MOLECULAR STRATIFICATION AND PROGNOSTIC DETERMINANTS OF ADULT AND PEDIATRIC CROHN'S DISEASE

Benjamin Paul Keith

A dissertation submitted to the faculty at the University of North Carolina at Chapel Hill in particial fulfillment of the requirements for the degree of Doctor of Philosophy in the Curriculum in Bioinformatics and Computational Biology in the School of Medicine.

Chapel Hill
2020

Approved by

Terrence Furey

Shehzad Sheikh

Praveen Sethupathy

Joel Parker

Francisco Sylvester

ABSTRACT

Benjamin Paul Keith: Molecular Stratification and Prognostic Determinants of Adult and Pediatric Crohn's Disease
(Under the direction of Terry Furey, Praveen Sethupathy, and Shehzad Sheikh)

Crohn's disease (CD) is a chronic relapsing gastrointestinal inflammatory disorder with heterogeneous clinical presentation. Current clinical diagnostic methods involve invasive procedures and are ineffective towards predicting disease progression and response to therapy. High-throughput sequencing technologies have been utilized in various complex disorders to identify disease subtypes based on molecular signatures that are associated with clinical parameters. Given the success of molecular subtyping, particularly within the cancer field, there is substantive interest in implementing this methodology to predict disease progression and treatment response in CD. By analyzing differences in colonic gene expression between CD patients, our group revealed 2 subsets of patients—a group characterized by genes more highly expressed in the colon (colon-like CD) and a group with increased expression of ileum marker genes (ileum-like CD). Building on these initial findings, we aimed to validate molecular subtypes associated with CD through analysis of RNA-sequencing and small RNA-sequencing data in adult and pediatric cohorts. In the following chapters, I show that CD molecular subtypes can be detected using genome-wide expression of microRNAs, a class of small non-coding RNAs that post-transcriptionally regulate gene expression. Further, I show that microRNA-31 (miR-31) is a molecular driver of CD subtypes in both adult and pediatric patient cohorts and, through association analyses of clinical patient phenotypes, show that miR-31 expression levels are associated with the development of distinct disease outcomes. Using gene expression data from a large cohort of adult colon-like patients, I identified heterogeneous expression signatures that suggests an additional molecular subtype of CD associated with the expression of Paneth cell markers. Together, the results presented in the dissertation provide novel insights into the molecular heterogeneity of CD that can be used to guide future molecular subtyping research within the field by our group and the wider IBD research community.

ACKNOWLEDGEMENTS

Honestly, I've been dreading writing this section of my dissertation almost as much as any other. Not due to a lack of positive things to say or people to thank, but because I don't think I can adequately convey my thanks to so many people who have done so much for me over the past 4 and a bit years. If you're reading this, chances are you've permanently shaped how I think as a scientist and as a person, and I can't thank you enough.

The work presented here would not have been possible without the constant support of my advisors, Terry Furey, Shehzad Sheikh, and Praveen Sethupathy. The three of you make a brilliant team and it's always a pleasure to be in the same room as you guys, whether that be in person or on Zoom. Terry, aside from the first time we met and you asked me whether I supported Manchester United or Manchester City (a cardinal sin in the part of the world I'm from), you always seem to know what to say to really get the best out of me. Shehzad, your energy and passion toward every project we work on is honestly infectious and I hope I carry that passion to future projects. Praveen, when I arrived at UNC I struggled to give a self-introduction at the first BBSP session but you've since given me confidence to stand and speak in front of an audience, due in large part to the emphasis you place on training your students in the art of science communication. You have all been instrumental in this journey and I will always be tremendously grateful for the opportunity for work alongside you.

I would also like to thank my thesis committee: Joel Parker and Francisco Sylvester. In the few times that we've all met together for my committee meeting, you've provided new perspectives and helpful feedback that helped me to push my research further and gain further appreciation for the potential impacts our work could have on real people.

Thank you to my colleagues throughout the Furey, Sheikh, and Sethupathy labs, for your advice and for pulling me away from my little cubicle every now and again, particularly (but definitely not limited

to): Nur Shahir, Takahiko Toyonaga, Michelle Hoffner O'Connor, Raulie Raulerson, Rowan Beck, Mike Shanahan, Matt Kanke, Paul Cotney, Jeremy Simon, Bryan Quach, and Jeremy Wang.

It wouldn't be a proper acknowledgements section without thanking John Cornett and Cara Marlow. It wouldn't be at all dramatic to say that you guys have made this whole experience 90% less stressful than it could have been. You're always there when somebody in the department needs you, and you're always happy to help us. Thank you so much for everything over my time at UNC.

Finally, thank you to my family back in the UK and over here in the US. Although many of you still have absolutely no idea what I do, your support has never wavered. You've helped me through countless difficult times throughout my academic career to get to this point. Thank you for everything.

PREFACE

The research described in chapter 2 published in JCI Insight on 4 October 2018 and has been modified for this dissertation. Benjamin Keith acquired, analyzed and interpreted data, prepared figures, and served are the primary author of the manuscript. Jasmine Barrow acquired, analyzed and interpreted data. Nevzat Kazgan acquired, analyzed and interpreted data. Greg Gipson, Matthew Schaner, Michelle Hoffner O'Connor, Shruti Saxena, Omar Trad, Paul Cotney, Takahiko Toyonaga, and Nancy Allbritton acquired data; and Wendy Pitman and Matthew Kanke analyzed data. Neil Shah, Elisabeth Wolber, Nicole Chaumont, Timothy Sadiq, Mark Koruda, Dimitri Trembath and Francisco Sylvester provided help with tissue acquisition and patient phenotyping. Terrence Furey and Praveen Sethupathy designed the study, analyzed and interpreted the data, and obtained funding. Shehzad Sheikh conceptualized and designed the study, acquired the data, interpreted data, obtained funding, acted as study sponsor, and supervised the study.

The work discussed chapter 3 describes ongoing efforts using RNA-sequencing data to expand upon molecular CD subtypes. Benjamin Keith acquired, analyzed and interpreted data, and prepared figures. Matthew Schaner acquired data. Mike Shanahan aided data interpretation. Elisabeth Wolber Caroline Beasley, Nicole Chaumont, Timothy Sadiq, Mark Koruda, and Dimitri Trembath provided help with tissue acquisition and patient phenotyping. Terrence Furey, Praveen Sethupathy, and Shehzad Sheikh designed the study, analyzed and interpreted the data, obtained funding, acted as study sponsor, and supervised the study.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

AMP	Antimicrobial peptide
CC	Consensus clustering
CD	Crohn's disease
CL	Colon-like
DE	Differentially expressed
FDR	False discovery rate
FFPE	Formalin-fixed paraffin-embedded
GI	Gastrointestinal
GWAS	Genome wide association study
HTS	High-throughput sequencing
IBD	Inflammatory bowel disease
IEC	Intestinal epithelial cell
IL	Ileum-like
lncRNA	Long non-coding RNA
miR-31	MicroRNA-31
miRNA	MicroRNA
mRNA	Messenger RNA
NIBD	Non-IBD
PCA	Principal components analysis
PD	Paneth depleted
PE	Paneth enriched
qRT-PCR	Quantitative reverse transcription polymerase chain reaction
RNA-seq	RNA-sequencing
RPMMM	Reads per million mapped to microRNAs
smRNA-seq	Small RNA-sequencing
TCGA	The Cancer Genome Atlas
UC	Ulcerative colitis

CHAPTER I: INTRODUCTION

The release of the human genome in 2001, and the advances in high-throughput DNA sequencing technologies that followed, revolutionized our ability to detect changes in gene expression and regulatory mechanisms that cause complex disease. In the subsequent decade, large international consortia, such as ENCODE (1), Roadmap Epigenomics Project (2), and FANTOM (3), aiming to identify all functional elements within the human genome have provided vital insights into mechanisms of gene regulation. Initial efforts to apply information from these consortia to human disease through genome-wide association studies (GWAS) suggested that non-coding regulatory elements of the genome contribute to the development of complex diseases (4). This has presented significant opportunities within biomedical research to find the molecular causes of disease. The utilization of genomic technologies will provide new insights into gene expression and its heterogeneous regulatory modalities in a disease-specific context, revolutionizing disease diagnosis and providing novel disease classifications informed by changes in gene expression and gene regulation.

Post-transcriptional regulation of gene expression by microRNAs

The years following the initial draft of the human genome sought to further understand the ~99% of the genome that did not code for protein. Initially referred to as “junk DNA”, due to their inability to code for protein, consortium projects such as ENCyclopedia of DNA Elements (ENCODE) revealed that these non-coding regions are evolutionarily conserved serving important and diverse functions in regulating gene expression (5). Coordinating gene expression through intricate regulatory mechanisms ensures that cells maintain homeostasis through changing physiological conditions as well as control the fate of stem cell differentiation in distinct cells and tissues (6). The regulation of gene expression occurs through a broad range of mechanisms. Chapter II of this dissertation focusses on a specific class of small non-

coding RNAs, called microRNAs, which bind to and suppress the translation of target genes with wide-ranging functional utility and important roles in various disease contexts.

MicroRNAs (miRNAs) are a class of small interfering ncRNAs (siRNAs) ~18-22 nucleotides in length that post-transcriptionally regulate the translation of messenger RNA (mRNA) through interactions with their 3' untranslated region (UTR) (7). miRNA genes reside within intronic sequences of protein-coding genes, as well as long ncRNAs (lncRNAs), and predominantly utilize RNA polymerase II for their transcription (8). These immature transcripts, "pri-miRNAs", contain at least one hairpin structure that is recognized and cleaved by RNase III Microprocessor liberating the "pre-miRNA" which is subsequently transported to the cytoplasm. In the cytoplasm, another RNase III endonuclease, termed Dicer, cleaves the pre-miRNA which is finally processed into a mature miRNA sequence (8). This mature sequence is loaded onto a protein called Argonaute (Ago) to form a multiprotein complex, the RNA-induced silencing complex (RISC) (8). A miRNA loaded onto RISC will scan for target sites typically within the 3'UTRs of mRNAs that are complementary to a miRNAs seed region, nucleotides 2-8 of a miRNA from the 5' end. Gene silencing is subsequently achieved through the effector complexes of RISC preventing translation or destabilizing the mRNA target (9).

Since their initial discovery in nematodes in 1993 (10) (11), miRNAs have emerged as key regulators of biological processes. Early studies of miRNAs suggested that they are broadly conserved across diverse animal lineages (12) (7) and are found within virtually all cells and tissues (13), owing to their importance in gene regulation. A single miRNA can repress potentially hundreds of genes, although the singular effect of each miRNA is generally mild (13), and a single gene can have multiple miRNA target sites (9). Largely acting as fine-tuners of gene translation, miRNAs buffer against fluctuations in gene expression, shaping the topology of the transcriptome (7). Collectively, miRNAs operate complex gene regulatory networks contributing to a cell's ability to adapt and provide robustness against diverse environmental stimuli (14). In the early stages of development, miRNAs have been known to act as binary switches, conferring roles as master regulators of gene expression programs (9) (15).

According to the most recent update of the miRNA database miRbase (12) (October 2018), 2,654 mature miRNAs have been discovered in humans with more than 60% of protein-coding genes found to be under a selective pressure to maintain miRNA target sites (16). miRNAs are implicated in almost all

cellular processes but are essential in developmental processes, cellular differentiation, immune response, and maintaining homeostasis (13). Intercellular communication may also be further facilitated through the packaging of miRNAs into exosomes for the regulation of gene expression in distant cells (17). Deregulated miRNAs have been associated with numerous diseases, such as various cancers, cardiovascular disease, neurodevelopmental disorders, autoimmune diseases, skin diseases, and inflammatory bowel diseases (18) (19). Although elucidating the role of miRNAs in complex disease may provide a tool for identifying key genes and pathways, the development of therapeutics that utilize the gene silencing ability of miRNAs offer novel strategies for tackling disease (20). As recently as 2018, a novel treatment for a rare polyneuropathy became the first siRNA drug to receive FDA-approval (21) and numerous active early phase trials for therapeutic miRNA mimics and siRNAs that target specific miRNAs are currently ongoing (22). Further, due to their secretion from cells, miRNAs are found and can be isolated from biological fluids making them ideal biomarker candidates and potential prognostic markers of disease progression within clinical settings (23) (22).

High-throughput characterization of gene expression and gene regulation

Over the past decade, rapid advances in sequencing technologies have revolutionized the ability of biomedical researchers to tackle the complexities of genomes, gene regulation, and alterations in both resulting in disease. The introduction of massively parallel sequencing platforms in the mid-2000s supplied cost-effective, high-throughput methods to study various molecular characteristics of the cell genome-wide. Although the downstream chemistry employed by various platforms differ (24), high-throughput sequencing (HTS) is generally performed through the preparation of sequencing libraries followed by DNA amplification and the identification of DNA sequences in a platform-specific manner at nucleotide resolution (24). Modern sequencing platforms perform millions of sequencing reactions simultaneously, thus identifying millions of DNA molecules in parallel making these approaches high-throughput. The output data from these platforms, termed sequencing reads, subsequently require bioinformatic approaches to process, analyze, and interpret their biological significance.

Sequencing platforms are now used for a wide variety of applications. Through the isolation of specific fragments of DNA or RNA from cells, we can obtain genome-wide transcriptomic and regulatory

information. Transcriptomic information is obtained through the isolation of RNA transcripts and, for the compatibility for sequencing technologies, converted to a library of cDNA fragments before being sequenced (25). Following sequencing, the resulting transcriptomic readouts, referred to as “reads”, are aligned to the genome of interest to produce a genome-scale transcriptomic snapshot of the transcriptional landscape along with the level of expression for each gene and associated isoforms. RNA-sequencing (RNA-seq) can be utilized to assay all species of mRNAs and lncRNAs, whereas small RNA-sequencing (smRNA-seq) captures the small RNA content (<50 nucleotides) within the cell. By employing a size-exclusion step before sequencing, smRNA-seq facilitates the identification of small regulatory ncRNAs primarily comprising, but not limited to (26), miRNAs (27) (28). Compared with common alternatives for transcriptomic analyses, such as quantitative PCR and microarrays, RNA-seq provides a more accurate genome-wide quantitative determination of RNA abundance, does not require *a priori* sequence information, facilitating the discovery of novel RNAs and RNA isoforms, and can be used within species for which the genome has not been fully mapped (24) (27).

The integration of various ‘omics’ data allows researchers to untangle the interconnectivity between transcription and its complex regulatory processes (29). HTS has contributed novel insights into the regulation of the human genome through large-scale consortium projects such as ENCODE (1), and more recently the Roadmap Epigenomics Project (2), revealing the importance of epigenomics and transcriptomics in cell-type identity and disease development. As technological advances have driven us in the era of the sub-$1000 genome (24), HTS has become ubiquitous within biomedical research. Our greater understanding of the possible molecular causes of disease, together with the decreasing costs of HTS, has facilitated new methods to study human disease with the potential to revolutionize clinical decision-making tools.

Translational applications of high-throughput sequencing: biomarker discovery and molecular subtyping

Through the analysis of patient samples from diseased individuals compared with healthy controls, we can further understand the molecular basis of disease (30). HTS is now frequently being utilized for the identification of biomarkers and drivers of disease, which can be detected at the DNA (31), RNA (32), and protein levels (33) (34). Although biomarkers typically serve to differentiate affected and

healthy individuals, they can serve multiple purposes (35). In breast cancer alone, specific biomarkers are used to estimate the risk of disease development, determine prognosis, predict response to therapy, and monitor progression in metastatic disease (36). A major benefit of biomarker use, compared to more traditional clinical distinctions of disease, is their potential detection in circulation (whole blood, serum, or plasma) preventing the need for repeated invasive procedures (36). In addition to identifying molecular profiles from diseased and non-diseases individuals, HTS has proven useful in the discovery of molecular subtypes of disease.

Molecular subtyping involves the classification of samples using molecular data into clusters with distinct molecular profiles. Subtyping of disease across the various high-throughput molecular data allows for the classification of disease that associates better with clinical outcomes than traditional clinical methods. This has provided enhanced diagnostic, prognostic, and therapeutic options to treat disease while uncovering disease mechanisms that define disease heterogeneity (37) (38). Molecular subtyping in various cancer, largely driven by The Cancer Genome Atlas (TCGA) and the data made publicly available through their studies (39), has proven successful to classify patients into more homogeneous groups. A prominent example is the stratification of breast cancer into four distinct subtypes exhibiting distinct clinical outcomes (40), with subsequent studies largely recapitulating the identified subtypes using DNA methylation, copy number variation, and miRNA expression (41). Other notable examples include colorectal cancer (42), lung cancer (43) (44), leukemia (43) (45), pancreatic cancer (45), and multiple carcinomas (46) (47) (48) (49), resulting in the development of more targeted and personalized therapeutics. Advances in HTS, along with applications in the discovery of disease classifications and clinically relevant biomarkers, will supply novel breakthroughs in numerous fields of biomedical disease research. The inflammatory bowel diseases, particularly Crohn's disease, are disorders that would greatly benefit from approaches that have become common in the cancer field.

Crohn's disease

Crohn's disease (CD), one of the major categories of inflammatory bowel disease (IBD), is a chronic autoimmune disorder of the gastrointestinal (GI) tract caused by an abnormal immune response to luminal gut contents in genetically susceptible individuals (50). In 2015, it was estimated that

approximately 3.1 million adults received a diagnosis of IBD in the United States alone (51), with around half being attributed to CD (52). Compared with the other principal IBD, ulcerative colitis (UC), CD exhibits transmural inflammation that can occur at any point along the GI tract contributing to the highly heterogeneous nature of the disease in terms of location and progression (50). Due to this heterogeneity, diagnosing CD is challenging due to widespread and cryptic manifestations, with clinical features varying according to disease location typified by periods of relapse and remission with cycles of intestinal inflammation (50) (53). Current methods of diagnosing CD involve a combination of often invasive procedures, utilizing a combination of endoscopic, histological, and clinical findings to differentiate CD from UC and irritable bowel syndrome, which can often be difficult based on early symptoms (54). Therapeutic intervention is tailored through monitoring of disease presentation and progression with surgical interventions being required in up to two-thirds of CD patients during their lifetime (53). There is currently no cure for CD, warranting further characterization of the underlying cause of disease and the implementation of methods that describe the heterogeneity in disease presentation and progression.

Although the cause of CD is unknown, the manifestations of the disorder are highly heterogeneous with multifactorial etiology involving interactions between genetics, enteric microbiota, and the immune system (55). Since the publication of the first GWAS in CD in 2005 (56), 242 susceptibility loci have been associated with the presence of IBD (57). Further supported by twin studies (58), these findings have provided strong evidence of a strong genetic contribution to the disease. Attempts to provide molecular mechanisms associated with these variants have proved successful. Examples include studies of differential NOD2 expression suggesting a 2-4 times increased risk for IBD (59), IL23R expression conferring resistance against the development CD5 (60), and the association of HLA alleles with the development of colonic disease (61) (62).

Consistent with other large scale GWAS studies (5), the majority of CD-associated loci are located within non-coding regions of the genome (63) suggesting the importance of gene regulation in the development of CD. miRNA expression has been found to be dynamic in both tissue (64) (65) and peripheral blood (66) (67), suggesting miRNAs as biomarkers for CD diagnosis and differentiation from UC, especially in early-onset and pediatric CD (68). miRNAs have also provided novel molecular insights through crucial roles in regulating autoimmunity and inflammation, with specific miRNAs exhibiting

regulatory roles in the maintenance of intestinal epithelia (69) (70) and the differentiation, maturation, and function of innate and adaptive immune cells (71). Recent studies using lncRNAs suggest important roles in immune responses associated CD pathogenesis with unique expression signatures in IBD relative to normal controls (72) and specific expression in pediatric CD (73). Although beneficial as a research tool to understand the molecular and genetic determinants of CD, these studies have not aided advancements in CD diagnosis, treatment selection, or prognosis (74).

Genome-wide profiling studies of IBD have aided the identification of distinct CD and UC signatures using gene expression (75) (76), lncRNA expression (72), DNA regulatory elements (77), histone modifications (78), and miRNA expression (79) (80). Specific profiling of CD may provide insight into disease heterogeneity through the identification of CD subtypes that associate with clinical measures of disease activity and predict disease progression, which commonly used clinical activity scores fail to achieve (53) (81). A previous study by our group assessed transcriptional and regulatory landscapes of CD using RNA-seq to quantify gene expression and Formaldehyde-Assisted Isolation of Regulatory Elements sequencing (FAIRE-seq) to identify regions of accessible chromatin (82). Using principal component analysis, adult CD patients stratified into two specific subgroups. Differential expression revealed a contrasting enrichment for markers of normal colon-tissue (colon-like) and ileum-tissue (ileum-like), despite all tissue samples originating from colonic tissue (82). Pathway enrichment indicated that the ileum-like subclass was characterized by upregulated lipid and xenobiotic metabolism pathways, whereas colon-like samples generally showed increased gene expression in energy metabolism pathways. Importantly, applying clinical phenotype data to identified clusters revealed that ileum-like patients were more likely to require biologics post-surgery while an increased incidence of severe rectal disease and need for colectomy was associated with the colon-like subgroup (82), indicating the potential translational utility of these identified subclasses.

In this dissertation, I further explore the molecular subtypes introduced by Weiser et al. (82). In chapter II, I investigate the contribution of miRNAs to colon-like and ileum-like subtypes in adult and pediatric CD cohorts using high-throughput genome-wide miRNA profiles (83). Through the identification of a specific miRNA, miR-31, by differential analysis and correlating miRNA activity with target genes, I show the utility of miRNAs in CD subtype classification and their potential utility in the clinic through the

application of clinical data to the two molecular subtypes (83). In Chapter III, I expand upon our established CD subtypes using genome-wide gene expression profiles in a large cohort of adult CD patients followed by clinical phenotype association testing. Through unsupervised clustering analyses and differential gene expression analysis, I identify a third CD molecular subtype. In Chapter IV, I discuss the contributions of CD molecular subtypes in furthering our understanding CD pathogenicity and I conclude by discussing the potential significance of my findings for future studies using genome-wide expression profiling as well as the downstream translations impacts of my findings.

CHAPTER II: COLONIC EPITHELIAL MIR-31 EXPRESSION ASSOCIATES WITH THE DEVELOPMENT OF CROHN'S DISEASE PHENOTYPES IN ADULT AND PEDIATRIC POPULATIONS[1]

INTRODUCTION

Crohn's disease (CD), one of the primary inflammatory bowel diseases (IBD), is a chronic inflammatory condition of the gastrointestinal tract resulting from an aberrant immune response to the enteric microbiota in a genetically susceptible host. CD is highly heterogeneous in disease location, behavior, and progression. Using gene expression and chromatin accessibility profiles in colon tissue, we previously identified two molecular subtypes in adult CD associated with unique phenotypes (82). Recent studies validate the premise that specific genetic and molecular profiles are associated with, and may contribute to, disease heterogeneity and behavior. Over 200 genetic loci have been significantly associated with CD risk (84). A study of 29,838 adult individuals did not identify DNA variants predictive of CD behavior over time, but did associate genetic variants in IBD with disease location (74). Notably, a longitudinal inception cohort study of treatment-naïve pediatric CD patients revealed lipid metabolism and extracellular matrix gene expression signatures in the ileum as predictive of response to steroids and fibrostenotic ileal CD, respectively (75, 85). However, a more complete set of robust prognostic determinants for CD phenotypes, especially incorporating non-coding RNAs, is still lacking. As such, there remains active, substantive interest in the CD research community to identify specific genetic and molecular factors that mark disease subtypes, and more importantly, inform on disease progression and outcome.

Distinct disease outcomes of CD are likely due in large part to variability in cellular processes that underlie the natural history of CD. Disruption of the intestinal epithelial barrier and loss of tolerance by

[1] This chapter originally appeared in the Journal of Biological Chemistry. The original citation is as follows: Keith BP et al. Colonic epithelial miR-31 associates with the development of Crohn's phenotypes. JCI Insight 2018;3(19). doi:10.1172/JCI.INSIGHT.122788

immune cells to the enteric microbiota are critical cellular events that lead to chronic inflammation seen in CD. Precise cell type-specific mechanisms leading to these dysfunctions are poorly understood. Recently, microRNAs (miRNAs) that confer post-transcriptional regulation of gene expression have emerged as key modulators of intestinal epithelial cell (IEC) biology (69, 86) and of pathways that underlie the pathogenesis of CD (87, 88). Mice deficient for miRNAs in the intestinal epithelium exhibit altered intestinal architecture and increased barrier permeability (69), which leads to immune cell infiltration and severe intestinal inflammation.

In this study, we identified miRNA-31 (miR-31) as the primary contributor to our previously identified two major molecular subtypes of adult CD patients. We determined that the upregulation of miR-31 in colonic tissue of CD patients is driven in large part by increased expression specifically in IECs. Importantly, we expanded our study to incorporate a large cohort of 234 formalin-fixed paraffin embedded (FFPE) index biopsies of colon and ileum tissue from 127 treatment-naïve pediatric patients and non-IBD (NIBD) controls. In medically refractory adult CD patients undergoing surgical resection, one subtype with a lower, more typical level of colonic miR-31 expression at the time of surgery was associated with a worse post-operative outcome (as measured by recurrence in the neo-terminal ileum at the anastomotic site) and need for subsequent colectomy. In pediatric patients, the same lower colonic miR-31 expression subtype in index biopsies was associated with progression to fibrostenotic ileal disease. Our study shows that miR-31 is a candidate prognostic determinant of CD behavior in adult and pediatric patients and highlights the potential role of miR-31 in the pathobiology of CD.

RESULTS

MicroRNAs and lncRNAs stratify adult CD patients into two molecular subtypes

Previously, we demonstrated that medically refractory Crohn's disease (CD) patients undergoing surgery clustered into two distinct groups using principal component analysis (PCA) of Mrna expression by RNA-seq on uninflamed colonic mucosa from 21 adult patients with CD and 11 adult control patients (NIBD) (82). Analysis of genes differentially expressed between these two groups revealed that genes more highly expressed in the colon of one group were enriched for previously identified NIBD colonic marker genes, while genes more highly expressed in the second group were enriched for normal ileum

marker genes. We labeled these groups colon-like (CL) and ileum-like (IL). We showed by a prospective analysis that these CL and IL CD subgroups exhibit colonic CD and ileal inflammation, respectively. To evaluate further whether this molecular stratification was evident within non-coding RNAs, we analyzed small RNA-seq data from most of the same CD and NIBD patients (18 CD, 12 NIBD) to quantify the expression of microRNAs (miRNAs), which we previously showed was able to distinguish CD patients from NIBD controls (79). We also re-interrogated the RNA-seq data from the same patients to quantify long, non-coding RNAs (lncRNAs). PCA on each of the miRNA and lncRNA datasets (Figure 2.1A and 2.1B; Table 2.1) revealed that CD samples clustered into the same distinct CL and IL groups as initially defined with the Mrna data, which we also recapitulated in this study using updated gene annotations (Figure 2.2). These data demonstrate that the CL and IL CD subtypes are defined by expression profiles of several types of RNA molecules, which perform diverse functions within the cell.

MiR-31 is the primary driver of molecular stratification and is associated with post-operative outcome in adult CD patients

To identify the miRNAs that contribute most to the stratification of the two molecular CD subtypes, we initially compared genome-wide miRNA expression profiles between the 9 CL and 9 IL CD patients. We found that 19 miRNAs were significantly differentially expressed between the two groups ($|\log_2(FC)| > 1$, FDR < 0.05). Strikingly, we observed a 13.5-fold change in miR-31-5p (miR-31; $P_{adj} = 1.43 \times 10^{-18}$) between CL and IL samples. Analysis of PCA components revealed that miR-31 is the top contributor to the variance observed for principle component (PC)-2 that separates the CL and IL patients (Table 2.1). These findings suggest that miR-31 expression can stratify CD into two major molecular subtypes (Figure 2.3A).

We and others have identified miR-31 as a discriminant more generally of CD and NIBD patients (80, 89). We hypothesized that this difference is driven primarily by CD patients in the IL group. To test this hypothesis, we compared the levels of miR-31 in each of IL and CL groups relative to NIBD. We observed a dramatic and highly significant up-regulation (~60-fold) of miR-31 in IL patients compared with NIBD patients ($P_{adj} = 2.59 \times 10^{-51}$; Figure 2.3B). We also detected a significant difference in expression between CL and NIBD (~4-fold, $P_{adj} = 7.66 \times 10^{-06}$; Figure 2.3B); however, the magnitude of the difference is much lower. These findings support the above-stated hypothesis, indicating that while miR-31 is a

strong marker of disease presence in all CD patients, this signal is driven predominantly by those patients of the IL subtype.

Expression levels of mature miRNAs can be altered in several ways, including changes to the rate of transcription, efficacy of the maturation (biogenesis) process, and RNA stability. To determine whether the miR-31 locus is subject to enhanced transcription in IL CD patients, we quantified the normalized density of RNA-seq reads mapping to the primary transcript of miR-31 (*MIR31HG*) across all samples. We observed that transcription levels of *MIR31HG* are indeed dramatically elevated in the IL subgroup relative to both the CL subgroup and NIBD patients (Figure 2.3C). These data indicate that increased level of transcription is one major contributor to the observed difference in miR-31 levels between IL patients and the NIBD and CL patients. Notably, RNA-seq data from the ileum of an NIBD patient (Figure 2.3C) revealed a signal at the *MIR31HG* locus that closely resembles the signal from the colon of IL CD patients.

MiRNAs regulate gene expression by binding to recognition elements in the 3' untranslated regions of target mRNAs and marking the mRNAs for translational repression and degradation (90). Therefore, we sought to determine, using our published tool miRhub (91), whether genes that are down-regulated in IL relative to NIBD are enriched for predicted target sites of miR-31 or any other miRNA shown to be upregulated in the colon of IL patients. Notably, we found that miR-31 is the only upregulated miRNA whose target genes are significantly enriched among the genes downregulated in IL patients compared to both CL and NIBD patients (empirical $P < 0.05$; Figure 2.4). This indicates that miR-31 is not only dramatically elevated in the IL subtype of CD, but also a candidate master regulator of genes that are downregulated in that subtype.

To validate the differential expression of miR-31 between the IL and CL subtypes of CD, we measured colon miR-31 levels in an independent cohort of 40 adult CD and 29 NIBD patients using Qrt-PCR. Biopsies were obtained at the time of surgical resection for medically refractory disease. We first recapitulated the finding that miR-31 levels are significantly up-regulated overall in CD relative to NIBD ($P = 3.27 \times 10^{-4}$, 2-tailed unpaired Student's *t* test; Figure 2.5A). As expected, we also found that miR-31 expression levels stratify CD patients into two subgroups, "high" and "low", which we hypothesized reflect the IL and CL molecular subtypes, respectively. To test this hypothesis, we measured Mrna levels of

APOA1 (Figure 2.5B), a marker gene in ileum, and *CEACAM7* (Figure 2.5C), a marker gene in colon, both of which we previously showed can stratify IL and CL patients (92, 93). We found that the patients with high colonic miR-31 expression also show high *APOA1* expression and low *CEACAM7* expression, and we observed the opposite trend for the patients with low colonic miR-31 ex pression. Altogether, these data confirm that miR-31 expression levels stratify CD patients into two molecular subgroups.

We then studied prospectively the clinical characteristics of adult patients after surgery for medically refractory disease. Since all patients had disease removed at initial surgery, we followed post-surgery disease recurrence based on Rutgeerts post-operative endoscopic scoring (94) of the neo-terminal ileum or the need for an end ileostomy due to severe refractory disease within a year after the initial surgery. Post-operative management as well as timing of endoscopy for reassessment was determined by the managing IBD specialist. Most patients had a post-operative staging colonoscopy within one year of surgery. Recurrence was defined as having a Rutgeerts score of i2, i3, i4 or the need for an end ileostomy within a year after the initial surgery. No recurrence was defined as a Rutgeerts score if i0, i1. Strikingly, despite similar patient demographics at time of surgery as well as no significant differences in post-operative management between the two subtypes (Table 2.3), the CL subtype of CD patients demonstrated a worse post-operative course compared to the IL subtype (Table 2.3, p=0.030). While, this patient population is not anti-TNF treatment naïve, to our knowledge, this data provides the first evidence for the potential clinical utility of miRNA profiling to predict a poor post-operative outcome of CD.

MiR-31 is dramatically up-regulated in intestinal epithelial cells and crypt derived colonoids established from adult CD patients

Colon tissue is composed of several distinct cell types, and expression studies in tissue do not reveal from which particular cells transcripts originated. To measure miR-31 expression in specialized cell types of the colon, we isolated intestinal epithelial cells (IECs; CD326+) and matched lamina propria immune cells (CD3+ T cells, CD20+ B cells, CD33+CD14- resident intestinal macrophages, CD33+CD14+ infiltrating inflammatory intestinal macrophages) by flow cytometry from macroscopically uninflamed tissue from adult patients with CD (N=11-20) and NIBD controls (N=8-16). While relative miR-31 expression levels based on Qrt-PCR were increased in B cells and resident macrophages isolated

from CD patients compared to NIBD controls ($P < 0.05$, 2-tailed unpaired Student's t test), these results were dwarfed in comparison to the increase seen in IECs (~52-fold difference, $P = 1.28 \times 10^{-8}$; Figure 2.5D).

To evaluate this finding further, we established three-dimensional epithelial colonoids from crypts isolated from both CD patients and NIBD individuals. These structures contain crypt-like domains reminiscent of the gut epithelium, and they continuously produce all cell types found normally within the intestinal epithelium (95). We found that colonoids from CD patients express significantly higher levels of miR-31 compared to NIBD controls, similar to the primary tissue from which the colonoids were derived (Day 2 $P = 0.041$, 2-tailed unpaired Student's t test; Day 6 $P = 0.0095$, 2-tailed unpaired Student's t test; Figure 2.5E). These results suggest upregulated miR-31-5p is not a transient result due to external signalling but is a predisposing factor in IECs of CD patients. Disruption of the intestinal epithelial barrier is a critical determinant of the predisposition to chronic inflammation and fibrosis seen in CD. Going forward these data open up the potential to understand the impact of miR-31 on barrier function.

MiR-31 expression in formalin-fixed paraffin-embedded (FFPE) tissue from treatment-naïve pediatric CD patients also defines two subtypes and is associated with development of ileal fibrostenotic disease

The molecular profiles we have generated and analyzed in fresh tissue and cells from adult CD represent a fundamental advance in understanding adult CD heterogeneity. At the time of this analysis, though, these adult patients had progressed to medically refractory disease, each with individual treatment histories that could potentially confound results. Therefore, as a next step, we performed smRNA-seq on microscopically uninflamed FFPE mucosal tissue from ascending colon and terminal ileal biopsies in age-matched treatment-naïve pediatric patients with CD (n=76) and NIBD controls (n=51) obtained at the time of diagnosis (index colonoscopies). It is important to note that this is not a validation cohort of the adult CD, but rather a completely independent analysis that offers at least five unique advantages. Firstly, as noted above, these samples are from treatment-naïve individuals, which greatly mitigates the potential confounding effects of treatment history that may be present in adults. Secondly, the samples are FFPE as opposed to fresh frozen tissue. Successful molecular subtyping of CD patients using FFPE tissue will greatly expand our ability in the future to analyze retrospectively the clinical

characteristics associated with subtypes, given that most tissue biopsies are bioarchived as FFPE. Thirdly, the number of samples is substantially greater than in our adult CD study, affording additional power for molecular subtyping. Fourthly, we have matched ileum and colon biopsies from the same patient allowing for the interrogation of site-specific changes and impact on disease phenotype. Finally, these tissue samples are index biopsies, obtained at the time of diagnosis and prior to significant disease progression, which provides a unique opportunity to determine whether miR-31 expression is associated with the development of CD phenotypes.

As in the adult cohort, we found that the levels of miR-31 expression in the colon are significantly upregulated in CD patients relative to NIBD controls (~7.8-fold, $P = 4.64 \times 10^{-7}$, 2-tailed unpaired Student's *t* test; Figure 2.6A and 2.7). We observed that miR-31 expression in the ileum is also significantly upregulated in CD patients ($P = 9.97 \times 10^{-7}$, ~1.5-fold), however the effect is not nearly as pronounced as in the colon (Figure 2.6B). This may be due in part to significantly higher baseline miR-31 expression levels in the ileum of unaffected (NIBD) individuals compared to in the colon ($P = 5.71 \times 10^{-28}$; Figure 2.8).

Using miRNA expression data from the 100 most variable miRNAs, we independently performed PCA on the colon (Figure 2.6C) and ileum (Figure 2.6D) pediatric samples and observed a robust separation of NIBD and CD patients. Notably, miR-31 is the largest contributor to this stratification in the colon, but not in the ileum (Table 2.4 and 2.5). This indicates that specifically colonic miR-31 is a primary marker of disease presence.

We investigated whether colonic miR-31 levels were associated with the eventual development of specific CD phenotypes and tested for association with clinical features both at the time of diagnosis and across disease course (Table 2.6). We first analyzed pediatric NIBD samples and found that all colon samples but one had miR-31 levels < 150 RPMMM and all ileum samples had mir-31 levels > 150 RPMMM (Figure 2.9). Using this threshold, we defined two distinct subgroups within our colonic pediatric CD samples as "miR-31-low" (n = 46) and "miR-31-high" (n = 30). MiR-31 expression was validated in our two subgroups through Qrt-PCR of a subset of low- (n = 7) and high-miR-31 (n = 7) samples ($r = 0.94$, $P = 3.89 \times 10^{-7}$; Figure 2.10).

We then studied prospectively the clinical characteristics of only pediatric patients that presented with inflammation at time of diagnosis (i.e., no initial stricturing, penetrating disease). Since all patients

were treatment naïve, we defined stricturing CD as primary (not anastomotic) fibrostenotic stricture of the terminal ileum where medical treatment would be ineffective, and therefore, surgical resection was considered a reasonable treatment option. These were diagnosed based on physician preference of using standard endoscopy and/or computed tomography (enterography) (CTE) or magnetic resonance imaging (MRI) and correlation with patient symptoms. Low miR-31 expression was significantly associated with the eventual development of ileal stricturing (P = 0.001) and having surgery involving an anastomosis (P = 0.048). Remarkably, we found that no miR-31-high patients progressed to develop a stricturing phenotyping. To our knowledge, this data provides the first evidence for the potential clinical utility of miRNA profiling to predict increased risk of the development of stricturing phenotype in patients with Crohn's disease.

DISCUSSION

We identified colonic miR-31 expression as central to clinically-relevant molecular subtypes found in independent cohorts of adult and treatment-naïve pediatric patients. Notably, low levels of miR-31 in medically refractory adult CD patients at the time of surgical resection are indicative of a worse post-operative outcome as measured by recurrence in the neo-terminal ileum. Similarly, lower miR-31 expression in pediatric patients at the time of diagnosis is indicative of increased risk for development of ileal stricturing complications. Our study introduces small RNAs as potential predictors of disease phenotype and, with use of FFPE samples, offers distinct advantages over Mrna studies in the context of fresh tissue. These findings are reminiscent of early descriptions of transcriptomic signatures in breast cancer (40). Further large-scale studies of gene expression profiles in breast tumors, including those of The Cancer Genome Atlas (TCGA) project (96), eventually established four major molecular classes that vary in their aggressiveness and respond differently to therapies. Similarly, diffuse large B cell lymphoma (97), glioblastoma (98), endometrial cancer (49), and lung cancer subtypes (44) have been identified by genomic profiling, facilitating the development and application of targeted therapies (https://cancergenome.nih.gov).

Our study includes two distinct populations of patients with disease at different stages of development. It is unclear how clinical associations are related to patient age and/or disease state. For

instance, molecular levels at initial diagnosis that predict disease progression may not be maintained once disease has actually progressed (99). Long-term longitudinal studies will need to be conducted with serial quantification of genomic profiles over the course of disease evolution in pediatric patients transitioning into adulthood. It is also imperative we understand the unique characteristics of disease presentation and evolution in adult patients, impacted by major life style and environmental factors, each uniquely contributing to colonic miR-31 regulation and its impact on phenotype. Medical management of fibrostenotic ileal disease is unpredictable and in many cases, is not long lasting and requiring surgery. Thus, results from these longitudinal studies may eventually impact treatment designs for these difficult disease phenotypes. The robust establishment of CD subtypes may also influence future design of clinical trials where subtypes can be considered during patient randomization, allowing for better evaluation of subtype identification when making therapeutic decisions.

Recent studies have started to unravel the molecular mechanisms associated with distinct IBD phenotypes. Genetic variants in *NOD2*, *MHC*, and *MST1* 3p21 were shown to be associated with disease location (colonic CD, ileal CD and ulcerative colitis (UC)) but not disease behavior (74). But, the genetic contribution to CD pathogenesis has been shown to be disproportionate, ranging from most impactful in very early onset IBD (100) (VEOIBD) to modest significance in older pediatric and adult IBD patients (101–104). In rectal tissue from pediatric patients, expression patterns of *IL-13*, *IL23A*, and *IL17* distinguished colonic CD from UC (105). Also, a lipid metabolism related gene expression signature in the ileum of pediatric CD patients accurately predicted 6-month steroid-free remission (75). Follow-up studies of these same ileal samples showed a distinct collagen and extracellular matrix gene expression signature present at time of diagnosis in a subset of patients who developed fibrostenotic ileal disease (85). Interestingly, our prior analysis of these patients identified an association between these same pathways in the ileum and the CL molecular phenotype (82). Moving forward, the challenge is to define molecular subtypes while also uncovering the cell type-specific genetic, molecular, and environmental contributors to each subtype.

This current study along with our previous study have now shown that whole genome Mrna, miRNA, and lncRNA transcript levels, along with the open chromatin landscape, define two molecular adult CD subtypes. In addition, miRNA expression patterns can stratify pediatric CD. Together, these

findings suggest that across CD patients, colonic tissue is altered in different ways at a cellular level supporting the idea of multiple Crohn's diseases. This also underscores the necessity for a more complete molecular characterization of CD across larger populations to uncover additional distinct subtypes. We advanced our work into FFPE tissue which opens the possibility to increase sample numbers, perform longitudinal follow-up studies, and facilitate the association of molecular markers to disease course.

We demonstrate miR-31 to be specifically dysregulated in colonic epithelial cells. Breakdown in the intestinal barrier is critical to intestinal chronic inflammation; a hallmark of CD. MiRNAs, including miR-31, are known to have significant contributions to gastrointestinal epithelial barrier function (70). *Dicer1* deficient mice display colonic barrier integrity dysfunction as evidenced by lymphocyte and neutrophil infiltration as well as mis-localization of the tight junction protein Claudin-7 (69). In the esophagus, Hussey et al. found that miR-31 is one of only a few differentially expressed miRNAs in post-ablation epithelium with increased barrier permeability (106, 107). In the colon, Wu et al. postulate that lowly expressed miR-31 plays a protective role after hypothermic ischemia induced barrier dysfunction in the colon, perhaps aiding in post-injury healing, specifically by targeting the hypoxia inducible factor (HIF)-factor inhibiting *HIF* (*FIH-1*) pathway (108). Using combinational computational methods to predict miR-31 target-pathways, one group found a connection specifically between miR-31 and tight junctions in lung epithelium (109). Most recently, Yu et al. demonstrated using *in vivo* knock-in and knock-out models that miR-31 plays a role in regulating intestinal stem cell behavior during regeneration after radiation injury (110). We show that patient crypt-derived colonoids in a sterile environment retain the aberration in miR-31 expression present in the tissue of origin, which supports a cellular defect that is intrinsic and not secondary to inflammation or other external signals due to the presence of disease. The colonoid experimental system will enable future studies to interrogate the role(s) of specific factors in driving a fibrostenotic phenotype, especially in the context of co-culture with lamina propria immune cells, mesenchymal cells, as well as stimulation with commensal and/or colitogenic bacteria.

In summary, we provide the most comprehensive molecular characterization of CD to date. We uncover miR-31 as an identifier of CD, but more importantly a molecular stratifier of both pediatric and adult patients, an indicator of established disease phenotype in adult patients, and a predictor of clinical

phenotype at the time of diagnosis in pediatric patients. These findings represent significant progress in molecularly defining the Crohn's disease(s), moving closer toward potential personalization of therapy and improving outcomes.

MATERIALS AND METHODS

Patient populations and phenotyping

Adult and pediatric patients with CD and NIBD related illnesses diagnosed at The University of North Carolina hospitals (UNC) were included in this study. Clinical phenotypes considered in this study include demographic and clinical variables such as age, sex, disease duration, age at diagnosis, age at sample acquisition, disease location, and disease behavior. Summarized (Table 2.7) and detailed information of patient demographics and phenotypes for the adult and pediatric cohorts are provided. This study was not blinded, and all authors had access to the study data and reviewed and approved the final manuscript.

Tissue isolation and characterization

For our adult cohort, all CD and NIBD mucosal biopsies were obtained from macroscopically unaffected sections of the ascending colon at the time of surgery and flash-frozen. No samples showed signs of active microscopic inflammation or disease, as confirmed by an independent pathologist. Treatment-naïve pediatric patients were diagnosed at UNC. From formalin-fixed, paraffin-embedded (FFPE) tissue, mucosal sections from both macroscopically and microscopically non-inflamed sections of the ascending colon and terminal ileum from the time of initial diagnosis (index biopsy) were identified by a pathologist, and scrolls were obtained for small RNA isolation. Absence of acute (active) inflammation, including neutrophilic inflammation of crypt epithelium and crypt abscess formation, and chronic inflammation, including architectural distortion and basal lymphoplasmacytosis of the lamina propria, was determined after review of each H&E stained slide (Figure 2.11).

RNA isolation, sequencing, and analysis

RNA was isolated from flash-frozen adult samples from surgical resections using the Qiagen RNeasy Mini Kit (Valencia, CA) following the manufacturer's protocol. This kit uses column-based DNase

treatment to eliminate DNA contamination, and allows the miRNA and mRNA content to be preserved. miRNA was enriched from FFPE tissue for pediatric samples using the Roche High Pure miRNA Isolation Kit (Penzberg, Germany). RNA purity and integrity were assessed with Thermo Scientific NanoDrop 2000 (Waltham, MA) and Agilent 2100 Bioanalyzer (Santa Clara, CA), respectively. For all clinical categories of flash frozen adult samples, we observed average RNA integrity (RIN) values above 7.

RNA-seq libraries were prepared using the Illumina TruSeq polyA+ Sample Prep Kit. Paired-end (50 bp) sequencing was performed on the Illumina HiSeq 2500 platform (GEO accession GSE85499). Reads were aligned to the GRCh38 genome assembly using STAR (111) with default parameters. Transcript expression was quantified with Salmon (112) using default parameters. Post-alignment normalization and differential analysis was performed using DESeq2 (113) with GENCODE_V25 gene annotations requiring base mean expression >10 and an FDR <0.05.

Small RNA libraries were generated using Illumina TruSeq Small RNA Sample Preparation Kit (San Diego, CA). Single-end (50 bp) sequencing was performed on the Illumina HiSeq 2500 platform (GEO accession GSE101819). miRquant 2.0 (114) was used for miRNA annotation and quantification. Samples with less than 3 million reads mapping to miRNAs were excluded. Differential analysis was performed using DESeq2 (113).

PCA was performed using the prcomp function in R on DESeq2 normalized VST transformed counts for mRNAs ("protein_coding" in GENCODE_V25) and lncRNAs ("lincRNA" or "antisense" in GENCODE_V25) with an expression base mean > 10. For miRNA expression data, PCA was performed using reads per million miRNAs mapped (RPMMM) normalized $\log_2$ transformed counts for the 100 miRNAs with the highest standard deviation values across all samples and a normalized expression level of 500 RPMMM across at least 20% of samples. For pediatric samples, we eliminated 18 miRNAs not found in the adult samples to remove potential artifacts due to FFPE preservation. Candidate master regulator miRNAs were detected using miRHub (91), using "non-network" mode and requiring a predicted target site to be conserved between human and at least two other species.

Quantitative reverse transcriptase PCR

For miR-31, total RNA was isolated from tissues using Norgen's Total RNA Purification Kit (Thorold, ON, Canada). 50ng of RNA was used for reverse transcription with the Life Technologies

TaqMan MicroRNA Reverse Transcription Kit (Grand Island, NY). MiRNA qRT-PCR were performed using the TaqMan Universal PCR Master Mix per Life Technologies' protocol, on Bio-Rad Laboratories CFX96 Touch Real Time PCR Detection System (Richmond, CA). Reactions were performed in triplicate using RNU48 as the normalizer. For *APOA1* and *CEACAM7*, total RNA was isolated as described above. cDNA was derived from 1µg RNA by reverse transcriptase using the BioRad iScript cDNA Synthesis kit. RT-qPCR was then performed on these cDNA samples using the BioLine Hi-ROX SYBR kit.

LPMCs and IECs were isolated from intestinal specimens using modifications of previously described techniques (115). LPMCs were isolated from human colon by an enzymatic method, followed by Percoll (GE Healthcare, Piscataway, NJ) density-gradient centrifugation. LPMCs were further separated into CD33+14+ peripheral macrophages, CD33+CD14- intestinal resident macrophage, CD20+ B cells, and CD3+ T cells corresponding antibody labeled microbeads (Miltenyi Biotec, Auburn, CA). IECs were isolated from human colon mucosa using Ethylenediaminetetraacetic acid (EDTA) followed by magnetic bead sorting via CD326 labeled microbeads. Purity was >90% by flow cytometric analysis.

Colonoid generation and analysis

Epithelial colonoid cultures were generated from non-inflamed regions of colon tissue from NIBD controls and CD patients. The intestinal tissues were washed and mucosectomy performed with surgical scissors. Minced colonic mucosal fragments were incubated at 37°C in 5 ml of digestion media (1 mg/ml collagenase VIII in Advanced Dulbecco's modified Eagle medium/F12 (ADF), 10% FBS, 15mM HEPES buffer, penicillin/streptomycin, 2mM Glutamax, 100ug/ml Primocin (Invivogen, antibiotic/antimitotic), 10uM Y-27632) for 30 minutes with mechanical disruption. The digested tissue/crypts were centrifuged at 200*g* for 5 minutes to separate crypts from single cells. Pelleted colonic crypts were resuspended in 5 ml of digestion media and centrifuged again at 200*g* for 5 minutes. Volume of crypts needed for 40-50 crypts per 96-well well was centrifuged in 1.5 mL tubes at 2500 RPM for 5 minutes. Crypts were embedded in appropriate volume of Growth Factor Reduced Matrigel (Corning) on ice and seeded at 10uL per 96-well. Basal stem culture medium (50% WNT3a conditioned media, 50% R-spondin 2 conditioned media, supplemented with 1 mM HEPES, 2mM Glutamax, 1X N2, 1X B27, and 1 mM *N*-acetylcysteine, 100ug/ml Primocin, with growth factors 50ng/mL murine EGF, 100ng/mL murine noggin, 1 ug/mL gastrin, 0.01uM

PGE2, 10mM nicotinamide, and small molecule inhibitors 500 nM LY2157299, 10 uM SB202190) with 10 uM Y-27632 was added at 100uL per well. At selected timepoints, colonoids embedded in matrigel were lifted from wells with cold ADF. For miRNA analysis, day 2 and 6 reverse transcription and quantitative real time PCR for miR-31 and *RNU-48* (housekeeping) were performed using predesigned TaqMan miRNA assays (Life Technologies). The relative expression was calculated by the comparative CT method and normalized to the expression of *RNU-48*.

Statistics

Differential expression analyses of RNA-seq and small RNA-seq data were performed using DESeq2 (113), with FDR adjusted p-values being used to measure statistical significance. MicroRNA target enrichment was determined using miRHub, which generates empirical p-values through Monte Carlo simulations (91). Significance of differential expression in RT-qPCR and colonoid assays was assessed using Student's t test (unpaired, 2 tailed) to compare 2 groups of independent samples. Significance of association with patient phenotype data was determined using Fisher's exact test (categorical data) or a 2-tailed unpaired Student's t test (continuous data). For all tests, $P_{adj} < 0.05$, empirical $P < 0.05$, or $P < 0.05$ was considered statistically significant.

Study Approval

Both the adult and pediatric sections of this study received Institutional Review Board approval at UNC (protocol 10-0355 and 15-0024). Written informed consent was received from all participants prior to inclusion in the study. All participants are identified by number and not by name or any Protected Health Information (PHI).

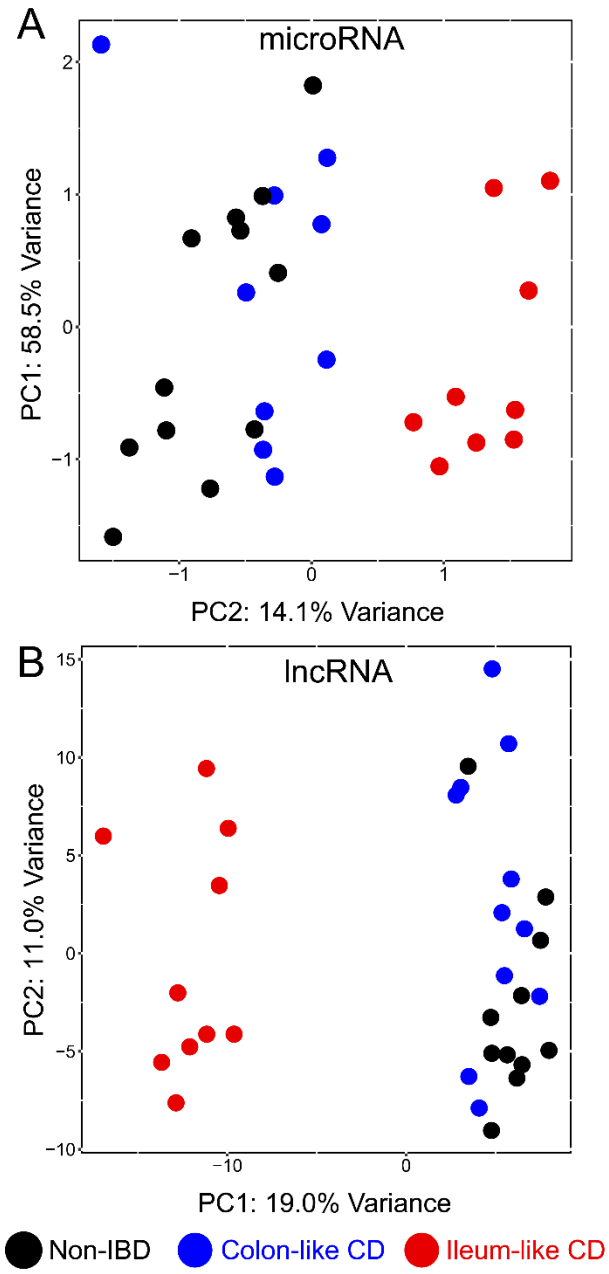


Figure 2.1 Two distinct molecular subtypes across multiple data types in adult Crohn's disease (CD). Principal components analysis (PCA) of microRNA (A) and long non-coding RNA (B) expression profiles for patients with CD and patients with NIBD exhibit (black, n=11-12) distinct clusters; one enriched for colon-like CD patients (blue, n=9-11), and another enriched for ileum-like CD patients (red, n=9-10).

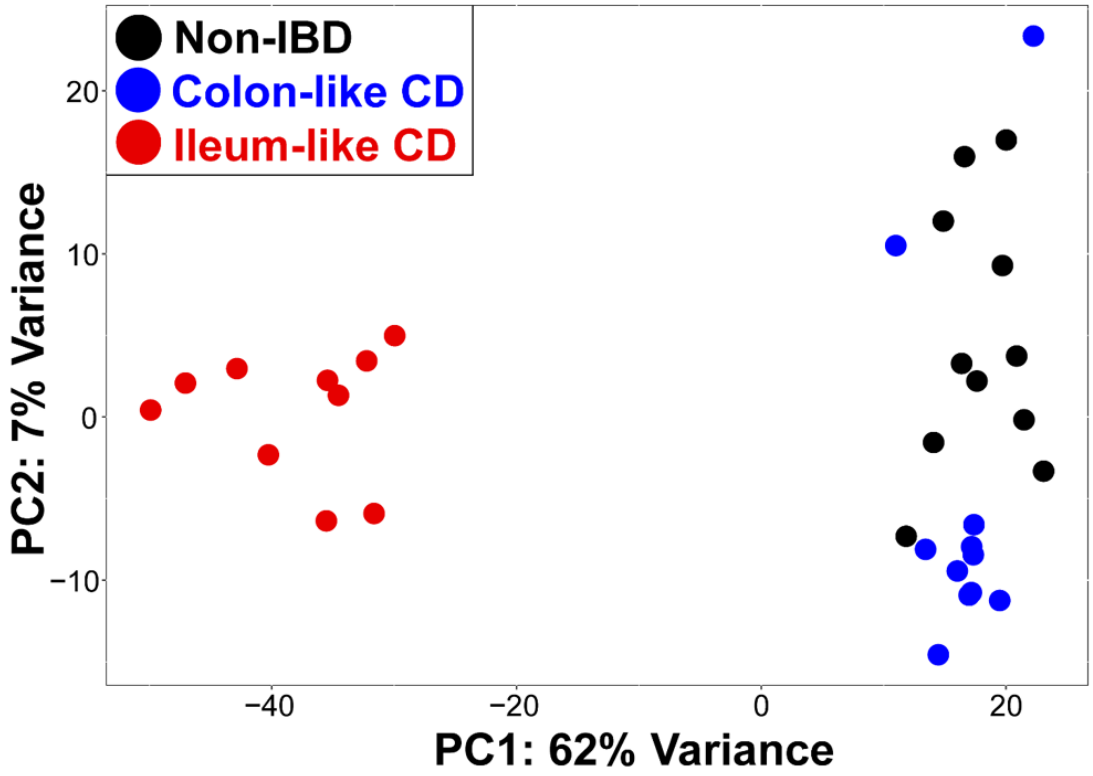


Figure 2.2 mRNA expression profiles segregate Crohn's disease (CD) samples into two distinct molecular subtypes. Using updated gene annotations from GENCODE, we recapitulated our previous analysis (82) to show that clustering of colon-like CD patients (blue, n=11) with non-IBD patients (black, n=11) is distinct from ileum-like CD patients (red, n=10).

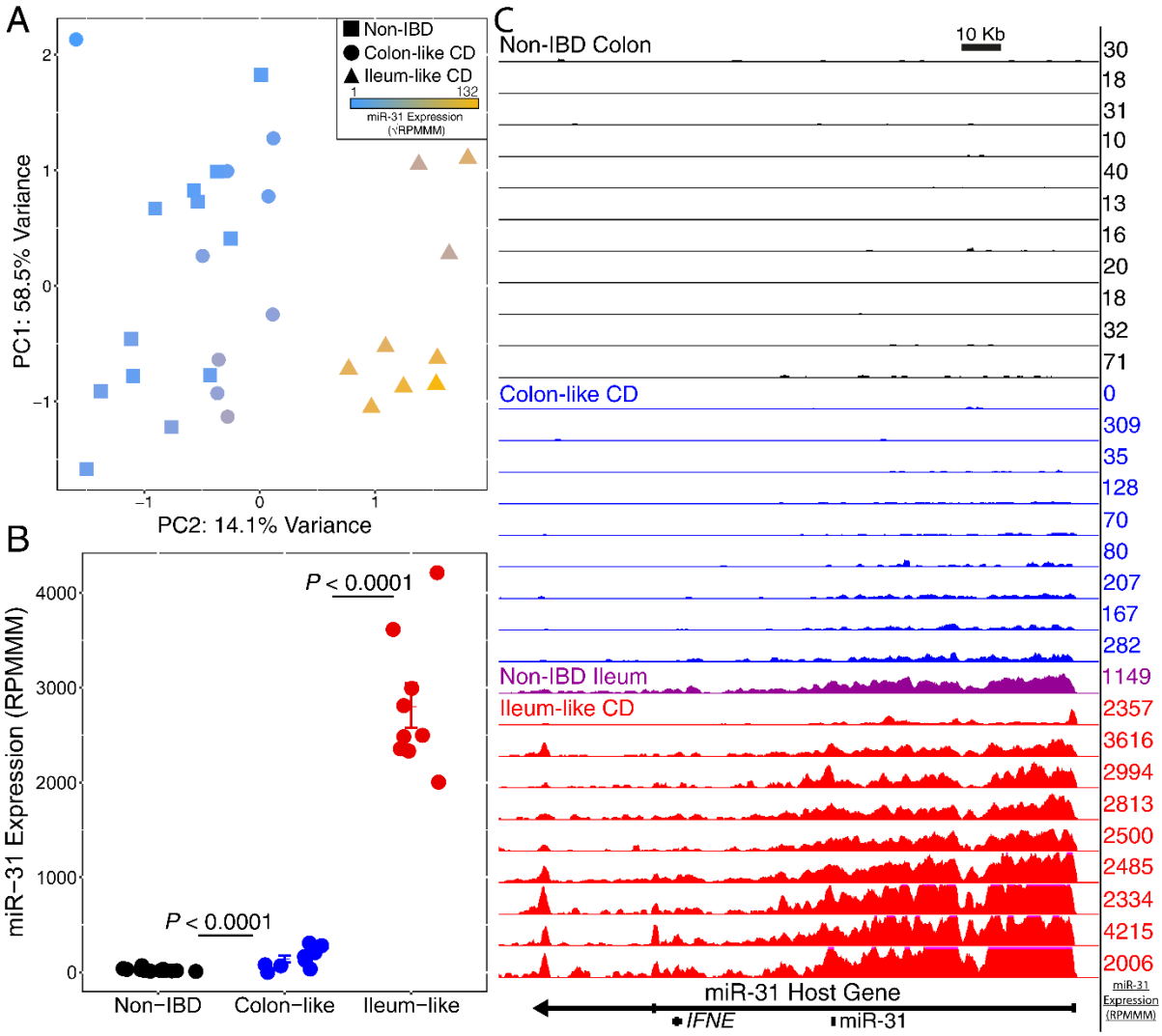


Figure 2.3 miR-31 is a driver of colon-like and ileum-like stratification. (A) Principal components analysis of miRNA expression data from small RNA-seq. MiR-31 expression (blue-gold, low-high) appears to distinguish the NIBD (n=12) and colon-like samples (n=9) from ileum-like samples (n=9). (B) Normalized miR-31 expression exhibits a significant upregulation in Crohn's disease sub-groups (colon-like, n=9; ileum-like, n=9) samples compared with NIBD samples (n=12). (C) UCSC browser representation of normalized RNA-seq reads mapping to the miR-31 host gene for NIBD colon (black, n=11), colon-like CD (blue, n=9), NIBD Ileum (purple, n=1) and ileum-like CD (red, n=9). miR-31 transcript expression levels from small RNA-seq (RPMMM) are displayed to the right of each track. FDR adjusted p-values determined using DESeq2, with data presented as mean RPMMM ± SE.

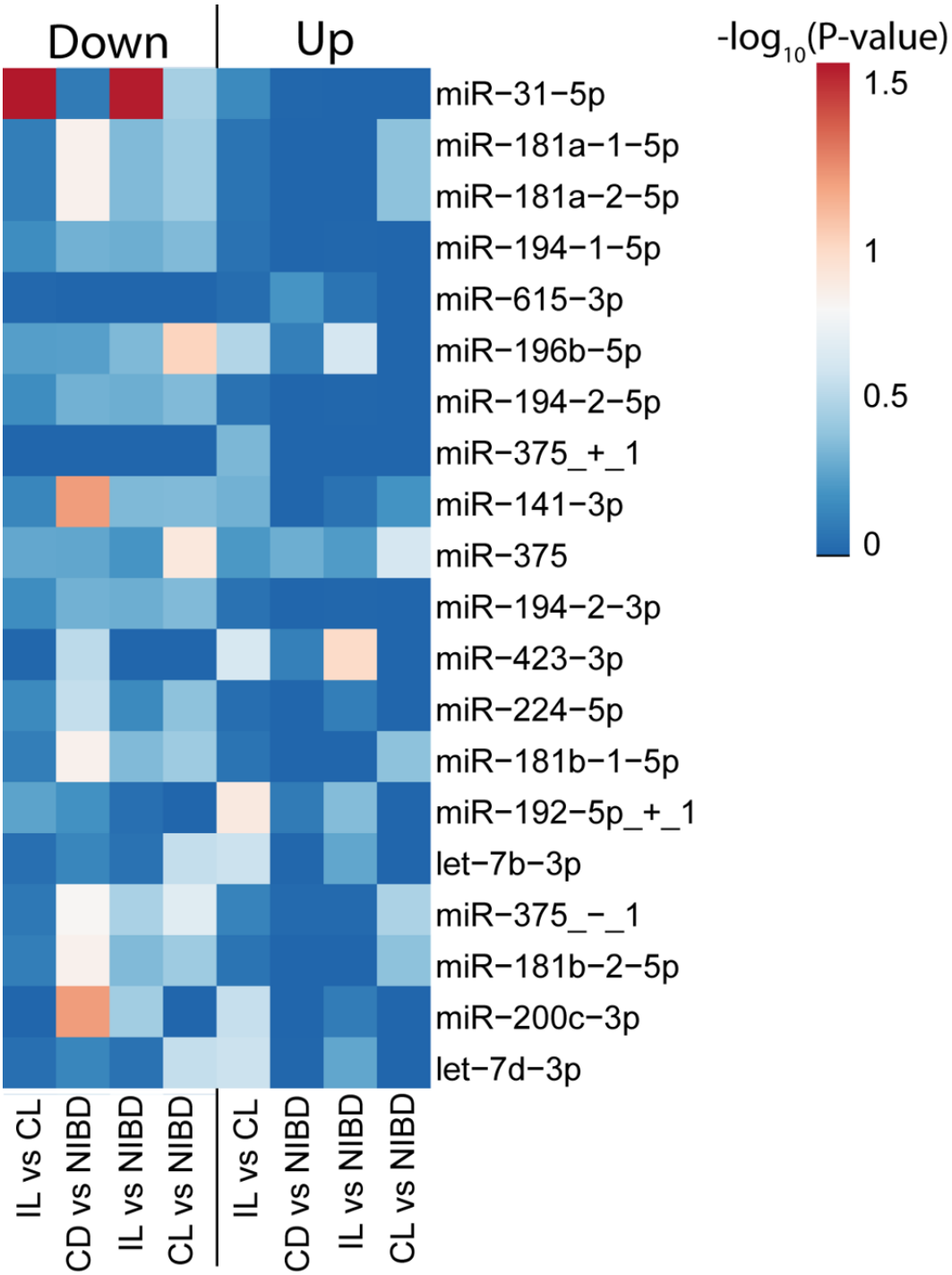


Figure 2.4 microRNA-31 is potential master regulator of pathways disrupted in CD pathogenesis. Using genes that were differentially expressed between IL-CD and CL-CD samples, CD and NIBD samples, IL-CD and NIBD samples, and CL-CD and NIBD samples, we used miRHub to test whether the top differentially expressed microRNAs significantly targeted differently expressed genes within four different conditions. miR-31 significantly targets genes that are downregulated in IL-CD.

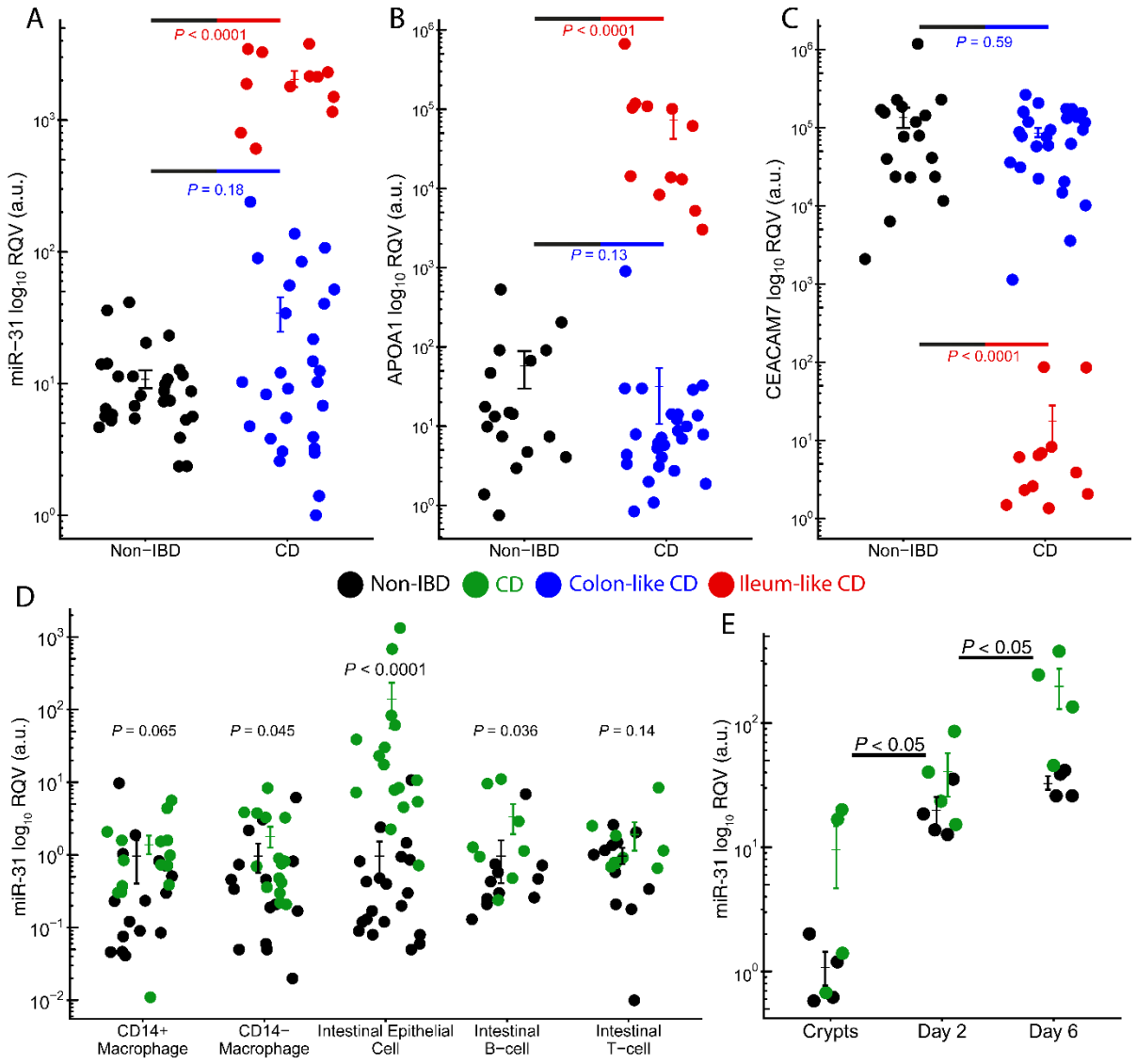


Figure 2.5 miR-31 is specifically upregulated in intestinal epithelial cells. qRT-PCR for miR-31 (A), *APOA1* (B), and *CEACAM7* (B) in an independent adult cohort displays colon-like (blue; n=27-28) and ileum-like (red; n=12) clustering patterns for CD samples compared with NIBD samples (black; n=18-29). (D) qRT-PCR of five colon-specific cell types reveal significant miR-31 upregulation in intestinal epithelial cells isolated from CD patients (n=11-20) relative to NIBD controls (n=8-16). (8 NIBD matched and 6 CD matched across all cell types). (E) Relative miR-31 expression by qPCR of in colonoid cultures generated from NIBD controls (n=4) compared with CD patients (n=4). miR-31 expression is increased in fresh crypts and remains higher at day 2 and day 6 of colonoid culture. miRNA levels are relative to RNU-48 expression compared to fresh NIBD crypts. Significance values determined by a 2-tailed unpaired Student's t test. Data are presented as mean ± SE.

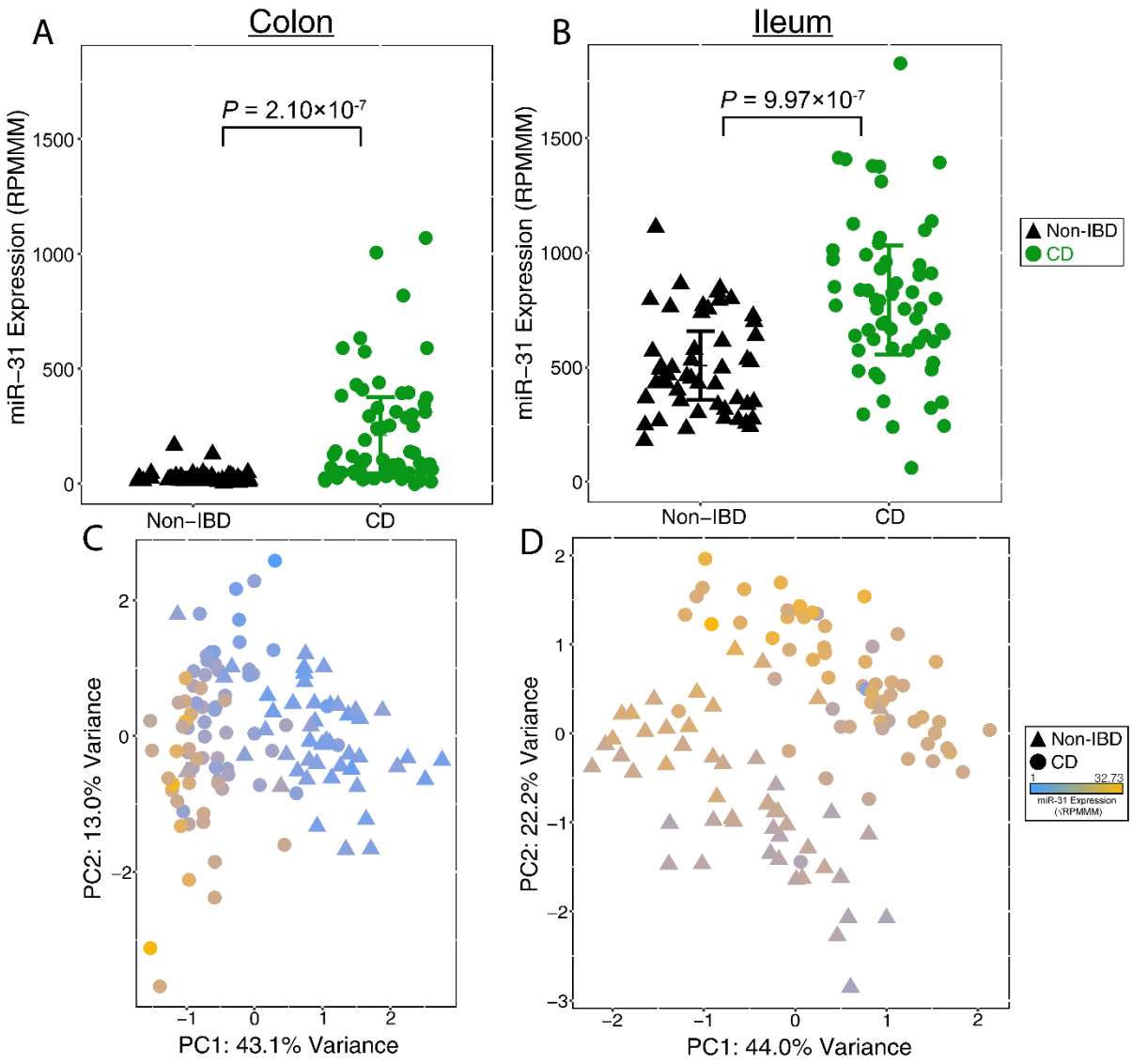


Figure 2.6 miR-31 is differentially expressed in treatment-naïve pediatric Crohn's disease (CD) samples. miR-31 expression is significantly upregulated in the colon (A) and ileum (B) of treatment-naïve pediatric CD samples (colon, n=76; ileum, n=60) compared with pediatric NIBD samples (colon, n=48; ileum, n=50). Principal components analysis of miRNA expression profiles from small RNA-seq results in distinct clusters of NIBD and CD patients for colon (C) and ileum (D) samples. Points are colored according to miR-31 expression (blue-gold; low-high). Significance determined by a 2-tailed unpaired Student's t test where $P < 0.05$. Data presented as mean ± SE

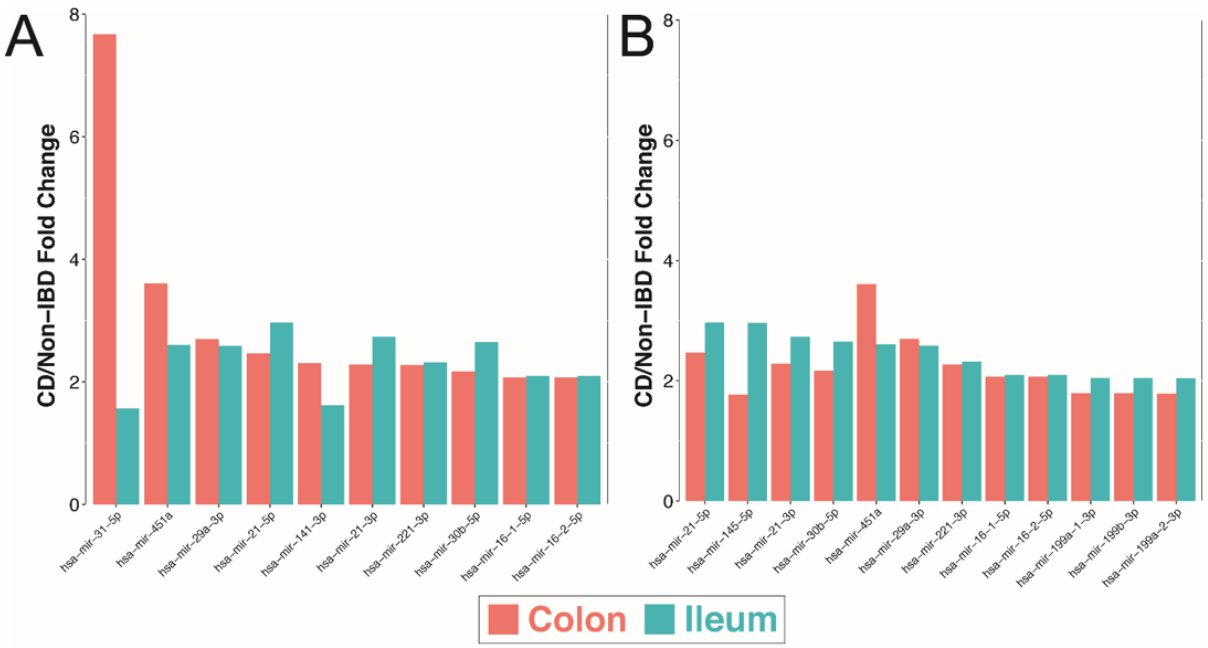


Figure 2.7 miR-31 is uniquely upregulated in the colon of pediatric CD patients relative to NIBD patients. The average expression of the 82 microRNAs used for principal component analysis were compared between CD and NIBD patients for colon samples (CD, n=76; NIBD=48) and ileum samples (CD, n=60; NIBD, n=50). microRNAs with a fold change greater than 2 when comparing CD expression with non-IBD expression are shown above for colonic microRNA expression (A), and ileal microRNA expression (B).

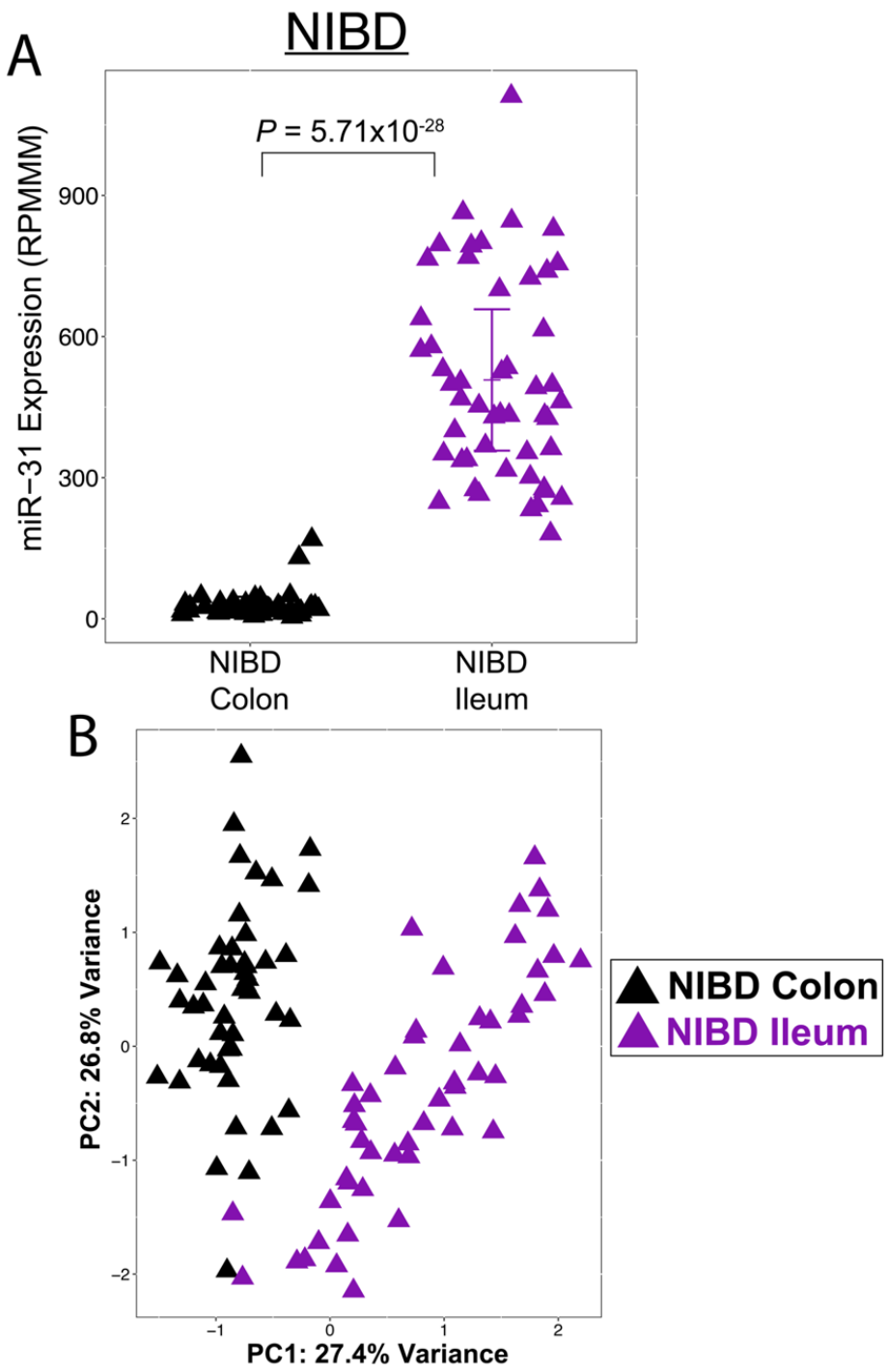


Figure 2.8 Pediatric NIBD microRNA expression profiles are significantly different between colon and ileum tissue. (A) miR-31 is significantly differentially expressed between NIBD colon samples (n=48) and NIBD ileum samples (n=50). (B) More broadly, PCA using microRNA expression profiles reveals that PC1 splits NIBD colon and ileum samples, with miR-31 being the highest contributor to the variance explained along this axis. Data is mean RPMMM ± SEM with significance determined by 2-tailed unpaired Student's test

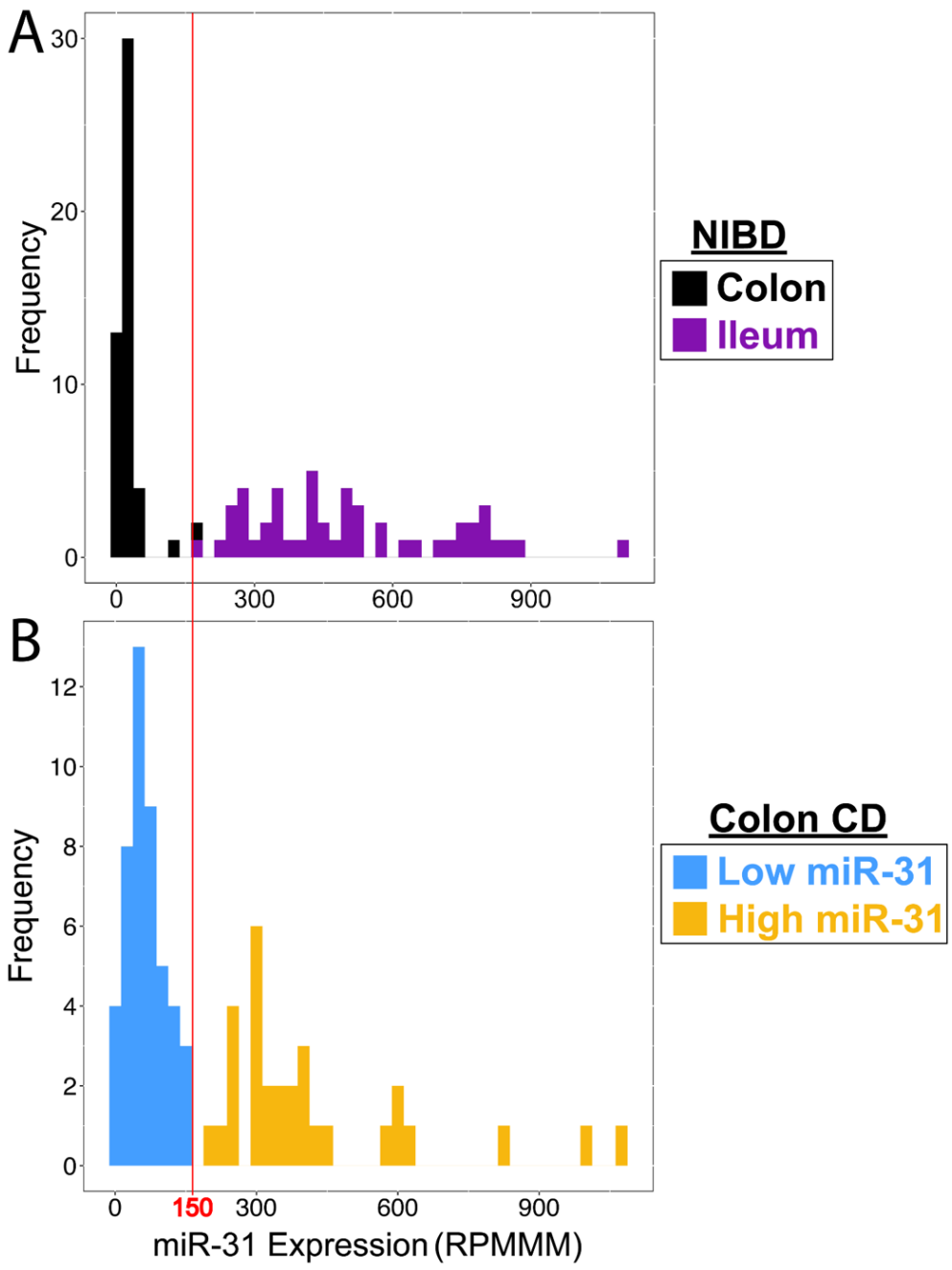
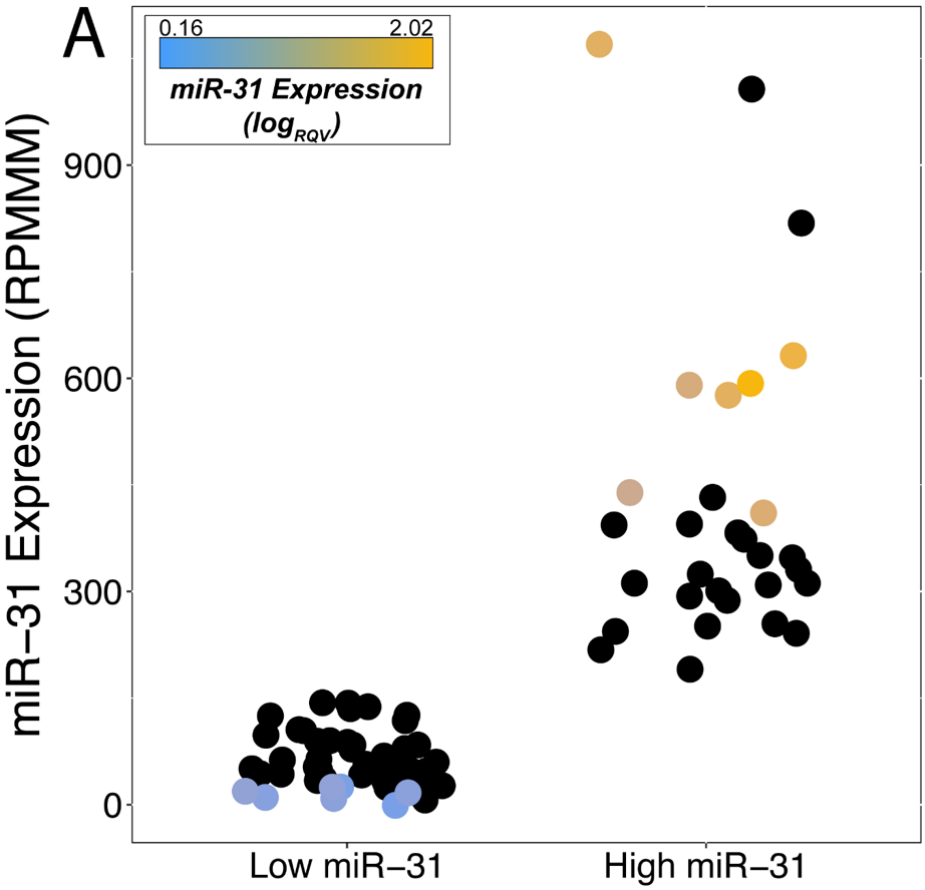


Figure 2.9 Stratification of pediatric colon CD samples guided by NIBD miR-31 expression. Normalized miR-31 expression of 150 reads per million mapped to miRs (RPMMM) can distinguish pediatric NIBD colon (n=48) and ileum samples (n=50). (A) Pediatric miR-31 expression is lower in the colon compared with miR-31 expression the ileum. The lowest ileal miR-31 expression that we observed in the ileum was 181 RPMMM. (B) Using a threshold of 150 RPMMM, we segregate pediatric CD samples into a low miR-31 group (<150 RPMMM; n=46), and a high miR-31 group (≥150 RPMMM; n=30).



B

Sample	Subgroup	miR-31 smRNA-seq	miR-31 qPCR Ct	miR-31 qPCR RQV
PB305	Low	0	37.78	0.43
PB223	Low	10.48	35.54	2.38
PB109	Low	11.92	33.39	1.70
PB102	Low	17.98	32.31	2.20
PB205	Low	19.65	38.95	3.59
PB227	Low	23.6	33.76	3.11
PB140	Low	25.37	32.77	0.56
PB107	High	411.6	28.68	47.32
PB103	High	439.45	28.90	30.29
PB126	High	575.55	29.05	59.96
PB115	High	590.29	29.36	41.27
PB213	High	592.02	27.47	103.03
PB132	High	632.62	27.03	78.32
PB209	High	1070.25	25.88	58.07

Figure 2.10 RT-qPCR confirmation of miR-31 expression in FFPE pediatric colonic mucosal samples. (A) Pediatric miR-31 expression of FFPE colon samples according to small RNA-sequencing is confirmed through RT-qPCR of a subset of miR-31 low (n = 7) and miR-31 high (n = 7) samples from each group ($r = 0.94$, $p = 3.89 \times 10^{-7}$). Points are colored according to RQV obtained through RT-qPCR of the same samples (blue-gold, low-high). (B) Table of expression values of small RNA-sequencing and RT-qPCR matched samples. Significance determined through a test for association between paired samples using Pearson's product moment correlation coefficient.

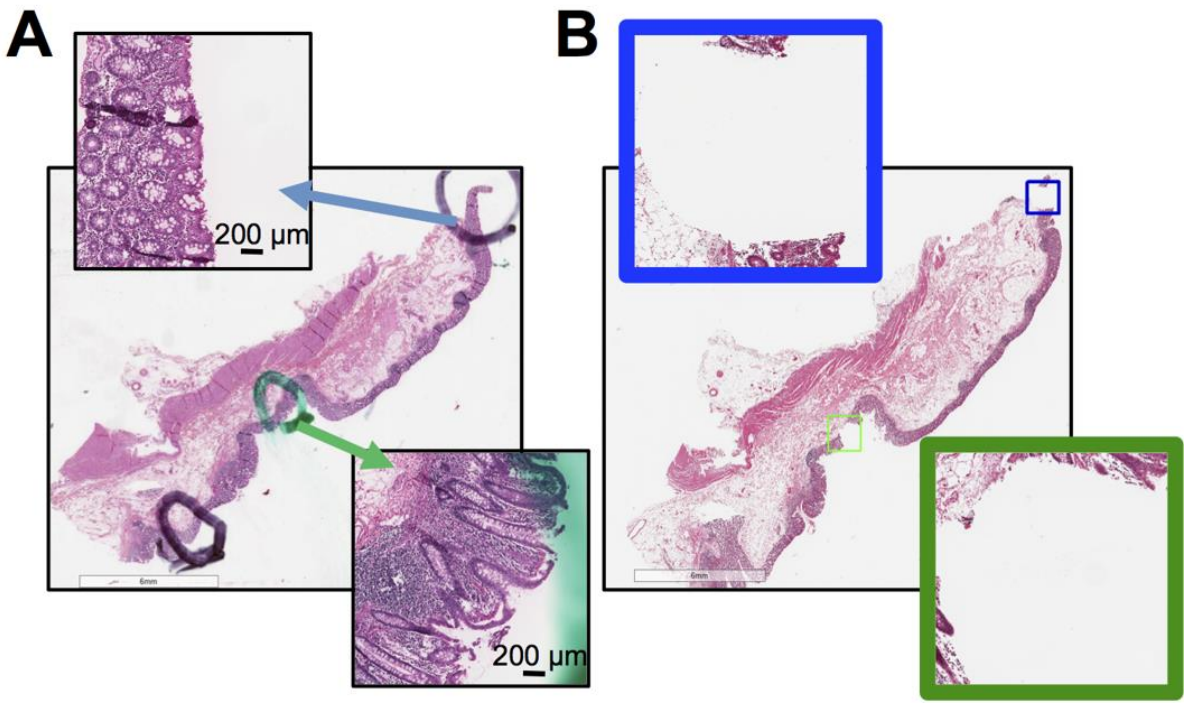


Figure 2.11 Hematoxylin and eosin (H&E) staining of isolated tissue from FFPE histological sections. H&E staining for an FFPE sample from a pediatric Crohn's disease patient, and mucosal region selected for miRNA isolation and qRT-PCR. A) Section before and B) after selection of the mucosal region. Circles indicate the selected area selected for RNA extraction for small RNA-sequencing and qRT-PCR.

microRNAs	PC1	PC2	PC3
miR-31-5p	-0.203	0.648	0.023
miR-215	-0.236	0.301	-0.027
miR-194-1-5p	-0.143	0.185	0.068
miR-194-2-5p	-0.141	0.117	0.067
miR-146b-5p	-0.093	0.102	-0.252
miR-141-3p	-0.127	0.092	0.105
miR-192-5p_+_1	-0.149	0.086	0.036
miR-192-5p	-0.151	0.067	-0.011
miR-181a-2-5p	-0.050	0.059	0.037
miR-181a-1-5p	-0.050	0.059	0.037
miR-200c-3p	-0.085	0.039	0.114
miR-197-3p	-0.052	0.035	0.114
miR-574-3p	-0.051	0.034	0.110
miR-22-3p	-0.097	0.027	-0.034
miR-146a-5p	-0.081	0.024	-0.264
miR-21-3p	-0.109	0.023	-0.078
miR-484	-0.102	0.023	0.126
miR-342-3p	-0.138	0.015	0.015
miR-142-5p_-_2	-0.209	0.012	-0.207
miR-221-3p	-0.092	0.011	0.006
miR-200a-3p	-0.122	0.007	-0.007
miR-222-3p	-0.085	0.007	0.077
miR-423-3p	-0.016	0.005	0.129
miR-181b-1-5p	-0.046	0.004	0.039
miR-181b-2-5p	-0.046	0.004	0.036
miR-486-5p	-0.056	-0.005	0.071
miR-21-5p	-0.159	-0.006	-0.219
miR-375	-0.058	-0.007	0.186
miR-25-3p	-0.112	-0.008	-0.001
miR-92a-1-3p	-0.051	-0.009	0.155
miR-191-5p	-0.084	-0.011	-0.008
miR-99b-5p	-0.126	-0.011	0.139
miR-92a-2-3p	-0.051	-0.012	0.150
miR-423-5p	-0.018	-0.012	0.115
miR-24-1-3p	-0.083	-0.016	-0.100
miR-24-2-3p	-0.083	-0.016	-0.101
miR-93-5p	-0.076	-0.016	0.058
miR-451a	-0.145	-0.019	-0.200
let-7d-3p	-0.035	-0.019	0.174
miR-140-3p	-0.105	-0.019	0.029
miR-27a-3p	-0.104	-0.020	-0.019
miR-182-5p	-0.133	-0.021	-0.088

miR-191-5p_+_1	-0.082	-0.024	0.012
miR-200b-3p_+_1	-0.098	-0.024	0.088
miR-30b-5p	-0.101	-0.024	-0.027
miR-151a-3p	-0.103	-0.026	0.025
miR-30d-5p	-0.058	-0.030	0.032
miR-148a-3p	-0.158	-0.031	-0.108
miR-29a-3p	-0.109	-0.032	-0.024
miR-126-3p	-0.135	-0.032	-0.106
miR-125a-5p	-0.084	-0.035	0.141
miR-16-1-5p	-0.145	-0.036	-0.055
miR-16-2-5p	-0.145	-0.036	-0.055
miR-127-3p	-0.071	-0.039	0.104
miR-26a-1-5p	-0.085	-0.041	0.031
miR-26a-2-5p	-0.085	-0.041	0.031
miR-140-3p_+_1	-0.077	-0.044	0.047
miR-26b-5p	-0.100	-0.045	-0.003
miR-103a-2-3p	-0.087	-0.046	0.013
miR-103a-1-3p	-0.087	-0.047	0.013
let-7i-5p	-0.071	-0.049	-0.049
miR-200b-3p	-0.121	-0.049	0.090
miR-150-5p	-0.116	-0.049	0.138
miR-30a-5p	-0.085	-0.049	-0.026
miR-320a	-0.016	-0.050	0.087
miR-92b-3p	-0.041	-0.052	0.146
miR-27b-3p	-0.083	-0.053	-0.022
miR-28-3p	-0.056	-0.054	0.096
miR-30c-2-5p	-0.098	-0.054	0.010
miR-30c-1-5p	-0.098	-0.054	0.010
let-7g-5p	-0.075	-0.056	0.008
let-7b-5p	-0.022	-0.056	0.126
miR-155-5p	-0.099	-0.057	-0.065
miR-186-5p	-0.101	-0.059	-0.060
miR-30e-5p	-0.106	-0.059	-0.059
miR-100-5p	-0.160	-0.061	0.116
miR-125b-1-5p	-0.096	-0.075	0.157
miR-125b-2-5p	-0.096	-0.075	0.157
miR-23b-3p	-0.064	-0.079	0.055
miR-23a-3p	-0.089	-0.081	-0.016
miR-497-5p	-0.091	-0.082	0.120
let-7a-3-5p	-0.071	-0.083	0.077
let-7a-1-5p	-0.071	-0.083	0.077
let-7a-2-5p	-0.071	-0.083	0.077
miR-10a-5p_+_1	-0.087	-0.083	-0.112

let-7f-2-5p	-0.102	-0.083	-0.046
let-7f-1-5p	-0.103	-0.084	-0.046
miR-199a-1-3p	-0.099	-0.085	-0.001
miR-199b-3p	-0.099	-0.085	-0.001
miR-199a-2-3p	-0.098	-0.086	0.000
miR-378a-3p	-0.045	-0.087	0.090
miR-10a-5p	-0.096	-0.088	-0.120
let-7e-5p	-0.036	-0.089	0.122
miR-145-5p	-0.054	-0.092	0.103
let-7d-5p	-0.059	-0.094	0.066
miR-143-3p	-0.087	-0.108	-0.032
let-7c	-0.029	-0.117	0.155
miR-10b-5p_+_1	-0.087	-0.128	-0.166
miR-10b-5p	-0.098	-0.132	-0.189
miR-196b-5p	0.007	-0.359	-0.126

Table 2.1 Adult miRNA PCA loadings. First three loadings from PCA analysis of miRNA data from all NIBD and CD adult samples (Figure 2.1A). These cumulatively explain 82.0% of total variation.

Phenotype	Colon-like (n=39)	Ileum-like (n=22)	*P* Value
Patient Characteristics			
Age at Sample Collection (years)	37.64	34.73	0.456
Male	18	10	1.000
Female	21	12	1.000
Smoker (current or previous)	15	7	0.782
Location			
Ileum-only	8	4	1.000
Colon-only	11	3	0.225
Ileum+Colon	20	15	0.282
Upper GI	5	4	0.710
Phenotypes and Involvement			
Perianal	17	5	0.165
Ileal Disease	28	19	0.225
Inflammatory	7	0	**0.000**
Stricturing	17	15	0.108
Penetrating	4	4	0.443
Disease Duration (years)	11.79	10.50	0.702
Pre-operative treatment history			
Steroids	22	12	1.000
5-ASA	11	11	0.104
Immunomodulation	11	12	0.056
Anti-TNF	22	13	1.000
Non-anti-TNF biologic	5	1	0.404

Table 2.2 Demographics of adult CD patients. CD patients were classified into colon-like and ileum-like CD molecular subtypes. Phenotype associations with subtypes were assessed using Fisher's exact test (categorical data) and 2-tailed unpaired Student's t test (continuous data). $P < 0.05$ (bolded) was considered statistically significant.

Phenotype	Colon-like (n=39)	Ileum-like (n=22)	*P* Value
Surgery	33	22	0.079
Biologic Use	=21/33	=16/22	0.566
End ileostomy	=11/33	=1/22	**0.006**
Second resection	=16/33	=8/22	0.418
Time to first resection (years)	6.91	8.82	0.349
Time from first to second resection (years)	9.4 (=15/33)	5 (=8/22)	0.292
Post-op Colonoscopy	=32/33	=17/22	NA
Remission (i0,i1)	=9/32	=10/17	**0.030**
Recurrence (i2,i3,i4) or end ileostomy	=23/32	=7/17	**0.030**

Table 2.3 Post-operative clinical characteristics of adult CD patients. Associations between CD molecular subtypes with post-operative phenotypes were assessed using Fisher's exact test (categorical data) and 2-tailed unpaired Student's t test (continuous data). $P < 0.05$ (bolded) was considered a statistically significant association. Recurrence was defined as having a Rutgeerts score of i2, i3, i4 or the need for an end ileostomy within a year after the initial surgery. Remission was defined as a Rutgeerts score if i0, i1.

microRNAs	PC1	PC2	PC3
hsa-mir-31-5p	0.3016	-0.3615	0.1582
hsa-mir-100-5p	0.0237	-0.2002	0.0119
hsa-mir-125b-2-5p	0.0712	-0.1962	-0.0386
hsa-mir-125b-1-5p	0.0716	-0.1958	-0.0400
hsa-mir-146b-5p	-0.0992	-0.1708	0.1077
hsa-mir-155-5p	0.1276	-0.1466	-0.0824
hsa-mir-142-5p_-_2	0.1203	-0.1457	-0.0524
hsa-mir-486-5p	-0.0415	-0.1448	-0.0408
hsa-mir-146a-5p	-0.0294	-0.1326	0.1265
hsa-let-7i-5p	0.1052	-0.1247	-0.0669
hsa-mir-127-3p	-0.0223	-0.1242	-0.0431
hsa-mir-342-3p	0.1622	-0.1185	-0.0053
hsa-mir-99b-5p	-0.0216	-0.1164	-0.0066
hsa-mir-143-3p	-0.0331	-0.1084	0.0236
hsa-let-7e-5p	-0.1734	-0.1028	-0.2720
hsa-mir-222-3p	-0.0621	-0.0972	-0.2130
hsa-mir-21-3p	0.1425	-0.0960	0.0068
hsa-mir-150-5p	0.1715	-0.0921	-0.2049
hsa-mir-125a-5p	-0.0363	-0.0809	-0.0506
hsa-mir-126-3p	0.1165	-0.0808	-0.0938
hsa-mir-221-3p	0.1655	-0.0742	-0.0092
hsa-mir-92b-3p	-0.0987	-0.0728	-0.1797
hsa-mir-181a-1-5p	0.0303	-0.0588	-0.1138
hsa-mir-181a-2-5p	0.0303	-0.0588	-0.1138
hsa-mir-30a-5p	0.0352	-0.0548	-0.0246
hsa-mir-21-5p	0.1848	-0.0536	-0.0131
hsa-mir-29a-3p	0.2077	-0.0513	-0.0347
hsa-mir-27a-3p	0.0881	-0.0511	-0.0833
hsa-mir-199a-1-3p	0.1187	-0.0494	-0.0877
hsa-mir-199b-3p	0.1187	-0.0494	-0.0877
hsa-mir-199a-2-3p	0.1186	-0.0485	-0.0884
hsa-mir-10b-5p_+_1	-0.0771	-0.0432	0.0673
hsa-mir-148a-3p	0.0478	-0.0423	-0.0463
hsa-mir-27b-3p	0.0523	-0.0372	-0.0259
hsa-let-7f-1-5p	-0.1088	-0.0345	-0.1806
hsa-mir-25-3p	0.0660	-0.0344	-0.0581
hsa-mir-10b-5p	-0.0877	-0.0327	0.0574
hsa-let-7f-2-5p	-0.1090	-0.0317	-0.1810
hsa-mir-451a	0.2300	-0.0310	0.0150
hsa-let-7a-3-5p	-0.0843	-0.0253	-0.1997
hsa-let-7a-1-5p	-0.0845	-0.0252	-0.1999
hsa-let-7a-2-5p	-0.0850	-0.0251	-0.2001

hsa-let-7g-5p	0.0676	-0.0202	-0.0738
hsa-mir-423-5p	-0.1294	-0.0196	-0.2563
hsa-mir-26b-5p	0.0949	-0.0106	-0.0700
hsa-let-7b-5p	-0.0637	-0.0106	-0.1841
hsa-mir-145-5p	0.1503	-0.0101	0.0681
hsa-mir-10a-5p_+_1	-0.0227	-0.0015	0.1161
hsa-mir-320a	0.0377	0.0006	-0.1335
hsa-mir-10a-5p	-0.0372	0.0029	0.1007
hsa-mir-22-3p	0.1272	0.0042	-0.0391
hsa-mir-92a-2-3p	-0.0545	0.0089	-0.2035
hsa-mir-103a-2-3p	0.1191	0.0093	-0.0782
hsa-mir-103a-1-3p	0.1189	0.0093	-0.0782
hsa-mir-16-2-5p	0.1693	0.0095	-0.0490
hsa-mir-16-1-5p	0.1693	0.0096	-0.0488
hsa-mir-92a-1-3p	-0.0549	0.0127	-0.2052
hsa-mir-26a-1-5p	0.0828	0.0154	-0.0461
hsa-mir-26a-2-5p	0.0829	0.0155	-0.0462
hsa-mir-191-5p	0.0051	0.0202	0.0492
hsa-mir-182-5p	-0.0018	0.0206	0.0891
hsa-mir-423-3p	-0.0158	0.0226	-0.1806
hsa-mir-186-5p	0.0985	0.0295	-0.0190
hsa-mir-28-3p	-0.0529	0.0339	-0.0902
hsa-mir-30e-5p	0.1083	0.0434	-0.0485
hsa-mir-30d-5p	0.0344	0.0459	-0.0909
hsa-mir-30c-1-5p	0.1170	0.0901	-0.1760
hsa-mir-30c-2-5p	0.1170	0.0901	-0.1761
hsa-mir-196b-5p	0.0146	0.1053	-0.0554
hsa-mir-30b-5p	0.2464	0.1173	-0.0940
hsa-mir-141-3p	0.1438	0.1357	0.1066
hsa-mir-192-5p	-0.0248	0.1501	0.0364
hsa-mir-200b-3p	0.0357	0.1530	-0.1058
hsa-mir-378a-3p	0.0432	0.1701	-0.0626
hsa-mir-200a-3p	0.1710	0.1742	-0.0164
hsa-mir-215	0.0307	0.1810	0.0630
hsa-mir-375	-0.0738	0.1829	-0.1513
hsa-mir-200b-3p_+_1	0.0626	0.1841	-0.0862
hsa-mir-192-5p_+_1	0.0020	0.1852	0.0257
hsa-mir-200c-3p	0.1313	0.2138	-0.0205
hsa-mir-194-1-5p	0.1438	0.2363	-0.0050
hsa-mir-194-2-5p	0.1755	0.2492	-0.0400

Table 2.4 Pediatric miRNA colon tissue PCA loadings. First three microRNA PCA loadings from pediatric CD and non-IBD colon samples (Figure 2.6C)**.** These cumulatively explain 67.7% of total variation.

microRNAs	PC1	PC2	PC3
hsa-mir-150-5p	0.2897	-0.0660	0.1074
hsa-mir-142-5p_-_2	0.2386	-0.1092	0.0487
hsa-mir-342-3p	0.2326	0.0033	-0.0736
hsa-mir-155-5p	0.2289	-0.1013	0.0449
hsa-mir-29a-3p	0.2156	0.0630	0.0705
hsa-mir-21-5p	0.1879	0.1005	0.0547
hsa-mir-125b-1-5p	0.1830	-0.0837	-0.1763
hsa-mir-125b-2-5p	0.1828	-0.0839	-0.1759
hsa-mir-21-3p	0.1799	0.0568	0.0225
hsa-mir-100-5p	0.1681	-0.1133	-0.3045
hsa-let-7i-5p	0.1676	-0.0216	0.0737
hsa-mir-16-1-5p	0.1601	0.0637	0.0922
hsa-mir-16-2-5p	0.1601	0.0635	0.0922
hsa-mir-221-3p	0.1598	0.0792	0.0451
hsa-mir-145-5p	0.1566	0.1991	-0.1190
hsa-mir-199a-1-3p	0.1472	0.0326	0.1227
hsa-mir-199b-3p	0.1472	0.0326	0.1227
hsa-mir-199a-2-3p	0.1467	0.0323	0.1231
hsa-mir-30b-5p	0.1405	0.2144	0.0958
hsa-mir-126-3p	0.1332	0.0306	0.0618
hsa-mir-146a-5p	0.1278	-0.1993	-0.1700
hsa-mir-186-5p	0.1202	0.0498	0.0266
hsa-mir-99b-5p	0.1139	-0.0556	-0.2917
hsa-mir-27a-3p	0.1049	0.0234	0.0659
hsa-mir-451a	0.1037	0.1861	0.0732
hsa-mir-103a-2-3p	0.1003	0.0628	0.0916
hsa-mir-103a-1-3p	0.1001	0.0625	0.0918
hsa-mir-26a-1-5p	0.0959	0.0116	0.0327
hsa-mir-26a-2-5p	0.0959	0.0117	0.0327
hsa-mir-30e-5p	0.0956	0.0523	0.0645
hsa-let-7g-5p	0.0952	-0.0257	0.0725
hsa-mir-181a-1-5p	0.0898	-0.0672	0.0562
hsa-mir-181a-2-5p	0.0898	-0.0672	0.0562
hsa-mir-30c-2-5p	0.0878	0.0680	0.1255
hsa-mir-30c-1-5p	0.0878	0.0680	0.1255
hsa-mir-125a-5p	0.0868	-0.0755	-0.1776
hsa-mir-25-3p	0.0847	-0.0213	0.0452
hsa-mir-26b-5p	0.0814	0.0222	0.0361
hsa-mir-30a-5p	0.0781	-0.0101	-0.0191
hsa-mir-27b-3p	0.0746	0.0151	-0.0044
hsa-mir-191-5p	0.0726	0.0031	-0.0980
hsa-mir-127-3p	0.0664	-0.0422	-0.0624

hsa-mir-148a-3p	0.0599	0.0269	-0.0053
hsa-mir-146b-5p	0.0585	-0.1648	-0.1995
hsa-mir-320a	0.0559	0.0174	0.1025
hsa-mir-143-3p	0.0528	-0.0375	-0.0468
hsa-mir-22-3p	0.0482	0.0839	0.1250
hsa-mir-30d-5p	0.0441	0.0046	0.0433
hsa-mir-10a-5p_+_1	0.0423	0.0108	-0.1020
hsa-mir-10b-5p_+_1	0.0392	-0.0723	-0.1094
hsa-mir-92b-3p	0.0294	-0.1995	0.0646
hsa-mir-10b-5p	0.0283	-0.0775	-0.1211
hsa-mir-10a-5p	0.0274	-0.0009	-0.1294
hsa-mir-423-3p	0.0180	-0.0792	0.2023
hsa-mir-92a-2-3p	0.0134	-0.1304	0.0867
hsa-mir-486-5p	0.0121	-0.1203	0.0152
hsa-mir-182-5p	0.0108	0.0237	-0.0893
hsa-mir-92a-1-3p	0.0102	-0.1283	0.0895
hsa-mir-222-3p	-0.0015	-0.1338	0.1592
hsa-mir-28-3p	-0.0038	-0.0762	0.0432
hsa-mir-31-5p	-0.0112	0.1615	0.1179
hsa-mir-196b-5p	-0.0115	0.0509	0.0131
hsa-mir-200a-3p	-0.0189	0.1887	0.0416
hsa-mir-200b-3p	-0.0189	0.1336	-0.0432
hsa-mir-378a-3p	-0.0222	0.1172	0.0767
hsa-mir-200c-3p	-0.0224	0.1892	0.0622
hsa-let-7b-5p	-0.0251	-0.1169	0.1303
hsa-mir-141-3p	-0.0251	0.1620	-0.0245
hsa-let-7a-3-5p	-0.0359	-0.1530	0.1507
hsa-let-7a-1-5p	-0.0361	-0.1532	0.1509
hsa-let-7a-2-5p	-0.0366	-0.1537	0.1512
hsa-let-7e-5p	-0.0398	-0.2765	0.2285
hsa-let-7f-1-5p	-0.0400	-0.1820	0.1722
hsa-mir-423-5p	-0.0401	-0.2089	0.1330
hsa-mir-200b-3p_+_1	-0.0406	0.1462	0.0596
hsa-let-7f-2-5p	-0.0420	-0.1808	0.1712
hsa-mir-194-2-5p	-0.0444	0.2281	0.0809
hsa-mir-194-1-5p	-0.0905	0.1734	0.0266
hsa-mir-375	-0.1101	0.0053	0.0639
hsa-mir-192-5p	-0.1234	0.0433	-0.0345
hsa-mir-192-5p_+_1	-0.1255	0.0580	-0.0142
hsa-mir-215	-0.1509	0.0527	-0.0294

Table 2.5 Pediatric miRNA ileum tissue PCA loadings. First three microRNA PCA loadings from pediatric CD and non-IBD ileum samples (Figure 2.6D)**.** These cumulatively explain 77.3% of total variation.

Phenotype	Low miR-31 (n=46)	High miR-31 (n=30)	*P* Value
Patient Characteristics			
Age at Diagnosis	12.1	11.6	0.524
Male	30	21	0.804
Female	16	9	0.804
INITIAL PHENOTYPES			
Location			
Ileum-only	10	1	**0.042**
Colon-only	6	8	0.225
Ileum + Colon	30	21	0.804
Upper GI*	19	7	0.134
Phenotypes and Involvement			
Perianal	15	8	0.620
Ileal Disease	40	22	0.225
Inflammatory	35	20	0.436
Stricturing	4	2	1.000
Penetrating	1	0	1.000
SUBSEQUENT PHENOTYPES			
Location			
Ileum-only	8	1	0.078
Colon-only	4	7	0.100
Ileum + Colon	34	22	1.000
Upper GI*	26	13	0.348
Phenotypes and Involvement			
Perianal	16	12	0.808
Ileal Disease	42	23	0.100
Inflammatory	24	19	0.356
Stricturing	16	2	**0.005**
Penetrating	2	2	0.645
Disease Duration (years)	6.2	6.7	0.404
Phenotypes and Involvement (Progression)			
Ileal Disease (no initial complications)	=37/41	=21/28	0.106
Inflammatory	=23/41	=19/28	0.452
Stricturing	=12/41	=0/28	**0.001**
Penetrating	=2/41	=2/28	1.000
Time to Stricturing (years)	2.9	NA	NA
Surgical History			
Surgery with Anastomosis	20	6	**0.048**
Peri-anal Surgery	8	3	0.511
Temporary Ileostomy	4	3	1.000
Permanent Ileostomy	1	0	1.000

Table 2.6 Clinical phenotypes of pediatric CD patients. Pediatric CD patients were classified into low miR-31 expression (<150 RPMMM; n=46) and high miR-31 expression (≥150 RPMMM; n=30) groups. Clinical phenotypes were recorded at time of initial diagnosis when miR-31 expression was determined, and at subsequent time points after these initial diagnoses. Location of disease in the upper gastrointestinal (GI) tract is in addition to colonic and/or ileal disease. Only patients that initially presented with inflammation only and no complications were considered when assessing progression to disease complications. Associations between molecular subtypes and clinical phenotypes were assessed using Fisher's exact test and were performed only on categories with at least 8 patients across both subtypes. Significant associations ($p<0.05$) are bolded.

CHAPTER III: UPREGULATION OF PANETH CELL-ASSOCIATED ANTIMICROBIAL PEPTIDE EXPRESSION WITHIN COLONIC IECS DEFINES A NOVEL MOLECULAR CD SUBTYPE

INTRODUCTION

The number and scale of disease-specific sequencing studies continue to increase, improving our ability to accurately and robustly determine molecular subtypes of disease. As well as providing the potential to identify novel subtypes, this increase in power allows for a more nuanced distinction of the molecular drivers underpinning distinct disease subtypes. Subtyping of disease through biologically relevant molecular similarities and differences has been well documented in various cancers where this methodology has resulted in a more accurate prognosis, better optimized individualized patient management, and improved therapeutic selection (37). The molecular subtyping of pancreatic cancer (PDAC), where subtypes do not currently inform clinical decision making (116), is still in its infancy. The first classification studies of PDAC using transcriptomic data discovered three prognostic subtypes that have been further revised to four subgroups after numerous subsequent studies (117) (118) (119), reminiscent of early breast cancer studies (120). The established subtypes of breast cancer identified through molecular profiles are now routinely used to guide decision making in the clinic and to facilitate the development of novel therapeutics (37).

Previously, our group identified distinct molecular subtypes of Crohn's disease (CD), a genetically and clinically heterogeneous disorder of the gastrointestinal (GI) tract characterized by abnormal immune responses to luminal gut contents, and one of the primary inflammatory bowel diseases (IBD) (50). Using non-inflamed samples of the colonic mucosa from 21 adult patients, striking differences in molecular profiles were observed across genome-wide gene expression, open chromatin, and microRNA data (82) (83). Differential gene expression analysis revealed a contrasting enrichment for markers of normal colon-tissue (colon-like; CL) and ileum-tissue (ileum-like; IL) between the two groups. Importantly, we identified clinically relevant associations with these molecular subtypes, including stricturing ileal disease, severe rectal disease, and need for anti-TNF therapy post-surgery (82) (83).

Although the exact cause of CD is unknown, numerous studies have suggested that the dysfunction of intestinal barriers is fundamental to its pathogenesis (121) (122). Intestinal epithelial cells (IECs) are at the interface of the enteric microbiota and the intestinal mucosa, mediating complex crosstalk to provide a physical and chemical barrier that protects the intestine from pathogenic and commensal microbial species (123). The IEC monolayer consists of various cell types with specialized functions whose proportions change according to specific segments of the GI tract, protecting host tissue through mucus secretion, tight junctions between cells, and antimicrobial peptide (AMP) production (122). Paneth cells are a specialized cell type within IECs of the small intestine that produces antimicrobial peptides (AMPs) to protect stem cell populations residing within the base of intestinal crypts from the enteric microbiota (121) (124). In the colon, the exact mechanisms driving the protection of the intestinal epithelial are unclear, but a comparatively thicker mucus layer due to increased abundances of goblet cells is thought to play a critical role in host protection from the intestinal microbiota (125). Although Paneth cells are not detected within the colonic mucosa of healthy individuals, several studies suggest that inflammation induces Paneth cell metaplasia in the colon (126) (127) (128).

Validation of previously established subtypes using a larger patient cohort will provide a more complete understanding of the molecular and cellular drivers of CD subtypes to enable novel development of therapeutics. Due to the heterogeneity of CD and transcriptomic profiles of the colonic mucosa (82) (77) (129), increasing the study size may facilitate the discovery of additional molecular subtypes. In this study, we performed RNA-sequencing (RNA-seq) on a large cohort of non-inflamed mucosal samples from the ascending colon consisting of 90 CD and 27 non-IBD controls (NIBD). Through deconvolution analysis of bulk colonic mucosal tissue, we first confirmed that the primary cellular faction of our samples was IECs. After correcting for varying proportions of IECs across our cohort, we repeat the stratification of CD samples into the previously established CL and IL subtypes (82), and further confirm our previous finding that used microRNA expression data to suggest that the stratification is driven by gene expression profiles within IECs (83). Using CL samples, we performed consensus clustering analysis to further subdivide the subtype into two distinct clusters. Through differential gene expression and pathway analyses, we determined that specific Paneth cell markers drive the separation of these novel CD subtypes suggesting the presence of undefined Paneth-like cells with AMP function.

Our study underscores the importance of molecular profiling in heterogeneous disorders such as CD. These results may reflect a novel mechanism for the colon to respond to an increased inflammatory and microbial burden that is absent in a subset of patients.

RESULTS

Covariate correction and deconvolution analyses reveal bulk colon mucosa transcriptomic profiles driven by intestinal epithelial cells

In contrast to our previous study that demonstrated distinct clusters of CD patients using Mrna expression profiles (82), the patient cohort utilized here incorporated an additional 78 medically refractory CD patients undergoing surgery as well as CD patients undergoing routine endoscopies at UNC hospitals. This presented additional challenges in the processing of expression data for downstream analyses. Tissue samples obtained for expression analysis were ascending colon mucosal biopsies and expected to predominantly consist of intestinal epithelial cell (IEC) populations. To confirm this, we performed deconvolution analysis of our bulk RNA-seq data using expression profiles from 12 cell types expected within mucosal tissue, along with control expression profiles from pancreas and heart tissue (Figure 3.1A), to predict IEC representation within tissue samples. As expected, IECs were the primary cell type across our samples with the remainder mainly consisting of T-cells and macrophages. For skeletal muscle, heart, and pancreas data, deconvolution did not attribute these expression profiles to any of our colonic samples. Although predicted IEC proportion varied across samples (mean: 78.9%, SD: 0.073), differences in IEC proportions were not significant between CD and NIBD ($P = 0.195$) or surgical and endoscopic ($P = 0.541$) samples. These findings suggest that variations in expression profiles across samples are largely representative of changes in IEC gene expression.

To facilitate the detection of changes in gene expression driven by CD and molecular subtypes of CD, we corrected for several covariates in differential gene expression and downstream clustering analyses. Samples were sequenced across 9 batches which separated when performing principal components analysis (PCA) (Figure 3.1B). Subsequent PCA after correcting for batch resulted in significant correlations between top principal components and patient sex (PC2: $r = -0.686$, $P = 2.2 \times 10^{-16}$; Figure 3.1C), IEC proportion (PC1: $r = 0.400$, $P = 6.5 \times 10^{-06}$; Figure 3.1D), RNA degradation (Transcript Integrity Number, TIN; PC1: $r = 0.339$, $P = 1.6 \times 10^{-04}$), and age at sample acquisition (PC2: *r*

= -0.289, $P = 0.002$; Table 3.1). Significant correlations were also observed between these covariates and top principal components when analyzing CD subtypes independently. Further, we detected significant differences between patient age at sample acquisition and disease status ($P = 2.3 \times 10^{-08}$), in addition to TIN and disease subtype ($P = 1.9 \times 10^{-06}$; Table 3.2). Although differences in patient age between our NIBD and CD patients are not surprising due to an increased prevalence of GI complications in later life compared to the presentation of CD predominantly in young adults (130) (131), age needed to be corrected to prevent the detection of variation due to differences in age across sampled patients. To ensure consistency, gene expression data throughout our analyses were corrected for batch, sex, age at sample acquisition, IEC proportion, and TIN.

Transcriptomic data from a large adult cohort recapitulates previously established molecular subtypes

Our previous two studies demonstrated the utility in genome-wide Mrna, lncRNA, microRNA, and chromatin accessibility for the stratification of CD into distinct molecular subtypes (82) (83). To strengthen these findings, we combined our previous cohort of 13 adult mucosal biopsies from the ascending colon (12 CD and 1 NIBD) and further introduced 107 adult mucosal biopsies (78 CD and 29 NIBD) to identify CD molecular subtypes using Mrna and lncRNA expression. PCA of Mrna (Figure 3.2A) and lncRNA (Figure 3.2B) expression profiles revealed the characteristic stratification of CD samples as observed previously, with one subgroup of CD samples more closely clustering with NIBD controls and the other clustering independently. Differential expression (DE) analysis was performed between these two CD subgroups, revealing an 8.5 fold increase in the number of DE genes between the two groups than observed previously (849 genes at FDR < 0.05 previously (82) versus 7230 genes at FDR < 0.05 in the current study), primarily due to the increase in power obtained by our increased sample size in the present study. Colon-like (CL) and ileum-like (IL) expression signatures were confirmed through the comparison of colon-specific and ileum-specific marker expression used in our previous analysis (Figure 3.2C) (93) (82). In our previous studies (82) (83), we used the ileum-specific marker APOA1 (Figure 3.2D) and colon-specific marker CA2 (Figure 3.2E) to designate samples as CL or IL. By contrasting the expression of these markers across our two CD groups, we observed a clear distinction between samples that strongly associate with ileum or colonic transcriptomic signatures.

To interrogate genome-wide colon and ileum expression signatures further between the identified CD subgroups, gene set enrichment analysis (GSEA) was performed using weighted gene scores based on DE significance and fold change. Custom phenotypes consisting of colon-specific and ileum-specific tissue markers from an analysis of gene expression signatures from various regions of the gastrointestinal tract (93) were tested across our weighted gene set. As expected, we discovered a strong enrichment of ileum-specific genes for upregulated genes within our isolated CD subgroup (FDR < 0.0001) compared with highly significant enrichment of colon-specific marker genes within the upregulated gene set of our other CD subgroup (FDR < 0.0001) (Figure 3.2F). Pathway enrichment analysis further recapitulated our previous findings. IL CD samples were enriched in genes involved in metabolism (FDR = 5.67×10^{-23}), specifically lipid metabolism (FDR = 2.60×10^{-18}) and xenobiotic metabolism (FDR = 0.0014), whereas CL CD samples exhibited enrichment for energy production through the TCA cycle (FDR = 4.37×10^{-06}). Together, these results demonstrate that our novel cohort of CD samples from non-inflamed regions of ascending colonic mucosa display distinct gene expression profiles replicating the findings of our previous studies.

The identification of molecular CD subtypes through PCA resulted in 11 IL CD samples (4 novel IL samples) and 79 CL CD samples (74 novel CL samples). In contrast to our previous two studies (82) (83), where samples used were identical across both publications, we observe unequal proportions of our two CD subtypes (11.5% IL CD in the present study versus 47.6%-50% IL CD in previous studies). There may be several reasons for this discrepancy. Firstly, our original two molecular subtypes may represent extremes within the CD patient population that are not fully captured in the cohort selected for this study. A total of 21 patients were selected for our original study, with tissue samples being taken from various regions of the colon, compared with 96 samples taken specifically from the ascending colon of patients discussed here. These differences in both sample size along with the anatomical location of the sampled tissue may contribute to the altered molecular subtype proportions we observed. Secondly, the samples acquired here consist of surgical and endoscopic biopsies. In general, patients who undergo surgery might exhibit more active disease in comparison to non-surgical patients, which we may expect to be reflected at the molecular level when looking at gene expression with our previous studies associating molecular subtypes with disease course (82) (79) (83). Overall, 0/40 CD samples obtained via an

endoscopy were identified as IL CD compared with 11/56 IL CD samples from surgical procedures. These data again confirm that at least two distinct molecular subtypes exist within colonic CD driven by GI location-specific expression profiles. Further studies of these molecular subtypes, in particular IL CD, will require careful study design to ensure sufficient sample sizes for statistical power and further refinement of the IL CD subtype.

Differential gene expression analysis reveals ileal gene signatures in colon-like CD samples

To evaluate further whether additional CD molecular subtypes could be identified using transcriptomic signatures from non-inflamed colonic tissue, we filtered samples identified as IL CD. Due to the extreme variation in gene expression profiles between IL and CL CD samples, dimensionality reduction through PCA was sufficient to confidently assign samples into subtypes for downstream analyses. PCA of gene expression data from CL CD and NIBD samples (Figure 3.3A) indicated that variation among CL CD samples was largely driven by principal component 1 (PC1; $r = 0.56$, $P = 7.91\times10^{-08}$). Interestingly, PC2 was strongly correlated with disease status ($r = 0.52$, $P = 1.26\times10^{-08}$) suggesting that variability across these samples is primarily driven by CL CD samples rather than changes in gene expression due to disease status. Through differential gene expression analysis of CL CD and NIBD, we detected 210 differentially expressed genes (FDR < 0.05; Figure 3.3B). After setting a magnitude threshold ($\log_2$ fold change > 1.5) for genes significantly upregulated in CL CD, 5/9 genes (REG3A, DEFA6, DEFA5, REG1B, ITLN2) were specifically elevated in the terminal ileum according to GTEx (132). These same genes are upregulated further in IL CD when compared against CL CD samples (FDR < 1×10^{-06}, $\log_2$ fold change > 5.1), suggesting a discrete, but detectable, ileal signature within a subset CL CD samples.

Consensus clustering identifies colon-like CD clusters driven by Paneth cell-associated expression signatures

To stratify CL CD into further subgroups, we employed the class discovery methodology consensus clustering (CC) (133) and the partitioning around medoids clustering algorithm across 79 CL CD samples. Compared to common clustering approaches that produce varying results either due to a randomized start procedure (K-means) or a user-defined number of clusters (K-means/hierarchical

clustering) (134), CC relies on multiple iterations of a clustering algorithm on subsamples of the data to gain consensus subgroup assignments (133). After first removing rare and ubiquitously expressed genes from the dataset that would not be informative for identifying novel sample clusters, and further identifying a gene set of the most variable genes across all CL CD samples, unsupervised CC analysis was performed. To identify the optimal number of clusters from CC analysis of CL CD samples, we used the cumulative distribution function (CDF; Figure 3.4A). Using consensus index (CI) values in the range [0.2, 0.8], we defined the optimal and stable partitioning of samples as the flattest CDF curve within our CI range. The proportion of ambiguous clusters (PAC) score was also employed as a robust alternative to identify the correct number of clusters (135), with the lowest PAC score of *k* clusters indicating optimal cluster number. We found that 2 clusters inferred the optimal number of clusters within our dataset, defined by the most stable CDF curve (Figure 3.4A, red) and the lowest PAC score (2 clusters = 0.181).

To identify the genes driving CC clusters, we first performed hierarchical clustering on the 100 most variably expressed genes across 79 CL CD samples (Figure 3.4B). Across one of our CC derived clusters we observed a group of genes that are more robustly expressed within the small intestine (132) (136). Additionally, we determined that a subset of these genes were markers of Paneth cells, a specialized cell type found at the base of small intestinal crypts that produce antimicrobial peptides (AMPs), which is not usually within the colon in healthy individuals (121) (124). The variability in AMP expression across CD samples from the colonic mucosa was also recently observed in a study using Genome-wide 5'-RNA sequencing of capped RNAs (CAGE) (77). To further confirm this Paneth cell signature, we performed differential gene expression analysis between CC-derived CL CD subgroups revealing 452 differentially expressed (DE) genes (FDR < 0.05, Figure 3.4C). By considering significantly DE genes with the largest changes in magnitude between CC groups ($\log_2$ fold change > 1.5), 6/16 genes (REG1A, DEFA6, REG3A, DMBT1, REG1B, DEFA5) were established Paneth cell markers (FDR < 0.003, $\log_2$ fold change > 1.7; Table 3.3). These results suggest that the upregulation of ileum-associated genes, especially marking Paneth cells, are driving the differences across our two CC clusters. Due to the upregulation of this Paneth-associated signature, we referred to CC clusters as Paneth enriched CD (PE) and Paneth depleted CD (PD).

To more broadly evaluate differential gene signatures, we compared pathway-level expression patterns of CC subgroups. We first identified several significant associations (FDR < 0.05) with innate immune responses, including interferon signaling, defensins, and antigen presentation in the PE subgroup (Figure 3.4E, orange). Reanalysis of our deconvolution results for predicted immune cell fractions revealed no significant differences in T-cell, macrophage, or B-cell proportions between CC groups. With the majority of colonic samples previously being shown to consist of IECs, upregulated immune-associated pathways may reflect an increased immune capacity of PE IECs. Pathway analysis also revealed a significant enrichment of metabolism-associated pathways in PE samples, resembling the upregulation metabolic functions observed in IL CD (82). In the PD subgroup, we observed a significant association of pathways involved in BMP signaling along with guanylate cyclase and sialic acid metabolism (Figure 3.4E, blue). BMP signaling is part of the transforming growth factor (TGF)-beta family of signaling molecules and contributes to a range of biological functions across various tissues (137). Interestingly, a recent study by our group discovered that BMP signaling restricts the stemness of colonic IECs, with upregulated BMP signaling diving differentiation towards colonocytes (138), a key cell type in maintaining colonic barrier function (139). These associations may, therefore, suggest that differences in gene expression profiles between our CC groups point to an altered stemness capacity within IECs promoting elevated immune function.

Together, these results suggest that the heterogeneity among CL samples is indicative of additional CD molecular subtypes. Similar to the identification of the IL subtype, differences between CL subgroups are driven by gene expression signatures found outside the colon in healthy individuals. The upregulation of AMP gene expression may reflect a primed state of PE IECs within a subset of CL CD patients for defense against infiltrating bacterial populations within the colonic lumen.

Upregulation of Paneth marker genes may suggest the differential proliferation of an undefined Paneth-like cell type between colon-like CD groups.

To contrast the expression of Paneth-associated genes across our 3 CD subgroups and between NIBD controls from the colon (Cnibd) and ileum (Inibd), we compared the expression of three AMPs that are critical to the antimicrobial function of Paneth cells, DEFA5 (Figure 3.5A), REG1A (Figure 3.5B), and LYZ (Figure 3.5C). LYZ is an important enzyme that is secreted by Paneth cells to regulate intestinal inflammation (140) (141) and is often used as a marker of Paneth cells (142) (143). We observed a

striking gradient of increasing DEFA5 and REG1A expression from Cnibd to Inibd samples that are not shared by LYZ. LYZ is not differentially expressed between PD and PE subtypes (FDR < 0.324) but is significantly elevated in IL samples compared to PD and PE subtypes (FDR < 5.65×10^{-07} and FDR < 4.23×10^{-03} for PD and PE comparisons respectively). We further performed a supervised hierarchical clustering analysis using a combination of CL, IL, PD, and PE specific genes selected through differential analysis (Figure 3.5D). The result from this analysis again reveals the extreme variations in expression profiles between IL and CL subtypes and displays the underlying heterogeneity among CL CD samples. Although a clear stratification of IL CD and CL CD samples was observed, we did not observe a clean stratification of PE and PD subtypes. Interestingly, the only subset of genes used in this analysis that are consistently upregulated in both IL and PE CD samples are the Paneth cell AMPs identified through PCA. While these results generally agree with the classification obtained through CC analysis, further refinement of these subtypes will be necessary to understand the molecular mechanisms driving Paneth signature displayed by a subset of CL CD samples.

To characterize the impact of CL CD subgroups on disease presentation, we analyzed the clinical phenotypes of the same 50 PE and 29 PD samples defined through CC analysis (Table 3.4). Although we detected no significant differences in disease location, disease behavior, or treatment history at the time of sample collection, we found that disease duration at the time of sampling was significantly shorter in PE compared with PD ($P = 0.026$). However, after further correlation analysis of mean AMP expression against disease duration, we found that the variability in disease duration among CL CD patients resulted in a no correlation between AMP expression and disease duration ($r = 0.08$, $P = 0.943$). Compared with IL and CL molecular subtypes, our newly discovered PD and PE CD subtypes display more subtle changes in gene expression profiles. Longitudinal prospective monitoring of clinical characteristics will be necessary to relate these molecular subtypes to clinical presentation.

Taken together, these results suggest that differences in AMP expression drive the stratification of CL CD samples. The specific combination of AMP genes driving this expression profile point to an undefined Paneth-like cell type that may form part of a response to inflammation of the colonic mucosa in a subset of CL CD patients. Further confirmation of AMP expression in colonic IECs will be vital to

understand this expression signature and its function within CL CD. More extensive clinical characteristics will also become important in follow up studies to translate these findings to the clinic.

DISCUSSION

We confirmed our previous findings that genome-wide transcriptomic profiles of colon tissue from adult CD patients stratify into two distinct molecular subtypes. While further confirming the enrichment of pathways involved in lipid and xenobiotic metabolism in the IL subtype, our study builds on this result by performing deconvolution analyses to confirm that molecular signatures are driven by expression profiles in the colonic epithelia. Surprisingly, we found far fewer IL samples in this larger patient cohort. Compared to the adult cohort used across our previous studies (82) (83), here we recruited samples from patients undergoing surgical procedures and routine endoscopies. We found that 0/40 endoscopy samples displayed IL gene expression signatures suggesting that disease course may be inherently linked to the presentation of CL and IL molecular subtypes. Although additional analyses focusing on IL samples was not performed in this study due to limitations in power, it will become vital in the future to design studies that ensure greater numbers of samples to refine this molecular subtype further.

Through unsupervised consensus clustering analysis of 79 CL subtype samples, we identified two distinct clusters that were driven by the upregulation of small intestinal genes. Further, differential analysis suggested that the antimicrobial peptides (AMPs), specially REG family genes and α-defensins, and DMBT1 indicate Paneth cell signatures within a subset of CL CD patients that is not observed within the colonic epithelia of healthy individuals. Generally, these genes function to regulate the enteric microbiota within the small intestine, representing key players of innate immunity (144) (145) (146). The REG family of genes represent a group of secretory proteins that play a wide range of roles to promote proliferation, differentiation, and prevent apoptosis (144) with expression mainly localized to the pancreas, liver, brain and GI tract (147). A previous study suggested that REG is upregulated in the colon in response to inflammation (148) and further investigations by *in situ* hybridization showed that REG1A, REG1B, and REG3A are detected in the colonic mucosa of inflamed and uninflamed IBD patients, localizing to metaplastic Paneth cells (128). Various studies have shown that ileal CD is characterized by decreased α-defensin expression compromising the antimicrobial activity of the gut mucosa (149) (150).

Although normal colonic mucosa does not express α-defensins, they have been detected within the crypt of inflamed and uninflamed IBD samples due to Paneth cell metaplasia (126) (127). DMBT1 has been proposed to play roles in various epithelial cancer types with studies suggesting potential functions in the differentiation of epithelial and stem cells for tumor suppression (151) (152). In the intestine, upregulation of DMBT1 has been observed primarily within Paneth cells of various epithelial cell lines (153), human ileal Paneth cells correlating with disease activity in IBD patients (154), and in the IEC of the colonic mucosa biopsies in response to intestinal inflammation (155). In contrast to our sample cohort, which consists of non-inflamed sections of the ascending colon, recent studies have suggested that inflammation plays a key role in the upregulation of Paneth cell markers in the colonic epithelial. Follow-up studies will be required to determine whether the upregulation of Paneth cell marker genes in colon IECs is the result of aberrant regulatory mechanisms or represents a priming response to intestinal inflammation.

While our results are consistent with previous studies that have discovered Paneth-like gene expression profiles in colonic crypts in rats (156), mice (157), and humans (158), a major discrepancy between studies performed using human colon samples are the genes marking Paneth-like cell populations. Studies in mice have discovered populations of goblet cells marked by the stem cell growth factor cKit (cKit+ cells) located at the base of colonic crypt and interspaced between Lgr5+ stem cells, reminiscent of Paneth cells in the small intestine (157). Through the secretion of several important factors, such as defensins, these cells were hypothesized to play an important role in maintaining crypt homeostasis in the colon. A more recent study suggested that cKit+ cell differentiation is promoted by Stat5 upregulation in response to intestinal injury (159). In human colonic crypts, significant upregulation of DEFA5, DEFA6, and LYZ has been attributed to metaplastic Paneth-like cells in the colon of UC patients (160). Single-cell RNA-seq (scRNA-seq) of non-inflamed colonic mucosa sections from colon cancer patients revealed a Paneth-like population of cells marked by a number of Paneth marker genes, including LYZ (158). In contrast to our findings, the Paneth-like cell signature did not include defensins and REG family genes. Through reanalysis of this scRNA-seq dataset, we found that cell clusters identified as Paneth cells in ileal mucosa samples were not specifically marked by DEFA5 or reg family genes, raising further question about the reliability of this data in identifying Paneth-like gene signatures.

Although Paneth signatures in the colon reported in the literature and our study appear different in terms of the genes marking this cell population, they provide significant evidence to suggest a secretory Paneth-like cell type that is generated in response to intestinal injury to protect the colonic mucosa from the enteric microbiota.

In summary, we have recapitulated previously established molecular CD subtypes and provided evidence that Paneth cell signatures drive the stratification of a third CD subtype across surigcally resected and ensocopic samples from human patients. In agreement with previous studies, we find that the upregulation of AMPs in the colonic mucosa is upregulated and associated with intestinal injury and inflammation. We hypothesize that the differentiation of cells with Paneth-like function within colonic crypts is in response to the infiltration of enteric microbiota, although further investigation will be required to determine the causal stimulus that facilitates the proliferation of this cell population. Future longitudinal studies of patient phenotypes of Paneth-enriched and Paneth-depleted CL patients will be required to understand the clinical utility of these novel CD subtypes. High-throughput single-cell assays that assess the transcriptomic and regulatory landscape of Paneth-like cell types within colonic crypts of CD patients will become vital for future follow-up studies of molecular CD subtypes.

MATERIALS AND METHODS

Patient populations and phenotyping

Patients with CD and NIBD-related illnesses diagnosed at The UNC hospitals were included in this study. Clinical phenotypes considered in this study include demographic and clinical variables such as age, sex, disease duration, age at diagnosis, age at sample acquisition, disease location, and disease behavior. A Summarized table of patient demographics and phenotypes for clinical associations are provided in Table 3.4 and a summary of samples used in the study is provided in Table 3.5. This study was not blinded, and all authors had access to the study data.

Tissue isolation and characterization

For our adult cohort, all CD and NIBD mucosal biopsies were obtained from macroscopically unaffected sections of the ascending colon or terminal ileum at the time of surgery or endoscopy and

were flash-frozen. The ascending colon was chosen specifically to avoid the detection of colon sublocation-specific gene expression differences. No samples showed signs of active microscopic inflammation or disease, as confirmed by an independent pathologist. Absence of acute (active) inflammation, including neutrophilic inflammation of crypt epithelium and crypt abscess formation, and chronic inflammation, including architectural distortion and basal lymphoplasmacytosis of the lamina propria, was determined after a review of each H&E-stained slide.

RNA isolation, sequencing, and processing

RNA was isolated from flash-frozen adult samples from surgical resections and endoscopic biopsies using the Qiagen RNeasy Mini Kit following the manufacturer's protocol. This kit uses column-based DNase treatment to eliminate DNA contamination, and it allows the miRNA and mRNA content to be preserved. miRNA was enriched from FFPE tissue for pediatric samples using the Roche High Pure miRNA Isolation Kit. RNA purity and integrity were assessed with Thermo Scientific NanoDrop 2000 and Agilent 2100 Bioanalyzer, respectively. For all clinical categories of flash-frozen adult samples, we observed average RNA integrity (RIN) values above 7. RNA-seq libraries were prepared using the Illumina TruSeq polyA+ Sample Prep Kit. Paired-end (50 bp) sequencing was performed on the Illumina HiSeq 2500 and 4000 platforms.

Before alignment of sequencing reads to the reference genome, sequencing adapters were removed using TagDust (161). Further filtering of sequencing reads based on quality was performed using FASTX-Toolkit (161) using the parameter '-q 33'. Quality metrics were retrieved using FastQC (162), RSeQC (163), and curated via MultiQC (164). Samples with an average GC content greater than 60% were excluded from the analysis. Sequencing reads were aligned to the hg19 genome assembly using STAR (111) using default parameters. Quantification of gene transcripts was performed through RSEM (165) with default parameters. Measures of RNA degradation were obtained using median Transcript Integrity Number (TIN) (166) from the RSeQC (163) package.

Tissue deconvolution

Deconvolution of colonic gene expression profiles to estimate cell type proportions was performed using the function 'unmix' within the DESeq2 package v1.22.2 (113). Gene expression profiles

of 12 distinct cell types from 3 different sources were used to deconvolute colonic gene expression. CD326 Intestinal epithelial cell (IECs), CD3 T-cell, and CD33 macrophage profiles were obtained from in-house, unpublished data of antibody-based magnetic cell sorted cells from NIBD individuals. B-cell Neutrophil, and Natural killer cell expression data were taken from Linsley *et al.* (167) (Gene Expression Omnibus [GEO] accession no. GSE60424). Finally, fibroblast, adipose, nerve, skeletal muscle, pancreas, and heart expression profiles were obtained through GTEx (132). Pancreas and heart were intended as control expression profiles and no cell type proportions were attributed to either for any sample analyzed.

All cell-specific expression profiles were normalized using DESeq2 and were corrected for study-specific biases using the function 'removeBatchEffects' from the limma package v3.38.3 (168). Unmix was performed using the parameter 'shift=0.5'. Samples with a predicted IEC proportion of less than 60% correlated well with obvious PCA outliers and were therefore removed from downstream analyses.

RNA analysis

Correction of sequencing batches and sex, age, IEC proportion, and TIN covariates was performed using the 'removeBatchEffects' function in limma on DESeq2 normalized variance stabilizing transformation (VST; 'blind=TRUE') transformed counts. mRNAs were defined as "protein_coding" and lncRNAs as "lincRNA"," sense_intronic"," sense_overlapping", or "antisense" using GENCODE GRCh37 version 19 biotype annotations. PCA was performed using normalized, transformed, and corrected counts using the prcomp function in R. Correlations with principal components for quality control was performed using the 'cor.test' function in R.

Consensus clustering was performed using the ConsensusClusterPlus package v1.46.0 in R. 'partitioning around medoids' and 'maximum' were used for as the clustering algorithm and distance metric, respectively, across 10000 iterations of the 500 most variable protein-coding genes across 79 CD samples. For each iteration, pFeature=1 and pItem=0.8, resulting in clustering of 80% of samples across all 500 protein-coding genes. The optimal value of *k*-clusters was selected by visual inspection of the cumulative distribution function (CDF) and the lowest proportion of ambiguous clustering (PAC) using consensus index values of 0.2 and 0.8 for lower and upper bounds, respectively. The 'PAC' function in diceR package v.0.6.0 was used to calculate PAC values. Hierarchical clustering analysis was performed

using the ComplexHetmap (169) v2.1.0 package v in R. ‘ward.D2’ was selected as the clustering method and ‘euclidean’ was chosen as the method to measure distance for both rows and columns.

For differential gene expression analysis, we used DESeq2. Sequencing batches and covariates corrected for downstream analyses were included in the design formula. Numeric covariates were centered and scaled before differential analysis to prevent convergence issues with the generalized linear model employed by DESeq2. Volcano plots were generated using the EnhancedVolcano (170) v.1.3.5 package in R following DESeq2 analysis. R version 3.5.0 was used for all processing and visualization of data.

Pathway analysis was performed using gene set enrichment analysis (GSEA) (171) through WebGestalt (172) and Enrichr (173). Gene weightings for GSEA were calculated using DESeq2 FDR and log2 fold change. Colon-specific and ileum-specific genes used as GSEA phenotypes in Figure 3.2F were taken from Weiser *et al.* (82).

Statistics

Statistical significance for differential analyses using DESeq2 was measured using FDR adjusted p-values. Correlations were performed using Pearson’s distance and correlation significance was assessed using a correlation test for association using Pearson’s distance. Patient phenotypes were tested for significant associations using a 2-tailed unpaired Student’s *t* test (continuous data) or a Fisher’s exact test (categorical data). For all tests, $Padj < 0.05$ or $P < 0.05$ was considered statistically significant, unless otherwise stated.

Study approval

This study received IRB approval at UNC (protocol 10-0355). Written informed consent was received from all participants before inclusion in the study. All participants are identified by number and not by name or any protected health information (PHI).

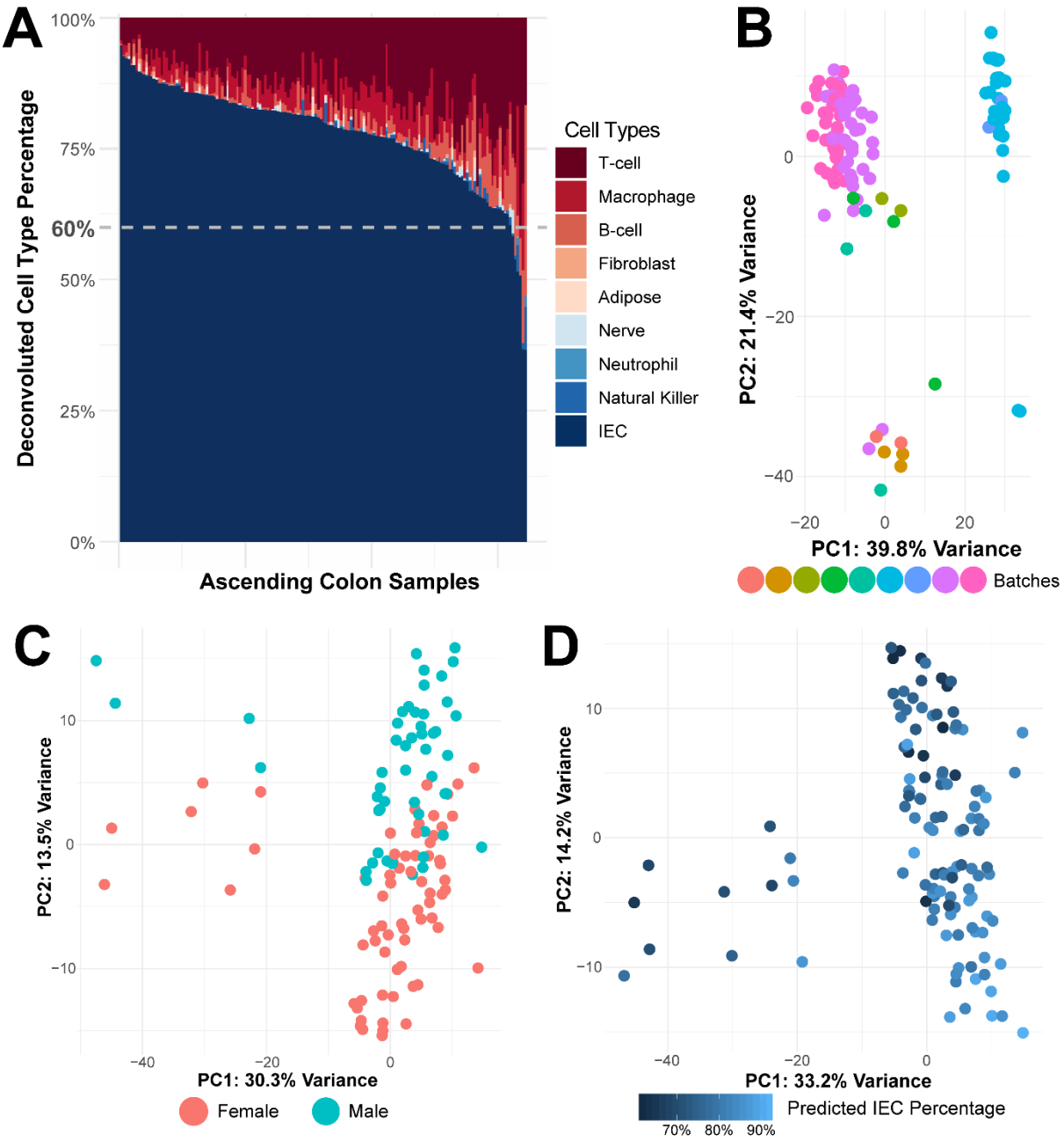


Figure 3.1. Covariate correction and deconvolution of colonic mucosa samples. (A) Deconvolution analysis of RNA-seq data of 145 non-inflamed mucosal biopsies from the ascending colon using cell-type-specific profiles from 12 distinct cell types. Predicted cell type percentages are displayed for cell types in which percentages were assigned. IEC percentage (navy blue) was identified as the top cell type across most samples. Samples with a predicted IEC percentage of less than 60% was removed from downstream analyses. PCA analysis of 136 colonic mucosa samples resulted in significant correlations with sequencing batch (B), patient sex (C), or predicted IEC percentage (D) and PC1 or PC2. As a result, these covariates were included in design formulas for differential analyses and downstream count correction procedures.

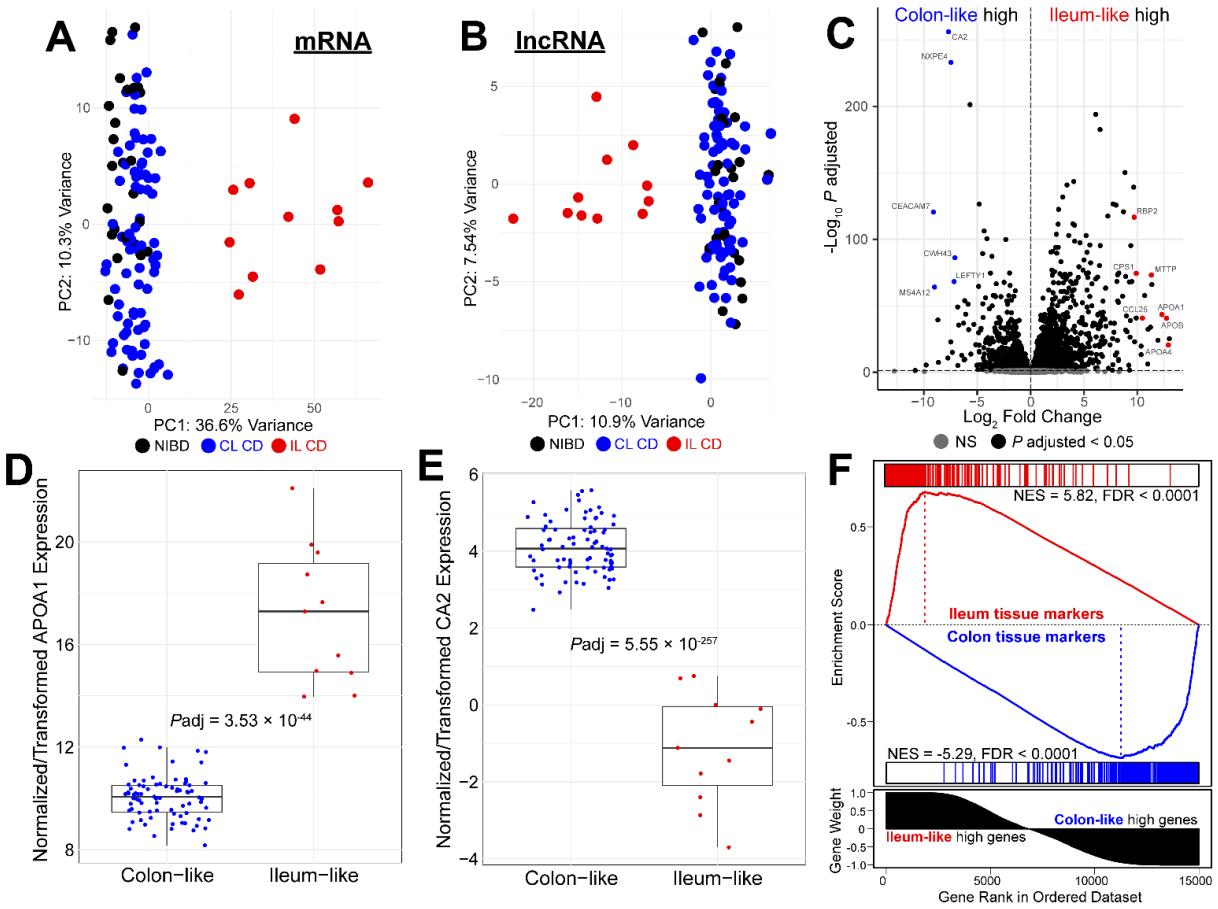


Figure 3.2. Transcriptomic profiles recapitulate colon-like and ileum-like molecular subtypes. PCA analysis reveals the stratification of previously identified CD molecular subtypes using 136 non-inflamed mucosal biopsies from the ascending colon using mRNA (A) and lncRNA (B) data. (C) Volcano plot of CD molecular subtypes after differential analysis using DESeq2. Markers of normal ileum (red; n=11) and colon (blue; n=79) tissue for significantly differential genes (black) between molecular subtypes indicate characteristic colon and ileum-specific expression profiles (82). Expression of APOA1 (D) and CA2 (E) across CD samples reveals a perfect stratification of molecular subtypes. (F) GSEA using ileum and colon tissue marker sets from (93) as custom phenotypes further suggest tissue-specific profiles across colon-like and ileum-like molecular subtypes. FDR adjusted P values and log2 fold changes were determined using DESeq2.

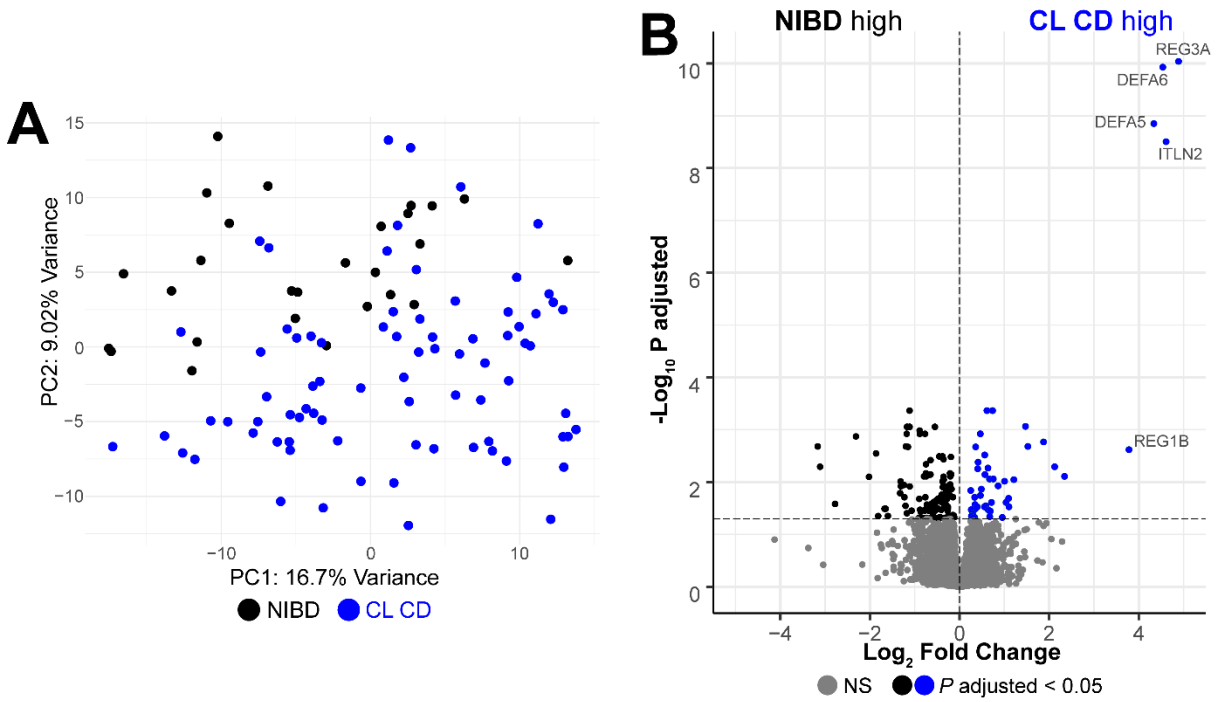


Figure 3.3. Ileal gene expression profiles drive colon-like CD and NIBD differences. PCA of colon-like CD (blue; n=79) and NIBD (black; n=27) samples reveals heterogeneity across colon-like CD in PC1 (r = 0.56, P = 7.91×10-08) and a significant correlation with disease status in PC2 (r = 0.52, P = 1.26×10-08). (B) Volcano plot after differential gene expression using DESeq2. Labelled genes meeting significance and $\log_2$ fold change thresholds (*P*adj < 0.05, |Log_2 fold change| > 1.5) were specifically elevated in the terminal ileum according to GTEx (132). FDR adjusted P values and log2 fold changes were determined using DESeq2.

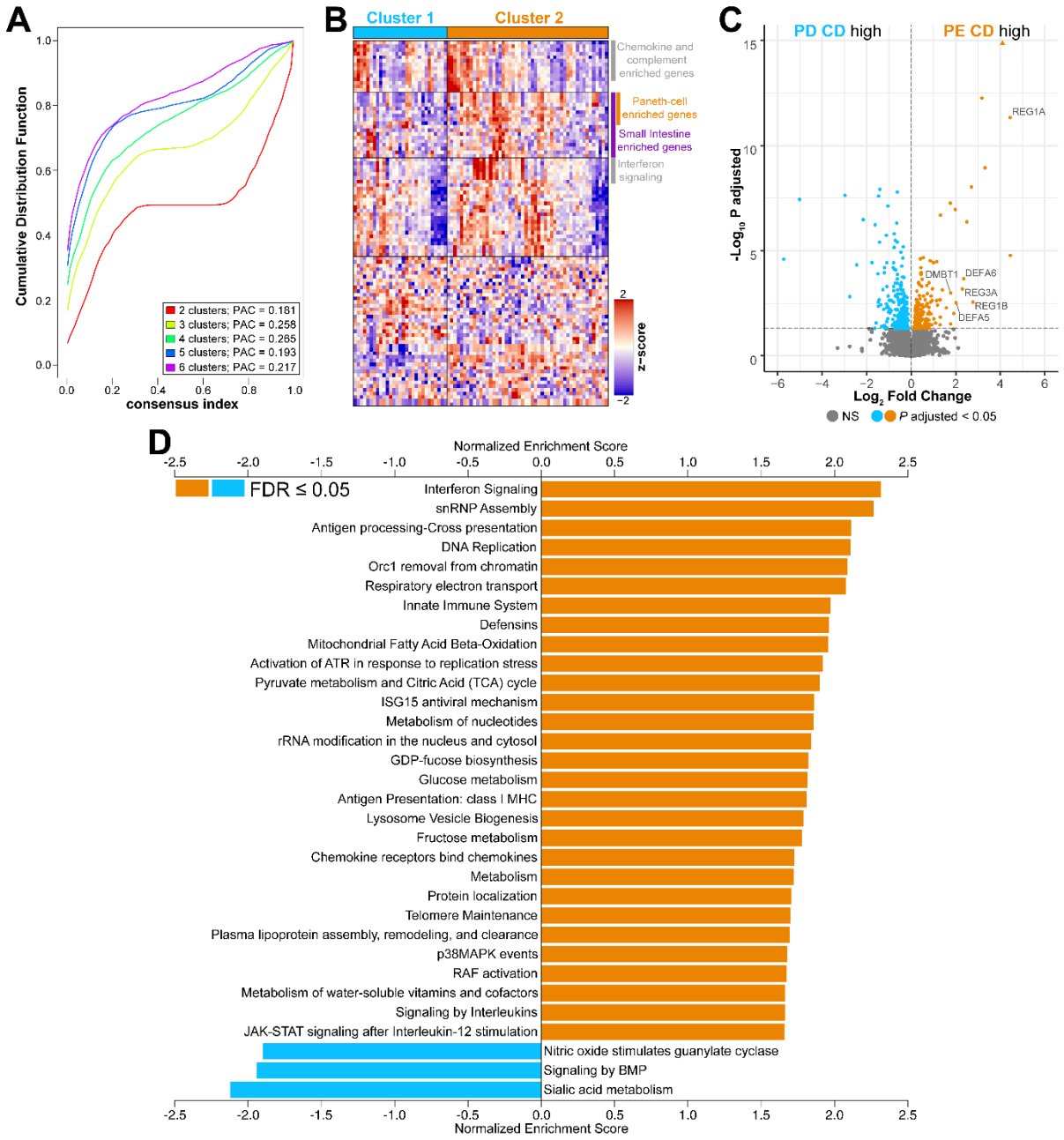


Figure 3.4. Consensus clustering followed by a differential analysis of colon-like subgroups reveals Paneth cell-driven signature. (A) Consensus clustering (CC) analysis using the 500 most variable genes across 79 colon-like CD samples identified *k*=2 (red) as the correct number of clusters. The proportion of ambiguous clusters (PAC) score was used to further quantify cluster stability and select the correct *k* clusters. (B) Hierarchical clustering analysis across CC subgroups, cluster 1 (light blue; n=29) and cluster 2 (orange; n=50), reveals the upregulation of small intestine markers in cluster 2. A subset of small intestinal-associated genes were identified as markers of Paneth cells, a cell type usually found in the small intestine. Columns (samples) were fixed according to CC assigned cluster labels and rows (genes) were clustered using ward's method and Euclidean distance measure. (C) Volcano plot of significantly differential genes between Paneth depleted (PD; light blue) and Paneth enriched (PE; orange) CD samples. Paneth marker genes REG1A, REG3A, REG1B, DEFA5, DEFA6, and DMBT1 are significantly upregulated within one of the CC assigned subgroups. (D) Pathway analysis through GSEA using the Reactome database revealed an enrichment of immune and metabolic-related pathways among genes upregulated in PE CD genes (orange, FDR ≤ 0.05) compared with BMP signaling enrichment among upregulated PD CD genes (light blue, FDR ≤ 0.05). FDR adjusted *P* values and $\log^2$ fold changes were determined using DESeq2.

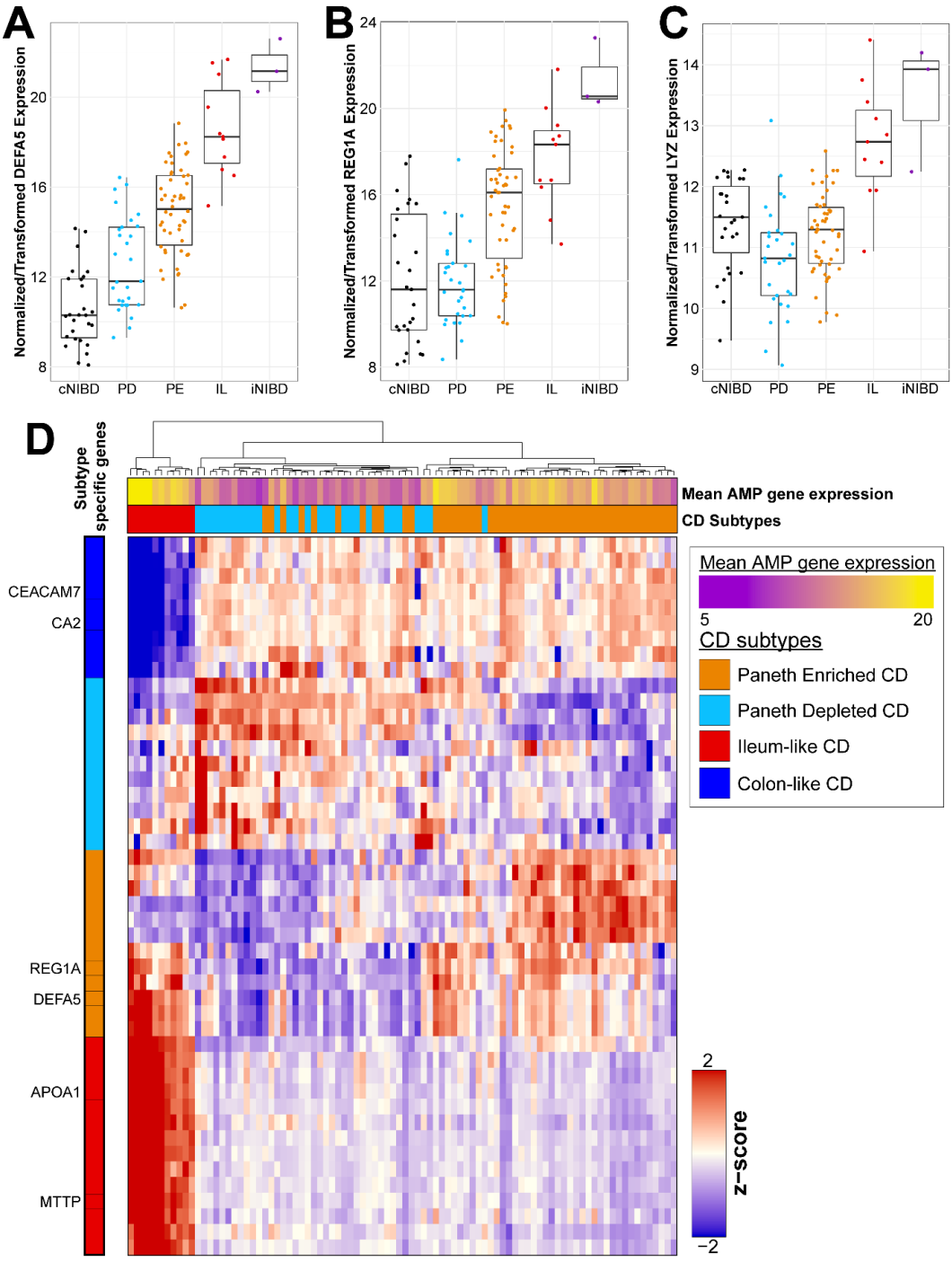


Figure 3.5 Upregulation of specific antimicrobial peptides indicates an undefined Paneth-like cell type in Paneth enriched CD samples. Boxplots across colonic NIBD (cNIBD; n=27), Paneth depleted CD (PD; n=29), Paneth enriched CD (PE, n=50), ileum-like CD (IL; n=11), and ileal NIBD samples (iNIBD; n=3) for DEFA5 (A) and REG1A (B) expression reveals a general gradient in expression from NIBD colon to NIBD ileal samples. LYZ (C), a well-established Paneth cell marker, is not differentially expressed between CL CD subclusters but is generally increased in IL CD samples. (D) Supervised hierarchical clustering using subtype-specific markers of CD molecular subtypes reveals the extreme variation between IL and CL subtypes while displaying underlying heterogeneity among PE (orange) and PD (light blue) subgroups. Mean AMP expression (purple-gold; low-high) using DEFA5, DEFA6, REG1A, REG3A, and REG1B suggests that expression of these genes largely separates CL subgroups, but that this stratification is likely driven by additional gene expression signatures.

	All Samples		IL Samples Removed	
	PC1	**PC2**	**PC1**	**PC2**
Sex	r = -0.066 (P = 0.469)	***r* = -0.686 (*P* = 2.2E-16)**	***r* = 0.415 (*P* = 4.05E-06)**	***r* = 0.875 (*P* = 2.2E-16)**
IEC proportion	***r* = 0.400 (*P* = 6.54E-06)**	***r* = -0.492 (*P* = 1.25E-08)**	***r* = -0.617 (*P* = 2.01E-13)**	***r* = 0.549 (*P* = 2.16E-10)**
TIN Score	***r* = 0.339 (*P* = 1.64E-04)**	r = 0.079 (P = 0.391)	r = 0.107 (P = 0.253)	r = -0.051 (P = 0.587)
Age at sample collection	r = 0.027 (P = 0.766)	***r* = -0.276 (*P* = 0.002)**	***r* = 0.289 (*P* = 0.001)**	r = -0.093 (P = 0.32)

Table 3.1. Principal component correlations indicate significant covariate-driven variation across samples. Correlation (r) values are reflective of Pearson's distance and significance was determined using a correlation test between a covariate and a principal component (PC) value assigned to a sample. $P < 0.05$ (bold) was considered statistically significant.

	Sex	IEC proportion	Age at sample collection	RNA degradation
Status	*P* = 0.6721	*P* = 0.1953	***P* = 2.35E-08**	*P* = 0.9270
Subtype	*P* = 0.7537	*P* = 0.0477	*P* = 0.8935	***P* = 1.90E-06**
CL NIBD	*P* = 0.6672	*P* = 0.2680	*P* = 0.0571	*P* = 0.0967
Procedure	*P* = 0.1305	*P* = 0.5408	*P* = 0.9227	***P* = 0.0439**
Mean	NA	0.7892	43.6849	63.2367
Standard Deviation	NA	0.0731	15.3932	4.2379

Table 3.2. Significant differences between sample groups and covariates warrant correction. Associations between covariates and various sample stratifications were assessed using Fisher's exact test (categorical data) and 2-tailed unpaired Student's t test (continuous data). P < 0.05 (bolded) was considered statistically significant.

Gene ID	Mean Normalized Expression	log_2 Fold Change	FDR	
FAM3B	380.080	5.041	8.58E-39	
GATA5	25.119	4.453	1.68E-05	
REG1A	3650.597	4.443	4.63E-12	
SLC14A2	125.483	3.313	1.14E-09	
SLC37A2	2066.635	3.173	5.53E-13	
REG1B	153.326	2.774	0.003	
POPDC3	26.818	2.704	9.26E-09	
DRD5	56.233	2.495	4.26E-07	**Paneth enriched high**
DEFA6	576.288	2.356	2.14E-04	
REG3A	932.356	2.298	6.61E-04	
DEFA5	1435.380	2.011	0.003	
GPC3	73.037	1.977	1.09E-07	
FOSB	972.589	1.908	0.010	
DMBT1	10688.210	1.775	1.03E-03	
SLC34A2	2.824	1.772	0.031	
DPEP1	58.903	1.756	5.49E-08	
TDRD1	10.951	-1.576	0.027	
GLDN	32.683	-1.620	5.84E-07	
HOXA13	370.622	-1.638	0.049	
ST3GAL4	445.753	-1.771	3.69E-05	
CLDN8	128.254	-2.157	3.32E-07	**Paneth depleted high**
SLC13A1	4.429	-2.450	4.70E-05	
EVX1	2.421	-2.762	1.54E-03	
INSL5	22.443	-2.977	2.33E-08	
HOXB13	54.135	-5.022	3.63E-08	
PRAC1	30.549	-5.728	2.54E-05	

Table 3.3. Top differential genes for Paneth enriched and Paneth depleted comparisons meeting FDR and fold change thresholds. An FDR adjusted *P*adj < 0.05 and absolute log_2 fold change > 1.5 was used to filter genes after differential analysis using DESeq2. 6/16 genes identified upregulated within the Paneth enriched subgroup (bolded) were identified as Paneth cell markers.

	Paneth Enriched	Paneth Depleted	*p* value
Total number	50	29	NA
Sample acquisition, Surgical/Endoscopy, n	27/23	12/17	0.352
Sex, Male/Female, n	20/30	13/16	0.813
Age at sampling, mean (SD)	38.18 (13.14)	41.38 (13.50)	0.300
Disease duration at sampling (years), mean (SD)	10.56 (8.26)	16.24 (14.10)	**0.026**
Former/Current smoker, n (%)	29 (58%)	19 (66%)	0.634
Disease extent at sampling			
Ileum-only, n (%)	17 (34%)	12 (41%)	0.634
Colon-only, n (%)	11 (22%)	4 (14%)	0.387
Ileum + Colon, n (%)	20 (40%)	13 (45%)	0.816
Upper GI, n (%)	8 (16%)	2 (7%)	0.304
Disease behavior at sampling			
B1 - Inflammatory, n (%)	8 (16%)	5 (17%)	1.000
B2 - Stricturing, n (%)	27 (54%)	16 (55%)	1.000
B3 - Penetrating, n (%)	13 (26%)	8 (28%)	1.000
B2/B3 - Stricturing/Penetrating, n (%)	40 (80%)	24 (83%)	1.000
Perianal disease, n (%)	14 (28%)	9 (31%)	1.000
Stricture/Penetration, n (%)	40 (80%)	24 (83%)	1.000
Ileal involvement disease behavior at sampling			
Ileal involvement, n (%)	37 (74%)	25 (86%)	0.387
B1 - Inflammatory, n (% ileal involvement)	6 (16%)	4 (16%)	1.000
B2 - Stricturing, n (% ileal involvement)	24 (65%)	16 (64%)	1.000
B3 - Penetrating, n (% ileal involvement)	7 (19%)	5 (20%)	1.000
B2/B3 - Stricturing/Penetrating, n (% ileal involvement)	31 (84%)	21 (84%)	1.000
Treatment history until sampling			
Aminosalicylates, n (%)	19 (38%)	9 (31%)	0.629
Oral steroids, n (%)	31 (62%)	14 (48%)	0.250
Immunosappressants, n (%)	21 (42%)	11 (38%)	0.814
Anti-TNF alpha agents, n (%)	28 (56%)	20 (69%)	0.340
Anti-integlin, n (%)	6 (12%)	0 (0%)	0.080
Anti-IL12/23p40, n (%)	5 (10%)	2 (7%)	1.000
Antibiotics, n (%)	31 (62%)	16 (55%)	0.637
Surgery, n (%)	19 (38%)	13 (45%)	1.000
1 Surgery, n (% surgery)	13 (26%)	6 (21%)	0.786
More than 1 Surgery, n (% surgery)	6 (12%)	7 (24%)	0.211
Time to 1st Surgery (years), mean (SD)	5.00 (5.55)	9.00 (12.19)	0.236

Table 3.4. Demographics and clinical phenotypes of colon-like CD patients. Clinical phenotypes of Paneth enriched (n=50) and Paneth depleted (n=29) were recorded at the time of initial diagnosis with Crohn's disease and at the time of sample acquisition. Location of disease in the upper gastrointestinal tract is in addition to colonic and/or ileal disease. Associations between molecular subtypes and clinical phenotypes were assessed using Fisher's exact test (discrete data) and 2-tailed unpaired Student's t test (continuous data). Significant associations ($P < 0.05$) are bolded.

	Sample location/subtype						Sample acquisition		Sample cohort	
Disease status	**Ascending colon**	**Ileum-like**	**Colon-like**	**Paneth depleted**	**Paneth enriched**	**Terminal ileum**	**Surgery**	**Endoscopy**	**Weiser/Keith *et al.***	**Present study**
Non-IBD	27	-	-	-	-	3	25	5	1	29
CD	90	11	79	29	50	-	50	40	12	78

Table 3.5. Summary of patient sample numbers.

CHAPTER IV: DISCUSSION

The advances in and the availability of high-throughput sequencing technologies are changing the landscape of disease classification in the lab and the clinic. In heterogeneous and complex disorders, such as Crohn's disease (CD), where standardized treatment may not be effective in all patients, the molecular stratification of disease is a methodology that may translate into novel intervention and prognostic strategies. Through pioneering studies in breast cancer (40), which provided a novel method to study complex disease resulting in enhanced therapeutic strategies, molecular subtyping is now being applied throughout biomedical disease research (174) (37) (175) (176) (177).

The overall theme of the research presented in this dissertation centered on further investigating and refining previously identified molecular subtypes of CD (82) by utilizing high-throughput microRNA (miRNA) and mRNA expression data. The results from our studies can be used to guide future molecular subtyping research within the field by our group and the wider IBD research community.

Stratification of CD molecular subtypes by miR-31 and association with clinical phenotypes

In chapter II, I primarily utilized small RNA-sequencing (smRNA-seq) to evaluate the regulatory impact of miRNA expression on CD subtypes using human mucosal tissue samples from adult and pediatric patient cohorts. Motivated by the lack of personalized therapeutic approaches in CD we aimed to discover novel markers of CD subtypes. With the ability to detect miRNA expression using blood samples, future studies can build on our findings to identify markers of disease behavior within a clinical setting without the need for invasive procedures. The combination of bioinformatic, experimental, and clinical expertise within our group allowed us to find potential microRNAs (miRs) of interest through bioinformatic analyses, experimentally validating expression *in vitro* and *ex vivo*, and accessing patient phenotypes to evaluate the clinical utility of our findings.

By performing RNA-sequencing (RNA-seq) on a cohort of 12 non-IBD (NIBD) and 18 CD patient samples, our group discovered two distinct molecular subtypes of CD that exhibited differential gene expression profiles associated with normal colon-like (CL) and ileum-like (IL) expression patterns. Using the same cohort of patient samples, I showed that miRNA expression profiles stratified CD into the same distinct molecular subtypes. Additionally, through the reanalysis of our RNA-sequencing (RNA-seq) data, we showed that long non-coding RNA (lncRNA) expression recapitulated CD subtype stratification. Differential analysis of miRNA expression between CL and IL subtypes revealed several significantly differential miRs. In particular, miR-31 exhibited a 13.5-fold change between CL and IL samples and was the only significantly enriched miR among genes downregulated in IL patients, suggesting a role as a candidate master regulator of downregulated genes within the IL subtype. Although only miR-31 was selected for further downstream analysis and was the only statistically significant miR identified through miRHub analysis, 3 additional miRs (miR-196b-5p, miR-194-1-5p, and miR-615-3p) were identified with at least a 4-fold increase in the IL subtype compared to CL CD. In isolation, these additional miRs do not appear to significantly target downregulated IL genes, but the cumulative effect of multiple miRs on gene expression pathways was not investigated. Due to the complex nature of CD pathogenesis, an interesting potential follow-up analysis would involve considering the contribution of other IL-upregulated miRs in supplementing miR-31 inhibition of gene expression pathways. In addition, future studies that assess the utility of miR-31 as a non-invasive indicator of CD subtypes through analysis of patient serum samples will benefit from a selection of multiple miRs. While miR-31 provides a clear distinction of CD subtypes in tissue samples, this may not be the case in serum where other miRs may provide better resolution for this separation.

Validation of miR-31 upregulation in IL CD was confirmed using an independent cohort of 40 CD and 29 NIBD patients, although proportionally fewer IL CD samples were identified compared to CL samples. Through isolation of colon-specific cell types, we found that intestinal epithelial cells (IECs) appeared to drive miR-31 upregulation in mucosal colonic tissue. Using *ex vivo* patient-derived colonoids, we further found that miR-31 expression in colonoids from CD patients was significantly higher compared with NIBD controls. Together, these results suggested that miR-31 upregulation in our colonic tissue data was driven by a cell type critical to the barrier function within the colonic mucosa and that IECs were

predisposed to high miR-31 expression in CD patients. These results served as the basis of a functional follow up study that investigated the role of miR-31 targets in disrupting the IEC barrier in CD (138). In this study, miR-31 was identified to target and suppress the expression of ALK1, increasing the stemness of colonic IECs through TGF-beta signaling. Through barrier permeability assays, we discovered that decreased colonic ALK1 expression resulted in disrupted IEC barrier integrity and was associated with an increased risk of surgery in CD patients (138). Together, the results of our experimental validation assays in chapter II combined with further mechanistic studies performed in (138) suggest that miR-31 upregulation within IEC populations in the colon function to increase the expression of stemness-related genes in-part through interactions with ALK1. High expression of miR-31, and therefore low ALK1 expression, results in increased barrier permeability associated with worse disease course in CD patients. Although these findings highlight the importance of miR-31 interactions with ALK1 within IECs in the colon, additional targets of miR-31 may still play critical roles in maintaining barrier integrity within the colonic crypt.

Using a pediatric cohort consisting of formalin-fixed paraffin-embedded (FFPE) tissue from 76 treatment-naïve CD patients and 51 NIBD controls, I discovered that miR-31 again exhibited significant upregulation in CD samples compared with NIBD controls. By considering patients that presented with inflammation at the time of diagnosis, we discovered that low levels of miR-31 in the ascending colon were associated with the development of ileal stricturing, a severe phenotype involving narrowing and obstruction of the gastrointestinal tract. The design of this cohort offered distinct advantages compared with our adult cohort, discussed further in chapter II. However, a limitation of this analysis was the lack of matching RNA-seq to facilitate CL and IL subtype stratification, although this was overcome through the utilization of ileal miR-31 expression from NIBD controls. RNA extracted from FFPE tissue is generally of lower quality compared with fresh frozen tissue as a result of the fixation procedure that modifies, cross-links and degrades RNA (178). Transcriptomic analyses on full-length RNAs is therefore difficult due to the quality of data that is often obtained (179). At this time the analyses were conducted, there was a lack of robust methods to extract RNA from FFPE tissue to produce high-quality data but these methods have recently improved providing an opportunity to revisit this sample cohort using addition high-throughput methods (180) (179). In addition, methods for high-quality chromatin extraction from FFPE

tissue are continually improving, providing an opportunity for additional studies into regulatory signatures using cohorts consisting of FFPE tissue (181) (182).

Paneth-like expression profiles further stratify colon-like CD into distinct molecular subtypes

By incorporating additional adult CD and NIBD samples, we increased our adult patient cohort in chapter III to gain further insights into CD subtypes and increase our power to investigate the heterogeneity within CD subtypes. Probing the transcriptome of colonic CD samples through RNA-seq allowed us to identify altered gene expression pathways within the colonic mucosa and deconvolution analyses confirmed a largely IEC-driven expression signature within our samples. After first correcting for variation in sample quality and predicted IEC proportions across our dataset, we increased our confidence in determining differential gene sets between CL and IL subtypes. Our increased sample size resulted in an 8.5-fold increase in the number of significant differentially expressed genes detected, with pathway enrichment analyses revealing a large degree of overlap with our initial molecular subtyping study (82). Consistent with experimental validation studies conducted in chapter II, we detected proportionally fewer IL CD samples compared with CL samples. An important future direction for this project will involve using larger IL cohorts to potentially refine the IL subtype further and to understand cell type-specific mechanisms driving the ileal expression signature we observe in this subset of samples. By restricting patients in a future study cohort to those exhibiting phenotypes associated with IL patients in our previous studies, such as an absence of rectal disease (82) and surgery without the need for an end ileostomy (83), we would be better poised to identify larger numbers of IL patients allowing us to more finely study IL CD.

The remainder of chapter III centered on investigating the heterogeneity across CL samples. Through consensus clustering analyses, we discovered two distinct subgroups there were associated with differential small intestinal gene expression signatures. Further investigation revealed distinct upregulation of several antimicrobial peptide (AMP) genes within one subgroup, specifically reg and defensin-α family genes, which initially suggested metaplastic Paneth cell differential within colonic crypts of a subset of CL CD patients, consistent with several previous studies (128) (126) (127). Based on these expression profiles, novel subgroups of CL CD were therefore referred to as Paneth enriched (PE) or

Paneth depleted (PD) samples. However, due to an absence of well-established Paneth cell markers, such as lysozyme (LYZ), we hypothesized that this Paneth profile was associated with the differentiation of cells with the capacity to secrete AMPs in response to intestinal inflammation. Similar to Paneth cells in the small intestinal and Paneth-like cells discovered previously within the colon (157), we expect Paneth-like cells discovered within PE CD patients to be interspaced between stem cell populations at the base of colonic crypts. A vital next step for our group will involve experimental validation of AMP presence within the colonic crypt of PE samples through immunohistochemistry and *ex vivo* assays. As well as AMP presence suggesting an increased antimicrobial function of this undefined cell type, the position of these cells relative to stem cell populations within the crypt will offer additional clues to their potential function. While missing traditional Paneth cell markers, I hypothesize that these AMP marked cells function to protect stem cell populations due to the breakdown of the colonic mucus layer that acts as the main antimicrobial barrier in healthy individuals.

Although additional work remains to accurately define our novel expression signature within the PE subset of CL patients, our findings provide further evidence of distinct molecular subtypes within CD. Through further longitudinal clinical association analyses of PE and PD patients, we will be able to assess the clinical utility of our newly defined subclasses. Future studies will benefit from single-cell assays to determine the presence of Paneth expression profiles within CL samples and accurately pinpoint the cell populations expressing AMPs. Through comparisons with single-cell data generated in other colonic disease contexts (158), we will better understand the disease-specific effects of Paneth-like signatures in the colon.

Future directions and closing thoughts

In chapters II and III, we discovered variation across CD samples attributed to non-coding RNA (ncRNA) regulation through miRNAs and lncRNAs, but there is still plenty of room for exploration in the contribution of ncRNAs in establishing CD subtypes. LncRNAs are now emerging as important functional regulators of a range of diverse biological functions (183) (184) and are now appreciated as critical regulators that play a contributing role in disease development (185) (186). In the IBD field, various studies have indicated that lncRNAs are critical to the pathogenesis of IBDs (187) (188). In IL and our two

distinct CL subtypes, the role of lncRNAs in driving separation of the subtypes along with their mechanistic roles within IEC population warrants further attention due to the promise of lncRNAs as therapeutic targets (189) (190) (191) and functions in molecular disease mechanisms across various disorders. Through analysis of smRNA-seq, a novel class of small ncRNAs called tRNA-derived RNAs (tDRs) can be robustly detected (192). tDRs have recently generated excitement due to their association with human disease (193) (192) and detection within serum samples, facilitating their use as disease biomarkers (194). To date, no papers have been published with a primary focus of tDRs within IBD, although a previous study by our group suggests that tDRs are detectable and variable among CD patients. Using the smRNA-seq data generated in chapter II, we found that genome-wide tDR expression did not separate adult CD patients in molecular subtypes (data not shown) but this was not investigated further in our pediatric patient cohort. With evidence in other disease fields suggesting important roles for tDRs in disease development, it would be interesting to revisit these non-coding regulators in the context of CD.

In the cancer field, integrating molecular information across various high-throughput molecular assays has become essential in characterizing consensus subtypes as well as facilitating a greater biological understanding of the molecular drivers underpinning disease subtypes (96). In our studies, employing additional high-throughput assays to gain insights into gene regulation will help us develop a more accurate picture of the complex regulatory network that underlie subtype-specific gene expression patterns. Our group has performed assay for transposase-accessible chromatin with high-throughput sequencing (ATAC-seq) on most of the adult samples used in chapter III, allowing us to study genome-wide changes in chromatin accessibility within CD and between CD subtypes. By integrating our findings from analyses of ATAC-seq data with transcriptomic data through RNA-seq and smRNA-seq, we will identify subtype-specific dysregulation at sites through the genome that correlates with subtype-specific transcriptomic profiles. Although analysis is currently underway, this data will supplement our RNA-seq data and provide novel regulatory insights into Paneth-driven molecular signatures in CL subtype samples. In addition, we have also performed genotyping of matched RNA-seq and ATAC-seq samples for quantitative trait locus (QTL) and allelic imbalance analyses to further map the genetic variation that underlies CD subtypes.

In conclusion, there remains active, substantive interest in the CD research community to identify molecular factors that distinguish subtypes of CD to develop more accurate diagnostic methods and identify more effective therapeutic strategies. By utilizing high-through molecular assays, we have identified reproducible signatures that are reflective of the underlying molecular biology driving CD heterogeneity and disease presentation. Through additional experimental validation and ongoing analyses of the regulatory signatures that underpin the findings discussed in the dissertation, our understanding of CD subtypes will continue to develop in the coming years. Although there is still much work to be done before CD subtypes become a clinical utility, this research demonstrates the value of molecular subtyping approaches in CD and provides another step towards the identification of novel diagnostic and prognostic indicators of disease.

REFERENCES

1. Feingold EA et al. The ENCODE (ENCyclopedia of DNA Elements) Project. *Science (80-.).* 2004;306(5696):636–640.

2. Bernstein BE et al. The NIH Roadmap Epigenomics Mapping Consortium. *Nat. Biotechnol.* 2010;28(10):1045–1048.

3. Carninci P et al. Genome-wide analysis of mammalian promoter architecture and evolution. *Nat. Genet.* 2006;38(6):626–635.

4. Roadmap Epigenomics Consortium et al. Integrative analysis of 111 reference human epigenomes. *Nature* 2015;518(7539):317–329.

5. The ENCODE Project Consortium et al. An integrated encyclopedia of DNA elements in the human genome.. *Nature* 2012;489(7414):57–74.

6. Lee TI, Young RA. Transcriptional regulation and its misregulation in disease. *Cell* 2013;152(6):1237–1251.

7. Bartel DP. Metazoan MicroRNAs.. *Cell* 2018;173(1):20–51.

8. Liu X, Fortin K, Mourelatos Z. MicroRNAs: Biogenesis and molecular functions. *Brain Pathol.* 2008;18(1):113–121.

9. Bartel DP. MicroRNAs: Target Recognition and Regulatory Functions. *Cell* 2009;136(2):215–233.

10. Lee RC, Feinbaum RL, Ambros V. The C. elegans heterochronic gene lin-4 encodes small RNAs with antisense complementarity to lin-14. *Cell* 1993;75(5):843–854.

11. Wightman B, Ha I, Ruvkun G. Posttranscriptional regulation of the heterochronic gene lin-14 by lin-4 mediates temporal pattern formation in C. elegans. *Cell* 1993;75(5):855–862.

12. Griffiths-Jones S, Saini HK, Van Dongen S, Enright AJ. miRBase: Tools for microRNA genomics. *Nucleic Acids Res.* 2008;36(SUPPL. 1). doi:10.1093/nar/gkm952

13. Gebert LFR, MacRae IJ. Regulation of microRNA function in animals. *Nat. Rev. Mol. Cell Biol.* 2019;20(1):21–37.

14. Bhajun R, Guyon L, Gidrol X. MicroRNA degeneracy and pluripotentiality within a Lavallière-tie architecture confers robustness to gene expression networks. *Cell. Mol. Life Sci.* 2016;73(15):2821–2827.

15. Ivey KN, Srivastava D. microRNAs as developmental regulators. *Cold Spring Harb. Perspect. Biol.* 2015;7(7):1–9.

16. Friedman RC, Farh KKH, Burge CB, Bartel DP. Most mammalian mRNAs are conserved targets of microRNAs. *Genome Res.* 2009;19(1):92–105.

17. Boon RA, Vickers KC. Intercellular transport of MicroRNAs. *Arterioscler. Thromb. Vasc. Biol.* 2013;33(2):186–192.

18. Ardekani AM, Naeini MM. The role of microRNAs in human diseases. *Avicenna J. Med. Biotechnol.* 2010;2(4):161–179.

19. Li Y, Kowdley K V. MicroRNAs in Common Human Diseases. *Genomics, Proteomics Bioinforma.*

2012;10(5):246–253.

20. Drury RE, O'Connor D, Pollard AJ. The clinical application of MicroRNAs in infectious disease. *Front. Immunol.* 2017;8(SEP). doi:10.3389/fimmu.2017.01182

21. Kristen A V et al. Patisiran, an RNAi therapeutic for the treatment of hereditary transthyretin-mediated amyloidosis. *Neurodegener. Dis. Manag.* 2019;9(1):5–23.

22. Hanna J, Hossain GS, Kocerha J. The Potential for microRNA Therapeutics and Clinical Research. *Front. Genet.* 2019;10(MAY):478.

23. De Guire V et al. Circulating miRNAs as sensitive and specific biomarkers for the diagnosis and monitoring of human diseases: Promises and challenges. *Clin. Biochem.* 2013;46(10–11):846–860.
24. Goodwin S, McPherson JD, McCombie WR. Coming of age: Ten years of next-generation sequencing technologies. *Nat. Rev. Genet.* 2016;17(6):333–351.

25. Wang Z, Gerstein M, Snyder M. RNA-Seq: a revolutionary tool for transcriptomics.. *Nat. Rev. Genet.* 2009;10(1):57–63.

26. Vickers KC, Roteta LA, Hucheson-Dilks H, Han L, Guo Y. Mining diverse small RNA species in the deep transcriptome.. *Trends Biochem. Sci.* 2015;40(1):4–7.

27. Lowe R, Shirley N, Bleackley M, Dolan S, Shafee T. Transcriptomics technologies. *PLoS Comput. Biol.* 2017;13(5). doi:10.1371/journal.pcbi.1005457

28. Giraldez MD et al. Comprehensive multi-center assessment of small RNA-seq methods for quantitative miRNA profiling. *Nat. Biotechnol.* 2018;36(8):746–757.

29. Huang S, Chaudhary K, Garmire LX. More Is Better: Recent Progress in Multi-Omics Data Integration Methods. *Front. Genet.* 2017;8:84.

30. Lightbody G et al. Review of applications of high-throughput sequencing in personalized medicine: barriers and facilitators of future progress in research and clinical application. *Brief. Bioinform.* [published online ahead of print: July 31, 2018]; doi:10.1093/bib/bby051

31. Ziegler A, Koch A, Krockenberger K, Großhennig A. Personalized medicine using DNA biomarkers: A review. *Hum. Genet.* 2012;131(10):1627–1638.

32. Xi X et al. RNA biomarkers: Frontier of precision medicine for cancer. *Non-coding RNA* 2017;3(1). doi:10.3390/ncrna3010009

33. Mueller C, Haymond A, Davis JB, Williams A, Espina V. Protein biomarkers for subtyping breast cancer and implications for future research. *Expert Rev. Proteomics* 2018;15(2):131–152.

34. Le Tourneau C et al. Treatment Algorithms Based on Tumor Molecular Profiling: The Essence of Precision Medicine Trials. *J. Natl. Cancer Inst.* 2016;108(4):djv362.

35. Strimbu K, Tavel JA. What are biomarkers?. *Curr. Opin. HIV AIDS* 2010;5(6):463–466.

36. Henry NL, Hayes DF. Cancer biomarkers. *Mol. Oncol.* 2012;6(2):140–146.

37. Zhao L, Lee VHF, Ng MK, Yan H, Bijlsma MF. Molecular subtyping of cancer: Current status and moving toward clinical applications. *Brief. Bioinform.* 2019;20(2):572–584.

38. Furey TS, Sethupathy P, Sheikh SZ. Redefining the IBDs using genome-scale molecular phenotyping. *Nat. Rev. Gastroenterol. Hepatol.* 2019;1.

39. Tomczak K, Czerwińska P, Wiznerowicz M. Review The Cancer Genome Atlas (TCGA): an immeasurable source of knowledge. *Współczesna Onkol.* 2015;1A:68–77.

40. Perou CM et al. Molecular portraits of human breast tumours. *Nature* 2000;406(6797):747–752.

41. Cancer Genome Atlas Network TCGA. Comprehensive molecular portraits of human breast tumours.. *Nature* 2012;490(7418):61–70.

42. Wang W et al. Molecular subtyping of colorectal cancer: Recent progress, new challenges and emerging opportunities. *Semin. Cancer Biol.* 2019;55(May 2018):37–52.

43. Garber ME et al. Diversity of gene expression in adenocarcinoma of the lung. *Proc. Natl. Acad. Sci. U. S. A.* 2001;98(24):13784–13789.

44. Collisson EA et al. Comprehensive molecular profiling of lung adenocarcinoma. *Nature* 2014;511(7511):543–550.

45. Figueroa ME et al. DNA Methylation Signatures Identify Biologically Distinct Subtypes in Acute Myeloid Leukemia. *Cancer Cell* 2010;17(1):13–27.

46. Cancer Genome Atlas Research Network TCGAR. Comprehensive molecular characterization of clear cell renal cell carcinoma.. *Nature* 2013;499(7456):43–9.

47. Cancer Genome Atlas Research Network TCGAR. Comprehensive molecular characterization of urothelial bladder carcinoma.. *Nature* 2014;507(7492):315–22.

48. Cancer Genome Atlas Research Network WM et al. Comprehensive Molecular Characterization of Papillary Renal-Cell Carcinoma.. *N. Engl. J. Med.* 2016;374(2):135–45.

49. Getz G et al. Integrated genomic characterization of endometrial carcinoma. *Nature* 2013;497(7447):67–73.

50. Kalla R, Ventham NT, Satsangi J, Arnott IDR. Crohn's disease. *Bmj* 2014;349(nov19 13):g6670–g6670.

51. Dahlhamer JM, Zammitti EP, Ward BW, Wheaton AG, Croft JB. Prevalence of inflammatory bowel disease among adults aged ≥18 years — United States, 2015. *Morb. Mortal. Wkly. Rep.* 2016;65(42):1166–1169.

52. Malik TA. Inflammatory Bowel Disease. *Surg. Clin. North Am.* 2015;95(6):1105–1122.
53. Gajendran M, Loganathan P, Catinella AP, Hashash JG. A comprehensive review and update on Crohn's disease. *Disease-a-Month* 2018;64(2):20–57.

54. Mowat C et al. Guidelines for the management of inflammatory bowel disease in adults. *Gut* 2011;60(5):571–607.

55. Ananthakrishnan AN. Epidemiology and risk factors for IBD. *Nat. Rev. Gastroenterol. Hepatol.* 2015;12(4):205–217.

56. Yamazaki K et al. Single nucleotide polymorphisms in TNFSF15 confer susceptibility to Crohn's disease. *Hum. Mol. Genet.* 2005;14(22):3499–3506.

57. Mirkov MU, Verstockt B, Cleynen I. Genetics of inflammatory bowel disease: beyond NOD2. *Lancet Gastroenterol. Hepatol.* 2017;2(3):224–234.

58. Gordon H, Trier Moller F, Andersen V, Harbord M. Heritability in inflammatory bowel disease: from the

first twin study to genome-wide association studies.. *Inflamm. Bowel Dis.* 2015;21(6):1428–34.

59. Cuthbert AP et al. The contribution of NOD2 gene mutations to the risk and site of disease in inflammatory bowel disease.. *Gastroenterology* 2002;122(4):867–74.

60. Duerr RH et al. A genome-wide association study identifies IL23R as an inflammatory bowel disease gene.. *Science* 2006;314(5804):1461–3.

61. Silverberg MS et al. A population- and family-based study of Canadian families reveals association of HLA DRB1*0103 with colonic involvement in inflammatory bowel disease.. *Inflamm. Bowel Dis.* 2003;9(1):1–9.

62. Newman B et al. CARD15 and HLA DRB1 alleles influence susceptibility and disease localization in Crohn's disease.. *Am. J. Gastroenterol.* 2004;99(2):306–15.

63. Meddens CA, Van Der List ACJ, Nieuwenhuis EES, Mokry M. Non-coding DNA in IBD: From sequence variation in DNA regulatory elements to novel therapeutic potential. *Gut* [published online ahead of print: 2019]; doi:10.1136/gutjnl-2018-317516

64. Alexandru V. Olaru, MD1, Florin M. Selaru, MD1, Yuriko Mori MP. Dynamic changes in the expression of microRNA-31 during inflammatory bowel disease-associated neoplastic transformation2012;100(2):130–134.

65. Schaefer JS et al. MicroRNA signatures differentiate Crohn's disease from ulcerative colitis.. *BMC Immunol.* 2015;16:5.

66. Wu F et al. Peripheral blood microRNAs distinguish active ulcerative colitis and Crohn's disease.. *Inflamm. Bowel Dis.* 2011;17(1):241–50.

67. Wang H et al. Circulating MicroRNA223 is a New Biomarker for Inflammatory Bowel Disease. *Medicine (Baltimore).* 2016;95(5):e2703.

68. Béres NJ et al. Role of Altered Expression of miR-146a, miR-155, and miR-122 in Pediatric Patients with Inflammatory Bowel Disease. *Inflamm. Bowel Dis.* 2016;22(2):327–335.

69. McKenna LB et al. MicroRNAs control intestinal epithelial differentiation, architecture, and barrier function.. *Gastroenterology* 2010;139(5):1654–64, 1664.e1.

70. Cichon C, Sabharwal H, Rüter C, Schmidt MA. MicroRNAs regulate tight junction proteins and modulate epithelial/endothelial barrier functions. *Tissue Barriers* 2014;2(4):e944446.

71. Xu XM, Zhang HJ. MiRNAs as new molecular insights into inflammatory bowel disease: Crucial regulators in autoimmunity and inflammation. *World J. Gastroenterol.* 2016;22(7):2206–2218.

72. Mirza AH et al. Transcriptomic landscape of lncRNAs in inflammatory bowel disease. *Genome Med.* 2015;7(1):39.

73. Haberman Y et al. Long ncRNA Landscape in the Ileum of Treatment-Naive Early-Onset Crohn Disease. *Inflamm. Bowel Dis.* 2018;24(2):346–360.

74. Cleynen I et al. Inherited determinants of Crohn's disease and ulcerative colitis phenotypes: A genetic association study. *Lancet* 2016;387(10014):156–167.

75. Haberman Y et al. Pediatric Crohn disease patients exhibit specific ileal transcriptome and microbiome signature. *J. Clin. Invest.* 2014;124(8):3617–3633.

76. Han L et al. A probabilistic pathway score (PROPS) for classification with applications to inflammatory bowel disease. *Bioinformatics* 2018;34(6):985–993.

77. Boyd M et al. Characterization of the enhancer and promoter landscape of inflammatory bowel disease from human colon biopsies.. *Nat. Commun.* 2018;9(1):1661.

78. Mokry M et al. Many inflammatory bowel disease risk loci include regions that regulate gene expression in immune cells and the intestinal epithelium. *Gastroenterology* 2014;146(4):1040–1047.

79. Peck BCE et al. MicroRNAs Classify Different Disease Behavior Phenotypes of Crohn's Disease and May Have Prognostic Utility.. *Inflamm. Bowel Dis.* 2015;21(9):2178–87.

80. Béres NJ et al. Altered mucosal expression of microRNAs in pediatric patients with inflammatory bowel disease. *Dig. Liver Dis.* [published online ahead of print: 2016];(2016). doi:10.1016/j.dld.2016.12.022

81. Satsangi J, Silverberg MS, Vermeire S, Colombel J-F. The Montreal classification of inflammatory bowel disease: controversies, consensus, and implications. *Gut* 2006;55(6):749–753.

82. Weiser M et al. Molecular classification of Crohn's disease reveals two clinically relevant subtypes. *Gut* 2016;gutjnl-2016-312518.

83. Keith BP et al. Colonic epithelial miR-31 associates with the development of Crohn's phenotypes. *JCI Insight* 2018;3(19). doi:10.1172/JCI.INSIGHT.122788

84. Jostins L et al. Host-microbe interactions have shaped the genetic architecture of inflammatory bowel disease.. *Nature* 2012;491(7422):119–24.

85. Kugathasan S et al. Prediction of complicated disease course for children newly diagnosed with Crohn's disease: a multicentre inception cohort study. *Lancet* [published online ahead of print: March 2017]; doi:10.1016/S0140-6736(17)30317-3

86. Peck BCE et al. Functional Transcriptomics in Diverse Intestinal Epithelial Cell Types Reveals Robust MicroRNA Sensitivity in Intestinal Stem Cells to Microbial Status.. *J. Biol. Chem.* 2017;292(7):2586–2600.

87. Chapman CG, Pekow J. The emerging role of miRNAs in inflammatory bowel disease: a review. *Therap. Adv. Gastroenterol.* 2015;8(1):4–22.

88. Pekow JR, Kwon JH. MicroRNAs in inflammatory bowel disease.. *Inflamm. Bowel Dis.* 2012;18(1):187–93.

89. Lin J et al. Novel specific microRNA biomarkers in idiopathic inflammatory bowel disease unrelated to disease activity.. *Mod. Pathol.* 2014;27(4):602–8.

90. Kim D et al. General rules for functional microRNA targeting. *Nat. Genet.* [published online ahead of print: October 24, 2016]; doi:10.1038/ng.3694

91. Baran-Gale J, Fannin EE, Kurtz CL, Sethupathy P. Beta Cell 5'-Shifted isomiRs Are Candidate Regulatory Hubs in Type 2 Diabetes. *PLoS One* 2013;8(9). doi:10.1371/journal.pone.0073240

92. Uhlen M et al. Tissue-based map of the human proteome. *Science (80-.).* 2015;347(6220):1260419–1260419.

93. Comelli EM et al. Biomarkers of human gastrointestinal tract regions. *Mamm. Genome* 2009;20(8):516–527.

94. Rutgeerts P et al. Predictability of the postoperative course of Crohn's disease.. *Gastroenterology* 1990;99(4):956–63.

95. Zachos NC et al. Human Enteroids/Colonoids and Intestinal Organoids Functionally Recapitulate Normal Intestinal Physiology and Pathophysiology.. *J. Biol. Chem.* 2016;291(8):3759–66.

96. Tomczak K, Czerwińska P, Wiznerowicz M. The Cancer Genome Atlas (TCGA): an immeasurable source of knowledge.. *Contemp. Oncol. (Poznan, Poland)* 2015;19(1A):A68-77.

97. Lenz G et al. Molecular subtypes of diffuse large B-cell lymphoma arise by distinct genetic pathways.. *Proc. Natl. Acad. Sci. U. S. A.* 2008;105(36):13520–5.

98. Verhaak RGW et al. Integrated genomic analysis identifies clinically relevant subtypes of glioblastoma characterized by abnormalities in PDGFRA, IDH1, EGFR, and NF1.. *Cancer Cell* 2010;17(1):98–110.

99. Hoadley KA et al. Tumor Evolution in Two Patients with Basal-like Breast Cancer: A Retrospective Genomics Study of Multiple Metastases. *PLOS Med.* 2016;13(12):e1002174.

100. Glocker E-O et al. Inflammatory bowel disease and mutations affecting the interleukin-10 receptor.. *N. Engl. J. Med.* 2009;361(21):2033–45.

101. Ostrowski J et al. Genetic architecture differences between pediatric and adult-onset inflammatory bowel diseases in the Polish population. *Sci. Rep.* 2016;6(1):39831.

102. Jakobsen C et al. Differences in phenotype and disease course in adult and paediatric inflammatory bowel disease - a population-based study. *Aliment. Pharmacol. Ther.* 2011;34(10):1217–1224.

103. Duricova D et al. Age-related differences in presentation and course of inflammatory bowel disease: an update on the population-based literature. *J. Crohn's Colitis* 2014;8(11):1351–1361.

104. De Greef E et al. Diagnosing and treating pediatric Crohn's disease patients: is there a difference between adult and pediatric gastroenterologist's practices ? Results of the BELCRO cohort.. *Acta Gastroenterol. Belg.* 2014;77(1):25–9.

105. Rosen MJ et al. Mucosal Expression of Type 2 and Type 17 Immune Response Genes Distinguishes Ulcerative Colitis From Colon-Only Crohn?s Disease in Treatment-Naive Pediatric Patients. *Gastroenterology* 2017;152(6):1345-1357.e7.

106. Jovov B, Shaheen NJ, Orlando GS, Djukic Z, Orlando RC. Defective barrier function in neosquamous epithelium.. *Am. J. Gastroenterol.* 2013;108(3):386–91.

107. Sreedharan L et al. MicroRNA profile in neosquamous esophageal mucosa following ablation of Barrett's esophagus.. *World J. Gastroenterol.* 2017;23(30):5508–5518.

108. Lin W-B et al. MicroRNA profiling of the intestine during hypothermic circulatory arrest in swine.. *World J. Gastroenterol.* 2015;21(7):2183–90.

109. Gao W et al. A systematic analysis of predicted MiR-31-targets identifies a diagnostic and prognostic signature for lung cancer. *Biomed. Pharmacother.* 2014;68(4):419–427.

110. Tian Y et al. Stress responsive miR-31 is a major modulator of mouse intestinal stem cells during regeneration and tumorigenesis. *Elife* 2017;1–30.

111. Dobin A et al. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* 2013;29(1):15–21.

112. Patro R, Duggal G, Love MI, Irizarry RA, Kingsford C. Salmon provides fast and bias-aware

quantification of transcript expression. *Nat. Methods* 2017;14(4):417–419.

113. Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2.. *Genome Biol.* 2014;15(12):550.

114. Kanke M, Baran-Gale J, Villanueva J, Sethupathy P. miRquant 2.0: an Expanded Tool for Accurate Annotation and Quantification of MicroRNAs and their isomiRs from Small RNA-Sequencing Data doi:10.2390/biecoll-jib-2016-307

115. Kamada N et al. Unique CD14 intestinal macrophages contribute to the pathogenesis of Crohn disease via IL-23/IFN-gamma axis.. *J. Clin. Invest.* 2008;118(6):2269–80.

116. Collisson EA, Bailey P, Chang DK, Biankin A V. Molecular subtypes of pancreatic cancer. *Nat. Rev. Gastroenterol. Hepatol.* 2019;16(4):207–220.

117. Collisson EA et al. Subtypes of pancreatic ductal adenocarcinoma and their differing responses to therapy. *Nat. Med.* 2011;17(4):500–503.

118. Bailey P et al. Genomic analyses identify molecular subtypes of pancreatic cancer. *Nature* 2016;531(7592):47–52.

119. Veenstra VL, Garcia-Garijo A, Van Laarhoven HW, Bijlsma MF. Extracellular influences: Molecular subclasses and the microenvironment in pancreatic cancer. *Cancers (Basel).* 2018;10(2). doi:10.3390/cancers10020034

120. Dai X et al. Breast cancer intrinsic subtype classification, clinical use and future trends. *Am. J. Cancer Res.* 2015;5(10):2929–2943.

121. Okumura R, Takeda K. Roles of intestinal epithelial cells in the maintenance of gut homeostasis. *Exp. Mol. Med.* 2017;49(5):e338–e338.

122. Martini E, Krug SM, Siegmund B, Neurath MF, Becker C. Mend Your Fences: The Epithelial Barrier and its Relationship With Mucosal Immunity in Inflammatory Bowel Disease. *CMGH* 2017;4(1):33–46.

123. Soderholm AT, Pedicord VA. Intestinal epithelial cells: at the interface of the microbiota and mucosal immunity. *Immunology* 2019;158(4):267–280.

124. Gassler N. Paneth cells in intestinal physiology and pathophysiology. *World J. Gastrointest. Pathophysiol.* 2017;8(4):150–160.

125. Schroeder BO. Fight them or feed them: how the intestinal mucus layer manages the gut microbiota.. *Gastroenterol. Rep.* 2019;7(1):3–12.

126. Cunliffe RN, Mahida YR. Expression and regulation of antimicrobial peptides in the gastrointestinal tract. *J. Leukoc. Biol.* 2004;75(1):49–58.

127. Perminow G et al. Defective paneth cellmediated host defense in pediatric ileal crohn's disease. *Am. J. Gastroenterol.* 2010;105(2):452–459.

128. van Beelen Granlund A et al. REG gene expression in inflamed and healthy colon mucosa explored by in situ hybridisation. *Cell Tissue Res.* 2013;352(3):639–646.

129. Haberman Y et al. Ulcerative colitis mucosal transcriptomes reveal mitochondriopathy and personalized mechanisms underlying disease severity and treatment response. *Nat. Commun.* 2019;10(1):1–13.

130. Shivashankar R, Tremaine WJ, Harmsen WS, Loftus E V. Incidence and Prevalence of Crohn's Disease and Ulcerative Colitis in Olmsted County, Minnesota From 1970 Through 2010. *Clin. Gastroenterol. Hepatol.* 2017;15(6):857–863.

131. Dumic I et al. Gastrointestinal tract disorders in older age. *Can. J. Gastroenterol. Hepatol.* 2019;2019. doi:10.1155/2019/6757524

132. Lonsdale J et al. The Genotype-Tissue Expression (GTEx) project. *Nat. Genet.* 2013;45(6):580–585.

133. Monti S, Tamayo P, Mesirov J, Golub T. Consensus clustering: A resampling-based method for class discovery and visualization of gene expression microarray data. *Mach. Learn.* 2003;52(1–2):91–118.

134. Oyelade J et al. Clustering algorithms: Their application to gene expression data. *Bioinform. Biol. Insights* 2016;10:237–253.

135. Şenbabaoğlu Y, Michailidis G, Li JZ. Critical limitations of consensus clustering in class discovery. *Sci. Rep.* 2014;4(1):1–13.

136. Rouillard AD et al. The harmonizome: a collection of processed datasets gathered to serve and mine knowledge about genes and proteins. *Database* 2016;2016. doi:10.1093/database/baw100

137. Hata A, Chen YG. TGF-β signaling from receptors to smads. *Cold Spring Harb. Perspect. Biol.* 2016;8(9):a022061.

138. Toyonaga T et al. Decreased Colonic Activin Receptor Like Kinase 1 Disrupts Epithelial Barrier integrity and is associated with a poor clinical outcome in Crohns disease. *bioRxiv* 2020;2020.02.21.960070.

139. Parikh K et al. Colonic epithelial cell diversity in health and inflammatory bowel disease. *Nature* 2019;567(7746):49–55.

140. Zhang Q et al. Commensal bacteria direct selective cargo sorting to promote symbiosis. *Nat. Immunol.* 2015;16(9):918–926.

141. Wang H et al. Rip2 Is Required for Nod2-Mediated Lysozyme Sorting in Paneth Cells. *J. Immunol.* 2017;198(9):3729–3736.

142. Ragland SA, Criss AK. From bacterial killing to immune modulation: Recent insights into the functions of lysozyme. *PLoS Pathog.* 2017;13(9):e1006512.

143. Holly MK, Smith JG. Paneth cells during viral infection and pathogenesis. *Viruses* 2018;10(5):225.

144. Zhang YW, Ding LS, Lai M De. Reg gene family and human diseases. *World J. Gastroenterol.* 2003;9(12):2635–2641.

145. Sankaran-Walters S, Hart R, Dills C. Guardians of the gut: Enteric defensins. *Front. Microbiol.* 2017;8(APR):647.

146. Madsen J, Mollenhauer J, Holmskov U. Gp-340/DMBT1 in mucosal innate immunity. *Innate Immun.* 2010;16(3):160–167.

147. Chen Z, Downing S, Tzanakakis ES. Four Decades After the Discovery of Regenerating Islet-Derived (Reg) Proteins: Current Understanding and Challenges. *Front. Cell Dev. Biol.* 2019;7:235.

148. Van Beelen Granlund A et al. Activation of REG family proteins in colitis. *Scand. J. Gastroenterol.*

2011;46(11):1316–1323.

149. Ramasundara M, Leach ST, Lemberg DA, Day AS. Defensins and inflammation: The role of defensins in inflammatory bowel disease. *J. Gastroenterol. Hepatol.* 2009;24(2):202–208.

150. Coretti L et al. The Interplay between Defensins and Microbiota in Crohn's Disease. *Mediators Inflamm.* 2017;2017. doi:10.1155/2017/8392523

151. Mollenhauer J et al. DMBT1 encodes a protein involved in the immune defense and in epithelial differentiation and is highly unstable in cancer. *Cancer Res.* 2000;60(6):1704–1710.

152. Braidotti P et al. DMBT1 expression is down-regulated in breast cancer. *BMC Cancer* 2004;4:46.

153. Rosenstiel P et al. Regulation of DMBT1 via NOD2 and TLR4 in Intestinal Epithelial Cells Modulates Bacterial Recognition and Invasion . *J. Immunol.* 2007;178(12):8203–8211.

154. Renner M et al. DMBT1 Confers Mucosal Protection In Vivo and a Deletion Variant Is Associated With Crohn's Disease. *Gastroenterology* 2007;133(5):1499–1509.

155. Diegelmann J et al. Intestinal DMBT1 Expression Is Modulated by Crohn's Disease-Associated IL23R Variants and by a DMBT1 Variant Which Influences Binding of the Transcription Factors CREB1 and ATF-2. *PLoS One* 2013;8(11):e77773.

156. Mantani Y et al. Ultrastructural and histochemical study on the paneth cells in the rat ascending colon. *Anat. Rec.* 2014;297(8):1462–1471.

157. Rothenberg ME et al. Identification of a cKit+ colonic crypt base secretory cell that supports Lgr5+ stem cells in mice. *Gastroenterology* 2012;142(5). doi:10.1053/j.gastro.2012.02.006

158. Wang Y et al. Single-cell transcriptome analysis reveals differential nutrient absorption functions in human intestine. *J. Exp. Med.* 2019;jem.20191130.

159. Liu R et al. Constitutive STAT5 activation regulates Paneth and Paneth-like cells to control Clostridium difficile colitis. *Life Sci. Alliance* 2019;2(2). doi:10.26508/lsa.201900296

160. Fahlgren A, Hammarström S, Danielsson Å. Increased expression of antimicrobial peptides and lysozyme in colonic epithelial cells of patients with ulcerative colitis. *Clin. Exp. Immunol.* 2003;131(1):90–101.

161. FASTX-Toolkithttp://hannonlab.cshl.edu/fastx_toolkit/index.html. cited January 31, 2020

162. Andrews S. FastQC A Quality Control tool for High Throughput Sequence Datahttps://www.bioinformatics.babraham.ac.uk/projects/fastqc/. cited February 7, 2020

163. Wang L, Wang S, Li W. RSeQC: Quality control of RNA-seq experiments. *Bioinformatics* 2012;28(16):2184–2185.

164. Ewels P, Magnusson M, Lundin S, Käller M. MultiQC: summarize analysis results for multiple tools and samples in a single report. *Bioinformatics* 2016;32(19):3047–3048.

165. Li B, Dewey CN. RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome.. *BMC Bioinformatics* 2011;12:323.

166. Wang L et al. Measure transcript integrity using RNA-seq data. *BMC Bioinformatics* 2016;17(1):58.

167. Linsley PS, Speake C, Whalen E, Chaussabel D. Copy number loss of the interferon gene cluster in

melanomas is linked to reduced T cell infiltrate and poor patient prognosis. *PLoS One* 2014;9(10). doi:10.1371/journal.pone.0109760

168. Ritchie ME et al. limma powers differential expression analyses for RNA-sequencing and microarray studies.. *Nucleic Acids Res.* 2015;43(7):e47.

169. Gu Z, Eils R, Schlesner M. Complex heatmaps reveal patterns and correlations in multidimensional genomic data. *Bioinformatics* 2016;32(18):2847–2849.

170. lighe K, Rana S LM. EnhancedVolcano: Publication-ready volcano plots with enhanced colouring and labeling2019;https://github.com/kevinblighe/EnhancedVolcano. cited February 8, 2020

171. Subramanian A et al. Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl. Acad. Sci. U. S. A.* 2005;102(43):15545–15550.

172. Liao Y, Wang J, Jaehnig EJ, Shi Z, Zhang B. WebGestalt 2019: gene set analysis toolkit with revamped UIs and APIs. *Nucleic Acids Res.* 2019;47(W1):W199–W205.

173. Chen EY et al. Enrichr: Interactive and collaborative HTML5 gene list enrichment analysis tool. *BMC Bioinformatics* 2013;14. doi:10.1186/1471-2105-14-128

174. Castrillo JI, Lista S, Hampel H, Ritchie CW. Systems Biology Methods for Alzheimer's Disease Research Toward Molecular Signatures, Subtypes, and Stages and Precision Medicine: Application in Cohort Studies and Trials. In: *Methods in Molecular Biology*. Humana Press Inc.; 2018:31–66

175. Wang C, Baer HM, Gaya DR, Nibbs RJB, Milling S. Can molecular stratification improve the treatment of inflammatory bowel disease?. *Pharmacol. Res.* 2019;148:104442.

176. Tam OH et al. Postmortem Cortex Samples Identify Distinct Molecular Subtypes of ALS: Retrotransposon Activation, Oxidative Stress, and Activated Glia. *Cell Rep.* 2019;29(5):1164-1177.e5.

177. Yin L, Chau CKL, Sham P-C, So H-C. Integrating Clinical Data and Imputed Transcriptome from GWAS to Uncover Complex Disease Subtypes: Applications in Psychiatry and Cardiology. *Am. J. Hum. Genet.* 2019;105(6):1193–1212.

178. Gaffney EF, Riegman PH, Grizzle WE, Watson PH. Factors that drive the increasing use of FFPE tissue in basic and translational cancer research. *Biotech. Histochem.* 2018;93(5):373–386.

179. Zhao Y et al. Robustness of RNA sequencing on older formalin-fixed paraffin-embedded tissue from high-grade ovarian serous adenocarcinomas. *PLoS One* 2019;14(5):e0216050.

180. Li J, Fu C, Speed TP, Wang W, Symmans WF. Accurate RNA Sequencing From Formalin-Fixed Cancer Tissue to Represent High-Quality Transcriptome From Frozen Tissue. *JCO Precis. Oncol.* 2018;2018(2):1–9.

181. Amatori S et al. Epigenomic profiling of archived FFPE tissues by enhanced PAT-ChIP (EPAT-ChIP) technology. *Clin. Epigenetics* 2018;10(1):143.

182. Zhong J et al. Enhanced and controlled chromatin extraction from FFPE tissues and the application to ChIP-seq. *BMC Genomics* 2019;20(1):249.

183. Kopp F, Mendell JT. Functional Classification and Experimental Dissection of Long Noncoding RNAs. *Cell* 2018;172(3):393–407.

184. Yao RW, Wang Y, Chen LL. Cellular functions of long noncoding RNAs. *Nat. Cell Biol.* 2019;21(5):542–551.

185. DiStefano JK. The Emerging Role of Long Noncoding RNAs in Human Disease. In: *Methods in Molecular Biology*. Humana Press Inc.; 2018:91–110

186. Zhang X, Hong R, Chen W, Xu M, Wang L. The role of long noncoding RNA in major human disease. *Bioorg. Chem.* 2019;92:103214.

187. Yarani R, Mirza AH, Kaur S, Pociot F. The emerging role of lncrnas in inflammatory bowel disease. *Exp. Mol. Med.* 2018;50(12). doi:10.1038/s12276-018-0188-9

188. Jabandziev P et al. The Emerging Role of Noncoding RNAs in Pediatric Inflammatory Bowel Disease. *Inflamm. Bowel Dis.* [published online ahead of print: February 3, 2020]; doi:10.1093/ibd/izaa009

189. Kumar MM, Goyal R. LncRNA as a Therapeutic Target for Angiogenesis. *Curr. Top. Med. Chem.* 2017;17(15):1750.

190. Zhou Q, Chen W, Yu X-Q. Long non-coding RNAs as novel diagnostic and therapeutic targets in kidney disease. *Chronic Dis. Transl. Med.* 2019;5(4):252–257.

191. Arun G, Diermeier SD, Spector DL. Therapeutic Targeting of Long Non-Coding RNAs in Cancer. *Trends Mol. Med.* 2018;24(3):257–277.

192. Jehn J, Rosenkranz D. tRNA-Derived Small RNAs: The Good, the Bad and the Ugly. *Med One* 2019;1–30.

193. Oberbauer V, Schaefer MR. tRNA-derived small RNAs: Biogenesis, modification, function and potential impact on human disease development. *Genes (Basel).* 2018;9(12). doi:10.3390/genes9120607

194. Balatti V, Pekarsky Y, Croce CM. Role of the tRNA-Derived Small RNAs in Cancer: New Potential Biomarkers and Target for Therapy. In: *Advances in Cancer Research*. Academic Press Inc.; 2017:173–187