Title

Gardening Effect of Host Genetics on Human Intestinal Mucosal Microbiome and Its Link to Inflammatory Bowel Disease

Author
Tong, Maomeng

Publication Date
2014

UNIVERSITY OF CALIFORNIA

Los Angeles

Gardening Effect of Host Genetics on Human Intestinal Mucosal Microbiome and Its Link to Inflammatory Bowel Disease

A dissertation submitted in partial satisfaction of the requirements for the degree Doctor of Philosophy in Molecular and Medical Pharmacology

by

Maomeng Tong

2014

ABSTRACT OF THE DISSERTATION

Gardening Effect of Host Genetics on Human Intestinal Mucosal Microbiome and Its Link to Inflammatory Bowel Disease

by

Maomeng Tong
Doctor of Philosophy in Molecular and Medical Pharmacology
University of California, Los Angeles, 2014
Professor Jonathan Braun, Chair

Inflammatory bowel disease (IBD) is a set of chronic, relapsing inflammatory diseases of the intestine. The two major subtypes of IBD are Crohn's disease (CD) and ulcerative colitis (UC). Although the pathogenesis of IBD remains largely unknown, Crohn’s disease is considered to result from the interaction of environmental factors, including intestinal microbiota, with host immune mechanisms in genetically susceptible individuals. Recent advances in sequencing technologies have allowed us to characterize the IBD associated dysbiosis in unprecedented depth. However, phylogenetic profiling can only provide limited information on the functional implication of these alterations. To address this analytical challenge, we developed the novel mucosal lavage sampling approach, which enabled the profiling of multi’omic molecular features including

microbiome, metaproteome and metabolome. Combined with host genomic information, these tools can provide us with unprecedented understanding of the dynamics of host–microbial interaction, and help us to investigate the pathogenesis of inflammatory bowel diseases.

Another analytical challenge to identify microbial taxa consistently representing IBD associated dysbiosis is the high complexity and low inter-individual overlap of intestinal microbial composition. This difficulty can be overcome by an ecologic analytic strategy to identify modules of interacting bacteria (rather than individual bacteria) as quantitative reproducible features of microbial composition in normal and IBD mucosa. We developed the strategy to analyze microbial composition using microbial co-occurrence network approach. This strategy uncovered 5 reproducible functional microbial communities (FMCs) detectable in the mucosa of all individuals. The quantitative levels of two FMCs were significantly associated with IBD states. Imputed metagenome analysis indicated the functional importance of the disease associated modules reflected by the enrichment of virulent and pathogenic pathways. Thus, these modules appear to define novel microbial communities within the intestinal microbial ecology, some of which are commonly and stably modified by the IBD disease state, and may be of particular relevance for microbial pathogenesis and intervention.

Using this experimental and bioinformatic framework, we investigated the microbial gardening effect of *FUT2* gene and its link to Crohn's disease. Fucosyltransferase 2 (*FUT2*) is an enzyme that is responsible for the synthesis of the H antigen in body fluids and on the intestinal mucosa. Non-secretors, who are homozygous

for the loss-of-function alleles of *FUT2* gene (*sese*), have increased susceptibility to Crohn's disease. In healthy individuals, imputed metagenomic analysis revealed perturbations of energy metabolism in the microbiome of non-secretor and heterozygote individuals, notably the enrichment of carbohydrate and lipid metabolism, cofactor and vitamin metabolism, and glycan biosynthesis and metabolism related pathways; and, the depletion of amino acid biosynthesis and metabolism. Similar changes were observed in mice bearing the *FUT2*$^{-/-}$ genotype. Metabolomic analysis of human specimens revealed concordant as well as novel changes in the levels of several metabolites. Human metaproteomic analysis indicated that these functional changes were accompanied by sub-clinical levels of inflammation in the local intestinal mucosa. In an extended cohort containing both healthy and CD individuals, the phylogenetic composition of intestinal mucosal microbiota was affected by an interaction of Crohn's disease status and *FUT2* genotype. Decreased abundances of Firmicutes were associated with both CD and *FUT2* risk allele. At metagenomic level, a distinct signature of amino acid metabolism deficiency was identified in CD and non-secretor microbiome. Such changes were also reflected at metabolomic level in the proximal gut region. Taken together, *FUT2* gene increased the risk of Crohn's disease by changing the microbial composition and function to a disease-like state. The CD associated perturbations of metagenome and metabolome were driven by the *FUT2* risk allele.

The same experimental and bioinformatic approach can also be applied to study the composition and functional changes of mucosal associated microbiota in other chronic inflammatory disease, namely HIV-1 infection. In the rectal mucosa, microbial composition and imputed function in HIV-positive individuals not receiving cART was significantly different from HIV-negative individuals. Genera including *Roseburia, Coprococcus, Ruminococcus, Eubacterium, Alistipes* and *Lachnospira* were depleted in HIV-infected subjects not receiving cART, while *Fusobacteria, Anaerococcus, Peptostreptococcus* and *Porphyromonas* were significantly enriched. HIV-positive subjects receiving cART exhibited similar depletion and enrichment for these genera, but were of intermediate magnitude and did not achieve statistical significance. Imputed metagenomic functions, including amino acid metabolism, vitamin biosynthesis, and siderophore biosynthesis differed significantly between healthy controls and HIV-infected subjects not receiving cART. In the cervicovaginal mucosa, significant differences in alpha and beta diversity were observed between HIV-negative and HIV-positive women, with the latter enriched of organisms associated with bacterial vaginosis and depleted of Lactobacilli. These ecologic changes occurred concomitantly with significant metagenomic and immunologic differences. Such functional pathways may represent novel interventional targets for HIV therapy if normalizing the microbial composition or functional activity of the microbiota proves therapeutically useful.

COMMITTEE PAGE

The dissertation of Maomeng Tong is approved.

Heather R. Christofk

Thomas A. Drake

Huiying Li

Jonathan Braun, Committee Chair

University of California, Los Angeles

2014

DEDICATION

To my always supporting parents

Zhiying Bao and Zuming Tong

and beloved life partner

Lingzi Liu

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS

AA	Amino acid
cART	Combination anti-retroviral therapy
CCA	Constrained correspondence analysis
CD	Crohn’s disease
CVL	Cervical-vaginal lavage
FMC	Functional microbial communities
FUT2	Fucosyltransferase 2
GWAS	Genome-wide association study
HIV	Human immunodeficiency virus
HNP	Human neutrophil peptide
HUMAnN	Human Microbiome Project unified metabolic analysis network
IBD	Inflammatory bowel disease
KEGG	Kyoto Encyclopedia of Genes and Genomes
MALDI-TOF MS	Matrix-assisted laser desorption/ionization time-of-flight mass spectrometry
OTU	Operational taxonomic unit
PCA	Principal component analysis
PCoA	Principal coordinate analysis
PICRUSt	Phylotypic investigation of communities by reconstruction of unobserved states
QIIME	Quantitative Insights Into Microbial Ecology

Q-TOF MS	Quadrupole time-of-flight mass spectrometry
SCFA	Short-chain fatty acid
UC	Ulcerative colitis
WGCNA	Weighted gene correlation network analysis

ACKNOWLEDGEMENT

I would like to first and foremost thank my thesis advisor Dr. Jonathan Braun. His endless and unconditioned support created the unique risk-free environment where I can wholeheartedly study the most challenging, yet equally interesting and innovative problems in the field of microbiome research, without worrying about making mistakes. I cannot be grateful enough about what I have learned from him in, among other things, scientific research and leadership skills. I would like to thank Drs. Lynn Gordon, Bo Wei, and other current and past Braun/Gordon lab members for their guidance and support throughout the 5-year journey of my graduate study. I would also like to thank my graduate thesis committee, Drs. Heather R. Christofk, Thomas A. Drake and Huiying Li, for their insightful discussions on my research, and the advice for career development.

Chapters 2 and 3 have been published in *PLOS ONE* in 2013 (1). I would like to thank Dr. Michelle Xiaoxiao Li for establishing the patient cohort and sampling methodology, Drs. Andrew Ippoliti from Cedars-Sinai Medical Center and Bennett Roth from UCLA School of Medicine for clinical sample collections, Drs. Laura W Parfrey, Dan Knights and Rob Knight from University of Colorado Boulder for generating sequencing data and guiding me through the analysis of next-generation sequencing data and microbial ecology, Dr. Steve Horvath for his help with WGCNA, Drs. Daniel N. Frank from University of Colorado and Ellen Li from Stony Brook University for providing their IBD 16S sequencing datasets for this collaboration, and UCLA ATS Statistical Consulting Group for the help and suggestion of statistical analysis.

Chapter 4 has been published in *The ISME Journal* (2); Supplemental Chapter 2 has been published in *Microbiome* (3); and the manuscript for Chapter 5 is in preparation. These three chapters are part of the CCFA-MLI Consortium Project, which is an incredible collaborative effort involving several labs across the US. I would like to thank Drs. Dermot P.B. McGovern, Michelle Xiaoxiao Li, Talin Haritunians and Mr. Gildardo Barron from Cedars-Sinai Medical Center for establishing the patient cohort and performing the genetic analysis, Drs. Paul Ruegger and James Borneman from UC Riverside for generating the 16S rRNA sequencing data, Drs. Maryam Goudarzi and Albert J. Fornace Jr. from Georgetown University for generating the metabolomic data, Drs. Ian H. McHardy and Thomas G. Graeber for generating the meta-proteomic data, Drs. Purna C. Kashyap and Justin L. Sonnenburg from Stanford University for sharing their *FUT2*$^{-/-}$ mice dataset for this collaboration, and Dr. Curtis Huttenhower and Ms. Emma Schwager from Harvard University for the help with metagenome imputation and inter-omic analysis.

Supplemental Chapter 1 has been published in the *Journal of Clinical Investigation* (4). I would like to thank Drs. Bo Wei from UCLA and Lijun Xia from University of Oklahoma for providing me with the opportunity to participate in this study, and Dr. Ling Chen for her help with the immunofluorescent staining.

Supplemental Chapter 3 has been published in *Microbiome* (5). I would like to thank Drs. Xiaoxiao Michelle Li and Peter Anton for establishing the patient cohort and collecting the samples, Drs. Paul Ruegger and James Borneman from UC Riverside for generating the 16S rRNA sequencing data, Dr. Ian H. McHardy for the opportunity to be part of this study, and Dr. Jonathan Jacobs for the insightful discussion.

Supplemental Chapter 4 is currently in submission. I would like to thank Dr. Ian H. McHardy for the opportunity to be part of this collaboration and for generating the 16S rRNA sequencing data, and Drs. Marla J. Keller and Betsy C. Herold from Albert Einstein College of Medicine for establishing the patient cohort and collecting biospecimens.

In addition, I am grateful to the patients at UCLA Medical Center and Cedars-Sinai Medical Center, whose participation and contribution had made my thesis possible. Lastly, my graduate study has been supported by Burroughs Wellcome Fund Inter-school Training Program in Metabolic Diseases (BWF-IT-MD). I would like to thank the former program co-directors, Drs. Simin Liu and Thomas Drake, for such a rewarding opportunity to be part of the BWF family.

VITA

EDUCATION

2005-2009 Tsinghua University, Bachelor of Biological Sciences and Biotechnology

PUBLICATION

- **Tong M**, McHardy IH, Ruegger P, Goudarzi M, Kashyap P, Li X, Schwager E, Huttenhower C, Fornace A, Sonnenburg J, McGovern D, Borneman J, Braun J. Reprograming of Gut Microbiome Energy Metabolism by the *FUT2* Crohn's Disease Risk Polymorphism. ***ISME J.*** 29 April 2014; doi:10.1038/ismej.2014.64
- McHardy IH, Li X, **Tong M,** Ruegger P, Jacobs J, Borneman J, Anton P, Braun J. HIV Infection is associated with compositional and functional shifts in the rectal mucosal microbiota. ***Microbiome.*** 2013 Oct 12;1(1):26.
- McHardy IH, Goudarzi M, **Tong M**, Ruegger PM, Schwager E, Weger JR, Graeber TG, Sonnenburg JL, Horvath S, Huttenhower C, McGovern DP, Fornace AJ Jr, Borneman J, Braun J. Integrative analysis of the microbiome and metabolome of the human intestinal mucosal surface reveals exquisite inter-relationships. ***Microbiome.*** 2013 Jun 5;1(1):17.
- **Tong M**, Li X, Parfrey LW, Roth B, Ippoliti A, Wei B, Borneman J, McGovern DP, Frank DN, Li E, Horvath S, Knight R, Braun J. A modular organization of the human intestinal mucosal microbiota and its association with inflammatory bowel disease. ***PLoS One.*** 2013 Nov 19;8(11):e80702.
- Fu J, Wei B, Wen T, Johansson ME, Liu X, Bradford E, Thomsson KA, McGee S, Mansour L, **Tong M**, McDaniel JM, Sferra TJ, Turner JR, Chen H, Hansson GC, Braun J, Xia L. Loss of intestinal core 1-derived *O*-glycans causes spontaneous colitis in mice. ***J Clin Invest.*** 2011 Apr 1;121(4):1657-66.
- Wang X, Song X, Zhuo W, Fu Y, Shi H, Liang Y, **Tong M**, Chang G, Luo Y. The regulatory mechanism of Hsp90alpha secretion and its function in tumor malignancy. ***Proc Natl Acad Sci U S A.*** 2009 Dec 15;106(50):21288-93.

PUBLICATION - SUBMITTED

- **Tong M**, Jacobs JP, McHardy IH, Braun J. Microbial ecological analysis of mucosal associated microbiome in human using 16S rRNA sequencing. ***Under review (Curr Protoc Immunol)***
- McHardy IH, **Tong M,** LeBlanc J, Liu Z, Keller MJ, Herold BC, Braun J. Multi-omic Analysis of Cervicovaginal Lavage from HIV-negative and HIV-positive Women. ***Under review (mBio)***
- Li X, LeBlanc J, Elashoff D, McHardy IH, **Tong M,** Roth B, Ippoliti A, McDonald KG, Newberry RD, McGovern D, Graeber TG, Goodglick L, Braun J. Micro-geographic functional communities of the human colonic mucosa and their association with inflammatory bowel disease. ***Under review (Gastroenterology)***

PATENT

- Anton P, Braun J, McHardy IH, Jacobs J, **Tong M**. Rectal Mucosa Sampling Tool. Tech ID: 22740 / UC Case 2012-535-0

INVITED TALKS

- Gardening of Human Intestinal Mucosal Microbiome by Fucosyltransferase 2 (*FUT2*) Gene. In: Annual Retreat of Department of Molecular and Medical Pharmacology, Nov. 2013, Huntington Beach, CA.
- Redefining the Ecology of Intestinal Microflora through Mucosal Functional Neighborhoods. In: The 19th International Microbial Genomes Conference, Sept. 2012, Lake Arrowhead, CA.

HONORS AND AWARDS

- 2012-2014 Fellowship, Burroughs Wellcome Fund Inter-school Program in Metabolic Diseases
- 2012 National Institute of Health (NIH) Award for the International Human Microbiome Congress

CHAPTER 1

Introduction

Inflammatory bowel disease (IBD) is a set of chronic, relapsing inflammatory diseases of the intestine. The two major subtypes of IBD are Crohn's disease (CD) and ulcerative colitis (UC). Although the pathogenesis of IBD remains largely unknown, Crohn’s disease is considered to result from the interaction of environmental factors, including intestinal microbiota, with host immune mechanisms in genetically susceptible individuals (6, 7).

Recent advances in next-generation sequencing technologies have allowed researchers to assess the composition and variation of human microbiome in an unprecedented depth. Since the launch of major microbiome research initiatives including the Human Microbiome Project (HMP) and Metagenomics of the Human Intestinal Tract Project (MetaHIT), our knowledge have expanded significantly regarding how the composition and function of gut microbiome affect human health (8, 9). However, several analytical issues emerged for sequencing based microbiome study. The first challenge is the high complexity and low inter-individual overlap of intestinal microbial composition (8, 10, 11). This variability has complicated the association of microbial phylogenetic composition with disease, in that it is challenging to determine if the absence of a given phylotype in a healthy or disease subject is due to the pathogenic physiology or simply temporal or inter-individual stochastic fluctuations. Although a “core microbiome” at the gene level is identifiable (12), the core feature at the organismal lineage level, which resolves functionally redundant phylotypes into distinct communities, has not yet been defined. In the context of IBD microbial pathogenesis, due to the limitations of current microbial analysis, reproducible microbial features established for human IBD are quite limited: reduced alpha diversity, and a small number of elevated or reduced taxa detectable at the level of patient categories but only sporadic at the level of individual patients. Accordingly, these findings are sufficient neither to test the current pathogenesis concept, nor to provide a strategy to classify and monitor individual patients for disease-associated microbial taxa.

The second challenge is that phylogentic profiling of microbiota only provide limited insight into the metagenomic functional outcomes of such dysbiosis. The intestinal mucosal ecosystem harbors an assortment of host factors, microbiota, and metabolites. A central goal and methodologic challenge in human-associated microbial ecology is to identify dietary, metabolic, and host and microbial factors that drive microbial community structure (8, 13). Identification of such relationships is fundamental for us to understand the host-microbial interaction in IBD pathogenesis and for interventional strategies to alter microbiota composition and function in the context of dysbiosis. Indeed, direct analysis of metabolic output by and interactions between microbial species is a burgeoning investigative field, but challenging methodologically (14, 15).

Dysbiosis, which refers to perturbations of the normally stable intestinal microbiota, has been associated with the development and progression of inflammatory bowel diseases (16-18). The reasons for such associations are not yet clear and may reflect either causal or secondary processes due to the impact on microbial composition and function of inter-individual variability and the contributions of environment and host genetics (19). In the meantime, genome-wide association studies have identified a complex set of polymorphisms that confer varying levels of genetic risk for IBD (20).

Functional annotation of the host genes tagged by these loci suggests that impaired handling of commensal microbes and pathogens is a prominent factor in disease development. Determining the extent of host genetic influence on the composition and function of the gut microbiome is an important next step in understanding the mechanisms linking these genetic traits with microbial function and disease biology.

CHAPTER 2

Lavage Sampling Enables Multi’omic Analysis of Mucosal Associated Microbiota

Abstract

Background

To study the dysbiosis of host-associated intestinal mucosal microbiota in chronic diseases, appropriate collection and pre-processing of biospecimens from humans is necessary for accurate analysis of microbial composition.

Methods

64 subjects were recruited, including 32 normal subjects (NM), 16 Crohn's disease patients in remission (CD), and 16 ulcerative colitis patients in remission (UC). 190 mucosal lavage samples were collected from different anatomical regions of these subjects. To profile phylogenetic composition, the hyper-variable region 4 of the 16S ribosomal RNA gene was then amplified and sequenced on an Illumina HiSeq 2000. The soluble metabolites of the same samples were analyzed using quadrupole time-of-flight (Q-TOF) mass spectrometry.

Results

Microbial composition and phylotype richness of lavage samples were comparable with the 16S sequence datasets generated from fecal or tissue samples. High yields of soluble fraction proteins and metabolites were produced from lavage samples, which enabled the identification of robust, disease-specific biochemical features of the mucosal surface.

Conclusion

Lavage sampling, by permitting microbial and biochemical analysis from the same mucosal site, is a novel strategy for integrated multi’omic analysis to functionally characterize the intestinal microbial ecosystem.

Introduction

The importance of host-associated commensal microbiota in maintaining the health of humans has been well appreciated. Dysbiosis of microbiota is implicated in various chronic diseases including obesity, diabetes, inflammatory bowel disease, HIV, vaginosis, asthma and other immune related chronic disorders (Kinross et al., 2011). Recent advances in sequencing technologies have enabled the profiling of microbial compositions with unprecedented depth and coverage at significantly lower cost, and therefore substantially improved our understanding of host-associated microbiome in different habitats.

To study the dysbiosis of host-associated intestinal mucosal microbiota in chronic diseases, appropriate collection and pre-processing of biospecimens from humans is necessary for accurate analysis of microbial composition. The two widely used approaches of sampling intestinal microbiota are stool samples and biopsy. Compared with fecal samples, collection of lavage samples is warranted for two reasons: first, the phylogenetic composition of mucosa-associated microbiota is distinct from that of the luminal compartment (1, 21); second, fecal samples are a mixture of products from all intestinal regions, which may obscure the unique biogeography of host-bacteria interactions along intestine (22). In addition to the invasiveness, biopsy specimens also cannot be used to assess microbial metabolites and proteins as much of the material would derive from human cells.

Routine screening colonoscopy can be slightly modified to collect mucosal lavage samples that can be used for combined phylogenetic, proteomic, and metabolomic analysis. Prior to the procedure, patients scheduled for colonoscopy are consented per IRB requirements. Samples may then be collected any time during endoscopic examination, though physicians have historically preferred to collect samples upon completion and with retraction of the endoscope. During a typical colonoscopy, mucosal washes are routinely collected and discarded. Therefore, only slight modifications are necessitated for implementation of this protocol. Collected mucosal lavage samples can then be processed further and analyzed using many potential methods, including high-throughput sequencing or mass spectrometry for proteomic and metabolomic analysis.

Materials and Methods

Patient cohorts and lavage sample collection

A previously assembled patient cohort of 64 subjects was examined (Table 2-1) in accord with human subject protocols approved by the institutional review boards of University of California Los Angeles and Cedars Sinai Medical Center. All enrolled subjects were prepared for colonoscopy by taking Golytely the day before the procedure. The mucosal lavage samples representing the mucosal luminal interface were collected from different intestinal regions as described previously (22). All the lavage samples in this cohort were collected from non-involved intestinal regions, which excluded the potential influence of active inflammation on the mucosal microbiota as much as possible. Subjects metadata, including diagnosis, gender, age, and colon regions sampled, were recorded. The influence of medication on the microbiome was not evaluated, due to the unavailability of data.

16S rRNA gene sequencing and microbial composition analysis

After collection, the sample was centrifuged at 3,500g for 15 minutes to separate the microbiota from the soluble fraction. Genomic DNA was extracted as described in Costello *et al*.(23). The hyper-variable region 4 of the 16S ribosomal RNA gene was then amplified and sequenced on an Illumina HiSeq 2000 as described in Caporaso *et al.* (24). The sequence data is deposited in European Bioinformatics Institute [EMBL: ERP001780]. The median read length of sequences that passed quality filtering is 90 bp and the average read length is 88 bp with a filtering threshold of 75bp. For quality control, all the singletons were removed, and samples with fewer than 3,000 reads were excluded from the following analyses. The 97% OTUs were picked against the Greengenes reference database (February 4th, 2011) first, then reads that did not match a Greengenes sequence at 97% or greater sequence identity were clustered *de novo* using uclust (25). Taxonomy of each OTU was assigned by blasting the representative sequence against Greengenes reference database (26) (http://greengenes.lbl.gov/cgi-bin/nph-index.cgi). These steps were performed using Quantitative Insights Into Microbial Ecology (QIIME) v1.4.0 (27).

Alpha rarefaction was performed using the Phylogenetic Diversity index. Ten sampling repetitions were performed at each sampling depth ranging from 10 to 3,000 reads. The comparison of alpha diversity between two groups at certain sampling depths was performed using a two-sided Student *t* test. Significance was defined as a *P* value of less than 0.05. Beta diversity was estimated by computing unweighted UniFrac distances between samples using QIIME. Principal coordinates analysis (PCoA) was applied to reduce the dimensionality of the resulting distance matrix.

Mass spectrometry analysis

For metabolomics analysis, each human lavage samples was subjected to solid-phase extraction to eliminate a polymeric contaminant believed to originate from the lubricant used during colonoscopy preparation. The eluate was dried and reconstituted in 2% acetonitrile in water prior to MS analysis. A 5 µL aliquot of extracted metabolites from each sample was injected onto a reverse-phase 50 × 2.1 mm ACQUITY 1.7-µm C18 column (Waters Corp, Milford, MA) using an ACQUITY UPLC system (Waters Corp, Milford, MA). A Waters Q-TOF Premier was operated in negative-ion (ESI-) or positive-ion (ESI+) electrospray ionization mode with a capillary voltage of 3200 V and a sampling cone voltage of 20 V in negative mode and 35 V in positive mode. Data were acquired in centroid mode with a mass window of 50 to 850 *m/z*, and processed using MassLynx software (Waters Corp, Milford, MA).

Results

Microbiome dataset from lavage samples recaptured IBD associated dysbiosis

To study the host-microbial interaction at the mucosal luminal interface, 179 lavage samples were collected from different intestinal regions of 64 subjects; this cohort and these samples were previously used in a metaproteomic study of IBD (22) (Table 2-1). The microbiota from these samples were profiled by multiplex sequencing, and a total of 1,236,641 reads (6,909/sample on average) were generated after quality control. 10,208 species level OTUs were then generated by collapsing the reads at a 97% sequence similarity threshold. At the phylum level, the bacterial community from lavage sample mainly consisted of Bacteroidetes (44.29%), Firmicutes (35.48%), Proteobacteria (6.76%), Tenericutes (1.63%) and Verrucomicrobia (1.35%) (Figure 2-1). Other phyla were also detected at relatively low abundances (<1%) including Actinobacteria and Fusobacteria. We compared our dataset with other 16S sequence datasets generated from fecal or tissue samples after re-processing them using the same OTU picking and taxonomy assignment algorithms. Given the difference of sequencing platforms, primer sets and colon regions, the Tong dataset was comparable with other intestinal microbial datasets in terms of microbial composition and phylotype richness.

The IBD-associated dysbiosis of mucosal microbiota has been delineated in detail in several investigations (16, 28, 29). Specifically, IBD patients have fewer Firmicutes and a concomitant increase in Proteobacteria, validated in several independent cohorts (30, 31). To determine whether previously reported alterations were also observed in our dataset, we compared the relative abundances of each phylum between disease states using analysis of variance (ANOVA). In contrast to controls, IBD patients harbored relatively more abundant Actinobacteria (FDR corrected $P = 0.006$ for UC, < 0.0001 for CD), accompanied with the depletion of Firmicutes (FDR corrected $P = 0.056$ for UC, 0.25 for CD) in these subjects (Figure 2-2 A). The increases of Proteobacteria (FDR corrected $P = 0.254$ for UC, 0.143 for CD) and Tenericutes (FDR corrected $P = 0.115$ for UC, 0.157 for CD) were also observed in IBD patients, although not statistically significant. Taken together, microbial composition represented by this study cohort, and captured by lavage sampling, reflected the changes of relative abundances of enteric microbiota in IBD subjects at phylum level observed in other datasets and sampling methods.

The reduction in bacterial diversity in IBD patients is a consistent finding across studies (16, 32, 33), although it is still unknown whether this alteration is causative or a secondary effect of IBD. Compared with controls, the phylogenetic diversities of UC and CD subjects at 97% OTU level were significantly lower (Figure 2-2 B), and the difference was more evident in CD (t-test, *P* = 0.0003) than that in UC subjects (t-test, *P* = 0.0056) at the depth of 3,000 reads per sample. This data indicates that the lower microbial diversity previously observed in patients with active IBD also persists in clinically quiescent phases of disease.

To evaluate the similarity between microbial communities in lavage samples from control and IBD subjects, the beta-diversity measured by unweighted distance matrix was calculated for each sample. The principal coordinate analysis (PCoA) plot showed that the samples clustered by diagnosis (Figure 2-2 C). The IBD-associated dysbiosis of mucosal microbiota was reflected by the cluster of samples enriched for IBD, especially CD subjects. 55% of the IBD samples (49% of UC and 61% of CD) were in this cluster, whereas only 8% of the controls were in this IBD enriched clusters. The clustering was evident considering the heterogeneity of the pathogenesis of IBD (34), although control samples can be observed in the IBD enriched cluster and not all the IBD samples were grouped into this subset.

Metabolomic profiling of mucosal luminal interface using lavage samples

We profiled the soluble metabolites of the same lavage samples using Q-TOF MS. The analysis generated a rich metabolomic dataset consisting of 649 and 576 spectral features in the cecum and sigmoid regions, respectively. Putative IDs were assigned to 372 ions by comparing their *m/z* values to those available in online databanks using a predefined mass error window of 20 ppm. The putative IDs were then used to map out the ions to various metabolomic pathways in the KEGG dataset. In accord with our previous study, ~50% of all metabolites were located at the terminal end of metabolic pathways, suggesting enrichment for end-products (3).

Discussion

Mucosal sampling provided robust differentiation at the individual patient level that has not been achieved previously by analysis of the fecal compartment and conventional analyses (individual phylotypes levels, community alpha-diversity, or principal component analysis) (30, 31). How might the design features of the study have contributed to this outcome? One distinction was the first use of mucosal lavage for depth microbial analysis. Lavage samples microbiota embedded in the superficial mucin, but may also include luminal fecal residue remaining after intestinal preparation. Mucosa-associated microbial composition varies across segments of the intestine, and distinct as well from the fecal compartment (23, 35). Due to the predominant inter-individual signal, it is uncertain whether the lavage compartment yields a distinct microbial composition from mucosal or fecal sampling.

Nonetheless, compared to fecal samples, mucosal lavage is from a defined (~1 cm^2) area of mucosal surface (22), and therefore captures a local microbial community more homogeneous for local metabolic exchange and interaction in the microenvironment. Indeed, since the local habitat modifies the functional state of the microbial community, lavage samples can be analytically extended to define microbial state (by transcriptional and metagenomic analysis of the bacterial pellet) and the habitat (by biochemical analysis of supernatant proteins and metabolites). We have recently reported high yields of soluble fraction proteins and metabolites by lavage sampling, and have uncovered robust, disease-specific biochemical features of the mucosal surface (22). Lavage sampling, by permitting microbial and biochemical analysis from the same mucosal site, could be extended to integrated multi'omic analysis to functionally characterize the intestinal microbial ecosystem. And, owing to its noninvasiveness, lavage sampling in contrast to biopsy or surgery permits longitudinal sampling which is an important barrier to monitoring the mucosal microbiota and its dynamic temporal state (10, 23).

Figures and Tables

Figure 2-1

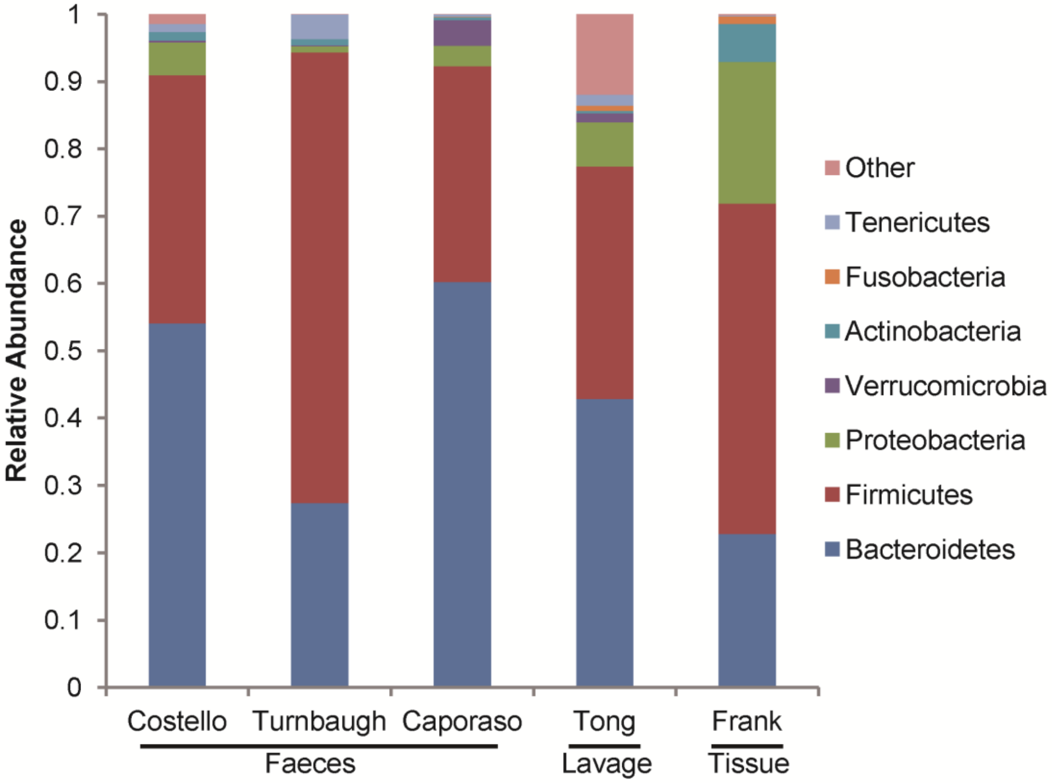


Figure 2-1. Phylum level microbial compositions of faeces, lavage and tissue samples. Biospeciemens from faeces (Costello (23), Turnbaugh (12) and Caporaso (10)), lavage samples (Tong) and tissue samples (Frank (16)) were compared. Only predominant phyla with relative abundances higher than 0.1% in Tong dataset were depicted in the bar graph, and the phyla with low abundances were grouped together. For Costello and Caporaso datasets, only the fractions of intestinal microbiota were shown here.

Figure 2-2

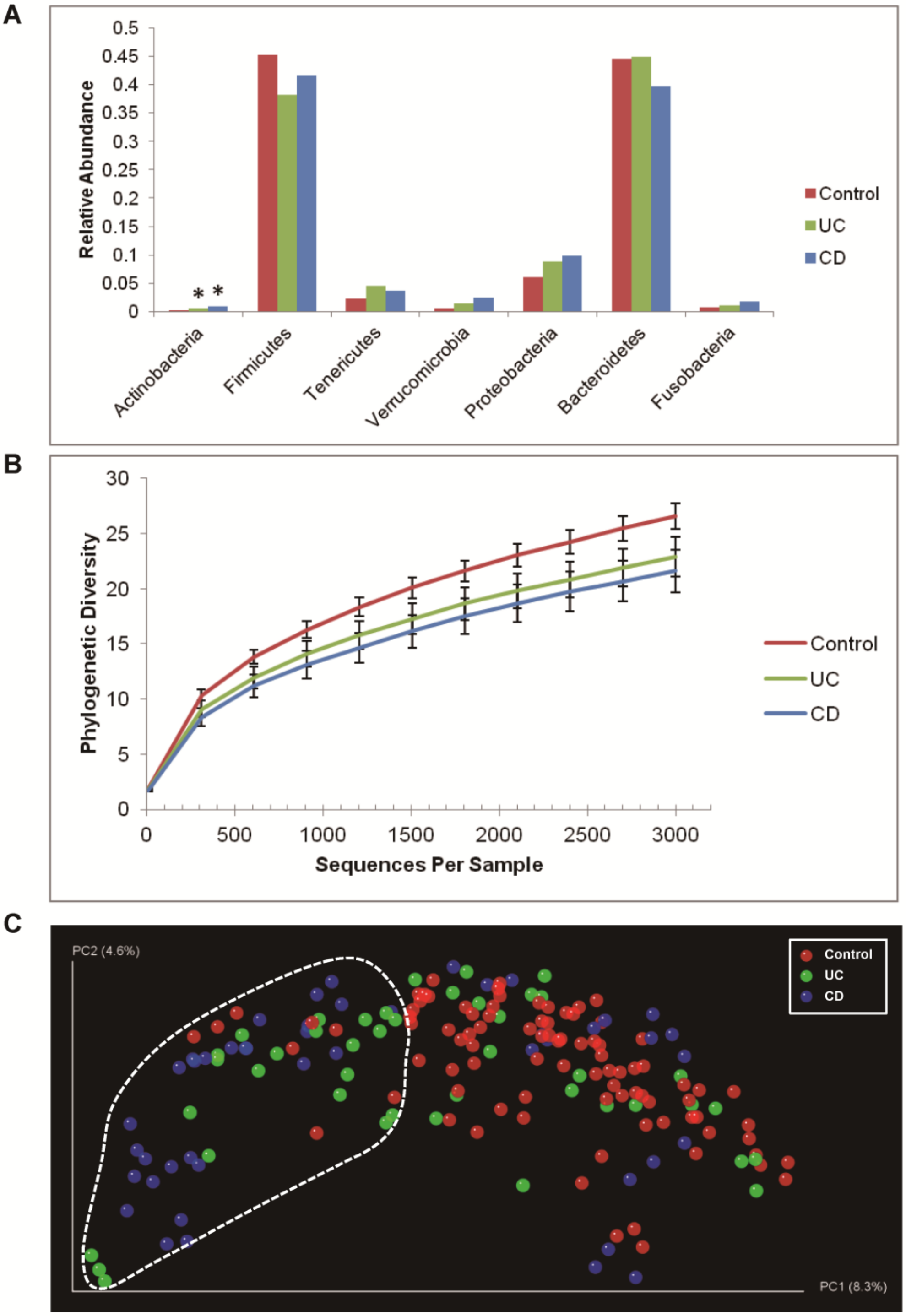

A
Relative Abundance
0.5
0.45
0.4
0.35
0.3
0.25
0.2
0.15
0.1
0.05
0
* *
Actinobacteria
Firmicutes
Tenericutes
Verrucomicrobia
Proteobacteria
Bacteroidetes
Fusobacteria
Control
UC
CD
B
Phylogenetic Diversity
30
25
20
15
10
5
0
0
500
1000
1500
2000
2500
3000
Sequences Per Sample
Control
UC
CD
C
PC2 (4.6%)
PC1 (8.3%)
Control
UC
CD

Figure 2-2. Shifts of mucosal microbial composition in IBD patients in remission. **(A)** The change of relative abundance between disease states at phylum level. *: $P <$ 0.05 compared to control, ANOVA. **(B)** Phylogenetic diversity curves for the microbiota from lavage samples. Mean ± 95% CI was shown. **(C)** Communities clustered using PCoA of the unweighted UniFrac distance matrix. Each point corresponds to a sample colored by disease phenotype. The dotted line indicated the cluster of samples enriched for IBD subjects.

Table 2-1. Demographic information of Tong dataset

		Control	UC	CD
Subject (64)		32	16	16
Gender	Male	20	9	11
	Female	12	7	5
Age (Average ± SD)		60 ± 12	36 ± 12	41 ± 12
Sample (179)		90	43	46
Anatomical region	CE (1)	1	0	0
	AS (47)	25	15	7
	TR (22)	4	8	10
	DE (57)	31	12	14
	RE (52)	29	8	15

Note: on average, 3 samples of different intestinal regions were collected from each subject. UC, ulcerative colitis; CD, Crohn's disease; CE, cecum; AS, ascending colon; TR, transverse colon; DE, descending colon; RE, rectum.

CHAPTER 3

Bacterial Co-Occurrence Network at the Mucosal-Luminal Interface

Abstract

Background

Abnormalities of the intestinal microbiota are implicated in the pathogenesis of Crohn's disease (CD) and ulcerative colitis (UC), two spectra of inflammatory bowel disease (IBD). However, the high complexity and low inter-individual overlap of intestinal microbial composition are formidable barriers to identifying microbial taxa representing this dysbiosis. These difficulties might be overcome by an ecologic analytic strategy to identify modules of interacting bacteria (rather than individual bacteria) as quantitative reproducible features of microbial composition in normal and IBD mucosa.

Methods

To investigate the change of microbial ecology on the intestinal mucosal surface and its role in the pathogenesis of IBD, 16s ribosomal DNA (rDNA) extracted from the bacterial pellets were sequenced by Illumina HiSeq 2000 to determine the phylogenetic distribution of the microbiomes. The microbial co-occurrence network was constructed according to the abundance profiles of each genus, and the functional microbial communities (FMCs) were detected *via* weighted correlation network analysis. The gene content of 1,119 KEGG reference genomes was used to infer the approximate gene content of the detected 97% OTUs in our dataset using Phylogenetic Investigation of Communities by Reconstruction of Unobserved States (PICRUSt) (v0.1), and the metabolic pathways were re-constructed using HUMAnN (v0.98)

Results

Analysis of weighted co-occurrence network revealed 5 microbial modules. These modules were unprecedented, as they were detectable in all individuals, and their composition and abundance was recapitulated in an independent, biopsy-based mucosal dataset. Two modules were associated with healthy, CD, or UC disease states. Imputed metagenome analysis indicated that these modules displayed distinct metabolic functionality, specifically the enrichment of oxidative response and glycan metabolism pathways relevant to host-pathogen interaction in the disease-associated modules. The highly preserved microbial modules accurately classified IBD status of individual patients during disease quiescence, suggesting that microbial dysbiosis in IBD may be an underlying disorder independent of disease activity. Microbial modules thus provide an integrative view of microbial ecology relevant to IBD.

Conclusion

These modules appear to define novel microbial communities within the intestinal microbial ecology, some of which are commonly and stably modified by the IBD disease state, and may be of particular relevance for microbial pathogenesis and intervention.

Introduction

Inflammatory bowel disease (IBD), a spectrum of chronic, relapsing inflammatory intestinal diseases, results from the interaction of environmental factors, including intestinal microbiota, with host immune mechanisms in genetically susceptible individuals (6, 7). Human and animal studies demonstrate the involvement of intestinal microbiota in the onset or perpetuation of inflammation, and intensive efforts have search for individual bacterial species and specific bacterial products in the pathogenesis of IBD (36-38). However, rather than revealing a single agent responsible for disease, these studies have uncovered a variety of bacterial taxa and products that can either promote or attenuate the inflammatory disease state. Moreover, the relevant microbiota differ in accord with the genetic susceptibility traits of the host (29, 39-41). These insights have shifted the concept of microbial pathogenesis in IBD away from specific pathogens and towards ecologic, community-level change (42), and raised concomitant challenges of establishing coherent concepts and analytic strategies to identify microbiota relevant to disease risk or disease activity in individual IBD patients.

In recent years, the phylogenetic and functional characterizations of the human enteric microbiota in IBD have been elucidated with the help of second-generation sequencing platforms. One striking feature of human intestinal microbiome is its great inter-individual phylotypic variation (8, 10, 11). This variability has complicated the association of microbial phylogenetic composition with disease, in that it is challenging to determine if the absence of a given phylotype in a healthy or disease subject is due to the pathogenic physiology or simply temporal or inter-individual stochastic fluctuations.

Although a “core microbiome” at the gene level is identifiable (12), the core feature at the organismal lineage level, which resolves functionally redundant phylotypes into distinct communities, has not yet been defined. In the context of IBD microbial pathogenesis, this has prompted the current concept that an individual’s distinct microbial composition (shaped by host genetics, founder effects, and diet) may create a disease-susceptible ecology prone to blooms of pathobionts (and/or busts of protective taxa) when stressed by environmental, metabolic, or viral disturbances (43). However, due to the limitations of current microbial analysis, reproducible microbial features established for human IBD are quite limited: reduced alpha diversity, and a small number of elevated or reduced taxons detectable at the level of patient categories but only sporadic at the level of individual patients. Accordingly, these findings are sufficient neither to test the current pathogenesis concept, nor to provide a strategy to classify and monitor individual patients for disease-associated microbial taxa.

To validate this concept and allow clinical translation, we must move beyond existing studies of taxon and/or gene composition to instead quantify relevant features of the microbial community at the ecological level. Extensive inter-species interactions exist in the highly complex intestinal microbial ecosystem (44, 45). Investigating the hundreds of thousands of possible pairwise inter-species interactions in a defined system is not feasible (46), especially because few known intestinal microbes are cultivable. 16S rRNA gene profiling allows us infer inter-species correlations from relative abundance profiles. Several benchmarking studies have documented microbial co-occurrence in different environments (47-50), but the role of inter-species interactions during the pathogenesis of chronic disease remains largely unexplored. Here we adopted a methodology for phylogenetic network analysis to search for such interactions, suggesting that the human mucosal surface bacterial community is organized into 5 highly preserved modules. Two of these modules are reciprocally associated with inflammatory bowel disease.

Materials and Methods

Construction of microbial co-occurrence network

We first defined a co-occurrence similarity measure which was used to define the network. Assume that the vector x_i specifies the abundance of the *i*-th genus across the samples, the pair-wise Sparse Correlations for Compositional data (SparCC) ρ_{ij} was inferred from the abundance profile of each genus x_i and x_j as the measurement of co-occurrence relationship. A signed weighted adjacency matrix (network) was defined by raising ρ_{ij} to a power $a_{ij} = (0.5 + 0.5\rho_{ij})$ ^ β, with $\beta = 4$ (51). The power is a soft threshold that preserves the continuous nature of the underlying co-occurrence information. The relatively low power of 4 (chosen with the scale free topology criterion) likely reflected the fact that the network was comprised of relatively few nodes (263 genera). Once the network was constructed, modules were then defined as branches of a hierarchical clustering tree based on the topological overlap measure, because it is a highly robust measure of network interconnectedness. The modules were detected after applying the dynamic tree cut method (52). These network modules (clusters) were interpreted as functional microbial communities (FMCs). To summarize the profiles of co-occurrence modules, we calculated the eigengenus, which provides a mathematically optimal way of summarizing the co-occurrence patterns of all genera belonging to each module. To identify modules (FMCs) that were correlated with clinical traits, we used correlation tests to relate each eigengenus to the clinical traits. These steps were performed using WGCNA package (version 1.13) in R (version 2.13.1) (53).

Module preservation analysis

Meta-analysis was performed with two mucosal microbial datasets: the previously published “Frank” dataset (16), and the “Tong” dataset presented in this paper (using the same biospecimens described in a recently reported patient cohort (22)). Prior to meta-analysis, the taxonomy of each OTU in Frank dataset was re-assigned by blasting the representative sequence against Greengenes reference database. The common phylotypes at genus level that were present in both dataset were then identified. After this filtering step, 1,196,466 out of 1,236,641 reads (96.8%), or 129 out of 263 genera in Tong dataset, and 13,165 out of 15,172 reads (86.8%), or 129 out of 263 genera in Frank dataset were included in the following analysis. To determine whether a FMC found in the reference dataset was also present in the test dataset, we used a powerful module preservation statistic implemented in the R software function modulePreservation (54). For each module, the aggregate measure of module preservation was termed the preservation Z-summary statistic. The higher the value of the Z-summary statistic is for a given module, the stronger the evidence that the module is preserved in the test dataset. Comprehensive simulation studies led to the following thresholds: a module shows no evidence of preservation if its Z-summary statistic is smaller than 2; a Z-summary statistic larger than 5 (or 10) indicates moderate (strong) module preservation.

Imputation of microbial gene content and metagenomes of FMCs

The OTU table of the 5 FMCs in Tong dataset was generated with 1 count for each 97% OTU in a given FMC. The gene content of 1,119 KEGG reference genomes was used to infer the approximate gene content of the detected 97% OTUs in our dataset using Phylogenetic Investigation of Communities by Reconstruction of Unobserved States (PICRUSt) (v0.1). The functional traits copy numbers of the reference genomes represented in the format of KEGG KO functions can be downloaded from the PICRUSt website (http://picrust.github.com). To predict the functional traits of non-sequenced microbial genomes (i.e. 97% OTUs) in Tong dataset, a phylogenetic tree of 97% OTUs in Greengenes database was constructed using 16S marker gene. The tree has tips representing both sequenced referenced genomes and non-sequenced genomes. Then the ancestral state reconstruction (ASR) was run for this tree to make predictions for each KO functions for every internal node and unsequenced tips in the phylogenetic tree. The program output the inferred metagenome represented by KEGG Orthology for each FMC. Taking the PICRUST KO gene abundance inferences as inputs, the metabolic pathways were re-constructed using HUMAnN (v0.98) (55).

Results

Defining a microbial co-occurrence network at the intestinal mucosal surface

Extensive inter-species interactions are likely to operate among mucosal-associated microbiota residing in the complex and functionally diverse ecosystem of the intestines, either locally through formation of biofilms or through diffusion of nutrients and metabolites longitudinally along the intestine (56, 57). Such interactions can thus be potentially reflected by the co-occurrence and co-exclusion patterns inferred from abundance profiles of phylotypes (58). Therefore, in addition to individual phylotypes, we must identify IBD-associated microbial community structures. To test the hypothesis that interactions among microbes increase our ability to classify samples according to clinical state, we constructed the microbial co-occurrence network using an approach specifically tailored for the 16S profiling data (Figure 3-1). The edge connecting each pair of nodes was the co-occurrence estimate inferred from the relative abundance profiles of genera using the sparse correlation measure SparCC (59), which ranged from -0.541 to 0.774, suggesting strong co-exclusion and co-occurrence relationships between phylotypes. As described in Methods, we transformed the SparCC correlation measure into a weighted network.

Identification of highly preserved functional microbial communities (FMCs)

To understand the topological structure of a network, one crucial step is to define modules, which are groups of highly connected nodes. In biological networks, modules can correspond to functional subunits such as protein complexes (60) or molecular pathways (61). There is an extensive literature on clustering procedures, including simple k-means, partitioning around medoid, hierarchical clustering, message passing and model-based methods (62-65). To determine if the genera in the microbial co-occurrence network can form network modules, we adapted weighted correlation network analysis (implemented in the WGCNA package) to construct microbial modules which can be interpreted as functional microbial communities (FMCs). WGCNA uses a measure of shared protein neighbors (based on the topological overlap measure) as input of hierarchical clustering. The height in the dendrogram is a measure of dissimilarity based on the topological overlap matrix; modules are defined as branches of a hierarchical cluster tree (51, 53). WGCNA is attractive in our study since it provides module preservation statistics that allowed us to assess the reproducibility of modules across different data sets; provides a measure of intramodular connectivity that can be used to define intramodular hub genera (66); and, allows us to summarize each module by its module eigengenus.

We first calculated the pair-wise topological overlap matrix after the soft-thresholding step to reduce the noise-level weak correlations (Methods). After grouping the nodes based on their topological overlaps using hierarchical clustering, we identified 5 functional microbial communities (Figure 3-2 A), which consisted of 5 to 167 phylogenetically diverse genera. Using the same method, analysis of the Frank dataset (263 genera) identified 6 microbial modules (Figure 3-2 B), with a similar numerical range of genera per module. The Tong and Frank datasets shared 129 genera. This reduced, common set of shared phylotypes yielded a similar module organization: 4 modules in the Tong dataset, and 2 modules in the Frank dataset.

To quantitatively evaluate the degree of module preservation, we carried out a Z-summary test. Alternative statistics are available to assess the quality and reproducibility of clusters among datasets (54, 67-71). An advantage of the Zsummary statistic is that it allows for significance thresholds: Z-summary <2 indicates no significant module preservation; 2<Z-summary<10 indicates moderate preservation; and, Z-summary>10 indicates strong preservation. Also, in previous work comparing WGCNA's module preservation statistics to a robust alternative method (the in-group proportion test of Kapp and Tibshirani (72)), both tests were highly correlated under a number of simulation conditions, and the Zsummary statistic had distinct advantages for studying the preservation of network modules (53, 54, 73).

As expected, all 4 modules of the shared 129 genera from the Tong dataset were highly preserved in the Frank dataset (Figure 3-2 C), with the blue FMC demonstrating the strongest preservation. Conversely, all the modules in the Frank dataset were also well-preserved in the Tong dataset (Figure 3-2 D). Given the methodological differences between Tong and Frank datasets, the co-occurrence pattern of these genera can still be observed at mucosal surface. We expected even more significant preservation when comparing datasets collected from same compartment and analyzed using same methodology. Indeed, when using another lavage sample dataset, referred hereafter as the mucosal luminal interface or MLI dataset, as the reference, 4 modules of the shared 233 genera from the Tong dataset were highly preserved in the MLI dataset, with much higher Z-summary statistics. Thus, there were 2 core modules (turquoise and blue) in these datasets, despite the difference of sampling methods. These results indicated that the FMCs identified using our approach were not dataset specific, but robust and reproducible ecological structures commonly existing at the intestinal mucosal surface.

Identification of functional microbial communities (FMCs) associated with IBD

An optimal summary of the genus abundance profiles of a given FMC is the module eigengenus. In the Tong dataset, we found that the turquoise FMC was significantly associated with Crohn's disease state (P = 4 × 10^{-5}, Pearson correlation) (Figure 3-2 E). The blue FMC was negatively associated with IBD states, although not statistically significant. If the turquoise FMC was merely a group of individual CD-associated genera, it would include most of the 17 genera that were significantly enriched in CD samples (ANOVA, FDR corrected $P < 0.05$). However, only 7 of them were assigned to the turquoise FMC, indicating that FMCs also captured other intricate and underlying ecological relationships. Strikingly, classification of IBD status using the two core FMCs as quantitative microbial biomarkers achieved higher accuracy (17/47, or 36.2%) compared to using individual genera (Figure 3-3), indicating that the microbial modules allowed quantitative and reproducible microbial monitoring of the intestinal mucosa. Because the two core FMCs were highly preserved in both datasets, the same associations were also observed in the Frank dataset. The turquoise FMC was positively associated with IBD states, most significantly with UC (P = 9 × 10^{-7}, Pearson correlation), whereas the blue FMC was negatively associated with CD (P = 1 × 10^{-9}, Pearson correlation) (Figure 3-2 F). The associations were stronger in the Frank dataset than those in the Tong dataset, possible because the samples were from patients with active disease. Consistent with previous observation (29), the blue FMC was also negatively associated with the *NOD2* risk allele in the Frank dataset, supporting the hypothesis that the CD-associated dysbiosis was driven by the *NOD2* risk allele (39, 74).

After defining modules, we sought to analyze them by intuitive topological concepts such as intramodular connectivity, to better describe the network structure. Therefore, we determined the kME value (intramodular connectivity based on the module eigengenus) to define the correlation between each genus and the respective module eigengenus. Because nodes with high connectivity, i.e. the hubs, are centrally located within the module, they may be functionally essential as keystone species in the context of biological networks (75) and during the assemblage of a disease associated FMC. Indeed, in the turquoise FMC, the intramodular connectivities of the genera enriched in CD samples were significantly higher than those of the other members (t-test, P < 0.001). Potential pathobiont genera such as *Enterococcus* (76) and *Escherichia* (including adherent-invasive *Escherichia coli* (77)) can also be observed among the hub genera of CD-associated turquoise FMC. In the blue module, one of the intramodular hub genera was *Faecalibacterium*, a genus including the anti-inflammatory commensal bacterium *Faecalibacterium prausnitzii*, that is negatively associated with Crohn's disease (28, 78, 79). Accordingly, the relative abundance of *Faecalibacterium* decreased by 2-fold in Crohn's disease samples (ANOVA, FDR corrected P = 0.006). Other short-chain fatty acid (SCFA) producing bacteria including *Eubacterium*, *Roseburia*, *Faecalibacterium* and *Coprococcus* were also observed in the blue FMC (80-82). Taken together, these data demonstrated the functional importance of the FMCs associated with CD.

Metabolic inference and reconstruction of functional microbial communities

The disease associations of the well preserved FMCs suggest that these co-occurred microbial communities represent distinct functional units at the mucosal surface. To profile the metabolic capabilities of FMCs, the approximate gene contents of the detected phylotypes in each FMCs were inferred using the 1,119 KEGG reference genomes. After aggregating the individual inferred genomes according to module membership, the relative abundances of metabolic pathways in each FMC were re-constructed. The functional profiles of FMCs were significantly variable (Figure 3-4). The representation of the functional groups that were likely essential for life in the gut was highly consistent across FMCs including those for carbohydrate and amino-acid metabolism (for example glycolysis/gluconeogenesis (KO00010), pyruvate metabolism (KO00620) and glycine, serine and threonine metabolism (KO00260)). In contrast, several virulent pathways including bacterial invasion of epithelial cells (KO05100) and pathogenic *Escherichia coli* infection (KO05130) were only present in the IBD-associated turquoise FMC. Variably represented pathways included glycan degradation (KO00511) and glycosaminoglycan degradation (KO00531), which were over-represented in the UC-associated brown FMC; and, glutathione metabolism (KO00480), which was enriched in turquoise FMC. With respect to the former, murine defects in mucosal barrier function due to depletion of intestinal *O*-glycans causes spontaneous colitis (4). Regarding the latter, an increase in glutathione metabolism is a feature of intestinal microbiome in inflammatory bowel disease (18). These observations, combined with the disease association, indicated that the imputed virulent metabolic functions carried out by the disease associated FMCs contributed to the pathogenic and chronic inflammatory state of intestinal mucosal surface.

Discussion

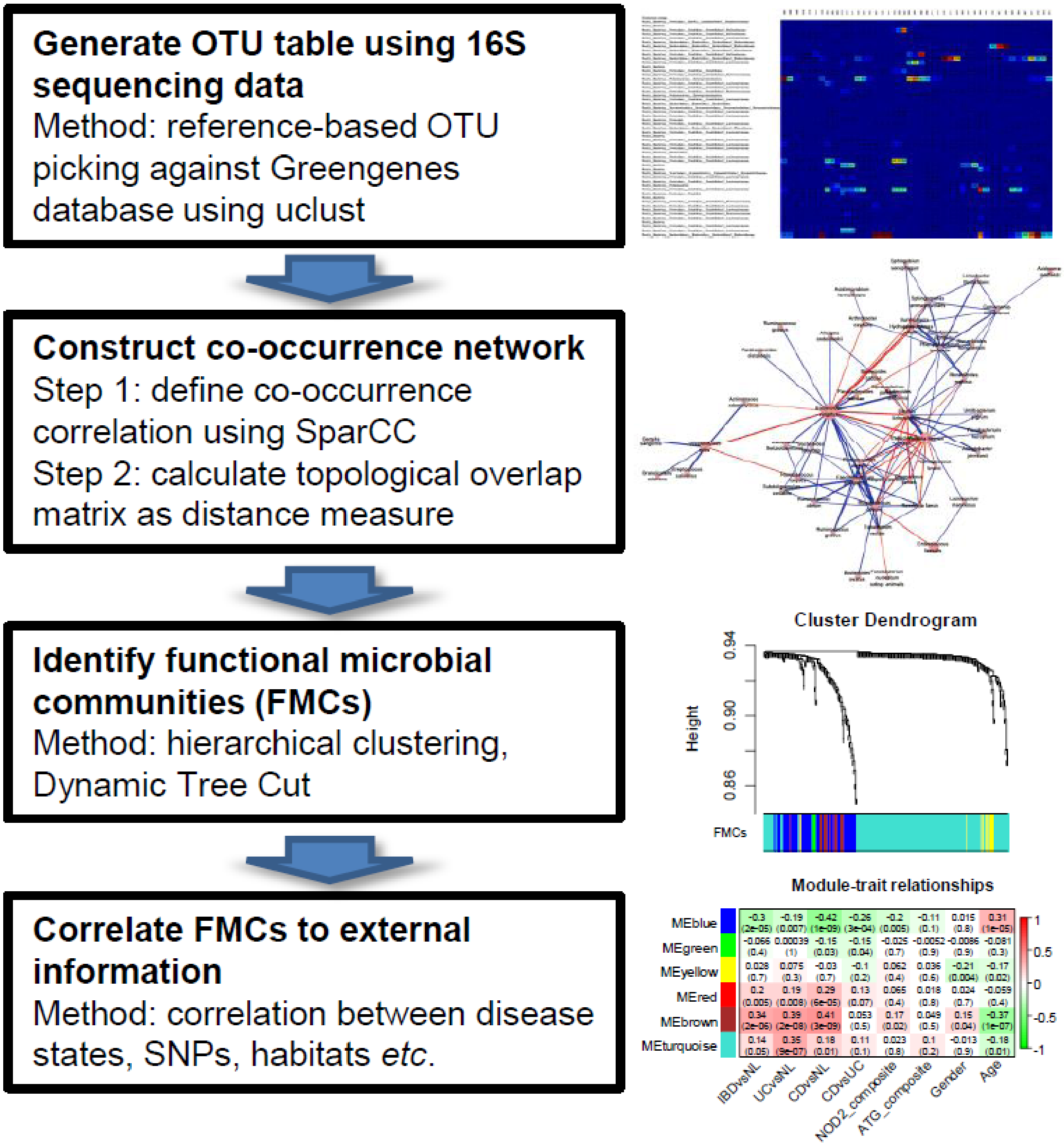
We have developed a novel strategy using an ecologic mucosal microbial framework, minimally invasive mucosal sampling, short-read Illumina sequencing, network analysis, and imputed metagenomics. This strategy uncovered 5 microbial modules detectable in the mucosa of all individuals, and reproducible in an independent mucosal resection dataset (Figure 3-2). The quantitative levels of two modules were significantly associated with disease states (Figure 3-2 E). More than 70% of the subjects can be correctly classified as control or IBD patients using genera from rectal sampling alone (Figure 3-3), which is a minimally invasive procedure compared to endoscopic biopsy, and thus notable for clinical translation. Imputed metagenome analysis indicated the functional importance of the disease associated modules reflected by the enrichment of virulent and pathogenic pathways (Figure 3-4). Thus, these modules appear to define novel microbial communities within the intestinal microbial ecology, some of which are commonly and stably modified by the IBD disease state, and may be of particular relevance for microbial pathogenesis and intervention.

The present study uncovered similar networks of co-occurrent and co-exclusive microbiota, confirming shared features detected in the fecal and mucosal lavage compartments. In addition, extending the analysis to WGCNA uncovered 5 reproducible microbial modules, each comprised of distinct but phylogenetically mixed group of organisms, and a blend of positive and negative microbial interactions. We have termed them functional microbial communities (FMCs), with the speculation that they reflect a physically localized and biologically integrated microbial network. Since each module is defined by both positively and negatively microbial interactions, we speculate that they will be defined by a distinctive ensemble of biologic factors, such as host microenvironment and microbial gardening, microbial cross-feeding and competition, and microbial small molecule and environmental modification (83-86).

Therapeutic intervention targeting microbial dysbiosis in inflammatory bowel disease is an important prospect for changing the natural history for patients with inflammatory bowel disease. However, the heterogeneity and temporal variation of microbial composition requires new concepts to define the target of microbial intervention, and analytic tools to accurately sub-stratify and monitor individual patients. In our study, all 5 of the FMCs identified were present in all the subjects, but with different overall abundances that varied with disease states. In the CD-associated turquoise FMC, the difference of intramodular connectivity suggested that the pathogenic microbes were more likely to be the core members of the microbiota, rather than opportunistic pathogens. Direct evidence for the physical localization of such ecological structures could be validated using methods such as fluorescence *in situ* hybridization, whereas functional features of these communities would require comprehensive metagenomic or biochemical analysis. In this respect, a recent metaproteomic study of the mucosa surface detected a physical microgeographic mosaic of proteins, which might represent a biochemical counterpart to the microbial modules.

Figures and Tables

Figure 3-1



Generate OTU table using 16S sequencing data
Method: reference-based OTU picking against Greengenes database using uclust

Construct co-occurrence network
Step 1: define co-occurrence correlation using SparCC
Step 2: calculate topological overlap matrix as distance measure

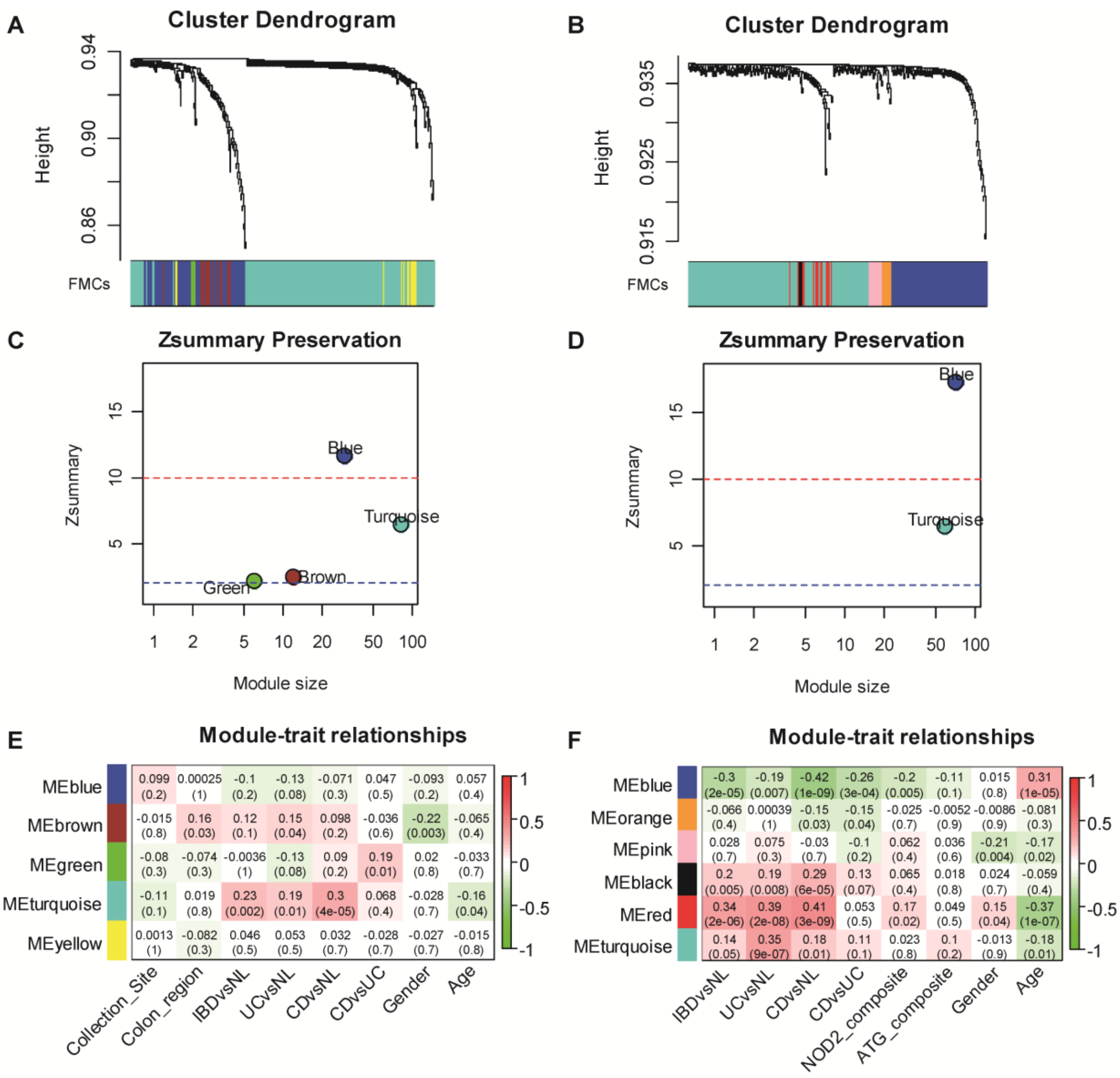
Identify functional microbial communities (FMCs)
Method: hierarchical clustering, Dynamic Tree Cut

Correlate FMCs to external information
Method: correlation between disease states, SNPs, habitats *etc*.

	IBDvsNL	UCvsNL	CDvsNL	CDvsUC	NOD2_composite	ATG_composite	Gender	Age
MEblue	-0.3 (2e-05)	-0.19 (0.007)	-0.42 (1e-09)	-0.26 (3e-04)	-0.2 (0.005)	-0.11 (0.1)	0.015 (0.8)	0.31 (1e-05)
MEgreen	-0.066 (0.4)	0.00039 (1)	-0.15 (0.03)	-0.15 (0.04)	-0.025 (0.7)	-0.0052 (0.9)	-0.0086 (0.9)	-0.081 (0.3)
MEyellow	0.028 (0.7)	0.075 (0.3)	-0.03 (0.7)	-0.1 (0.2)	0.062 (0.4)	0.036 (0.6)	-0.21 (0.004)	-0.17 (0.02)
MEred	0.2 (0.005)	0.19 (0.008)	0.29 (6e-05)	0.13 (0.07)	0.065 (0.4)	0.018 (0.8)	0.024 (0.7)	-0.059 (0.4)
MEbrown	0.34 (2e-06)	0.39 (2e-08)	0.41 (3e-09)	0.053 (0.5)	0.17 (0.02)	0.049 (0.5)	0.15 (0.04)	-0.37 (1e-07)
MEturquoise	0.14 (0.05)	0.35 (9e-07)	0.18 (0.01)	0.11 (0.1)	0.023 (0.8)	0.1 (0.2)	-0.013 (0.9)	-0.18 (0.01)

Figure 3-1. Overview of the methodology for inferring microbial co-occurrence network and identifying functional microbial communities.

Figure 3-2



E Module-trait relationships

	Collection_Site	Colon_region	IBDvsNL	UCvsNL	CDvsNL	CDvsUC	Gender	Age
MEblue	0.099 (0.2)	0.00025 (1)	-0.1 (0.2)	-0.13 (0.08)	-0.071 (0.3)	0.047 (0.5)	-0.093 (0.2)	0.057 (0.4)
MEbrown	-0.015 (0.8)	0.16 (0.03)	0.12 (0.1)	0.15 (0.04)	0.098 (0.2)	-0.036 (0.6)	-0.22 (0.003)	-0.065 (0.4)
MEgreen	-0.08 (0.3)	-0.074 (0.3)	-0.0036 (1)	-0.13 (0.08)	0.09 (0.2)	0.19 (0.01)	0.02 (0.8)	-0.033 (0.7)
MEturquoise	-0.11 (0.1)	0.019 (0.8)	0.23 (0.002)	0.19 (0.01)	0.3 (4e-05)	0.068 (0.4)	-0.028 (0.7)	-0.16 (0.04)
MEyellow	0.0013 (1)	-0.082 (0.3)	0.046 (0.5)	0.053 (0.5)	0.032 (0.7)	-0.028 (0.7)	-0.027 (0.7)	-0.015 (0.8)

F Module-trait relationships

	IBDvsNL	UCvsNL	CDvsNL	CDvsUC	NOD2_composite	ATG_composite	Gender	Age
MEblue	-0.3 (2e-05)	-0.19 (0.007)	-0.42 (1e-09)	-0.26 (3e-04)	-0.2 (0.005)	-0.11 (0.1)	0.015 (0.8)	0.31 (1e-05)
MEorange	-0.066 (0.4)	0.00039 (1)	-0.15 (0.03)	-0.15 (0.04)	-0.025 (0.7)	-0.0052 (0.9)	-0.0086 (0.9)	-0.081 (0.3)
MEpink	0.028 (0.7)	0.075 (0.3)	-0.03 (0.7)	-0.1 (0.2)	0.062 (0.4)	0.036 (0.6)	-0.21 (0.004)	-0.17 (0.02)
MEblack	0.2 (0.005)	0.19 (0.008)	0.29 (6e-05)	0.13 (0.07)	0.065 (0.4)	0.018 (0.8)	0.024 (0.7)	-0.059 (0.4)
MEred	0.34 (2e-06)	0.39 (2e-08)	0.41 (3e-09)	0.053 (0.5)	0.17 (0.02)	0.049 (0.5)	0.15 (0.04)	-0.37 (1e-07)
MEturquoise	0.14 (0.05)	0.35 (9e-07)	0.18 (0.01)	0.11 (0.1)	0.023 (0.8)	0.1 (0.2)	-0.013 (0.9)	-0.18 (0.01)

Figure 3-2. Identification of preserved functional microbial communities (FMCs) associated with disease phenotype across studies. Hierarchical clustering dendrograms of genera based on microbial co-occurrence network using the Tong dataset **(A)** and the Frank dataset **(B)** are shown. In the dendrograms, each color represents one FMC, and each branch represents one genus. The Z-summary statistic

plots (y-axis) as a function of the module size are shown for the Tong dataset **(C)** and the Frank dataset **(D)**. Each point represents a module labeled by color. The dashed blue and red lines indicate the thresholds Z = 2 and Z = 10, respectively. FMC-trait correlations and *P* values of the Tong dataset **(E)** and the Frank dataset **(F)**. Each cell reports the correlation coefficient (and *P* value) derived from correlating FMC eigenvectors (rows) to traits (columns). The table is color-coded by correlation according to the color legend. Collection site: University of California Los Angeles or Cedars Sinai Medical Center; Colon region: 5 anatomical regions coded from 0 to 5, which are cecum, ascending colon, transverse colon, descending colon and rectum.

Figure 3-3

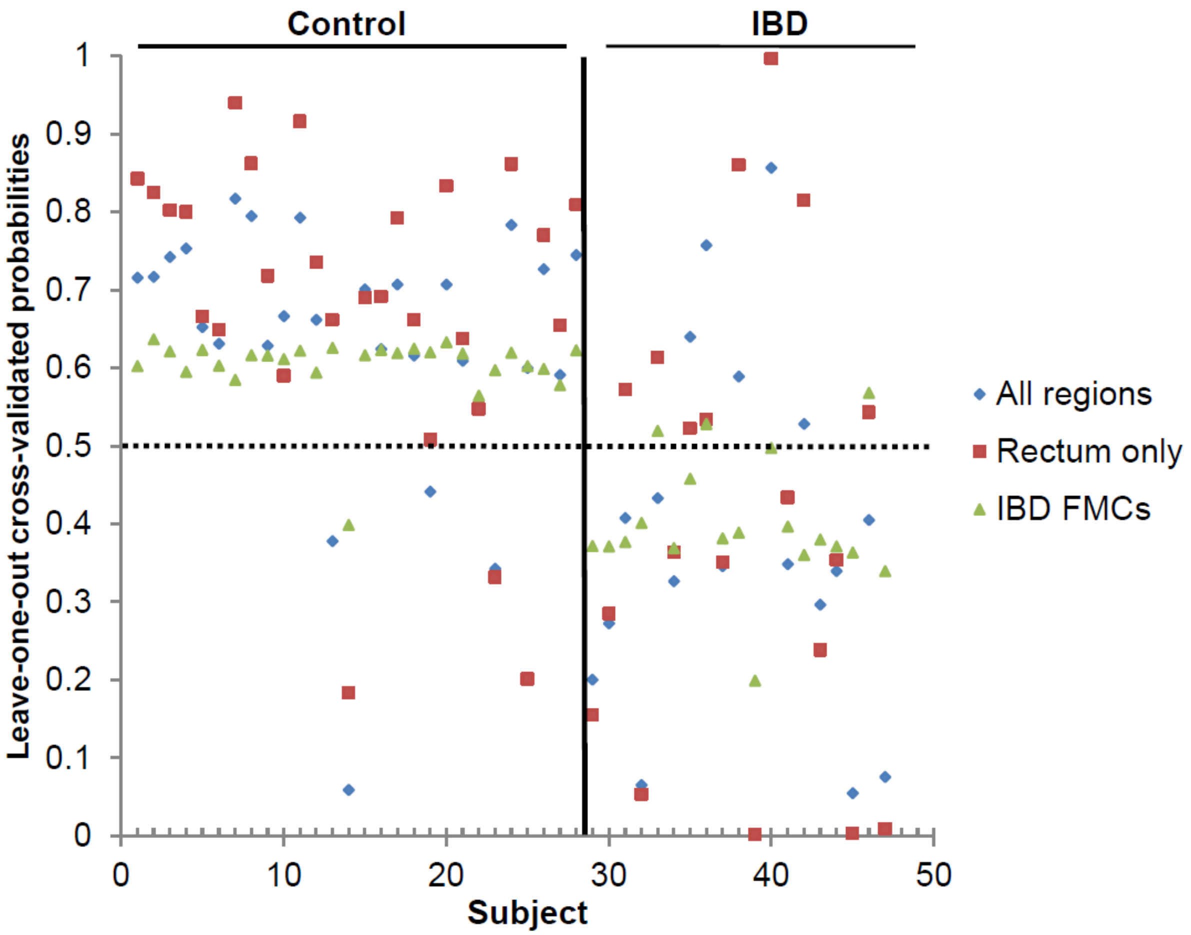


Figure 3-3. Classification of control and IBD subjects using nearest shrunken centroids analyses of the relative abundances of bacterial genera and FMCs from lavage samples. Only subjects (n = 47) that had matched samples from both descending colon and rectum regions were included in the analysis. Control and IBD samples with leave-one-out cross-validated probabilities higher than 50% were considered correctly classified. Diamond, classification using 30 genus-region variables (error = 18/47, or 38.3%); Square: classification using 39 rectum genera variables (error

= 14/47, or 29.8%); Triangle, classification using 4 FMC-region variables (error = 17/47, or 36.2%).

Figure 3-4

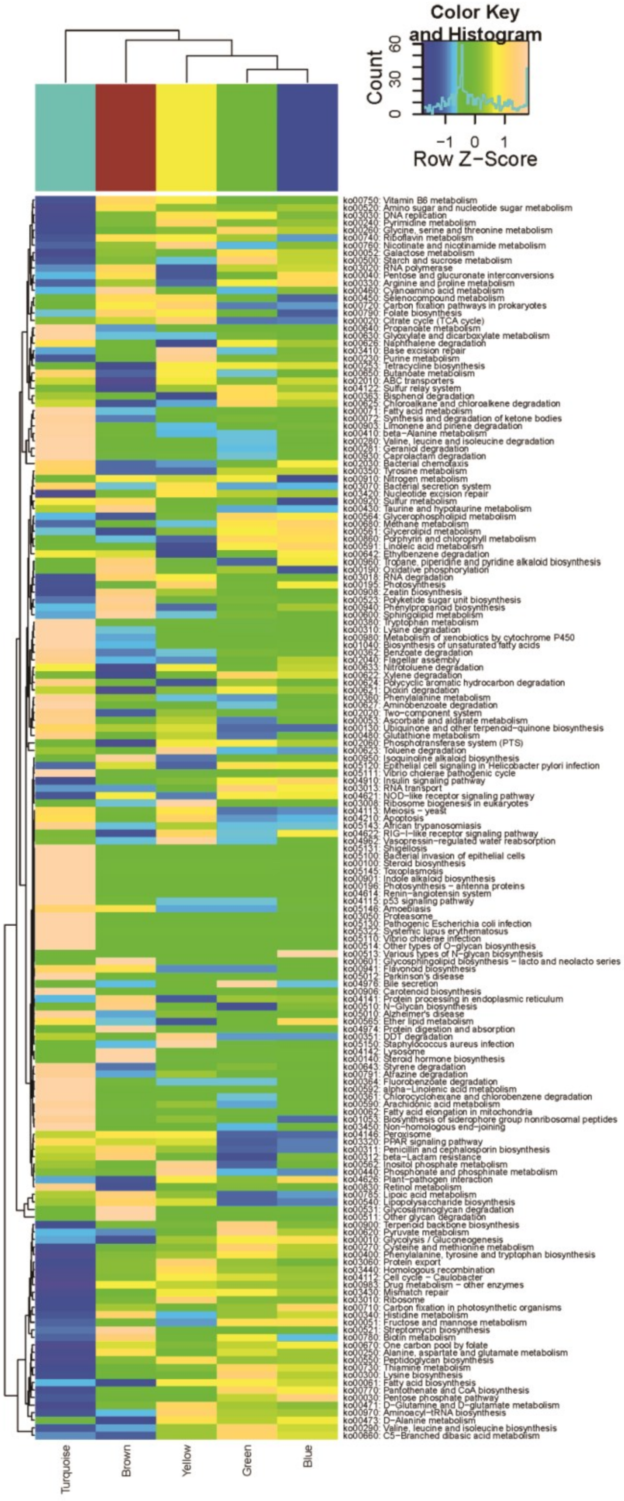

Color Key
and Histogram
Count
0
30
60
−1
0
1
Row Z−Score
ko00750: Vitamin B6 metabolism
ko00520: Amino sugar and nucleotide sugar metabolism
ko03030: DNA replication
ko00240: Pyrimidine metabolism
ko00260: Glycine, serine and threonine metabolism
ko00740: Riboflavin metabolism
ko00760: Nicotinate and nicotinamide metabolism
ko00052: Galactose metabolism
ko00500: Starch and sucrose metabolism
ko03020: RNA polymerase
ko00040: Pentose and glucuronate interconversions
ko00330: Arginine and proline metabolism
ko00460: Cyanoamino acid metabolism
ko00450: Selenocompound metabolism
ko00720: Carbon fixation pathways in prokaryotes
ko00790: Folate biosynthesis
ko00020: Citrate cycle (TCA cycle)
ko00640: Propanoate metabolism
ko00630: Glyoxylate and dicarboxylate metabolism
ko00626: Naphthalene degradation
ko03410: Base excision repair
ko00230: Purine metabolism
ko00253: Tetracycline biosynthesis
ko00650: Butanoate metabolism
ko02010: ABC transporters
ko04122: Sulfur relay system
ko00363: Bisphenol degradation
ko00625: Chloroalkane and chloroalkene degradation
ko00071: Fatty acid metabolism
ko00072: Synthesis and degradation of ketone bodies
ko00903: Limonene and pinene degradation
ko00410: beta−Alanine metabolism
ko00280: Valine, leucine and isoleucine degradation
ko00281: Geraniol degradation
ko00930: Caprolactam degradation
ko02030: Bacterial chemotaxis
ko00350: Tyrosine metabolism
ko00910: Nitrogen metabolism
ko03070: Bacterial secretion system
ko03420: Nucleotide excision repair
ko00920: Sulfur metabolism
ko00430: Taurine and hypotaurine metabolism
ko00564: Glycerophospholipid metabolism
ko00680: Methane metabolism
ko00561: Glycerolipid metabolism
ko00860: Porphyrin and chlorophyll metabolism
ko00591: Linoleic acid metabolism
ko00642: Ethylbenzene degradation
ko00960: Tropane, piperidine and pyridine alkaloid biosynthesis
ko00190: Oxidative phosphorylation
ko03018: RNA degradation
ko00195: Photosynthesis
ko00908: Zeatin biosynthesis
ko00523: Polyketide sugar unit biosynthesis
ko00940: Phenylpropanoid biosynthesis
ko00600: Sphingolipid metabolism
ko00380: Tryptophan metabolism
ko00310: Lysine degradation
ko00980: Metabolism of xenobiotics by cytochrome P450
ko01040: Biosynthesis of unsaturated fatty acids
ko00362: Benzoate degradation
ko02040: Flagellar assembly
ko00633: Nitrotoluene degradation
ko00622: Xylene degradation
ko00624: Polycyclic aromatic hydrocarbon degradation
ko00621: Dioxin degradation
ko00360: Phenylalanine metabolism
ko00627: Aminobenzoate degradation
ko02020: Two−component system
ko00053: Ascorbate and aldarate metabolism
ko00130: Ubiquinone and other terpenoid−quinone biosynthesis
ko00480: Glutathione metabolism
ko02060: Phosphotransferase system (PTS)
ko00623: Toluene degradation
ko00950: Isoquinoline alkaloid biosynthesis
ko05120: Epithelial cell signaling in Helicobacter pylori infection
ko05111: Vibrio cholerae pathogenic cycle
ko04910: Insulin signaling pathway
ko03013: RNA transport
ko04621: NOD−like receptor signaling pathway
ko03008: Ribosome biogenesis in eukaryotes
ko04113: Meiosis − yeast
ko04210: Apoptosis
ko05143: African trypanosomiasis
ko04622: RIG−I−like receptor signaling pathway
ko04962: Vasopressin−regulated water reabsorption
ko05131: Shigellosis
ko05100: Bacterial invasion of epithelial cells
ko00100: Steroid biosynthesis
ko05145: Toxoplasmosis
ko00901: Indole alkaloid biosynthesis
ko00196: Photosynthesis − antenna proteins
ko04614: Renin−angiotensin system
ko04115: p53 signaling pathway
ko05146: Amoebiasis
ko03050: Proteasome
ko05130: Pathogenic Escherichia coli infection
ko05322: Systemic lupus erythematosus
ko05110: Vibrio cholerae infection
ko00514: Other types of O−glycan biosynthesis
ko00513: Various types of N−glycan biosynthesis
ko00601: Glycosphingolipid biosynthesis − lacto and neolacto series
ko00941: Flavonoid biosynthesis
ko05012: Parkinson's disease
ko04976: Bile secretion
ko00906: Carotenoid biosynthesis
ko04141: Protein processing in endoplasmic reticulum
ko00510: N−Glycan biosynthesis
ko05010: Alzheimer's disease
ko00565: Ether lipid metabolism
ko04974: Protein digestion and absorption
ko00351: DDT degradation
ko05150: Staphylococcus aureus infection
ko04142: Lysosome
ko00140: Steroid hormone biosynthesis
ko00643: Styrene degradation
ko00791: Atrazine degradation
ko00364: Fluorobenzoate degradation
ko00592: alpha−Linolenic acid metabolism
ko00361: Chlorocyclohexane and chlorobenzene degradation
ko00590: Arachidonic acid metabolism
ko00062: Fatty acid elongation in mitochondria
ko01053: Biosynthesis of siderophore group nonribosomal peptides
ko03450: Non−homologous end−joining
ko04146: Peroxisome
ko03320: PPAR signaling pathway
ko00311: Penicillin and cephalosporin biosynthesis
ko00312: beta−Lactam resistance
ko00562: Inositol phosphate metabolism
ko00440: Phosphonate and phosphinate metabolism
ko04626: Plant−pathogen interaction
ko00830: Retinol metabolism
ko00785: Lipoic acid metabolism
ko00540: Lipopolysaccharide biosynthesis
ko00531: Glycosaminoglycan degradation
ko00511: Other glycan degradation
ko00900: Terpenoid backbone biosynthesis
ko00620: Pyruvate metabolism
ko00010: Glycolysis / Gluconeogenesis
ko00270: Cysteine and methionine metabolism
ko00400: Phenylalanine, tyrosine and tryptophan biosynthesis
ko03060: Protein export
ko03440: Homologous recombination
ko04112: Cell cycle − Caulobacter
ko00983: Drug metabolism − other enzymes
ko03430: Mismatch repair
ko03010: Ribosome
ko00710: Carbon fixation in photosynthetic organisms
ko00340: Histidine metabolism
ko00051: Fructose and mannose metabolism
ko00521: Streptomycin biosynthesis
ko00780: Biotin metabolism
ko00670: One carbon pool by folate
ko00250: Alanine, aspartate and glutamate metabolism
ko00550: Peptidoglycan biosynthesis
ko00730: Thiamine metabolism
ko00300: Lysine biosynthesis
ko00061: Fatty acid biosynthesis
ko00770: Pantothenate and CoA biosynthesis
ko00030: Pentose phosphate pathway
ko00471: D−Glutamine and D−glutamate metabolism
ko00970: Aminoacyl−tRNA biosynthesis
ko00473: D−Alanine metabolism
ko00290: Valine, leucine and isoleucine biosynthesis
ko00660: C5−Branched dibasic acid metabolism
Turquoise
Brown
Yellow
Green
Blue

Figure 3-4. Variations of KEGG metabolic pathways in the functional microbial communities. The heatmap shows the functional profiles of FMCs (columns) based on the relative abundance of KEGG metabolic pathways (rows) after *z* score transformation. The color bar on top shows module membership. The dendrograms show the hierarchical clustering of columns and rows respectively using Euclidean distance.

CHAPTER 4

Reprograming of Gut Microbiome Energy Metabolism by the *FUT2* Crohn's Disease Risk Polymorphism

Abstract

Background

Fucosyltransferase 2 (*FUT2*) is an enzyme that is responsible for the synthesis of the H antigen in body fluids and on the intestinal mucosa. The H antigen is an oligosaccharide moiety that acts as both an attachment site and carbon source for intestinal bacteria. Non-secretors, who are homozygous for the loss-of-function alleles of *FUT2* gene (*sese*), have increased susceptibility to Crohn's disease.

Methods

To characterize the effect of *FUT2* polymorphism on the mucosal ecosystem, we profiled the microbiome, meta-proteome and meta-metabolome of 75 endoscopic lavage samples from the cecum and sigmoid of 39 healthy subjects (12 *SeSe*, 18 *Sese* and 9 *sese*). To investigate the change of microbial ecology on the intestinal mucosal surface and its role in the pathogenesis of IBD, 16s ribosomal DNA (rDNA) extracted from the bacterial pellets were sequenced by Illumina HiSeq 2000 to determine the phylogenetic distribution of the microbiomes. The gene content of 1,119 KEGG reference genomes was used to infer the approximate gene content of the detected 97% OTUs in our dataset using Phylogenetic Investigation of Communities by Reconstruction of Unobserved States (PICRUSt) (v0.1), and the metabolic pathways were re-constructed using HUMAnN (v0.98). To profile the meta-proteome of lavage samples, matrix-assisted laser desorption/ionization time-of-flight mass spectrometry (MALDI-TOF MS) was performed using the soluble fraction of the lavage samples. The soluble metabolites of the same samples were analyzed using quadrupole time-of-flight (Q-TOF) mass spectrometry.

Results

Imputed metagenomic analysis revealed perturbations of energy metabolism in the microbiome of non-secretor and heterozygote individuals, notably the enrichment of carbohydrate and lipid metabolism, cofactor and vitamin metabolism, and glycan biosynthesis and metabolism related pathways; and, the depletion of amino acid biosynthesis and metabolism. Similar changes were observed in mice bearing the *FUT2*$^{-/-}$ genotype. Metabolomic analysis of human specimens revealed concordant as well as novel changes in the levels of several metabolites. Human metaproteomic analysis indicated that these functional changes were accompanied by sub-clinical levels of inflammation in the local intestinal mucosa.

Conclusion

The colonic microbiota of non-secretors is altered at both the compositional and functional levels, affecting the host mucosal state and potentially explaining the association of *FUT2* genotype and Crohn's disease susceptibility.

Introduction

The human intestinal microbiome contributes vital biological functions to healthy hosts, including maintenance of immune homeostasis, modulation of intestinal development, and enhanced metabolic capabilities (8, 9, 12, 87, 88). Dysbiosis, which refers to perturbations of the normally stable intestinal microbiota, has been associated with the development and progression of many conditions, including inflammatory bowel diseases (IBD) (16-18, 89), type 2 diabetes (90), and obesity (12). The reasons for such associations are not yet clear, and may reflect either causal or secondary processes due to the impact on microbial composition and function of inter-individual variability and the contributions of environment and host genetics (19). The contributions of such factors on human microbial composition are beginning to emerge through dietary and environmental studies (18, 91-93), as well as twin studies (12, 93-95), respectively. However, much work is still necessary to fully understand the extent of host genetic influence on the composition and function of the gut microbiome, and the mechanisms linking these genetic traits with microbial function and disease biology.

A recent genome-wide association study (GWAS) published by our group identified *Fucosyltransferase 2 (FUT2)* gene as a Crohn's disease (CD) risk locus (96), a finding that has been validated in a meta-analysis of Crohn's disease and ulcerative colitis genome-wide association scans (20). However, the molecular mechanism of the association between non-secretor status and CD remains unknown. Mucin 2 (muc2), the predominantly secreted mucin in the colon, plays an important barrier role in intercepting and excluding bacteria from the mucosal cell surface, therefore reducing host susceptibility to colitis (97-99). Research from our group has shown that aberrant glycosylation of muc2 core proteins causes spontaneous colitis in mice (4). Both the core 1- and core 3-derived *O*-glycans of mucin core proteins are terminally fucosylated, which serve as interceptive binding structures for bacteria (100). Moreover, a subset of human intestinal microbiota produce glycosidases capable of hydrolyzing α-1,2-fucosyl linkages present in various mucin-type glycoproteins, as well as mucus glycan structures that are not capped by fucose (101). A mass spectrometry based analysis of insoluble colonic mucin of both *Fut2*-null and wild type mice (102) identified 17 different oligosaccharides with up to eight sugar residues of which 11 were neutral, five sulfated and one sialylated. The most abundant structures were composed of core 2 (Galβ1-3(GlcNAc β1-6)GalNAc-) glycan sequence with some based on core 1 (Galβ1-3GalNAc-) glycan structures. The primary difference in oligosaccharides was the presence of terminal fucose residues forming the blood group H-type epitope in most of the oligosaccharides in wild type mice. In contrast, all the peaks for oligosaccharides carrying blood group H type epitopes were absent in the *Fut2*-null mice. Therefore, *FUT2* deficiency may alter the composition of intestinal microbiota by affecting either microbial adhesion and/or utilization of host derived glycans, potentially leading to dysbiosis.

The phylogenetic composition in non-secretor individuals have been characterized in several studies recently (40, 103, 104), showing that the *FUT2* genotype was associated both with deviations of overall community ecology and with altered abundances of specific microbes. However, these descriptions did not address the degree to which these alterations were functional, nor their potential mechanisms of action in IBD risk. Both questions are of particular interest because microbial composition can exhibit large inter-individual variations compared to function-based analyses even in healthy individuals (9). This may also be one of the reasons for existing between-study discrepancies (40, 103), other than the difference between measurements of the luminal/fecal microbiota and those at the mucosal surface (21). Since bacterial colonization largely occurs in the outer mucous layers (98) where the residual glycans that fuel bacterial growth are degraded, lavage sampling of the mucosal surface compartment coupled with functional and metabolic profiling is arguably more biologically relevant to host-microbial glycan metabolism.

We present here a comprehensive description of the mucosal luminal interface of healthy individuals distinguished by secretor status, capturing multiple aspects of the microbial ecosystem including microbiome composition, imputed function, metabolome and proteome. 16S rRNA gene sequencing can be used to characterize the composition and diversity of the microbiota, and with recent advances it allows us to impute functionality of the microbiome. Deeper insight into microbial functionality can be provided by combining 16S rRNA gene sequencing with proteomic and metabolomic data (105). We detailed the phylogenetic and functional profiles of the mucosal microbiome associated with *FUT2* polymorphism, indicating a strong effect of host genetics on the re-programing of energy metabolisms into dysbiotic setting. The combination of multi’omic analysis also provided us with unprecedented understanding of the dynamics of host-microbial interaction.

Materials and Methods

Subject Cohort and Lavage Sample Collection

A subject cohort of 131 individuals (Table 4-1) was recruited from patients presenting for screening colonoscopy at Cedars-Sinai Medical Center. Following institutional review board approval, subjects were consented and then included in the study if colonoscopy did not reveal any mucosal abnormalities. Enrolled subjects were prepared for colonoscopy by taking Golytely the day before the procedure. The mucosal lavage samples representing the mucosal luminal interface were collected from cecum and sigmoid colon as described previously (22).

Animals

All animal protocols were in accordance with Administrative Panel on Laboratory Animal Care, the Stanford Institutional Animal Care and Use Committee. Conventionally housed *Fut2-/-* mice (B6.129X1-Fut2tm1Sdo/J; backcrossed 12 generations with C57BL/6J) were re-derived as germ free (GF) and maintained in gnotobiotic isolators. Eight week old non littermate GF Fut2-deficient *Fut2*$^{-/-}$ (n=10), wild-type *Fut2*$^{+/+}$ (n=10; C57BL/6J) and heterozygous *Fut2*$^{+/-}$ (n=8) mice were colonized with feces obtained from a healthy human donor (secretor) by oral gavage of 200µl of human fecal sample (male, age 38, American diet). The sample was prepared by mixing stored frozen human fecal sample with filter-sterilized pre-reduced phosphate buffered saline. Mice were singly housed and maintained in gnotobiotic isolators on a strict 12h light cycle for the experiment. Mice were fed a standard autoclaved mouse diet (Purina LabDiet 5K67).

Fecal samples were collected 4 weeks after humanization for 16S rRNA gene sequencing using the 454 titanium platform.

Genotyping

The SNP rs601338 (G > A) defines secretor status in Europeans and Africans (106). We used rs516246, which is in strong linkage disequilibrium with rs601338, to infer secretor status. Estimate of linkage between rs516246 and rs601338 is 100%. These 2 SNPs tag each other perfectly as they are in perfect LD with one another, with R-square of 1.0 and D-prime of 1.0, according to Hapmap3 release 2, CEU population. The individuals with the homozygote A/A genotype are defined as nonsecretors. In this cohort, 97% (38/39) of the subjects are Caucasian. One subject is African American, who is heterozygous G/A genotype for rs516246, and therefore categorized as *Sese*. Mouse genomic DNA was prepared from ear tissue obtained by ear punch. PCR amplification using three primers (F: 5'CCTGCCATGCTTTCTTTCCTG3' R: 5'ATTCCTTCTCTGACAGGGTTTGG3' (WT), 5' TGGGTAACGCCAGGGTTTTC3' (KO)) yielded either a 191bp band (*Fut2*$^{-/-}$) or 154 bp band (*Fut2*$^{+/+}$) or both (*Fut2*$^{+/-}$).

16S rRNA Gene Sequencing and Microbial Composition Analysis

Genomic DNA was extracted as previously described (3). The V4 region of 16S ribosomal RNA genes were amplified and sequenced on an Illumina HiSeq 2000 as previously described (3). HiSeq reads were processed using QIIME v1.5.0 (27) with parameters of: minimum Q-score considered high quality: 20, maximum number of consecutive low-quality base calls allowed before truncating: 3, maximum number of N characters allowed: 0. All filtered reads had a length of 101 bp. The number of reads per sample ranged from 326,481 to 1,021,473, with a mean of 646,140 and totaling 48,460,491. Sequence sub-sampling was performed for each sample at the depth of 300,000 reads/sample. This normalized dataset was used for all the following analysis including alpha-diversity analysis, beta-diversity analysis, and imputed metagenomic analysis. For mouse fecal pellets, after DNA isolation (MoBio fecal DNA kit), 626 bp amplicons spanning 16S variable regions 3-5 (V3-V5) were generated using barcoded forward primer (338F, 906R) (107). Samples were sent for pyrosequencing to Duke ISGP using the Roche 454 titanium platform. Operational taxonomic units (OTUs) were picked against the February 4th, 2011 version of the Greengenes database (http://greengenes.lbl.gov/cgi-bin/nph-index.cgi) (26), pre-filtered at 97% identity. For quality control, all the singletons were removed. After reference-based OTU picking, 97.5% of the total reads were successfully mapped to the reference Greengenes database. These steps were performed using QIIME v1.5.0 (27). Alpha rarefaction was performed using the number of observed species, Chao1 and phylogenetic diversity. The comparison of alpha diversity between two groups at certain sampling depths was performed using a two-sided Student *t* test. Beta diversity of 16S rRNA gene and imputed metagenomic datasets was estimated by computing unweighted UniFrac and Bray-Curtis distances between samples respectively using QIIME. Ordination of the resulting distance matrix was performed using principal coordinate analysis (PCoA). Pair-wise comparisons between *SeSe*, *Sese* and *sese* individuals were conducted using the Kruskal-Wallis test to identify differentially abundant bacterial phylotypes at phylum and 97% OTU levels. Multiple hypothesis tests were adjusted to produce a final Benjamini and Hochberg false discovery rate (108), and significant association was considered below a FDR q-value threshold of 0.25.

Imputation of microbial gene content and metagenomes

This study takes advantage of PICRUSt, a program that infers the metagenome of a sample from its phylogenetic composition and was recently validated against conventional deep-sequencing metagenomics (109). The OTU table including all 75 samples was used as the input file for metagenome imputation of individual human and mouse samples. For the metagenomic profiling of FMCs, the OTU table of the 6 FMCs was generated with 1 count for each 97% OTU in a given FMC. The gene content of 2,590 KEGG (Kyoto Encyclopedia of Genes and Genomes) reference genomes was used to infer the approximate gene content of the detected phylotypes using PICRUSt (v0.1) (http://picrust.sourceforge.net/) (109). The program output the inferred metagenome represented by KEGG Orthology (KO) for each FMC. Taking the PICRUSt KO gene abundance inferences as inputs, the metabolic pathways were re-constructed using HUMAnN (v0.98) (55). We restricted our analysis to the KEGG pathways that were present in at least 90% of the samples. Pair-wise Kruskal-Wallis tests between *SeSe*, *Sese* and *sese* individuals were performed to identify imputed KEGG pathways with differential relative abundance. Multiple hypothesis tests were adjusted to produce a final Benjamini and Hochberg false discovery rate (108), and significant association was considered below a FDR q-value threshold of 0.05.

Mass spectrometry analysis

For metabolomics analysis, each human lavage samples was subjected to solid-phase extraction to eliminate a polymeric contaminant believed to originate from the lubricant used during colonoscopy preparation. The eluate was dried and reconstituted in 2% acetonitrile in water prior to MS analysis. A 5 μL aliquot of extracted metabolites from each sample was injected onto a reverse-phase 50 × 2.1 mm ACQUITY 1.7-μm C18 column (Waters Corp, Milford, MA) using an ACQUITY UPLC system (Waters Corp, Milford, MA). A Waters Q-TOF Premier was operated in negative-ion (ESI-) or positive-ion (ESI+) electrospray ionization mode with a capillary voltage of 3200 V and a sampling cone voltage of 20 V in negative mode and 35 V in positive mode. Data were acquired in centroid mode with a mass window of 50 to 850 *m/z*, and processed using MassLynx software (Waters Corp, Milford, MA). To profile the meta-proteome of lavage samples, matrix-assisted laser desorption/ionization time of flight mass spectrometry (MALDI-TOF MS) was performed using the soluble fraction of samples as previously described in (22). The abundances of metabolomic and proteomic peaks were compared using ANOVA to identify features associated with *FUT2* genotype. Multiple hypothesis tests were adjusted to produce a final Benjamini and Hochberg false discovery rate (108), and significant association was considered below a FDR q-value threshold of 0.25. The relatively high FDR was used to avoid excessively strict filtering of metabolomics and proteomic features.

Results

Whole-community microbiome ecology differs according to secretor status

To study the host-microbial interaction at the mucosal luminal interface, 75 lavage samples were collected from the cecum and sigmoid colons of 39 healthy subjects (Table 4-1). We assessed differences in overall microbial ecology between secretors (both homozygous *SeSe* and heterozygous *Sese* for the functional allele) and non-secretors (*sese*).

We first examined the microbial composition in these samples, to affirm that the present cohort matched previously reported differences between secretors and non-secretors in microbial composition (40, 103). The microbiota from these samples was profiled by multiplex sequencing, and a total of 47,171,132 reads (628,948 ± 130,744 s.d. reads per sample) were generated after quality control. 4,074 OTUs were then identified by grouping reads at a 97% sequence similarity threshold. Compared with *SeSe* individuals, both *Sese* and *sese* individuals exhibited lower alpha-diversity based on the number of observed species (*t* test, *P* = 0.012 and 0.085 respectively), although the difference between *SeSe* and *sese* individuals was not statistically significant (Figure 4-1 A). We also measured other diversity indexes including Chao1 and phylogenetic diversity. Compared with *SeSe* individuals, *Sese* individuals exhibited significantly lower alpha-diversity as indicated by Chao1 and phylogenetic diversity indexes at the depth of 300,000 reads per sample (*t* test, *P* = 0.019 and 0.02 respectively). The same trend was observed in *sese* individuals, although not statistically significant (*t* test, *P* = 0.10 for Chao1 and 0.18 for phylogenetic diversity) (Figure 4-1 B and C).

The beta-diversity measured by unweighted UniFrac distance matrix was calculated for each sample to evaluate the similarity between microbial communities.

Principal coordinate analysis (PCoA) demonstrated that the phylogenetic compositions of *SeSe* microbiomes were significantly different from those of *Sese* (Adonis test, *P* = 0.016), but not *sese* individuals (Adonis test, *P* = 0.092) (Figure 4-2 A). The significant difference in phylotype abundances reported previously, namely the increase in Bacteroidetes among nonsecretors, was confirmed at the phylum level (40) (Figure 4-2 B).

To analyze at lower taxonomic levels, we filtered out low-abundant 97% OTUs based on the criteria of 1) minimum total observation count of 30 across all samples and 2) being observed in at least 60% of the samples, reducing the number of OTUs from 4,074 to 419. Among these OTUs, 19 (4.5%) of them were depleted in *Sese* and *sese* compared to *SeSe* individuals (Kruskal-Wallis, FDR q < 0.25). In summary, the *FUT2* polymorphism was significantly associated with selected phylotypes of colonic microbiota in *Sese* and *sese* individuals, and the alterations in *Sese* individuals resulted in a significant shift of microbial composition compared to *SeSe*. These data revealed the gardening effect of *FUT2* polymorphism on phylogenetic composition of the colonic microbiota.

Non-secretor associated functional changes revealed by imputed metagenomes

We hypothesized that these compositional changes result in selectively augmented or deficient functional capabilities that may be relevant to Crohn's disease susceptibility. To test this idea, we inferred the metabolic capacities of mucosal microbiota associated with secretor status, using a recently developed bioinformatic pipeline centering on the PICRUSt (109) and HUMAnN tools (55). In the 'gene content inference' step, the gene contents and 16S rRNA gene copy number of the detected phylotype were predicted based on its evolutionary similarity with the 1,119 KEGG (Kyoto Encyclopedia of Genes and Genomes) reference genomes. In the subsequent 'metagenome inference' step, the resulting gene content predictions for all microbial taxa with the relative abundance of 16S rRNA genes in each samples are combined, corrected for expected 16S rRNA gene copy number, to generate the expected abundances of gene families in the entire community represented by KEGG Orthology (KOs). The prediction accuracy of PICRUSt has been validated using human and mammalian gut microbiome with paired 16S rRNA gene and metagenome sequencing data (109). The relative abundances of KEGG pathways in each sample were then reconstructed by mapping KOs to these pathways using HUMAnN. At our routine sampling depth, both *Sese* and *sese* individuals harbored 15% fewer microbial genes on average than *SeSe* individuals (Figure 4-3 A), which is consistent with the significant lower compositional diversity observed in *Sese* individuals compared to *SeSe* individuals.

The similarity of functional states of the microbiomes from secretors and non-secretors was evaluated by the composition of imputed metagenomes. Principal coordinate analysis (PCoA) using Bray–Curtis distance demonstrated separation of the samples from *SeSe*, *Sese* and *sese* individuals along PC1. The clustering of *SeSe* was significant compared to both *Sese* and *sese* individuals (Adonis test, *P* = 0.004 and 0.004 respectively), suggesting that the variations at metagenomic level were more profound than those at the compositional level, and that *FUT2* exhibited haploinsufficiency in programing the metagenomic functions (Figure 4-3 B). In support of the distinction between the phylogenetic and imputed metagenomic datasets, Procrustes analysis of the

Bray-Curtis PCoA plots derived from 16S rRNA gene and imputed metagenome datasets showed that the clustering of samples across datasets was not significant (*P* = 0.510). Among the 154 imputed metabolic pathways, 23 (14.9%) were differentially abundant between *SeSe* and *Sese* individuals. The alterations in *sese* individuals were greater, as shown by the changes of abundances of 43 (27.9%) of the imputed pathways compared to *SeSe* individuals (Kruskal-Wallis, FDR q < 0.05).

The *FUT2*-associated changes were more robust at the imputed metagenomic versus the phylotypic level. As compared to *SeSe* individuals, a diverse consortium of metabolic functions represented by 27 KEGG pathways were depleted in *sese* individuals, including amino acid metabolism pathways, cofactors and vitamins pathways and genetic information processing pathways. A broad-based decrease in amino-acid biosynthesis was observed, including lower abundances of lysine (KO00300), valine, leucine and isoleucine (KO00290), and phenylalanine, tyrosine and tryptophan (KO00400) biosynthesis pathways. Accompanying the depletion, 16 microbial pathways were enriched in these individuals, highlighted by aspects of energy metabolism including carbohydrate and lipid metabolism, cofactors and vitamins metabolism, and glycan biosynthesis and metabolism. These data suggest that *FUT2* gene polymorphism acted in a haploinsufficient manner to perturb metabolic pathways such as amino acid biosynthesis encoded by the gut microbiome at the mucosal interface.

To validate these findings, we performed the same analysis on a 16S rRNA gene dataset of the fecal samples collected from humanized *FUT2*$^{-/-}$ mice (germ free *FUT2*$^{-/-}$ mice colonized with human feces from a healthy secretor) (107). Among the 146 metabolic pathways reconstructed, 47 (32.2%) of them were differentially abundant between the *FUT2*$^{+/+}$ and *FUT2*$^{-/-}$ mice, comparable to our findings with the human samples. After cross comparing the two datasets, we identified 13 pathways that were consistently enriched or depleted with *FUT2* haploinsufficiency (Kruskal-Wallis, FDR $q < 0.05$) (Figure 4-3 C). Specifically, the carbohydrate and lipid metabolism and glycan biosynthesis related pathways were over-represented in *Sese/sese* individuals and *FUT2*$^{+/-}$/*FUT2*$^{-/-}$ mice, whereas the relative abundances of 5 amino-acid and vitamin metabolism related pathways were enriched in *SeSe* individuals and *FUT2*$^{+/+}$ mice (Figure 4-3 C). These findings suggest that *FUT2* genotype had a similar impact on imputed microbial metagenomic functions in humans and mice, notably in reduced amino acid synthesis capabilities.

Functional microbial communities associated with non-secretor status

It is well-understood that even healthy individuals differ remarkably in their fecal and mucosal surface gut microbial composition, especially at the genus and species level. Mucosal microbiota can be further assessed for functional relatedness based on their co-occurrence patterns (50, 110). To determine whether such ecological structures can be observed in this dataset, we developed a methodology to infer microbial co-occurrence networks. Nodes (OTUs) of these networks were grouped based on their topological overlaps using hierarchical clustering. Using this approach, six modules, ranging from 39 to 97 OTUs, were identified. These modules of OTUs represent putatively key ecological features, which we term as functional microbial communities (FMC). Multi-dimensional scaling was used to depict module structure and network connections (Figure 4-4 A).

Phylogenetically related OTUs were clustered into the same FMC, presumably because preferences for ecological niche are more likely to be shared between more closely related microbes. However, each FMC also included phylogenetically distinct OTUs from different phyla, suggesting that in addition to phylogenetic relatedness, the formation of FMC depended upon additional ecological affinities, which could range from syntrophic dependences to convergent functionality between distinct phylotypes.

The "abundance" of each FMC can be quantitated by defining the OTU abundance profiles of each FMC. One can thus correlate the abundance of the FMCs with metadata including host genotype, disease phenotype, age, *etc*. When using an additive genetic model, the abundances of turquoise and blue FMCs significantly associated with the copy number of the *FUT2* loss-function-allele reciprocally (*P* = 0.04 and 0.05 respectively with rs516246 by Pearson correlation) (Figure 4-4 B). This observation was concordant with results at the individual OTU level: when examining the membership of the FMCs, we found that 12 of the 19 OTUs enriched in *SeSe* individuals were assigned to the blue FMC. Gender was another subject phenotype that significantly associated with microbial composition: the turquoise (*P* = 0.004, Pearson correlation) FMC was enriched in females, whereas the blue and red FMCs (*P* = 0.04 and 0.006 respectively, Pearson correlation) were more abundant in males. The gender effect on intestinal microbiome has also been reported previously in humans and murine model (111, 112). Due to the difference of male/female ratio in *SeSe*, *Sese* and *sese* individuals, the gender effect could potentially contribute to the association of turquoise and blue FMCs with *FUT2* genotype.

To determine whether these co-occurring microbial communities represented distinct functional units at the mucosal surface, we profiled the metabolic capabilities of FMCs using the approximate gene contents imputed previously. After aggregating the individual inferred genomes according to module membership, the relative abundances of metabolic pathways in each FMC were re-constructed. The functional profiles of FMCs were highly variable (Figure 4-5). The pathways associated with *FUT2* clustered into two groups that were overrepresented in the *FUT2* loss-of-function allele-associated turquoise FMC and functional allele-associated blue FMC respectively. Moreover, the pathways enriched in *SeSe* individuals tended to cluster into amino acid metabolism class and cofactors and vitamins metabolism class, whereas the pathways enriched in *Sese* and *sese* individuals highlighted amino acid metabolism, lipid metabolism and biosyntehesis of secondary metabolites (Figure 4-5). These metabolically specialized microbial communities were therefore responsible for the imputed metagenomic alterations associated with *FUT2* polymorphism.

Non-Secretor Associated Metagenomic Changes Reflected by Metabolomic and Proteomic Profiling

To determine if the imputed metagenomic alterations associated with *FUT2* polymorphism correlate with changes in metabolic activities of mucosal microbiota, we profiled the soluble metabolites of the same lavage samples using Q-TOF MS. The analysis generated a rich metabolomic dataset consisting of 649 and 576 spectral features in the cecum and sigmoid regions, respectively. Putative IDs were assigned to 372 ions by comparing their *m/z* values to those available in online databanks using a predefined mass error window of 20 ppm. The putative IDs were then used to map out the ions to various metabolomic pathways in the KEGG dataset. In accord with our previous study, ~50% of all metabolites were located at the terminal end of metabolic pathways, suggesting enrichment for end-products (3). In the cecum, 48 metabolites were mapped to the 13 KEGG pathways associated with non-secretor individuals in the human and murine datasets, and were present in more than 90% of the samples. Of these, 13 (27.1%) were differentially abundant among secretors (*SeSe* and *Sese*) and non-secretors (*sese*) (ANOVA, FDR q < 0.25). In the sigmoid, 30 metabolites were mapped to the non-secretor associated pathways; 4 (13.3%) were differentially abundant (ANOVA, FDR q < 0.25) (Figure 4-6 A). A less stringent q value, up to 0.25, was used to avoid missing significant associations, as shown in recent comparable study designs (18, 92). When using more stringent q-value threshold (FDR q-value < 0.1), these metabolites did not show significant association, which is comparable with the results from study with similar design (92). Thus, differences in imputed microbial metagenome content corresponded to abundances of metabolic end-products directly detected in the same samples.

To determine how the host reacted to the changes of the functional state of the microbiota, we also profiled the proteomic features of the same samples using MALDI-TOF MS. We focused our analysis on peaks of human origin. Of the 453 peaks included, the abundances of 16 (3.5%) were significantly different among secretors and non-secretors (ANOVA, FDR q < 0.25). Although only 17 (3.8%) molecular features could be identified, we found that the expression levels of human neutrophil peptide 1 and 2 (HNP-

1 and -2) were significantly higher in non-secretors, consistent with a subclinical inflammatory state at the mucosal surface (Figure 4-6 B). This difference could be driven by one particular *sese* sample that had high expression of HNP 1 and 2 relative to the other samples (Figure 4-6 B). To exclude such possibility, we repeated the analysis excluding this sample, and found that the levels of HNP1 and HNP2 were still significantly higher in non-secretors based on nominal *P* values (ANOVA, nominal *P* = 0.007 and 0.047 respectively), although FDR q values were higher than 0.25 (FDR q = 0.27 and 0.49 respectively). Our clinical records did not indicate that this individual had GI symptoms or chronic inflammation in the intestine at the time of sampling. These data suggest that the non-secretor state of the mucosa, *via* alteration of the mucosal surface ecosystem, changes the inflammatory state of the human intestinal mucosa.

Discussion

We combined 16S rRNA gene sequencing, metagenome imputation, meta-metabolomic, and meta-proteomic profiling, to delineate the integrated landscape of the mucosal surface ecosystem. The phylogenetic diversity and composition of intestinal mucosal microbiota in non-secretor individuals were significantly different from that of secretors. Compared with *SeSe*, the metabolic functions encoded and expressed by the gut microbiome in non-secretors were enriched for carbohydrate and lipid metabolism, cofactors and vitamins metabolism, and glycan biosynthesis and metabolism, and depleted for 7 pathways related to amino acid metabolism. These alterations in humans were highly consistent with analogous changes in the murine genetic counterpart.

Changes in the imputed metagenomes were reflected by concordant metabolite pools as determined by meta-metabolomic assays, providing validation for certain imputed metagenomic changes as functionally consequential. Moreover, these microbial functional changes were accompanied by sub-clinical intestinal inflammation. *FUT2* therefore appears to play a role in shaping the functional state of the mucosal surface by affecting not only microbial composition, but also the resulting functional state of the microbiota at the human intestinal mucosal surface.

It was surprising that difference with the *SeSe* group were in several cases (e.g., alpha-diversity) greater in the *Sese* rather than *sese* group. This raises two issues. One is the extent of glycan difference produced by haploinsufficiency. Since there was a strong haploinsufficiency phenotype in all facets of this study, we surmise that a substantial glycan change is produced by haploinsufficiency. However, to our knowledge, the glycan profile in heterozygous individuals has not been well described in the literature. The other issue is why the *Sese* group had a larger and more significant difference than the *sese* group. One possible explanation is that the inter-individual variation of gut microbial phylogenetic composition is inherently large (8, 11). In this context, it is notable that the sizes of the test groups were modest (9 for *sese*, 18 for *Sese*). It is possible that the particularly small size of the *sese* group made it particularly prone to outliers, and potentially underpowered for establishing mean phenotypes and robust statistical comparisons (e.g., the 95% confidence interval (CI) for alpha-diversity was larger for the *sese* (4.4) versus the *Sese* (1.8) group).

The imputed metagenome represents an accurate but approximate inference of the reference microbial genomes currently available. Potential bias may result from unmappable 16S rRNA gene reads and lack of sufficient reference genomes. After the reference-based OTU picking, 97.5% of the total reads were successfully mapped to the reference Greengenes database. The 2.5% unmappable 16S rRNA gene reads might cause the loss of metagenomic content that can be captured by shot-gun sequencing. Also, only the 2,590 microbial genomes that had identifiers in the Greengenes reference tree were used as the reference to predict unknown genomes. Despite these bias and limitations, PICRUSt predictions usually reach high agreement with metagenomically measured gene content (Spearman r = 0.8-0.9) (109). In this study, the findings of imputed metagenomic analysis in human subjects were validated by metabolomics and proteomic data as well as comparison to an independent 16S rRNA gene dataset from humanized *FUT2*$^{-/-}$ mice. In the future, adding meta-transcriptomic data will further enrich our understanding of microbial functional capability at any given time.

Inconsistencies of *FUT2* associated imputed metagenomic changes were also observed between human and murine datasets. In the humanized mouse gut microbiota, there are changes of microbial compositions as compared to the donor (113). The fecal sample used for humanization was from only one healthy donor, which might cause inherent bias due to the limited sample size. Also, taxonomic biases between the two different 16S rRNA gene datasets may exist due to different PCR primer sequences, amplicon lengths and sequencing technologies (114).

One of the key drivers of gut microbiota composition and function is the type and quantity of complex carbohydrates, which are typically derived from either diet or host mucus (92, 115). These polysaccharides serve as a primary metabolic input for the abundant carbohydrate-fermenting bacteria within the microbiota (116). However, different microbes within the gut are differentially endowed with abilities to use specific types of glycans (117), and so differences in carbohydrate availability, such as the presence or absence of fucose in mucosal glycans, translate into selective and regulatory events that result in discrete alterations in the microbiota's functional properties (107). Individual members of the microbiota can alter gene expression to accommodate the absence of fucose, while other members may be lost, and others recruited (107). Additionally a change in carbohydrate utilization by relatively few members can cascade into ecosystem-wide alterations given the interconnectedness of metabolic functions within the microbiome. Fucose processing by a gut resident symbiont allows expansion of pathogenic species such as *Salmonella typhimurium*. Similarly increased sialic acid release by certain bacterial species (*Bacteroides thetaiotaomicron*) allows expansion of *Clostridium difficile* (118). Differences in host glycan fucosylation result in distinct microbial ecosystems at the mucosal interface, the compositions and metabolic activities of which are likely precursors and predictors of ensuing disease phenotypes. This concept has been validated by the upstream role of microbial fucose-processing for the expansion of the enteric pathogen *Salmonella typhimurium* (118).

Among the phylotypes that are depleted in *Sese* and *sese* compared to *SeSe* individuals, *Roseburia* and *Faecalibacterium* are both short-chain fatty acid (SCFA) producing bacteria (81, 82), and reported to be anti-inflammatory (79, 119). Moreover, depletion of Firmicutes and expansion of Proteobacteria members are also characteristic of changes associated with the IBD microbiome (16, 120). Both *Sese* and *sese* individuals harbored 15% fewer microbial genes on average than *SeSe* individuals (Figure 4-3 A). The lower metagenome diversity is an unfavorable feature for the host, which has also been observed in inflammatory bowel disease (9) and obesity (121). Among the pathways that were consistently depleted or enriched in *Sese* and *sese* individuals, the increase in glutathione metabolism and decrease in amino acid biosynthesis (particularly lysine) pathways have been reported as a feature of the metagenome in IBD patients (18). These data indicated that the non-secretor associated changes of microbial composition and imputed metagenomic functions are also characteristics of other chronic inflammatory conditions, and therefore are unfavorable for the host.

FUT2$^{-/-}$ mice have a marked alteration in gastric mucosa glycosylation, characterized by diminished expression of alpha(1,2)fucosylated structures (122). Since gut microbes have developed the ability to degrade host derived glycans (101, 123), the deprivation of terminal fucosylation may affect the metabolic activity of the gut microbiota and thus its fermentation products potentially available to the host. Recent work reported that enterohaemorrhagic *Escherichia coli* encodes a two-component system, termed FusKR, which responds to fucose and controls metabolic gene expression (124). The imputed metagenomic changes in non-secretors highlighted the depletion of indispensable amino acid biosynthesis. This group of metagenomic functions complements that encoded by the host genome (9, 125, 126). In the IBD metagenome, amino acid biosynthesis and carbohydrate metabolism are reduced in favor of nutrient uptake (18). Such changes might reflect compensation by the microbiota for the lower availability of carbon sources. Amino acid starvation can lead to host stress response and the induction of autophagy of intestinal epithelial cells (127), which may increase risk of IBD. Although the current imputed metagenomic analysis is limited to the KEGG pathway level, further insights could be gained by extending the analysis to individual KEGG module or enzyme. It would be helpful to further identify the individual genes or reactions that are differentially abundant in these pathways, which would serve as potential candidates for therapeutic manipulation.

Low richness of gut microbiota is a well-known feature of patients with IBD (17, 32), and other chronic conditions such as obesity (12) and elderly patients with inflammation (91). A recent study defined two groups of individuals that differed by the number of gut microbial gene as low gene count (LGC) and high gene count (HGC) (121). LGC individuals exhibited an imbalance of pro- and anti-inflammatory bacterial species and evidence of low-grade inflammation. We have shown that both *Sese* and *sese* individuals harbored 15% fewer microbial genes on average than *SeSe* individuals, and therefore exhibited low metagenomic richness. Similarly, in *Sese* and *sese* individuals, genera associated with LGC individuals, including *Faecalibacterium* and *Coprococcus*, were also more dominant. Moreover, the subclinical inflammatory state at the mucosal surface was reflected by the higher expression levels of HNP-1 and -2 in non-secretors.

The data presented here supports the hypothesis that the *FUT2* loss-of-function allele increased the risk of Crohn’s disease by shaping the functional states of mucosal microbiota. Meta-analysis of genome-wide association studies (GWAS) has increased the number of confirmed IBD (both Crohn’s disease and ulcerative colitis) susceptibility loci to 167 (20), indicating that IBD is biologically heterogeneous. The analysis presented in this study focused on individuals without clinical symptoms. It will be important to extend the same analysis to patients with Crohn’s disease to determine to what extent the changes we present are recapitulated in the disease setting. In addition to *FUT2*, other risk genes have also been shown to affect the gut microbial composition, such as *NOD2* (128) and several defensin genes (29, 129, 130). It is currently unclear whether the microbiota associated with genes of similar functions has the same compositional and functional signatures. The stratification of gut microbiome by host genetics is a crucial step for elucidating the pathogenic mechanism of IBD as well as the design of personalized therapeutic interventions.

To achieve unprecedented understanding of the ecological structures and biomolecular activities of the gut microbiome, it is necessary to extend the analysis to multiple levels of biological organization – genome content, gene expression, protein expression, and metabolism (14, 86). In this study, we used multiple ’omic approaches to disentangle the complex host-microbial metabolic interplay. Meta-proteomic analysis in this case focused on the host side, but metabolomics could be extended to incorporate richer microbial analysis in the future. Although only a limited number of integrative 'omics profiles of the gut microbiota currently exist (105, 131), multi’omic studies have shown great potential in providing a holistic picture of the metabolic status of the gut microbiota and the host response to functional changes.

Figures and Tables

Figure 4-1

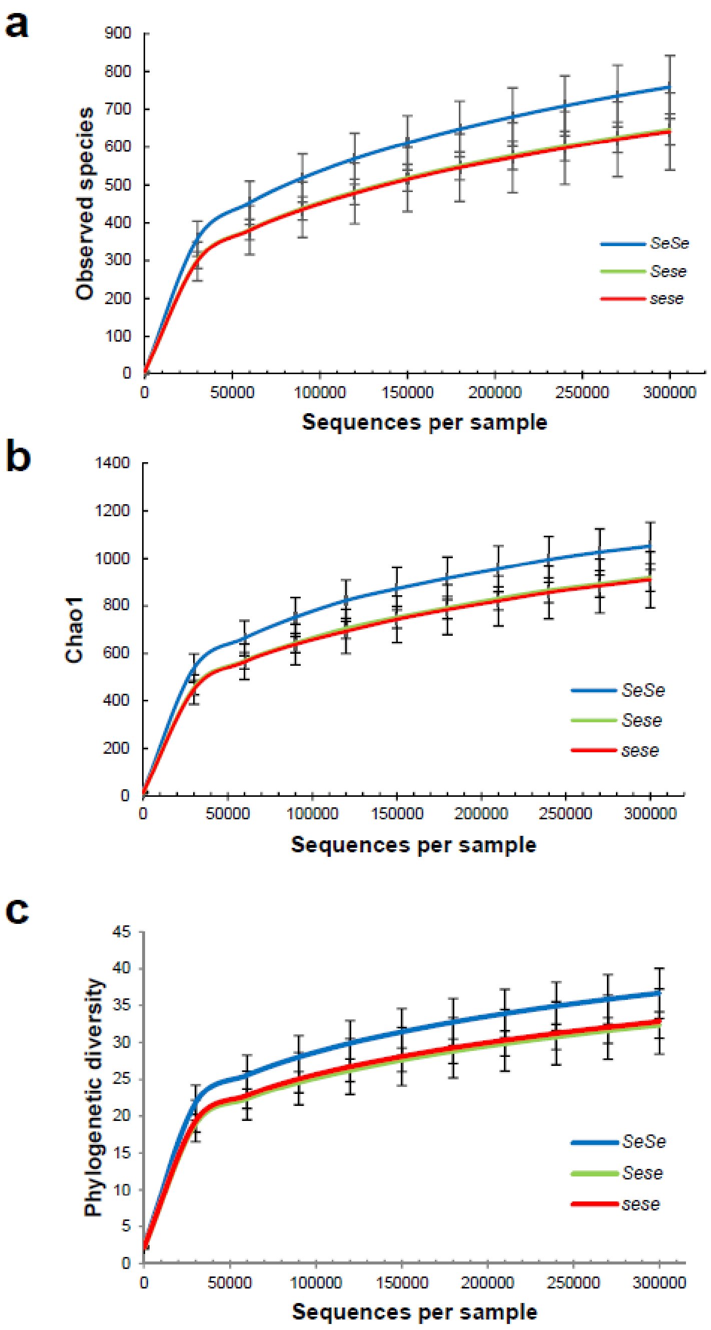


Figure 4-1. Rarefaction curve of microbial diversity for the microbiota from lavage samples. Rarefaction curves of (a) observed species, (b) Chao1 and (c) phylogenetic diversity for microbiota (mean ± 95% CI) were plotted at different sequencing depths.

Figure 4-2

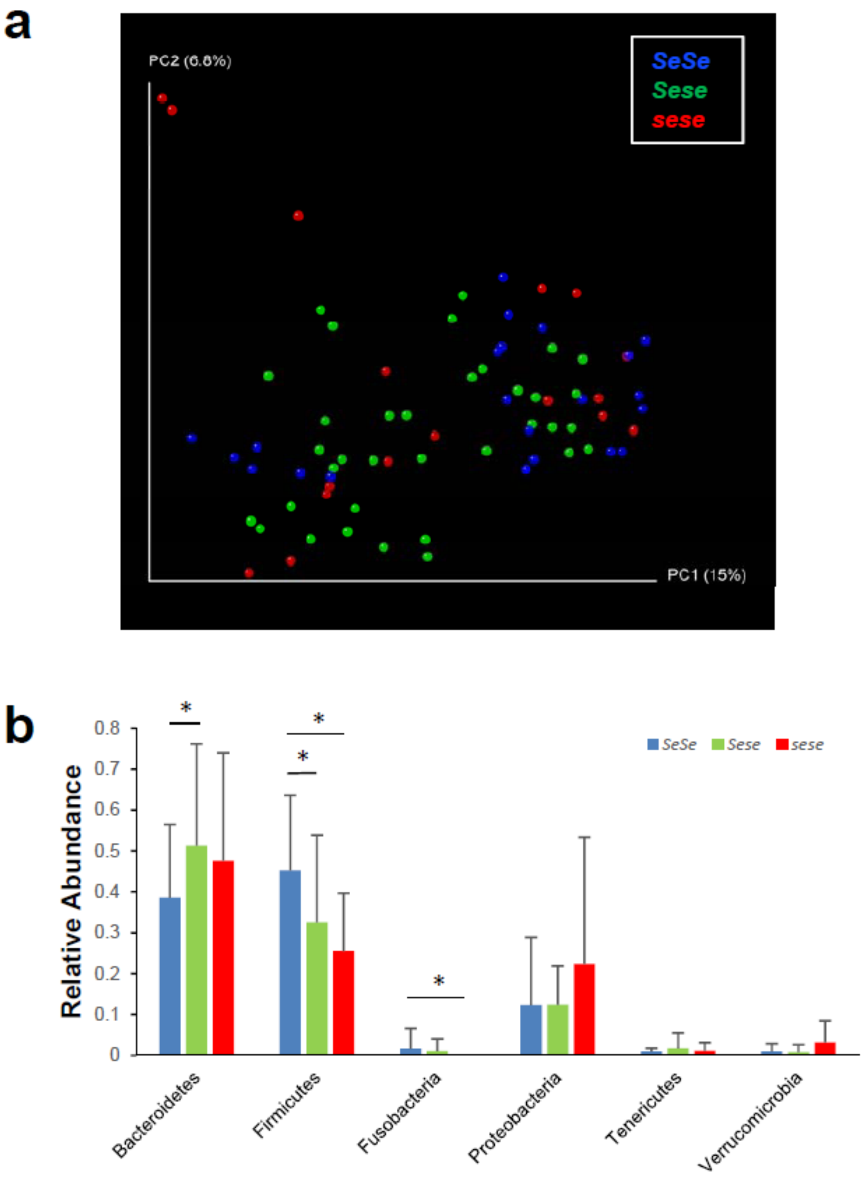


Figure 4-2. Shifts of mucosal microbial composition in secretor and non-secretor individuals. (a) Communities clustered using PCoA of the unweighted UniFrac distance matrix. Each colored point corresponds to a sample. **(b)** Phylum level microbial compositions of *SeSe*, *Sese* and *sese* individuals (mean ± s.d.). The bacterial community

from lavage sample mainly consisted of Bacteroidetes (46.59%), Firmicutes (34.8%), Proteobacteria (14.6%), Verrucomicrobia (1.4%) and Tenericutes (1.4%). Only predominant phyla with relative abundances higher than 1% were depicted in the bar graph. *: Kruskal-Wallis test, FDR-corrected *P* < 0.25. Shifts at the whole-phylum level were observed in *SeSe* individuals, including decreased relative abundances of Bacteroidetes, accompanied by increase of Firmicutes as compared with *Sese* individuals. When compared with *sese* individuals, *SeSe* subjects harbored more Firmicutes and Fusobacteria. In secretors, the trend of lower abundances of Proteobacteria can also be observed.

Figure 4-3

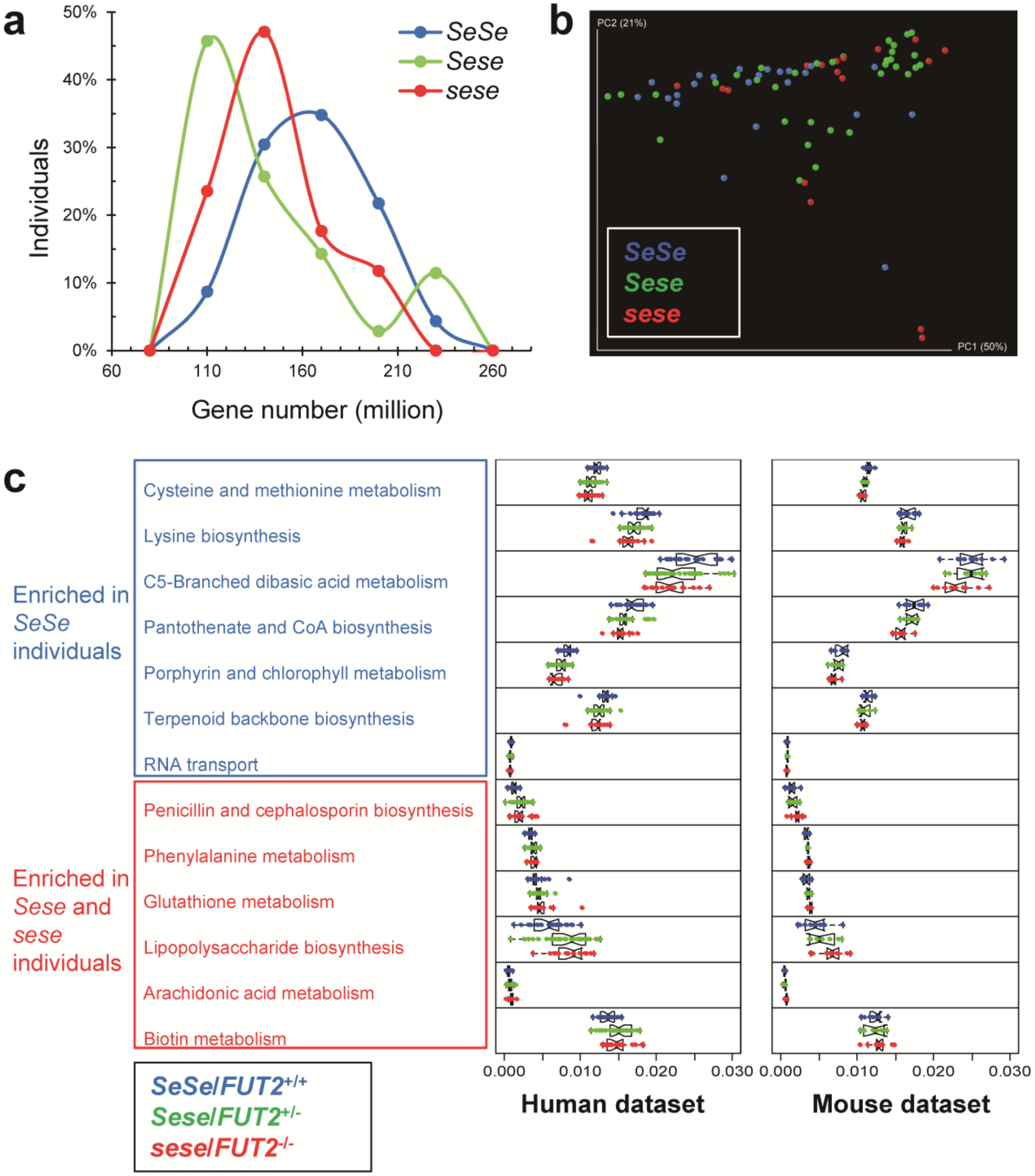
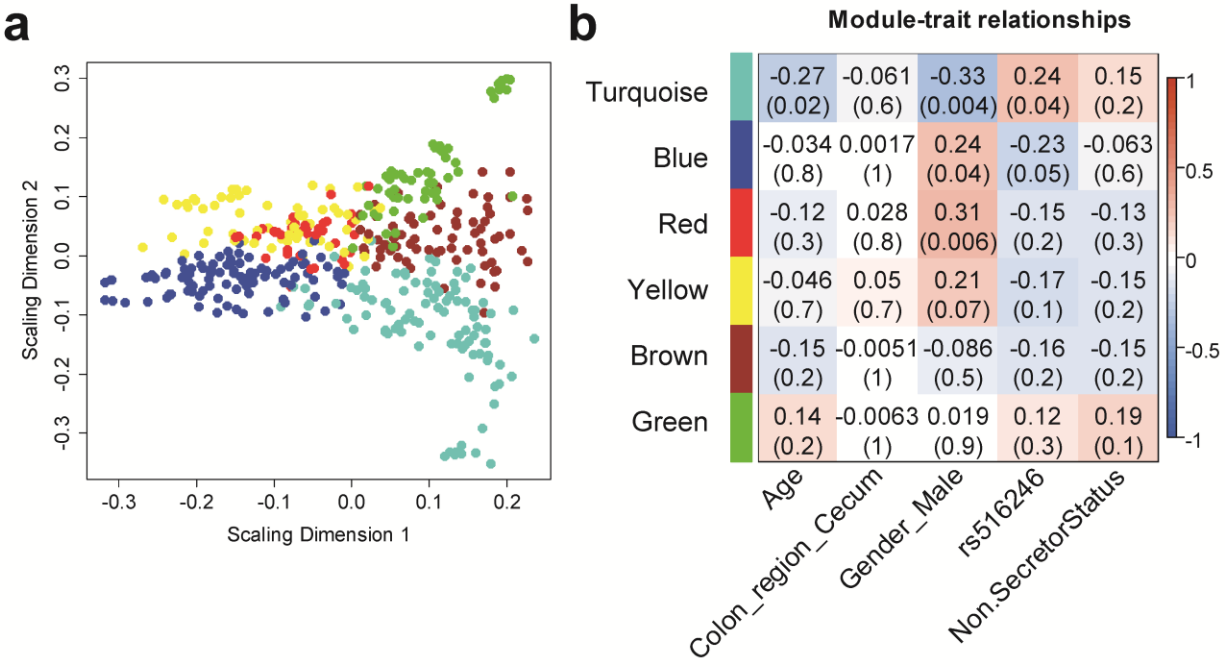

a
50%
40%
30%
20%
10%
0%
Individuals
60
110
160
210
260
Gene number (million)
SeSe
Sese
sese
b
PC2 (21%)
PC1 (50%)
SeSe
Sese
sese
c
Enriched in SeSe individuals
Cysteine and methionine metabolism
Lysine biosynthesis
C5-Branched dibasic acid metabolism
Pantothenate and CoA biosynthesis
Porphyrin and chlorophyll metabolism
Terpenoid backbone biosynthesis
RNA transport
Enriched in Sese and sese individuals
Penicillin and cephalosporin biosynthesis
Phenylalanine metabolism
Glutathione metabolism
Lipopolysaccharide biosynthesis
Arachidonic acid metabolism
Biotin metabolism
0.000
0.010
0.020
0.030
Human dataset
0.000
0.010
0.020
0.030
Mouse dataset
SeSe/FUT2+/+
Sese/FUT2+/-
sese/FUT2-/-

Figure 4-3. Imputed metagenomes reveal the significant enrichment of KEGG pathways in secretors and non-secretor individuals. (a) Distribution of bacterial genes in *SeSe, Sese, and sese* individuals. The proportion of individuals having a given number of genes was shown. (b) Communities clustered using PCoA of the Bray-Curtis distance matrix. Each colored point corresponds to a sample. The clustering of *SeSe* was significant compared to both *Sese* and *sese* individuals (Adonis test, *P* = 0.004 and 0.004 respectively). (c) Relative abundance of KEGG metabolic pathways in microbiome samples was colored by secretor status. Only the 13 pathways showing concordant alterations in both human and murine datasets were plotted.

Figure 4-4



Module-trait relationships

	Age	Colon_region_Cecum	Gender_Male	rs516246	Non.SecretorStatus
Turquoise	-0.27 (0.02)	-0.061 (0.6)	-0.33 (0.004)	0.24 (0.04)	0.15 (0.2)
Blue	-0.034 (0.8)	0.0017 (1)	0.24 (0.04)	-0.23 (0.05)	-0.063 (0.6)
Red	-0.12 (0.3)	0.028 (0.8)	0.31 (0.006)	-0.15 (0.2)	-0.13 (0.3)
Yellow	-0.046 (0.7)	0.05 (0.7)	0.21 (0.07)	-0.17 (0.1)	-0.15 (0.2)
Brown	-0.15 (0.2)	-0.0051 (1)	-0.086 (0.5)	-0.16 (0.2)	-0.15 (0.2)
Green	0.14 (0.2)	-0.0063 (1)	0.019 (0.9)	0.12 (0.3)	0.19 (0.1)

Figure 4-4. Functional microbial communities (FMCs) associated with non-secretor status. (a) Classical multi-dimensional scaling plot in which OTUs in each FMC represented by colored dots tend to form distinct clusters. **(b)** FMC-trait correlations and *P* values. Each cell reports the Pearson correlation coefficient (and *P* value) derived from correlating FMC eigenvectors (rows) to traits (columns). For the association with non-secretor status, the *SeSe* and *Sese* individuals were group together as secretor. For the association with *FUT2* genotype (rs516246), additive genetic model was used. The table was color-coded by correlation according to the color legend.

Figure 4-5

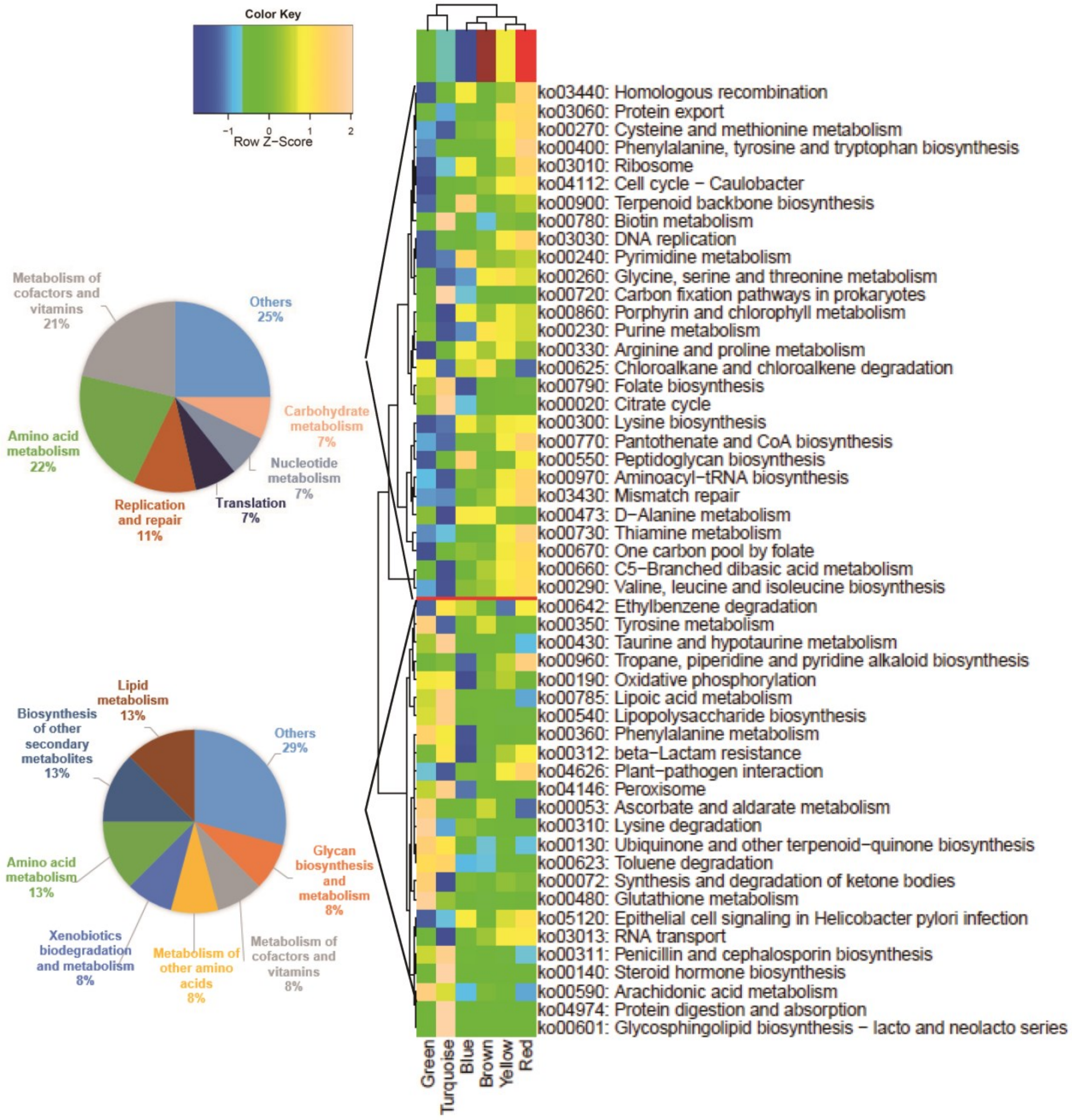


Figure 4-5. Variations of KEGG metabolic pathways in the functional microbial communities. The heatmap shows the functional profiles of FMCs (columns) based on the relative abundance of *FUT2* associated metabolic pathways (rows) after *z* score

transformation. The color bar on top shows module membership. The dendrograms show the hierarchical clustering of columns and rows respectively using Euclidean distance. The two pie-charts show the number of pathways in each functional class for the cluster associated with turquoise and blue FMC respectively.

Figure 4-6

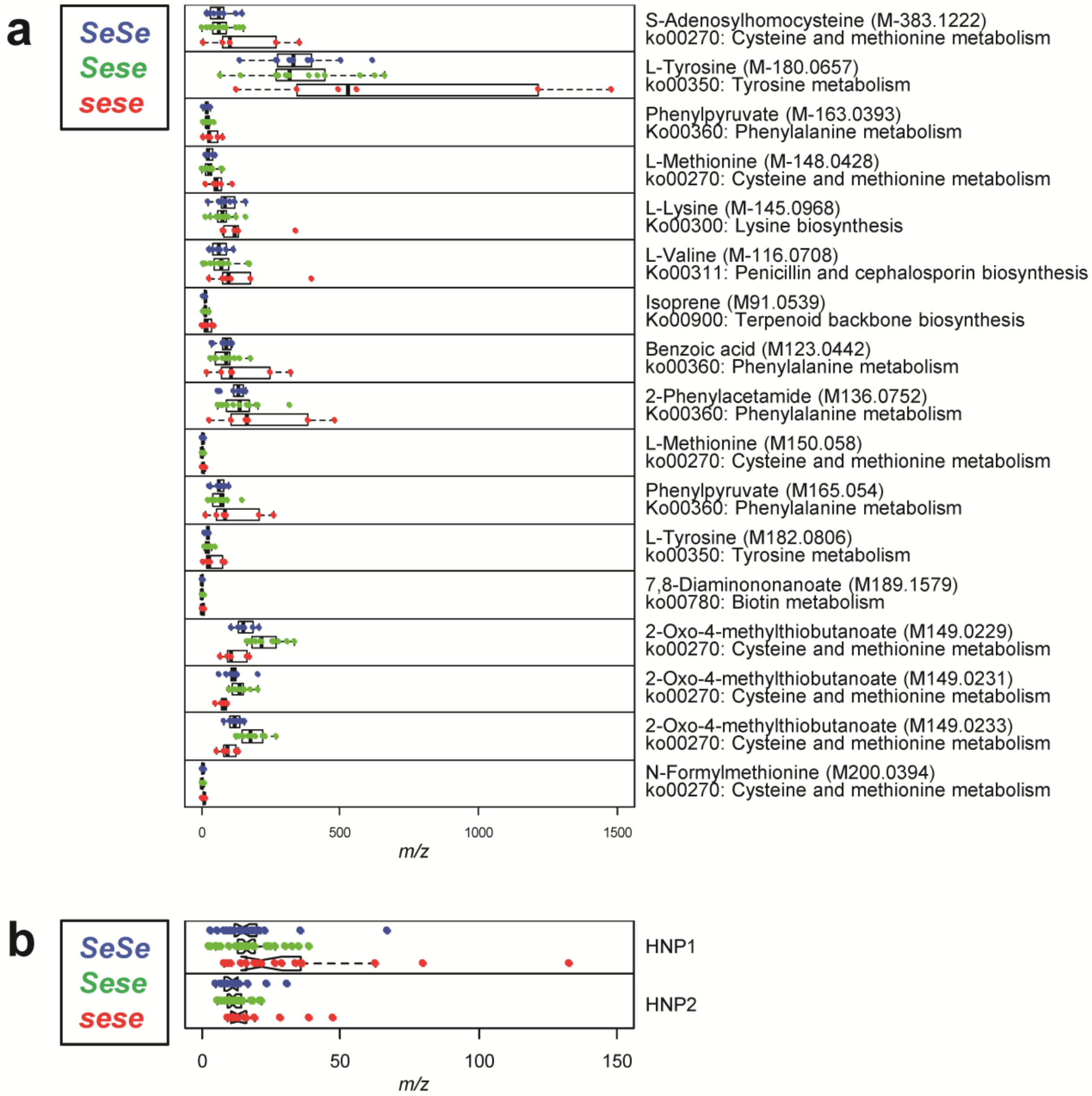


Figure 4-6. Meta-metabolomic and meta-proteomic features that differentiate secretors and non-secretors. Relative abundance of meta-metabolomic **(a)** and meta-proteomic **(b)** features in lavage samples is colored by secretor status.

Table 4-1. Demographic information of the cohort

Metadata of Dataset				
		SeSe	*Sese*	*sese*
Total Subject (39)		12 (31%)	18 (46%)	9 (23%)
Total Sample (75)		23	35	17
Gender	Male (24)	9	12	3
	Female (15)	3	6	6
Age (Average ± s.d.)		63.4 ± 9.7	63.9 ± 13.8	58.9 ± 5.6
Anatomical region	Cecum (36)	11	17	8
	Sigmoid (39)	12	18	9

CHAPTER 5

Microbial Gardening by *FUT2* Gene and Its Link to Crohn’s Disease

Abstract

Background

FUT2 non-secretor status (*sese*), is associated with increased susceptibility to Crohn's disease. Previous study in our group has revealed significant alterations of the colonic microbiome associated with healthy non-secretors. The changes are pervasive at both the compositional and functional levels, highlighting perturbations of energy metabolism including the enrichment of carbohydrate and lipid metabolism and the depletion of amino acid biosynthesis and metabolism related pathways in non-secretors. Such gardening effect, however, has not been validated in Crohn's disease (CD) patients.

Methods

To characterize the effect of *FUT2* polymorphism on the mucosal ecosystem in Crohn's disease patients, we profiled the microbiome and meta-metabolome of 252 endoscopic lavage samples from the cecum and sigmoid of 98 healthy subjects (35 *SeSe*, 45 *Sese* and 18 *sese*) and 33 CD patients (9 *SeSe*, 14 *Sese* and 10 *sese*). To profile the phylogenetic composition of microbiota on the intestinal mucosal surface, 16s ribosomal DNA (rDNA) extracted from the bacterial pellets were sequenced by Illumina HiSeq 2000. The genome contents of the detected 97% OTUs were imputed using (PICRUSt) (v0.1), and the metabolic pathways were re-constructed using HUMAnN (v0.98). The soluble metabolites of the same samples were analyzed using quadrupole time-of-flight (Q-TOF) mass spectrometry.

Results

The phylogenetic composition of intestinal mucosal microbiota was affected by an interaction of Crohn’s disease status and host genetics, specifically *FUT2* genotype. Decreased abundances of Firmicutes were associated with both CD and *FUT2* risk allele. At metagenomic level, a distinct signature of amino acid metabolism deficiency was identified in CD and non-secretor microbiome. Such changes were also reflected at metabolomic level in the proximal gut region.

Conclusion

Our data supported the hypothesis that *FUT2* gene increased the risk of Crohn’s disease by changing the microbial composition and function to a disease-like state. The CD associated perturbation of metagenome and metabolome was driven by the *FUT2* risk allele.

Introduction

FUT2 encodes for fucosyltransferase 2, which is an enzyme that is responsible for the synthesis of the H antigen in body fluids and on the intestinal mucosa. Recent genome-wide association studies have identified *FUT2* as a Crohn's disease risk locus (20, 96), although the molecular mechanism of the association between non-secretor status and CD remains unknown. Dysbiosis, which refers to perturbations of the normally stable intestinal microbiota, has been associated with the development and progression of many conditions, including inflammatory bowel diseases (IBD) (16-18, 89). The phylogenetic composition in non-secretor individuals have been documented in several studies recently (40, 103, 104), showing that the *FUT2* genotype was associated both with deviations of overall community ecology and with altered abundances of specific microbes. Thus, one hypothesis is that the *FUT2* loss-of-function allele increased the risk of Crohn's disease by shaping the compositional and functional states of mucosal microbiota. However, previous studies only characterized the compositional changes associated with *FUT*2, and provided only limited information on the metagenomic functional changes and the potential mechanisms of action in IBD risk.

Combining multi'omics profiling approaches, we have recently presented a comprehensive characterization of different aspects of the microbial ecosystem including microbiome composition, imputed function, metabolome and proteome (2). Imputed metagenomic analysis revealed perturbations of energy metabolism in the microbiome of non-secretor and heterozygote individuals, notably the enrichment of carbohydrate and lipid metabolism, cofactor and vitamin metabolism, and glycan biosynthesis and metabolism related pathways; and, the depletion of amino acid biosynthesis and metabolism. Similar changes were observed in mice bearing the *FUT2*$^{-/-}$ genotype. Metabolomic analysis of human specimens revealed concordant as well as novel changes in the levels of several metabolites. Human metaproteomic analysis indicated that these functional changes were accompanied by sub-clinical levels of inflammation in the local intestinal mucosa.

We presented here the follow-up study of the analysis in healthy individuals, aiming to characterize the effect of *FUT2* on the functional states of intestinal microbiota in Crohn's disease patients. The composition and diversity of the microbiota were characterized by 16S rRNA gene sequencing, and the metagenomic contents were imputed using PICRUSt. So far, no metabolic profiling has been done to study the effect of *FUT2* on intestinal metabolome in CD patients. It allowed us to move from observing patterns to understanding mechanisms by combining 16S rRNA gene sequencing and metagenomic profiling with metabolomics data.

Materials and Methods

Subject Cohort and Lavage Sample Collection

A subject cohort of 131 individuals (Table 5-1) was recruited from patients presenting for screening colonoscopy at Cedars-Sinai Medical Center. Following institutional review board approval, subjects were consented and then included in the study if colonoscopy did not reveal any mucosal abnormalities. Enrolled subjects were prepared for colonoscopy by taking Golytely the day before the procedure. The mucosal lavage samples representing the mucosal luminal interface were collected from cecum and sigmoid colon as described previously (22).

Genotyping

The SNP rs601338 (G > A) defines secretor status in Europeans and Africans (106). We used rs516246, which is in strong linkage disequilibrium with rs601338, to infer secretor status. Estimate of linkage between rs516246 and rs601338 is 100%. These 2 SNPs tag each other perfectly as they are in perfect LD with one another, with R-square of 1.0 and D-prime of 1.0, according to Hapmap3 release 2, CEU population. The individuals with the homozygote A/A genotype are defined as nonsecretors.

16S rRNA Gene Sequencing and Microbial Composition Analysis

Genomic DNA was extracted as previously described (3). The V4 region of 16S ribosomal RNA genes were amplified and sequenced on an Illumina HiSeq 2000 as previously described (3). HiSeq reads were processed using QIIME v1.5.0 (27) with parameters of: minimum Q-score considered high quality: 20, maximum number of consecutive low-quality base calls allowed before truncating: 3, maximum number of N characters allowed: 0. All filtered reads had a length of 101 bp. The number of reads per sample ranged from 304,675 to 1,008,302, with a mean of 591,912. Sequence sub-sampling was performed for each sample at the depth of 300,000 reads/sample. This normalized dataset was used for all the following analysis including alpha-diversity analysis, beta-diversity analysis, and imputed metagenomic analysis. Operational taxonomic units (OTUs) were picked against the February 4th, 2011 version of the Greengenes database (http://greengenes.lbl.gov/cgi-bin/nph-index.cgi) (26), pre-filtered at 97% identity. For quality control, all the singletons were removed. After reference-based OTU picking, 97.5% of the total reads were successfully mapped to the reference Greengenes database. These steps were performed using QIIME v1.5.0 (27). Alpha rarefaction was performed using phylogenetic diversity. The comparison of alpha diversity between two groups at certain sampling depths was performed using a two-sided Student *t* test.

Imputation of microbial gene content and metagenomes

This study takes advantage of PICRUSt, a program that infers the metagenome of a sample from its phylogenetic composition and was recently validated against conventional deep-sequencing metagenomics (109). The OTU table including all 252 samples was used as the input file for metagenome imputation of individual human samples. The gene content of 2,590 KEGG (Kyoto Encyclopedia of Genes and Genomes) reference genomes was used to infer the approximate gene content of the detected phylotypes using PICRUSt (v0.1) (http://picrust.sourceforge.net/) (109). Taking the PICRUSt KO gene abundance inferences as inputs, the metabolic pathways were re-constructed using HUMAnN (v0.98) (55). We restricted our analysis to the KEGG pathways that were present in all of the samples.

Bioinformatic and statistical analyses

Constrained correspondence analysis (CCA): CCA was performed using the **cca** function in the vegan R package. CCA was performed on the imputed metagenome and metabolome with collected subject and sample metadata. Model fitting was performed using the **envfit** function in the vegan R package. The constrained model was defined with anatomical region, gender, ethnicity, *FUT2* genotype and CD phenotype as constrained variables. The significance of each variable was assessed using permutation tests.

LDA effect size analysis: LDA effect size analysis was performed using the LEfSe tool (http://huttenhower.sph.harvard.edu/galaxy/). Parameters were set as: α for pairwise tests: 0.05 for both class normality and subclass tests; threshold on the logarithmic score of LDA analysis: 2.0. All LDA scores are determined by bootstrapping over 30 cycles, each sampling two-thirds of the data with replacement, with the maximum influence of the LDA coefficients in the LDA score of three orders of magnitude.

Results

Phylogenetic composition of intestinal mucosal microbiota associated with both Crohn's disease phenotype and *FUT2* genotype

To define the effect of disease phenotype and host genetics on mucosal microbiota, we collected 252 lavage samples from the cecum and sigmoid colons of 98 healthy subjects and 33 CD patients. (Table 5-1). We assessed shift in overall microbial ecology between secretors (both homozygous *SeSe* and heterozygous *Sese* for the functional allele) and non-secretors (*sese*) in each subject group respectively. To examine the microbial composition in these samples, the microbiota was profiled by multiplex sequencing, and on average 591,912 reads per sample were generated after quality control. 5,648 OTUs were then identified by grouping reads at a 97% sequence similarity threshold.

Dysbiosis of commensal microbes and pathogens is a prominent factor in the development of inflammatory bowel disease (1, 16, 39). One of the well-defined feature is the reduction in diversity of the colonic mucosa associated bacterial microflora in IBD patients. Indeed, compared with healthy individuals, microbiota from CD patient exhibited significantly lower alpha diversity as measured by phylogenetic diversity (Figure 5-1) (*t*-test, $P < 0.001$ at the depth of 300,000 reads). Moreover, individuals carrying the loss-of-function allele of *FUT2* (*Sese* and *sese*) also showed the trend of decreased microbial diversity in both control and CD subjects respectively (Figure 5-1), although only the difference between Non-IBD *SeSe* and *Sese* individuals was significant (*t*-test, $P < 0.001$). Such haploinsufficiency effect of *FUT2* on microbial composition and function has also been observed before in our pilot cohort (2). The effect of *FUT2* loss-of-function allele was masked by disease status in CD patient. The intermediate low microbial diversity in *Sese* and *sese* individuals supported the hypothesis that *FUT2* gene increased the risk of Crohn’s disease by changing the microbial composition to a disease-like state.

To evaluate the effect of disease phenotype and *FUT2* genotype, we used constrained correspondence analysis to evaluate the phylogenetic similarity between samples. In constrained ordination, only the variation that can be explained by the environmental variables was evaluated and displayed. In this model, we used subject and sample metadata including anatomical region, gender, ethnicity, *FUT2* genotype and CD phenotype as constrained variables. Crohn’s disease phenotype remained to be a strong factors that shaped microbial composition (Figure 5-2) ($P = 0.001$). In addition, subjects also clustered by *FUT2* genotype in both control and CD groups, with the clustering of

SeSe individuals most distinct. The association between microbial composition and *FUT2* genotype was less significant than that with disease status (*P* = 0.023).

To identify the phylotypes that explained the difference between these clusters, we applied the linear discriminant analysis (LDA) effect size (LEfSe) analysis to OTUs from phylum to genus level (132). Using this method, we identified 72 associations between microbial phylotypes and Crohn's disease at different taxonomic levels in total. The changes were even prominent at phylum level, with increased abundances of Proteobacteria and decreased abundances of Firmicutes in CD patients (Figure 5-3 A), the microbial signature that has been validated in several independent cohorts (1, 30, 31). At lower taxonomic level, the *Enterobacteriaceae* family containing potential pathobiont genera such as *Escherichia* (including adherent-invasive *Escherichia coli* (77)) and *Klebsiella* genera was enriched in CD patients, whereas the genus including the anti-inflammatory commensal bacterium *Faecalibacterium prausnitzii* (79), along with other short-chain fatty acid (SCFA) producing bacteria including *Eubacterium*, *Roseburia* and *Coprococcus,* were enriched in non-IBD subjects (80, 82). Interestingly, some of these phylotypes were also identified to be associated with *FUT2* genotype. Using the same approach, we identified 15 phylotype at different taxonomic level that were enriched in *SeSe*, *Sese* or *sese* individuals (Figure 5-3 B). Among those, the associations between Firmicutes, specifically *Lachnospiraceae* and *Veillonellaceae* families, and *SeSe* individuals were also observed in our previous healthy-only cohort (2). Taken together, these data indicated that the phylogenetic composition of intestinal mucosal microbiota was affected by an interaction of Crohn's disease status and host genetics, specifically *FUT2* genotype.

Inefficiency of amino acid metabolism in CD and non-secretor microbiome

As the result of phylogenetic dysbiosis, microbial function was consistently perturbed in IBD patients (18). Major shifts in metagenome have been identified, including increased oxidative stress pathways, as well as decreased carbohydrate metabolism and amino acid biosynthesis in favor of nutrient transport and uptake. To determine if consistent changes at metagenomic level can be observed in our dataset, we inferred the metabolic capacities of mucosal microbiota associated with secretor status, using a recently developed bioinformatic pipeline centering on the PICRUSt (109) and HUMAnN tools (55). Among the 152 imputed metabolic pathways that were present in all the samples, 62 were differentially abundant between CD and non-IBD control (Figure 5-4). There was an evident signature of amino acid metabolism deficiency in CD microbiome, as represented by the decreased abundances of lysine biosynthesis (ko00300), valine, leucine and isoleucine biosynthesis (ko00290), phenylalanine, tyrosine and tryptophan biosynthesis (ko00400) and cysteine and methionine metabolism (ko00270) pathways. Similarly, 17 pathways were identified to be associated with *FUT2* genotype (Figure 5-5), among which cysteine and methionine metabolism (ko00270) and phenylalanine metabolism (ko00360) pathways were enriched in *SeSe* and *Sese* individuals respectively. These data indicated that both CD phenotype and *FUT2* genotype played important roles in shaping the functional capacity of mucosal microbiome. A consistent pattern of AA metabolism deficiency can be observed in CD patients and individuals with loss-of-function allele of *FUT2*.

Metabolomic shift associated with both disease phenotype and host genotype

Metabolomic analyses enabled direct assessment of the effects of metagenomic changes on the metabolic outcomes of gut microbiota. Microbes are syntropic with mucosal metabolome composition, and highly influence the colonic luminal and mucosal metabolome (3, 15). Previous metabolomics profiling studies in IBD patients have identified pathways with differentiating metabolites, including those involved in the metabolism and/or synthesis of amino acids, fatty acids, bile acids and arachidonic acid (133-135). It is expected then that the perturbation of imputed metagenome associated with CD phenotype and *FUT2* genotype would also be reflected at metabolomics level. Indeed, samples from both cecum and sigmoid regions with same diagnosis and *FUT2* genotype formed distinct clusters based on their metabolomic composition (Figure 5-6 A and B). Crohn's disease remained to be a stronger factor as shown by the separation of CD and non-IBD control samples. Interestingly, a strong biogeographic effect of *FUT2* gene on the gut metabolome was observed, as indicated by the clustering of *sese* samples from cecum but not sigmoid region (Figure 5-6 A). These data, combined with the findings at the phylogenetic and imputed metagenomic levels, indicated that *FUT2* gene and Crohn's disease phenotype functioned together to shape the functional states of mucosal associated microbiota in human.

Discussion

IBD has long been known to have genetic risk factors due to the increased risk in first-degree relatives of affected individuals. Meta-analysis of genome-wide association studies (GWAS) has increased the number of confirmed IBD (both Crohn's disease and ulcerative colitis) susceptibility loci to 167 (20), indicating that IBD is biologically heterogeneous. As post-GWAS functional characterization of risk loci in human subjects, our data suggests that the CD associated perturbation of metagenome and metabolome was driven by the *FUT2* risk allele. Among the 167 susceptibility loci, several other genes were involved in the genetically impaired immune regulation and epithelial barrier function, such as *NOD2* in the sensing of bacterial products, *JAK2-STAT3* pathway in immune responses, and the IL-2-Th17 pathway in microbial defense mechanisms (20, 29, 136). Defining the dysbiosis and functional changes in individuals with different genetic risk profile will help us formulate the personalized therapeutic interventions with better efficacy.

We have previously reported the metabolomic changes concordant with metagenomic alterations in healthy non-secretors (2). Similar associations were confirmed in healthy individuals, and were also observed in CD patients as well. Interestingly, the effect of *FUT2* gene on the gut metabolome was only evident in the proximal cecum region, but not the distal part. This may in part be explained by the different glycosylation landscape in human colonic epithelium, since the availability of host fucosylated glycans directly affects the functions that microbiota expresses in the distal gut (107). The biogeographic pattern was also observed in *FUT2*$^{-/-}$ mice, although in distal rather than proximal gut (107). Further detailed characterization of glycosylation profile of mucin along human gut will help us explain the differential effect of *FUT2* on metabolome.

Figures and Tables

Figure 5-1

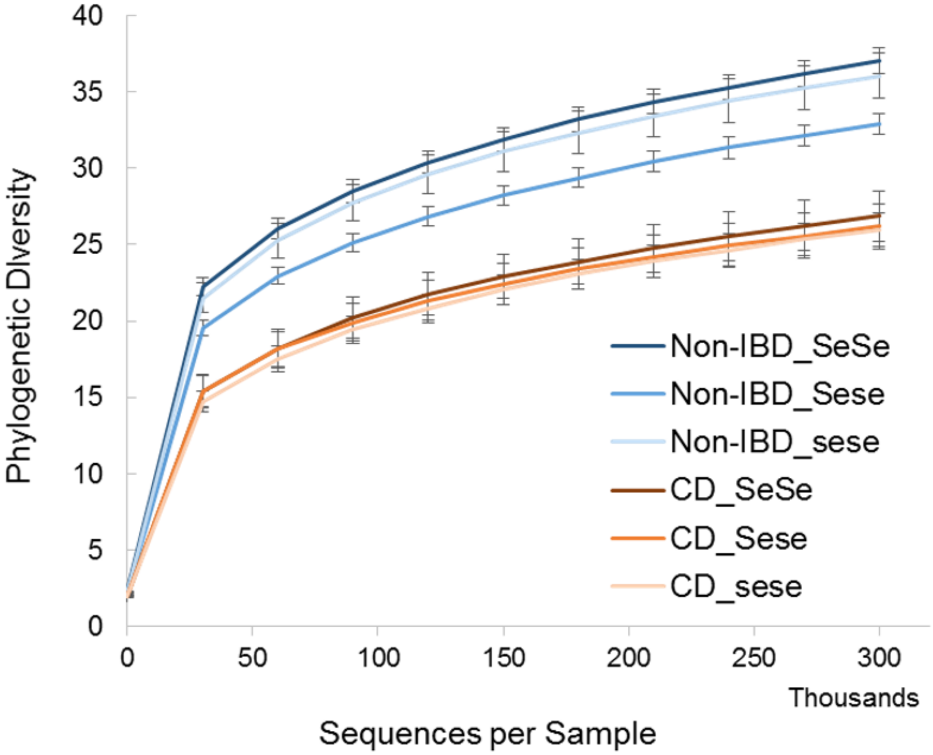


Figure 5-1. Rarefaction curve of microbial diversity for the microbiota from lavage samples. Rarefaction curves of phylogenetic diversity for microbiota (mean ± 95% CI) were plotted at different sequencing depths for individuals with different disease phenotype and *FUT2* genotype.

Figure 5-2

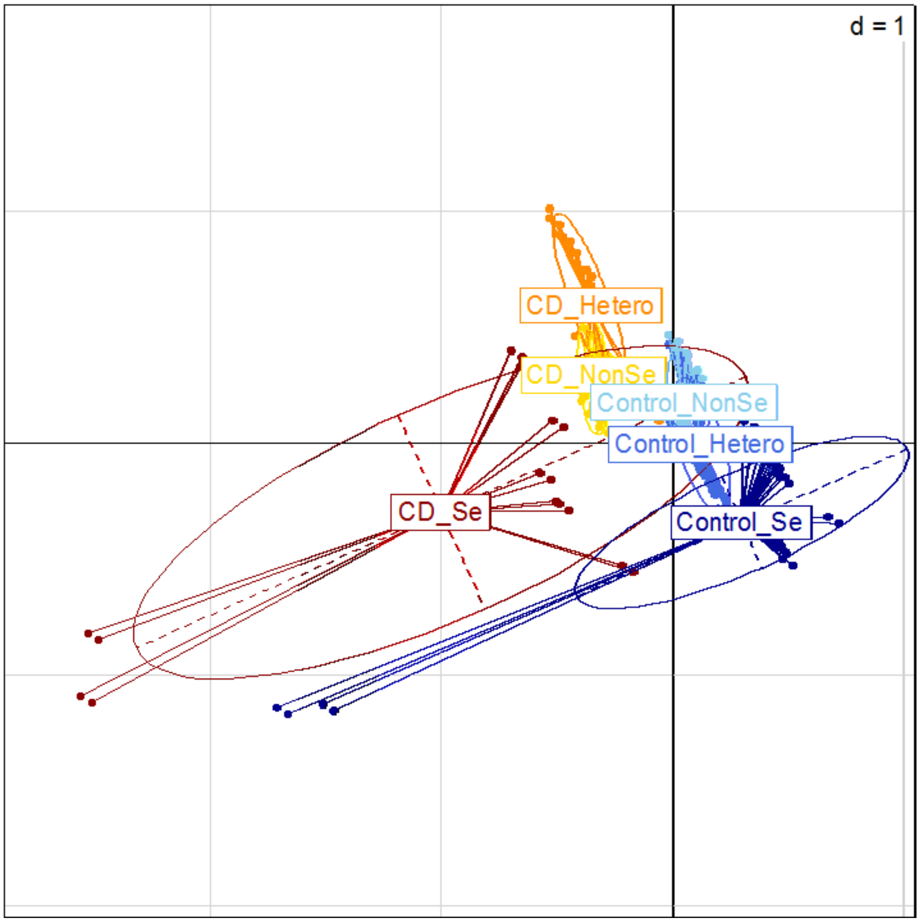


Figure 5-2. Shifts of mucosal microbial phylogenetic composition in individuals with different disease phenotype and FUT2 genotype. Ordination of samples using CCA of disease status and *FUT2* genotype, conditioned by environmental factors including anatomical region, gender and ethnicity. The centre of gravity for each cluster was marked by a rectangle.

Figure 5-3

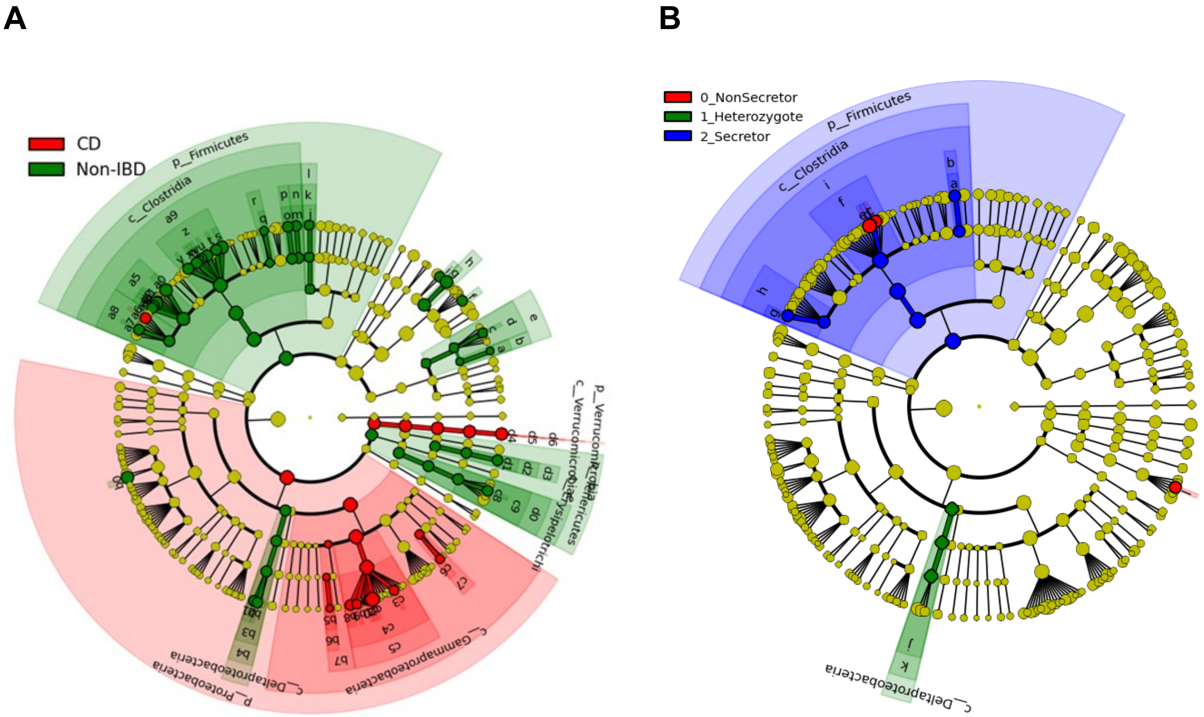


Figure 5-3. Cladogram of phylotypes associated with Crohn's disease and *FUT2* genotype. Phylogenetic taxonomy representation of statistically and biologically consistent differences between (A) Crohn's disease and non-IBD controls, and (B) *SeSe*, *Sese*, and *sese* individuals. Differences were represented in the color of the most abundant class. Each circle's diameter is proportional to the taxon's abundance.

Figure 5-4

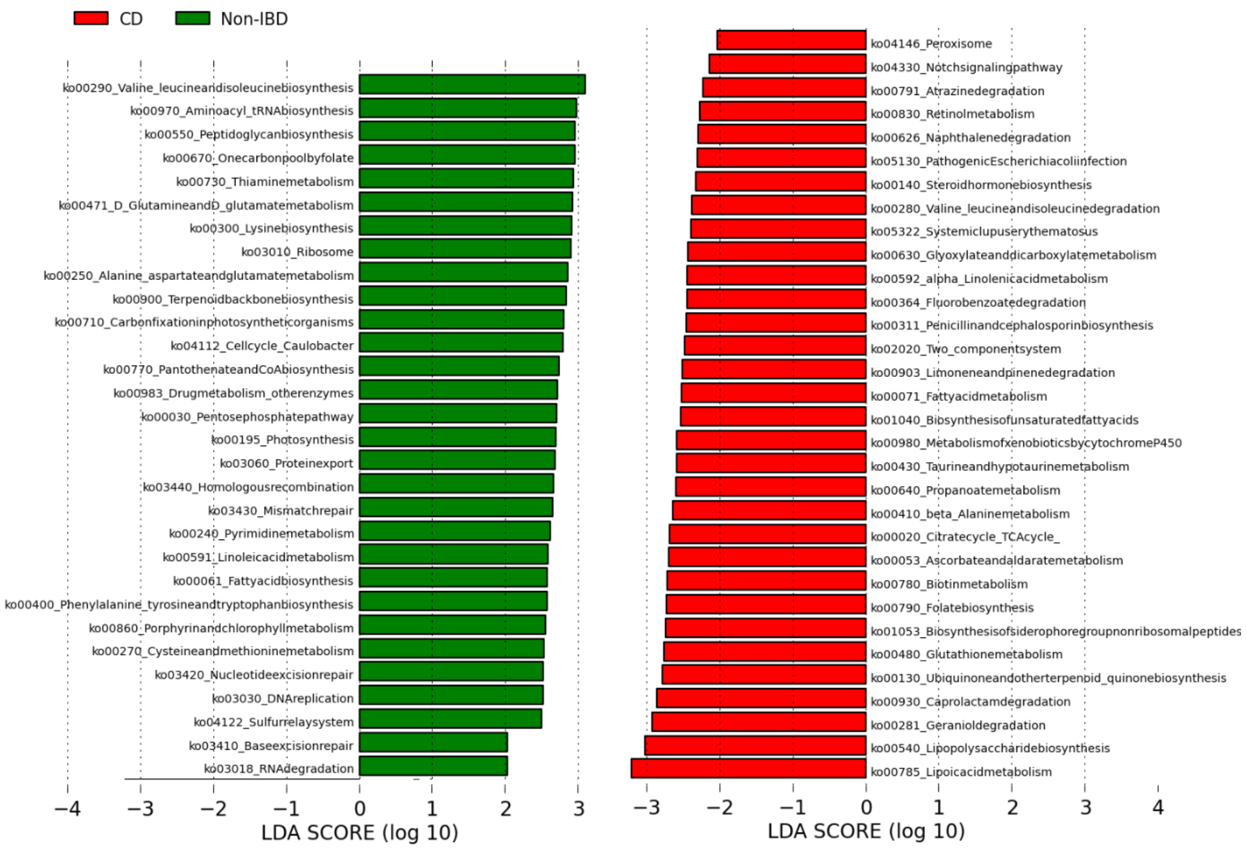


Figure 5-4. Crohn’s disease associated KEGG metabolic pathways. Histogram of the LDA scores computed for metagenomic features differentially abundant between CD (red) and non-IBD controls (green) was shown. LEfSe scores can be interpreted as the degree of consistent difference in relative abundance between features in the two classes of analyzed microbial communities. The histogram thus identifies which pathways among all those detected as statistically and biologically differential explain the greatest differences between communities.

Figure 5-5

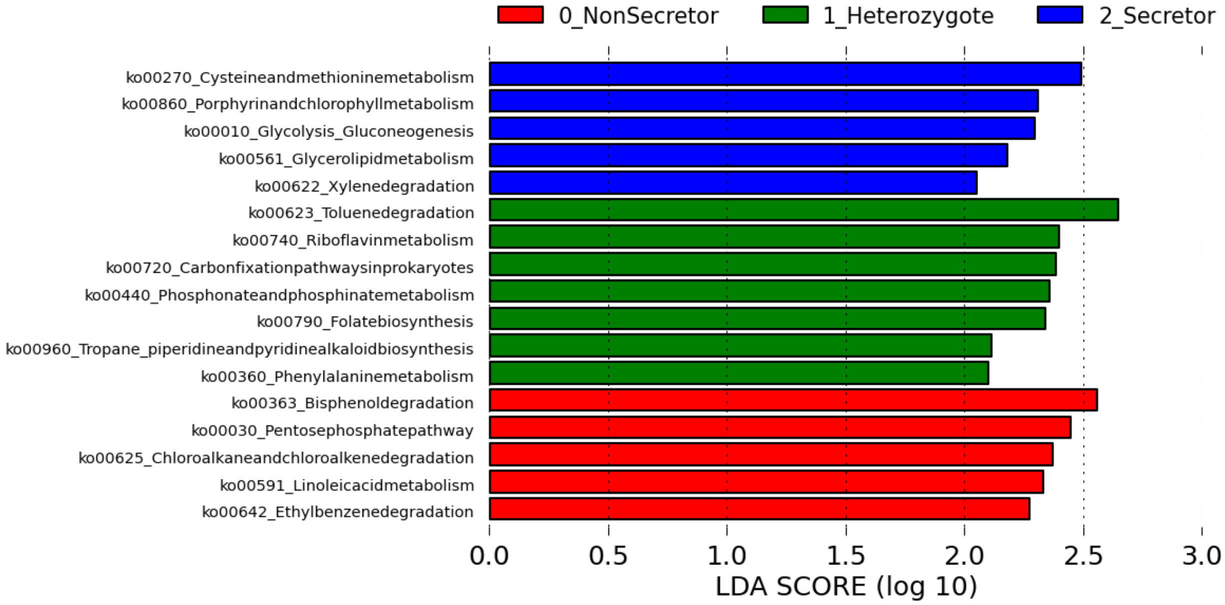


Figure 5-5. ***FUT2*** **genotype associated KEGG metabolic pathways.** Histogram of the LDA scores computed for metagenomic features differentially abundant between *SeSe* (blue), *Sese* (green) and *sese* (red) individuals was shown.

Figure 5-6

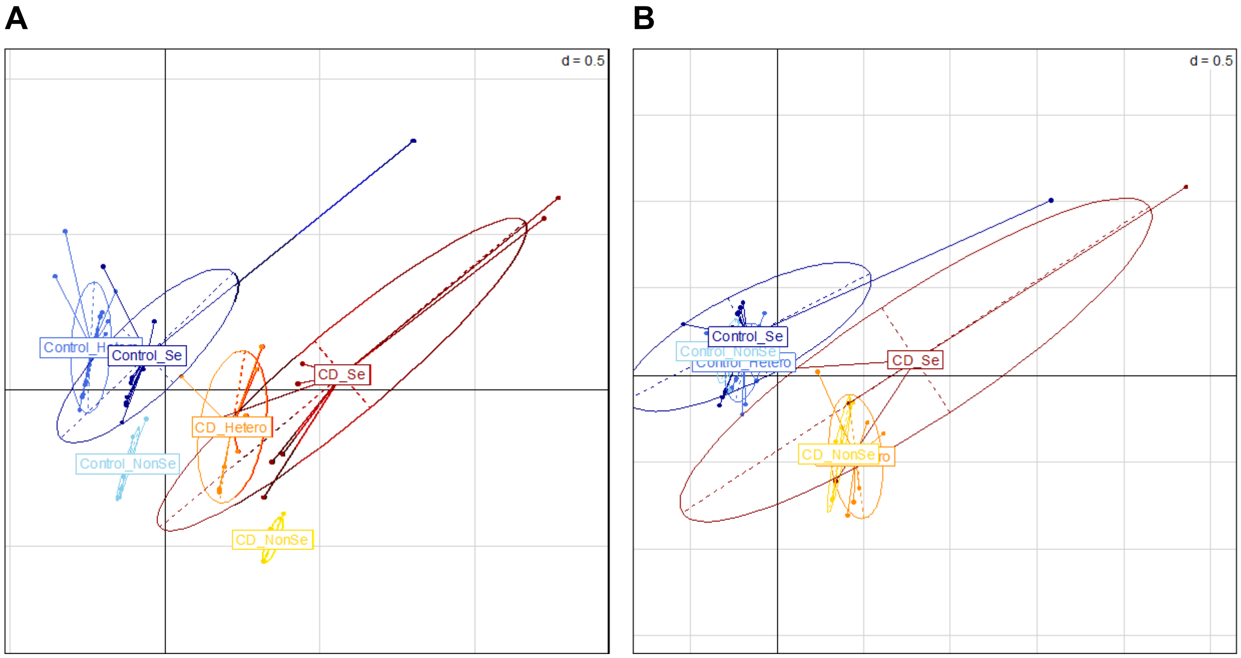


Figure 5-6. Shifts of mucosal metabolomic composition in individuals with different disease phenotype and *FUT2* genotype. Ordination of samples from (A) cecum and (B) sigmoid using CCA of disease status and *FUT2* genotype. The centre of gravity for each cluster was marked by a rectangle.

Table 5-1. Demographic information of the Crohn's disease patient cohort

		Control (98)			CD (33)		
		SeSe	*Sese*	*sese*	*SeSe*	*Sese*	*sese*
	Total Subject (131)	35	45	18	9	14	10
Gender	Male (85)	25	33	12	5	5	5
	Female (46)	10	12	6	4	9	5
	Age (Average ± s.d.)	63.5 ± 10.0	63.9 ± 12.5	64.4 ± 8.4	36.8 ± 13.8	39.3 ± 16.1	41.5 ± 8.6
	Total Sample (252)	67	87	35	16	27	20
Anatomical region	Cecum (126)	34	43	17	9	13	10
	Sigmoid (126)	33	44	18	7	14	10

CHAPTER 6

Conclusions

Mucosal lavage sampling: a novel way to study host-microbial interaction

We developed the novel mucosal lavage sampling approach, which enabled the profiling of multi'omic molecular features including microbiome, metaproteome and metabolome. Combined with host genomic information, these tools can provide us with unprecedented understanding of the dynamics of host–microbial interaction, and help us to investigate the pathogenesis of inflammatory bowel diseases. The same framework also represents a powerful tool to study the mucosal biology in other chronic inflammatory diseases such as HIV infection.

Functional microbial communities: a new insight into microbial ecology

We have developed a novel strategy using an ecologic mucosal microbial framework, minimally invasive mucosal sampling, short-read Illumina sequencing, network analysis, and imputed metagenomics. This strategy uncovered 5 reproducible functional microbial communities (FMCs) detectable in the mucosa of all individuals. The quantitative levels of two FMCs were significantly associated with IBD states. Imputed metagenome analysis indicated the functional importance of the disease associated modules reflected by the enrichment of virulent and pathogenic pathways. Thus, these modules appear to define novel microbial communities within the intestinal microbial ecology, some of which are commonly and stably modified by the IBD disease state, and may be of particular relevance for microbial pathogenesis and intervention.

In the meantime, alternative analytical tools have been developed by other research groups following the same rationale of microbial co-occurrence network analysis. Claesson *et al.* established co-abundance associations of genera and then clustered correlated genera into six co-abundance groups (CAGs) (91). The distinctive dominances of these CAGs accompanied the transition of residence location of elderly subjects and the dietary and lifestyle distinctions between African and western population (91, 137). An study of the initial Human Microbiome Project (HMP) cohort constructed a global network of 3,005 significant co-occurrence and co-exclusion relationships between 197 clades occurring throughout the human microbiome (50). This network revealed strong niche specialization, with most microbial associations occurring within body sites and a number of accompanying inter-body site relationships. These tools combined together will provide an integrative view of microbial ecology relevant to chronic diseases.

FUT2 and Crohn's disease: a better understanding of IBD pathogenesis

Using this experimental and bioinformatic framework, we investigated the microbial gardening effect of *FUT2* gene and its link to Crohn's disease. In healthy individuals, imputed metagenomic analysis revealed perturbations of energy metabolism in the microbiome of non-secretor and heterozygote individuals, notably the enrichment of carbohydrate and lipid metabolism, cofactor and vitamin metabolism, and glycan biosynthesis and metabolism related pathways; and, the depletion of amino acid biosynthesis and metabolism. Similar changes were observed in mice bearing the *FUT2*$^{-/-}$ genotype. Metabolomic analysis of human specimens revealed concordant as well as novel changes in the levels of several metabolites. Human metaproteomic analysis indicated that these functional changes were accompanied by sub-clinical levels of inflammation in the local intestinal mucosa.

In an extended cohort containing both healthy and CD individuals, the phylogenetic composition of intestinal mucosal microbiota was affected by an interaction of Crohn’s disease status and *FUT2* genotype. Decreased abundances of Firmicutes were associated with both CD and *FUT2* risk allele. At metagenomic level, a distinct signature of amino acid metabolism deficiency was identified in CD and non-secretor microbiome. Such changes were also reflected at metabolomic level in the proximal gut region. Taken together, *FUT2* gene increased the risk of Crohn’s disease by changing the microbial composition and function to a disease-like state. The CD associated perturbations of metagenome and metabolome were driven by the *FUT2* risk allele.

So what’s next?

The findings here should be further tested and validated in other human microbiome datasets and in mouse models. Based on cross-sectional cohort studies, we have identified a set of metagenomic pathways involved in the pathogenesis of IBD. By comparing with other IBD microbiome dysfunction study (18), several pathways were highlighted that were consistently associated with disease phenotype and IBD risk gene in different datasets (Table 6-1). These pathways should be further validated by cross-comparing with any future IBD or chronic disease microbiome studies. Also, during the past several years, there has been a shift from association analysis in human to mechanistic study of observed dysbiosis and dysfunction. Such investigation would be enabled by longitudinal cohort study or colitis mouse models, which will allow us to determine which alterations are causal for IBD.

The metagenomic and metabolomic changes associated with IBD or different genetic background serve as promising targets for therapeutic efforts. Intervention strategies should be immediately developed to reverse the metagenomic and metabolomic changes in patients to alleviate the symptoms, and to prevent onset of diseases in individuals with high genetic risk. Combined with genetic testing such as Immunochip profiling, these efforts will be the first step towards personalized treatment of IBD.

The study design and analytic pipelines are transferrable for other IBD risk genes, and other chronic diseases with both genetic and microbiome components such as metabolic syndrome. IBD has long been known to have genetic risk factors due to the increased risk in first-degree relatives of affected individuals. Meta-analysis of genome-wide association studies (GWAS) has increased the number of confirmed IBD (both Crohn’s disease and ulcerative colitis) susceptibility loci to 167 (20), indicating that IBD is biologically heterogeneous. The experimental and analytic approaches presented here represent a powerful and valuable framework of post-GWAS functional characterization of IBD risk loci in human subjects. Among the 167 susceptibility loci, several other genes were involved in the genetically impaired immune regulation and epithelial barrier function, such as *NOD2* in the sensing of bacterial products, *JAK2-STAT3* pathway in immune responses, and the IL-2-Th17 pathway in microbial defense mechanisms (20, 29, 136). Defining the dysbiosis and functional changes in individuals with different genetic risk profile will help us formulate the personalized therapeutic interventions with better efficacy.

Table 6-1. Meta-analysis of metagenomic pathways associated with IBD activity and risk gene

Note: The associations of each KEGG pathways with IBD activity and risk gene in different 16S rRNA microbiome datasets are listed. The Morgan *et al*. (18), MLI (presented in Chapter 4) and Tong *et al.* (1) datasets consisted of IBD patients and control subjects, and the associations between pathways and disease phenotype were identified. The *FUT2* dataset (2) presented the association with protecting or IBD risk allele of *FUT2*. The score counts the number of times that the same pathway was identified in different studies, and only the pathways with a score higher than 1 are listed.

KEGG Pathways	**Morgan *et al.***	**MLI**	**Tong *et al.***	***FUT2***	**Score**
ko00300-Lysine biosynthesis	TRUE	Control	Control	Protecting	4
ko00480-Glutathione metabolism	TRUE	CD	IBD	Risk	4
ko00730-Thiamine metabolism	TRUE	Control	Control	Protecting	4
ko00860-Porphyrin and chlorophyll metabolism	TRUE	Control	Control	Protecting	4
ko00270-Cysteine and methionine metabolism	TRUE	Control	Control	Protecting	4
ko00311-Penicillin and cephalosporin biosynthesis	TRUE	CD	IBD	Risk	4
ko00240-Pyrimidine metabolism		Control	Control	Protecting	3
ko00900-Terpenoid backbone biosynthesis		Control	Control	Protecting	3
ko00785-Lipoic acid metabolism		CD	IBD	Risk	3
ko00281-Geraniol degradation	TRUE	CD	IBD		3
ko00430-Taurine and hypotaurine metabolism		CD	IBD	Risk	3

ko00053-Ascorbate and aldarate metabolism		CD	IBD	Risk	3
ko00130-Ubiquinone and other terpenoid quinone biosynthesis		CD	IBD	Risk	3
ko00540-Lipopolysaccharide biosynthesis		CD	IBD	Risk	3
ko00290-Valine, leucine and isoleucine biosynthesis		Control	Control	Protecting	3
ko00020-Citrate cycle-TCA cycle		CD	IBD	Risk	3
ko00280-Valine, leucine and isoleucine degradation	TRUE	CD	IBD		3
ko00330-Arginine and proline metabolism		Control	Control	Protecting	3
ko00623-Toluene degradation		CD	IBD	Risk	3
ko00140-Steroid hormone biosynthesis	TRUE	CD		Risk	3
ko00720-Carbon fixation pathways in prokaryotes		CD	IBD	Risk	3
ko04146-Peroxisome		CD	IBD	Risk	3
ko00195-Photosynthesis		Control	Control		2
ko00550-Peptidoglycan biosynthesis		Control		Protecting	2
ko05120-Epithelial cell signaling in Helicobacter pylori infection		Control		Protecting	2
ko03440-Homologous recombination		Control		Protecting	2
ko03010-Ribosome		Control		Protecting	2
ko00970-Aminoacyl tRNA biosynthesis		Control		Protecting	2
ko00471-D Glutamine and D glutamate metabolism		Control	Control		2
ko03060-Protein export		Control		Protecting	2
ko00250-Alanine, aspartate and glutamate metabolism		Control	Control		2
ko00670-One carbon pool by folate		Control		Protecting	2

ko00790-Folate biosynthesis		CD	IBD		2
ko03030-DNA replication		Control		Protecting	2
ko00770-Pantothenate and CoA biosynthesis		Control		Protecting	2
ko04112-Cell cycle-Caulobacter		Control		Protecting	2
ko00030-Pentose phosphate pathway	TRUE	Control			2
ko00640-Propanoate metabolism	TRUE	CD			2
ko03430-Mismatch repair		Control		Protecting	2
ko00590-Arachidonic acid metabolism		CD		Risk	2
ko00400-Phenylalanine, tyrosine and tryptophan biosynthesis		Control		Protecting	2
ko04621-NOD like receptor signaling pathway		Control	Control		2
ko00930-Caprolactam degradation	TRUE	CD			2
ko00500-Starch and sucrose metabolism		Control	Control		2
ko04626-Plant pathogen interaction		Control		Protecting	2
ko00410-beta Alanine metabolism		CD	IBD		2
ko00051-Fructose and mannose metabolism	TRUE	Control			2
ko00361-Chlorocyclohexane and chlorobenzene degradation		CD	IBD		2
ko00061-Fatty acid biosynthesis		Control	Control		2
ko00780-Biotin metabolism		CD		Risk	2
ko03013-RNA transport		Control		Protecting	2
ko00362-Benzoate degradation	TRUE		Control		2
ko00625-Chloroalkane and chloroalkene degradation	TRUE		Control		2
ko00450-Selenocompound metabolism	TRUE		IBD		2
ko00350-Tyrosine metabolism	TRUE			Protecting	2
ko00230-Purine metabolism	TRUE			Protecting	2

SUPPLEMENTAL CHAPTER 1

O-glycan Structure and Intestinal Epithelial Barrier Function

Abstract

Background

Mucin-type O-linked oligosaccharides (*O*-glycans) are primary components of the intestinal mucins that form the mucus gel layer overlying the gut epithelium. Impaired expression of intestinal *O*-glycans has been observed in patients with ulcerative colitis (UC), but its role in the etiology of this disease is unknown.

Methods

We generated intestinal epithelial cell–specific *C1galt1*$^{-/-}$ (IEC *C1galt1*$^{-/-}$) mice by crossing mice with loxP sites flanking *C1galt1* (*C1galt1*$^{f/f}$ mice) with an intestinal epithelium–specific Cre-expressing transgenic line (VillinCre mice). Peripheral granulocytes and monocytes in IEC *C1galt1*$^{-/-}$ mice were analyzed by flow cytometry. The infiltration of TNF-producing granulocytes and monocytes/macrophages in IEC *C1galt1*$^{-/-}$ colon tissues was analyzed by immunofluorescent staining.

Results

Mice with intestinal epithelial cell–specific deficiency of core 1–derived *O*-glycans, the predominant form of *O*-glycans, developed spontaneous colitis that resembled human UC, including massive myeloid infiltrates and crypt abscesses. The colitis manifested in these mice was also characterized by TNF-producing myeloid infiltrates in colon mucosa in the absence of lymphocytes, supporting an essential role for myeloid cells in colitis initiation.

Conclusion

These data indicate a causal role for the loss of core 1–derived *O*-glycans in colitis.

Myeloid cells are essential for the initiation of colitis in IEC C1galt1$^{-/-}$ mice.

Introduction

Ulcerative colitis (UC) is an immune-mediated disorder that results from an abnormal interaction between colonic bacteria and mucosal immune cells in a genetically susceptible host (6, 138). The mechanisms underlying this interaction remain to be defined (6, 138).

The colon mucus layer comprises the polymerized mucins, primarily Muc2, that are produced by goblet cells (139, 140). Mucins are glycoproteins that carry large numbers of *O*-linked oligosaccharides (*O*-glycans), which account for up to 80% of the mass of the mucin molecules and are responsible for many of the properties of mucins. *O*-glycans are synthesized post-translationally in the Golgi apparatus (141-143). All *O*-glycans are initiated with a primary structure referred to as Tn antigen (GalNAcα-*O*-Ser/Thr), which is normally masked by additional glycosylation to form the main type of *O*-glycans, core 1–derived structures (143). The biosynthesis of core 1 is controlled by core 1 β1,3-galactosyltransferase (C1galt1, also known as T-synthase) (143), for which expression requires the specific molecular chaperone C1GALT1C1 in the ER (144).

Most UC patients have colitis only in the distal colon (145). Although the reasons for this regional variation are unknown, UC patients have deterioration of the mucus layer and abnormal mucin expression in the distal colon (138, 146-148). Altered intestinal O-glycosylation also occurs in patients with UC. However, the nature of the impaired O-glycosylation in UC patients is unclear (149). Whether abnormal O-glycosylation impairs the mucus inner layer and causes spontaneous colitis is unknown.

Materials and Methods

Mice

IEC *C1galt1*$^{-/-}$ mice and TM-IEC *C1galt1*$^{-/-}$ mice were generated by breeding *C1galt1*$^{f/f}$ mice with VillinCre transgenic mice (Tg[Vil-cre]997Gum; Jackson Laboratory) and VillinCre-ER^{T2} transgenic mice (provided by Sylvie Robine, CNRS–Institut Curie, Paris, France, and Robert Coffey, Vanderbilt University, Nashville, Tennessee, USA), respectively. To induce a postnatal deficiency in intestinal *O*-glycans, 6- to 8-week-old TM-IEC *C1galt1*$^{-/-}$ mice were injected intraperitoneally with 1 mg TM (MP Biomedicals) in an ethanol/sunflower seed oil mixture (1:9 [v/v]) for 5 consecutive days. *Rag1*$^{-/-}$ mice (Jackson Laboratory), *Tlr4*$^{-/-}$ mice, and *Myd88*$^{-/-}$ mice (originally developed by Shizuo Akira, Osaka University, Osaka, Japan) (150) were crossed to IEC *C1galt1*$^{-/-}$ mice to establish mice with combined gene deficiencies. The animal studies were conducted with protocols that had been approved by the Institutional Animal Care and Use Committee of the Oklahoma Medical Research Foundation and UCLA. Mice were raised in a specific pathogen–free barrier facility and genotyped routinely by PCR assay on genomic DNA isolated from tail clips. The primers used for genotyping include Flox1: 5′-TGACAGCCAGGAATGGAACTTG-3′ and Flox2: 5′-GCCTCTTCTCGCAACAAAATACTC-3′; CRE1: 5′-AGGTGTAGAGAAGGCACTTAGC-3′ and CRE2: 5′-CTAATCGCCATCTTCCAGCAGG-3′. Sex- and age-matched littermates were used as controls in all experiments. Unless specified, all mice were on the C57BL/6J congenic background.

Flow cytometry

Single-cell suspensions were obtained by passing the spleen or mesenteric lymph node through a 100-μm cell strainer. Red blood cells were lysed by a 15-min incubation in ammonium chloride lysing reagent (BD Biosciences). Intestinal intraepithelial lymphocytes from colons were isolated as described (151). All antibodies were obtained from BD Biosciences unless specified otherwise. For myeloid cell analysis, peripheral blood leukocytes that were positive for myeloid marker CD11b but negative for the rest of the lineage markers were defined as monocytes. mAb against Ly-6C (AL-21) was used to classify monocytes into subsets with either high expression of Ly-6C (Ly-6Chi) or low expression of Ly-6C (Ly-6Clo). Cells from indicated compartments were incubated with antibodies. Two-color analyses were performed using the FACSCalibur system (BD Biosciences).

Immunofluorescence staining

For the detection of infiltrated inflammatory cells, distal colons were dissected from mice and fixed in 4% paraformaldehyde at 4°C overnight, followed by cryoprotection in 20% sucrose, and then embedded in a mixture of OCT compound (Sakura Finetek) and tissue freezing medium (Electron Microscopy Sciences). Cryosections (20 μm thick) were incubated with rat mAb against murine Ly6G (5 μg/mL, clone 1A8; BD) or rat mAb against murine F4/80 (5 μg/mL, clone C1:A3-1; AbD Serotec) combined with rabbit anti–mouse TNF (polyclonal; BD) overnight at 4°C, developed with phycoerythrin-labeled goat anti–rabbit IgG (1:50; Jackson ImmunoResearch Laboratories Inc.) and Alexa Fluor 488–conjugated donkey anti–rat IgG (1:100; Invitrogen), and mounted with ProLong Gold mounting medium with DAPI (Invitrogen). Specimens were analyzed by epifluorescence imaging using a Nikon C1 confocal laser-scanning unit equipped with a 3-laser launcher mounted on an Eclipse TE200-U inverted microscope. Species-specific, isotype control antibodies were used for negative controls. For each section, 5 different high-powered microscopic fields (×20) were chosen randomly to count the TNF$^+$, Ly6G$^+$, and TNF$^+$/Ly6G$^+$ cells. DAPI-positive areas were quantified as pixels per high-powered microscopic field (×20) using Photoshop (Adobe). The cell numbers were normalized to DAPI staining.

Results

To establish the role of myeloid cells in incipient disease, we first analyzed peripheral granulocytes and monocytes in IEC *C1galt1*$^{-/-}$ mice by flow cytometry. Notably, peripheral blood from IEC *C1galt1*$^{-/-}$ mice had elevated numbers of peripheral granulocytes (Ly6G$^+$CD11b$^+$) at 2 weeks and had a higher level of the inflammatory subset of monocytes (Ly-6G$^-$CD11b$^+$Ly-6C^{hi}) at 3 weeks (Figure S1-1 A). Blockage of P- and E-selectins, which inhibit myeloid cell transmigration into inflammatory sites, or of TNF significantly improved colitis in IEC *C1galt1*$^{-/-}$ mice as compared with IEC *C1galt1*$^{-/-}$ mice treated with control agents (Figure S1-1 B), indicating that myeloid cells and TNF are key inflammatory initiators in our mouse model.

Further immunostaining revealed substantial infiltration of TNF-producing granulocytes and monocytes/macrophages in IEC *C1galt1*$^{-/-}$ colon tissues (2.5-week-old; Figure S1-1 C) but not before the onset of colitis (1 week of age), suggesting that these are major cell types that sense the early microbial intrusion and initiate inflammation. We bred IEC *C1galt1*$^{-/-}$ with mice lacking Myd88, a universal adaptor protein used by most TLRs, or TLR4, which recognizes bacterial antigens, and found that neither Myd88 nor TLR4 deficiency protected IEC *C1galt1*$^{-/-}$ mice from developing colitis (Figure S1-2), suggesting that TLRs are not essential for recognition of bacterial components in our mouse models.

Discussion

Activation and recruitment of myeloid cells in colon mucosa are hallmarks of active UC and are associated with epithelial injury and clinical disease activity (138). However, the significance of myeloid cells in the initiation of colitis and in disease progression remains unclear. Our data indicate that myeloid cells are among the major early responders to bacterial intrusion when the mucus barrier is breached and contribute significantly to tissue damage in the *C1galt1*$^{-/-}$ colon. In general, microbial recognition is achieved through pattern recognition receptors, such as TLR and Nod-like receptors (NLRs) (6, 138, 152). The activation of myeloid cells in our models is Myd88 independent, which is similar to that in T-bet–deficient mice that exhibit colitis in the background of Rag1 deficiency (153), although, unlike in T-bet–deficient mice, spontaneous colitis in*C1galt1*$^{-/-}$ mice is not dependent on the Rag1 deficiency. The underlying mechanism of colitis susceptibility in T-bet–deficient mice includes an enteric dysbiosis resulting from their host genetic deficiency in mucosal defense, with resultant recruitment and activation of proinflammatory myeloid cells. This is different from the mechanism in *C1galt1*$^{-/-}$ mice. In *C1galt1*$^{-/-}$ mice, increased association of intestinal bacteria with colonic epithelia may result from impaired ability of the *O*-glycan–dependent mucus barrier to impede nonspecific bacterial invasion. Recent work has demonstrated that some symbiotic bacteria have developed *O*-glycan sensing and degradation pathways (154). Based on this, it can be anticipated that the changes in the *O*-glycans will alter the composition of the gut microbiota. The ability of bacteria to penetrate the mucus barrier may also contribute to the pathogenesis of colitis in our models. Nevertheless, in both models, Myd88 independence is a surprising feature, suggesting that the relevant microbial sensing in this setting of innate inflammation predominates with Myd88-independent TLR, NLR, or other families of pathogen-associated molecular products (138, 152).

Figures and Tables

Figure S1-1

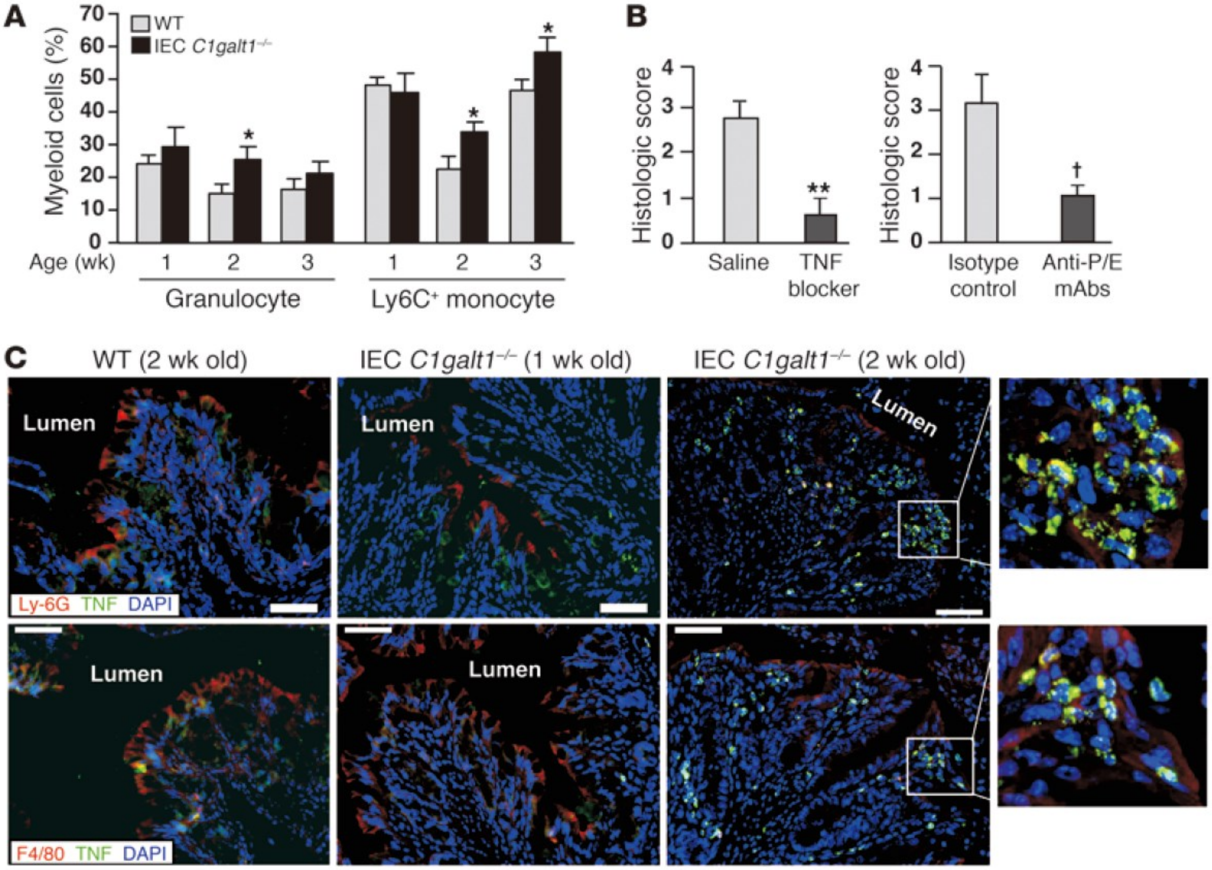


Figure S1-1. Innate immune cells are a critical initiator of colitis in IEC C1galt1−/− mice. (**A**) Peripheral myeloid cells analyzed by flow cytometry (mean ± SD, *n* = 8 mice/group). Granulocytes are defined as Ly-6G$^+$CD11b$^+$. Ly-6G$^-$CD11b$^+$ monocytes were further analyzed for Ly-6C expression. **P* < 0.02. (**B**) Histologic scores of H&E-stained colon sections of 2.5-week-old IEC*C1galt1*$^{-/-}$ mice treated with etanercept (TNF blocker; saline as control, mean ± SD, *n* = 5 mice/group) or blocking antibodies to mouse P- and E-selectin (anti-P/E, rat IgG; isotype IgG used as controls; mean ± SD, *n* = 6

mice/group). $^{**}P < 0.002$, $^{\dagger}P < 0.01$. (**C**) Cryosections of IEC *C1galt1*$^{-/-}$ colons at different ages stained with mAbs to granulocytes (Ly-6G), macrophages (F4/80), and TNF. Insets (original magnification, ×400) highlight TNF^{+} granulocytes and macrophages. Scale bars: 50 μm. Data are representative of at least 3 experiments.

Figure S1-2

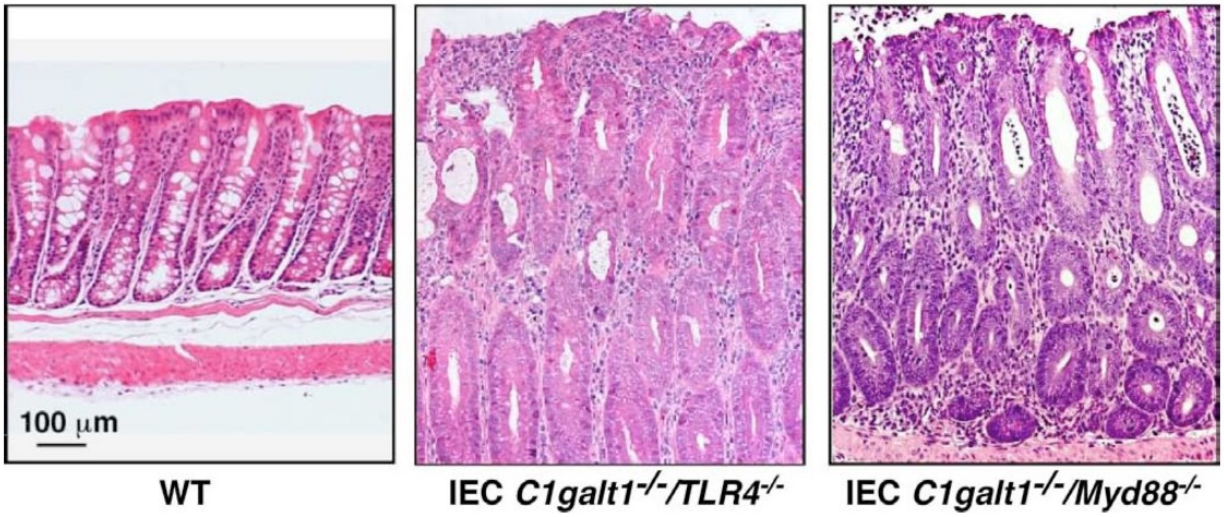

100 μm
WT
IEC *C1galt1*$^{-/-}$/*TLR4*$^{-/-}$
IEC *C1galt1*$^{-/-}$/*Myd88*$^{-/-}$

Figure S1-2. Myd88 and TLR4 are dispensable in the pathogenesis of colitis in *O*-glycan-deficient mice. We bred IEC *C1galt1*$^{–/–}$ with mice lacking Myd88, an universal adaptor protein used by most TLRs, or TLR4, which recognizes bacterial antigens, and found that neither Myd88 nor TLR4 deficiency protected IEC *C1galt1*$^{–/–}$ mice from developing colitis. H&E-stained representative colonic sections indicate similar colonic inflammation in 20-week-old IEC *C1galt1*$^{–/–}$ mice with or without Myd88 or TLR4 deficiencies. Scale bar, 100 μm.

SUPPLEMENTAL CHAPTER 2

System Biology Integrating Microbiome and Metabolome Profiling of the Gut Ecosystem

Abstract

Background

Consistent compositional shifts in the gut microbiota are observed in IBD and other chronic intestinal disorders and may contribute to pathogenesis. The identities of microbial biomolecular mechanisms and metabolic products responsible for disease phenotypes remain to be determined, as do the means by which such microbial functions may be therapeutically modified.

Methods

The composition of the microbiota and metabolites in gut microbiome samples in 47 subjects were determined. Samples were obtained by endoscopic mucosal lavage from the cecum and sigmoid colon regions, and each sample was sequenced using the 16S rRNA gene V4 region (Illumina-HiSeq 2000 platform) and assessed by UPLC mass spectroscopy. Spearman correlations were used to identify widespread, statistically significant microbial-metabolite relationships.

Results

Procrustes and coinertia analysis indicated that inter-omic syntropy existed between mucosal microbiome and metabolome. Such inter-omic relationship was stronger in the cecum than in the sigmoid. OTUs from the Firmicutes and Proteobacteria clades were particularly influential to the inter-omic relationship. The corresponding metabolome analysis indicated metabolites associated with amino acid, porphyrin, and chlorophyll metabolism are important to the inter-omic relationship.

Conclusion

The results suggest that microbes are syntropic with mucosal metabolome composition and therefore may be the source of and/or dependent upon gut epithelial metabolites. The finding that certain metabolites strongly correlate with microbial community structure raises the possibility of targeting metabolites for monitoring and/or therapeutically manipulating microbial community function in IBD and other chronic diseases. This study represents one of the first successful integrations of different microbiome components in the adult colonic mucosa.

Introduction

The intestinal mucosal surface is the site of a complex orchestration of immunologic, metabolic and ecological forces that drive microbial community structure. In most cases, these forces balance the composition of the gut microbiota with mucosal health, facilitating normal nutrient absorption, local and systemic endocrinology, epithelial barrier function, immune development and gut homeostasis (100, 155-157). However, the immunological and functional state of the mucosa is influenced by the microbiota, and it is therefore susceptible to detrimental interactions with changes in luminal bacteria (35, 158). The microbial composition is typically well controlled; however, in certain genetically and environmentally susceptible individuals, control of microbial composition is compromised, leading to (or resulting from) clinical manifestations in immune and inflammatory diseases (18, 90, 159, 160).

The intestinal mucosal ecosystem harbors an assortment of host factors, microbiota, and metabolites. The microbial ecology in the context of this molecular milieu is an area of intense study, but to this point it has mainly been probed by the potential (versus expressed) functionality represented by the microbial metagenome (8, 10, 23, 92). A central goal and methodologic challenge in human-associated microbial ecology is to identify dietary, metabolic, and host and microbial factors that drive microbial community structure. Recent work by Jansson and colleagues (105, 161) and our group (162) indicates that components of the mucosal proteome correlate with certain microbial species and reveals intriguing differences between the potential and expressed biochemical pathways detected in microbial communities *in vivo* (105). In twin-pair studies, Crohn's disease-associated differences in fecal metabolites have been detected in parallel with microbial compositional and metagenomic differences in this compartment, and represented biomarkers related to disease state, presumably in part as products of the disease-associated changes in microbial metagenomic function (28, 133, 163). Identification of such relationships is fundamental for interventional strategies to alter microbiota composition in the context of dysbiosis, and have been highlights of recent landmark studies of environment and diet in human fecal microbial composition (18, 92, 163). Indeed, direct analysis of metabolic output by and interactions between microbial species is a burgeoning investigative field, but challenging methodologically, particularly *in vivo (14, 15)*.

Materials and Methods

Sample collection and pre-processing

All enrolled subjects were consented under an approved Institutional Review Board (IRB) protocol from Cedars Sinai Medical Center prior to routine colonoscopy. All subjects underwent bowel preparation with Miralax® prior to colonoscopy. For each sample region, approximately 30mL of sterile water was endoscopically flushed onto the mucosal surface and recollected via aspiration. Samples were obtained from the cecum and sigmoid colon region of each subject. Samples were kept on ice for the duration of the pre-processing that immediately followed sample collection. Samples were centrifuged at 4,000 × g for 10 minutes at 4°C. The supernatant was aliquoted into three 50-mL tubes with equal volumes and frozen at −80°C. The pellets were resuspended in

2 mL of RNAprotect Bacteria Reagent (Qiagen, Valencia, CA, USA), aliquoted into three separate 15-mL conical tubes, centrifuged at 4,000 × g for 10 minutes at 4°C, separated from the supernatant and frozen at −80°C.

High-throughput 16S analysis

DNA was extracted from 93 samples using the PowerSoil DNA Isolation Kit (MO BIO Laboratories, Carlsbad, CA, USA), and a 30-second beat-beating step using a Mini-Beadbeater-16 (BioSpec Products, Bartlesville, OK, USA). High-throughput sequencing analysis of bacterial rRNA genes was performed using extracted genomic DNA as the templates. One hundred-microliter amplification reactions were performed in an MJ Research PTC-200 thermal cycler (Bio-Rad Inc., Hercules, CA, USA) and contained 50 mM Tris (pH 8.3), 500 µg/mL BSA, 2.5 mM $MgCl_2$, 250 µM of each deoxynucleotide triphosphate (dNTP), 400 nM of each primer, 4 µL of DNA template, and 2.5 units JumpStart *Taq* DNA polymerase (Sigma-Aldrich, St. Louis, MO, USA). The PCR primers (F515/R806) targeted a portion of the 16S rRNA gene containing the hypervariable V4 region, with the reverse primers including a 12-bp barcode (164). Thermal cycling parameters were 94°C for 5 minutes; 35 cycles of 94°C for 20 seconds, 50°C for 20 seconds, and 72°C for 30 seconds, and followed by 72°C for 5 minutes. PCR products were purified using a MinElute 96 UF PCR Purification Kit (Qiagen). DNA sequencing was performed using an Illumina HiSeq 2000 (Illumina, Inc., San Diego, CA, USA). Clusters were created using template concentrations of 1.9 pM and PhiX at 65 K/mm^2, which is recommended by the manufacturer for samples with uneven distributions of A, C, G and T. One hundred base-sequencing reads of the 5' end of the amplicons and seven base barcode reads were obtained using the F515/R806 sequencing primers. De-multiplexing, quality control, and operational taxonomic unit (OTU) binning were performed using quantitative insights into microbial ecology (QIIME) (27). The total initial number of sequencing reads was 70,278,364. Low-quality sequences were removed using the following parameters: Q20, minimum number of consecutive high-quality base calls = 100 bp, maximum number of N characters allowed = 0, maximum number of consecutive low-quality base calls allowed before truncating a read = 3. Numbers of sequences removed using the aforementioned quality control parameters were: barcode errors (5,199,568), reads too short after quality truncation (5,545,570), and too many Ns (38,358). Then, 59,494,868 remaining reads were then used to pick OTUs from the GreenGenes reference database, which automatically bins OTUs at 97% identity: 1,536,002 reads were discarded during OTU picking due to alignment failure. After OTU picking, 57,958,866 reads remained.

Solid phase extraction (SPE)

Before cecum and sigmoid lavage aliquots were subjected to metabolomic analysis, they were cleaned with SPE due to the presence of a polymer presumably derived from bowel preparation (bowel preparation often involves polyethylene glycol). The SPE protocol was adopted, modified and made compatible for the downstream mass spectrometry (MS) analysis. MCX cartridges (Waters Corp. Milford, MA, USA) were conditioned with methanol and phosphoric acid prior to use. Each sample was diluted 1:2 in 2% phosphoric acid and loaded on to the MCX cartridge. Samples were incubated with the mix-mod polymer sorbent in the cartridges. The application of vacuum throughout the procedure was kept to the minimum to allow for ample sample/sorbent interaction. The sorbent was then washed with 2% formic acid in water and 10 mL of water. The metabolites were then eluted off the column by subsequent washes with methanol and 5% ammonium hydroxide, dried, and reconstituted in 2% acetonitrile in water.

Mass spectrometry analysis

A 5-μL aliquot of extracted metabolites from each sample was injected onto a reverse-phase 50 × 2.1 mm ACQUITY 1.7-μm C18 column (Waters Corp.) using an ACQUITY UPLC system (Waters Corp.) with a gradient mobile phase consisting of 2% acetonitrile in water containing 0.1% formic acid (A) and 2% water in acetonitrile containing 0.1% formic acid (B). Each sample was resolved for 10 minutes at a flow rate of 0.5 ml/minute. The gradient consisted of 100% A for 0.5 minutes, then a ramp of curve 6 to 100% B from 0.5 minutes to 10 minutes. The column eluent was introduced directly into the mass spectrometer by electrospray. MS was performed on a Q-TOF Premier (Waters Corp.) operating in either negative-ion (ESI-) or positive-ion (ESI+) electrospray ionization mode with a capillary voltage of 3,200 V, and a sampling cone voltage of 20 V in negative mode and 35 V in positive mode. The desolvation gas flow was set to 800 L/h and the temperature was set to 350°C. The cone gas flow was 25 L/h, and the source temperature was 120°C. Accurate mass was maintained by introduction of LockSpray interface of sulfadimethoxine (311.0814 (M+H) + or 309.0658 9M-H)−) at a concentration of 250 pg/μL in 50% aqueous acetonitrile and a rate of 150 μL/minute. Data were acquired in centroid mode from 50 to 850 m/z in MS scanning. Centroided and integrated MS data from the UPLC-TOFMS were processed to generate a multivariate data matrix using

MarkerLynx (Waters Corp.). The data were normalized to total protein and processed using an array of statistical tools such as R, SIMCA P, and an in-house statistical script. The statistically significant metabolites were putatively identified using several online databses such as HMDB, MMCD, KEGG, and Lipidmaps.

Coinertia analysis

Coinertia analysis identifies successive axes of covariance between two datasets involving the same test subjects. Coinertia analysis was performed using the *coinertia* function from the *ade4* R package, applied to eigenvalues of the metabolome and microbiome (165). The significance of RV scores, which are indicative of global similarity, was estimated using the *RV.rtest* function, which performs a Monte Carlo-based estimation on the sum of eigenvalues from a coinertia analysis.

Procrustes analysis

Procrustes analysis analyzes the congruence of two-dimensional shapes produced from superimposition of principal component analyses from two datasets. To remain consistent, we performed Procrustes analysis on the Euclidian distances of eigenvalues for both the microbiome and metabolome using the *Procrustes* function in the *vegan* R package (R package version 2.0-4. http://cran.r-project.org/web/packages/vegan/index.html).

Putative metabolite identity determination

An in-house script called StandAlone BioIdentifier was used to putatively identify ions based on their biological relevance via incorporation of four major small molecule databases: KEGG, HMDB, LipidMaps, and BioCyc. This metabolomic tool has the unique ability to distinguish mammalian metabolites from those of bacterial and plant origin providing an extra degree of confidence in the ions' putative IDs. This user-friendly script allows one to choose from several positive and negative adducts at a user-predefined mass tolerance. For our UPLC/MS setup we chose H+ and Na+ adducts for the ESI+ mode and H- and Cl- for the ESI- mode at a predefined mass window of 20 ppm.

Results

To measure the composition, function and interdependence of the colonic microbiome and metabolome, a serial cross-sectional study was performed on human subjects undergoing screening colonic endoscopy: 93 mucosal water-lavage samples from the sigmoid and cecum regions of 47 subjects between the ages of 20 and 83 years (mean 61, SD 14.2) (Figure S2-1, Table S2-1). For this pilot study, we included forty-two healthy subjects and five subjects with clinically quiescent Crohn's disease. We opted to collect mucosal rather than fecal samples, due to the distinct composition and intimate relationship of the mucin-associated microbiota with the colonic epithelium (18, 23, 166); lavage sampling permitted efficient recovery of extracellular biosynthetic products present at the mucosal surface (22). Bacteria were separated from supernatants via centrifugation and the two sample components were analyzed separately for 16S microbiome and metabolome composition (22, 162). However, contaminating polymer, presumably polyethylene glycol from bowel preparation, from metabolic aliquots was removed using SPE. Cell-free supernatants were analyzed for metabolic content via UPLC-MS with concomitant *in silico* filtering and thresholding. Cell pellets were analyzed for microbial abundance and composition using the Illumina-HiSeq 2000 platform in combination with the QIIME software suite. Microbial OTUs were rarified to 30,000 reads per sample to reduce noise in downstream analyses. Phylotypic analysis revealed the relative abundance of 2,473 cecum and 2,595 sigmoid GreenGenes reference database-picked OTUs binned at 97% sequence similarity, and thresholded on detection in at least two samples. Metabolome analysis revealed 649 and 576 metabolite peaks detected in the cecum and sigmoid colon, respectively. Stringent comparison of conserved metabolite masses and retention times between the cecum and sigmoid datasets revealed 342 metabolites present in both the cecum and sigmoid datasets. Furthermore, using putative metabolic IDs from mass, many putative metabolites observed were located at the terminus of metabolic pathways, suggesting enrichment for end products. However, many metabolites had more than one possible putative ID, making precise quantification of end products difficult. However, previous studies have suggested the colonic microbiota contributes to a large representation of metabolic end products (167, 168). Given the large number of metabolites observed, we did not attempt to biochemically validate molecular identities. Instead, we investigated if any inter-omic syntropy could be detected using computational approaches.

Overview of the measured mucosal microbiome and metabolome

To determine whether any inter-omic syntropy existed, we first generated two-dimensional principal component distribution plots (PC1 and PC2), with either red (microbiome) or green (metabolome) spots representing each study participant in each data set, and then measured their inter-omic relatedness using Procrustes analysis

(Figure S2-2). Procrustes analysis superimposes and scales principal component plots and allows quantification of non-random congruence between two different measurements from a single group of subjects. To simplify comparisons, Euclidean distances were used in calculating principal components of the microbiome and metabolome constituents. We then performed inter-omic Procrustes analysis on the microbiome and metabolome. Inter-omic Procrustes on cecal samples revealed a strong similarity (Figure S2-2 C: Monte Carlo P ≤0.007), while the sigmoid microbiome and metabolome were less similar, though still significant (Figure S2-2 F: Monte Carlo P ≤0.045). These findings are consistent with recent studies involving the fecal compartment (91, 105). To further confirm this, we also performed coinertia analysis (165). The coinertia *RV* coefficient is a number between 0 and 1; higher numbers are indicative of more global similarity between two datasets (and for which significance values can be determined). The *RV* scores and Monte Carlo *P*-values were 0.67 and 0.01 for the cecum data, and 0.6 and 0.07 for the sigmoid data, respectively. Excluding subjects <50 years old or subjects with inflammatory bowel disease (IBD) did not increase significance of the inter-omic comparisons using Procrustes or co-inertia analysis, suggesting that age and disease status were not strong drivers of the inter-omic relationship in this cohort. Thus, both Procrustes and coinertia metrics indicated that the inter-omic relationship was stronger in the cecum than in the sigmoid. In addition, microbes with the strongest inter-omic covariance, as predicted by the coinertia analysis, are shown in Figure S2-2 G and H. This analysis suggested OTUs from the Firmicutes and Proteobacteria clades were particularly influential to the inter-omic relationship. The corresponding metabolome analysis is available in Figure S2-3. Despite the lack of KEGG assignments for metabolites, this analysis indicated metabolites associated with amino acid, porphyrin, and chlorophyll metabolism are important to the inter-omic relationship. Overall, these results supported our hypothesis that significant inter-omic interdependence existed between the metabolome and microbiome. We therefore sought to more clearly define this relationship.

Discussion

With the advent of next-generation sequencing platforms, a major influx of studies have sought to identify microbial composition differences in various habitats. However, such studies rarely consider environmental variables, such as metabolites or proteins, resulting in incomplete systemic clarity and potentially erroneous assumptions. This study represents one of the first successful attempts to integrate components of the adult gut mucosal ecosystem. We chose to perform analysis on two distinct colonic regions to ensure reproducibility of findings. Notably, all mucosal samples were collected from subjects who had undergone bowel preparation. While standard for both clinical and research endoscopy, bowel preparation is known to alter microbial alpha and beta diversity (169). Accordingly, such depletion of mucosal microbiota is likely to reduce the scope of detectable inter-omic relationships. However, we reason that the observed relationships are representative of the native mucosa. Also, bowel preparation should result in less dietary and enteric secretion input from the proximal intestine, thereby increasing biogeographic resolution and decreasing noise from dietary metabolites.

Nonetheless, it is possible that bowel preparation introduces metabolic changes in the microbial community that elicits non-physiologic inter-omic relationships. Therefore, the scope and quality of these inter-omic relationships merit additional assessment in undisturbed mucosal sites.

A central finding of this study was the rich network of significant correlations between the microbiome and metabolome. Such correlation structure likely arises from a combination of two general processes: 1) catabolism and anabolism of metabolites by microbes, and 2) stimulation and inhibition of microbial growth by metabolites. Indeed, it is widely accepted that dietary alteration is accompanied by shifts in gut microbiome composition and that microbial composition influences the intestinal metabolome (14, 91, 92). However, the metabolites and metabolic pathways involved in such processes are unknown. Therefore, while it is difficult to conclusively assign cause and effect to correlation data, a central goal of this study was to determine whether bioinformatic signatures of either process could be detected.

Figures and Tables

Figure S2-1

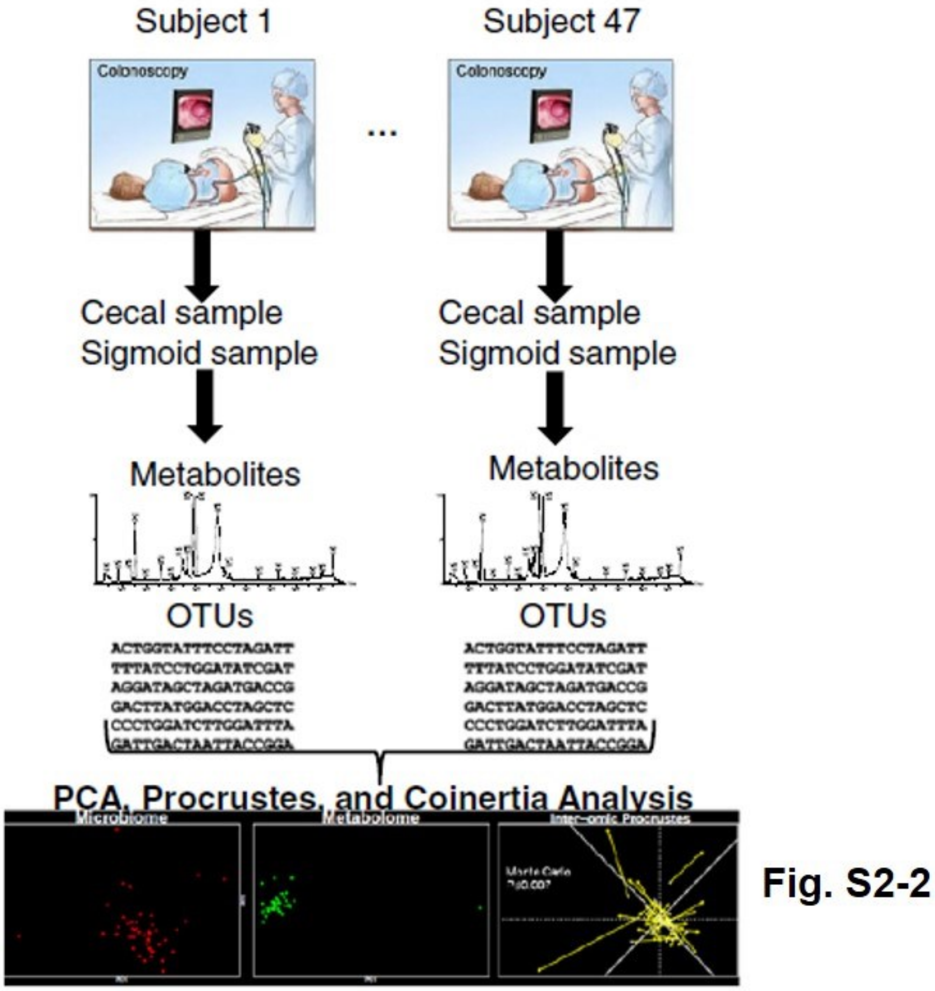


Figure S2-1. Procedural schematic. Endoscopic lavage samples were collected from the cecum and sigmoid colon of each subject. The microbial and metabolic components of each sample were analyzed using Illumina-HiSeq 2000 and ultra-performance liquid chromatography (UPLC)-mass spectrometry (MS), respectively. The analytic pipeline thereafter is shown. See methods for additional details. OTU: operational taxonomic unit; PCA, Principal Component Analysis.

Figure S2-2

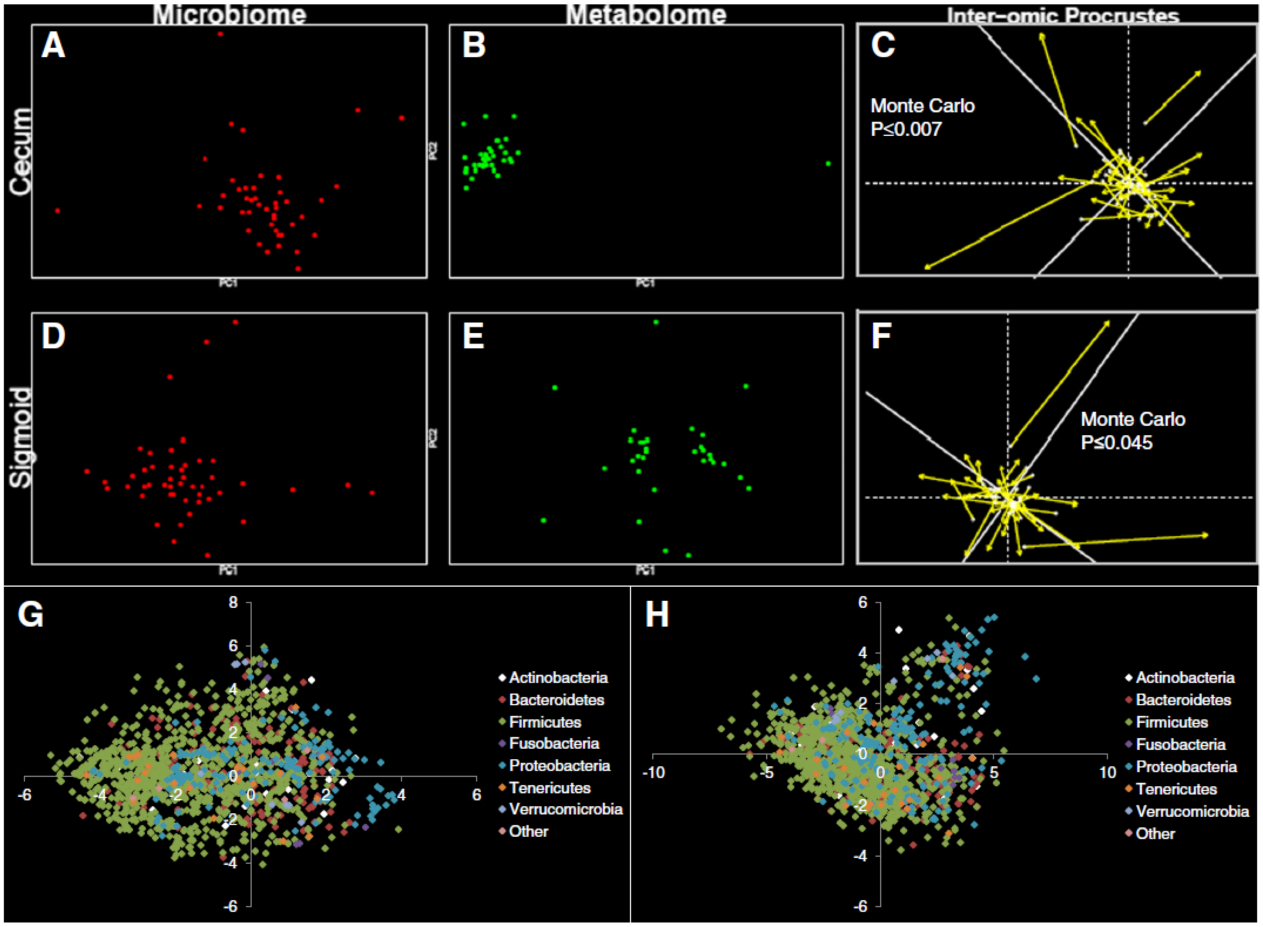


Figure S2-2. Principal component, Procrustes and coinertia analysis. The first column of plots contains microbiome data (red spots) and the second column contains metabolome data (green spots). The first row contains cecal data and the second row contains sigmoid data. Principal component analysis was performed on the cecum microbiome (A), cecum metabolome (B), sigmoid microbiome (D) and sigmoid metabolome (E). Inter-omic (C and F) Procrustes analysis was then performed. Longer lines on Procrustes plots indicate more within-subject dissimilarity of the microbiome and metabolome. Significance values shown were calculated using the protest function from

the vegan R package, which performs repeated symmetric Procrustes analysis to estimate significance. (G) and (H) show operational taxonomic unit (OTU)-level coinertia analysis. Individual OTUs are plotted based on their cointeria-predicted covariance with the metabolome from the cecum (A) and sigmoid (B). To reduce noise in this visualization, the data were thresholded such that only OTUs measured above background in ≥18% of samples are shown. Distance from the center is indicative of the strength of covariance.

Figure S2-3

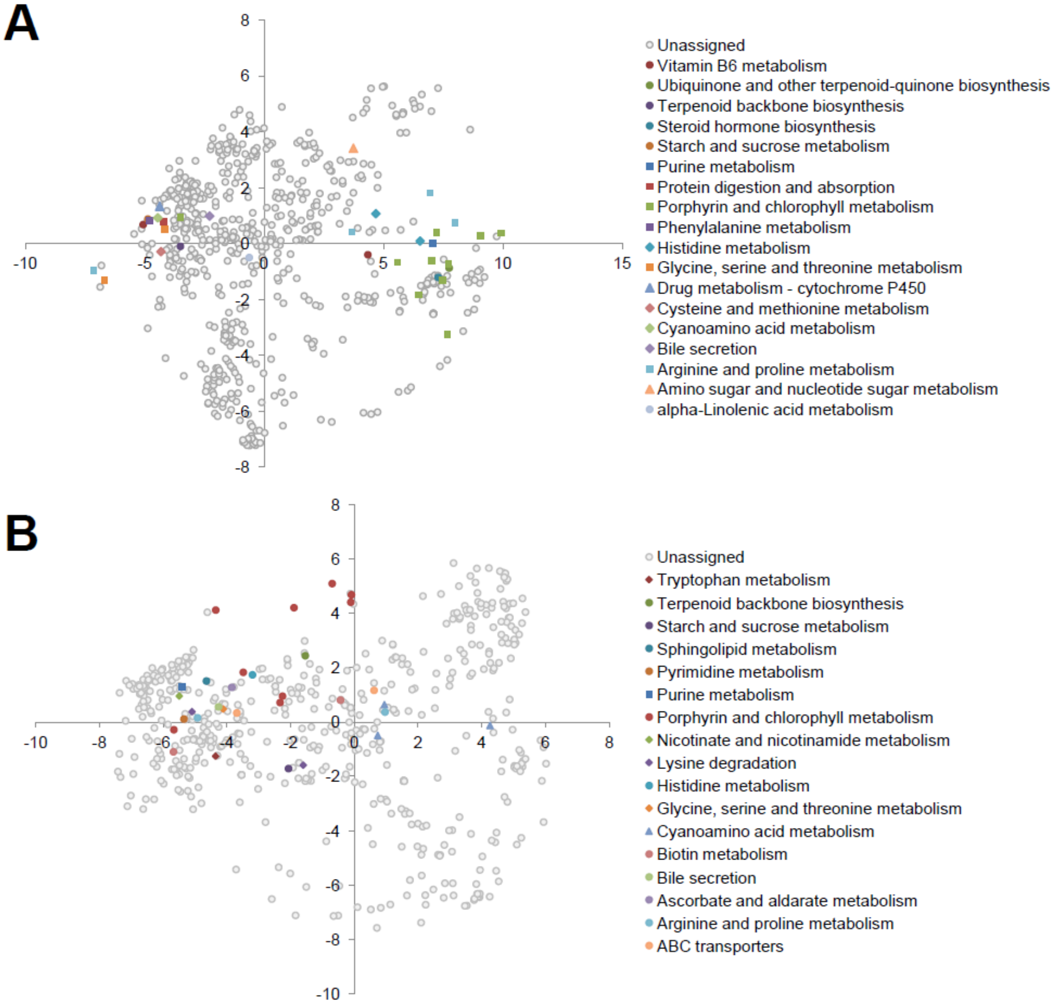


Figure S2-3. Metabolite-level coinertia analysis. Individual metabolites are plotted based on their coinertia-predicted covariance with the microbiome from the cecum (A) and sigmoid (B). To reduce noise, data were first thresholded such that only metabolites measured above background in ≥18% of samples were analyzed. Distance from the center is indicative of the strength of covariance. Technical limitations limited assignment of putative IDs, and thus Kyoto Encyclopedia of Genes and Genomes (KEGG) pathways, to metabolites, so only a minority of metabolites are labeled.)

Table S2-1

Table S2-1. Samples and subjects used for analysis

		Men	**Women**	**Total**
Subjects	Healthy, number	28	14	42
	Crohn's Disease, number	0	5	5
	Average age, years	62.4	58.7	60.9
Samples	Cecum samples, number	27	19	46
	Sigmoid samples, number	28	19	47

SUPPLEMENTAL CHAPTER 3

HIV-Infection Associated Dysbiosis and Functional Changes of Intestinal Microbiome

Abstract

Background

Regardless of infection route, the intestine is the primary site for HIV-1 infection establishment and results in significant mucosal CD4+ T lymphocyte depletion, induces an inflammatory state that propagates viral dissemination, facilitates microbial translocation, and fosters establishment of one of the largest HIV reservoirs. Here we test the prediction that HIV infection modifies the composition and function of the mucosal commensal microbiota.

Methods

Rectal mucosal microbiota were collected from human subjects using a sponge-based sampling methodology. Samples were collected from 20 HIV-positive men not receiving combination anti-retroviral therapy (cART), 20 HIV-positive men on cART and 20 healthy, HIV-negative men. Microbial composition of samples was analyzed using barcoded 16S Illumina deep sequencing (85,900 reads per sample after processing). Microbial metagenomic information for the samples was imputed using the bioinformatic tools PICRUST and HUMAnN.

Results

Microbial composition and imputed function in HIV-positive individuals not receiving cART was significantly different from HIV-negative individuals. Genera including *Roseburia, Coprococcus, Ruminococcus, Eubacterium, Alistipes* and *Lachnospira* were depleted in HIV-infected subjects not receiving cART, while *Fusobacteria, Anaerococcus, Peptostreptococcus* and *Porphyromonas* were significantly enriched. HIV-positive

subjects receiving cART exhibited similar depletion and enrichment for these genera, but were of intermediate magnitude and did not achieve statistical significance. Imputed metagenomic functions, including amino acid metabolism, vitamin biosynthesis, and siderophore biosynthesis differed significantly between healthy controls and HIV-infected subjects not receiving cART.

Conclusion

HIV infection was associated with rectal mucosal changes in microbiota composition and imputed function that cART failed to completely reverse. HIV infection was associated with depletion of some commensal species and enrichment of a few opportunistic pathogens. Many imputed metagenomic functions differed between samples from HIV-negative and HIV-positive subjects not receiving cART, possibly reflecting mucosal metabolic changes associated with HIV infection. Such functional pathways may represent novel interventional targets for HIV therapy if normalizing the microbial composition or functional activity of the microbiota proves therapeutically useful.

Introduction

HIV-1 transmission and replication occur primarily at mucosal sites. There is increasing recognition that HIV-1 infection is substantially a mucosal disease with systemic manifestations(170). Regardless of infection route, gut-associated lymphoid tissue (GALT) is the major site of virus replication early in HIV infection due to local retention, enhanced activation states and increased memory immune function of GALT CD4+ T lymphocytes as compared to peripheral blood mononuclear cells (PBMCs) (171-173). Accordingly, acute HIV infection results in massive depletion of GALT CD4+ T lymphocytes with slow and only partial reconstitution with combination anti-retroviral therapy (cART) (170, 174-177). Other reports have indicated complete intestinal reconstitution of CD4+ T cells in subjects receiving cART that have sustained undetectable HIV replication for many years (178). In health, lymphocyte-mediated inflammatory processes naturally occur, in part, as a response to ongoing luminal antigenic stimulation (179) and help shape immune and inflammatory responses (180, 181). However, HIV-induced immunological imbalance results in pathological manifestations including first acute and then chronic mucosal tissue inflammation, systemic immunological activation, increased epithelial permeability and systemic microbial translocation (182-185). Indeed, local mucosal (and distant, peripheral) inflammation has emerged as a key process in HIV infection, dissemination, pathogenesis and possibly perpetuation (186). Thus, strategies to reverse or reduce HIV-specific as well as more generalized subsequent inflammation could help prevent HIV infection sequelae and dissemination.

One potential source of such inflammation is intestinal bacteria. Commensal microbiota have major effects on the biologic state of the host cell types in the mucosal compartment. They modulate epithelial processes controlling stem cell replenishment, barrier permeability and microbial intrusion (187, 188), mucosal lymphocyte development and IL-17- and IL-22- dependent immune surveillance to microbial challenge (187, 189, 190), and mucosal myeloid (macrophage, dendritic cell, and neutrophil) microbial surveillance and immune regulation (191-194). Given that HIV is associated with intestinal inflammation and that intestinal microbiota can be altered in inflammatory diseases, intestinal microbial compositional aberrations might be expected in HIV-infected individuals (18, 90). However, only minor differences in abundance of a few specific pathogens have been observed in HIV-infected human feces (195), and larger-scale studies of fecal microbiota involving simian immunodeficiency virus (SIV) indicate no significant compositional differences from healthy controls (196, 197). The discrepancy between expectations and published observations could indicate that: 1) no significant changes in microbial composition occur in HIV-infected subjects, and 2) any major changes in microbial composition of HIV-infected subjects are minor and/or easily masked. The latter might be possible if, for example, mucosal, as opposed to luminal, bacteria were differentially abundant in HIV-infected subjects.

Therefore, to investigate our hypotheses that HIV infection is associated with altered intestinal microbial composition, we utilized a rectal mucosal sampling strategy involving small, anally-inserted sponges to absorb mucosal-derived bacteria. These samples, collected from three human cohorts, enabled assessment of whether the intestinal mucosal microbiome composition or function varied with HIV infection and allowed investigation of the ecologic influence of cART.

Materials and Methods

Subject recruitment

The protocol was designed by the investigators and approved by the UCLA Office of the Human Research Protection Program Institutional Review Board (UCLA IRBs #11-001592 and #10-000750) with all participants providing written informed consent. Protocol-based inclusion criteria required that only men age ≥18 years were recruited into this pilot study. Subjects were divided into three groups: 20 healthy HIV-negative control subjects; 20 healthy HIV-positive subjects on chronic cART; and 20 healthy HIV-positive subjects not on cART (cART-naïve or not on cART for ≥3 months). Exclusion criteria included: being female; having history of inflammatory bowel disease (IBD); having any active inflammatory conditions affecting the rectum; and use of rectally administered medications, including over-the-counter enemas, within 48 hours.

Sample procurement

All subjects (as specified above) were seen once in the UCLA Digestive Diseases Clinic for sample collection. Following a brief history, physical examination and confirmation of inclusion/exclusion criteria, as well as confirmation of subject-reported HIV serostatus (PCR and HIV-1 antibody test), subjects received a clinician-applied preparatory enema (118-ml saline enema). Subjects were asked to retain the fluid for at least 5 minutes and then expel the fluid into a toilet. While this procedure could conceivably disturb mucosal surface contents, it was deemed necessary to eliminate stool that might otherwise interfere with sponge placement. Following a 15 minute rest, mucosal sampling by sponge collection took place, using previously reported methods used for secreted antibodies and cytokines (198, 199). Briefly, two ophthalmic eye spears (Beaver Visitec, Waltham, MA, USA) were simultaneously inserted into the rectum via anoscope, as previously reported (198, 199), and allowed to absorb mucosal material for 5 minutes. Samples were immediately placed on ice and transported to the laboratory for immediate processing.

Sample preparation and 16S V4 sequencing analysis

Sponges were removed from their plastic stems and individually placed in 0.5-ml tubes (Eppendorf, Hauppauge, NY, USA), which had the distal end previously pierced using a sterile 18-gauge needle (BD Biosciences, San Jose, CA, USA); each of these individual tubes were then placed into 2-ml tubes (Eppendorf, Hauppauge, NY, USA). Bacteria were quickly eluted and pelleted by adding 100 µl of 25 mM HEPES, 50 mM NaCl, 1% Triton-X, 1 mM DTT, and 5 mM EDTA and centrifuging in an Eppendorf 5415D centrifuge (Eppendorf, Hauppauge, NY, USA) for 30 s. This collection step was repeated with another 100 µl of the elution buffer above. Supernatant was immediately removed and pellets were frozen in a -80° freezer (Model: ELT1786-9-D40, Thermo Scientific, Asheville, NC, USA) with a backup phone system, until further processing.

Genomic DNA was extracted from the 60 samples using the PowerSoil DNA Isolation Kit (MO BIO Laboratories, Carlsbad, CA, USA), and a 30-second beat-beating step using a Mini-Beadbeater-16 (BioSpec Products, Bartlesville, OK, USA). High-throughput sequencing analysis of bacterial rRNA genes was performed using extracted genomic DNA as the templates. One hundred microliter amplification reactions were performed in an MJ Research PTC-200 thermal cycler (Bio-Rad Inc., Hercules, CA, USA) and contained: 50 mM Tris (pH 8.3), 500 µg/ml bovine serum albumin (BSA), 2.5 mM MgCl2, 250 µM of each deoxynucleotide triphosphate (dNTP), 400 nM of each primer, 4 µl of DNA template, and 2.5 units JumpStart Taq DNA polymerase (Sigma-Aldrich, St Louis, MO, USA). The PCR primers (F515/R806) targeted a portion of the 16S rRNA gene containing the hypervariable V4 region, with the reverse primers including a 12-bp barcode (164). Thermal cycling parameters were 94°C for 5 minutes; 35 cycles of 94°C for 20 seconds, 50°C for 20 seconds, and 72°C for 30 seconds, followed by 72°C for 5 minutes. PCR products were purified using a MinElute 96 UF PCR Purification Kit (Qiagen, Valencia, CA, USA). DNA sequencing was performed using an Illumina HiSeq 2000 (Illumina, Inc., San Diego, CA, USA). Clusters were created using template concentrations of 1.9 pM and PhiX at 65 K/mm^2, (manufacturer's recommendations for samples with uneven distributions of A, C, G and T). One hundred base sequencing reads of the 5′ end of the amplicons and seven base barcode reads were obtained. De-multiplexing, quality control, and operational taxonomic unit (OTU) binning were performed using Quantitative Insights into Microbial Ecology (QIIME) (27).

The total initial number of sequencing reads was 71,581,480. Low quality sequences were removed using the following parameters: Q20, minimum number of consecutive high-quality base calls = 100 bp, maximum number of N characters allowed = 1, maximum number of consecutive low-quality base calls allowed before truncating a read = 3. Numbers of sequences removed using the aforementioned quality control parameters were: barcode not in mapping file (35,296,547), reads too short after quality truncation (4,926,462), and too many Ns (5,431). Remaining reads numbered 31,353,040, which were then used to pick OTUs from the GreenGenes reference database (May 18, 2012 database); this database automatically bins OTUs at 97% identity, ensuring the resulting data were compatible with phylotypic investigation of communities by reconstruction of unobserved states (PICRUSt) analysis. Due to alignment failure, an additional 1,511,116 reads were discarded during OTU picking, providing 29,841,924 reads for downstream analysis.

Rarefaction and diversity analysis

After picking OTUs from the GreenGenes reference database, rarefaction was performed to 85,900 (corresponding to the sample with the fewest reads) reads per sample using the QIIME software suite (version 1.6) running on an Ubuntu virtual machine (27). Alpha diversity metrics used included Phylogenetic Diversity, Chao1, observed species and Shannon index. For all sampling depths, each plotted point represents the average of ten random samplings. The comparison of alpha diversity between the three groups was performed using the two-sided Student *t*-test. Beta diversity analysis was performed in QIIME and utilized unweighted UniFrac distances to estimate sample distributions. Adonis significance analysis was performed for each pairwise comparison of sample groups using the Adonis function from the vegan R package.

Taxonomic analyses

Microbial composition at each taxonomic level was defined using the summarize_taxa function in QIIME. Prior to all analysis of variance (ANOVA), taxa at each taxonomic level were thresholded such that any taxon present in fewer than 20% of samples was discarded.

Statistical analysis

All statistical analyses were conducted using R (http://www.r-project.org/). HIV- and cART-associated microbial changes were calculated using ANOVA with multiple comparison correction using *q*-values (R package qvalue). Associations between imputed metagenomic functions and HIV infection were calculated using Kruskal-Wallis ANOVA and corrected for multiple comparisons using*q*-values. All taxonomic associations reported were significant q <0.15 unless stated otherwise. All metagenomic associations reported were significant at q <0.25.

Metagenomic imputation

PICRUSt is a well-documented tool designed to impute metagenomic information based on 16S input data (http://picrust.github.io/picrust). Sample metagenomic imputation was performed using the default settings of PICRUSt (version 0.9.1). The resulting metagenomic data were then entered into the HMP unified metabolic analysis network (HUMAnN) pipeline (version 0.98) (55) to sort individual genes into Kyoto encyclopedia of genes and genomes (KEGG) pathways representing varying proportions of each imputed sample metagenome. Both PICRUSt and HUMAnN analyses were performed using the terminal interface of a QIIME virtual machine running the Ubuntu operating system.

Results

Rectal mucosa microbial sampling strategy

In place of bowel preparation, all subjects first received a saline enema to void solid fecal contents from the rectal vault. Two absorbent ophthalmic sponges were then applied, under direct vision via anoscope, to opposing sides of the rectal mucosa (Figure S3-1), enabling collection of measureable levels of bacteria, protein and metabolites. Despite their small volume, eluate from each sponge allowed recovery of approximately 5 × 10^7 ± 4.4 × SD 10^7 of bacterial cells and approximately 10^5 μg ± SD 75.9 of protein via bacterial hemocytometer and Bradford assay, respectively. To account for possible micro-biogeographic variations in microbial composition or abundance, material obtained from two sponges from each subject were pooled for further analyses. We did not attempt to determine whether this sampling methodology artificially enriched or depleted certain microbes, proteins, or metabolites, so we could not eliminate the possibility.

Samples were collected from 20 healthy controls (HC), 20 healthy HIV-positives on cART (cART(+)) and 20 healthy HIV-positive subjects off cART (cART(-)) for at least three months (Table S3-1). All collected subject metadata including age, ethnicity, viral loads, serum CD4+ T cell counts, durations of infection, and prescribed cART drug classes are tabulated in Table S3-1. Importantly, to eliminate the potential influences of gender, only men were recruited for this pilot study. Microbial components were lysed and analyzed for V4 16S composition using barcoded Illumina-HiSeq 2000 V4 sequencing. OTUs were then picked at a similarity threshold of 97% using the

GreenGenes reference database, such that 497,365 ± SD 119,950 reads were retained for each sample. Sequenced reads were then rarified to 85,900 reads per sample, yielding a total of 3,281 OTUs.

HIV infection alters mucosal microbial diversity

To determine whether HIV infection altered microbial diversity of the rectal mucosa, alpha and beta diversity metrics were analyzed. As groups, the HC and cART(+) subjects revealed very similar alpha diversity rarefaction profiles. However, cART(-) subjects exhibited significant reduction of alpha diversity using the Chao1 diversity metric, which accurately estimates OTU richness for microbial communities, compared with HC, indicating HIV infection is associated with a potential collapse in alpha diversity (two tailed*t*-test, $P \leq 0.05$ at all sampling depths >17,188) (Figure S3-2 A). A similar, though insignificant trend was observed between cART(+) and cART(-) subjects (two tailed *t*-test, $P = 0.05, 0.1$ for all sampling depths >10). Together, these derivative data suggest HIV infection resulted in a slight reduction of alpha diversity in cART(-) subjects that was reversed to near equivalence with HC in cART(+) subjects. Beta diversity analysis was then performed using the unweighted unifrac distance metric to determine whether HC, cART(+) or cART(-) subjects differed in their microbial composition. Adonis analysis of the resulting unifrac distance matrix suggested that the microbial compositions of: 1) cART(-) subjects were significantly different from HC subjects ($P = 0.017$); 2) cART(+) and cART(-) subjects overlapped but had slightly different trends ($P = 0.053$); and 3) cART(+) subjects were not statistically different from HC subjects ($P = 0.1$). Principal coordinate analysis (PCoA) of the first and fourth principal components of the unweighted unifrac distance matrix allowed visualization of these differences (Figure S3-2 B). The variation captured in the first, third and fourth principal components (14%, approximately 5%, and 4% of total variation, respectively) of the unweighted unifrac analysis varied significantly between cART(-) and HC subjects (Kruskal-Wallis, all *P* ≤0.05). Excluding age (Pearson correlation with principle coordinate 1 (PC1), *P* <0.001), which is known to correlate with changes in intestinal microbiota composition (8, 11, 18), no other subject-reported metadata (including serum CD4 levels, serum viral titers, duration of infection, and ethnicity) significantly correlated with any of the first five principal components of the combined data. Thus, both alpha and beta diversity analysis suggested that HIV infection resulted in ecological changes relative to healthy controls that were partially normalized in cART(+) subjects.

As an initial confirmation of the beta diversity predicted differences, the microbial composition of samples from HIV-infected (both cART(+) and cART(-)) subjects was compared with that of HC subjects at the phylum level. Significant phylum level differences were corrected for multiple comparisons using a significance cutoff of *q* <0.15. Samples obtained from cART(-) subjects were enriched with Fusobacteria (ANOVA, *q* = 0.1) and depleted of Firmicutes (ANOVA, *q* = 0.058), compared to HC samples (Figure S3-2 C). However, cART(+) subjects displayed only intermediate enrichment of Fusobacteria (ANOVA, *q* = 0.27) and depletion of Firmicutes (ANOVA, *q* = 0.27). Therefore, this analysis also suggested microbial composition in cART(-) subjects was significantly different from HC and that the composition of cART(+) was partially, though incompletely, normalized to that of HC.

Imputed metagenomic metabolic functions significantly vary with HIV infection in the absence of cART

Having identified distinct microbial changes in cART(-) subjects compared to HC subjects, concomitant functional metagenomic differences might also be expected. Metagenomic composition is traditionally defined using shotgun sequencing. In the absence of measured metagenomic sequencing data, PICRUSt (http://picrust.github.io/picrust *webcite*) in combination with HUMAnN was used to bioinformatically impute sample metagenomes and determine relative genomic abundances of KEGG metabolic pathways from all subjects, respectively. PICRUSt allows imputation of most microbial genomes present in each sample based on sequence similarity of input GreenGenes sequences to sequenced reference genomes. When combined with HUMAnN, a separate bioinformatic tool that organizes metagenomic data into relative abundances of KEGG pathways per sample, the resulting data are highly comparable to sequenced metagenomic data and observed metabolomic data (3, 18).

Using these bioinformatic tools, metagenomic functions were compared between cART(-), cART(+) and HC subjects. For these metagenomic analyses, the *q*-value threshold for correcting multiple comparisons was relaxed to include $q < 0.25$ comparisons in agreement with previous studies using this methodology (18). Like the analyses above, no significant differences were observed between cART(+) and HC subjects or between cART(+) and cART(-) subjects after correction for multiple comparisons. However, 10 KEGG pathways were significantly different between cART(-) and HC subjects (Kruskal-Wallis, $q < 0.25$). Plotting these pathways with respect to cART status revealed that the distribution of these pathways was similar to the distribution of enriched and depleted genera (Figure S3-3); While cART(-) subjects exhibited the most enrichment or depletion of each pathway, cART(+) subjects had intermediate levels of metagenomic pathway abundance relative to HC (Figure S3-3). Overall, the metagenomes of cART(-) subjects tended to be depleted of amino acid production, amino acid metabolism, CoA biosynthesis, and fructose/mannose metabolism compared with HC subjects. Instead, the microbiota of cART(-) subjects were metagenomically enriched for glutathione metabolism, selenocompound metabolism, folate biosynthesis and siderophore biosynthetic genes. These results indicate that HIV infection in the absence of cART results in significant functional metagenomic differences that are not fully restored with cART. Such functional differences may reflect the functions that HIV-infected mucosa select for and could have downstream implications on vitamin and nutrient availability for the host.

Discussion

This study represents one of the first attempts to define changes in the mucosal microbiota composition in HIV infection and revealed several HIV-dependent changes in the mucosal microbiota. Significant functional and compositional differences in rectal microbiota were observed between cART(-) and HC subjects; these were incompletely normalized by cART in cART(+) subjects. However, an HIV-associated reduction in alpha diversity was adequately normalized by cART. The differing imputed metagenomic compositions between cART(-) subjects and HC subjects suggested HIV infection altered the rectal ecosystem and selected for different microbial metagenomic functions.

This study also revealed several important, imputed metagenomic functional pathways that varied in abundance between cART(-) and HC subjects. Although previous studies have suggested that intra-subject metagenomic content tends to remain relatively stable over time independent of microbial composition, new studies suggest metagenomic content can vary in the context of certain diseases (8, 12, 18). These results, combined with the observed microbial compositional changes, suggest differing metabolic functions arise based on the different microbial communities more pronounced in HIV-infected subjects. Genes encoding amino acid biosynthesis and metabolism, CoA biosynthesis, selenocompound metabolism, glutathione metabolism and folate biosynthesis were compositionally altered in cART(-) subjects. This imbalance might indicate that free vitamins and nutrients available to HIV-infected hosts might be altered. Interestingly, siderophore biosynthetic genes were enriched in cART(-) subjects. Siderophores act as quorum sensing molecules for gram-negative organisms, so the enrichment of this pathway could be indicative of increased intra- or inter-species communication in cART(-) subjects. cART(-) subjects encoded fewer genes for fructose and mannose metabolism, which may be reflective of altered environmental nutrient availability or the differing metabolic potential of bacterial species that best adapt to such environments. Besides illuminating the metabolic potential of the underlying bacterial community, these pathways might be exploitable to help normalize microbial composition of HIV-infected subjects if microbial remediation strategies prove warranted. Given the role of diet in driving microbial composition, one could imagine exploiting such differences in metabolic function by dietary optimization to enrich for preferred bacteria (91, 92, 113).

Figures and Tables

Figure S3-1

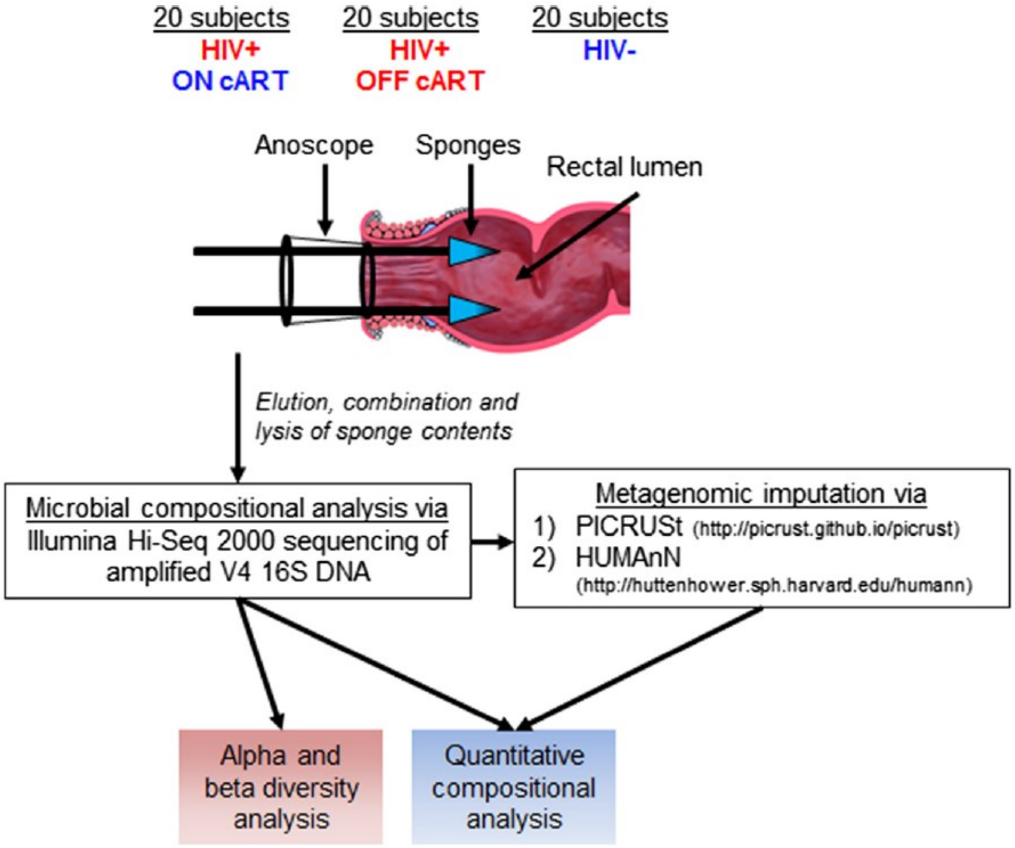

20 subjects
HIV+
ON cART
20 subjects
HIV+
OFF cART
20 subjects
HIV-
Anoscope
Sponges
Rectal lumen
Elution, combination and lysis of sponge contents
Microbial compositional analysis via Illumina Hi-Seq 2000 sequencing of amplified V4 16S DNA
Metagenomic imputation via
1) PICRUSt (http://picrust.github.io/picrust)
2) HUMAnN (http://huttenhower.sph.harvard.edu/humann)
Alpha and beta diversity analysis
Quantitative compositional analysis

Figure S3-1. Schematic of sampling and bioinformatic methodology. Rectal mucosa secretions were collected from 60 human subjects as shown and subjected to high-throughput deep V4 16S sequencing. Alpha and beta diversity analyses were then performed, which suggested microbial composition was altered in HIV-infected subjects who were not receiving combination anti-retroviral therapy (cART). Microbial differences in these subjects were identified and compared with HIV-infected subjects receiving cART. In addition, the different classes of cART were analyzed to determine whether any class

was significantly associated with differences in microbial composition. Imputed metagenomic differences between HIV-infected subjects not receiving cART and healthy control subjects were then identified and compared between the three patient cohorts. PICRUSt, phylotypic investigation of communities by reconstruction of unobserved states; HUMAnN, HMP unified metabolic analysis network.

Figure S3-2

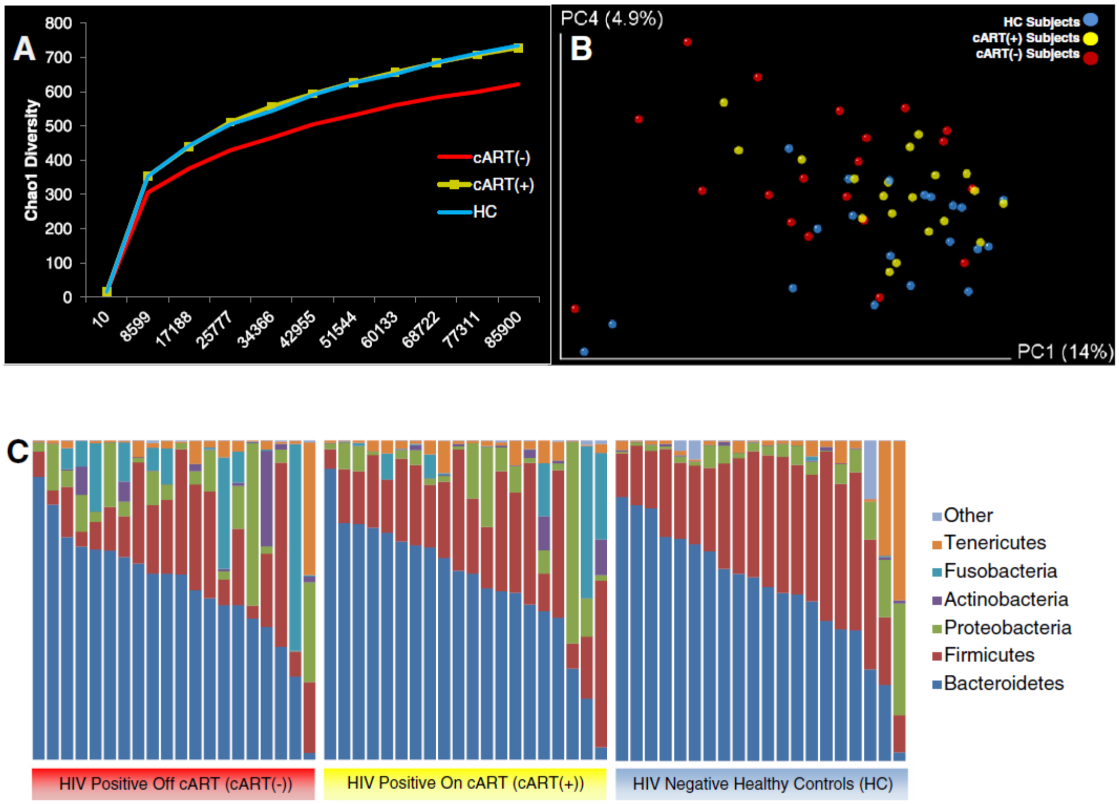


Figure S3-2. Alpha and beta diversity. (A) Chao1 alpha diversity indicated subjects on combination anti-retroviral therapy (cART(-)) had reduced species richness than healthy controls (HC) and subjects on (cART(+)), though only the comparison with HC was significant (*t*-test, *P* ≤0.05 at all sampling depths >17,188). Abundance curves for HC and cART(+) subjects were nearly indistinguishable. **(B)** Beta diversity was analyzed by unweighted unifrac analysis using the first and fourth principal components. These principal components were selected because principal coordinate 1 (PC1) was significantly different between cART(-) and cART(+) subjects (Kruskal-Wallis, *P* = 0.02) and both PC1 and PC4 were significantly different between cART(-) and HC subjects

(Kruskal-Wallis, both *P* <0.05). HC subjects (blue) clustered relatively tightly with cART(+) subjects (yellow), whereas cART(-) subjects (red) were more diffusely scattered along PC1 and PC4 (Adonis, for cART(+) vs cART(-) *P* = 0.06, and for cART(-) versus HC *P* = 0.02). **(C)** Phylum level composition of each subject was sorted based on HIV and cART status. The abundance of Firmicutes was significantly reduced in cART(-) subjects compared with healthy subjects (analysis of variance (ANOVA), *q* = 0.06) while Fusobacteria were significantly enriched (ANOVA, *q* = 0.11).

Figure S3-3

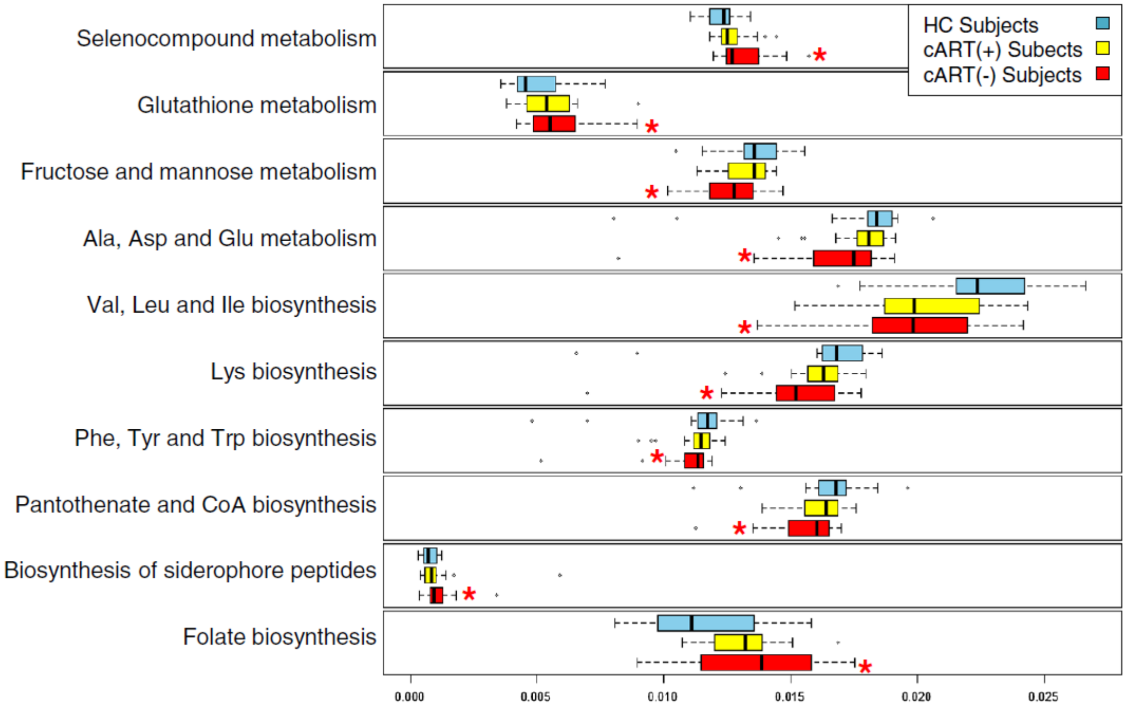


Figure S3-3. Imputed metagenomic differences between subjects not on combination anti-retroviral therapy (cART(-)) and healthy control (HC) subjects. The relative abundance of metabolic pathways encoded in each imputed sample metagenome was analyzed by HIV infection status of each subject using box plots. From these box plots, clear differences are observed between the relative abundance of several imputed metagenomic functions between cART(-) subjects and HC. Significance of each comparison was determined using Kruskal-Wallis one way analysis of variance. Box plots of subjects on cART (cART (+)) are included to provide context for each comparison. Vertical black bars represent group averages. The x-axis represents

the percent abundance of pathways for each imputed sample metagenome. Whiskers represent the interquartile ranges multiplied by 1.5; **q* <0.25 relative to HC.

Table S3-1

Table S3-1. Human subject metadata

	cART(-)	**cART(+)**	**HC**	
n	20	20	20	
Gender	All male	All male	All male	Subject metadata
Age, years	40.9 (± 11.4)	46.3 (± 9.2)	48.6 (±12)	
Viral Loads	158,147 (± 366,197)	3,563 (± 14,333)	N/A	
Years infected	8.9 (± 8.8)	14.2 (± 6.5)	N/A	
Serum CD4 levels	439.6 (± 271.8)	534 (± 246)	NT	
Hispanic	4	1	5	Ethnicity
Black	13	14	9	
Caucasian	3	5	5	
Other	0	0	1	
NNRTI	N/A	9	N/A	cART drugs prescribed
NRTI	N/A	18	N/A	
PI	N/A	12	N/A	
II (raltegravir)	-	4	-	

Note: Results are presented as number or mean (± SD). (cART(-), healthy HIV-positive men off combination anti-retroviral therapy (cART) for at least three months; cART(+), healthy HIV-positive men on (cART); HC, healthy controls; NT, not tested; N/A, not applicable; NNRTI, non-nucleoside reverse transcriptase inhibitor; NRTI, nucleoside reverse transcriptase inhibitor; PI; protease inhibitor; II, integrase inhibitor.

SUPPLEMENTAL CHAPTER 4

HIV-Infection Associated Perturbation of Cervical Vaginal Microbiome

Abstract

Background

HIV infection is associated with bacterial vaginosis, but the relationship between HIV, the vaginal microbiome and mucosal immunity is unclear.

Methods

We profiled the cervicovaginal lavage from 21 HIV-positive and 20 -negative women for microbial composition and imputed metagenomic functions. Microbial composition of samples was analyzed using barcoded 16S Illumina deep sequencing (996,727 ± 193,037 SD reads after processing). Microbial metagenomic information for the samples was imputed using the bioinformatic tools PICRUST and HUMAnN.

Results

Significant differences in alpha and beta diversity were observed between HIV-negative and HIV-positive women, with the latter enriched of organisms associated with bacterial vaginosis and depleted of Lactobacilli. These ecologic changes occurred concomitantly with significant metagenomic and immunologic differences.

Conclusion

These results demonstrate that HIV infection results in a dysregulated mucosal ecosystem where Lactobacilli are replaced or outcompeted by other organisms.

Introduction

Compared with the microbiome of other human anatomic habitats, the vaginal microbiome is highly unique in diversity, structure, and function (8, 200, 201). Unlike most habitats, vaginal microbial communities are highly modular in structure, and women are separated into distinct groups based on the abundance or absence of certain *Lactobacillus* species (110, 200). Five vaginal microbial communities have been documented; four are characterized by their composition of specific *Lactobacillus* species, while the fifth is notably lacking a dominant *Lactobacillus* species and is instead defined by its high species diversity (200). The functional significance of these communities is incompletely understood, though community composition is significantly correlated with vaginal pH and bacterial vaginosis (BV), as defined by the Nugent score (200). Low vaginal pH (3.5 - 4.5), commonly observed in women colonized by *Lactobacillus*-dominant communities, is thought to originate from microbial production of organic acid, particularly lactic acid, and is believed to contribute to regional homeostasis by inhibiting growth of pathobionts (202, 203). Lactobacillus predominance is associated with bactericidal activity of genital tract secretions against *E. coli ex vivo* (204-206). While higher vaginal pH (>4.5) does not preclude health, it is correlated with microbial communities that lack a dominant *Lactobacillus* species and thus may respond differently to environmental stimuli and pathogen invasion (207). Host factors, ranging from pregnancy to ethnicity, appear to play important roles in defining vaginal microbial community structure (200, 208, 209). Indeed, as many as 40% of US Black and Hispanic women are colonized by microbial communities lacking a dominant *Lactobacillus* species (200).

Despite technological and bioinformatic improvements facilitated by high-throughput sequencing, the definition of a “healthy” vaginal microbiota has actually become less clear. For example, BV is a malodorous condition that afflicts millions of women in the US, yet its etiological agent(s) remain uncertain despite numerous correlated species of bacteria (207, 210, 211). Instead, evidence suggests that BV results from, or is potentially caused by, disruption of mucosal homeostasis that allows subsequent enrichment for specific phylotypes (207, 211). Such disruptions likely have relevance in shaping systemic health, as susceptibility to HIV infection has been linked with BV (212) and prevalence of asymptomatic BV is higher in HIV-positive subjects (213, 214).

While a few studies have explored the vaginal microbiome composition in the context of HIV infection, much remains unknown regarding its effects on the vaginal ecosystem (215-217). Microbiome composition of several anatomical habitats (particularly the gut) has been shown to vary with HIV-infection, suggesting commensal microbiome composition is generally altered during HIV infection (5, 218-222).

Materials and Methods

Subject Recruitment

The protocol was designed by the investigators and approved by the Offices of the Human Research Protection Program Institutional Review Board (Einstein IRB #07-469 and 09-547, UCLA IRB #10-000750) with all participants providing written informed consent. Cervicovaginal lavage (CVL) samples were obtained from 21 healthy, HIV-negative women who enrolled in a microbicide safety study and 20 HIV-positive women who participated in a mucosal immunology study. Subjects were ≥ 18 years of age and were excluded for pregnancy, breastfeeding, and active genitourinary infection, including symptomatic BV diagnosed by Amsel's criteria at the time of genital tract sampling.

Sample Procurement

Vaginal pH was measured from a swab of the lateral vaginal wall (Whatman pH paper, pH 3.8–5.5). CVL was performed by washing the cervix and posterior fornix with 10mL of normal saline (pH,5.0). The samples were transported on ice to the laboratory, centrifuged at 700g for 10 minutes and the supernatants divided into aliquots and stored at -80°C.

Sample Preparation and 16S V4 Sequencing Analysis

Bacterial portions of CVL samples were separated from supernatants by centrifugation and frozen in a -80 freezer (Model: ELT1786-9-D40, Thermo Scientific, Asheville, NC, USA) with a back-up phone system until further processing.

Genomic DNA was extracted from the 41 samples using the PowerSoil DNA Isolation Kit (MO BIO Laboratories, Carlsbad, CA, USA), and a 30-second vortexing step using a vortex Genie-2 adapter (MO BIO Laboratories, Carlsbad, CA, USA). High throughput sequencing analysis of bacterial rRNA genes was performed using extracted genomic DNA as the templates. Thirty microliter amplification reactions were performed in triplicate in an GeneAmp PCR System 9700 (Applied Biosystems, Grand Island, NY, USA) and contained: 50 mM Tris (pH 8.3), 2.5 mM MgCl2, 250 μM of each deoxynucleotide triphosphate (dNTP), 400 nM of each primer, 1 μL of DNA template, and

1 unit JumpStart Taq DNA polymerase (Sigma-Aldrich, St. Louis, MO, USA). The PCR primers (F515/R806) targeted a portion of the 16S rRNA gene containing the hypervariable V4 region, with the reverse primers including a 12-bp barcode (164). Thermal cycling parameters were 94°C for 5 minutes; 35 cycles of 94°C for 20 seconds, 50°C for 20 seconds, and 72°C for 30 seconds, followed by 72°C for 5 minutes. Triplicate PCR reactions were then combined and products were purified using a Qiagen PCR Purification Kit (Qiagen, Valencia, CA, USA). DNA sequencing was performed using an Illumina HiSeq 2000 (Illumina, Inc., San Diego, CA) at the UCLA Clinical Microarray Core. One hundred base sequencing reads of the 5' end of the amplicons were obtained. De-multiplexing, quality control, and OTU binning were performed using QIIME (27).

The total initial number of sequencing reads was 62,836,169. Low quality sequences were removed using the default parameters of split_libraries_fastq.py in QIIME. Numbers of sequences removed using these quality control parameters were: reads too short after quality truncation (194,156) and too many Ns (81,011). 40,865,824 reads remained and were then used to pick OTUs from the GreenGenes reference database (May 18, 2012 database) at 97% identity, ensuring the resulting data was compatible with PICRUSt analysis. Due to alignment failure, an additional 3,688,289 reads were discarded during OTU picking resulting in 37,177,535 reads for downstream analysis.

Bioinformatic analyses

Rarefaction and Diversity Analysis: After picking OTUs from the GreenGenes reference database, rarefaction was performed to 500,000 reads per sample, resulting in

2,547 OTUs using the QIIME software suite (version 1.6) running on an Ubuntu virtual machine (27). All analyses involving the microbiome were based on this rarified data.

Metagenomic Imputation: PICRUSt is a well-documented tool designed to impute metagenomic information based on 16S input data (http://PICRUSt.github.com/PICRUSt/). Sample metagenomic imputation was performed using the default settings of PICRUSt (version 0.9.1). The resulting metagenomic data was then input into the HUMAnN pipeline (version 0.98) (55) to sort individual genes into KEGG pathways representing varying proportions of each imputed sample metagenome. Both PICRUSt and HUMAnN analyses were performed using the terminal interface of a QIIME virtual machine running the Ubuntu operating system.

Canonical Correspondence Analysis (CCA): CCA was performed using the cca function in the vegan R package. CCA was performed on the microbiome, imputed metagenome, and proteome with collected subject and sample metadata.

Principal Coordinate Analysis (PCoA): Principal coordinate analysis for the imputed metagenome, microbiome and proteome was performed in R (ape software package (223)) to estimate sample distributions using Bray-Curtis distance matrices. Microbial beta diversity analysis was also performed using weighted UniFrac distances to estimate sample distributions in QIIME. Adonis significance analysis (vegan R package) was utilized to compare subject distribution patterns of HIV-negative and HIV infected subjects.

Iterative analysis of enriched and depleted genera: Microbial composition at the genus level was first defined using the “summarize_taxa” function in QIIME. Genera were then thresholded such that any genus present in less than 20% of all samples was discarded. Iterative ANOVA analysis (which simplified to Student's T-test) was then performed on all genera with respect to HIV-infection status. Multiple comparisons were corrected for using q values. Only comparisons with q values <0.05 were retained and reported.

Iterative analysis of imputed metagenomic functions: Metagenomic functions were defined using PICRUSt and HUMAnN as described above. Pathways not represented in any samples were discarded prior to analysis leaving a total of 179 pathways, down from 186. Iterative Kruskal-Wallis analysis (which simplified to Mann–Whitney U test) was then performed on all pathways with respect to HIV-infection status. Multiple comparisons were corrected for using q values. Only comparisons with q values <0.02 were retained and reported.

Results

To begin to dissect the host and environmental factors that drive systemic function and microbial representation, CVL was examined from 21 HIV-positive and 20 HIV-negative women who were screened for participation in studies of genital tract mucosal immunity. Multiple physiologic and contextual parameters were measured on women and collected samples, respectively. Subject characteristics, CVL immune mediator and anti-HSV and anti-*E. coli* activity data are summarized in Table S4-1.

Bacteria and soluble protein were separated by centrifugation and analyzed for bacterial composition using barcoded Illumina-HiSeq 2000 V4 16S sequencing.

Operational Taxonomic Units (OTUs) were then picked at a similarity threshold of 97% using the GreenGenes reference database such that 996,727 ± 193,037 SD reads were retained for each sample. Sequenced reads were then rarified to 500,000 reads per sample, yielding a total of 2,547 OTUs. Despite the deep sequencing depth, attempts to define vaginal microbial community structure using methods described by Koren *et al*. were unsuccessful, likely due to the relatively small number of samples collected compared with the large cohorts used in studies defining such structure (110).

Dysbiosis of the cervical vaginal microbiome associated with HIV infection

To help determine whether microbiota composition co-varied with host factors, including HIV-infection, canonical correspondence analysis (CCA) was performed. As shown in Figure S4-1 A, microbial composition seemed to vary with HIV-infection, vaginal wall pH, and the abundances of IL-1α, IL-1ra, IL-1β, and lactoferrin. Notably, microbial composition also appeared to vary with CVL inhibitory activity of HSV and *E. coli* in the opposite direction of HIV-infection, suggesting HIV infection negatively correlated with these activities; ANOVA analysis supported this observation (*E. coli* inhibition $p=2.1*10^{-8}$, HSV inhibition $p=2*10^{-4}$). Despite the numerous strong dependencies indicated by Figure S4-1 A, only HIV infection and IL-1ra significantly co-varied with the microbiome using permutation based significance analysis (Table S4-2).

To define the extent of vaginal microbial composition variation between HIV-positive and HIV-negative women, principal coordinate analysis (PCoA) was performed on a weighted unifrac distance matrix. The resulting plot, shown in Figure S4-1 B, revealed a striking separation of HIV-positive from HIV-negative women. Similar, yet slightly less dramatic results were obtained using the Bray-Curtis distance metric, which does not consider phylogenetic information when calculating distances. However, Adonis analysis of the significance of separation between groups was significant at $p \leq 1*10^{-4}$ for both distance metrics. Summarized, these results indicate that HIV infection was concomitant with drastic shifts in vaginal microbial composition.

To visualize the genera that were most differentially abundant between the groups, relative abundance bar plots were generated for all subjects containing genera that were present at ≥5% in ≥1 subject (Figure S4-1 C). As expected, *Lactobacillus* dominated the composition of most HIV-negative women, while 10% of women were colonized by a diverse group of organisms that would place them into the diverse "group IV" vaginal community defined by Ravel *et al*. (200). In contrast, *Lactobacillus* dominated the composition of 28% of HIV-positive women, with most women having variants of "group IV" vaginal microbial communities, potentially reflecting the increased prevalence of subclinical BV in HIV-positive women (213, 214). Indeed, many of the organisms that dominated the vaginal microbiome of HIV-positive women were correlated with BV, including *Gardnerella*, *Atopobium*, and *Prevotella* (211). To define differences that were statistically significant, iterative Kruskal–Wallis analysis was performed with respect to HIV infection on microbiota that had been binned by genus. Predictably, highly and semi-abundant genera, including *Gardnerella*, *Prevotella*, *Atopobium*, and *Adlercreutzia*, were enriched in HIV-positive women whereas *Lactobacillus* species were depleted. In contrast, with the exception of *Bacteroides* and an unclassified Bifidobacteriaceae, genera at low and very low abundance tended to be enriched in HIV-negative subjects compared with

HIV-positive subjects, suggesting a potential collapse in alpha diversity of rare species in HIV-positive subjects despite the increased diversity of “group IV” communities that predominated in HIV-positive women.

Combined, these analyses suggest strong shifts in microbiome composition accompanying HIV-infection. We therefore determined whether these shifts were concomitant with changes in the vaginal habitat and metabolic functionality of the microbial communities.

Imputed metagenomic function of the microbiome strongly varies with HIV infection

To obtain imputed metagenomic information from the collected 16S phylotypic data, a new bioinformatic tool called PICRUSt was used. PICRUSt aligns GreenGenes 16S sequences to >1,000 available reference genomes to reconstruct highly representative, albeit imputed, sample metagenomes (http://picrust.github.io/picrust/). To reduce noise, resulting data was then sorted into KEGG metabolic pathways using the bioinformatic tool HUMAnN.

To identify the specific functional metagenomic differences between the two cohorts, iterative Kruskal-Wallis analysis was performed on all metagenomic functions with respect to HIV infection, which was then corrected for multiple comparisons. Eliminating all comparisons with q values ≥0.05 indicated that 106, or 60% of all analyzed metagenomic pathways were significantly different between the two cohorts. Therefore, to reduce potential noise associated with imputation, a more stringent threshold of $q<0.02$ was selected such that only 80, or 45% of all included metagenomic pathways were considered significantly different (Figure S4-2). Notably, imputed microbial metagenomes of HIV-negative subjects were enriched for pathways involved in xenobiotics biodegradation and metabolism, lipid metabolism and glycan biosynthesis and metabolism. Meanwhile, imputed microbial metagenomes of HIV-positive subjects were enriched for amino acid metabolism, cofactors and vitamins metabolism, and fatty acid metabolism (Figure S4-2). This figure revealed the potential key difference that microbial metagenomes from HIV-positive subjects tended to be relatively enriched for organic acid metabolism, whereas metagenomes derived from HIV-negative subjects tended to be enriched for pathways that biosynthesize organic acids. This was concordant with observed vaginal pH differences of HIV-positive and HIV-negative subjects, with HIV-positive subjects having significantly (before correction for multiple metadata comparisons) higher vaginal pH measurements (Table S4-1).

Discussion

Differences in vaginal microbiome in HIV infected and uninfected women have been described in several small studies. The present study is an important advance, by defining the scope of HIV-associated microbial compositional and metagenomic change, and how it is integrated with host physiologic conditions, and mucosal functional state. Combined, these analyses reveal an integrated network of interactions between the microbiome and host mucosal state in HIV-negative subjects that is disrupted with HIV infection even in the context of ongoing anti-retroviral therapy.

While microbial metagenomic function would be suspected to vary between women with different microbial community structure, the dramatic differences observed between HIV-negative and HIV-positive subjects was surprising. Indeed, despite wide variations in community composition, metagenomic functions of most human anatomic microbial communities remain relatively static when viewed cross-sectionally (8, 207, 224). In contrast, increased metagenomic diversity has been reported in vaginal communities with lower *Lactobacillus* representation, suggesting variation of metagenomic content might be more common in vaginal communities (8). In this study, analysis of metabolic pathway representation appeared to reflect opposing tendencies for organic acid production. This divergence correlated with the difference in *Lactobacillus* representation between the two cohorts. However, it is notable that organic acid production is not simply related to *Lactobacillus* composition in normal individuals (200, 208, 209). Thus, future studies should seek to confirm the imputed metagenomic changes, and their association with measured metabolite changes in the mucosal lumen.

Figures and Tables

Figure S4-1

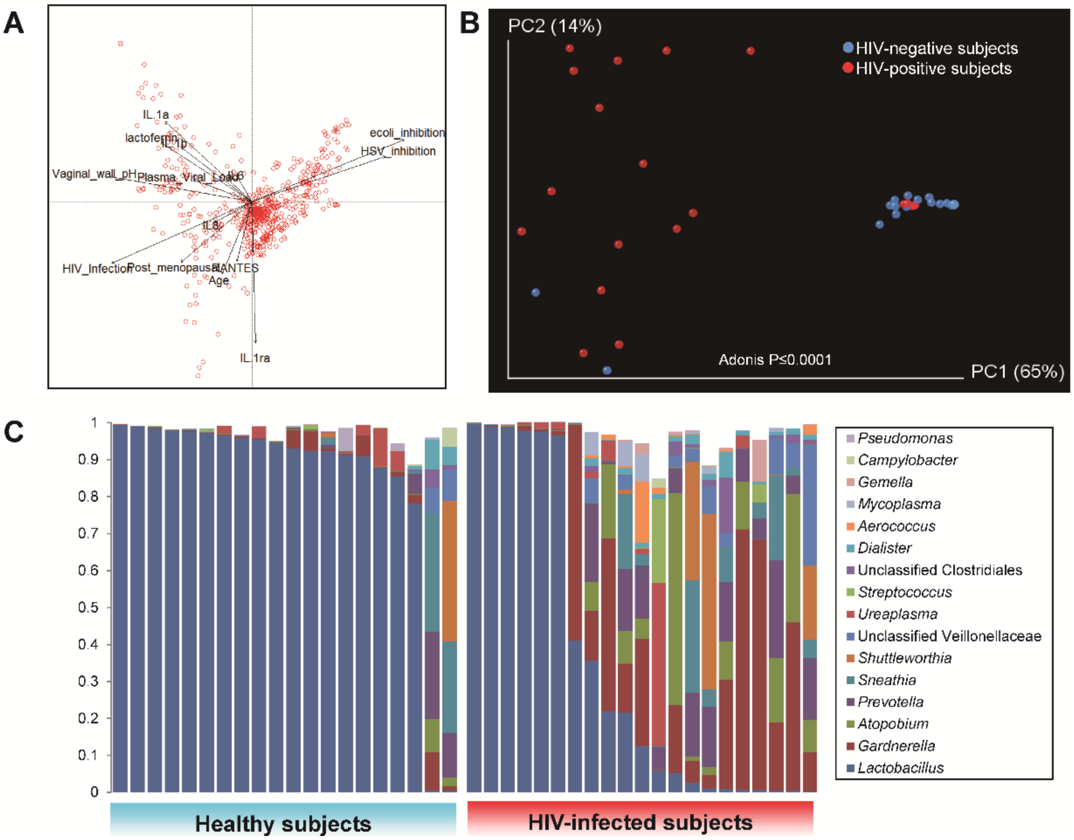


Figure S4-1. Canonical correspondence, principal coordinate, and compositional analysis of the vaginal microbiome. (**A**) CCA analysis was performed on OTUs with respect to HIV infection, HSV inhibition, E. coli inhibition, age, vaginal wall pH, and the abundance of several cytokines: IL-1α, IL-1β, IL-1ra, lactoferrin, IL-6, IL-8, and RANTES. Red spots represent individual OTUs. Arrows point in the direction of OTUs that most co-vary with the labeled factor and arrow length is indicative of the correlation strength between the ordination and factor tested. The data qualitatively indicates that IL-1ra, IL-1α, HIV infection, vaginal wall pH, HSV inhibition and *E. coli* inhibition co-vary with

microbiome composition. (**B**) Principal Coordinate analysis of the microbiome using the weighted unifrac distance metric revealed distinct clusters of HIV-positive and HIV-negative individuals. Adonis significance analysis was performed on the weighted unifrac distance matrix and the resulting p value is shown on the plot to simplify interpretation. (**C**) Genus-level abundances of vaginal microbiota are shown for all subjects separated by HIV infection status. A clear depletion of *Lactobacillus* is apparent in HIV-positive versus HIV-negative subjects.

Figure S4-2

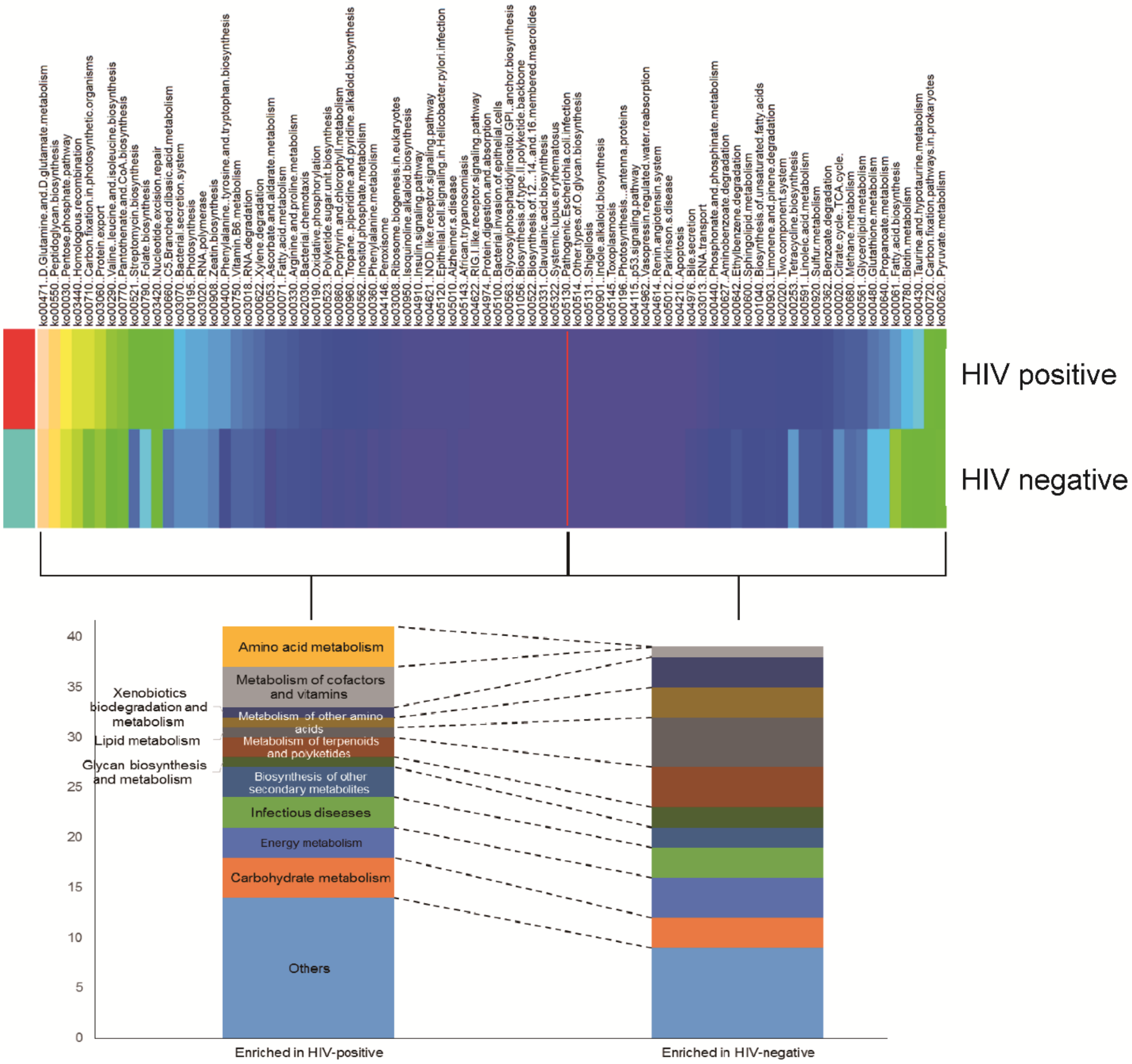


Figure S4-2. Metabolic heatmap highlighting metagenomic KEGG functions enriched or depleted in HIV-positive women. All functional metagenomic features that were significantly associated with HIV infection status are shown on the heatmap based on their relative abundance after *z* score transformation. The color bar on the left shows

HIV positive and negative. The stacked column chart shows the number of pathways in each functional class for the enriched and depleted pathways.

Table S4-1

Table S4-1. Subject characteristics, CVL immune mediator and antimicrobial activity data

	HIV-negative (n=20)	**HIV-positive (n=21)**	**p value**
Age (mean ± standard deviation (SD))	29.5 ± 7.2	39.3 ± 13.7	0.007
Race (number, %)			0.23
White	5 (25)	1 (5)	
Black	8 (40)	8 (38)	
Hispanic	5 (25)	10 (48)	
Other	2 (10)	2 (9)	
Postmenopausal (number, %)	0	6 (29)	0.02
Current cigarette smoker (number, %)	2 (10)	11 (52)	0.006
Receiving antiretroviral therapy (number, %)	0	17 (81)	
CD4 count in cells/mm^3 (median, range)		500 (25-1555)	
Plasma HIV-1 RNA in copies/mL (median, range)		40 (40-50,359)	
CVL HIV-1 RNA in copies/mL (median, range)		40 (40-25,924)	

History of douching (number, %)	4 (20)	12 (57)	0.02
HSV seropositivity (number, %)			
HSV-1 seropositive	9 (45)	14 (67)	0.21
HSV-2 seropositive	3 (15)	8 (38)	0.16
Vaginal pH (median, interquartile range (IQR))	4.9 (4.5-5)	5.2 (4.5-5.5)	0.02
CVL pH	4.5 (4.5-5)	5.0 (4.7-5.2)	0.001
***E. coli* inhibition (%)**	61 (50-89)	22 (4-42)	<0.0001
HSV inhibition (%)	37 (25-63)	10 (-4-18)	<0.0001
Total protein (µg/mL)	248 (161-388)	272 (181-522)	0.11
IL-1α (pg/mL)	42 (23-79)	103 (20-413)	0.07
IL-1β (pg/mL)	3 (1-15)	4 (1-23)	0.44
IL-1ra (ng/mL)	8 (6-9)	8 (6-10)	0.64
Lactoferrin (ng/mL)	499 (96-824)	918 (552-2165)	0.01
IL-6 (pg/mL)	11 (2-17)	4 (0.5-11)	0.13
IL-8 (pg/mL)	224 (93-926)	359 (196-1798)	0.28
RANTES (pg/mL)	3.6 (2-5)	3 (0.5-11)	0.95
MIP-1α (pg/mL)	9 (4-16)	6 (4-14)	0.56
MIP-1β (pg/mL)	6 (3-15)	11 (2-11)	0.16

IgG (pg/mL)	14 (6-33)	6 (3-19)	0.06
IgA (pg/mL)	1 (0.4-3)	2 (0.6-4)	0.32
HNP1-3 (ng/mL)	81 (8-402)	112 (25-382)	0.31
SLPI (ng/mL)	240 (104-331)	144 (60-360)	0.51
Lysozyme (ng/mL)	104 (82-484)	215 (53-451)	0.74

Table S4-2

Table S4-2. CCA p values for each metadata category

	Microbiome	**Metagenome**
HIV infection	0.005	0.005
HSV inhibition	0.125	0.075
E. coli inhibition	0.125	0.255
IL.1ra	0.005	0.005
IL.1a	0.095	0.07
IL.1b	0.085	0.185
IL6	0.255	0.555
IL8	0.325	0.59
RANTES	0.925	0.93
Vaginal wall pH	0.105	0.005
Lactoferrin	0.135	0.455
Age	0.09	0.025
Post menopausal	0.785	0.685
Plasma viral load	0.32	0.7

REFERENCES

1. Tong M, *et al.* (2013) A modular organization of the human intestinal mucosal microbiota and its association with inflammatory bowel disease. *PLoS One* 8(11):e80702.

2. Tong M, *et al.* (2014) Reprograming of gut microbiome energy metabolism by the FUT2 Crohn's disease risk polymorphism. *ISME J*.

3. McHardy I, *et al.* (2013) Integrative analysis of the microbiome and metabolome of the human intestinal mucosal surface reveals exquisite inter-relationships. *Microbiome* 1(1):17.

4. Fu J, *et al.* (2011) Loss of intestinal core 1-derived O-glycans causes spontaneous colitis in mice. *J Clin Invest* 121(4):1657-1666.

5. McHardy I, *et al.* (2013) HIV Infection is associated with compositional and functional shifts in the rectal mucosal microbiota. *Microbiome* 1(1):26.

6. Abraham C & Cho JH (2009) Inflammatory bowel disease. *New England Journal of Medicine* 361(21):2066-2078.

7. Braun J & Wei B (2007) Body traffic: ecology, genetics, and immunity in inflammatory bowel disease. *Annu Rev Pathol* 2:401-429.

8. Human Microbiome Project C (2012) Structure, function and diversity of the healthy human microbiome. *Nature* 486(7402):207-214.

9. Qin J, *et al.* (2010) A human gut microbial gene catalogue established by metagenomic sequencing. *Nature* 464(7285):59-65.

10. Caporaso JG, *et al.* (2011) Moving pictures of the human microbiome. *Genome Biol* 12(5):R50.

11. Yatsunenko T, *et al.* (2012) Human gut microbiome viewed across age and geography. *Nature* 486(7402):222-227.

12. Turnbaugh PJ, *et al.* (2009) A core gut microbiome in obese and lean twins. *Nature* 457(7228):480-484.

13. Human Microbiome Project C (2012) A framework for human microbiome research. *Nature* 486(7402):215-221.

14. Nicholson JK, *et al.* (2012) Host-gut microbiota metabolic interactions. *Science* 336(6086):1262-1267.

15. Matsumoto M, *et al.* (2012) Impact of intestinal microbiota on intestinal luminal metabolome. *Sci Rep* 2:233.

16. Frank DN*, et al.* (2007) Molecular-phylogenetic characterization of microbial community imbalances in human inflammatory bowel diseases. *Proc Natl Acad Sci U S A* 104(34):13780-13785.

17. Lepage P*, et al.* (2011) Twin study indicates loss of interaction between microbiota and mucosa of patients with ulcerative colitis. *Gastroenterology* 141(1):227-236.

18. Morgan XC*, et al.* (2012) Dysfunction of the intestinal microbiome in inflammatory bowel disease and treatment. *Genome Biol* 13(9):R79.

19. Spor A, Koren O, & Ley R (2011) Unravelling the effects of the environment and host genotype on the gut microbiome. *Nat Rev Microbiol* 9(4):279-290.

20. Jostins L*, et al.* (2012) Host-microbe interactions have shaped the genetic architecture of inflammatory bowel disease. *Nature* 491(7422):119-124.

21. Eckburg PB*, et al.* (2005) Diversity of the human intestinal microbial flora. *Science* 308(5728):1635-1638.

22. Li X*, et al.* (2011) A metaproteomic approach to study human-microbial ecosystems at the mucosal luminal interface. *PLoS One* 6(11):e26542.

23. Costello EK*, et al.* (2009) Bacterial community variation in human body habitats across space and time. *Science* 326(5960):1694-1697.

24. Caporaso JG*, et al.* (2012) Ultra-high-throughput microbial community analysis on the Illumina HiSeq and MiSeq platforms. *ISME J* 6(8):1621-1624.

25. Edgar RC (2010) Search and clustering orders of magnitude faster than BLAST. *Bioinformatics* 26(19):2460-2461.

26. DeSantis TZ*, et al.* (2006) Greengenes, a chimera-checked 16S rRNA gene database and workbench compatible with ARB. *Appl Environ Microbiol* 72(7):5069-5072.

27. Caporaso JG*, et al.* (2010) QIIME allows analysis of high-throughput community sequencing data. *Nat Methods* 7(5):335-336.

28. Willing BP*, et al.* (2010) A pyrosequencing study in twins shows that gastrointestinal microbial profiles vary with inflammatory bowel disease phenotypes. *Gastroenterology* 139(6):1844-U1105.

29. Frank DN*, et al.* (2011) Disease phenotype and genotype are associated with shifts in intestinal-associated microbiota in inflammatory bowel diseases. *Inflamm Bowel Dis* 17(1):179-184.

30. Nagalingam NA & Lynch SV (2012) Role of the microbiota in inflammatory bowel diseases. *Inflammatory Bowel Diseases* 18(5):968-980.

31. Peterson DA, Frank DN, Pace NR, & Gordon JI (2008) Metagenomic approaches for defining the pathogenesis of inflammatory bowel diseases. *Cell Host Microbe* 3(6):417-427.

32. Manichanh C*, et al.* (2006) Reduced diversity of faecal microbiota in Crohn's disease revealed by a metagenomic approach. *Gut* 55(2):205-211.

33. Gophna U, Sommerfeld K, Gophna S, Doolittle WF, & Veldhuyzen van Zanten SJ (2006) Differences between tissue-associated intestinal microfloras of patients with Crohn's disease and ulcerative colitis. *J Clin Microbiol* 44(11):4136-4141.

34. Khor B, Gardet A, & Xavier RJ (2011) Genetics and pathogenesis of inflammatory bowel disease. *Nature* 474(7351):307-317.

35. Dethlefsen L, McFall-Ngai M, & Relman DA (2007) An ecological and evolutionary perspective on human-microbe mutualism and disease. *Nature* 449(7164):811-818.

36. Kuhn R, Lohler J, Rennick D, Rajewsky K, & Muller W (1993) Interleukin-10-deficient mice develop chronic enterocolitis. *Cell* 75(2):263-274.

37. Uhlig HH & Powrie F (2009) Mouse models of intestinal inflammation as tools to understand the pathogenesis of inflammatory bowel disease. *European Journal of Immunology* 39(8):2021-2026.

38. Elson CO*, et al.* (2005) Experimental models of inflammatory bowel disease reveal innate, adaptive, and regulatory mechanisms of host dialogue with the microbiota. *Immunol Rev* 206:260-276.

39. Li E*, et al.* (2012) Inflammatory Bowel Diseases Phenotype, C. difficile and NOD2 Genotype Are Associated with Shifts in Human Ileum Associated Microbial Composition. *PLoS One* 7(6):e26284.

40. Rausch P*, et al.* (2011) Colonic mucosa-associated microbiota is influenced by an interaction of Crohn disease and FUT2 (Secretor) genotype. *Proc Natl Acad Sci U S A* 108(47):19030-19035.

41. Frank DN, Zhu W, Sartor RB, & Li E (2011) Investigating the biological and clinical significance of human dysbioses. *Trends in Microbiology* 19(9):427-434.

42. Costello EK, Stagaman K, Dethlefsen L, Bohannan BJ, & Relman DA (2012) The application of ecological theory toward an understanding of the human microbiome. *Science* 336(6086):1255-1262.

43. Virgin HW & Todd JA (2011) Metagenomics and personalized medicine. *Cell* 147(1):44-56.

44. Sansonetti PJ (2008) War and peace at the intestinal epithelial surface: an integrated view of bacterial commensalism versus bacterial pathogenicity. *Journal of Pediatric Gastroenterology & Nutrition* 46 Suppl 1:E6-7.

45. Parfrey LW, Walters WA, & Knight R (2011) Microbial eukaryotes in the human microbiome: ecology, evolution, and future directions. *Front Microbiol* 2:153.

46. Kato S, Haruta S, Cui ZJ, Ishii M, & Igarashi Y (2008) Network relationships of bacteria in a stable mixed culture. *Microb Ecol* 56(3):403-411.

47. Beman JM, Steele JA, & Fuhrman JA (2011) Co-occurrence patterns for abundant marine archaeal and bacterial lineages in the deep chlorophyll maximum of coastal California. *ISME J* 5(7):1077-1085.

48. Lozupone C*, et al.* (2012) Identifying genomic and metabolic features that can underlie early successional and opportunistic lifestyles of human gut symbionts. *Genome Research*.

49. Chaffron S, Rehrauer H, Pernthaler J, & von Mering C (2010) A global network of coexisting microbes from environmental and whole-genome sequence data. *Genome Research* 20(7):947-959.

50. Faust K*, et al.* (2012) Microbial Co-occurrence Relationships in the Human Microbiome. *PLoS Comput Biol* 8(7):e1002606.

51. Zhang B & Horvath S (2005) A general framework for weighted gene co-expression network analysis. *Stat Appl Genet Mol Biol* 4:Article17.

52. Langfelder P, Zhang B, & Horvath S (2008) Defining clusters from a hierarchical cluster tree: the Dynamic Tree Cut package for R. *Bioinformatics* 24(5):719-720.

53. Langfelder P & Horvath S (2008) WGCNA: an R package for weighted correlation network analysis. *BMC Bioinformatics* 9:559.

54. Langfelder P, Luo R, Oldham MC, & Horvath S (2011) Is my network module preserved and reproducible? *PLoS Comput Biol* 7(1):e1001057.

55. Abubucker S*, et al.* (2012) Metabolic reconstruction for metagenomic data and its application to the human microbiome. *PLoS Comput Biol* 8(6):e1002358.

56. Fukuda S*, et al.* (2011) Bifidobacteria can protect from enteropathogenic infection through production of acetate. *Nature* 469(7331):543-U791.

57. Summers ZM*, et al.* (2010) Direct exchange of electrons within aggregates of an evolved syntrophic coculture of anaerobic bacteria. *Science* 330(6009):1413-1415.

58. Foster JA, Krone SM, & Forney LJ (2008) Application of ecological network theory to the human microbiome. *Interdiscip Perspect Infect Dis* 2008:839501.

59. Friedman J & Alm EJ (2012) Inferring correlation networks from genomic survey data. *PLoS Comput Biol* 8(9):e1002687.

60. Krogan NJ*, et al.* (2006) Global landscape of protein complexes in the yeast Saccharomyces cerevisiae. *Nature* 440(7084):637-643.

61. Voineagu I*, et al.* (2011) Transcriptomic analysis of autistic brain reveals convergent molecular pathology. *Nature* 474(7351):380-+.

62. Nugent R & Meila M (2010) An overview of clustering applied to molecular biology. *Methods Mol Biol* 620:369-404.

63. Wang J, Li M, Deng Y, & Pan Y (2010) Recent advances in clustering methods for protein interaction networks. *BMC Genomics* 11 Suppl 3:S10.

64. Jay JJ*, et al.* (2012) A systematic comparison of genome-scale clustering algorithms. *BMC Bioinformatics* 13 Suppl 10:S7.

65. Rocke DM, Ideker T, Troyanskaya O, Quackenbush J, & Dopazo J (2009) Papers on normalization, variable selection, classification or clustering of microarray data. *Bioinformatics* 25(6):701-702.

66. Horvath S & Dong J (2008) Geometric interpretation of gene coexpression network analysis. *PLoS Comput Biol* 4(8):e1000117.

67. Handl J, Knowles J, & Kell DB (2005) Computational cluster validation in post-genomic data analysis. *Bioinformatics* 21(15):3201-3212.

68. Garge NR, Page GP, Sprague AP, Gorman BS, & Allison DB (2005) Reproducible clusters from microarray research: whither? *BMC Bioinformatics* 6 Suppl 2:S10.

69. Brock G, Datta S, Pihur V, & Datta S (2008) clValid: An R package for cluster validation. *Journal of Statistical Software* 25(4):1-22.

70. Giancarlo R, Scaturro D, & Utro F (2008) Computational cluster validation for microarray data analysis: experimental assessment of Clest, Consensus Clustering, Figure of Merit, Gap Statistics and Model Explorer. *Bmc Bioinformatics* 9.

71. Smolkin M & Ghosh D (2003) Cluster stability scores for microarray data in cancer studies. *BMC Bioinformatics* 4:36.

72. Kapp AV & Tibshirani R (2007) Are clusters found in one dataset present in another dataset? *Biostatistics* 8(1):9-31.

73. Horvath S (2011) *Weighted Network Analysis: Applications in Genomics and Systems Biology* (Springer-Verlag, Berlin).

74. Zhang TY*, et al.* (2011) Host genes related to Paneth cells and xenobiotic metabolism are associated with shifts in human ileum-associated microbial composition. *Inflammatory Bowel Diseases* 17:S80-S80.

75. Barabasi AL, Gulbahce N, & Loscalzo J (2011) Network medicine: a network-based approach to human disease. *Nat Rev Genet* 12(1):56-68.

76. Balish E & Warner T (2002) Enterococcus faecalis induces inflammatory bowel disease in interleukin-10 knockout mice. *Am J Pathol* 160(6):2253-2257.

77. Sepehri S*, et al.* (2011) Characterization of Escherichia coli isolated from gut biopsies of newly diagnosed patients with inflammatory bowel disease. *Inflammatory Bowel Diseases* 17(7):1451-1463.

78. Joossens M*, et al.* (2011) Dysbiosis of the faecal microbiota in patients with Crohn's disease and their unaffected relatives. *Gut* 60(5):631-637.

79. Sokol H*, et al.* (2008) Faecalibacterium prausnitzii is an anti-inflammatory commensal bacterium identified by gut microbiota analysis of Crohn disease patients. *Proc Natl Acad Sci U S A* 105(43):16731-16736.

80. Samuel BS*, et al.* (2008) Effects of the gut microbiota on host adiposity are modulated by the short-chain fatty-acid binding G protein-coupled receptor, Gpr41. *Proceedings of the National Academy of Sciences of the United States of America* 105(43):16767-16772.

81. Scheppach W (1994) Effects of short chain fatty acids on gut morphology and function. *Gut* 35(1 Suppl):S35-38.

82. Wong JM, de Souza R, Kendall CW, Emam A, & Jenkins DJ (2006) Colonic health: fermentation and short chain fatty acids. *J Clin Gastroenterol* 40(3):235-243.

83. Robinson CJ, Bohannan BJ, & Young VB (2010) From structure to function: the ecology of host-associated microbial communities. *Microbiol Mol Biol Rev* 74(3):453-476.

84. Smillie CS*, et al.* (2011) Ecology drives a global network of gene exchange connecting the human microbiome. *Nature* 480(7376):241-244.

85. Konopka A (2009) What is microbial community ecology? *ISME J* 3:1223-1230.

86. Raes J & Bork P (2008) Molecular eco-systems biology: towards an understanding of community function. *Nat Rev Microbiol* 6(9):693-699.

87. Lee YK & Mazmanian SK (2010) Has the microbiota played a critical role in the evolution of the adaptive immune system? *Science* 330(6012):1768-1773.

88. Li M*, et al.* (2008) Symbiotic gut microbes modulate human metabolic phenotypes. *Proc Natl Acad Sci U S A* 105(6):2117-2122.

89. Willing BP, Gill N, & Finlay BB (2010) The role of the immune system in regulating the microbiota. *Gut Microbes* 1(4):213-223.

90. Qin J*, et al.* (2012) A metagenome-wide association study of gut microbiota in type 2 diabetes. *Nature* 490(7418):55-60.

91. Claesson MJ*, et al.* (2012) Gut microbiota composition correlates with diet and health in the elderly. *Nature* 488(7410):178-184.

92. Wu GD*, et al.* (2011) Linking long-term dietary patterns with gut microbial enterotypes. *Science* 334(6052):105-108.

93. Smith MI*, et al.* (2013) Gut microbiomes of Malawian twin pairs discordant for kwashiorkor. *Science* 339(6119):548-554.

94. Stewart JA, Chadwick VS, & Murray A (2005) Investigations into the influence of host genetics on the predominant eubacteria in the faecal microflora of children. *J Med Microbiol* 54(Pt 12):1239-1242.

95. Erwin G. Zoetendal ADLA, Wilma M. Akkermans-van Vliet, J. Arjan G. M. de Visser, Willem M. de Vos (2001) The Host Genotype Affects the Bacterial Community in the Human Gastronintestinal Tract. *Microbial Ecology in Health and Disease* 13(3):129-134.

96. McGovern DP*, et al.* (2010) Fucosyltransferase 2 (FUT2) non-secretor status is associated with Crohn's disease. *Hum Mol Genet* 19(17):3468-3476.

97. Van der Sluis M*, et al.* (2006) Muc2-deficient mice spontaneously develop colitis, indicating that MUC2 is critical for colonic protection. *Gastroenterology* 131(1):117-129.

98. Johansson ME*, et al.* (2008) The inner of the two Muc2 mucin-dependent mucus layers in colon is devoid of bacteria. *Proc Natl Acad Sci U S A* 105(39):15064-15069.

99. McGuckin MA, Linden SK, Sutton P, & Florin TH (2011) Mucin dynamics and enteric pathogens. *Nat Rev Microbiol* 9(4):265-278.

100. Linden SK, Sutton P, Karlsson NG, Korolik V, & McGuckin MA (2008) Mucins in the mucosal barrier to infection. *Mucosal Immunol* 1(3):183-197.

101. Katayama T*, et al.* (2004) Molecular cloning and characterization of Bifidobacterium bifidum 1,2-alpha-L-fucosidase (AfcA), a novel inverting glycosidase (glycoside hydrolase family 95). *J Bacteriol* 186(15):4885-4893.

102. Hurd EA, Holmen JM, Hansson GC, & Domino SE (2005) Gastrointestinal mucins of Fut2-null mice lack terminal fucosylation without affecting colonization by Candida albicans. *Glycobiology* 15(10):1002-1007.

103. Wacklin P*, et al.* (2011) Secretor genotype (FUT2 gene) is strongly associated with the composition of Bifidobacteria in the human intestine. *PLoS One* 6(5):e20113.

104. Wacklin P*, et al.* (2014) Faecal Microbiota Composition in Adults Is Associated with the *FUT2* Gene Determining the Secretor Status. *PLoS ONE* 9(4):e94863.

105. Erickson AR*, et al.* (2012) Integrated metagenomics/metaproteomics reveals human host-microbiota signatures of Crohn's disease. *PLoS One* 7(11):e49138.

106. Ferrer-Admetlla A*, et al.* (2009) A natural history of FUT2 polymorphism in humans. *Mol Biol Evol* 26(9):1993-2003.

107. Kashyap PC*, et al.* (2013) Genetically dictated change in host mucus carbohydrate landscape exerts a diet-dependent effect on the gut microbiota. *Proc Natl Acad Sci U S A*.

108. Benjamini Y & Hochberg Y (1995) Controlling the False Discovery Rate - a Practical and Powerful Approach to Multiple Testing. *J Roy Stat Soc Meth* 57(1):289-300.

109. Langille MG*, et al.* (2013) Predictive functional profiling of microbial communities using 16S rRNA marker gene sequences. *Nat Biotechnol* 31(9):814-821.

110. Koren O*, et al.* (2013) A guide to enterotypes across the human body: meta-analysis of microbial community structures in human microbiome datasets. *PLoS Comput Biol* 9(1):e1002863.

111. Markle JG*, et al.* (2013) Sex differences in the gut microbiome drive hormone-dependent regulation of autoimmunity. *Science* 339(6123):1084-1088.

112. Mueller S*, et al.* (2006) Differences in fecal microbiota in different European study populations in relation to age, gender, and country: a cross-sectional study. *Appl Environ Microbiol* 72(2):1027-1033.

113. Turnbaugh PJ*, et al.* (2009) The effect of diet on the human gut microbiome: a metagenomic analysis in humanized gnotobiotic mice. *Sci Transl Med* 1(6):6ra14.

114. Claesson MJ*, et al.* (2010) Comparison of two next-generation sequencing technologies for resolving highly complex microbiota composition using tandem variable 16S rRNA gene regions. *Nucleic Acids Res* 38(22):e200.

115. Koropatkin NM, Cameron EA, & Martens EC (2012) How glycan metabolism shapes the human gut microbiota. *Nat Rev Microbiol* 10(5):323-335.

116. Martens EC*, et al.* (2011) Recognition and degradation of plant cell wall polysaccharides by two human gut symbionts. *PLoS Biol* 9(12):e1001221.

117. Sonnenburg ED*, et al.* (2010) Specificity of polysaccharide use in intestinal bacteroides species determines diet-induced microbiota alterations. *Cell* 141(7):1241-1252.

118. Ng KM, *et al.* (2013) Microbiota-liberated host sugars facilitate post-antibiotic expansion of enteric pathogens. *Nature* 502(7469):96-99.

119. Atarashi K, *et al.* (2011) Induction of colonic regulatory T cells by indigenous Clostridium species. *Science* 331(6015):337-341.

120. Sokol H & Seksik P (2010) The intestinal microbiota in inflammatory bowel diseases: time to connect with the host. *Curr Opin Gastroenterol* 26(4):327-331.

121. Le Chatelier E, *et al.* (2013) Richness of human gut microbiome correlates with metabolic markers. *Nature* 500(7464):541-546.

122. Magalhaes A, *et al.* (2009) Fut2-null mice display an altered glycosylation profile and impaired BabA-mediated Helicobacter pylori adhesion to gastric mucosa. *Glycobiology* 19(12):1525-1536.

123. Sonnenburg JL, *et al.* (2005) Glycan foraging in vivo by an intestine-adapted bacterial symbiont. *Science* 307(5717):1955-1959.

124. Pacheco AR, *et al.* (2012) Fucose sensing regulates bacterial intestinal colonization. *Nature* 492(7427):113-117.

125. Gill SR, *et al.* (2006) Metagenomic analysis of the human distal gut microbiome. *Science* 312(5778):1355-1359.

126. Metges CC (2000) Contribution of microbial amino acids to amino acid homeostasis of the host. *J Nutr* 130(7):1857S-1864S.

127. Tattoli I, *et al.* (2012) Amino acid starvation induced by invasive bacterial pathogens triggers an innate host defense program. *Cell Host Microbe* 11(6):563-575.

128. Petnicki-Ocwieja T, *et al.* (2009) Nod2 is required for the regulation of commensal microbiota in the intestine. *Proc Natl Acad Sci U S A* 106(37):15813-15818.

129. Salzman NH, *et al.* (2010) Enteric defensins are essential regulators of intestinal microbial ecology. *Nat Immunol* 11(1):76-83.

130. Ivanov, II, *et al.* (2008) Specific microbiota direct the differentiation of IL-17-producing T-helper cells in the mucosa of the small intestine. *Cell Host Microbe* 4(4):337-349.

131. Perez-Cobas AE, *et al.* (2012) Gut microbiota disturbance during antibiotic therapy: a multi-omic approach. *Gut*.

132. Segata N, *et al.* (2011) Metagenomic biomarker discovery and explanation. *Genome Biol* 12(6):R60.

133. Jansson J*, et al.* (2009) Metabolomics reveals metabolic biomarkers of Crohn's disease. *PLoS One* 4(7):e6386.

134. Le Gall G*, et al.* (2011) Metabolomics of fecal extracts detects altered metabolic activity of gut microbiota in ulcerative colitis and irritable bowel syndrome. *J Proteome Res* 10(9):4208-4218.

135. Marchesi JR*, et al.* (2007) Rapid and noninvasive metabonomic characterization of inflammatory bowel disease. *J Proteome Res* 6(2):546-551.

136. Abraham C & Cho J (2009) Interleukin-23/Th17 pathways and inflammatory bowel disease. *Inflamm Bowel Dis* 15(7):1090-1100.

137. Schnorr SL*, et al.* (2014) Gut microbiome of the Hadza hunter-gatherers. *Nat Commun* 5:3654.

138. Xavier RJ & Podolsky DK (2007) Unravelling the pathogenesis of inflammatory bowel disease. *Nature* 448(7152):427-434.

139. Johansson ME, Larsson JM, & Hansson GC (2011) The two mucus layers of colon are organized by the MUC2 mucin, whereas the outer layer is a legislator of host-microbial interactions. *Proc Natl Acad Sci U S A* 108 Suppl 1:4659-4665.

140. Gum JR, Jr., Hicks JW, Toribara NW, Siddiki B, & Kim YS (1994) Molecular cloning of human intestinal mucin (MUC2) cDNA. Identification of the amino terminus and overall sequence similarity to prepro-von Willebrand factor. *J Biol Chem* 269(4):2440-2446.

141. Ju T, Brewer K, D'Souza A, Cummings RD, & Canfield WM (2002) Cloning and expression of human core 1 beta1,3-galactosyltransferase. *J Biol Chem* 277(1):178-186.

142. Fu J*, et al.* (2008) Endothelial cell O-glycan deficiency causes blood/lymphatic misconnections and consequent fatty liver disease in mice. *J Clin Invest* 118(11):3725-3737.

143. Xia L*, et al.* (2004) Defective angiogenesis and fatal embryonic hemorrhage in mice lacking core 1-derived O-glycans. *J Cell Biol* 164(3):451-459.

144. Ju T, Aryal RP, Stowell CJ, & Cummings RD (2008) Regulation of protein O-glycosylation by the endoplasmic reticulum-localized molecular chaperone Cosmc. *J Cell Biol* 182(3):531-542.

145. Smithson JE*, et al.* (1997) Altered expression of mucins throughout the colon in ulcerative colitis. *Gut* 40(2):234-240.

146. Podolsky DK & Isselbacher KJ (1983) Composition of human colonic mucin. Selective alteration in inflammatory bowel disease. *J Clin Invest* 72(1):142-153.

147. Rhodes JM (1997) Colonic mucus and ulcerative colitis. *Gut* 40(6):807-808.

148. Corfield AP*, et al.* (1996) Colonic mucins in ulcerative colitis: evidence for loss of sulfation. *Glycoconj J* 13(5):809-822.

149. Podolsky DK & Fournier DA (1988) Alterations in mucosal content of colonic glycoconjugates in inflammatory bowel disease defined by monoclonal antibodies. *Gastroenterology* 95(2):379-387.

150. Akira S, Uematsu S, & Takeuchi O (2006) Pathogen recognition and innate immunity. *Cell* 124(4):783-801.

151. An G*, et al.* (2007) Increased susceptibility to colitis and colorectal tumors in mice lacking core 3-derived O-glycans. *J Exp Med* 204(6):1417-1429.

152. Strober W, Fuss I, & Mannon P (2007) The fundamental basis of inflammatory bowel disease. *J Clin Invest* 117(3):514-521.

153. Garrett WS*, et al.* (2007) Communicable ulcerative colitis induced by T-bet deficiency in the innate immune system. *Cell* 131(1):33-45.

154. Martens EC, Chiang HC, & Gordon JI (2008) Mucosal glycan foraging enhances fitness and transmission of a saccharolytic human gut bacterial symbiont. *Cell Host Microbe* 4(5):447-457.

155. Hooper LV & Gordon JI (2001) Commensal host-bacterial relationships in the gut. *Science* 292(5519):1115-1118.

156. Lundin A*, et al.* (2008) Gut flora, Toll-like receptors and nuclear receptors: a tripartite communication that tunes innate immunity in large intestine. *Cell Microbiol* 10(5):1093-1103.

157. Badman MK & Flier JS (2005) The gut and energy balance: visceral allies in the obesity wars. *Science* 307(5717):1909-1914.

158. Turnbaugh PJ*, et al.* (2007) The human microbiome project. *Nature* 449(7164):804-810.

159. DuPont AW & DuPont HL (2011) The intestinal microbiota and chronic disorders of the gut. *Nat Rev Gastroenterol Hepatol* 8(9):523-531.

160. Greenblum S, Turnbaugh PJ, & Borenstein E (2012) Metagenomic systems biology of the human gut microbiome reveals topological shifts associated with obesity and inflammatory bowel disease. *Proc Natl Acad Sci U S A* 109(2):594-599.

161. Verberkmoes NC*, et al.* (2009) Shotgun metaproteomics of the human distal gut microbiota. *ISME J* 3(2):179-189.

162. Presley LL*, et al.* (2012) Host-microbe relationships in inflammatory bowel disease detected by bacterial and metaproteomic analysis of the mucosal-luminal interface. *Inflamm Bowel Dis* 18(3):409-417.

163. Khoruts A, Dicksved J, Jansson JK, & Sadowsky MJ (2010) Changes in the composition of the human fecal microbiome after bacteriotherapy for recurrent Clostridium difficile-associated diarrhea. *J Clin Gastroenterol* 44(5):354-360.

164. Caporaso JG*, et al.* (2011) Global patterns of 16S rRNA diversity at a depth of millions of sequences per sample. *Proc Natl Acad Sci U S A* 108 Suppl 1:4516-4522.

165. Dray S & Dufour AB (2007) The ade4 package: Implementing the duality diagram for ecologists. *Journal of Statistical Software* 22(4):1-20.

166. Stearns JC*, et al.* (2011) Bacterial biogeography of the human digestive tract. *Sci Rep* 1:170.

167. Tannock GW (2001) Molecular assessment of intestinal microflora. *Am J Clin Nutr* 73(2 Suppl):410S-414S.

168. Hamer HM, De Preter V, Windey K, & Verbeke K (2012) Functional analysis of colonic bacterial metabolism: relevant to health? *Am J Physiol Gastrointest Liver Physiol* 302(1):G1-9.

169. Harrell L*, et al.* (2012) Standard colonic lavage alters the natural state of mucosal-associated microbiota in the human colon. *PLoS One* 7(2):e32545.

170. Veazey RS & Lackner AA (2005) HIV swiftly guts the immune system. *Nat Med* 11(5):469-470.

171. Ullrich R, Schieferdecker HL, Ziegler K, Riecken EO, & Zeitz M (1990) gamma delta T cells in the human intestine express surface markers of activation and are preferentially located in the epithelium. *Cell Immunol* 128(2):619-627.

172. Nabel G & Baltimore D (1987) An inducible transcription factor activates expression of human immunodeficiency virus in T cells. *Nature* 326(6114):711-713.

173. Schieferdecker HL, Ullrich R, Hirseland H, & Zeitz M (1992) T cell differentiation antigens on lymphocytes in the human intestinal lamina propria. *J Immunol* 149(8):2816-2822.

174. Kotler DP, Reka S, & Clayton F (1993) Intestinal mucosal inflammation associated with human immunodeficiency virus infection. *Dig Dis Sci* 38(6):1119-1127.

175. Lapenta C*, et al.* (1999) Human intestinal lamina propria lymphocytes are naturally permissive to HIV-1 infection. *Eur J Immunol* 29(4):1202-1208.

176. Mehandru S*, et al.* (2006) Lack of mucosal immune reconstitution during prolonged treatment of acute and early HIV-1 infection. *PLoS Med* 3(12):e484.

177. Mehandru S*, et al.* (2007) Mechanisms of gastrointestinal CD4+ T-cell depletion during acute and early human immunodeficiency virus type 1 infection. *J Virol* 81(2):599-612.

178. Sheth PM*, et al.* (2008) Immune reconstitution in the sigmoid colon after long-term HIV therapy. *Mucosal Immunol* 1(5):382-388.

179. Monteleone I, Vavassori P, Biancone L, Monteleone G, & Pallone F (2002) Immunoregulation in the gut: success and failures in human disease. *Gut* 50 Suppl 3:III60-64.

180. Macpherson AJ & Harris NL (2004) Interactions between commensal intestinal bacteria and the immune system. *Nat Rev Immunol* 4(6):478-485.

181. Abraham C & Medzhitov R (2011) Interactions between the host innate immune system and microbes in inflammatory bowel disease. *Gastroenterology* 140(6):1729-1737.

182. Dandekar S (2007) Pathogenesis of HIV in the gastrointestinal tract. *Curr HIV/AIDS Rep* 4(1):10-15.

183. Brenchley JM & Douek DC (2012) Microbial translocation across the GI tract. *Annu Rev Immunol* 30:149-173.

184. Brenchley JM*, et al.* (2006) Microbial translocation is a cause of systemic immune activation in chronic HIV infection. *Nat Med* 12(12):1365-1371.

185. Sandler NG & Douek DC (2012) Microbial translocation in HIV infection: causes, consequences and treatment opportunities. *Nat Rev Microbiol* 10(9):655-666.

186. Appay V & Sauce D (2008) Immune activation and inflammation in HIV-1 infection: causes and consequences. *J Pathol* 214(2):231-241.

187. Brown EM, Sadarangani M, & Finlay BB (2013) The role of the immune system in governing host-microbe interactions in the intestine. *Nat Immunol* 14(7):660-667.

188. Hanash AM*, et al.* (2012) Interleukin-22 protects intestinal stem cells from immune-mediated tissue damage and regulates sensitivity to graft versus host disease. *Immunity* 37(2):339-350.

189. Belkaid Y & Naik S (2013) Compartmentalized and systemic control of tissue immunity by commensals. *Nat Immunol* 14(7):646-653.

190. Brestoff JR & Artis D (2013) Commensal bacteria at the interface of host metabolism and the immune system. *Nat Immunol* 14(7):676-684.

191. Ng SC, Kamm MA, Stagg AJ, & Knight SC (2010) Intestinal dendritic cells: their role in bacterial recognition, lymphocyte homing, and intestinal inflammation. *Inflamm Bowel Dis* 16(10):1787-1807.

192. Abt MC*, et al.* (2012) Commensal bacteria calibrate the activation threshold of innate antiviral immunity. *Immunity* 37(1):158-170.

193. Rivollier A, He J, Kole A, Valatas V, & Kelsall BL (2012) Inflammation switches the differentiation program of Ly6Chi monocytes from antiinflammatory macrophages to inflammatory dendritic cells in the colon. *J Exp Med* 209(1):139-155.

194. Fournier BM & Parkos CA (2012) The role of neutrophils during intestinal inflammation. *Mucosal Immunol* 5(4):354-366.

195. Gori A*, et al.* (2008) Early impairment of gut function and gut flora supporting a role for alteration of gastrointestinal mucosa in human immunodeficiency virus pathogenesis. *J Clin Microbiol* 46(2):757-758.

196. McKenna P*, et al.* (2008) The macaque gut microbiome in health, lentiviral infection, and chronic enterocolitis. *PLoS Pathog* 4(2):e20.

197. Handley SA*, et al.* (2012) Pathogenic simian immunodeficiency virus infection is associated with expansion of the enteric virome. *Cell* 151(2):253-266.

198. Anton PA*, et al.* (2012) RMP-02/MTN-006: A phase 1 rectal safety, acceptability, pharmacokinetic, and pharmacodynamic study of tenofovir 1% gel compared with oral tenofovir disoproxil fumarate. *AIDS Res Hum Retroviruses* 28(11):1412-1421.

199. Anton PA*, et al.* (2011) First phase 1 double-blind, placebo-controlled, randomized rectal microbicide trial using UC781 gel with a novel index of ex vivo efficacy. *PLoS One* 6(9):e23243.

200. Ravel J*, et al.* (2011) Vaginal microbiome of reproductive-age women. *Proc Natl Acad Sci U S A* 108 Suppl 1:4680-4687.

201. Gajer P*, et al.* (2012) Temporal dynamics of the human vaginal microbiota. *Sci Transl Med* 4(132):132ra152.

202. Witkin SS, Linhares IM, & Giraldo P (2007) Bacterial flora of the female genital tract: function and immune regulation. *Best Pract Res Clin Obstet Gynaecol* 21(3):347-354.

203. Boskey ER, Telsch KM, Whaley KJ, Moench TR, & Cone RA (1999) Acid production by vaginal flora in vitro is consistent with the rate and extent of vaginal acidification. *Infect Immun* 67(10):5170-5175.

204. Kalyoussef S*, et al.* (2012) Lactobacillus proteins are associated with the bactericidal activity against E. coli of female genital tract secretions. *PLoS One* 7(11):e49506.

205. Ghartey JP*, et al.* (2012) Association of bactericidal activity of genital tract secretions with Escherichia coli colonization in pregnancy. *Am J Obstet Gynecol* 207(4):297 e291-298.

206. Valore EV, Wiley DJ, & Ganz T (2006) Reversible deficiency of antimicrobial polypeptides in bacterial vaginosis. *Infect Immun* 74(10):5693-5702.

207. Ma B, Forney LJ, & Ravel J (2012) Vaginal microbiome: rethinking health and disease. *Annu Rev Microbiol* 66:371-389.

208. Zhou X*, et al.* (2007) Differences in the composition of vaginal microbial communities found in healthy Caucasian and black women. *ISME J* 1(2):121-133.

209. Zhou X*, et al.* (2010) The vaginal bacterial communities of Japanese women resemble those of women in other racial groups. *FEMS Immunol Med Microbiol* 58(2):169-181.

210. Koumans EH*, et al.* (2007) The prevalence of bacterial vaginosis in the United States, 2001-2004; associations with symptoms, sexual behaviors, and reproductive health. *Sex Transm Dis* 34(11):864-869.

211. Fredricks DN, Fiedler TL, & Marrazzo JM (2005) Molecular identification of bacteria associated with bacterial vaginosis. *N Engl J Med* 353(18):1899-1911.

212. Atashili J, Poole C, Ndumbe PM, Adimora AA, & Smith JS (2008) Bacterial vaginosis and HIV acquisition: a meta-analysis of published studies. *AIDS* 22(12):1493-1501.

213. Nwadioha S, Egah D, Banwat E, Egesie J, & Onwuezobe I (2011) Prevalence of bacterial vaginosis and its risk factors in HIV/AIDS patients with abnormal vaginal discharge. *Asian Pac J Trop Med* 4(2):156-158.

214. Warren D*, et al.* (2001) A multicenter study of bacterial vaginosis in women with or at risk for human immunodeficiency virus infection. *Infect Dis Obstet Gynecol* 9(3):133-141.

215. Spear GT*, et al.* (2008) Comparison of the diversity of the vaginal microbiota in HIV-infected and HIV-uninfected women with or without bacterial vaginosis. *J Infect Dis* 198(8):1131-1140.

216. Spear GT*, et al.* (2011) Pyrosequencing of the genital microbiotas of HIV-seropositive and -seronegative women reveals Lactobacillus iners as the predominant Lactobacillus Species. *Appl Environ Microbiol* 77(1):378-381.

217. Hummelen R*, et al.* (2010) Deep sequencing of the vaginal microbiota of women with HIV. *PLoS One* 5(8):e12078.

218. Lozupone C*, et al.* (2013) Widespread Colonization of the Lung by Tropheryma whipplei in HIV Infection. *Am J Respir Crit Care Med* 187(10):1110-1117.

219. Dang AT*, et al.* (2012) Evidence of an increased pathogenic footprint in the lingual microbiome of untreated HIV infected patients. *BMC Microbiol* 12:153.

220. Iwai S*, et al.* (2012) Oral and airway microbiota in HIV-infected pneumonia patients. *J Clin Microbiol* 50(9):2995-3002.

221. Vujkovic-Cvijin I*, et al.* (2013) Dysbiosis of the gut microbiota is associated with hiv disease progression and tryptophan catabolism. *Sci Transl Med* 5(193):193ra191.

222. Lozupone CA*, et al.* (2013) Alterations in the Gut Microbiota Associated with HIV-1 Infection. *Cell Host Microbe* 14(3):329-339.

223. Paradis E, Claude J, & Strimmer K (2004) APE: Analyses of Phylogenetics and Evolution in R language. *Bioinformatics* 20(2):289-290.

224. Ochman H & Jones IB (2000) Evolutionary dynamics of full genome content in Escherichia coli. *EMBO J* 19(24):6637-6643.