
- If you have questions please send an email to jfridov@emory.edu or telephone 404-727-3924 and a member of our research team will respond.

You can print a copy of this consent form to keep. Feel free to take your time thinking about whether or not you would like to participate. You may wish to discuss your decision with your family or friends. Do not give your consent if you have questions or concerns that have not been answered. By giving your consent you will not give up any legal rights.

PROCEDURES

You will be asked to answer a short survey about your or your child's gastrointestinal health. These questions should take no more than about 10 minutes of your time to answer.

RISKS AND DISCOMFORTS

There are no physical risks or discomforts associated with participation in this study. There is a chance that someone could learn something about you or your child that you did not want them to know. However, we will do our best to protect your family's privacy. For more information you can contact the principal investigator of this study (Judith Fridovich-Keil, PhD, TEL 404-727-3924, EMAIL jfridov@emory.edu).

BENEFITS

Although participation in this study may not directly help you or your family, the results of this research will benefit future families whose infants are diagnosed with galactosemia.

CONFIDENTIALITY

We will not use or disclose information about you or your family in any ways other than the ways we describe in this form, or as required by law. Certain offices and people other than members of our research team may look at your information in our study records. For example, government agencies and Emory employees overseeing proper study conduct may look at our records. These offices include the Office for Human Research Protections, the Emory Institutional Review Board, and the Emory Office of Research Compliance. Study sponsors may also look at our study records. We will keep our research records, including any information received from you, as private as possible. For example, a study number rather than your or your child's name will be used on study records wherever possible. Your or your child's name and other identifying information will not appear when we present this study or publish its results.

Study records can be opened by court order. They may also be produced in response to a subpoena or a legal request for production of documents.

Information collected for this study will not go into your or your child's medical records unless you specifically ask us to forward the information to your doctor.

COSTS

It will not cost you or your insurance company any money to participate in this study.

WITHDRAWAL FROM THE STUDY

You have the right to leave this study at any time without penalty. If you or your child wish to withdraw from this study please send a written note to JL Fridovich-Keil, Emory University, Dept. of Human Genetics, 615 Michael Street, Atlanta, GA 30322.

CONTACT INFORMATION

Contact the principal investigator: Judith Fridovich-Keil, PhD at 404-727-3924 or jfridov@emory.edu

- **if you have any questions or concerns about this study or your part in it,**
- **if you feel you or your child have been harmed in any way by participating in this study, or**
- **if you have questions, concerns or complaints about the research.**

Contact the Emory Institutional Review Board at 404-712-0720 or 877-503-9797 or irb@emory.edu

- **if you have questions about your rights as a research participant or**
- **if you have questions, concerns or complaints about the research.**

You may also let the IRB know about your experience as a research participant through our Research Participant Survey at http://www.surveymonkey.com/s/6ZDMW75.

1. Do you agree to participate in this study?

yes	O
no	O

The purpose of this form is to collect information about the gastrointestinal (GI) health of volunteers and controls who have enrolled in our research study "Bases of Pathophysiology and Modifiers of Outcome in Galactosemia" (Emory IRB#00024933, formerly #618-99). The goal of this part of our study is to learn whether children and adults with classic galactosemia are at increased risk for GI disturbances compared with unaffected controls, and also to explore whether there may be some relationship between GI status and other symptoms or outcomes of galactosemia.

If you have questions or concerns, please contact Dr. Judy Fridovich-Keil at 404-727-3924 or jfridov@emory.edu

If you are taking this survey in hardcopy format (printed on paper) please return the completed survey, with any other requested information, to JL Fridovich-Keil PhD, Emory University School of Medicine, Dept. of Human Genetics, Room 325.2 Whitehead Bldg, 615 Michael St., Atlanta, GA 30322 (FAX 404-727-3949). You may also return your completed survey as a scanned PDF or JPG attachment to jfridov@emory.edu .

Thank you so much!

2. Volunteer's name:

volunteer's full name	________________________________

3. Date of birth:

	________________________________

4. Diagnosis:

classic galactosemia	○
control	○
other (explain)	○ ________________________________

5. Gender

Male	○
Female	○

6. Racial Group

White not of Hispanic origin	○
White of Hispanic origin	○
Black not of Hispanic origin	○
Black of Hispanic origin	○
Asian/Pacific Islander	○
Mixed	________________________________
Other	________________________________

7. Today's Date

	________________________________

8. Related to another study volunteer?

No	○
Yes (please explain)	○ ______________________________

9. Name of person filling out this form

	______________________________

10. Email address

	______________________________

11. Telephone number

	______________________________

12. Relationship to volunteer

	______________________________

13. Contact information for volunteer/family

Email	______________________________
Address line 1	______________________________
Address line 2	______________________________
City	______________________________
State	______________________________
Postal Code	______________________________
Country	______________________________
Telephone (if different from above)	______________________________

14. Is the volunteer currently on a galactose-related diet?

no	O
yes	O

15. If yes, which of the following are currently restricted/limited?

milk and milk-containing items like yogurt, soft cheeses, pudding	[ ]
aged hard cheeses, like parmesan	[ ]
tomatoes, tomato sauce, and ketchup	[ ]
legumes	[ ]
fruit (blueberries, strawberries, grapes, and kiwis)	[ ]
other (please explain)	[ ] ________________________________

16. Besides galactose, does the volunteer have any other known food sensitivities/allergies?

no	O
yes	O

17. To which foods?

	________________________________ ________________________________ ________________________________ ________________________________ ________________________________

18. Has the volunteer taken probiotic supplements (not including yogurt) in the past six months? Examples of probiotics include Align, Culturelle, generic Lactobacillus, Acidophilus, or Bifidobacterium supplements.

no	O
yes	O

19. Please list probiotics taken, approximate dates of use, and reason each probiotic was taken. If you need more space, please list additional probiotics or notes in the large text field below.

Probiotic 1	______________________________
Approximate dates of use	______________________________
Reason probiotic 1 was taken	______________________________
Probiotic 2	______________________________
Approximate dates of use	______________________________
Reason probiotic 2 was taken	______________________________
Additional probiotic usage or notes	______________________________ ______________________________ ______________________________ ______________________________ ______________________________

20. Has the volunteer taken antibiotics in the past six months?

no	O
yes	O

21. Please list antibiotics taken, approximate dates of use, and the reason each antibiotic was prescribed. If you need more space, please list additional antibiotics or notes in the large text field below.

Antibiotic 1	______________________________
Approximate dates of use	______________________________
Reason antibiotic 1 was prescribed	______________________________
Antibiotic 2	______________________________
Approximate dates of use	______________________________
Reason antibiotic 2 was prescribed	______________________________
Additional antibiotic usage or notes	______________________________ ______________________________ ______________________________ ______________________________ ______________________________

22. Has the volunteer ever been diagnosed with any of the following conditions (check all that apply)?

Gastroesophageal reflux disease (GERD)	[ ]
Gastroenteritis	[ ]
GI bleeding	[ ]
Appendicitis	[ ]
Colitis	[ ]
Irritable bowel syndrome or disease (IBS or IBD)	[ ]
Crohn&#39;s disease	[ ]
Gallstones	[ ]
Pancreatitis	[ ]
Peptic ulcer disease (PUD)	[ ]
Liver disease (cirrhosis, end-stage liver disease)	[ ]
Diverticulitis	[ ]
Celiac disease	[ ]

23. At what age(s) was the volunteer diagnosed?

	________________________________

24. Please describe any treatments used by the volunteer to manage these condition(s), including medications, supplements, procedures, and/or dietary modifications.

	________________________________ ________________________________ ________________________________ ________________________________ ________________________________

25. In the past year, how often has the volunteer experienced any the following conditions?

	never	less than once a month	at least once a month	weekly	daily
Abdominal pain	○	○	○	○	○
Bleeding gums	○	○	○	○	○
Constipation (defined as bowel movements that are hard to pass or more than three days apart)	○	○	○	○	○
Diarrhea (loose or watery stools)	○	○	○	○	○
Heartburn	○	○	○	○	○
Nausea	○	○	○	○	○
Vomiting	○	○	○	○	○

26. Do these problems occur:

Before meals	[ ]
During meals	[ ]
After meals	[ ]
After consumption of specific foods	[ ]
other (please specify)	[ ] ____________________

27. How often does the volunteer take the following medications/supplements?

	never	less than once a month	at least once a month	weekly	daily
antacids	O	O	O	O	O
laxatives	O	O	O	O	O
fiber supplements	O	O	O	O	O

28. How often does the volunteer move his/her bowels?

less than once a week	O
at least once a week	O
every few days	O
at least once a day	O
other -- please explain	____________________ ____________________ ____________________ ____________________ ____________________

29. Using the chart above, please select the choice that best represents the volunteer's most frequent stool type:

Type 1	O
Type 2	O
Type 3	O
Type 4	O
Type 5	O
Type 6	O
Type 7	O

30. What is the volunteer's next most frequent stool type?

Type 1	O
Type 2	O
Type 3	O
Type 4	O
Type 5	O

Type 6	O
Type 7	O

31. Is there anything else you would like to tell us?

	____________________________ ____________________________ ____________________________ ____________________________ ____________________________

THANK YOU

Please click the "Submit form" button below to save your responses and exit the survey.

CHAPTER III. Genetic variants and pathways implicated in a pediatric inflammatory bowel disease cohort

Coauthors: David J. Cutler, David Okou, Michael P. Epstein, Anne Dodd, Jennifer G. Mulle, Lee A. Denson, Subra Kugathasan, Michael E. Zwick

ABSTRACT

Background and aims: The two most common forms of inflammatory bowel disease (IBD) are Crohn's disease (CD) and ulcerative colitis (UC). CD and UC are severe chronic diseases characterized by relapsing-remitting gastrointestinal inflammation. Around 5% of existing IBD cases in the United States are patients under the age of 20. Studies of these pediatric cohorts can provide unique insights into the genetic architecture of IBD. Large genome-wide association studies of IBD have found more than 200 loci associated with disease but explain only 13.1% of variance in disease liability for CD and 8.2% for UC. In addition to environmental factors, other types of genetic variation such as rare variants likely contribute to disease development.

Methods: We compared exome sequencing of 368 pediatric IBD patients to publicly available exome sequencing (dbGaP) and aggregate frequency data (ExAC). With dbGaP data we performed logistic regression with common variants and optimal unified association tests (SKAT-O) for rare variants with combined annotation dependent depletion score >10. We compared rare variants in our data to ExAC with Fisher exact tests. We then did pathway enrichment analysis on the most significant genes from each comparison.

Results: Many common and rare variants overlapped with known IBD-associated genes (e.g. *NOD2*, *CARD9*). Rare variants were enriched in loci associated with CD (p=0.003) and showed a suggestive enrichment in neutrophil genes (p=0.08). Pathway enrichment analysis implicated many immune-related pathways consistent with our understanding of IBD, especially those involved in cell killing and apoptosis.

Conclusions: Our rare variant findings underscore the importance of genes involved in immune responses in the etiology of inflammatory bowel disease.

INTRODUCTION

Inflammatory bowel disease (IBD) is a disorder characterized by chronic remitting and relapsing gastrointestinal inflammation. The two most common forms of IBD, Crohn's disease (CD) and ulcerative colitis (UC), are most frequently diagnosed in young adults 20-29 years old[1]. However, IBD also frequently occurs in childhood and early adolescence. In the United States (US), the prevalence of IBD for children (<20 years old) was estimated to be 92 cases per 100,000 in 2009, accounting for approximately 5% of prevalent cases[2]. Increasing prevalence[2] and rates of hospitalization[3] for pediatric IBD have been observed in the US, mirroring the trend of increasing IBD incidence in both pediatric[4] and adult[1] populations worldwide.

IBD most frequently presents with abdominal pain and/or diarrhea, but other gastrointestinal symptoms like loss of appetite, nausea, and vomiting may also occur. One large study of 1009 pediatric IBD patients found that 17% presented with at least one extraintestinal manifestation (EIM) such as arthralgia, ankylosing spondylitis, arthritis, erythema nodosum, uveitis, or pancreatitis, with a 33% cumulative probability of experiencing an EIM over four years[5]. Pediatric patients can also experience disease-related growth impairment which is sometimes not recovered even with treatment for IBD[6,7]. Because IBD is a chronic disease, pediatric patients may also face years of medication, a high probability of surgery, and surveillance colonoscopy. For these reasons, better understanding of disease etiology and progression in this group is vital.

IBD is thought to have a strong genetic component, since family history of IBD is the greatest risk factor for disease at all ages. IBD patients with a family history of disease often present at a younger age[8–10], are more likely to experience EIM[8], have perforating disease, and require longer follow-up compared to patients without family history[8,9], likely reflecting an increased genetic liability to disease. Genetic analyses of pediatric cohorts are therefore useful in exploring genetic architecture of IBD.

Large genome-wide association studies (GWAS) of IBD have found more than 200 common loci associated with disease[11,12]. Pathway analysis of associated loci has found an enrichment of immune system genes, especially those related to host response to microbes, and a great deal of overlap with other immune diseases[11]. Findings of studies of common variation in pediatric IBD cohorts generally echo findings in adult populations. One study of greater than 1,000 pediatric-onset IBD cases and 1,600 controls found slightly increased odds ratios for risk alleles also found in adult populations (including the well-known *NOD2*), and greater burden of these common variants was weakly correlated with earlier age of onset in Crohn's disease[13].

A small proportion of disease liability has been explained by common variants in IBD—13.1% in CD and 8.2% in UC[11]—but the contribution of rare variants has not been assessed. This class of genetic variation is important because explosive growth of the human population in recent history has led to a corresponding excess of rare alleles[14], and most variants in protein-coding sequence are at low frequency[15–17]. The availability of public data sets allows us to compare whole exome sequencing (WES) of a pediatric IBD cohort to other WES data[18] and to large databases containing population allele frequency information[17,19]. We can further look at pathways implicated by genes annotated to these rare variants.

METHODS

Ethical approval and recruitment of study participants

Subjects for whole exome sequencing (WES) were selected from patients enrolled in the CCFA sponsored RISK cohort study and the NIH sponsored Emory African-American gene discovery study, for whom DNA had already been collected. RISK is the largest pediatric CD inception cohort in the world, with 1,813 subjects younger than 18 years old with suspected IBD enrolled at 28 North American sites, including Emory University, from November 2008 to June 2012 (ClinicalTrials.gov Identifier: NCT00790543). All patients underwent baseline colonoscopy and histological confirmation of chronic active colitis/ileitis prior to diagnosis and treatment. Once standard and published guidelines were met, patients were diagnosed with CD, UC or inflammatory bowel disease-undetermined (IBD-U). A consistent diagnosis of IBD was required during the one-year follow-up for inclusion into this study. At enrollment and during ongoing prospective follow-up, clinical and laboratory data were obtained for each enrolled patient and submitted to a centralized data management center. All patients were managed according to the dictates of their physicians, not by standardized protocols. The patient-based studies were approved by the Institutional Review Boards at each of the RISK sites. Consent was obtained from parents and adult subjects and assent from pediatric subjects age 11 and above.

Emory case sample collection, processing and exome sequencing

Genomic DNA was extracted from whole blood for a total of 567 early onset IBD samples, of which 553 (97.5 %) passed DNA QC. Library preparation and sequencing of the samples were performed at Broad Institute's Genomics Platform, Cambridge, USA. The libraries were prepared according to the manufacturer instructions using 1 μg of input DNA per sample. DNA was subjected to whole exome capture with the SureSelect Human All Exon 50-Mb Kit (Agilent Technologies) following standard protocols. Library validation was done with the KAPA Library Quantification Kit (KAPA Biosystems) and the whole exome capture libraries were then sequenced on the Illumina HiSeq platform according to standard protocols.

Publicly available datasets

Database of genotypes and phenotypes (dbGaP)[18] *data:* We identified and downloaded control data from the Epi4K (accession phs000653.v2.p1) and ARRA (accession phs000298.v3.p2) studies. SRA files were converted to fastq format using NCBI's SRA Toolkit[20].

Exome Aggregation Consortium (ExAC) (http://exac.broadinstitute.org/)[17,19] *data (version 0.3.1):* For this publicly available data set containing information on 60,706 individuals, we used liftOver to map all sites to hg38 for comparison with our data. We summed minor and total allele counts for the American, Finnish, and non-Finnish European groups and required a site to be typed in >90% of total chromosomes for these groups (at least 76,438 out of 84,930 chromosomes) for inclusion.

dbGaP (raw whole exome sequencing) analysis

We mapped Emory and dbGaP exome sequencing fastq files to hg38 using PEMapper and called variants using PECaller[21]. We then used SeqAnt[22] version 2.0 (Beta 3, https://seqant.genetics.emory.edu/) to get rsID numbers for plink and other annotation information for later analysis.

All following variant quality control (QC) was performed in PLINK 1.9[23–25]. Starting with 866,411 variants in 1,035 controls and 541 cases diagnosed with IBD before age 18, we filtered samples and variants using increasingly stringent completeness criteria until information for all remaining variants and samples was 99% complete. For each study individually (IBD, ARRA, Epi4k), we removed sites that were Bonferroni significant in a Hardy-Weinberg equilibrium test. We then performed a sex check of samples. Cases were removed if their sex was discordant with record review (N=9); other mislabeled sexes were corrected. We checked sample relatedness and removed 8 controls and 10 cases who were 2nd degree or more closely related to another study participant. Table 1A shows characteristics for the 517 remaining IBD patients who passed this first round of quality control.

To adjust for population stratification in our sample we used 10,913 common (minor allele frequency, a.k.a. MAF>0.05) SNPs to calculate principal components (PCs) using EIGENSTRAT[26] and anchoring with HapMap controls as described by Anderson et al[27] (Supplemental Figure 1A). We removed outliers (those with values greater or less than 3 standard deviations away from the mean) for any of the top 7 principal components (those which appeared meaningful with eigenvalues>2), recalculated principal components, and repeated outlier filtering with 4 meaningful PCs, leaving us with a final data set of 625 controls and 368 cases (Supplemental Figure 1B; Table 1B shows basic characteristics for these participants). PCs were recalculated again without HapMap samples (Supplemental Figure 1C) and the four principal components significant by Tracy-Widom tests were used as covariates in regressions.

As an additional filter, we removed variants that were most significantly different (top 2.5%) in Fisher's exact tests comparing our dbGaP controls to ExAC.

Common variant analysis: We performed logistic regression for sites with MAF>0.05 in plink with case/control status as outcome, genotype as predictor of interest, and sex and PCs as covariates. P-values were corrected with genomic control.

SKAT-O analysis: We used SKAT-O[28] in R[29], which performs optimized association tests unifying burden test and sequence kernel association test (SKAT) approaches, to analyze genes annotated to sites with MAF<0.05 and evidence of evolutionary conservation with combined annotation dependent depletion score (CADD) score>10. We tested for enrichment of variants in genes for any gene with 5 or more rare variants. We also lifted over loci associated with IBD from Jostins et al. 2012[11] and Liu et al. 2015[12] to hg38, yielding 201 loci, and tested for enrichment of rare variants 250kb upstream or downstream of CD, UC, or IBD loci as groups (Supplemental Table 1). We also examined whether these variants were enriched in a list of important neutrophil genes (Supplemental Table 2).

ExAC (aggregate allele count) analysis

Rare variant analysis: Using the same set of variants as in the dbGaP analysis (with sites most significantly different between dbGaP and ExAC filtered out), we used Fisher's exact tests to compare rare variant sites (MAF<0.05) between our IBD cases and ExAC. Genomic control was used to correct p-values.

Pathway enrichment analysis

To test for pathway enrichment, we used the ClueGO plugin version 2.3.2 for Cytoscape version 3.4.0. We performed right-sided hypergeometric tests for enrichment of level 3 to 8 biological process GO terms (using the Human GO database from January 25, 2017) with Benjamini-Hochberg p-value correction for multiple tests. GO Term Fusion was used to reduce pathway redundancy. For common and rare variants, the top 200 most significant genes were used to interrogate pathway enrichment in our sample.

RESULTS

Common variants (MAF>0.05)

Though no sites reached genome-wide significance after genomic control (p<2E-06, Figure 1 and Table 2), 14 out of the top 20 significant sites with MAF>0.05 in our logistic regression were near known CD- or IBD-associated loci. Nine variants are around the locus containing *CARD9*, a gene associated with both CD and UC, and three variants were near the locus containing CD-associated *NOD2*. Two protective variants also appeared at other CD loci in *ADAM30* and *NOTCH2*. Genes annotated to the top 20 sites that also appeared in our list of genes involved in neutrophil function included *NOD2*, *CARD9*, and *SNAPC4*.

Pathway enrichment: Many of the pathways implicated by the top 200 most significant annotated genes were immune-related (Table 3 and Figure 2). The largest network of significant GO terms included regulation of cell killing, natural killer cell mediated cytotoxicity, leukocyte mediated immunity, leukocyte apoptotic process, lymphocyte proliferation, and production of interferon-gamma and tumor necrosis factor. Other pathways with the same theme of cell killing included positive regulation of apoptotic cell clearance, regulation of complement activation, and cysteine-type endopeptidase activity involved in apoptotic process. Development of muscle cells and neural crest cells, along with Ras signaling, were also implicated.

Rare variants (MAF<0.05)

SKAT-O analysis of dbGaP analysis rare variants: The only genome-wide significant gene (p<2E-05) was the well-known *NOD2* (Table 4A). When we tested enrichment of variants in loci associated with IBD, the only significant list was the Crohn's-disease associated loci (p=0.002, Table 4B). We also found a suggestive relationship between case status and rare variants in neutrophil genes (p=0.08, Table 4C).

ExAC rare variant analysis: Using the carefully QC-ed list of coordinates from our dbGaP filtering and a minor allele frequency cutoff of less than 0.05, three sites were genome-wide significant (p<6E-07) including one annotated to *NOD2*. Two other of the top 20 most significant variants were annotated to known IBD loci: one other in *NOD2* and one in *D2HGDH*. Of our list of neutrophil genes, in addition to *NOD2* we found two variants in *PCDHA1* among the top 20 most significant rare variants.

Pathway enrichment: The top 200 most significant genes in our list of rare variants were enriched in a few pathways (Table 7 and Figure 3). Immune-related hits included negative regulation of the JAK-STAT cascade and modulation by host of viral transcription. Genes involved in ion transmembrane transport and negative regulation of axon extension were also significant.

DISCUSSION

Our findings did echo important aspects of previous genetic and pathway enrichment analyses. Crohn's-disease-associated loci had a strong showing in our results; two variants in *NOD2* were the most significant in our dbGaP common variant analysis, and 1 site was significant in our ExAC rare variant analysis. *NOD2* also emerged as significant in our gene-level SKAT-O analysis, and CD-associated genes as a group were also significant. This is not unexpected since the majority of our cohort was Crohn's patients. Of the top 20 most significant common variants, 9 were within a single 100kb region around *CARD9* (Supplemental Figure 3), a gene that has long been associated with IBD. This entire region looks equally associated with disease (OR ~1.5) in our cohort, reflecting that deep sequencing still can't solve problems regarding fine mapping of causative variants without sufficient recombination.

We also found intriguing variants in genes not yet associated with IBD. *KRTDAP*, one of our top common variant findings, is involved in keratinocyte differentiation. Keratinocytes are the most abundant component of the epidermis, at the interface between the body and environment. Capable of producing cytokines, these cells play an important role in immunomodulation, and overactivation or defects in that role could contribute to systemic inflammation.

LAMA5, another top hit in our common variant analysis, encodes a subunit of laminin. Laminins are extracellular matrix proteins which are a major component of the basement membrane, a matrix of tissue that separates the epithelium, mesothelium, and endothelium from underlying connective tissue. Because of the important role of laminins in the integrity of this layer, there could be a role for *LAMA5* in IBD pathogenesis. One study of transgenic mice overexpressing the *LAMA5* mouse homolog found an attenuated response to DSS-induced inflammation[30]. The two most significant genes in our SKAT-O rare variant analysis after *NOD2*, *VWA2* and *HAPLN3*, are also components of the extracellular matrix. The location and functions of the products of these genes are also linked to integrins, which have recently emerged in large IBD GWAS[31]. Further studies should be conducted to investigate the possibly interconnected roles of these extracellular matrix proteins in disease etiology.

We were additionally interested in testing enrichment of rare variants in neutrophil genes because children with inherited disorders of phagocyte function exhibit chronic intestinal inflammation similar to CD during the first decade of life[32,33]. Similarly, loss of function in neutrophil antimicrobial pathways could be one mechanism of pediatric CD pathogenesis. Though we did not find a significant association, we did find a suggestive relationship in SKAT-O between rare variants in genes involved in neutrophil function and case status (p=0.08). Positive regulation of leukocyte-mediated immunity was also one of the most significant pathways in our common variant analysis, supporting further study into the role of neutrophils in IBD.

Another important component of the immune system from our pathway analysis was complement; mutations in C2, C3, and CFB were among the top 200 most significant common variants associated with disease in our cohort. Though research into the role of complement has been somewhat lacking, evidence is growing for its potential relevance in disease pathophysiology (reviewed in [34]). A closely related theme, apoptosis, also appeared in several other significant pathways.

Ras signaling also emerged as a pathway of interest in our common variant analysis, and *SOS1*, one of the top hits in our rare variant SKAT-O analysis, is also a guanine nucleotide exchange factor for RAS proteins. In fact, this pathway was previously implicated by a large study drawing from over 30,000 cases and 50,000 controls in contributing to IBD etiology as part of growth factor signaling[35]. Because growth factor deficiencies have been found in patients with IBD, there has been substantial interest in their use as a potential therapeutic agent (reviewed in [36]). Other current targets of therapy that emerged in our analysis include interferon-gamma, a pro-inflammatory cytokine involved in intestinal homeostasis and linked to regulation of IL-23[37], another cytokine associated not only with IBD but other inflammatory diseases. In our rare variant analysis, we found negative regulation of the JAK-STAT cascade, another important inflammatory pathway targeted by recent therapies[38] which underscores the importance of neutrophil involvement in disease.

The primary limitation of this study is the lack of in-house controls for comparison to our cases. However, we performed stringent QC of our data to filter differences between data sets. We used the same processing pipeline for dbGaP as we used for our case data, and filtered to an ancestrally similar population. However, systematic calling differences between our pipeline and ExAC, such as calling or filtering of indels, could still be leading to inflation of p-values and odds ratios in our rare variant analysis.

While large genome-wide association studies have been performed in IBD, our study is the first to specifically investigate the contribution of rare, likely-damaging variants in pediatric-onset disease. Our findings provide further targets for exploring disease etiology—both at the gene and pathway level. Better understanding of the genetic architecture of IBD can hopefully improve disease prediction treatment.

ACKNOWLEDGEMENTS

We would like to thank our study participants as well as the providers and facilitators of publicly available data used in this project: the Exome Aggregation Consortium; the NCBI database of genotypes and phenotypes; the Epi4K Gene Discovery in Epilepsy study (NINDS U01-NS077303) and the Epilepsy Genome/Phenome Project (EPGP - NINDS U01-NS053998); and the ARRA Autism Sequencing Collaborative (an ARRA funded research initiative—R01-MH089208, R01-MH089175, R01-MH089025, R01-MH089004, and R01-MH089482).

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. Kappelman MD, Moore KR, Allen JK, et al. Recent trends in the prevalence of Crohn's disease and ulcerative colitis in a commercially insured US population. Dig Dis Sci 2013;58:519–525.

3. Sandberg KC, Davis MM, Gebremariam A, et al. Increasing hospitalizations in inflammatory bowel disease among children in the United States, 1988-2011. Inflamm Bowel Dis 2014;20:1754–1760.

4. Benchimol EI, Fortinsky KJ, Gozdyra P, et al. Epidemiology of pediatric inflammatory bowel disease: a systematic review of international trends. Inflamm Bowel Dis 2011;17:423–439.

5. Dotson JL, Hyams JS, Markowitz J, et al. Extraintestinal manifestations of pediatric inflammatory bowel disease and their relation to disease type and severity. J Pediatr Gastroenterol Nutr 2010;51:140–145.

6. Pfefferkorn M, Burke G, Griffiths A, et al. Growth abnormalities persist in newly diagnosed children with crohn disease despite current treatment paradigms. J Pediatr Gastroenterol Nutr 2009;48:168–174.

7. Markowitz J, Grancher K, Rosa J, et al. Growth failure in pediatric inflammatory bowel disease. J Pediatr Gastroenterol Nutr 1993;16:373–380.

8. Andreu M, Márquez L, Domènech E, et al. Disease severity in familial cases of IBD. J Crohns Colitis 2014;8:234–239.

9. Carbonnel F, Macaigne G, Beaugerie L, et al. Crohn's disease severity in familial and sporadic cases. Gut 1999;44:91–95.

10. Henriksen M, Jahnsen J, Lygren I, et al. Are there any differences in phenotype or disease course between familial and sporadic cases of inflammatory bowel disease? Results of a population-based follow-up study. Am J Gastroenterol 2007;102:1955–1963.

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

12. Liu JZ, Sommeren S van, Huang H, et al. Association analyses identify 38 susceptibility loci for inflammatory bowel disease and highlight shared genetic risk across populations. Nat Genet 2015;47:979–986.

13. Cutler DJ, Zwick ME, Okou DT, et al. Dissecting Allele Architecture of Early Onset IBD Using High-Density Genotyping. PloS One 2015;10:e0128074.

14. Keinan A, Clark AG. Recent explosive human population growth has resulted in an excess of rare genetic variants. Science 2012;336:740–743.

15. Tennessen JA, Bigham AW, O'Connor TD, et al. Evolution and functional impact of rare coding variation from deep sequencing of human exomes. Science 2012;337:64–69.

16. Marth GT, Yu F, Indap AR, et al. The functional spectrum of low-frequency coding variation. Genome Biol 2011;12:R84.

17. Lek M, Karczewski KJ, Minikel EV, et al. Analysis of protein-coding genetic variation in 60,706 humans. Nature 2016;536:285–291.

18. Mailman MD, Feolo M, Jin Y, et al. The NCBI dbGaP database of genotypes and phenotypes. Nat Genet 2007;39:1181–1186.

19. Anon. Exome Aggregation Consortium (ExAC). Camb MA 2015. Available at: http://exac.broadinstitute.org.

20. Anon. *NCBI SRA Toolkit*. Available at: https://trace.ncbi.nlm.nih.gov/Traces/sra/sra.cgi?cmd=show&f=software&m=software&s=software.

21. Johnston HR, Chopra P, Wingo T, et al. PEMapper / PECaller: A simplified approach to whole-genome sequencing. bioRxiv 2016:076968.

22. Shetty AC, Athri P, Mondal K, et al. SeqAnt: a web service to rapidly identify and annotate DNA sequence variations. BMC Bioinformatics 2010;11:471.

23. Chang CC, Chow CC, Tellier LC, et al. Second-generation PLINK: rising to the challenge of larger and richer datasets. GigaScience 2015;4:7.

24. Purcell S, Chang C. *PLINK 1.9*. Available at: https://www.cog-genomics.org/plink2.

25. Purcell S, Neale B, Todd-Brown K, et al. PLINK: a tool set for whole-genome association and population-based linkage analyses. Am J Hum Genet 2007;81:559–575.

26. Price AL, Patterson NJ, Plenge RM, et al. Principal components analysis corrects for stratification in genome-wide association studies. Nat Genet 2006;38:904–909.

27. Anderson CA, Pettersson FH, Clarke GM, et al. Data quality control in genetic case-control association studies. Nat Protoc 2010;5:1564–1573.

28. Lee S, Emond MJ, Bamshad MJ, 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 2012;91:224–237.

29. R Core team. R: A language and environment for statistical computing. R Found Stat Comput Vienna Austria 2015. Available at: http://www.R-project.org/.

30. Spenlé C, Lefebvre O, Lacroute J, et al. The laminin response in inflammatory bowel disease: protection or malignancy? PloS One 2014;9:e111336.

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

32. Rieber N, Hector A, Kuijpers T, et al. Current concepts of hyperinflammation in chronic granulomatous disease. Clin Dev Immunol 2012;2012:252460.

33. Yu JE, De Ravin SS, Uzel G, et al. High levels of Crohn's disease-associated anti-microbial antibodies are present and independent of colitis in chronic granulomatous disease. Clin Immunol Orlando Fla 2011;138:14–22.

34. Jain U, Otley AR, Van Limbergen J, et al. The complement system in inflammatory bowel disease. Inflamm Bowel Dis 2014;20:1628–1637.

35. Li J, Wei Z, Chang X, et al. Pathway-based Genome-wide Association Studies Reveal the Association Between Growth Factor Activity and Inflammatory Bowel Disease. Inflamm Bowel Dis 2016;22:1540–1551.

36. Krishnan K, Arnone B, Buchman A. Intestinal growth factors: potential use in the treatment of inflammatory bowel disease and their role in mucosal healing. Inflamm Bowel Dis 2011;17:410–422.

37. Sheikh SZ, Matsuoka K, Kobayashi T, et al. Cutting edge: IFN-gamma is a negative regulator of IL-23 in murine macrophages and experimental colitis. J Immunol Baltim Md 1950 2010;184:4069–4073.

38. Pedersen J, Coskun M, Soendergaard C, et al. Inflammatory pathways of importance for management of inflammatory bowel disease. World J Gastroenterol WJG 2014;20:64–77.

TABLES

Table 1A. Basic characteristics of all IBD samples with exome sequencing data used in analysis.

Age at diagnosis	Range	0-17
	Median	8
	Mean	7.5
Gender	Female	215 (42%)
	Male	302 (58%)
Diagnosis	CD	395 (76%)
	UC	89 (17%)
	IBD-other	33 (6%)
Self-identified race	African-American	83 (16%)
	Caucasian	360 (70%)
	Other	30 (6%)
	Not recorded	44 (9%)

Table 1B. Basic characteristics of samples with exome sequencing data used in analysis. Hyphen indicates not applicable.

		IBD cases of European ancestry	**ARRA controls of European ancestry**	**Epi4k controls of European ancestry**
Age at participation	Range	0-17	18-84	Ages not provided, but controls were parents of children with epilepsy
	Median	8	51	
	Mean	7.3	52	
Gender	Female	152 (41%)	118 (56%)	223 (53%)
	Male	216 (59%)	91 (44%)	199 (47%)
Diagnosis	CD	281 (76%)	-	-
	UC	61 (17%)	-	-
	IBD-other	26 (7%)	-	-

Table 2. Top 20 most significant loci found in our common variant logistic regression. Hyphens indicate not applicable or no.

Chrom	Position	ID	Alt	OR	Gene	p-value	Assoc. Diagnosis, Study	Neut gene list
chr16	50711288	rs2066843	T	1.6	NOD2	1E-05	CD, Jostins	Yes
chr16	50710713	rs2066842	T	1.6	NOD2	1E-05	CD, Jostins	Yes
chr9	136371953	rs10781499	A	1.5	CARD9	3E-05	IBD, Jostins	Yes
chr19	35488794	rs10410228	T	1.7	KRTDAP	3E-05	-	-
chr20	62346665	rs6143036	A	1.6	LAMA5	4E-05	-	-
chr1	119895261	rs2641348	G	0.5	ADAM30	4E-05	CD, Jostins	-
chr1	119915381	rs6685892	T	0.5	NOTCH2	5E-05	CD, Jostins	-
chr9	136372044	rs4077515	T	1.5	CARD9	6E-05	IBD, Jostins	Yes
chr16	50675812	rs6596	A	1.6	SNX20	6E-05	CD, Jostins	-
chr9	136395373	rs4266763	G	1.5	SNAPC4	6E-05	IBD, Jostins	Yes
chr9	136380752	rs3812570	C	1.5	SNAPC4	8E-05	IBD, Jostins	Yes
chr9	136380842	rs3812571	C	1.5	SNAPC4	8E-05	IBD, Jostins	Yes
chr9	136384721	rs10781510	A	1.5	SNAPC4	1E-04	IBD, Jostins	Yes
chr9	136404141	rs1051957	G	1.5	SDCCAG3	2E-04	IBD, Jostins	-
chr9	136477334	rs6560632	C	1.4	SEC16A	3E-04	IBD, Jostins	-
chr9	136432987	rs10781542	G	1.4	INPP5E	3E-04	IBD, Jostins	-
chr21	46246830	rs17183220	T	0.44	MCM3AP-AS1	4E-04	-	-
chr13	24799377	rs12865323	C	1.6	RNF17	5E-04	-	-
chr5	78885600	rs1071598	T	1.6	ARSB	6E-04	-	-
chr8	143867013	rs7839934	C	1.4	EPPK1	6E-04	-	-

Table 3. Significantly enriched pathways in the top 200 most significant genes in our common variant (dbGaP) analysis.

GO ID	GO Term	% pathway	p-value	Genes Found
GO:0001578	microtubule bundle formation	6	0.006	[CCDC40, DNAH5, MAP1B, RP1L1, SPAG16]
GO:0002703	regulation of leukocyte mediated immunity	4.3	0.007	[C3, HLA-A, IL2, LILRB1, NOD2, RASGRP1, SERPINB4]
GO:0002705	positive regulation of leukocyte mediated immunity	5.6	0.006	[C3, HLA-A, IL2, NOD2, RASGRP1]
GO:0010927	cellular component assembly involved in morphogenesis	5.2	0.007	[ANK2, DAG1, FHOD3, IGSF22, MYPN]
GO:0014032	neural crest cell development	6.7	0.007	[ERBB4, JAG1, LAMA5, RET]
GO:0014902	myotube differentiation	4.3	0.01	[GPX1, NOS1, RYR1, TANC1, XK]
GO:0030449	regulation of complement activation	7.9	0.01	[C2, C3, CFB]
GO:0031341	regulation of cell killing	8.1	0.003	[HLA-A, IL11, LILRB1, RASGRP1, SERPINB4]
GO:0032649	regulation of interferon-gamma production	4.2	0.02	[HLA-A, IL2, LILRB1, RASGRP1]
GO:0032760	positive regulation of tumor necrosis factor production	5.1	0.02	[CARD9, NOD2, RASGRP1]
GO:0042269	regulation of natural killer cell mediated cytotoxicity	12.1	0.003	[HLA-A, LILRB1, RASGRP1, SERPINB4]
GO:0043154	negative regulation of cysteine-type endopeptidase activity involved in apoptotic process	4.3	0.02	[ARRB1, GPI, GPX1, RPS6KA1]
GO:0045214	sarcomere organization	6.8	0.02	[FHOD3, IGSF22, MYPN]
GO:0046579	positive regulation of Ras protein signal transduction	5.7	0.02	[ARRB1, NOTCH2, RASGRP1]
GO:0048747	muscle fiber development	6.8	0.008	[GPX1, MYPN, RYR1, XK]
GO:0050672	negative regulation of lymphocyte proliferation	4.4	0.03	[IL2, KIAA0922, LILRB1]
GO:0055001	muscle cell development	4.2	0.006	[ANK2, FHOD3, GPX1, IGSF22, MYPN, RYR1, XK]
GO:2000106	regulation of leukocyte apoptotic process	4.5	0.02	[IL2, LILRB1, NOD2, TP53BP1]
GO:2000427	positive regulation of apoptotic cell clearance	33.3	0.002	[C2, C3, CCL2]

Table 4A. Top 15 results from SKAT-O analysis of enrichment of rare, conserved (CADD>10) variants in all genes.

SetID	p-value	N Variants
NOD2	8.4E-12	15
VWA2	0.0006	7
HAPLN3	0.0008	5
LMF1	0.002	5
SOS1	0.002	5
MAGI2	0.002	7
SRRM2	0.002	13
RGS12	0.003	10
SCAF4	0.003	5
STARD13	0.004	8
RHPN2	0.005	6
D2HGDH	0.005	6
G6PC2	0.005	6
NR4A1	0.005	5
EFEMP2	0.006	5

Table 4B. SKAT-O analysis for enrichment of rare variants with CADD scores>10 in loci associated with Crohn's disease (CD), inflammatory bowel disease (IBD), or ulcerative colitis (UC).

SetID	p-value	N Variants
CD	0.003	497
IBD	0.9	1782
UC	0.5	428

Table 4C. SKAT-O analysis for enrichment of rare, conserved variants in neutrophil genes (NEUT).

SetID	p-value	N Variants
NEUT	0.08	3334

Table 5. Top 20 most significant sites in our rare variant Fisher's exact tests. Hyphens indicate not applicable or no.

Chrom	Pos	ID	Alt	Type	OR	Gene	p-value	Assoc Diagnosis, Study	Neut gene list
chr11	294540	chr11_294540	GC	INS	123	ATHL1	6E-10	-	-
chr16	50729867	rs796661546	GC	INS	4.4	NOD2	6E-10	CD, Jostins	Yes
chr8	100712766	chr8_100712766	CA	INS	34	PABPC1	6E-10	-	-
chr9	101390469	chr9_101390469	GTA	INS	173	MRPL50	1E-06	-	-
chr21	44573789	rs9977039	G	SNP	5.8	TSPEAR	7E-06	-	-
chr10	29462394	chr10_29462394	AT	INS	Inf	SVIL-AS1	1E-05	-	-
chr16	50722629	rs2066845	C	MULTIALL.	3.4	NOD2	3E-05	CD, Jostins	Yes
chr4	56964497	rs17087307	C	SNP	0.34	NOA1	4E-05	-	-
chr7	72713798	rs146095374	A	SNP	0.26	TYW1B	5E-05	-	-
chr5	140822334	rs61730632	A	SNP	2.8	PCDHA1	6E-05	-	Yes
chr14	73953419	rs778985097	AT	INS	10	COQ6	8E-05	-	-
chr5	140875534	rs114654172	G	SNP	2.7	PCDHA1	1E-04	-	Yes
chr11	5544676	rs7934354	G	SNP	0.18	OR52H1	1E-04	-	-
chr6	31960262	rs11541400	G	SNP	5.2	SKIV2L	2E-04	-	-
chr6	31728544	rs139006870	A	SNP	5.2	DDAH2	2E-04	-	-
chr15	49588022	chr15_49588022	CT	INS	Inf	FAM227B	2E-04	-	-
chr3	51995472	rs371570896	A	SNP	77	RPL29	3E-04	-	-
chr2	241767780	rs143940595	A	SNP	0	D2HGDH	3E-04	CD, Liu	-
chr2	20034361	rs145912850	A	SNP	0.06	LAPTM4A	4E-04	-	-
chr3	114079955	rs772016664	G	SNP	348	QTRTD1	4E-04	-	-

Table 6. Significantly enriched pathways using the list of the top 200 most significant genes in our ExAC rare variant analysis.

GO ID	GO Term	% pathway	p-value	Genes Found
GO:0030517	negative regulation of axon extension	12	0.004	[BCL11A, RTN4R, SEMA5A]
GO:0043921	modulation by host of viral transcription	12	0.004	[HMGA2, POU2F3, PSG1]
GO:0046426	negative regulation of JAK-STAT cascade	5.8	0.02	[HMGA2, RTN4R, RTN4RL2]
GO:0098661	inorganic anion transmembrane transport	4.8	0.008	[ABCB11, ANKH, CLCN6, CLCNKB, SLC12A6, SLC26A2]
GO:1902476	chloride transmembrane transport	4.3	0.02	[CLCN6, CLCNKB, SLC12A6, SLC26A2]

FIGURES

Figure 1. Q-Q plot of p-values from logistic regression (with significant principal components and sex as covariates) comparing frequency of exome sequencing common variants in pediatric IBD cases to controls from dbGaP.

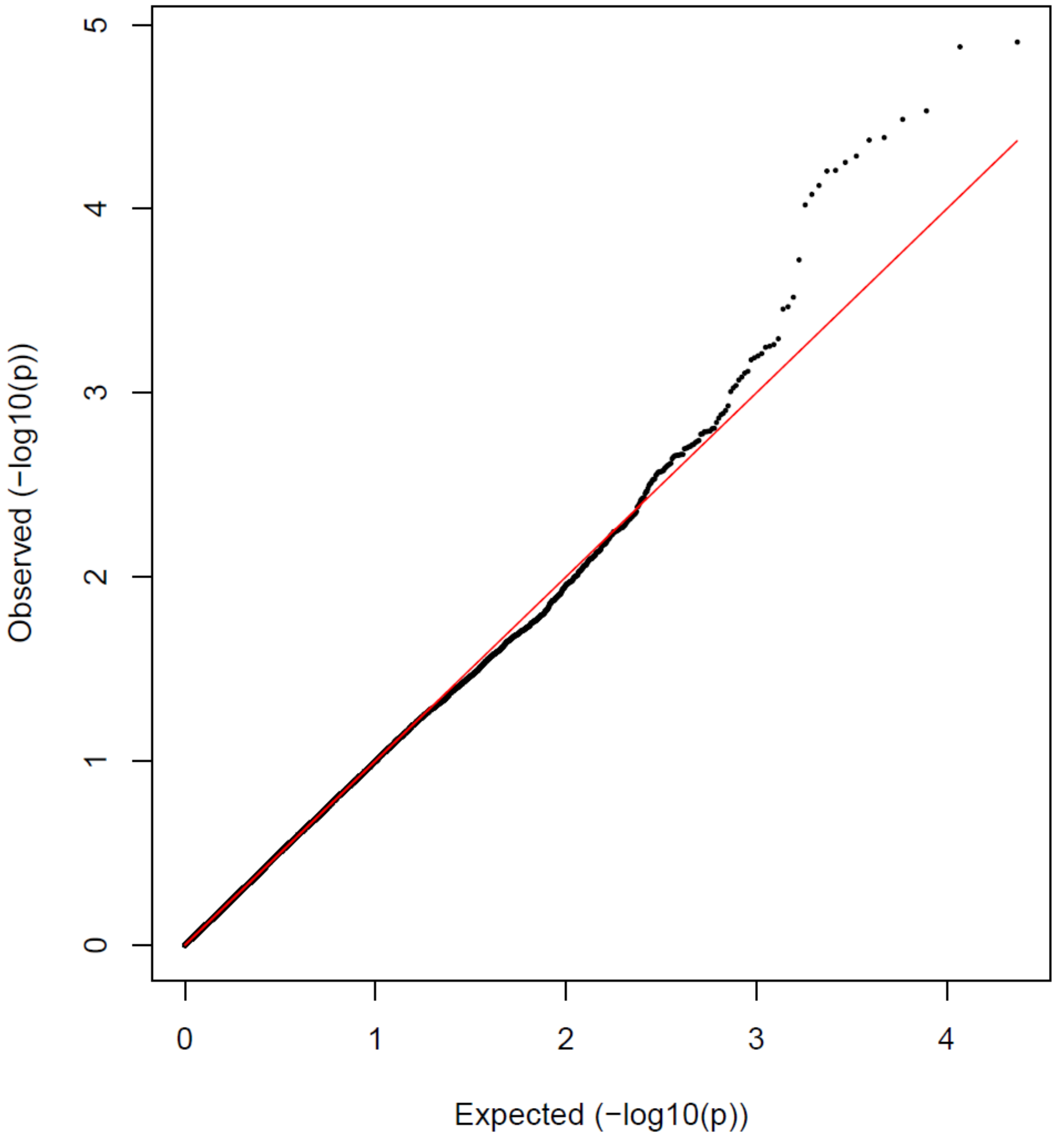


Figure 2. Pathway enrichment of the genes annotated to the top 200 most significant common variants tested in our logistic regression.

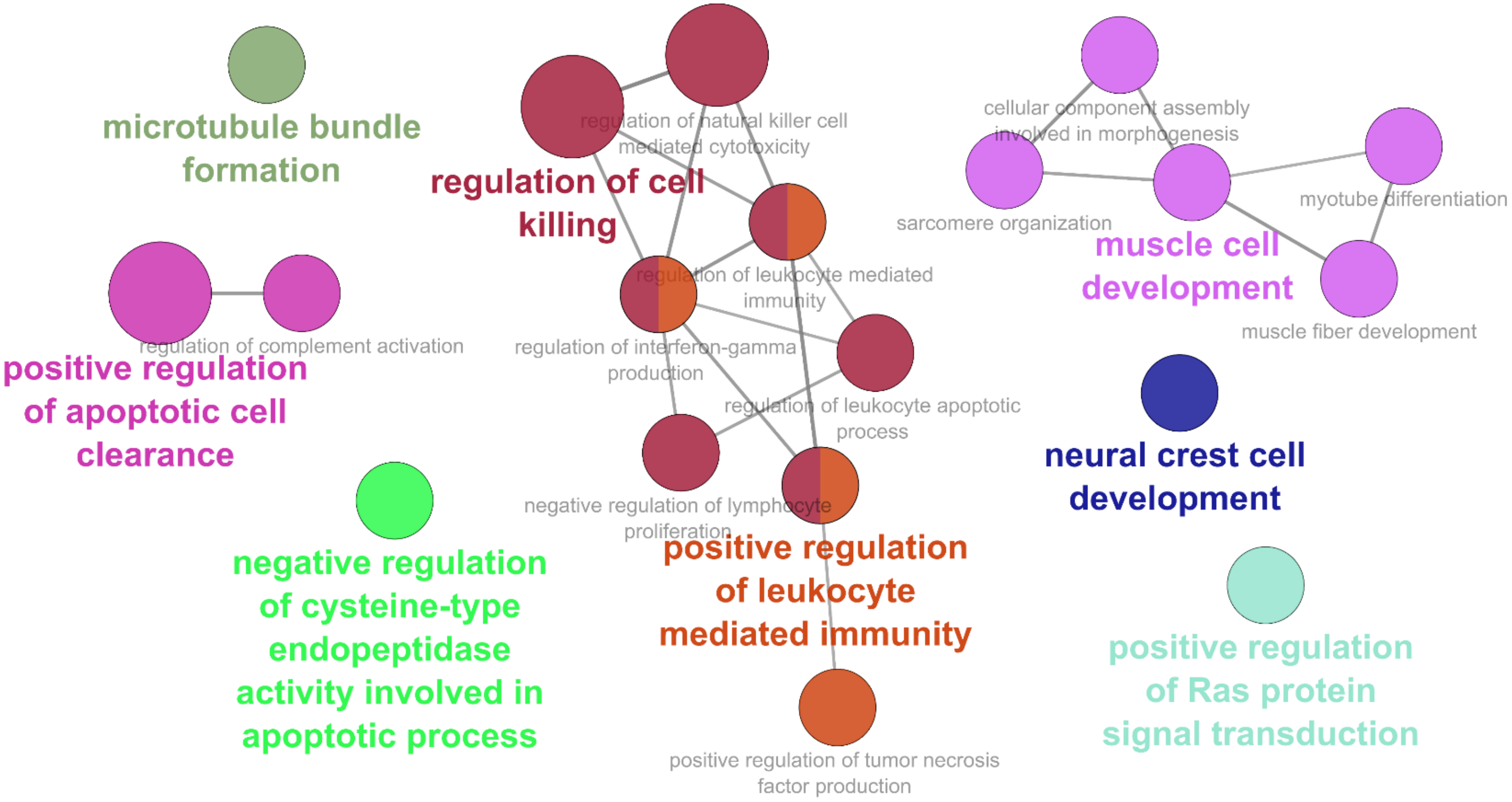
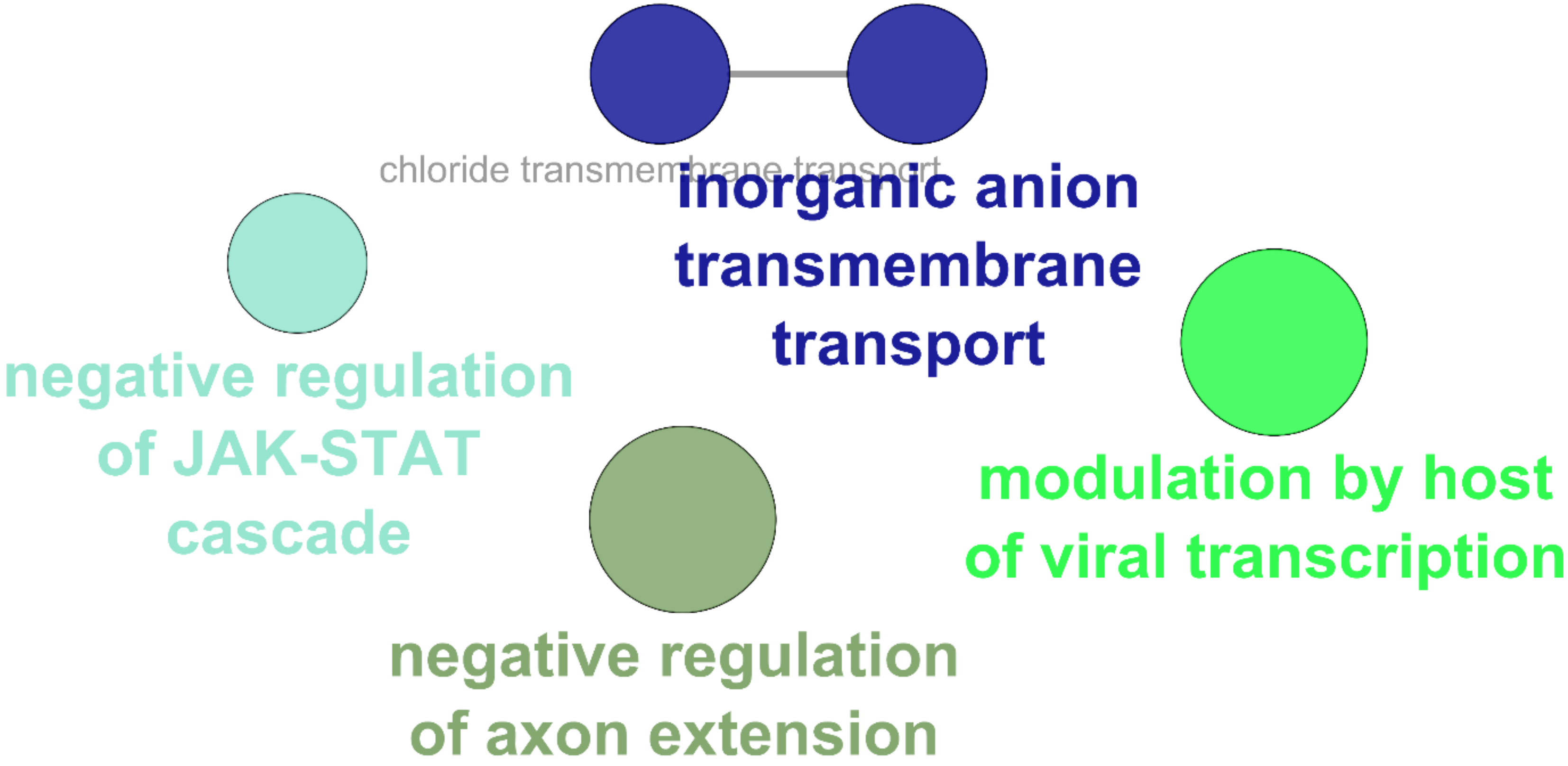

microtubule bundle formation
regulation of cell killing
regulation of natural killer cell mediated cytotoxicity
cellular component assembly involved in morphogenesis
sarcomere organization
myotube differentiation
muscle cell development
muscle fiber development
positive regulation of apoptotic cell clearance
regulation of complement activation
regulation of leukocyte mediated immunity
regulation of interferon-gamma production
regulation of leukocyte apoptotic process
neural crest cell development
negative regulation of lymphocyte proliferation
positive regulation of leukocyte mediated immunity
negative regulation of cysteine-type endopeptidase activity involved in apoptotic process
positive regulation of Ras protein signal transduction
positive regulation of tumor necrosis factor production

Figure 3. Pathway enrichment of the genes annotated to the top 200 most significant rare variants tested in our rare variant analysis.


chloride transmembrane transport
inorganic anion transmembrane transport
negative regulation of JAK-STAT cascade
modulation by host of viral transcription
negative regulation of axon extension

SUPPLEMENTAL TABLES

Supplemental Table 1. List of IBD-associated loci used in analysis.

Study	Type	hg19_chr	hg19_pos	hg38_chr	hg38_pos	minus250kb	plus250kb
Jostins_2012	CD	1	78620000	chr1	78154316	77904316	78404316
Jostins_2012	CD	1	114300000	chr1	113757378	113507378	114007378
Jostins_2012	CD	1	120450000	chr1	119907377	119657377	120157377
Jostins_2012	CD	1	172850000	chr1	172880860	172630860	173130860
Jostins_2012	CD	2	27630000	chr2	27407133	27157133	27657133
Jostins_2012	CD	2	62550000	chr2	62322865	62072865	62572865
Jostins_2012	CD	2	231090000	chr2	230225285	229975285	230475285
Jostins_2012	CD	2	234145000	chr2	233236354	232986354	233486354
Jostins_2012	CD	4	48360000	chr4	48357983	48107983	48607983
Jostins_2012	CD	4	102860000	chr4	101938843	101688843	102188843
Jostins_2012	CD	5	55430000	chr5	56134173	55884173	56384173
Jostins_2012	CD	5	72540000	chr5	73244173	72994173	73494173
Jostins_2012	CD	5	173340000	chr5	173912997	173662997	174162997
Jostins_2012	CD	6	21420000	chr6	21419769	21169769	21669769
Jostins_2012	CD	6	31270000	chr6	31302223	31052223	31552223
Jostins_2012	CD	6	127450000	chr6	127128855	126878855	127378855
Jostins_2012	CD	6	128240000	chr6	127918855	127668855	128168855
Jostins_2012	CD	6	159490000	chr6	159068968	158818968	159318968
Jostins_2012	CD	7	26880000	chr7	26840381	26590381	27090381
Jostins_2012	CD	7	28170000	chr7	28130381	27880381	28380381
Jostins_2012	CD	8	90870000	chr8	89857772	89607772	90107772
Jostins_2012	CD	8	129560000	chr8	128547754	128297754	128797754
Jostins_2012	CD	13	44450000	chr13	43875864	43625864	44125864
Jostins_2012	CD	15	38890000	chr15	38597799	38347799	38847799
Jostins_2012	CD	16	50660000	chr16	50626089	50376089	50876089
Jostins_2012	CD	17	25840000	chr17	27512974	27262974	27762974
Jostins_2012	CD	19	1120000	chr19	1120001	870001	1370001
Jostins_2012	CD	19	46850000	chr19	46346743	46096743	46596743
Jostins_2012	CD	19	49200000	chr19	48696743	48446743	48946743
Jostins_2012	CD	21	34770000	chr21	33397694	33147694	33647694
Jostins_2012	UC	1	2500000	chr1	2568561	2318561	2818561
Jostins_2012	UC	1	20150000	chr1	19823507	19573507	20073507
Jostins_2012	UC	1	200090000	chr1	200120872	199870872	200370872
Jostins_2012	UC	2	198650000	chr2	197785276	197535276	198035276
Jostins_2012	UC	2	199700000	chr2	198835276	198585276	199085276
Jostins_2012	UC	3	53050000	chr3	53015984	52765984	53265984
Jostins_2012	UC	4	103510000	chr4	102588843	102338843	102838843
Jostins_2012	UC	5	590000	chr5	589885	339885	839885
Jostins_2012	UC	5	134440000	chr5	135104310	134854310	135354310
Jostins_2012	UC	6	32595000	chr6	32627223	32377223	32877223

Jostins_2012	UC	7	2780000	chr7	2740366	2490366	2990366
Jostins_2012	UC	7	27220000	chr7	27180381	26930381	27430381
Jostins_2012	UC	7	107450000	chr7	107809555	107559555	108059555
Jostins_2012	UC	7	128570000	chr7	128929946	128679946	129179946
Jostins_2012	UC	11	96020000	chr11	96286836	96036836	96536836
Jostins_2012	UC	11	114380000	chr11	114509278	114259278	114759278
Jostins_2012	UC	15	41550000	chr15	41257802	41007802	41507802
Jostins_2012	UC	16	30470000	chr16	30458679	30208679	30708679
Jostins_2012	UC	16	68580000	chr16	68546097	68296097	68796097
Jostins_2012	UC	17	70640000	chr17	72643861	72393861	72893861
Jostins_2012	UC	19	47120000	chr19	46616743	46366743	46866743
Jostins_2012	UC	20	33800000	chr20	35212197	34962197	35462197
Jostins_2012	UC	20	43060000	chr20	44431360	44181360	44681360
Jostins_2012	IBD	1	1240000	chr1	1304620	1054620	1554620
Jostins_2012	IBD	1	8020000	chr1	7959940	7709940	8209940
Jostins_2012	IBD	1	22700000	chr1	22373507	22123507	22623507
Jostins_2012	IBD	1	67680000	chr1	67214317	66964317	67464317
Jostins_2012	IBD	1	70990000	chr1	70524317	70274317	70774317
Jostins_2012	IBD	1	151790000	chr1	151817524	151567524	152067524
Jostins_2012	IBD	1	155670000	chr1	155700209	155450209	155950209
Jostins_2012	IBD	1	160850000	chr1	160880210	160630210	161130210
Jostins_2012	IBD	1	161470000	chr1	161500210	161250210	161750210
Jostins_2012	IBD	1	197600000	chr1	197630870	197380870	197880870
Jostins_2012	IBD	1	200870000	chr1	200900872	200650872	201150872
Jostins_2012	IBD	1	206930000	chr1	206756655	206506655	207006655
Jostins_2012	IBD	2	25120000	chr2	24897131	24647131	25147131
Jostins_2012	IBD	2	28610000	chr2	28387133	28137133	28637133
Jostins_2012	IBD	2	43810000	chr2	43582861	43332861	43832861
Jostins_2012	IBD	2	61200000	chr2	60972865	60722865	61222865
Jostins_2012	IBD	2	65670000	chr2	65442866	65192866	65692866
Jostins_2012	IBD	2	102860000	chr2	102243540	101993540	102493540
Jostins_2012	IBD	2	163100000	chr2	162243490	161993490	162493490
Jostins_2012	IBD	2	191920000	chr2	191055274	190805274	191305274
Jostins_2012	IBD	2	219140000	chr2	218275277	218025277	218525277
Jostins_2012	IBD	2	241570000	chr2	240630583	240380583	240880583
Jostins_2012	IBD	3	18760000	chr3	18718508	18468508	18968508
Jostins_2012	IBD	3	48960000	chr3	48922567	48672567	49172567
Jostins_2012	IBD	4	74850000	chr4	73984283	73734283	74234283
Jostins_2012	IBD	4	123220000	chr4	122298845	122048845	122548845
Jostins_2012	IBD	5	10690000	chr5	10689888	10439888	10939888
Jostins_2012	IBD	5	40380000	chr5	40379898	40129898	40629898
Jostins_2012	IBD	5	96240000	chr5	96904296	96654296	97154296
Jostins_2012	IBD	5	130005000	chr5	130669307	130419307	130919307
Jostins_2012	IBD	5	131190000	chr5	131854307	131604307	132104307
Jostins_2012	IBD	5	141510000	chr5	142130435	141880435	142380435
Jostins_2012	IBD	5	150270000	chr5	150890438	150640438	151140438

Jostins_2012	IBD	5	158800000	chr5	159372992	159122992	159622992
Jostins_2012	IBD	5	176790000	chr5	177362999	177112999	177612999
Jostins_2012	IBD	6	14710000	chr6	14709769	14459769	14959769
Jostins_2012	IBD	6	20770000	chr6	20769769	20519769	21019769
Jostins_2012	IBD	6	90960000	chr6	90250281	90000281	90500281
Jostins_2012	IBD	6	106430000	chr6	105982125	105732125	106232125
Jostins_2012	IBD	6	111820000	chr6	111498797	111248797	111748797
Jostins_2012	IBD	6	138000000	chr6	137678863	137428863	137928863
Jostins_2012	IBD	6	143900000	chr6	143578863	143328863	143828863
Jostins_2012	IBD	6	167370000	chr6	166956512	166706512	167206512
Jostins_2012	IBD	7	50245000	chr7	50205404	49955404	50455404
Jostins_2012	IBD	7	98750000	chr7	99152377	98902377	99402377
Jostins_2012	IBD	7	100335000	chr7	100737377	100487377	100987377
Jostins_2012	IBD	7	116890000	chr7	117249946	116999946	117499946
Jostins_2012	IBD	8	126530000	chr8	125517758	125267758	125767758
Jostins_2012	IBD	8	130620000	chr8	129607754	129357754	129857754
Jostins_2012	IBD	9	4980000	chr9	4980000	4730000	5230000
Jostins_2012	IBD	9	93920000	chr9	91157718	90907718	91407718
Jostins_2012	IBD	9	117600000	chr9	114837720	114587720	115087720
Jostins_2012	IBD	9	139320000	chr9	136425548	136175548	136675548
Jostins_2012	IBD	10	6080000	chr10	6038037	5788037	6288037
Jostins_2012	IBD	10	30720000	chr10	30431071	30181071	30681071
Jostins_2012	IBD	10	35295000	chr10	35006072	34756072	35256072
Jostins_2012	IBD	10	59990000	chr10	58230239	57980239	58480239
Jostins_2012	IBD	10	64510000	chr10	62750240	62500240	63000240
Jostins_2012	IBD	10	75670000	chr10	73910242	73660242	74160242
Jostins_2012	IBD	10	81030000	chr10	79270243	79020243	79520243
Jostins_2012	IBD	10	82250000	chr10	80490244	80240244	80740244
Jostins_2012	IBD	10	94430000	chr10	92670243	92420243	92920243
Jostins_2012	IBD	10	101280000	chr10	99520243	99270243	99770243
Jostins_2012	IBD	11	1870000	chr11	1848770	1598770	2098770
Jostins_2012	IBD	11	58330000	chr11	58562527	58312527	58812527
Jostins_2012	IBD	11	60770000	chr11	61002528	60752528	61252528
Jostins_2012	IBD	11	61560000	chr11	61792528	61542528	62042528
Jostins_2012	IBD	11	64120000	chr11	64352528	64102528	64602528
Jostins_2012	IBD	11	65650000	chr11	65882529	65632529	66132529
Jostins_2012	IBD	11	76290000	chr11	76578956	76328956	76828956
Jostins_2012	IBD	11	87120000	chr11	87408958	87158958	87658958
Jostins_2012	IBD	11	118740000	chr11	118869291	118619291	119119291
Jostins_2012	IBD	12	12650000	chr12	12497066	12247066	12747066
Jostins_2012	IBD	12	40770000	chr12	40376198	40126198	40626198
Jostins_2012	IBD	12	48200000	chr12	47806217	47556217	48056217
Jostins_2012	IBD	12	68490000	chr12	68096220	67846220	68346220
Jostins_2012	IBD	13	27520000	chr13	26945863	26695863	27195863
Jostins_2012	IBD	13	40860000	chr13	40285863	40035863	40535863
Jostins_2012	IBD	13	99950000	chr13	99297746	99047746	99547746

Jostins_2012	IBD	14	69270000	chr14	68803283	68553283	69053283
Jostins_2012	IBD	14	75700000	chr14	75233297	74983297	75483297
Jostins_2012	IBD	14	88470000	chr14	88003656	87753656	88253656
Jostins_2012	IBD	15	67430000	chr15	67137662	66887662	67387662
Jostins_2012	IBD	15	91170000	chr15	90626768	90376768	90876768
Jostins_2012	IBD	16	11540000	chr16	11446144	11196144	11696144
Jostins_2012	IBD	16	23860000	chr16	23848679	23598679	24098679
Jostins_2012	IBD	16	28595000	chr16	28583679	28333679	28833679
Jostins_2012	IBD	16	86000000	chr16	85966394	85716394	86216394
Jostins_2012	IBD	17	32590000	chr17	34262981	34012981	34512981
Jostins_2012	IBD	17	37910000	chr17	39753747	39503747	40003747
Jostins_2012	IBD	17	40530000	chr17	42377982	42127982	42627982
Jostins_2012	IBD	17	57960000	chr17	59882639	59632639	60132639
Jostins_2012	IBD	18	12800000	chr18	12800001	12550001	13050001
Jostins_2012	IBD	18	46390000	chr18	48863629	48613629	49113629
Jostins_2012	IBD	18	67530000	chr18	69862764	69612764	70112764
Jostins_2012	IBD	19	10490000	chr19	10379324	10129324	10629324
Jostins_2012	IBD	19	33730000	chr19	33239094	32989094	33489094
Jostins_2012	IBD	19	55380000	chr19	54868545	54618545	55118545
Jostins_2012	IBD	20	30750000	chr20	32162197	31912197	32412197
Jostins_2012	IBD	20	31370000	chr20	32782194	32532194	33032194
Jostins_2012	IBD	20	44740000	chr20	46111361	45861361	46361361
Jostins_2012	IBD	20	48950000	chr20	50333463	50083463	50583463
Jostins_2012	IBD	20	57820000	chr20	59244945	58994945	59494945
Jostins_2012	IBD	20	62340000	chr20	63708648	63458648	63958648
Jostins_2012	IBD	21	16810000	chr21	15437681	15187681	15687681
Jostins_2012	IBD	21	40460000	chr21	39088074	38838074	39338074
Jostins_2012	IBD	21	45620000	chr21	44200117	43950117	44450117
Jostins_2012	IBD	22	21920000	chr22	21565711	21315711	21815711
Jostins_2012	IBD	22	30425000	chr22	30029011	29779011	30279011
Jostins_2012	IBD	22	39690000	chr22	39293995	39043995	39543995
Liu_2015	CD	1	63049593	chr1	62583922	62333922	62833922
Liu_2015	IBD	1	92554283	chr1	92088726	91838726	92338726
Liu_2015	UC	1	101466054	chr1	101000498	100750498	101250498
Liu_2015	IBD	1	169519049	chr1	169549811	169299811	169799811
Liu_2015	CD	1	186875459	chr1	186906327	186656327	187156327
Liu_2015	CD	1	198598663	chr1	198629533	198379533	198879533
Liu_2015	CD	2	145492382	chr2	144734815	144484815	144984815
Liu_2015	IBD	2	160794008	chr2	159937497	159687497	160187497
Liu_2015	UC	2	204592021	chr2	203727298	203477298	203977298
Liu_2015	IBD	2	228660112	chr2	227795396	227545396	228045396
Liu_2015	CD	2	242737341	chr2	241797926	241547926	242047926
Liu_2015	UC	3	46457412	chr3	46415921	46165921	46665921
Liu_2015	UC	3	101569726	chr3	101850882	101600882	102100882
Liu_2015	CD	3	141105570	chr3	141386728	141136728	141636728
Liu_2015	IBD	4	3444503	chr4	3442776	3192776	3692776

Liu_2015	IBD	4	26132361	chr4	26130739	25880739	26380739
Liu_2015	IBD	4	38325036	chr4	38323415	38073415	38573415
Liu_2015	UC	4	106075498	chr4	105154341	104904341	105404341
Liu_2015	IBD	5	38867732	chr5	38867630	38617630	39117630
Liu_2015	IBD	5	71693899	chr5	72398072	72148072	72648072
Liu_2015	IBD	5	172324978	chr5	172897975	172647975	173147975
Liu_2015	CD	6	382559	chr6	382559	132559	632559
Liu_2015	CD	6	3420406	chr6	3420172	3170172	3670172
Liu_2015	CD	6	149577079	chr6	149255943	149005943	149505943
Liu_2015	UC	7	17442679	chr7	17403055	17153055	17653055
Liu_2015	IBD	7	148220448	chr7	148523356	148273356	148773356
Liu_2015	IBD	8	27227554	chr8	27370037	27120037	27620037
Liu_2015	UC	8	49129242	chr8	48216682	47966682	48466682
Liu_2015	IBD	10	104232716	chr10	102472959	102222959	102722959
Liu_2015	CD	12	6491125	chr12	6381959	6131959	6631959
Liu_2015	IBD	12	112007756	chr12	111569952	111319952	111819952
Liu_2015	IBD	12	120146925	chr12	119709120	119459120	119959120
Liu_2015	CD	13	43018030	chr13	42443894	42193894	42693894
Liu_2015	CD	17	54880993	chr17	56803632	56553632	57053632
Liu_2015	UC	17	76737118	chr17	78741036	78491036	78991036
Liu_2015	CD	18	56879827	chr18	59212595	58962595	59462595
Liu_2015	CD	18	77220616	chr18	79460616	79210616	79710616
Liu_2015	CD	22	41867377	chr22	41471373	41221373	41721373

Supplemental Table 2. List of genes involved in neutrophil function.

AATF
ABCB1
ABCD4
ABCG1
ABCG5
ABR
ACE
ACP5
ACTL6A
ADAM10
ADAM17
ADORA3
AGA
AGR2
AGT
AGTR1
AIFM1
AK2
ALOX12
ALOX15
ALOX5
ALOX5AP
AMICA1
AMPD3
ANK1
ANXA1
ANXA3
AP3B1
ARHGAP15
ARHGDIA
ARHGEF1
ARHGEF4
ARHGEF5
ARID4A
ARNTL
ASXL1
ATP7B
ATRX
AZU1
B4GALT1
BACH2
BAG3

BCL2A1
BCL2L11
BCL6
BCL6B
BCR
BLNK
BPI
BRCA2
BTK
C3
C3AR1
C4A
C4B
C5AR1
CAMK1D
CAMP
CAPG
CASP1
CASP10
CASP4
CASP8
CAV1
CCL11
CCL13
CCL3L3
CCR1
CCR2
CCR3
CCR4
CD101
CD14
CD19
CD28
CD300A
CD300LB
CD34
CD3E
CD40
CD40LG
CD44
CD47
CD69

CD79A
CD93
CD97
CDH5
CEBPA
CEBPE
CERK
CFLAR
CHI3L1
CHRNA7
CHUK
CIITA
CISH
CLCA1
CLCN3
CLEC4E
CLEC6A
CLEC7A
CMKLR1
CNN2
COL1A1
CPA3
CR1
CR2
CREBBP
CSF1
CSF1R
CSF2
CSF2RA
CSF2RB
CSF3
CSF3R
CTNNB1
CTSC
CTSE
CTSG
CTSS
CTTN
CXCL12
CXCL6
CXCR2
CXCR4

CYB5R4
CYBA
CYBB
CYP1A1
DBA2
DCC
DDX58
DEFA1
DEFA1B
DEFA3
DEFA4
DIAPH1
DIDO1
DMD
DOCK2
DOK1
DOK2
DOT1L
DSG3
DUOX1
DUOX2
DUSP1
E2F4
EDIL3
EDN1
EGFR
EGR1
EIF2AK3
ELANE
ELMOD1
ENTPD1
EP300
ESR2
ETV6
F2RL1
F3
FAM104A
FANCA
FANCB
FANCC
FANCD2
FANCE

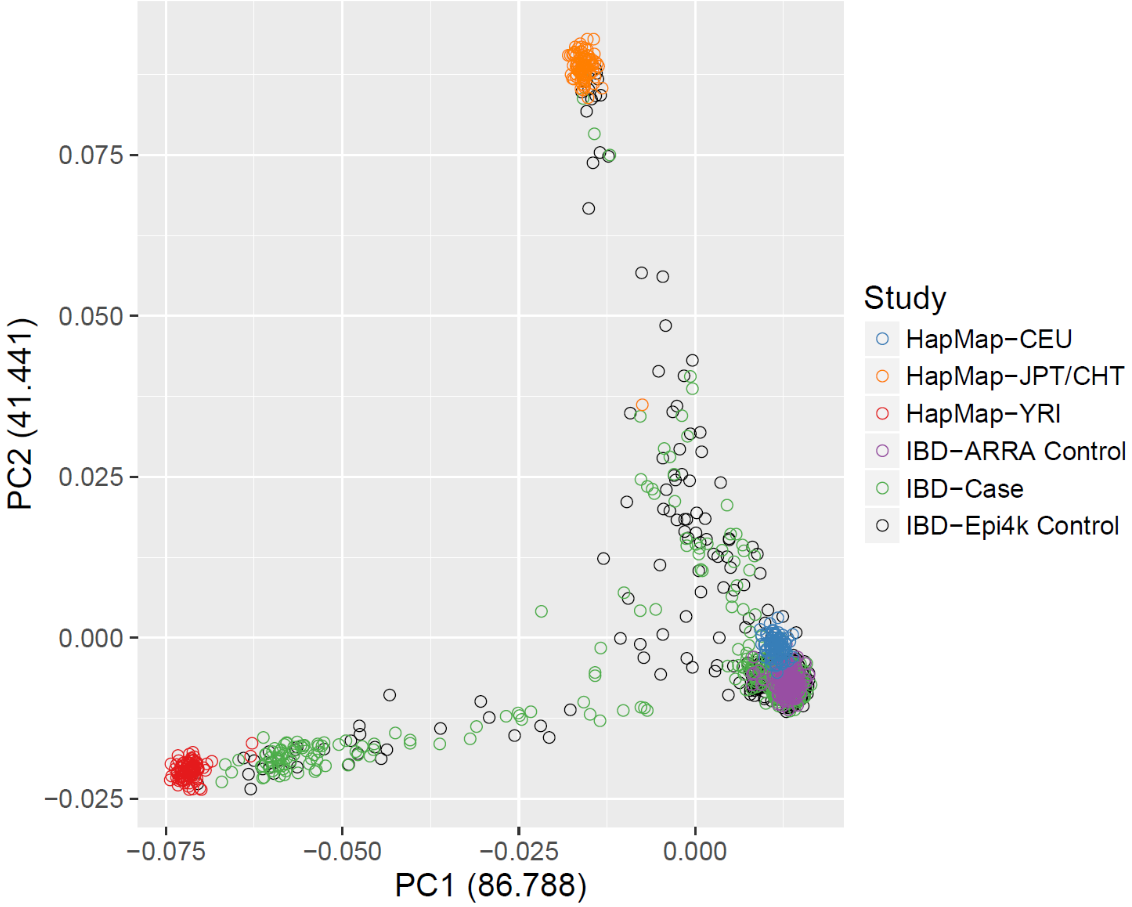
FANCF
FANCG
FAS
FASLG
FASTK
FBN1
FBXL4
FCAR
FCER1A
FCER1G
FCGR1A
FCGR2A
FCGR2B
FCGR3A
FCGR3B
FES
FFAR2
FGG
FGR
FHIT
FLI1
FLOT1
FLT3
FLT3LG
FMO3
FOXN1
FOXP3
FPR1
FPR2
FTCD
FUT4
FUT7
FZD9
G6PC3
G6PT1
GAB2
GADD45A
GALNT1
GATA1
GATA2
GBP5
GCNT1

GCNT4
GFI1
GIMAP1-GIMAP5
GIT2
GJA1
GMNN
GNAI2
GNMT
GPI
GPRC5C
GSE1
GSN
GSS
GSTP1
GSX1
HAX1
HCAR2
HCK
HDC
HLA-A
HLA-B
HLA-G
HMOX1
HPRT1
HSPB1
HVCN1
ICAM1
ICOS
ID1
IER3
IFNB1
IFNG
IGHM
IGLL1
IKBKB
IKZF1
IL10
IL13
IL13RA1
IL17A
IL17RA
IL17RB

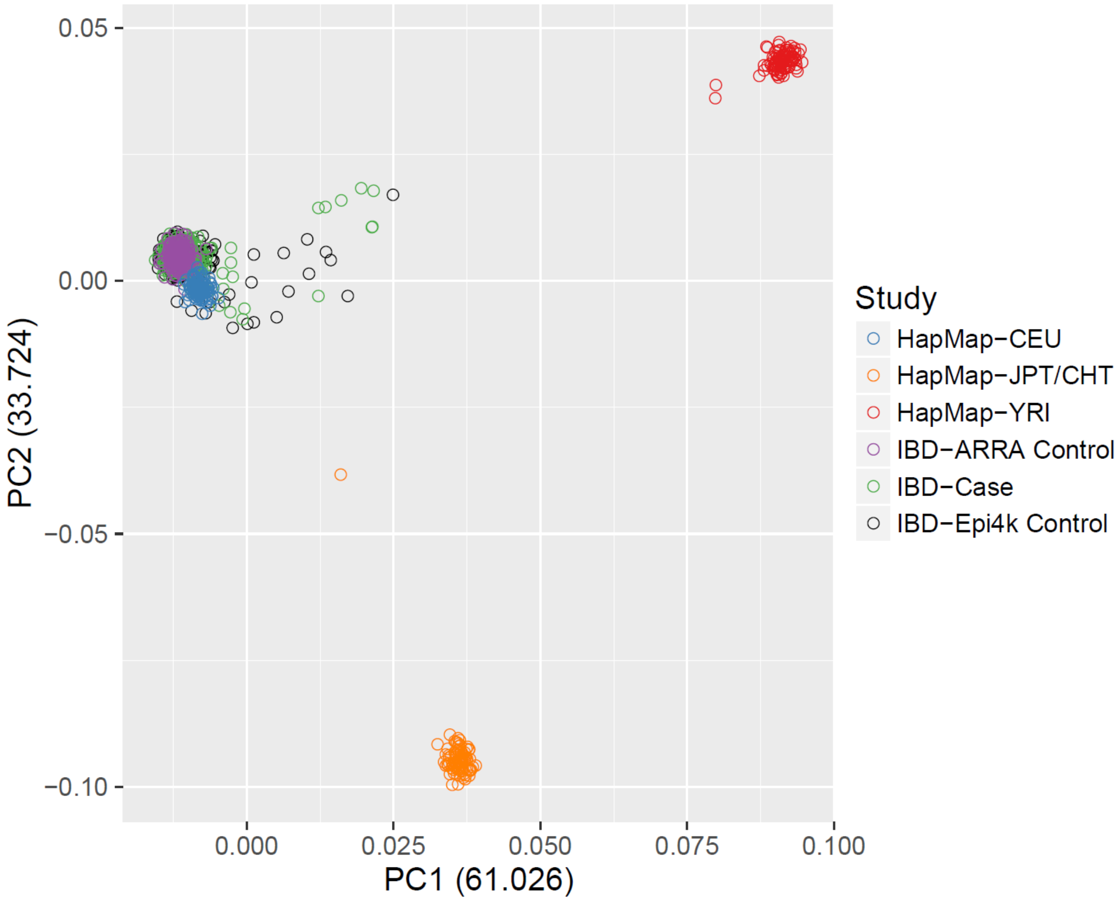
IL18
IL1B
IL1R1
IL1RL1
IL1RL2
IL21R
IL22
IL23A
IL25
IL27RA
IL2RB
IL2RG
IL33
IL36RN
IL4
IL4R
IL5
IL5RA
IL6
IL6R
IL6ST
IL9
INPP5D
INS
IRAK3
IRAK4
IRF8
ITGA1
ITGAL
ITGAM
ITGAX
ITGB1
ITGB2
ITGB7
JAGN1
JAK2
JAK3
JAM3
JDP2
KDM1A
KDM5A
KERA
KISS1R
KIT
KITLG
KMO
KMT2A
KMT2E
LAIR1
LAMTOR2
LAT
LBP
LBR
LCN2
LCP1
LDHA
LDLR
LEPR
LGALS3
LIF
LILRA1
LILRA2
LILRA4
LILRA5
LILRA6
LILRB2
LILRB4
LILRB5
LMBRD1
LMO2
LRRC8A
LSP1
LTB4R
LTB4R2
LTBR
LTF
LUM
LY96
LYN
LYST
LYZ
MAP3K14
MAPK1
MAPK3
MCL1
MDK
MDM2
MECOM
MEFV
MFAP5
MGAT4B
MIF
MINA
MITF
MMAA
MMAB
MMACHC
MMP28
MMP8
MMP9
MPO
MPP1
MRC1
MSI2
MTHFD1
MTOR
MUT
MXD1
MYB
MYBL2
MYD88
MYH9
MYL12A
MYO1F
MYSM1
NAMPT
NCF1
NCF1C
NCF2
NCF4
NCKAP1L
NDST2
NEDD4L
NF1
NFATC2IP
NFKB2
NFKBIA
NLRP12
NLRP3
NLRX1
NOD2
NOS1
NOS2
NOS3
NOX1
NOX3
NOX4
NOX5
NOXA1
NOXO1
PADI4
PCCA
PCCB
PECAM1
PGLYRP1
PGM3
PIK3CA
PIK3CB
PIK3CD
PIK3CG
PIK3R1
PLA2G1B
PLAU
PLAUR
PNP
PRAM1
PREX1
PRG3
PRKCD
PRTN3
RAB27A
RAC1
RAC2
RAG1
RAG2
RASGRP4
RBP1
RECQL4
REL
RELB
RFX5
RFXANK
RFXAP
RMRP
RPS19
RUNX2
S100A12
S100A8
S100A9
SBDS
SELL
SELPLG
SERPINB1
SERPINB2
SERPINE1
SFRP1
SFTPD
SIPA1
SLAMF6
SLC11A1
SLC35A1
SLC35C1
SLC37A4
SLC46A1
SMAD3
SMARCAL1
SOCS1
SOCS3
SOD1
SOD2
SPI1
SPRED1
ST3GAL6
ST6GAL1
STAT3
STAT5A
STAT5B
STK4
STX11
STXBP2
STXBP3
SYK
TAZ
TCIRG1
TCN2
TGFB1
THBS1
TIA1
TIMP1
TIMP2
TIMP3
TINF2
TIRAP
TLR2
TLR4
TLR5
TLX1
TNF
TNFAIP3
TNFRSF1A
TRAF2
TREM1
TRPM2
TSTA3
TUSC2
TWSG1
TXNRD1
TYROBP
UNC13D
USB1
VAMP7
VAV1
VAV2
VAV3
VPS13B
VPS45
WAS
ZFP36

SUPPLEMENTAL FIGURES

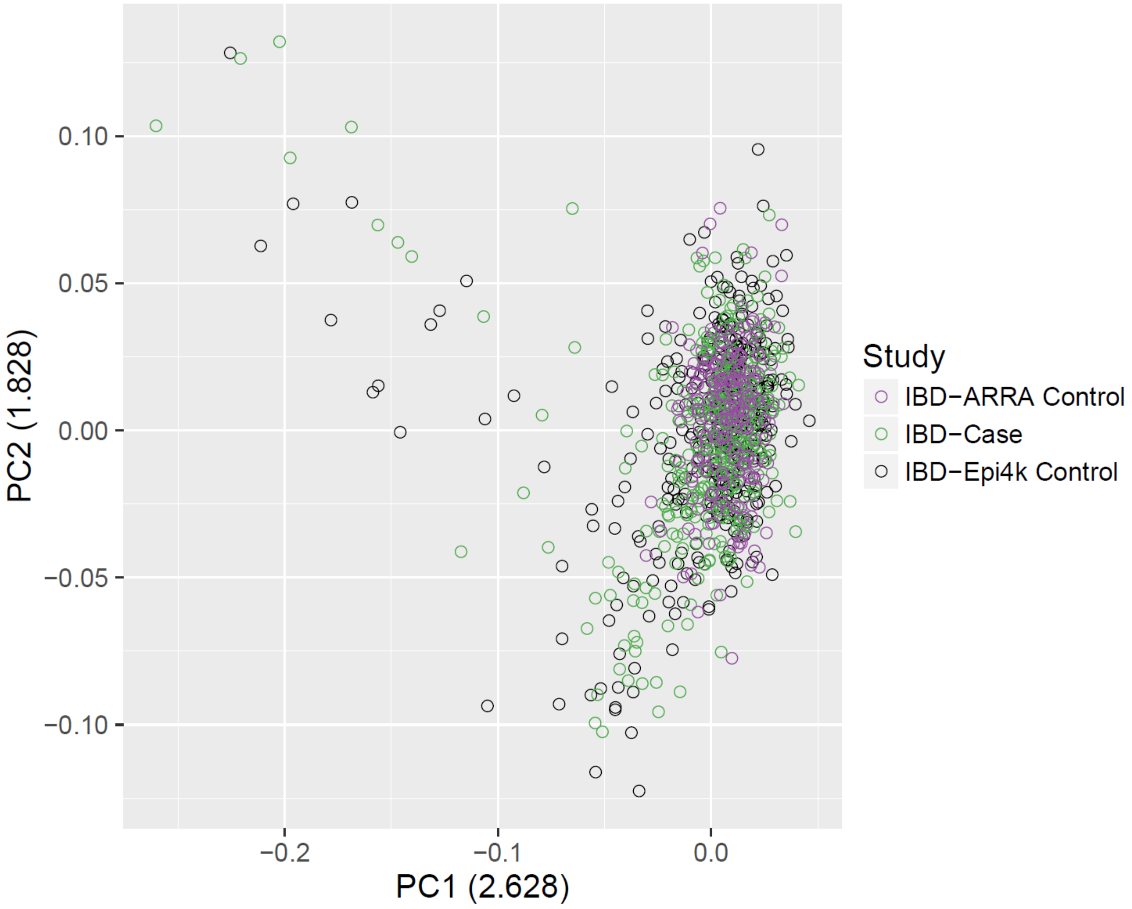
Supplemental Figure 1A. Principal component analysis of our cohort anchored with HapMap data before filtering outliers.



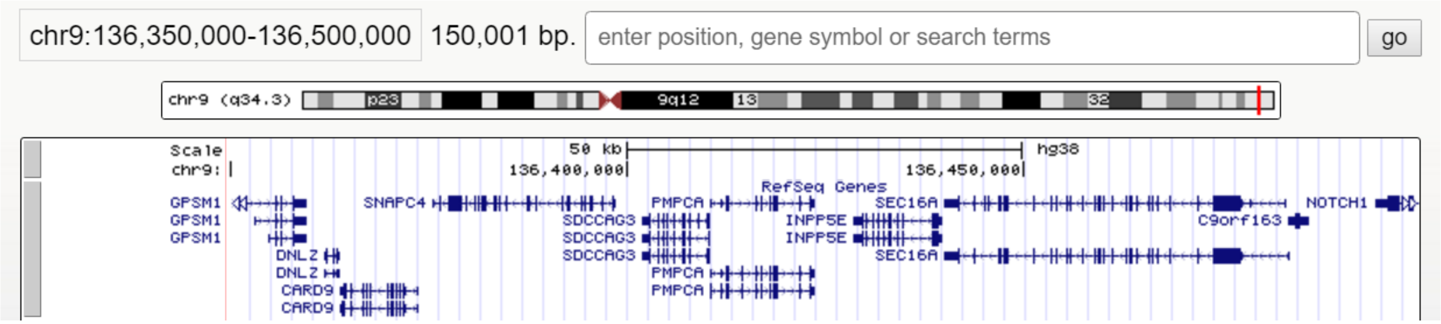
Supplemental Figure 1B. Principal component analysis of our cohort anchored with HapMap data after filtering outliers.



Supplemental Figure 1C. Post-filtering principal components of our cases and controls only. The first 4 principal components were significant by Tracy-Widom tests and were therefore used as covariates in our analyses of dbGaP data.



Supplemental Figure 2. Nine out of the top 16 most significant variants found in our logistic regression analysis were in this ~150kb region on chromosome 9.



chr9:136,350,000-136,500,000 150,001 bp.

CHAPTER IV: Dysbiosis, Inflammation, and Response to Treatment: a Longitudinal Study of Pediatric Subjects with Newly Diagnosed Inflammatory Bowel Disease

Coauthors: Madeline Bertha, Tatyana Hofmekler, Pankaj Chopra, Tommi Vatanen, Abhiram Srivatsa, Jarod Prince, Archana Kumar, Cary Sauer, Michael E. Zwick, Glen A. Satten, Aleksandar D. Kostic, Jennifer G. Mulle, Ramnik J. Xavier, and Subra Kugathasan

ABSTRACT

Background: Gut microbiome dysbiosis has been demonstrated in subjects with newly diagnosed and chronic inflammatory bowel disease (IBD). In this study we sought to explore longitudinal changes in dysbiosis and ascertain associations between dysbiosis and markers of disease activity and treatment outcome.

Methods: We performed a prospective cohort study of 19 treatment-naïve pediatric IBD subjects and 10 healthy controls, measuring fecal calprotectin and assessing the gut microbiome via repeated stool samples. Associations between clinical characteristics and the microbiome were tested using generalized estimating equations (GEE). Random forest classification was used to predict ultimate treatment response (presence of mucosal healing at follow-up colonoscopy) or non-response using patients' pre-treatment samples.

Results: Patients with Crohn's disease (CD) have increased markers of inflammation and dysbiosis compared to controls. Ulcerative colitis (UC) patients had even higher inflammation and dysbiosis compared to CD. For all cases, the gut microbial dysbiosis index associated significantly with clinical and biological measures of disease severity, but did not associate with treatment response. We found differences in specific gut microbiome genera between cases/controls and responders/non-responders including *Akkermansia, Coprococcus, Fusobacterium*, *Veillonella*, *Faecalibacterium*, and *Adlercreutzia*. Using pre-treatment microbiome data in a weighted random forest classifier we were able to obtain 76.5% accuracy for prediction of responder status.

Conclusions: Patient dysbiosis improved over time but persisted even among those who responded to treatment and achieved mucosal healing. Although dysbiosis index was not significantly different between responders and non-responders, we found specific genus-level differences. We found that pre-treatment microbiome signatures are a promising avenue for prediction of remission and response to treatment.

BACKGROUND

Inflammatory bowel disease (IBD), including Crohn's disease (CD) and ulcerative colitis (UC), is characterized by chronic remitting and relapsing inflammation of the gastrointestinal tract. Persistent inflammation and continuing insult lead to fibrosis, scarring, and the need for multiple surgeries. The pathogenesis of IBD is complex and poorly understood. A disturbance of intestinal mucosal homeostasis, influenced by genetic factors, the intestinal microbiome, the immune system, and environmental exposures, is believed to underlie IBD[1] [2]. While 200 distinct genetic loci have been associated with IBD in a recent report [3], many of these genes point to pathways involving bacterial recognition or host response to microbial infections, both clearly influenced by the environment. Although the prevalence of adult-onset IBD has plateaued in the Westernized world, recent population-based studies in IBD from Canada [4], USA [5], and Europe [6] suggest a rapid increase in pediatric-onset IBD, particularly in children younger than 10 years. Genetic causes are unlikely to account for these epidemiological findings. The risk of IBD among first-generation immigrants to the Western world from south Asia and Africa, as well as the prevalence of IBD in native Asia or Africa, are exceedingly low, yet second-generation immigrants have a greatly increased risk similar to the location to which they immigrated [7]. This emerging global rise of pediatric IBD incidence has fueled a quest to identify early life exposures including potential microbiome alterations due to lifestyle and diet that could explain the increasing risk for IBD among children [8, 9].

Several studies have described characteristic patterns within the gut microbiome of IBD patients [10-13]. In general, shifts in bacterial taxa and decreased community diversity have been found in treatment-naïve CD [14] and in IBD in general [15-17], with the extent of dysbiosis associated with severity of inflammation [18]; however, it is not clear whether these changes are a cause or consequence of IBD [2]. In one recent study involving a large number of subjects, the microbiome of treatment-naïve pediatric CD patients had a distinct signature compared to non-IBD subjects, as measured by both fecal and intestinal mucosa bacterial ecosystems [19]. However, this study used primarily mucosal biopsies and was limited to a single time point—it did not capture the dynamics of the gut microbiome over time. One recent study showed that dysbiosis results from independent effects of inflammation, diet, and antibiotics after selected pediatric Crohn's disease subjects were treated with enteral nutrition and some conventional medications [18]. Although this study measured bacterial community before and after intervention, the study only provided data for an 8-week study period and only 4 samples per patient. Long term data are still lacking regarding dysbiosis subjects who undergo standard of care treatment in clinical practice. Once IBD is diagnosed, patients undergo a series of treatments to induce clinical remission, in which mucosal healing is promoted by controlling mucosal inflammation. Some patients respond clinically to treatment with normalization of symptoms and evidence of mucosal healing seen in repeat colonoscopies ("responders" or "remitters"); other patients continue to have persistent inflammation or a remitting-relapsing disease course with a variable degree of mucosal inflammation (“non-responders” or “non-remitters”). It is critically important to study the intestinal microbiome over the course of treatment to identify whether there are microbial signatures that distinguish these different outcomes. This can be achieved with longitudinal microbiome analysis, starting at diagnosis and following up throughout treatment in parallel with clinical characterization. We hypothesize that distinct signatures of microbiota can be found and applied in clinical practice to assess ongoing inflammation and predict response to treatment. An important study by Kolho et al examined the treatment responses using fecal calprotectin in patients with median disease duration of 3.5 years after diagnosis [20]. Although our study was similar, our study design differed from Kolho et.al in that we used mucosal healing in addition to fecal calprotectin as measure of mucosal inflammation and used sequencing rather than phylogenetic microarray to classify species levels.

Here we report the results of a longitudinal investigation of 19 children diagnosed with IBD, of whom 15 had a final diagnosis of CD and 4 had a final diagnosis of UC. All 19 subjects were recruited from a single center, were treatment-naïve at the time of enrollment, were treated with current standards of practice guidelines, and were followed clinically for a median of 8 months. Treatment regimens were not protocolized, but treatment was escalated to maximal medical therapy or surgical resection was recommended if, upon clinical evaluation, the subject was categorized as a non-responder to previous treatment. We also recruited and followed 10 unaffected controls for comparison: 6 family members and 4 unrelated controls. We measured fecal calprotectin in all samples as an objective measure of inflammation as well as the subjective clinical disease activity indices (pediatric CD activity index (PCDAI) or pediatric UC activity index (PUCAI)). The strength of our study lies in the dense longitudinal data collection (217 total visits—a median of 8 time points for both cases and controls), thorough clinical characterization of our patients at each visit, measurement of clinical disease activity indices, and simultaneous use of fecal calprotectin as an objective measure of mucosal inflammation. We comprehensively analyzed inflammation, diversity and dysbiosis by standard methods including the previously described dysbiosis index, explored gut microbiome differences at the genus level among cases and controls and treatment responders and non-responders, and finally assessed the ability of pre-treatment samples to predict treatment response.

METHODS

Study Population

Potential participants were identified from Children's Healthcare of Atlanta inpatient wards and outpatient pediatric IBD clinics based on clinical suspicion of IBD based on symptoms or lab work. Criteria to participate in the study included CD or UC diagnosis confirmed by colonoscopy and/or magnetic resonance enterography, willingness to participate, and ability to maintain close follow-up. Patients and families gave informed consent and assent to participate in the study. Exclusion criteria included prior diagnosis of IBD, prior therapy with immunomodulators or biologics, or history of non-compliance with clinical appointments.

A total of 19 pediatric IBD cases (≤ 17 years old, 15 with CD and 4 with UC) were enrolled in this longitudinal prospective study between June 2013 and January 2014. Participants were followed at regular intervals beginning at the time of enrollment until the termination of the study in August 2014. All patients were phenotyped at the time of enrollment according to the Paris Classification [21] . Demographic and phenotypic characteristics were collected via patient interview and chart review at the time of sample delivery, and abbreviated PCDAI [22-24] or PUCAI was obtained at all clinical visits [25].

Medical treatment was not affected by joining this study. Patients started to receive treatment between their first and second clinical visits. Patients were treated with aggressive monotherapy of either immunomodulators or biologics with mucosal reassessment via colonoscopy approximately one year after diagnosis. Based on presence or absence of mucosal healing we dichotomized patients as responders (n = 6) or non-responders (n = 13), respectively, independent of any knowledge about microbiome composition. Since subjects received multiple treatments, we did not categorize based on the particular treatment exposures. Patients receiving surgery were classified as non-responders, and only pre-surgery time points were used in analyses. Family members of patients were recruited as related controls (n = 6), and unrelated controls ≤ 17 years old with no IBD diagnosis were also recruited (n = 4). Once enrolled, participants were followed no more frequently then weekly.

Specimen Collection and Processing

Fecal samples were obtained at regular intervals beginning at the time of diagnosis and throughout the study (Figure 1). Each fecal sample was collected and placed into two separate Para-Pak Vials: (i) with 100% ethanol (ii) without ethanol. The specimen with ethanol was submitted to the study coordinator at room temperature for processing within 24 hours of collection. The specimen was spun down, ethanol discarded, and the remaining stool was either stored at -20°C until ready for aliquoting, or immediately aliquoted to be stored at -80°C for fecal microbiome analysis. The specimen without ethanol was stored at -20°C until it was aliquoted and stored at -80°C for fecal calprotectin analysis. Fecal calprotectin was measured by Eagle Biosciences Calprotectin enzyme-linked immunosorbent assay (ELISA) kits according to manufacturer's guidelines.

Bioinformatic Processing

In collaboration with the Broad's Molecular Biology R&D (MBRD) Lab, we sequenced the V4 region of the bacterial 16S rRNA gene using the Illumina MiSeq platform according to manufacturer's specifications. Reads were demultiplexed into fastq files for each sample using sequence barcodes. Forward and reverse reads were joined with PANDASeq [26]. After samples with fewer than 3,000 reads were excluded, there was a median of 66,000 reads per sample used in the study. The joined sequence files were formatted using a Python script to add QIIME headers with the respective sample ID to each sequence before concatenating into one file for input into QIIME 1.8.0 [27]. Operational taxonomic units (OTUs) were picked using the QIIME pick_closed_reference_otus.py script with a threshold of 97% identity to the Greengenes v13_8 database. A median of 91% of reads per sample were classified successfully with this closed-reference OTU approach. Shannon alpha diversity was calculated on the unfiltered biom table using the alpha_diversity.py script, and weighted UniFrac distances were calculated with the beta_diversity.py script. The microbial dysbiosis index, initially described by Gevers 2013, was calculated in R for each sample. The microbial dysbiosis index is defined as the $\log_{10}$ of the total abundance in organisms increased in CD divided by the total abundance of organisms decreased in CD. The increased-in-CD taxa comprise *Enterobacteriaceae*, *Pasteurellaceae*, *Fusobacteriaceae*, *Neisseriaceae*, *Veillonellaceae*, and *Gemellaceae*. Decreased-in-CD taxa are *Bacteroidales*, *Clostridiales* (excluding *Veillonellaceae*), *Erysipelotrichaceae*, and *Bifidobacteriaceae* [19].

To test the robustness of our findings from these Shannon diversity and dysbiosis calculations, we repeated association tests between cases and controls using our data with 1) a *de novo* OTU clustering approach and 2) rarefying to even sequencing depth. Our *de novo* analysis was performed the same as our original closed-reference analysis with the exception that chimeras were first removed from each sample using USEARCH v6.1 [28], then OTUs were picked using the pick_de_novo_otus.py script. Taxonomic classification was performed using the same Greengenes database. The same median percentage of sequences were ultimately successfully classified (91%) using this *de novo* approach.

We randomly rarefied each sample in our original closed-OTU biom table to 3155 sequences, the lowest sequencing depth observed in our samples, using the rrarefy function in the R package vegan [29]. We then measured Shannon diversity using vegan's diversity function and calculated the dysbiosis index using the same R code described previously. We repeated this 10,000times and took the median of the results from these rarefactions for each sample; we then repeated our regression analyses using these values. For a complete summary of reads/sample, QC information, and calculated values, see "reads_microbiome_info" supplemental file.

Overall there were 7628 OTUs in our samples. For our genus-by-genus and random forest analyses we collapsed data to the genus level (combining OTUs belonging to the same genus) and converted counts to frequencies using the summarize_taxa.py QIIME script. There were 397 genus-level taxa in our 158 microbiome samples. To test for significance, we required a genus to be present at greater than 0.15% abundance in at least one sample, leaving 134 genera.

Statistical Analysis

We performed all data analyses in R. To account for the correlations within individuals over time, we performed linear regressions in a generalized estimating equations (GEE) framework [30] using the R package geepack [31]. We assumed an independent correlation structure and used the robust (sandwich) estimator for standard error. Subject observations were additionally inversely weighted by the total number of observations for that individual to ensure that results were not driven by individuals who were observed more frequently [32]. Wald tests were used to assess the significance of coefficients in our GEE. To compare marker levels between groups we modeled markers (calprotectin, dysbiosis, diversity) as a function of disease status (case vs control or UC vs CD). To assess differences between groups at baseline (all clinical outcomes as well as genus-by-genus analysis), or to measure changes over time we considered models with time since study enrollment. When comparing change over time between CD, UC and controls, time by diagnosis interactions were also considered. We used the same models without time to assess average differences between groups over the course of disease. For associations between pairs of markers (e.g. calprotectin and dysbiosis) throughout the course of our study, we modeled one marker (e.g. calprotectin) as a function of the other marker (e.g. dysbiosis).

Predictive Modeling

We used the R package randomForest [33] and genus frequency data from each subject's first pretreatment fecal sample (available for 5 responders and 12 non-responders) to train a random forest with 25,001 trees to predict response or non-response. Trees were grown to the maximum size possible; by default, 12 genera (the square root of the number of input genera) were considered as candidates at each split, and splitter importance was calculated as mean decrease in the Gini impurity, described in the randomForest documentation [33]. Because of the small sample size, we did not differentiate between UC and CD patients for this analysis. To assess if this was reasonable, we calculated the proportion of the variance in weighted Unifrac distances between patients' pretreatment samples explained by response/non-response status and IBD subtype using permutational ANOVA

(PERMANOVA) as implemented in the adonis function in the R package vegan [29]. To account for unequal sample sizes of responders and non-responders in our random forest, we used weights equal to the inverse of the sample size of each class; the cost of misclassifying responders therefore equaled the cost of misclassifying non-responders. We also performed the analysis with equal class sizes (5 each of responders and non-responders) to ensure our results were not the result of the class imbalance of our cohort. The receiver operating characteristic (ROC) curves and the area under the ROC curves (AUC) were generated using the ROCR package in R [34]. The significance of prediction accuracy and AUC was assessed by permuting response/non-response status 10,000 times.

RESULTS

Extensive Characterization of Gut inflammation and Microbiome in a Longitudinal Cohort of Children with IBD. Twenty-nine individuals were included in the longitudinal analysis, representing four groups: CD patients (n = 15), UC patients (n = 4), unaffected controls with a first-degree genetic relationship to an affected individual (family members, n = 6), and unaffected controls with no genetic relationship to any affected individual included in this study (unrelated, n = 4). Table 1 shows a summary of clinical characteristics and total number of visits used in analysis for all study participants. A more detailed summary of number of microbiome measures, calprotectin values, and PCDAI time points by case/control group is provided in Table S1. Figure 1 shows a comprehensive visualization of calprotectin measures for all patient and control time points used in all analyses. GEE comparison of familial and unrelated controls showed no significant differences at baseline, and no differences in average fecal calprotectin or alpha diversity between the two groups. However, on average unrelated controls had a higher dysbiosis index than related controls

(Table S2). These groups were pooled into one group of controls for all subsequent analyses, so our results were not inflated by the lower dysbiosis index apparent in related controls.

Subjects with IBD Have Increased Markers of Inflammation and Dysbiosis Compared to Controls. We first we tested general differences in inflammation, microbiome diversity, and microbial dysbiosis between IBD cases and controls using our weighted GEE approach to properly control for correlations within individuals. Significance of these coefficients was assessed via Wald tests. Table S3 summarizes beta and p-value information for comparisons of baseline values (including time since first sample as a covariate) and overall averages. Figure 2 shows calprotectin, alpha diversity, and dysbiosis for all timepoints for controls, CD patients, and UC patients (Figure S1 shows all time points summarized in box-and-whisker plots; Figure S2 shows controls, responders, and non-responders over time with a different color for each individual).

For controls, baseline calprotectin was 42 ± 99 μg/g. CD patients had fecal calprotectin values 313 μg/g higher at baseline than controls ($p = 0.0002$), and UC patients had values 1330 μg/g higher than controls ($p = 4E-11$; Table S3 summarizes all CD/UC/control comparisons). Over the entire course of our study the average difference in fecal calprotectin for CD and UC patients compared to controls was 181 μg/g ($p = 0.00002$) and 1100 μg/g ($p = 4E-10$), respectively. As seen in previous studies, IBD patients had overall lower alpha diversity as measured by the Shannon index. Shannon index at baseline for controls was 6.02 ± 0.58. CD patients had Shannon index values 0.94 lower at baseline ($p = 0.00001$) and 0.72 lower on average ($p = 0.007$) relative to controls. UC patients had Shannon values 1.31 lower at baseline ($p = 8E-05$) and 0.98 lower on average ($p = 0.002$).

Our sample of IBD patients also had significantly higher scores on the dysbiosis index than controls. At baseline, mean control dysbiosis index was -1.85 ± 0.55. Baseline dysbiosis was 0.86 points higher for CD patients (p = 6E-8) and 1.75 points higher for UC (p = 4E-15). Dysbiosis scores were on average 0.67 points higher in CD (p = 3E-07) and 1.38 points higher in UC (p = 3E-10).

Our microbiome findings of decreased Shannon diversity and increased dysbiosis did not change when we calculated these values after de novo OTU-picking, or after taking the median of 10,000 rarefactions to the lowest sequencing depth seen in our closed biom table (see "denovo_and_rarefy_analysis" supplemental file for a comparison of these approaches to results of our original closed-reference OTU approach).

UC patients had significantly higher calprotectin and dysbiosis indices than CD patients (Figure 2, Table S4). UC patients had fecal calprotectin levels 829 μg/g higher at baseline (p = 2E-05) and 917 μg/g higher on average (6E-06) compared to CD patients. The dysbiosis index was 0.49 points higher among UC patients at baseline (p = 0.02) and 0.70 points higher on average (0.0007) than CD patients. While Shannon diversity was lower in our UC patients this difference was not significant, possibly due to the relatively small sample size of our cohort.

Our longitudinal samples also show improvements in outcome measures over time for IBD patients (Figure 2), reflecting overall response to treatment, while these measures did not significantly change for controls over the course of the study (Table S3). Calprotectin declined in patients with CD relative to controls (p = 0.02), and in UC patients, calprotectin declined at around four times the rate of CD compared to controls (p = 3E-06). An increase in Shannon diversity relative to controls was not significant for CD patients, but Shannon diversity did improve over the course of the study for patients with UC compared to controls (p = 0.002). Both CD and UC patients showed improvements (decreases) in the microbial dysbiosis index compared to controls (p = 0.03 and p = 1E-13, respectively), with UC patients having a higher comparative rate of decline.

Dysbiosis Associates Significantly with Clinical and Biological Measures of Disease Severity. Our next aim was to test whether dysbiosis showed an association with calprotectin in our cohort. Using GEE, we found higher dysbiosis associated significantly with higher calprotectin (Table S5). In the overall dataset including both cases and controls, one unit increase in microbial dysbiosis (overall mean -1.3 ± 0.74) was associated with a 260-point increase in calprotectin (p = 0.0004). This finding also held true when examining cases only: a one-unit increase in dysbiosis (case mean -1.06 ± 0.66) associated with 286 μg/g higher calprotectin (p = 0.02, Figure S3A). This is the first time the dysbiosis characteristic of the CD gut microbiome has been linked to a clinical measure of inflammation, fecal calprotectin. In contrast, we found that Shannon alpha diversity did not show a relationship with calprotectin (Table S5). Our results were not impacted by using a de novo OTU-picking approach, or rarefying reads from each sample from the closed-OTU-picking biom file to even depth (see "denovo_and_rarefy_analysis" supplemental file).

For our Crohn's patients, dysbiosis also significantly associated with increased PCDAI, the current clinical measure of disease activity (p = 0.0001, Figure S3B). However, PCDAI did not associate significantly with calprotectin (Table S5, Figure S3C), suggesting that PCDAI is not a good stand-in for a direct measure of inflammation such as calprotectin.

Gut Microbiome Differences between Groups. While the dysbiosis index has predictive power of whether an individual has CD [19], we found that baseline dysbiosis index was not significantly different (p = 0.3) between treatment responders, who showed evidence of mucosal healing (n = 6), and non-responders (n = 13). This finding suggests that baseline dysbiosis may identify cases, but may not be the best tool for predicting actual response to treatment. Because the components of the dysbiosis index are broad categories (i.e., family- and order-level taxa), we next used GEE (again with Wald tests for coefficient significance) to test whether distinct microbiome signatures could be identified among responders and non-responders at the genus level. Using GEE allowed us to leverage the power of all of our time points to test differences, both between cases and controls and non-responders and responders.

We found 20 genera had nominally significantly different abundance (p ≤ 0.05) between cases and controls at baseline. Interestingly, 7 of these 20 genera were not captured by the dysbiosis index. We found also found 18 genera that differed significantly at baseline between responders and non-responders, 5 of which were not captured in the dysbiosis index. The taxa that differ between groups are summarized in Figure 3 and Table S6.

When we compared the list of significantly different genera between cases and controls to the significant genera from our non-responder/responder comparison, 11 of these taxa overlapped. The direction of effect in all overlapping taxa was the same in the two comparisons: if a genus was significantly increased in cases compared to controls, that genus was likewise increased in our non-responders compared to responders.

Because of our limited sample size, this analysis was largely exploratory: only 2 taxa, *Coprococcus* and *Adlercreutzia*, met the threshold for significance in the case/control comparison (no taxon met this threshold in our non-responder/responder comparisons) after conservative Bonferroni correction for multiple tests, with a significant p-value defined as <0.05/134. *Coprococcus* was decreased in cases compared to controls and further decreased in non-responders compared to responders. *Adlercreutzia* was also decreased in cases compared to controls but was at similar levels in non-responders and responders. While the association of *Coprococcus* with IBD has long been known, the association with *Adlercreutzia* has not been previously reported.

Predicting Future Response to Treatment via the Gut Microbiome Using Pre-Treatment Samples. We used a random forest classifier to determine if treatment response among cases could be predicted using microbiome data from the first pre-treatment sample from each individual. Five responders and twelve non-responders had pre-treatment samples for analysis. We combined UC and CD patients because IBD subtype explained only 4% of the variability in the weighted Unifrac distance between pretreatment samples after accounting for responder/non-responder status, which explained 23% of the variability (p=0.01 after 10,000 permutations). Our classifier attained an area under the ROC curve (AUC) of 0.75 (Figure 4A) and 76.5% accuracy of prediction (significant at p=0.04 and p=0.03, respectively, after 10,000 permutations of treatment response/nonresponse status). The confusion matrix and precision-recall curves for our random forest model can be found in Table S7 and Figure S4, respectively. Because the prediction error among responders in this model is high (60%) we were concerned that only non-responders had a distinctive pattern; this could also lead to a higher prediction error (lower accuracy) than reported here among populations having a higher proportion of responders. To investigate this, we additionally used a subsampling approach to fit our random forest classifier, so that each tree was fit using 5 responders and 5 non-responders. This model has the same overall prediction accuracy (76.5%) but the prediction error in responders (20%) and non-responders (25%) is more comparable, suggesting both responders and non-responders have distinct OTU profiles. These results also suggest that the prediction accuracy we report here is achievable even in populations with varying proportions of responders. The confusion matrix for the subsampled model can be found in Table S8 and the ROC and precision-recall curves can be found in Figure S5.

The abundances of genera with the top 15 highest variable importance scores in our weighted random forest (listed with importance scores in Table S9) are shown in Figure 4B. Figure S6 shows stacked bar charts for each sample used in the random forest (categorized by eventual response or non-response) summarizing those of the top 15 genera that were found above 1% average abundance. Four of the top fifteen genera (*Coprococcus*, *Adlercreutzia*, *Dialister*, and an unnamed genus of *Enterobacteriaceae*) overlapped with our GEE results. This overlap is denoted with asterisks in Figure 3A and Figure 4B. Three of these genera were the most significant in our GEE groupings, further implicating their significance in our IBD patients: *Coprococcus* was most significant of the genera in both case/control and responder/non-responder comparisons, *Adlercreutzia* was most significant in the case/control comparisons, and *Dialister* the most significant in responder/non-responder comparisons. Furthermore, *Coprococcus* and *Adlercreutzia* were the two genera that remained significant in our case/control analysis (both with decreased abundance) after Bonferroni correction of our GEE results. Importantly, fourteen of the top fifteen most important genera identified are identical between the weighted and equal sampling analyses (Table S10), supporting the conclusion these taxa are truly responsible for separating responders and non-responders in our cohort. Replication in a larger study will be needed to confirm the role of these taxa in treatment response.

DISCUSSION

We conducted the largest longitudinal study published to date following newly diagnosed IBD subjects in real time, collecting measures of disease activity, mucosal inflammation, and microbiome composition. Sample collection was initiated at diagnosis, prior to treatment, and continued throughout the medical and surgical management of these patients. Here we show that (1) longitudinal stool sampling was both feasible and robust; (2) microbial dysbiosis improved from baseline but persisted despite complete cessation of clinical disease activity among responders; (3) distinct microbiota signatures emerged among responders compared with non-responders at the genus level, but not dysbiosis index; (4) treatment-naïve analysis of the microbiome could potentially be used to predict whether a subject will respond to treatment. Our study was based on real day-to-day clinical practice, so study design did not impact treatment choices for the subjects. Using this approach, our patients could be treated in a manner consistent with standard-of-care. Our findings may prove clinically useful in tailoring therapies; if confirmed by a larger study, clinicians could, in the future, make microbiome-informed decisions about early escalation of medical therapies versus timely surgical interventions.

In our study, we focused on following patients over time using stool samples because obtaining repeated biopsy samples in a clinical setting is not feasible—it is invasive, expensive, and impractical for day-to-day clinical practice. We show that repeated stool samples can depict the diversity and dysbiosis of the microbiome. This is an important implication for future studies because it suggests that stool samples, which are relatively cheap and easy to acquire, are an appropriate substitute for biopsy samples to monitor the microbiome of IBD patients.

In terms of clinical outcomes, we assessed disease activity with PCDAI/PUCAI, the current standards in clinical use. These measures largely rely on clinician observation and patient self-report and are therefore indirect assessments of disease activity. Since inflammation impacts microbiome indices, many studies have been criticized for not having an objective measure of inflammation. To address this shortcoming, we measured fecal calprotectin as a proxy for mucosal inflammation [35, 36]. Fecal calprotectin is a quantitative measure of disease activity that is not affected by self-reporting bias and is a direct biomarker of mucosal inflammation, the trademark of IBD.

Previously, Gevers et al. [19] described the gut microbiome in treatment-naïve CD patients and created the dysbiosis index to reflect the distinct alteration of the microbiome in CD. We applied the dysbiosis index to our population and further showed it to be a useful and relevant tool: the dysbiosis index was significantly higher (indicating more dysbiosis) in both our CD and UC subjects compared to our unaffected subjects. Furthermore, the dysbiosis index decreased over the course of the study, consistent with treatment and subsequent clinical improvement. When it was created, the dysbiosis index showed strong correlation with clinical severity as measured by PCDAI, which we confirm in our study. We further share the novel finding that the dysbiosis index associates with the direct measure of inflammation, calprotectin. Because PCDAI does not show a similar association with higher calprotectin, the dysbiosis index may be more reflective of inflammatory status than the less direct disease activity measure.

Although our sample size is small, we showed that although the dysbiosis index was developed in CD patients, UC patients had significantly higher dysbiosis than CD patients did, along with increased calprotectin. Further, none of the responders in our study were UC patients. Additional studies in larger patient cohorts are needed to clarify any distinct features of the microbiome among IBD patients.

Our subjects were followed for an average of eight months and included patients who both responded and did not respond to treatment. Although the dysbiosis index improved over time in patients, it did not reach levels seen in controls. This finding has important implications for pathogenesis: it suggests that with aggressive treatment of inflammation and symptoms (as was the case in our population) disease activity will improve, but the gut microbiome may remain perturbed. This finding is in line with a recent paper by Forbes et al, who found that there was no clear difference between microbiota of inflamed and non-inflamed mucosa in either CD or UC, suggesting gut dysbiosis is the driver of inflammation rather than a result of it [37].

This pattern of persistent dysbiosis further emphasizes the need for prospective, longitudinal tracking with extensive follow-up: microbiome trends, microbiome resilience, and return to "healthy" composition may all be important to assess [38]. A larger study to investigate the impact of different treatments is also needed. Observations from such studies will open new therapeutic opportunities aimed at ameliorating dysbiosis in hopes of either preventing disease or limiting future complications.

At the individual genus level, several genera showed differences between groups in our GEE, random forest models, or both, with six bearing special mention: *Akkermansia, Coprococcus, Fusobacterium*, *Veillonella*, *Faecalibacterium*, and *Adlercreutzia*. In our sample, *Akkermansia* had a higher pretreatment abundance in non-responders compared to responders (Figure 4B). The genome of *Akkermansia*, identified in our random forest analysis, contains mucinase genes [39] and is considered to be a mucin-degrading bacterium [40]. In gnotobiotic mice, *Akkermansia* increases inflammation in mice co-infected with

Salmonella typhimurium [41]. We also found that *Coprococcus* (a genus identified in both GEE and random forest analyses) was diminished in cases compared to controls, and was further diminished in non-responders. In fact, agglutinating antibodies for *Coprococcus* were briefly considered as a biomarker for CD screening [42].

We have previously reported significantly higher abundance of *Fusobacterium* and *Veillonella* in the stool of treatment-naïve CD patients [19]. In our GEE analysis we again identified these two genera at increased abundance in cases, especially in non-responders to therapy. One recent study by Kelsen et al identified significantly increased levels of these two taxa, among others, in the subgingival microbiome of patients with CD who were not taking antibiotics [43]. This prompts the hypothesis that oral cavity microbiota, also seen in the guts of IBD patients may play a significant role in the pathogenesis and progression of IBD. Species of *Fusobacterium* are also associated with a wide variety of negative health outcomes, such as dental plaque, periodontal disease, Lemierre syndrome [44], head and neck infections [45], and especially colon cancer [46, 47].

Faecalibacterium, a genus of interest from our random forest analysis, includes the species *F. prausnitzii*. One particular strain of this species—A2-165—was recently found by Rossi et al to have an important role in anti-inflammatory processes. This bacterium was particularly adept at eliciting high levels of IL-10 production, enhancing ovalbumin-specific T cell proliferation, and reducing interferon-gamma-positive T cells. Treatment with A2-165 even attenuated inflammation in a murine model of chronic relapsing colitis [48]. Because *Faecalibacterium* abundance was found to be decreased in non-responders compared to responders, our study supports further investigation into the prognostic and therapeutic possibilities of this strain.

Another genus significant in both GEE and random forest analyses, *Adlercreutzia*, was found to be decreased in cases and further decreased in non-responders compared to responders. This genus was originally identified in human feces and found to play an important role in the metabolism of isoflavones to equol, a non-steroidal estrogen [49]. To our knowledge, the role of *Adlercreutzia* in IBD has not yet been explored; however, its appearance in the significant results of both our GEE and random forest analyses suggest it may be a future target of interest.

Genera from the families *Lachnospiraceae* and *Ruminococcaceae* appear several times in our GEE and random forest results. Though not included in the dysbiosis index, members of these families were found to be characteristic of tissue samples from Crohn's disease in a recent study by Tyler et al [50]. Four of the top fifteen most important genera identified by our classifier belong to the family *Lachnospiraceae* and are all reduced in non-responders compared to responders. Further research is needed into the possible contribution of members of this family to IBD pathophysiology.

Our study has several limitations. Some control subjects were related to affected subjects; however, the unrelated controls actually had significantly higher microbial dysbiosis than the related controls, suggesting shared environment did not overly inflate dysbiosis in the related study subjects. One factor that may have contributed to this trend is that some related controls were parents, and were hence older than the affected subjects. Additionally, there was variation in the number of samples obtained from each patient. To correct for this variation, we weighted samples for each study subject according to the number of samples they contributed to the study. Our sample population had a smaller number of UC subjects than CD subjects; although UC patients had higher measures of clinical activity, we combined these patients for predictive modeling because IBD disease type did not explain a large proportion of the variance between microbiome samples among IBD cases.

These unique data provide the first glimpse into the long-term dynamics of the gut microbiome of subjects with and without inflammatory bowel disease. The data show that the dysbiosis index captures alteration of the microbiome in IBD patients relative to controls, and associates with clinical and biochemical measures of disease activity. More importantly, the dysbiosis index did not decline to levels seen in unaffected individuals, even when patients were in remission. Distinct microbial signatures seen at the genus level among responders and non-responders may have clinical implications for therapeutics and risk stratification. The potential impact of this analysis is far-reaching, as it provides insight into how gut microbial dysbiosis changes with treatment and remission in IBD patients. Our results also lay the groundwork for predicting patients' ultimate response to therapy.

CONCLUSIONS

New findings:

- Markers of inflammation and dysbiosis are increased in IBD; microbial dysbiosis improves over time but persists despite cessation of clinical disease activity and mucosal healing among responders
- The dysbiosis index does associate with calprotectin, a measure of inflammation, but it does not distinguish treatment responders (those with mucosal healing) from non-responders. Other microbiome signatures do emerge at the genus level and warrant further investigation

Impact on clinical practice:

- Treatment-naïve analysis of the microbiome could potentially be used to predict whether a subject will respond to treatment
- Sustained and deep remission may require normalizing the gut dysbiosis

Disclaimer: The findings and conclusions in this report are those of the authors and do not necessarily represent the official position of the Centers for Disease Control and Prevention.

REFERENCES

1. Knights D, Silverberg MS, Weersma RK, Gevers D, Dijkstra G, Huang H, Tyler AD, van Sommeren S, Imhann F, Stempak JM, et al: Complex host genetics influence the microbiome in inflammatory bowel disease. *Genome Med* 2014, 6:107.
2. Manichanh C, Borruel N, Casellas F, Guarner F: The gut microbiota in IBD. *Nat Rev Gastroenterol Hepatol* 2012, 9:599-608.
3. Liu JZ, van Sommeren S, Huang H, Ng SC, Alberts R, Takahashi A, Ripke S, Lee JC, Jostins L, Shah T, et al: Association analyses identify 38 susceptibility loci for inflammatory bowel disease and highlight shared genetic risk across populations. *Nat Genet* 2015, 47:979-986.
4. Benchimol EI, Guttmann A, Griffiths AM, Rabeneck L, Mack DR, Brill H, Howard J, Guan J, To T: Increasing incidence of paediatric inflammatory bowel disease in Ontario, Canada: evidence from health administrative data. *Gut* 2009, 58:1490-1497.
5. Herrinton LJ, Liu L, Lewis JD, Griffin PM, Allison J: Incidence and prevalence of inflammatory bowel disease in a Northern California managed care organization, 1996-2002. *Am J Gastroenterol* 2008, 103:1998-2006.
6. Burisch J, Jess T, Martinato M, Lakatos PL, EpiCom E: The burden of inflammatory bowel disease in Europe. *J Crohns Colitis* 2013, 7:322-337.
7. Benchimol EI, Mack DR, Guttmann A, Nguyen GC, To T, Mojaverian N, Quach P, Manuel DG: Inflammatory bowel disease in immigrants to Canada and their children: a population-based cohort study. *Am J Gastroenterol* 2015, 110:553-563.
8. Molodecky NA, Soon IS, Rabi DM, Ghali WA, Ferris M, Chernoff G, Benchimol EI, Panaccione R, Ghosh S, Barkema HW, Kaplan GG: Increasing incidence and

prevalence of the inflammatory bowel diseases with time, based on systematic review. *Gastroenterology* 2012, 142:46-54 e42; quiz e30.

9. Malmborg P, Hildebrand H: The emerging global epidemic of paediatric inflammatory bowel disease - causes and consequences. *J Intern Med* 2015.

10. Kaakoush NO, Day AS, Huinao KD, Leach ST, Lemberg DA, Dowd SE, Mitchell HM: Microbial dysbiosis in pediatric patients with Crohn's disease. *J Clin Microbiol* 2012, 50:3258-3266.

11. Papa E, Docktor M, Smillie C, Weber S, Preheim SP, Gevers D, Giannoukos G, Ciulla D, Tabbaa D, Ingram J, et al: Non-invasive mapping of the gastrointestinal microbiota identifies children with inflammatory bowel disease. *PLoS ONE* 2012, 7:e39242.

12. Morgan XC, Tickle TL, Sokol H, Gevers D, Devaney KL, Ward DV, Reyes JA, Shah SA, Leleiko N, Snapper SB, et al: Dysfunction of the intestinal microbiome in inflammatory bowel disease and treatment. *Genome Biol* 2012, 13:R79.

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

14. Hansen R, Russell RK, Reiff C, Louis P, McIntosh F, Berry SH, Mukhopadhya I, Bisset WM, Barclay AR, Bishop J, 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, 107:1913-1922.

15. Sokol H, Seksik P: The intestinal microbiota in inflammatory bowel diseases: time to connect with the host. *Curr Opin Gastroenterol* 2010, 26:327-331.

16. Alipour M, Zaidi D, Valcheva R, Jovel J, Martinez I, Sergi C, Walter J, Mason AL, Wong GK, Dieleman LA, et al: Mucosal Barrier Depletion and Loss of Bacterial Diversity are Primary Abnormalities in Paediatric Ulcerative Colitis. *J Crohns Colitis* 2016, 10:462-471.

17. Knights D, Lassen KG, Xavier RJ: Advances in inflammatory bowel disease pathogenesis: linking host genetics and the microbiome. *Gut* 2013, 62:1505-1510.

18. Lewis JD, Chen EZ, Baldassano RN, Otley AR, Griffiths AM, Lee D, Bittinger K, Bailey A, Friedman ES, Hoffmann C, 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.

19. Gevers D, Kugathasan S, Denson LA, Vazquez-Baeza Y, Van Treuren W, Ren B, Schwager E, Knights D, Song SJ, Yassour M, et al: The treatment-naive microbiome in new-onset Crohn's disease. *Cell Host Microbe* 2014, 15:382-392.

20. Kolho KL, Korpela K, Jaakkola T, Pichai MV, Zoetendal EG, Salonen A, de Vos WM: Fecal Microbiota in Pediatric Inflammatory Bowel Disease and Its Relation to Inflammation. *Am J Gastroenterol* 2015, 110:921-930.

21. Levine A, Griffiths A, Markowitz J, Wilson DC, Turner D, Russell RK, Fell J, Ruemmele FM, Walters T, Sherlock M, et al: Pediatric modification of the Montreal classification for inflammatory bowel disease: the Paris classification. *Inflamm Bowel Dis* 2011, 17:1314-1321.

22. Turner D, Griffiths AM, Walters TD, Seah T, Markowitz J, Pfefferkorn M, Keljo D, Otley A, Leleiko NS, Mack D, et al: Appraisal of the pediatric Crohn's disease activity index on four prospectively collected datasets: recommended cutoff values and clinimetric properties. *Am J Gastroenterol* 2010, 105:2085-2092.

23. Hyams JS, Ferry GD, Mandel FS, Gryboski JD, Kibort PM, Kirschner BS, Griffiths AM, Katz AJ, Grand RJ, Boyle JT, et al.: Development and validation of a pediatric Crohn's disease activity index. *J Pediatr Gastroenterol Nutr* 1991, 12:439-447.

24. Shepanski MA, Markowitz JE, Mamula P, Hurd LB, Baldassano RN: Is an abbreviated Pediatric Crohn's Disease Activity Index better than the original? *J Pediatr Gastroenterol Nutr* 2004, 39:68-72.

25. Turner D, Otley AR, Mack D, Hyams J, de Bruijne J, Uusoue K, Walters TD, Zachos M, Mamula P, Beaton DE, et al: Development, validation, and evaluation of a pediatric ulcerative colitis activity index: a prospective multicenter study. *Gastroenterology* 2007, 133:423-432.

26. Masella AP, Bartram AK, Truszkowski JM, Brown DG, Neufeld JD: PANDAseq: paired-end assembler for illumina sequences. *BMC Bioinformatics* 2012, 13:31.

27. Caporaso JG, Kuczynski J, Stombaugh J, Bittinger K, Bushman FD, Costello EK, Fierer N, Pena AG, Goodrich JK, Gordon JI, et al: QIIME allows analysis of high-throughput community sequencing data. *Nat Methods* 2010, 7:335-336.

28. Edgar RC: Search and clustering orders of magnitude faster than BLAST. *Bioinformatics* 2010, 26:2460-2461.

29. Jari Oksanen FGB, Roeland Kindt, Pierre Legendre, Peter R. Minchin, R. B. O'Hara, Gavin L. Simpson, Peter Solymos, M. Henry H. Stevens and Helene Wagner: vegan: Community Ecology Package. 2016, R package version 2.3-4.

30. Zeger SL, Liang KY: Longitudinal data analysis for discrete and continuous outcomes. *Biometrics* 1986, 42:121-130.

31. Yan J: geepack: Yet Another Package for Generalized Estimating Equations. *R News* 2002, 2:12-14.

32. Williamson JM, Datta S, Satten GA: Marginal analyses of clustered data when cluster size is informative. *Biometrics* 2003, 59:36-42.

33. Wiener ALaM: Classification and Regression by randomForest. *R News* 2002, 2:18-22.

34. Sing T, Sander O, Beerenwinkel N, Lengauer T: ROCR: visualizing classifier performance in R. *Bioinformatics* 2005, 21:3940-3941.

35. Theede K, Holck S, Ibsen P, Ladelund S, Nordgaard-Lassen I, Nielsen AM: Level of Fecal Calprotectin Correlates With Endoscopic and Histologic Inflammation and Identifies Patients With Mucosal Healing in Ulcerative Colitis. *Clin Gastroenterol Hepatol* 2015.

36. Schoepfer AM, Vavricka S, Zahnd-Straumann N, Straumann A, Beglinger C: Monitoring inflammatory bowel disease activity: clinical activity is judged to be more relevant than endoscopic severity or biomarkers. In *J Crohns Colitis*, vol. 6. pp. 412-418; 2012:412-418.

37. Forbes JD, Van Domselaar G, Bernstein CN: Microbiome Survey of the Inflamed and Noninflamed Gut at Different Compartments Within the Gastrointestinal Tract of Inflammatory Bowel Disease Patients. *Inflamm Bowel Dis* 2016, 22:817-825.

38. Caporaso JG, Lauber CL, Costello EK, Berg-Lyons D, Gonzalez A, Stombaugh J, Knights D, Gajer P, Ravel J, Fierer N, et al: Moving pictures of the human microbiome. *Genome Biol* 2011, 12:R50.

39. van Passel MW, Kant R, Zoetendal EG, Plugge CM, Derrien M, Malfatti SA, Chain PS, Woyke T, Palva A, de Vos WM, Smidt H: The genome of Akkermansia muciniphila, a dedicated intestinal mucin degrader, and its use in exploring intestinal metagenomes. *PLoS One* 2011, 6:e16876.

40. Everard A, Cani PD: Diabetes, obesity and gut microbiota. *Best Pract Res Clin Gastroenterol* 2013, 27:73-83.

41. Ganesh BP, Klopfleisch R, Loh G, Blaut M: Commensal Akkermansia muciniphila exacerbates gut inflammation in Salmonella Typhimurium-infected gnotobiotic mice. *PLoS One* 2013, 8:e74963.

42. Wensinck F, van de Merwe JP, Mayberry JF: An international study of agglutinins to Eubacterium, Peptostreptococcus and Coprococcus species in Crohn's disease, ulcerative colitis and control subjects. *Digestion* 1983, 27:63-69.

43. Kelsen J, Bittinger K, Pauly-Hubbard H, Posivak L, Grunberg S, Baldassano R, Lewis JD, Wu GD, Bushman FD: Alterations of the Subgingival Microbiota in Pediatric Crohn's Disease Studied Longitudinally in Discovery and Validation Cohorts. *Inflamm Bowel Dis* 2015, 21:2797-2805.

44. Holm K, Svensson PJ, Rasmussen M: Invasive Fusobacterium necrophorum infections and Lemierre's syndrome: the role of thrombophilia and EBV. *Eur J Clin Microbiol Infect Dis* 2015.

45. Yusuf E, Halewyck S, Wybo I, Pierard D, Gordts F: Fusobacterium necrophorum and other Fusobacterium spp. isolated from head and neck infections: A 10-year epidemiology study in an academic hospital. *Anaerobe* 2015, 34:120-124.

46. Kostic AD, Gevers D, Pedamallu CS, Michaud M, Duke F, Earl AM, Ojesina AI, Jung J, Bass AJ, Tabernero J, et al: Genomic analysis identifies association of Fusobacterium with colorectal carcinoma. *Genome Res* 2012, 22:292-298.

47. Kostic AD, Chun E, Robertson L, Glickman JN, Gallini CA, Michaud M, Clancy TE, Chung DC, Lochhead P, Hold GL, et al: Fusobacterium nucleatum potentiates

intestinal tumorigenesis and modulates the tumor-immune microenvironment. *Cell Host Microbe* 2013, 14:207-215.

48. Rossi O, van Berkel LA, Chain F, Tanweer Khan M, Taverne N, Sokol H, Duncan SH, Flint HJ, Harmsen HJ, Langella P, et al: Faecalibacterium prausnitzii A2-165 has a high capacity to induce IL-10 in human and murine dendritic cells and modulates T cell responses. *Sci Rep* 2016, 6:18507.

49. Maruo T, Sakamoto M, Ito C, Toda T, Benno Y: Adlercreutzia equolifaciens gen. nov., sp. nov., an equol-producing bacterium isolated from human faeces, and emended description of the genus Eggerthella. *Int J Syst Evol Microbiol* 2008, 58:1221-1227.

50. Tyler AD, Kirsch R, Milgrom R, Stempak JM, Kabakchiev B, Silverberg MS: Microbiome Heterogeneity Characterizing Intestinal Tissue and Inflammatory Bowel Disease Phenotype. *Inflamm Bowel Dis* 2016, 22:807-816.

TABLES

Table 1. A summary of relevant characteristics is shown for study participants

Cases			
Diagnosis	Crohn's disease	15 (78.9%)	**Count (%)**
	Ulcerative colitis	4 (21.1%)	
Treatment outcome	Response/mucosal healing	6 (31.6%)	
	Non-response without surgery	8 (42.1%)	
	Non-response with surgery	5 (26.3%)	
Time points	microbiome	6 (1-12)	**Median (range)**
	calprotectin	6 (1-12)	
	PCDAI	7 (3-13)	

Controls			
Relatedness	Familial	6 (60%)	**Count (%)**
	Unrelated	4 (40%)	
Time points	microbiome	5 (1-8)	**Median (range)**
	calprotectin	6.5 (1-9)	
	PCDAI	NA	

FIGURES

Figure 1. Log$_{10}$(calprotectin+1) values for all study subjects used in analysis. Larger circle size reflects higher measured calprotectin. Time points where calprotectin was < 100 μg/g are shown in blue; time points where calprotectin was >100 μg/g are shown in red. CD, Crohn's disease; UC, ulcerative colitis; R, responder to treatment; NR, non-responder to treament; F, familial control; U, unrelated control. Patients are shown in order of decreasing length of followup. (See also Table 1 and S1.)

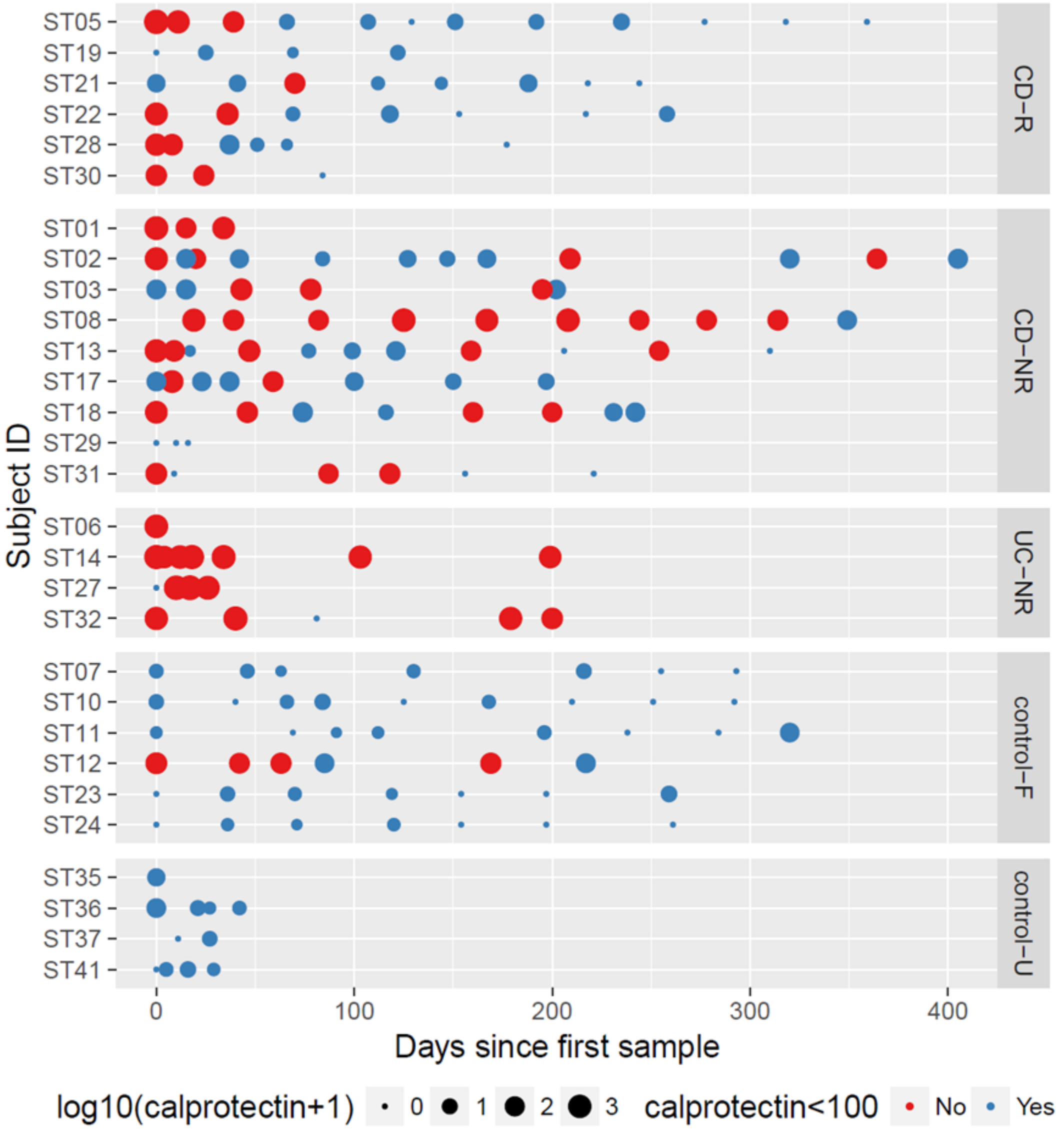


Figure 2. Clinical characteristics for all study subjects. (A-C) Characteristics for control subjects (black), Crohn's disease patients (CD, red), and ulcerative colitis patients (UC, blue) are plotted over time with unadjusted regression lines in black and 95% confidence intervals in grey. For CD and UC patients, calprotectin decreases (A), alpha diversity increases (B), and gut microbial dysbiosis decreases (C) over time, reflecting overall improvement following treatment. Additionally, calprotectin and microbial dysbiosis were significantly higher in our UC patients than in CD. (See also Figures S1 and S2, Tables S3 and S4.)

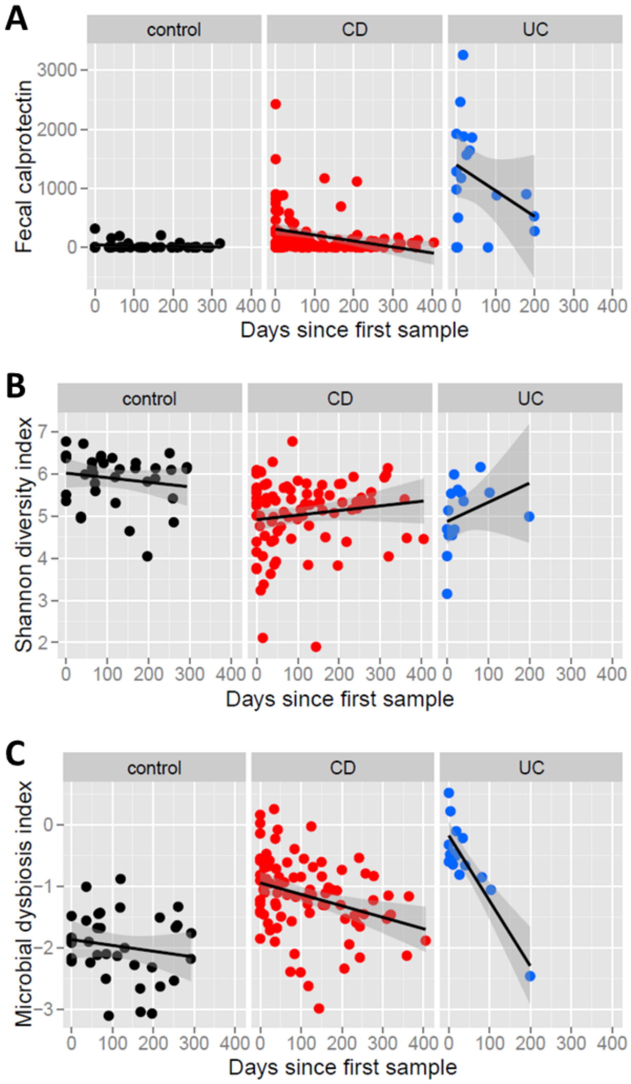


Figure 3. Genera with significant differences between cases and controls, non-responders and responders. (A) –log10(p value) from testing difference in abundance of each genus in cases compared to controls and non-responders compared to responders. Blue bars indicate taxa negatively associated with case or non-responder status, and red bars indicate a positive association. The line below 2 represents the threshold for nominal significance; the higher line is the significance level after Bonferroni adjustment for multiple tests. The asterisk denotes taxa that also appear in the results of our random forest classifier. (B-D) Example patterns representative of each of the three categories: (B) significant in both comparisons, (C) significant only between cases and controls, and (D) significant only between non-responders and responders. (See also Table S6.)

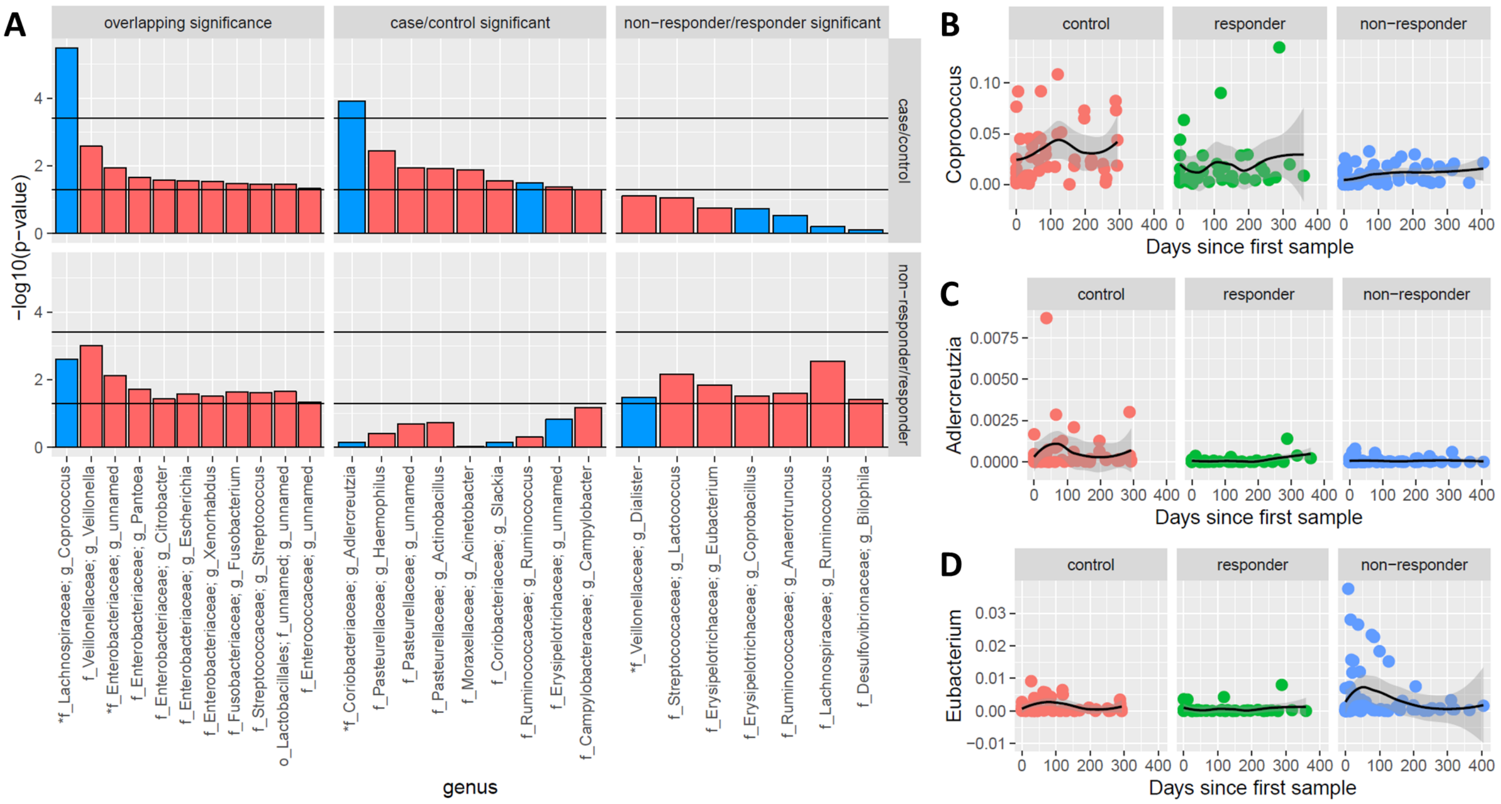
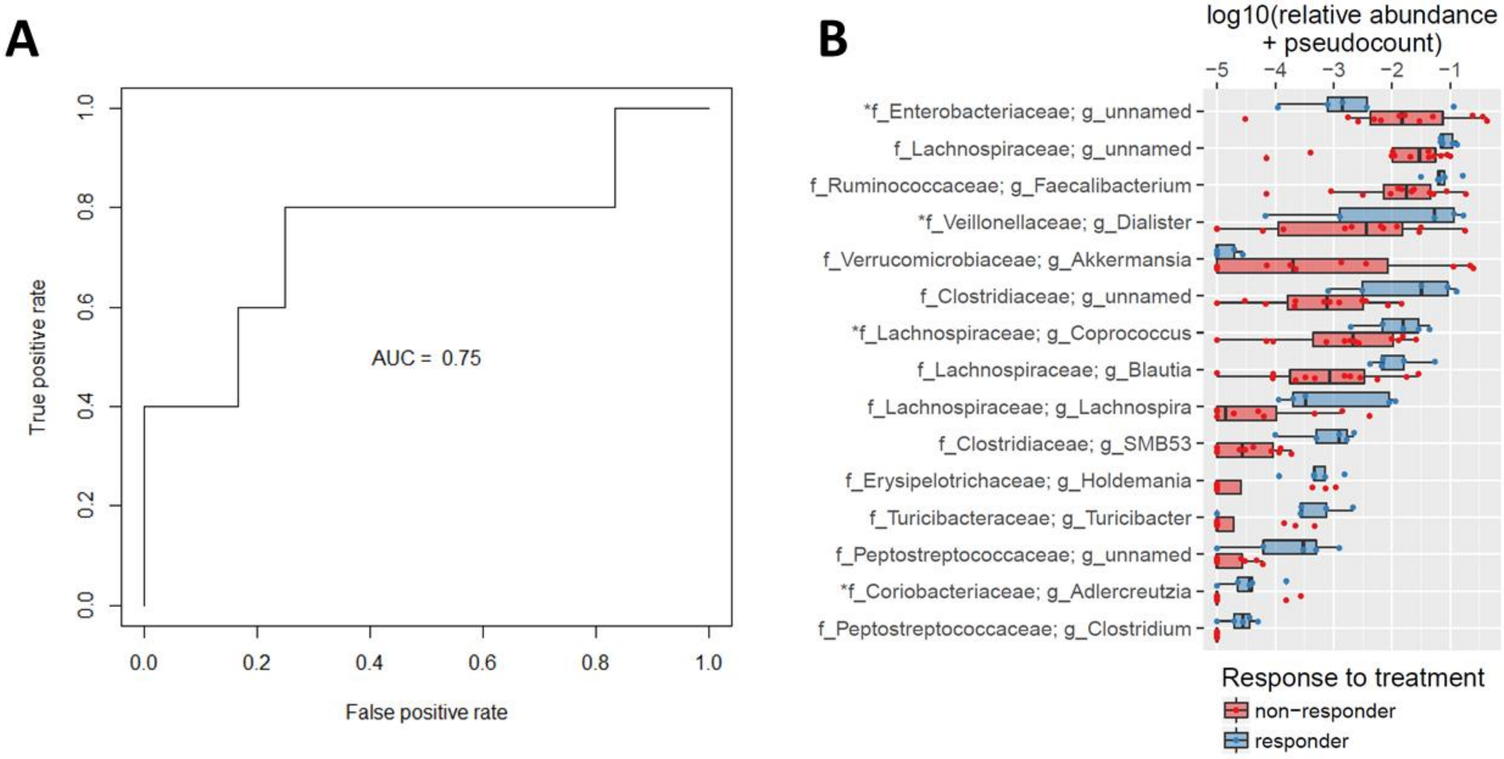


Figure 4. Use of genera to predict eventual response to treatment in pretreatment samples. (A) Our classifier classifies response status significantly better than random guess with AUC = 0.75 and overall accuracy of 76.5% for predicting treatment response/nonresponse. (B) Box plots of the $\log_{10}$ relative abundance plus pseudocount (1 x 10^{-5}) of the fifteen genera with highest importance scores in random forest analysis in responders and non-responders. The asterisk denotes taxa also identified as significant in our generalized estimating equations analysis. (See also Figure S4 and S6, Table S7 and S9.)



SUPPLEMENTARY TABLES

Table S1: summary of data available for all patients

		Total number of observations			Overlap w/ microbiome	
		microbiome	**calprotectin**	**PCDAI**	**calprotectin**	**PCDAI**
Total	**case**	111	125	120	103	97
	control	47	55	0	43	0
Median	**case**	6	6	7	5	6
	control	5	6.5	0	5	0

Table S2: statistical comparison of related (reference group) and unrelated controls

calprotectin mean ± SD = 23.7 ± 60

Shannon index mean ± SD = 5.85 ± 0.61

dysbiosis index mean ± SD = -1.87 ± 0.58

Difference at baseline:
y ~ relation + time

		X	
		beta	p-value
Y	calprotectin	-20.9	0.5
	Shannon	-0.098	0.7
	dysbiosis	0.32	0.1

Average difference:
y ~ relation

		X	
		beta	p-value
Y	calprotectin	-10.2	0.6
	Shannon	0.033	0.9
	dysbiosis	0.36	0.05

Table S3: statistical summary of differences between cases (stratified into CD and UC) and controls (reference group)

calprotectin mean ± SD = 240 ± 508; Shannon index mean ± SD = 5.24 ± 0.89; dysbiosis index mean ± SD = -1.28 ± 0.80

Difference at baseline:
y ~ diagnosis(control/CD/UC) + time + diagnosis*time

			diagnosis		time		diagnosis*time	
			beta	p-value	beta	p-value	beta	p-value
Y	calprotectin	CD	313	2E-04	-0.069	0.4	-1.03	0.02
		UC	1330	4E-11			-4.15	3E-06
	Shannon	CD	-0.94	1E-05	-1.1E-03	0.3	1.8E-3	0.1
		UC	-1.31	8E-05			6.3E-3	2E-03
	dysbiosis	CD	0.86	6E-08	-7.1E-04	0.2	-1.5E-3	0.03
		UC	1.75	4E-15			-0.011	1E-13
			(difference from controls)		(control change over time)		(change over time compared to controls)	

Average difference:
y ~ diagnosis(control/CD/UC)

			diagnosis	
			beta	p-value
Y	calprotectin	CD	181	2E-05
		UC	1100	4E-08
	Shannon	CD	-0.72	7E-03
		UC	-0.98	2E-03
	dysbiosis	CD	0.67	3E-07
		UC	1.38	3E-10
			(difference from controls)	

Table S4: statistical summary of differences between UC and CD (reference group)

calprotectin mean ± SD = 335 ± 584

Shannon index mean ± SD = 4.99 ± 0.86

dysbiosis index mean ± SD = -1.03 ± 0.75

Difference at baseline:
Y ~ diagnosis(UC/CD) + time

		beta	p-value
Y	calprotectin	829	2E-05
	Shannon	-0.18	0.5
	dysbiosis	0.49	0.02

Average difference:
Y ~ diagnosis(UC/CD)

		beta	p-value
Y	calprotectin	917	6E-06
	Shannon	-0.25	0.3
	dysbiosis	0.70	7E-04

Table S5: statistical summary of the association between Shannon/dysbiosis and calprotectin/PCDAI, and between PCDAI and calprotectin

ALL CASES AND CONTROLS

		X			
		Shannon		dysbiosis	
		beta	p-value	beta	p-value
Y	Calprotectin mean ± SD = 266 ± 548	-66.1	0.3	260	4E-04
		mean ± SD = 5.28 ± 0.86		mean ± SD = -1.3 ± 0.74	

CASES ONLY

		X			
		Shannon		dysbiosis	
		beta	p-value	beta	p-value
Y	Calprotectin mean ± SD = 366 ± 626	-13.3	0.9	286	3E-04
		mean ± SD = 5.04 ± 0.84		mean ± SD = -1.06 ± 0.66	
	PCDAI mean ± SD = 13.1 ± 12.1	-0.70	0.6	5.37	1E-04
		mean ± SD = 4.97 ± 0.88		mean ± SD = -1.06 ± 0.75	

CASES ONLY

		X	
		PCDAI	
		beta	p-value
Y	Calprotectin mean ± SD = 241 ± 491	11.0	0.06
		mean ± SD = 12.4 ± 11.7	

Table S6: Significant OTUs in case/control and/or responder/nonresponder comparisons. OTUs highlighted in red are "increased" (numerator) components of the dysbiosis index, OTUs in blue are "decreased" (denominator) dysbiosis index components, and OTUs in grey are not represented in the dysbiosis index.

	case/control		nonresponder/responder		
OTU	**β**	**pvalue**	**β**	**pvalue**	**comparisons significant**
k__Bacteria.p__Firmicutes.c__Clostridia.o__Clostridiales.f__Veillonellaceae.g__Veillonella	1.1E-02	2.6E-03	1.6E-02	9.9E-04	both
k__Bacteria.p__Proteobacteria.c__Gammaproteobacteria.o__Enterobacteriales.f__Enterobacteriaceae.g__	3.2E-02	1.2E-02	4.3E-02	7.4E-03	both
k__Bacteria.p__Proteobacteria.c__Gammaproteobacteria.o__Enterobacteriales.f__Enterobacteriaceae.g__Pantoea	3.1E-03	2.2E-02	4.5E-03	1.9E-02	both
k__Bacteria.p__Proteobacteria.c__Gammaproteobacteria.o__Enterobacteriales.f__Enterobacteriaceae.g__Citrobacter	2.3E-04	2.6E-02	3.1E-04	3.7E-02	both
k__Bacteria.p__Proteobacteria.c__Gammaproteobacteria.o__Enterobacteriales.f__Enterobacteriaceae.g__Escherichia	1.1E-03	2.7E-02	1.4E-03	2.6E-02	both
k__Bacteria.p__Proteobacteria.c__Gammaproteobacteria.o__Enterobacteriales.f__Enterobacteriaceae.g__Xenorhabdus	4.1E-05	3.0E-02	5.7E-05	3.1E-02	both
k__Bacteria.p__Fusobacteria.c__Fusobacteriia.o__Fusobacteriales.f__Fusobacteriaceae.g__Fusobacterium	9.0E-03	3.3E-02	1.3E-02	2.3E-02	both
k__Bacteria.p__Firmicutes.c__Clostridia.o__Clostridiales.f__Lachnospiraceae.g__Coprococcus	-2.2E-02	3.3E-06	-1.1E-02	2.5E-03	both
k__Bacteria.p__Firmicutes.c__Bacilli.o__Lactobacillales.f__Streptococcaceae.g__Streptococcus	1.4E-02	3.5E-02	2.0E-02	2.4E-02	both
k__Bacteria.p__Firmicutes.c__Bacilli.o__Lactobacillales.f__.g__	6.1E-05	3.5E-02	9.1E-05	2.2E-02	both
k__Bacteria.p__Firmicutes.c__Bacilli.o__Lactobacillales.f__Enterococcaceae.g__	4.1E-04	4.7E-02	5.9E-04	4.5E-02	both
k__Bacteria.p__Proteobacteria.c__Gammaproteobacteria.o__Pasteurellales.f__Pasteurellaceae.g__Haemophilus	7.4E-03	3.6E-03	4.0E-03	3.9E-01	case/control
k__Bacteria.p__Proteobacteria.c__Gammaproteobacteria.o__Pasteurellales.f__Pasteurellaceae.g__	8.5E-05	1.1E-02	7.0E-05	2.1E-01	case/control
k__Bacteria.p__Proteobacteria.c__Gammaproteobacteria.o__Pasteurellales.f__Pasteurellaceae.g__Actinobacillus	6.6E-05	1.2E-02	5.5E-05	1.8E-01	case/control
k__Bacteria.p__Firmicutes.c__Clostridia.o__Clostridiales.f__Ruminococcaceae.g__Ruminococcus	-2.4E-02	3.2E-02	2.0E-03	5.0E-01	case/control
k__Bacteria.p__Firmicutes.c__Erysipelotrichi.o__Erysipelotrichales.f__Erysipelotrichaceae.g__	5.4E-03	4.3E-02	-7.1E-03	1.5E-01	case/control
k__Bacteria.p__Actinobacteria.c__Coriobacteriia.o__Coriobacteriales.f__Coriobacteriaceae.g__Adlercreutzia	-4.8E-04	1.2E-04	-2.3E-05	7.0E-01	case/control
k__Bacteria.p__Proteobacteria.c__Gammaproteobacteria.o__Pseudomonadales.f__Moraxellaceae.g__Acinetobacter	6.4E-05	1.3E-02	5.2E-06	9.3E-01	case/control
k__Bacteria.p__Actinobacteria.c__Coriobacteriia.o__Coriobacteriales.f__Coriobacteriaceae.g__Slackia	7.5E-05	2.8E-02	-2.7E-05	7.1E-01	case/control
k__Bacteria.p__Proteobacteria.c__Epsilonproteobacteria.o__Campylobacterales.f__Campylobacteraceae.g__Campylobacter	5.4E-04	5.0E-02	7.2E-04	6.6E-02	case/control
k__Bacteria.p__Firmicutes.c__Clostridia.o__Clostridiales.f__Veillonellaceae.g__Dialister	2.7E-02	7.8E-02	-5.3E-02	3.3E-02	nonresponder/responder
k__Bacteria.p__Firmicutes.c__Erysipelotrichi.o__Erysipelotrichales.f__Erysipelotrichaceae.g__Eubacterium	1.7E-03	1.8E-01	3.7E-03	1.5E-02	nonresponder/responder
k__Bacteria.p__Firmicutes.c__Erysipelotrichi.o__Erysipelotrichales.f__Erysipelotrichaceae.g__Coprobacillus	-3.4E-04	1.8E-01	1.6E-04	3.0E-02	nonresponder/responder
k__Bacteria.p__Firmicutes.c__Clostridia.o__Clostridiales.f__Ruminococcaceae.g__Anaerotruncus	-8.1E-05	2.9E-01	1.1E-04	2.5E-02	nonresponder/responder
k__Bacteria.p__Firmicutes.c__Clostridia.o__Clostridiales.f__Lachnospiraceae.g__Ruminococcus	-1.4E-03	6.3E-01	3.0E-03	2.9E-03	nonresponder/responder
k__Bacteria.p__Firmicutes.c__Bacilli.o__Lactobacillales.f__Streptococcaceae.g__Lactococcus	5.5E-05	8.7E-02	1.0E-04	7.0E-03	nonresponder/responder
k__Bacteria.p__Proteobacteria.c__Deltaproteobacteria.o__Desulfovibrionales.f__Desulfovibrionaceae.g__Bilophila	-5.6E-04	7.8E-01	4.1E-03	3.8E-02	nonresponder/responder

Table S7: The WEIGHTED random forest confusion table is presented below.

Confusion table

	Nonresponder	Responder
Nonresponder	11	1
Responder	3	2

Table S8: The random forest EQUAL SAMPLING confusion table is presented below.

Confusion table

	Nonresponder	Responder
Nonresponder	9	3
Responder	1	4

Table S9: The genera shown below had the 15 highest importance scores for classifying patients into treatment responders/non-responders as determined by WEIGHTED random forest.

Taxon	Importance score (mean decrease Gini)
k__Bacteria;p__Firmicutes;c__Clostridia;o__Clostridiales;f__Clostridiaceae;g__SMB53	0.418091806
k__Bacteria;p__Firmicutes;c__Clostridia;o__Clostridiales;f__Peptostreptococcaceae;g__	0.355870067
k__Bacteria;p__Firmicutes;c__Clostridia;o__Clostridiales;f__Peptostreptococcaceae;g__[Clostridium]	0.354063882
k__Bacteria;p__Firmicutes;c__Clostridia;o__Clostridiales;f__Lachnospiraceae;g__Lachnospira	0.219463445
k__Bacteria;p__Firmicutes;c__Clostridia;o__Clostridiales;f__Clostridiaceae;g__	0.216848718
k__Bacteria;p__Firmicutes;c__Bacilli;o__Turicibacterales;f__Turicibacteraceae;g__Turicibacter	0.209579085
k__Bacteria;p__Firmicutes;c__Erysipelotrichi;o__Erysipelotrichales;f__Erysipelotrichaceae;g__Holdemania	0.205763239
k__Bacteria;p__Firmicutes;c__Clostridia;o__Clostridiales;f__Lachnospiraceae;g__Blautia	0.200571385
k__Bacteria;p__Firmicutes;c__Clostridia;o__Clostridiales;f__Lachnospiraceae;g__	0.190545727
k__Bacteria;p__Firmicutes;c__Clostridia;o__Clostridiales;f__Ruminococcaceae;g__Faecalibacterium	0.169648821
k__Bacteria;p__Proteobacteria;c__Gammaproteobacteria;o__Enterobacteriales;f__Enterobacteriaceae;g__	0.141463196
k__Bacteria;p__Actinobacteria;c__Coriobacteriia;o__Coriobacteriales;f__Coriobacteriaceae;g__Adlercreutzia	0.135978227
k__Bacteria;p__Firmicutes;c__Clostridia;o__Clostridiales;f__Lachnospiraceae;g__Coprococcus	0.133961682
k__Bacteria;p__Firmicutes;c__Clostridia;o__Clostridiales;f__Veillonellaceae;g__Dialister	0.120491045
k__Bacteria;p__Verrucomicrobia;c__Verrucomicrobiae;o__Verrucomicrobiales;f__Verrucomicrobiaceae;g__Akkermansia	0.111140619

Table S10: The genera shown below had the 15 highest importance scores for classifying patients into treatment responders/non-responders as determined by random forest with EQUAL SAMPLING. Genera in common with Table S8 are highlighted in red (*Dialister* was found previously in the top 15).

Taxon	Importance score (mean decrease Gini)
k__Bacteria;p__Firmicutes;c__Clostridia;o__Clostridiales;f__Clostridiaceae;g__SMB53	0.251949636
k__Bacteria;p__Firmicutes;c__Clostridia;o__Clostridiales;f__Peptostreptococcaceae;g__	0.187099754
k__Bacteria;p__Firmicutes;c__Erysipelotrichi;o__Erysipelotrichales;f__Erysipelotrichaceae;g__Holdemania	0.185306016
k__Bacteria;p__Firmicutes;c__Clostridia;o__Clostridiales;f__Peptostreptococcaceae;g__[Clostridium]	0.178959984
k__Bacteria;p__Firmicutes;c__Clostridia;o__Clostridiales;f__Lachnospiraceae;g__Blautia	0.176116098
k__Bacteria;p__Firmicutes;c__Clostridia;o__Clostridiales;f__Lachnospiraceae;g__Lachnospira	0.163797353
k__Bacteria;p__Firmicutes;c__Clostridia;o__Clostridiales;f__Lachnospiraceae;g__	0.156547738
k__Bacteria;p__Firmicutes;c__Clostridia;o__Clostridiales;f__Ruminococcaceae;g__Faecalibacterium	0.150600262
k__Bacteria;p__Firmicutes;c__Clostridia;o__Clostridiales;f__Clostridiaceae;g__	0.130824767
k__Bacteria;p__Firmicutes;c__Bacilli;o__Turicibacterales;f__Turicibacteraceae;g__Turicibacter	0.128628188
k__Bacteria;p__Verrucomicrobia;c__Verrucomicrobiae;o__Verrucomicrobiales;f__Verrucomicrobiaceae;g__Akkermansia	0.109253344
k__Bacteria;p__Proteobacteria;c__Gammaproteobacteria;o__Enterobacteriales;f__Enterobacteriaceae;g__	0.094807255
k__Bacteria;p__Bacteroidetes;c__Bacteroidia;o__Bacteroidales;f__Rikenellaceae;g__	0.090125252
k__Bacteria;p__Firmicutes;c__Clostridia;o__Clostridiales;f__Lachnospiraceae;g__Coprococcus	0.090089444
k__Bacteria;p__Actinobacteria;c__Coriobacteriia;o__Coriobacteriales;f__Coriobacteriaceae;g__Adlercreutzia	0.088743403

SUPPLEMENTARY FIGURES

Figure S1: All time points for calprotectin (panel A), Shannon alpha diversity (panel B), and gut microbial dysbiosis (panel C) for unaffected controls (black circles), Crohn's disease patients (CD, red circles), and ulcerative colitis patients (UC, blue circles) are shown. Overall, CD and UC patients have increased calprotectin, decreased alpha diversity, and increased gut microbial dysbiosis compared to controls.

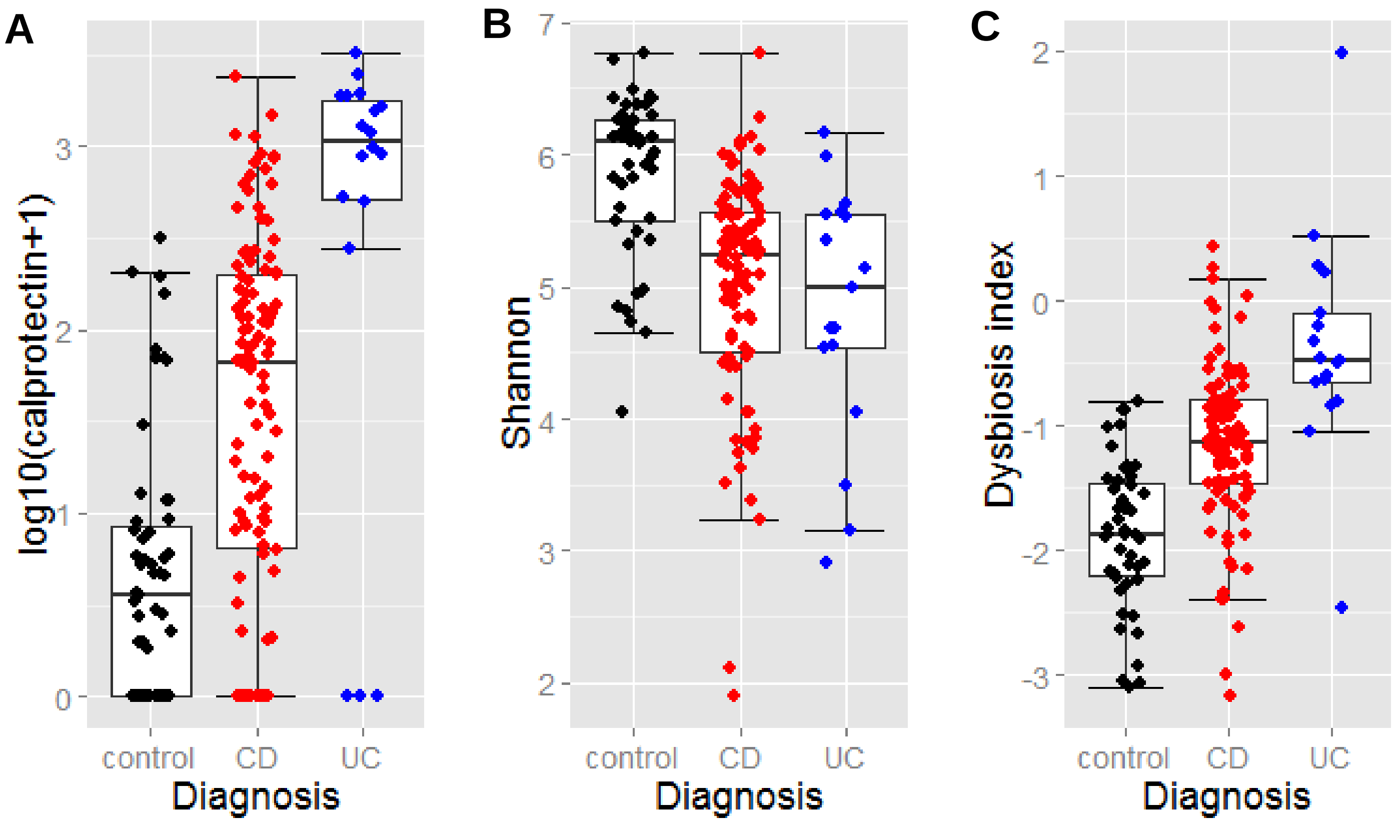


Figure S2: All control, responder, and non-responder calprotectin and microbiome time points further identified by individual.

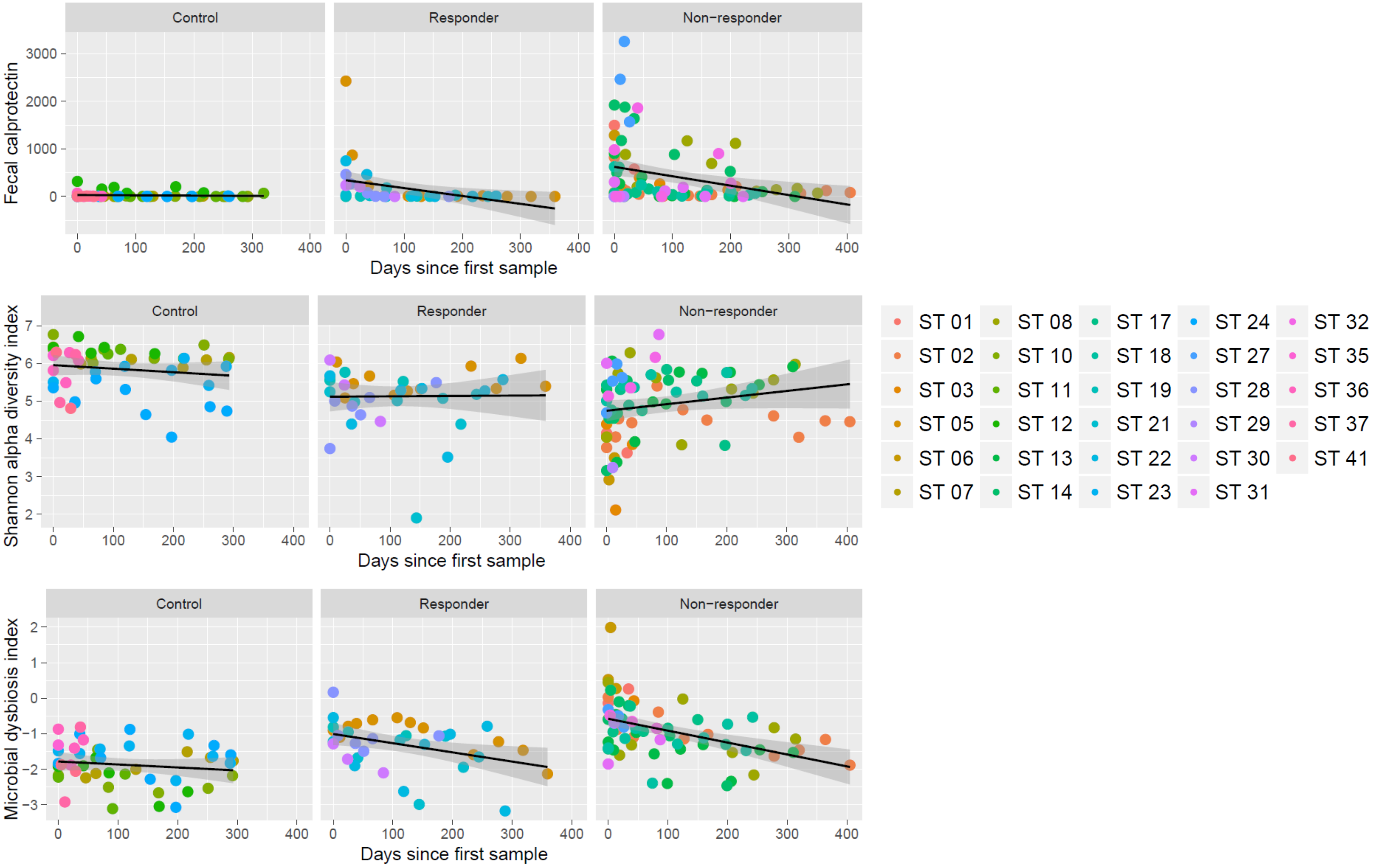


Figure S3: The relationships between relevant clinical measures for patients with IBD are shown. Regression lines, plotted in black, are adjusted for correlations within individuals. Dysbiosis and calprotectin are shown in panel A; panel B shows the relationship between dysbiosis index and PCDAI. Increased dysbiosis associates with increased calprotectin and higher PCDAI. Panel C shows the relationship between PCDAI and calprotectin. PCDAI does not significantly associate with increased calprotectin.

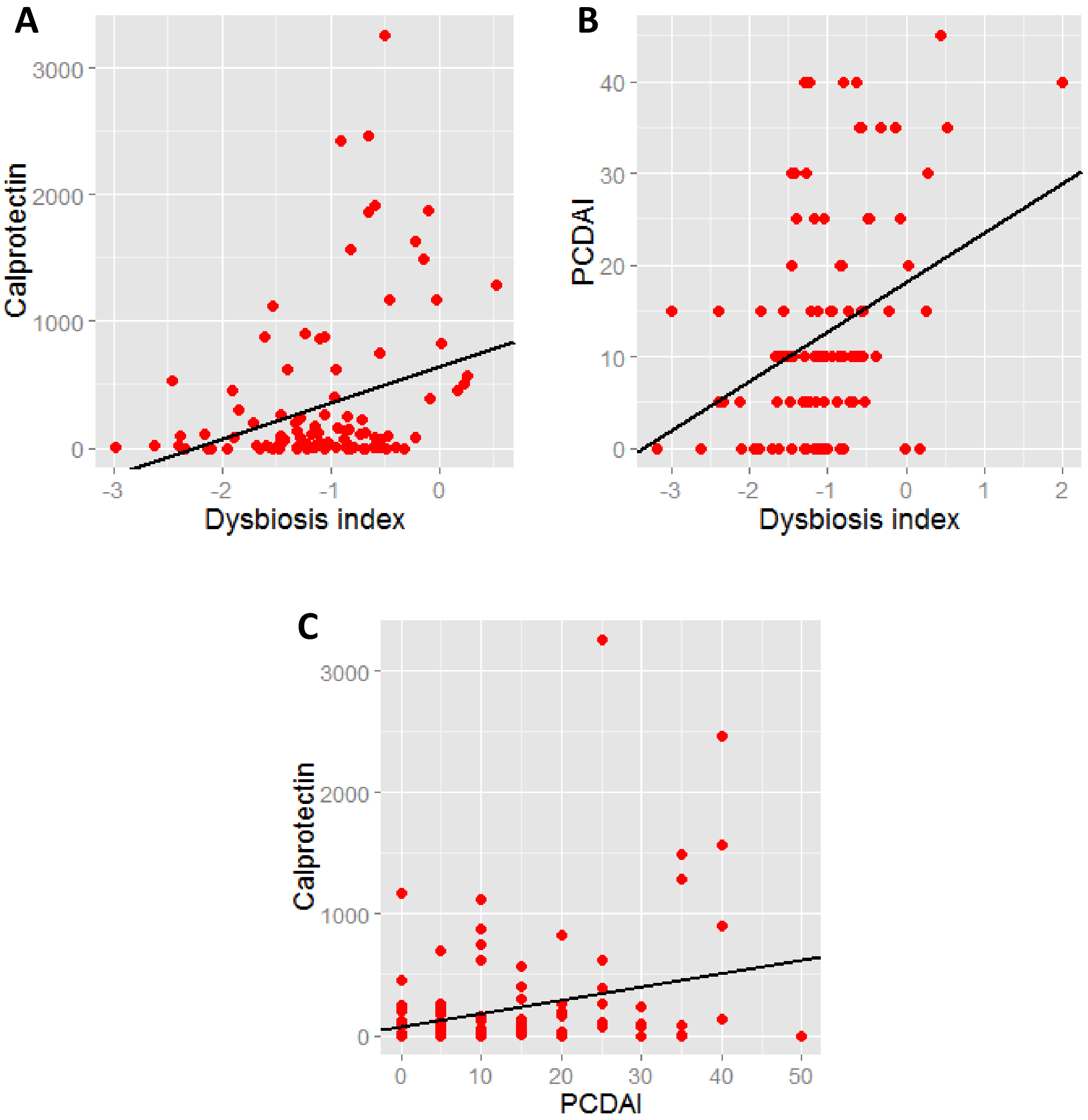


Figure S4: The precision (positive predictive value)/recall (sensitivity) curve for our weighted random forest model is shown below.

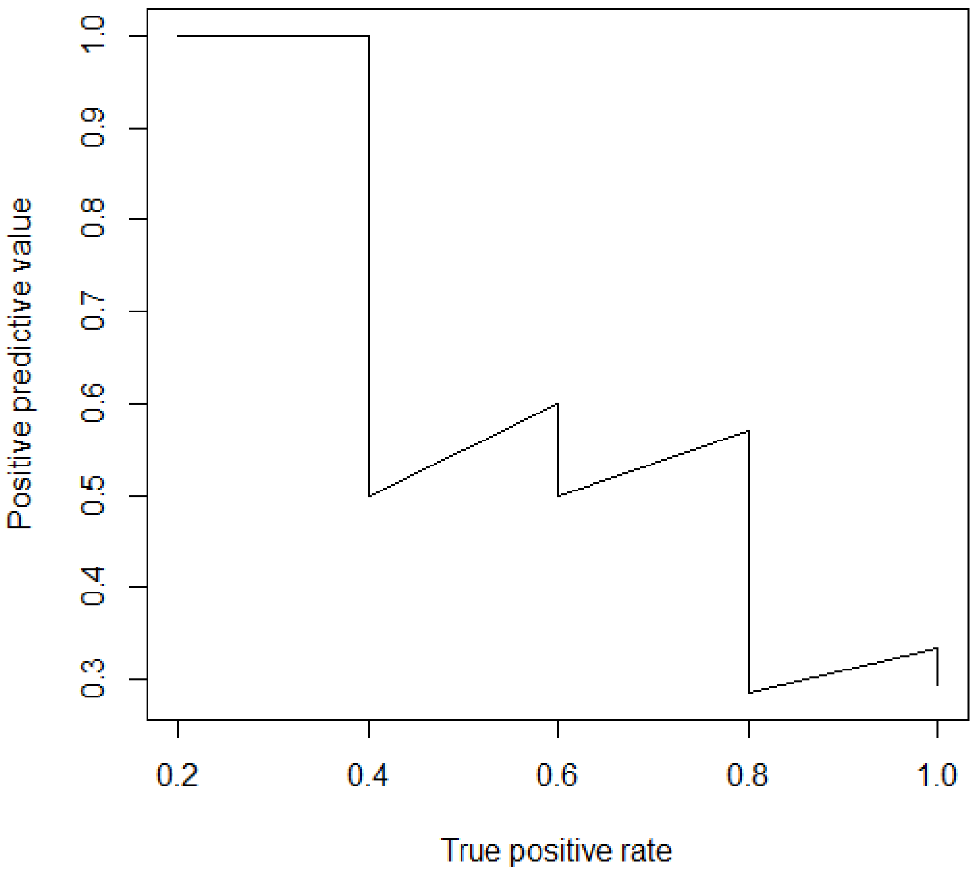


Figure S5: The receiver operating characteristic curve (ROC, panel A) and precision/recall curve (panel B) for our random forest analysis with equal sampling are shown below.

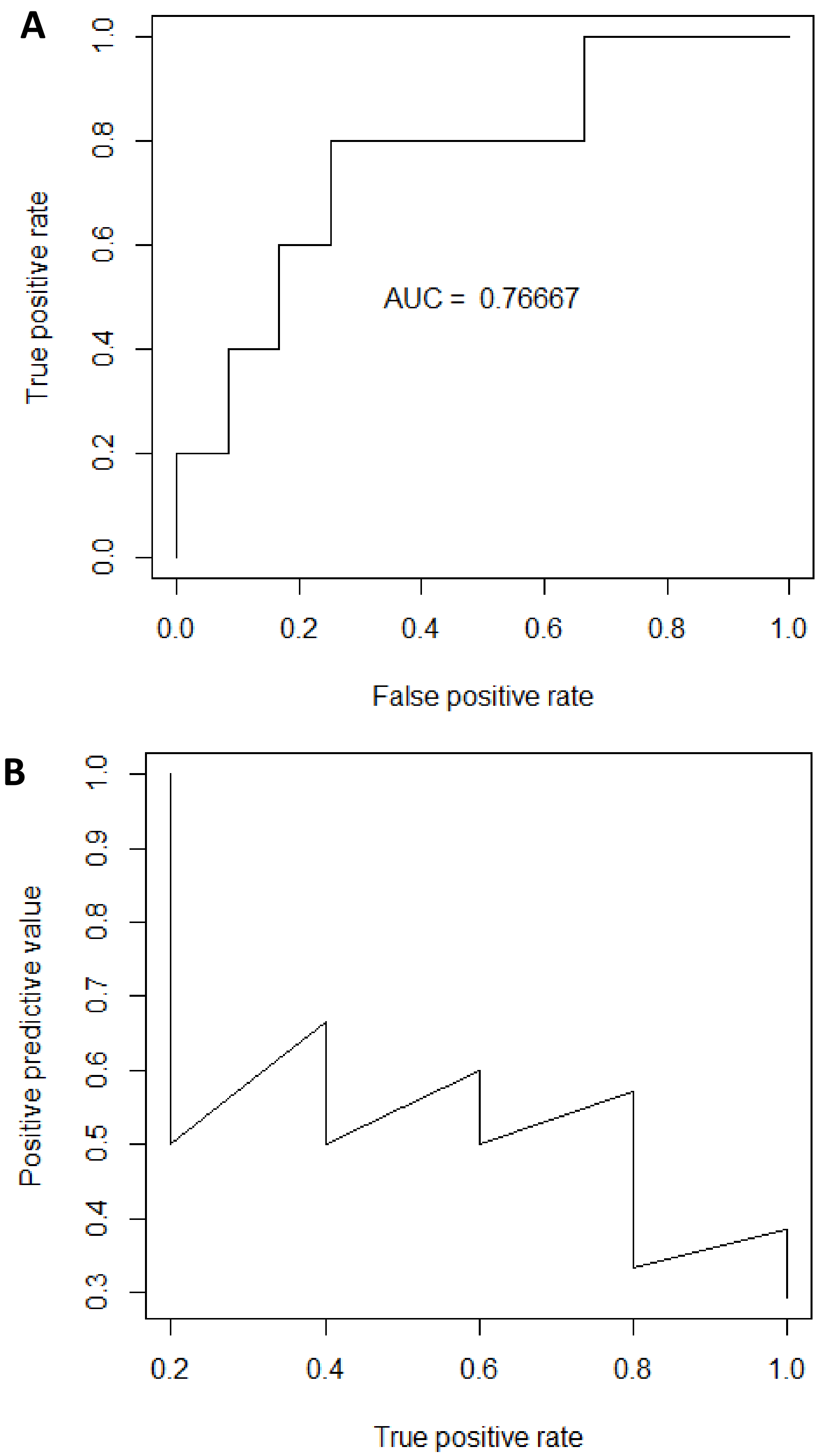
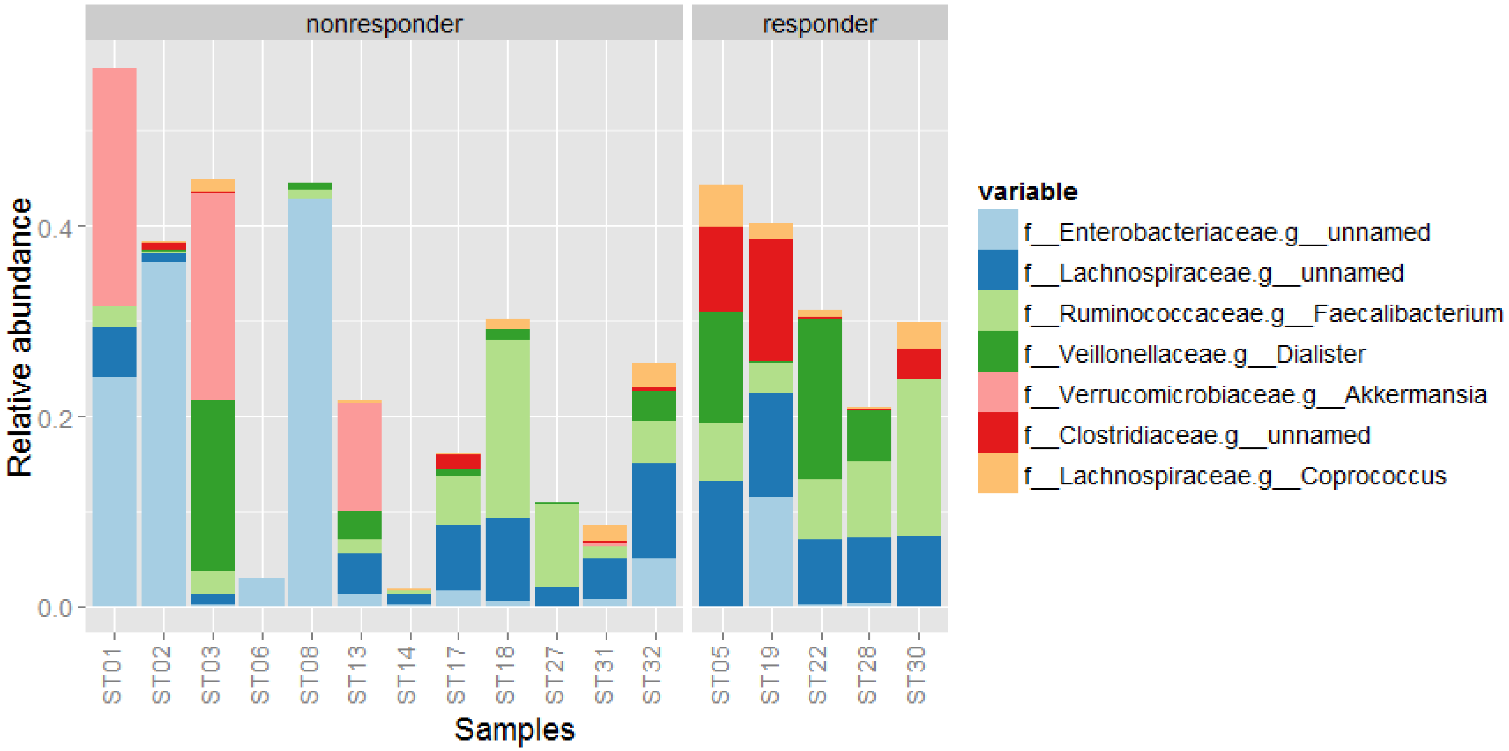


Figure S6: For those of the top 15 genera that were found above 1% average relative abundance, stacked bar charts are shown for each sample used in the random forest (categorized by response or non-response).



SUPPLEMENTARY FILES

Supplementary file 1: reads_microbiome_info.xlsx

New Emory ID	total reads	joined reads	alignment fail	ambiguous base	low quality	ok	pct ok
ST01.00	97396	42339	17	0	17	42305	99.92
ST01.03	105948	46074	16	0	25	46033	99.91
ST02.00	116100	50812	44	0	24	50744	99.87
ST02.01	109146	48568	38	0	12	48518	99.90
ST02.02	97052	42887	10	0	23	42854	99.92
ST02.03	114732	50692	28	0	26	50638	99.89
ST02.04	340028	152899	122	63	324	152390	99.67
ST02.05	155426	69325	22	0	41	69262	99.91
ST02.07	119416	52027	13	0	34	51980	99.91
ST02.10	90842	37660	25	0	33	37602	99.85
ST02.11	103020	43010	14	0	38	42958	99.88
ST02.12	66156	28019	26	0	43	27950	99.75
ST02.13	127598	54905	21	0	24	54860	99.92
ST03.00	175446	75028	68	0	103	74857	99.77
ST03.01	79820	34694	117	0	45	34532	99.53
ST03.02	159002	69498	36	0	20	69442	99.92
ST05.00	233794	103898	16	0	39	103843	99.95
ST05.01	83838	37696	13	0	20	37663	99.91
ST05.02	126504	59007	23	11	76	58897	99.81
ST05.03	157544	72117	55	21	116	71925	99.73
ST05.04	216326	102216	58	21	135	102002	99.79
ST05.05	129150	57117	42	0	35	57040	99.87
ST05.06	170628	76356	67	0	30	76259	99.87
ST05.07	138500	61282	99	0	56	61127	99.75
ST05.09	186022	86858	40	23	139	86656	99.77
ST05.10	128174	57298	63	18	171	57046	99.56
ST05.11	92026	39829	18	0	15	39796	99.92
ST05.12	232690	105237	130	40	294	104773	99.56
ST06.00	126732	52926	12	0	42	52872	99.90
ST06.01	5.85E+04	25153	18	0	41	25094	99.77
ST06.03	106384	47874	53	20	84	47717	99.67
ST07.01	138814	65642	33	23	97	65489	99.77
ST07.02	237562	110024	100	34	266	109624	99.64
ST07.03	2.90E+04	13043	7	0	10	13026	99.87
ST07.04	213950	93723	87	0	69	93567	99.83
ST07.05	218250	93562	36	0	60	93466	99.90
ST07.06	121544	56482	46	16	87	56333	99.74
ST07.07	161918	67632	42	0	57	67533	99.85
ST08.00	33184	13357	29	4	37	13287	99.48
ST08.01	244156	109335	117	66	373	108779	99.49
ST08.02	265254	125316	59	50	177	125030	99.77
ST08.03	188384	87830	50	33	127	87620	99.76
ST08.04	97082	42190	49	0	18	42123	99.84
ST08.06	173716	75188	82	0	51	75055	99.82
ST08.07	148696	63771	6	0	42	63723	99.92
ST08.08	144552	61106	54	0	58	60994	99.82
ST08.09	265138	118369	29	0	54	118286	99.93
ST10.01	10652	4741	8	3	11	4719	99.54
ST10.03	38524	14732	48	9	37	14638	99.36
ST10.04	234088	99367	94	0	61	99212	99.84

ST10.06	132838	54870	17	0	46	54807	99.89
ST10.08	206684	86612	32	0	72	86508	99.88
ST10.09	207790	89355	48	0	79	89228	99.86
ST11.00	166374	73305	90	33	237	72945	99.51
ST11.02	286536	131384	77	54	282	130971	99.69
ST11.03	200370	85340	108	0	57	85175	99.81
ST12.00	262350	123341	65	45	215	123016	99.74
ST12.01	184700	81699	132	46	296	81225	99.42
ST12.02	269382	119376	99	0	71	119206	99.86
ST12.03	268572	116905	128	0	97	116680	99.81
ST12.05	207838	86552	54	0	73	86425	99.85
ST12.06	89688	36901	19	0	21	36861	99.89
ST13.00	60610	25006	72	9	76	24849	99.37
ST13.01	102058	45072	50	11	119	44892	99.60
ST13.02	126748	54170	143	26	144	53857	99.42
ST13.04	199566	86944	102	38	229	86575	99.58
ST13.05	153422	68340	58	0	38	68244	99.86
ST13.06	9.55E+04	42601	51	0	28	42522	99.81
ST13.07	76248	33697	16	0	18	33663	99.90
ST13.08	182774	80965	49	0	55	80861	99.87
ST13.09	1.63E+05	70337	17	0	37	70283	99.92
ST13.10	132622	55206	37	0	29	55140	99.88
ST13.11	171412	73220	28	0	41	73151	99.91
ST14.00	40468	18716	20	5	29	18662	99.71
ST14.01	154646	70373	39	28	161	70145	99.68
ST14.02	3223248	1451511	966	509	2381	1447655	99.73
ST14.03	189242	82338	87	33	217	82001	99.59
ST14.04	209560	92168	133	47	254	91734	99.53
ST14.06	127370	55662	15	0	23	55624	99.93
ST14.09	172956	74755	19	0	42	74694	99.92
ST17.00	88196	40369	43	15	60	40251	99.71
ST17.01	314670	141582	182	75	399	140926	99.54
ST17.02	39674	15753	29	3	48	15673	99.49
ST17.03	172264	79798	89	23	202	79484	99.61
ST17.04	166040	78317	39	23	64	78191	99.84
ST17.05	218126	99692	29	0	34	99629	99.94
ST17.06	156614	67947	26	0	25	67896	99.92
ST17.07	111190	46871	15	0	19	46837	99.93
ST18.01	93272	44073	23	13	60	43977	99.78
ST18.02	59180	25487	14	0	24	25449	99.85
ST18.03	24622	10179	23	1	17	10138	99.60
ST18.04	48202	21333	11	0	5	21317	99.92
ST18.05	182206	80473	28	0	23	80422	99.94
ST18.06	91510	40415	7	0	18	40390	99.94
ST18.07	225908	95699	81	0	56	95562	99.86
ST18.08	131258	55843	22	0	41	55780	99.89
ST18.09	107112	48910	21	15	77	48797	99.77
ST19.00	145666	63803	104	36	205	63458	99.46
ST19.01	103476	48205	26	18	62	48099	99.78
ST19.03	1.67E+05	73713	70	0	30	73613	99.86
ST21.01	179408	76675	78	0	58	76539	99.82
ST21.02	181002	79191	109	0	71	79011	99.77
ST21.04	173736	77174	83	0	41	77050	99.84
ST21.05	34744	15031	31	0	32	14968	99.58
ST21.06	118476	49835	18	0	31	49786	99.90
ST21.07	172040	75543	102	40	357	75044	99.34
ST21.08	70664	31172	12	0	14	31146	99.92

ST22.00	102644	44553	41	14	126	44372	99.59
ST22.01	74642	32287	20	0	14	32253	99.89
ST22.03	130244	58587	17	0	23	58547	99.93
ST22.04	169050	68904	45	0	47	68812	99.87
ST22.05	180254	79818	117	48	222	79431	99.52
ST22.08	130876	55552	23	0	37	55492	99.89
ST22.09	2.03E+05	86686	87	0	58	86541	99.83
ST23.00	348350	153079	333	78	616	152052	99.33
ST23.01	205250	93293	147	42	272	92832	99.51
ST23.02	397640	178471	67	0	107	178297	99.90
ST23.03	2.21E+05	97438	46	0	68	97324	99.88
ST23.05	110106	47800	18	0	19	47763	99.92
ST23.06	67692	28668	11	0	25	28632	99.87
ST23.07	3.70E+05	158834	56	0	102	158676	99.90
ST23.08	158128	69720	28	0	24	69668	99.93
ST24.00	223786	102413	77	41	209	102086	99.68
ST24.01	1.83E+05	82153	26	0	36	82091	99.92
ST24.02	2.01E+05	90485	39	0	43	90403	99.91
ST24.03	171384	77392	31	0	32	77329	99.92
ST24.04	181668	75053	75	0	105	74873	99.76
ST24.05	7340	3265	0	0	1	3264	99.97
ST24.07	128232	54914	20	0	22	54872	99.92
ST24.08	1.39E+05	58644	41	0	36	58567	99.87
ST27.00	124786	55034	90	31	241	54672	99.34
ST27.01	178290	81728	64	28	205	81431	99.64
ST27.02	268516	114790	144	0	108	114538	99.78
ST27.03	250632	110689	98	0	90	110501	99.83
ST28.01	8.05E+04	35342	16	0	16	35310	99.91
ST28.02	186596	82378	13	0	37	82328	99.94
ST28.03	95628	43358	13	0	19	43326	99.93
ST28.04	72898	33026	17	0	13	32996	99.91
ST28.05	102358	45193	13	0	27	45153	99.91
ST28.06	153900	71620	33	22	108	71457	99.77
ST29.01	1.31E+05	58039	21	0	35	57983	99.90
ST30.00	256518	112990	74	0	68	112848	99.87
ST30.01	147908	63929	30	0	35	63864	99.90
ST30.02	126752	57945	27	21	95	57802	99.75
ST31.00	157716	68387	27	0	35	68325	99.91
ST31.02	205714	90880	25	0	32	90823	99.94
ST32.01	271532	117141	69	0	95	116977	99.86
ST32.02	163644	71005	305	0	75	70625	99.46
ST32.03	116812	50720	23	0	34	50663	99.89
ST35.00	1.26E+05	55344	22	0	18	55304	99.93
ST36.00	121122	53344	18	0	27	53299	99.92
ST36.01	121128	52138	18	0	34	52086	99.90
ST36.02	199530	88492	25	0	42	88425	99.92
ST36.03	198878	85363	31	0	55	85277	99.90
ST37.01	29292	12561	6	0	14	12541	99.84
ST37.03	8954	3720	15	0	17	3688	99.14
ST41.01	164450	70080	47	0	51	69982	99.86
ST41.03	126036	53245	32	0	53	53160	99.84

New Emory ID	closed sequences classified	closed pct classified	closed shannon	closed dysbiosis index	median of 10000 subsampled closed shannon	median of 10000 subsampled closed dysbiosis
ST01.00	39143	92.53	4.15	-0.14	3.04	-0.16
ST01.03	44114	95.83	3.63	0.26	2.59	0.25
ST02.00	47024	92.67	3.77	0.03	2.74	0.03
ST02.01	43474	89.60	4.05	-0.88	2.91	-0.88
ST02.02	40866	95.36	4.53	-0.67	3.20	-0.67
ST02.03	46853	92.53	4.43	-1.10	3.12	-0.63
ST02.04	134399	88.19	5.40	-0.39	3.89	-0.39
ST02.05	64708	93.42	4.77	-1.14	3.41	-1.15
ST02.07	49560	95.34	4.50	-1.02	3.17	-1.02
ST02.10	33906	90.17	4.61	-1.63	3.30	-1.63
ST02.11	40172	93.51	4.04	-1.46	2.87	-1.46
ST02.12	25668	91.84	4.48	-1.16	3.15	-1.16
ST02.13	51254	93.43	4.46	-1.88	3.16	-1.88
ST03.00	69542	92.90	4.39	-0.58	3.17	-0.60
ST03.01	30912	89.52	2.11	-0.47	1.82	-0.54
ST03.02	65243	93.95	3.85	-0.07	2.82	-0.08
ST05.00	99323	95.65	5.61	-0.90	4.00	-0.92
ST05.01	35464	94.16	6.04	-1.10	4.29	-1.11
ST05.02	54552	92.62	5.09	-0.79	3.60	-0.80
ST05.03	63321	88.04	5.46	-0.71	3.85	-0.72
ST05.04	94760	92.90	5.67	-0.61	4.00	-0.62
ST05.05	51649	90.55	5.16	-0.55	3.72	-0.57
ST05.06	70481	92.42	5.27	-0.68	3.72	-0.70
ST05.07	54061	88.44	5.32	-0.84	3.76	-0.85
ST05.09	76768	88.59	5.93	-1.59	4.19	-1.60
ST05.10	48857	85.64	5.33	-1.22	3.77	-1.23
ST05.11	36582	91.92	6.14	-1.47	4.43	-1.49
ST05.12	90130	86.02	5.40	-2.13	3.82	-2.14
ST06.00	49638	93.88	4.05	0.52	3.04	0.49
ST06.01	23338	93.00	2.91	1.99	2.40	1.94
ST06.03	43130	90.39	3.50	0.27	2.75	0.14
ST07.01	53579	81.81	6.44	-1.88	4.59	-1.90
ST07.02	90241	82.32	5.99	-2.24	4.24	-2.26
ST07.03	11146	85.57	6.14	-2.12	4.43	-2.15
ST07.04	80826	86.38	6.11	-2.00	4.35	-2.02
ST07.05	75299	80.56	5.89	-1.51	4.27	-1.54
ST07.06	43961	78.04	6.10	-1.67	4.38	-1.70
ST07.07	54957	81.38	6.16	-1.76	4.39	-1.78
ST08.00	10235	77.03	4.04	0.44	2.94	0.43
ST08.01	94280	86.67	5.74	-1.60	4.08	-1.62
ST08.02	114244	91.37	6.29	-1.32	4.42	-1.33
ST08.03	81902	93.47	5.62	-0.85	3.95	-0.86
ST08.04	38701	91.88	3.84	-0.02	2.73	-0.02
ST08.06	67660	90.15	5.32	-1.53	3.75	-1.53
ST08.07	57801	90.71	5.21	-2.16	3.68	-2.18
ST08.08	53716	88.07	5.56	-0.83	3.92	-0.84
ST08.09	111245	94.05	5.98	-1.15	4.22	-1.16
ST10.01	4226	89.55	6.77	-2.17	4.97	-2.20
ST10.03	11219	76.64	6.02	-1.45	4.35	-1.48
ST10.04	86061	86.74	6.38	-2.50	4.52	-2.53
ST10.06	49146	89.67	6.13	-2.66	4.39	-2.69
ST10.08	77699	89.82	6.49	-2.53	4.60	-2.57

ST10.09	80098	89.77	6.12	-2.18	4.34	-2.20
ST11.00	59323	81.33	6.37	-2.22	4.54	-2.24
ST11.02	109165	83.35	6.26	-3.11	4.46	-3.17
ST11.03	73157	85.89	6.38	-2.13	4.55	-2.15
ST12.00	109730	89.20	6.43	-1.92	4.54	-1.93
ST12.01	67343	82.91	6.72	-1.90	4.76	-1.91
ST12.02	108109	90.69	6.28	-1.68	4.42	-1.69
ST12.03	101818	87.26	6.43	-2.10	4.55	-2.12
ST12.05	74418	86.11	6.26	-3.04	4.45	-3.16
ST12.06	32554	88.32	6.13	-2.63	4.39	-2.63
ST13.00	18932	76.19	5.31	-1.23	3.79	-1.25
ST13.01	38861	86.57	6.06	-1.46	4.32	-1.48
ST13.02	47487	88.17	3.38	-0.56	2.50	-0.58
ST13.04	74741	86.33	3.92	-0.96	2.83	-0.97
ST13.05	62303	91.29	4.98	-1.57	3.55	-1.57
ST13.06	35946	84.54	4.93	-2.40	3.47	-2.41
ST13.07	29993	89.10	5.77	-1.43	4.08	-1.44
ST13.08	75446	93.30	5.74	-1.30	4.03	-1.31
ST13.09	64551	91.84	5.77	-2.34	4.27	-2.38
ST13.10	48494	87.95	5.44	-1.45	3.85	-1.47
ST13.11	66866	91.41	5.92	-1.53	4.20	-1.54
ST14.00	17149	91.89	3.16	-0.60	2.47	-0.48
ST14.01	63898	91.09	4.54	0.22	3.27	0.23
ST14.02	1325059	91.53	4.55	-0.46	3.42	-0.52
ST14.03	71984	87.78	4.68	-0.10	3.41	-0.35
ST14.04	78522	85.60	5.54	-0.21	3.95	-0.22
ST14.06	49484	88.96	5.56	-1.06	3.93	-1.15
ST14.09	69075	92.48	4.99	-2.46	3.58	-2.47
ST17.00	37325	92.73	5.43	-1.43	3.86	-1.44
ST17.01	123606	87.71	4.76	-0.95	3.51	-0.99
ST17.02	11569	73.81	5.31	-0.58	3.87	-0.60
ST17.03	70856	89.14	4.89	-0.22	3.49	-0.23
ST17.04	72623	92.88	4.75	-0.93	3.37	-0.94
ST17.05	93079	93.43	5.84	-0.85	4.15	-0.86
ST17.06	63543	93.59	5.54	-0.60	3.96	-0.62
ST17.07	44758	95.56	3.83	-1.30	2.83	-1.32
ST18.01	39913	90.76	5.01	-1.40	3.57	-1.42
ST18.02	24344	95.66	5.42	-1.13	3.83	-1.13
ST18.03	8491	83.75	5.37	-0.85	3.85	-0.86
ST18.04	19566	91.79	5.70	-2.39	4.16	-2.42
ST18.05	75797	94.25	5.24	-1.29	3.70	-1.29
ST18.06	38164	94.49	5.13	-1.11	3.64	-1.12
ST18.07	85996	89.99	5.76	-0.73	4.09	-0.74
ST18.08	51150	91.70	5.15	-1.47	3.70	-1.49
ST18.09	44051	90.27	5.33	-0.53	3.81	-0.55
ST19.00	56625	89.23	5.55	-0.81	3.94	-0.82
ST19.01	41882	87.07	5.77	-0.95	4.15	-0.96
ST19.03	64255	87.29	5.52	-1.05	3.96	-0.99
ST21.01	70543	92.17	5.25	-1.22	3.70	-1.23
ST21.02	67998	86.06	4.98	-1.68	3.54	-1.69
ST21.04	70947	92.08	5.01	-1.17	3.51	-1.18
ST21.05	13392	89.47	1.90	-2.99	2.33	-2.97
ST21.06	44247	88.87	5.07	-1.07	3.59	-1.08
ST21.07	64950	86.55	4.39	-1.94	3.09	-1.95
ST21.08	28177	90.47	5.18	-1.65	3.70	-1.66
ST22.00	38326	86.37	5.67	-0.55	4.02	-0.55
ST22.01	29102	90.23	4.39	-1.90	3.08	-1.90

ST22.03	54176	92.53	5.29	-2.62	3.88	-2.65
ST22.04	58905	85.60	5.34	-1.30	3.75	-1.30
ST22.05	70026	88.16	3.51	-1.01	2.62	-1.02
ST22.08	49815	89.77	5.27	-0.79	3.75	-0.74
ST22.09	78658	90.89	5.57	-3.17	4.06	-3.18
ST23.00	133939	88.09	5.51	-1.48	3.90	-1.49
ST23.01	81945	88.27	4.95	-1.56	3.49	-1.57
ST23.02	166395	93.32	5.78	-1.43	4.17	-1.45
ST23.03	90495	92.98	5.93	-1.34	4.25	-1.36
ST23.05	44173	92.48	5.82	-2.32	4.30	-2.35
ST23.06	25313	88.41	6.14	-1.01	4.52	-1.05
ST23.07	146015	92.02	5.42	-1.63	3.83	-1.65
ST23.08	65716	94.33	5.93	-1.83	4.24	-1.84
ST24.00	91472	89.60	5.36	-1.83	3.77	-1.84
ST24.01	77381	94.26	4.98	-1.01	3.57	-1.02
ST24.02	84727	93.72	5.59	-1.68	4.02	-1.70
ST24.03	74081	95.80	5.31	-0.88	3.81	-0.90
ST24.04	64505	86.15	4.64	-2.28	3.28	-2.28
ST24.05	3155	96.66	4.05	-3.07	2.97	-3.09
ST24.07	50866	92.70	4.85	-1.33	3.47	-1.35
ST24.08	54579	93.19	4.74	-1.59	3.40	-1.61
ST27.00	48852	89.35	4.69	-0.32	3.33	-0.32
ST27.01	73887	90.74	5.53	-0.64	3.92	-0.66
ST27.02	103298	90.19	5.99	-0.50	4.24	-0.51
ST27.03	101276	91.65	5.63	-0.81	3.95	-0.82
ST28.01	33666	95.34	3.74	0.17	2.43	0.13
ST28.02	77832	94.54	5.00	-1.05	3.51	-1.05
ST28.03	41908	96.73	4.87	-1.27	3.47	-1.28
ST28.04	30944	93.78	4.64	-1.49	3.28	-1.50
ST28.05	42674	94.51	5.10	-1.14	3.62	-1.15
ST28.06	65789	92.07	5.49	-1.06	3.93	-1.07
ST29.01	54224	93.52	3.23	-0.70	2.77	-0.72
ST30.00	104018	92.18	6.09	-1.28	4.34	-1.29
ST30.01	58343	91.36	5.43	-1.72	3.84	-1.71
ST30.02	53632	92.79	4.46	-2.10	3.13	-2.09
ST31.00	62720	91.80	6.01	-1.85	4.34	-1.89
ST31.02	82199	90.50	6.77	-1.17	4.73	-1.24
ST32.01	106170	90.76	5.13	-0.48	3.73	-0.49
ST32.02	57115	80.87	5.35	-0.65	3.79	-0.69
ST32.03	44144	87.13	6.16	-0.85	4.41	-0.86
ST35.00	49708	89.88	6.21	-1.32	4.50	-1.34
ST36.00	47851	89.78	5.82	-0.87	4.10	-0.93
ST36.01	47300	90.81	5.49	-1.88	3.86	-1.89
ST36.02	81994	92.73	6.29	-1.41	4.41	-1.41
ST36.03	77617	91.02	6.08	-1.17	4.26	-1.18
ST37.01	11380	90.74	4.96	-2.92	3.60	-2.79
ST37.03	3221	87.34	6.24	-0.81	4.49	-1.06
ST41.01	62249	88.95	6.30	-1.87	4.64	-1.92
ST41.03	47094	88.59	4.81	-2.05	3.52	-2.14

New	denovo	denovo		denovo
Emory	sequences	pct	denovo	dysbiosis
ID	classified	classified	shannon	index
ST01.00	38775	91.66	4.00	-0.07
ST01.03	44594	96.87	3.29	0.26
ST02.00	48593	95.76	3.95	0.03
ST02.01	45677	94.14	4.73	-0.87
ST02.02	41795	97.53	4.47	-0.66
ST02.03	48526	95.83	4.58	-0.62
ST02.04	133555	87.64	6.21	-0.28
ST02.05	66374	95.83	4.98	-1.14
ST02.07	49356	94.95	4.43	-1.00
ST02.10	33423	88.89	4.88	-1.61
ST02.11	38875	90.50	4.03	-1.43
ST02.12	23431	83.83	5.04	-1.11
ST02.13	51369	93.64	4.49	-1.87
ST03.00	72721	97.15	5.53	-0.50
ST03.01	33419	96.78	3.14	-0.47
ST03.02	66660	95.99	3.89	-0.01
ST05.00	96632	93.06	5.34	-0.86
ST05.01	36416	96.69	5.90	-1.10
ST05.02	53792	91.33	4.88	-0.74
ST05.03	64570	89.77	5.68	-0.67
ST05.04	92459	90.64	5.61	-0.51
ST05.05	49985	87.63	5.14	-0.44
ST05.06	69239	90.79	5.22	-0.61
ST05.07	54631	89.37	5.62	-0.81
ST05.09	77845	89.83	6.29	-1.61
ST05.10	51090	89.56	6.01	-1.24
ST05.11	34492	86.67	5.40	-1.47
ST05.12	96452	92.06	6.34	-2.13
ST06.00	51544	97.49	4.54	0.61
ST06.01	24469	97.51	4.13	1.74
ST06.03	45532	95.42	4.58	0.65
ST07.01	53617	81.87	6.97	-1.88
ST07.02	91903	83.83	6.92	-2.23
ST07.03	10451	80.23	5.94	-2.09
ST07.04	82501	88.17	6.56	-2.00
ST07.05	74308	79.50	6.42	-1.50
ST07.06	43782	77.72	6.38	-1.66
ST07.07	56542	83.72	6.81	-1.76
ST08.00	12725	95.77	6.05	0.53
ST08.01	99586	91.55	6.88	-1.59
ST08.02	114433	91.52	6.54	-1.31
ST08.03	82705	94.39	5.70	-0.82
ST08.04	40407	95.93	4.32	0.02
ST08.06	70254	93.60	6.03	-1.48
ST08.07	55669	87.36	5.00	-2.13
ST08.08	53838	88.27	5.61	-0.78
ST08.09	108013	91.32	5.65	-1.14
ST10.01	3952	83.75	6.28	-2.12
ST10.03	13638	93.17	7.65	-1.50
ST10.04	87387	88.08	6.74	-2.46
ST10.06	47152	86.03	5.94	-2.62
ST10.08	77678	89.79	6.48	-2.53
ST10.09	79869	89.51	6.28	-2.17
ST11.00	61289	84.02	7.45	-2.16

ST11.02	113325	86.53	6.96	-2.92
ST11.03	72018	84.55	6.60	-2.11
ST12.00	114029	92.69	7.02	-1.92
ST12.01	69238	85.24	7.49	-1.87
ST12.02	109764	92.08	6.42	-1.68
ST12.03	104986	89.98	7.05	-2.09
ST12.05	75463	87.32	6.65	-3.02
ST12.06	31184	84.60	5.99	-2.60
ST13.00	23410	94.21	7.74	-1.25
ST13.01	37390	83.29	6.45	-1.45
ST13.02	49792	92.45	4.56	-0.47
ST13.04	77563	89.59	5.96	-0.94
ST13.05	65474	95.94	5.59	-1.58
ST13.06	35772	84.13	5.09	-2.41
ST13.07	28216	83.82	5.74	-1.40
ST13.08	76855	95.05	5.86	-1.30
ST13.09	60261	85.74	5.03	-2.32
ST13.10	47483	86.11	5.82	-1.41
ST13.11	67736	92.60	6.63	-1.51
ST14.00	17059	91.41	4.71	-0.54
ST14.01	64210	91.54	4.81	0.71
ST14.02	1347919	93.11	5.31	-0.46
ST14.03	76195	92.92	5.76	-0.29
ST14.04	83037	90.52	6.34	-0.11
ST14.06	46708	83.97	5.24	-1.13
ST14.09	69491	93.03	5.38	-2.42
ST17.00	38252	95.03	6.32	-1.43
ST17.01	128378	91.10	6.12	-0.95
ST17.02	15005	95.74	6.96	-0.60
ST17.03	74625	93.89	5.52	-0.17
ST17.04	73490	93.99	4.41	-0.92
ST17.05	94118	94.47	5.71	-0.83
ST17.06	61789	91.01	4.93	-0.58
ST17.07	44548	95.11	3.49	-1.30
ST18.01	40784	92.74	4.79	-1.42
ST18.02	25090	98.59	4.99	-1.12
ST18.03	9840	97.06	6.01	-0.86
ST18.04	19135	89.76	4.98	-2.37
ST18.05	75867	94.34	4.65	-1.28
ST18.06	37898	93.83	4.40	-1.10
ST18.07	86517	90.53	5.83	-0.67
ST18.08	50242	90.07	4.73	-1.49
ST18.09	43077	88.28	5.19	-0.41
ST19.00	60297	95.02	6.01	-0.81
ST19.01	39325	81.76	5.56	-0.92
ST19.03	64046	87.00	5.68	-0.94
ST21.01	71899	93.94	5.10	-1.22
ST21.02	68505	86.70	5.38	-1.71
ST21.04	71120	92.30	4.98	-1.16
ST21.05	14703	98.23	3.95	-3.03
ST21.06	43226	86.82	5.21	-1.05
ST21.07	68011	90.63	5.35	-1.90
ST21.08	26143	83.94	4.76	-1.64
ST22.00	37708	84.98	6.19	-0.43
ST22.01	27307	84.66	4.29	-1.87
ST22.03	53336	91.10	5.08	-2.62
ST22.04	58279	84.69	5.87	-1.28

ST22.05	71460	89.96	4.64	-0.98
ST22.08	49839	89.81	5.61	-0.67
ST22.09	80502	93.02	5.78	-3.11
ST23.00	137141	90.19	6.37	-1.48
ST23.01	82447	88.81	5.77	-1.57
ST23.02	162901	91.36	5.65	-1.44
ST23.03	92838	95.39	5.95	-1.35
ST23.05	42252	88.46	5.57	-2.32
ST23.06	23466	81.96	5.86	-0.98
ST23.07	145305	91.57	5.84	-1.63
ST23.08	64980	93.27	5.74	-1.84
ST24.00	92906	91.01	6.09	-1.85
ST24.01	77259	94.11	4.79	-0.99
ST24.02	83549	92.42	5.31	-1.70
ST24.03	72862	94.22	5.00	-0.84
ST24.04	61270	81.83	5.75	-2.25
ST24.05	3123	95.68	3.53	-3.08
ST24.07	50201	91.49	5.00	-1.33
ST24.08	55101	94.08	4.92	-1.60
ST27.00	52346	95.75	5.45	-0.30
ST27.01	77233	94.84	6.52	-0.61
ST27.02	108736	94.93	6.61	-0.49
ST27.03	103921	94.05	6.13	-0.79
ST28.01	33625	95.23	3.44	0.14
ST28.02	75994	92.31	5.25	-1.02
ST28.03	41743	96.35	4.43	-1.27
ST28.04	31427	95.24	5.19	-1.50
ST28.05	42229	93.52	5.11	-1.14
ST28.06	65128	91.14	5.51	-1.06
ST29.01	51587	88.97	3.66	-0.69
ST30.00	104852	92.91	5.95	-1.26
ST30.01	55352	86.67	5.09	-1.67
ST30.02	53795	93.07	4.58	-2.05
ST31.00	63047	92.28	6.44	-1.87
ST31.02	78372	86.29	6.80	-1.21
ST32.01	108840	93.04	5.84	-0.40
ST32.02	63695	90.19	6.85	-0.71
ST32.03	42633	84.15	6.53	-0.83
ST35.00	46302	83.72	5.61	-1.28
ST36.00	45186	84.78	5.88	-0.86
ST36.01	45726	87.79	5.18	-1.86
ST36.02	79388	89.78	5.91	-1.40
ST36.03	77714	91.13	6.29	-1.15
ST37.01	11516	91.83	4.81	-2.25
ST37.03	3590	97.34	6.69	-1.06
ST41.01	59881	85.57	6.18	-1.90
ST41.03	45277	85.17	4.94	-2.12

Supplementary file 2: denovo_and_rarefy_analysis.xlsx

Analysis	**Extra notes**	**Outcome**	**Predictor**			**Microbiome method**		
						Closed	***Closed subsampled to 3155 seq***	***De novo***
Table S3 – cases vs. controls	at baseline	shannon	UC/CD/control diagnosis	CD	estimate	-0.94	-0.66	-0.91
					pvalue	1E-05	6E-07	7E-05
				UC	estimate	-1.31	-0.85	-0.79
					pvalue	8E-05	6E-06	4E-03
		dysbiosis	UC/CD/control diagnosis	CD	estimate	0.86	0.89	0.88
					pvalue	6E-08	2E-11	2E-11
				UC	estimate	1.75	1.73	1.8
					pvalue	4E-15	2E-16	2E-16
	average	shannon	UC/CD/control diagnosis	CD	estimate	-0.72	-0.52	-0.69
					pvalue	7E-03	3E-09	6E-06
				UC	estimate	-0.98	-0.65	-0.5
					pvalue	2E-03	1E-05	0.02
		dysbiosis	UC/CD/control diagnosis	CD	estimate	0.67	0.69	0.69
					pvalue	3E-07	9E-11	1E-10
				UC	estimate	1.38	1.36	1.41
					pvalue	3E-10	3E-09	5E-09
Table S5 - associations with calprotectin	cases + controls	calprotectin	dysbiosis		estimate	260	250	243
					pvalue	4E-04	9E-06	1E-05
	cases only	calprotectin	dysbiosis		estimate	286	274	256
					pvalue	3E-04	1E-04	2E-04
		PCDAI	dysbiosis		estimate	5.37	5.31	5.18
					pvalue	1E-04	9E-04	1E-03

CHAPTER V. Discussion

Common themes emerge in studies of disparate diseases with gastrointestinal involvement.

Inflammatory bowel disease (IBD) and classic galactosemia (CG) are very different diseases. Though monogenic forms of IBD exist, in most cases there is no single causative gene. Studies have therefore taken a multi-pronged approach to understanding IBD, investigating the underlying genetics but also focusing a great deal on environmental exposures that may contribute. The cause of CG—mutations in the galactose-1-phosphate uridylyltransferase gene that result in null or very low activity of GALT—has been known for decades, but the mechanism underlying the pathophysiology of disease is still unknown. It is likely CG could benefit from a broader inquiry to identify environmental exposures and genetic factors outside of the *GALT* gene that contribute to the range of secondary health outcomes.

In both diseases, since the cause of severity is unclear, successful prognosis or intervention is also difficult. More hypothesis-generating experiments should be conducted to survey different potential routes of toxicity. My dissertation work has provided new perspectives on CG and IBD research, and some common themes will be important to research moving forward.

Integration of multiple data sets

Though increases in sample size will continue to improve the power of genetic studies, efforts should be taken to integrate multiple -omics data sets including metabolomics, transcriptomics, metagenomics, and exposomics, to get a more complete picture of the biological processes involved in disease.

One example of this need is that microbiome studies in IBD should not be performed without considering what we know about the host genetic architecture of disease. Studies have so far focused on contributions of small numbers of candidate genes. An early study of *Nod2*-deficient mice found increased bacterial load in feces and terminal ileum as well as decreased resistance to colonization by pathogenic bacteria. The authors also found that expression of *Nod2* was influenced by the presence of commensal bacteria[1]. Increased bacterial load has also been found in Crohn's disease patients homozygous for *NOD2* mutations[2]. A study of CD patients with homozygous *FUT2* mutations, another IBD-associated gene, found shifts in microbiome structure explained by *FUT2* as well as disease-by-genotype interactions for several bacterial groups[3]. One study of a large pediatric IBD cohort with genotype and microbiome information investigated associations between the two, but the thousands of host genetic loci and thousands of bacterial species in the microbiome present substantial problems when correcting for multiple tests[4]. Until sample sizes grow large enough, current knowledge such as the findings in mouse studies should be leveraged to correct for host genotype as a potential confounder in analyses. Microbiome research should also not only be limited to bacteria—viruses and fungi are important components of the human microbiome and human health[5,6].

Diet is another important data point that often gets overlooked, likely in part because of the complexity of data collection. However this information is vital to collect because of the impact diet itself can have on GI function and symptoms, as well as the microbiome[7,8].

Longitudinal data will also be critical to meaningful findings in these multi -omics projects, to help understand these systems over time. For example, we found that while the

IBD-characteristic dysbiosis index decreased over time with treatment, it did not decrease to levels seen in controls. However, the therapeutic importance of this observation is unclear since the dysbiosis index did not clearly associate with treatment outcome[9]. This interestingly parallels recent genetic research which found that genetic loci that associate with treatment outcome are mostly distinct from the loci which associate with disease diagnosis[10]. But more importantly this suggests that addressing components of the dysbiosis index may not be enough to improve outcomes; there are likely microbiome components associated with treatment outcome and treatment should be focused on those groups rather than the dysbiosis index microbes.

Pursuit of the gut microbiome as an attractive therapeutic target with relatively simple interventions

An early goal of microbiome research has been manipulation of bacterial populations to treat dysbiosis relative to control individuals. Supplementation of healthy bacteria via probiotics has shown some beneficial effects in a variety of GI disorders[11]. A more radical treatment involving fecal microbiota transplant (FMT) derived from a healthy subject into another with a severely disrupted microbiome has shown success resolving antibiotic-induced *Clostridium difficile* infection in mouse models[12] as well as in the clinic. Studies of FMT used a treatment for patients with *C. diff* infection consistently show resolution of disease in more than 80-90% of cases[13,14]. The benefits shown to be possible via intervention targeted to the microbiome and the relative non-invasiveness of therapy have led to clinical trials of FMT in many diseases[15], even before the role of the microbiome is clearly understood. This is the case in IBD, where results in Crohn's disease—but not ulcerative colitis—have been promising, with a pooled estimate of clinical remission of 60%[16].

However, studies have been small and difficult to compare due to differences in approach, so increasing sample size and standardizing procedures will be important to interpret results.

The gut microbiome has not yet been studied in CG, but would be interesting for multiple reasons. Individuals with CG have a fundamentally different diet compared to those without CG due to the necessity to avoid galactose-containing foods. Beyond dietary differences, it is additionally possible that the specific metabolic defect in CG, GALT deficiency, further modifies the gut microbiome. UDP sugar substrate pools are disrupted in CG compared to controls, leading to defects in glycosylation which may impact the mucosal layer of the gut. This in turn could compromise gut barrier function and commensal bacterial population structure (reviewed in [17]). Beyond improving GI symptoms, studying and treating any abnormalities in gut microbiome in CG could potentially improve developmental outcomes. Experiments have shown effects of microbiome transfer on behavior[18,19], and one study using a maternal-immune-system-induced mouse model of autism even showed resolution of stressed and repetitive behaviors using a single bacterial species administered as a probiotic at weaning[20].

Need for mechanisms

Dextran sodium sulfate (DSS) has been used for years to induce IBD-like intestinal inflammation in mice. However, in classic galactosemia research, *Galt* knockout mice repeatedly failed to recapitulate acute or long-term disease symptoms despite high amounts of galactose exposure. With the advent of CRISPR as a reliable, simpler method of introducing knockouts, the Fridovich-Keil lab knocked out *GALT* in a rat strain and have seen phenotypes similar to humans emerge (data not published). In both CG and IBD, findings from studies of genetics, diet, and environmental factors like the gut microbiome should be examined in available model systems to better understand causal mechanisms.

From studying gastrointestinal health in multiple contexts, we can gain general knowledge of pathophysiology of GI issues; this can in turn improve disease prevention, prognosis, and treatment.

REFERENCES

1. Petnicki-Ocwieja T, Hrncir T, Liu Y-J, Biswas A, Hudcovic T, Tlaskalova-Hogenova H, Kobayashi KS. Nod2 is required for the regulation of commensal microbiota in the intestine. Proc Natl Acad Sci U S A. 2009 Sep 15;106(37):15813–15818. PMCID: PMC2747201

2. Rehman A, Sina C, Gavrilova O, Häsler R, Ott S, Baines JF, Schreiber S, Rosenstiel P. Nod2 is essential for temporal development of intestinal microbial communities. Gut. 2011 Oct;60(10):1354–1362. PMID: 21421666

3. Rausch P, Rehman A, Künzel S, Häsler R, Ott SJ, Schreiber S, Rosenstiel P, Franke A, Baines JF. Colonic mucosa-associated microbiota is influenced by an interaction of Crohn disease and FUT2 (Secretor) genotype. Proc Natl Acad Sci U S A. 2011 Nov 22;108(47):19030–19035. PMCID: PMC3223430

4. Knights D, Silverberg MS, Weersma RK, Gevers D, Dijkstra G, Huang H, Tyler AD, van Sommeren S, Imhann F, Stempak JM, Huang H, Vangay P, Al-Ghalith GA, Russell C, Sauk J, Knight J, Daly MJ, Huttenhower C, Xavier RJ. Complex host genetics influence the microbiome in inflammatory bowel disease. Genome Med. 2014;6(12):107. PMCID: PMC4292994

5. Hallen-Adams HE, Suhr MJ. Fungi in the healthy human gastrointestinal tract. Virulence. 2016 Oct 13;1–7. PMID: 27736307

6. Focà A, Liberto MC, Quirino A, Marascio N, Zicca E, Pavia G. Gut inflammation and immunity: what is the role of the human gut virome? Mediators Inflamm. 2015;2015:326032. PMCID: PMC4405218

7. David LA, Maurice CF, Carmody RN, Gootenberg DB, Button JE, Wolfe BE, Ling AV, Devlin AS, Varma Y, Fischbach MA, Biddinger SB, Dutton RJ, Turnbaugh PJ. Diet rapidly and reproducibly alters the human gut microbiome. Nature. 2014 Jan 23;505(7484):559–563. PMCID: PMC3957428

8. Turnbaugh PJ, Ridaura VK, Faith JJ, Rey FE, Knight R, Gordon JI. The effect of diet on the human gut microbiome: a metagenomic analysis in humanized gnotobiotic mice. Sci Transl Med. 2009 Nov 11;1(6):6ra14. PMCID: PMC2894525

9. Shaw KA, Bertha M, Hofmekler T, Chopra P, Vatanen T, Srivatsa A, Prince J, Kumar A, Sauer C, Zwick ME, Satten GA, Kostic AD, Mulle JG, Xavier RJ, Kugathasan S. Dysbiosis, inflammation, and response to treatment: a longitudinal study of pediatric subjects with newly diagnosed inflammatory bowel disease. Genome Med. 2016;8(1):75. PMID: 27412252

10. Lee JC, Biasci D, Roberts R, Gearry RB, Mansfield JC, Ahmad T, Prescott NJ, Satsangi J, Wilson DC, Jostins L, Anderson CA, UK IBD Genetics Consortium, Traherne JA, Lyons PA, Parkes M, Smith KGC. Genome-wide association study identifies distinct genetic contributions to prognosis and susceptibility in Crohn's disease. Nat Genet. 2017 Feb;49(2):262–268. PMID: 28067912

11. Ritchie ML, Romanuk TN. A meta-analysis of probiotic efficacy for gastrointestinal diseases. PloS One. 2012;7(4):e34938. PMCID: PMC3329544

12. Lawley TD, Clare S, Walker AW, Stares MD, Connor TR, Raisen C, Goulding D, Rad R, Schreiber F, Brandt C, Deakin LJ, Pickard DJ, Duncan SH, Flint HJ, Clark TG, Parkhill J, Dougan G. Targeted restoration of the intestinal microbiota with a simple, defined bacteriotherapy resolves relapsing Clostridium difficile disease in mice. PLoS Pathog. 2012;8(10):e1002995. PMCID: PMC3486913

13. Kassam Z, Lee CH, Yuan Y, Hunt RH. Fecal microbiota transplantation for Clostridium difficile infection: systematic review and meta-analysis. Am J Gastroenterol. 2013 Apr;108(4):500–508. PMID: 23511459

14. Rossen NG, MacDonald JK, de Vries EM, D'Haens GR, de Vos WM, Zoetendal EG, Ponsioen CY. Fecal microbiota transplantation as novel therapy in gastroenterology: A systematic review. World J Gastroenterol. 2015 May 7;21(17):5359–5371. PMCID: PMC4419078

15. Search of: fecal microbiome transfer - List Results - ClinicalTrials.gov [Internet]. [cited 2017 Feb 23]. Available from: https://clinicaltrials.gov/ct2/results?term=fecal+microbiome+transfer&Search=Search

16. Colman RJ, Rubin DT. Fecal microbiota transplantation as therapy for inflammatory bowel disease: a systematic review and meta-analysis. J Crohns Colitis. 2014 Dec;8(12):1569–1581. PMCID: PMC4296742

17. Bergstrom KSB, Xia L. Mucin-type O-glycans and their roles in intestinal homeostasis. Glycobiology. 2013 Jun 10;cwt045. PMID: 23752712

18. Bravo JA, Forsythe P, Chew MV, Escaravage E, Savignac HM, Dinan TG, Bienenstock J, Cryan JF. Ingestion of Lactobacillus strain regulates emotional behavior and central GABA receptor expression in a mouse via the vagus nerve. Proc Natl Acad Sci U S A. 2011 Sep 20;108(38):16050–16055. PMCID: PMC3179073

19. Bercik P, Denou E, Collins J, Jackson W, Lu J, Jury J, Deng Y, Blennerhassett P, Macri J, McCoy KD, Verdu EF, Collins SM. The intestinal microbiota affect central levels of brain-derived neurotropic factor and behavior in mice. Gastroenterology. 2011 Aug;141(2):599–609, 609.e1–3. PMID: 21683077

20. Hsiao EY, McBride SW, Hsien S, Sharon G, Hyde ER, McCue T, Codelli JA, Chow J, Reisman SE, Petrosino JF, Patterson PH, Mazmanian SK. Microbiota modulate behavioral and physiological abnormalities associated with neurodevelopmental disorders. Cell. 2013 Dec 19;155(7):1451–1463. PMCID: PMC3897394