S.T. Yau High School Science Award

Research Report

The Team

Name of team member: Stephanie Chang
School: Danbury Math Academy
City, Country: Danbury, United States of America

Name of team member: Adeethyia Shankar
School: Danbury Math Academy
City, Country: Danbury, United States of America

Name of supervising teacher: Xiaodi Wang
Job Title: Professor of Mathematics
School/Institution: Western Connecticut State University
City, Country: Danbury, Connecticut

Name of supervising teacher: Yongzhong Zhao
Job Title: Professor of Medical Science
School/Institution: Frontage Labs
City, Country: Exton, United States of America

Name of supervising teacher: Tong Liu
Job Title: PhD Student
School/Institution: Tsinghua University
City, Country: Beijing, China

Title of Research Report: A Wavelet-Based Approach Reveals Host-Cell-Type-Specific Multi-omics Networks in Inflammatory Bowel Disease

Date: August 20, 2023

*Names are listed in alphabetical order; both authors made equal contributions

A Wavelet-Based Approach Reveals Host-Cell-Type-Specific Multi-omics Networks in Inflammatory Bowel Disease

Stephanie Chang and Adeethyia Shankar

Abstract

The interplay between microbiome metabolites and human cells is crucial and mechanistically linked to human health and disease. Inflammatory bowel disease (IBD) is linked to microbiome metabolite and host gene expression. However, details of the microbiome and host interplays remain elusive. We carry out microbiome metabolome-wide and host transcriptome-wide association studies of IBD with microbiome metabolite targeted cell types discovery via leveraging the publicly available IBD data sets from Human Microbiome Project 2 (HMP2). By performing deconvolution on the transcriptomic data and applying discrete wavelet transform (DWT), we obtained cell type-metabolite correlations, which we visualize in the form of heatmaps and networks. We also carried out both targeted and untargeted approaches by mean of correlating the microbiome data matrix to host transcriptomic data. Given the limited sample size, in addition to visualizing a global picture of the interplay landscape between microbiome metabolites and host genes alongside distinct clusters of IBD and healthy controls with UMAP and t-SNE, we found a set of microbiome metabolites most likely linked to IBD and the transcriptomic signature of IBD. For the targeted approach, we also refer to the single-cell gene signature dataset, i.e., the MSigDB C8, uncovering a bile acid, namely, lithocholate, targeted cell types, including intestine lymphoid cells and enterocytes. Moreover, we utilized Mendelian Randomization causality tests with ursodeoxycholic acid (UDCA), RUNX1 gene, and IBD, resulting in a putative causality network of RUNX1, UDCA, and IBD. Taken together, our approaches shed light on the mechanistic interplay of microbiome metabolites and host cells in human health and disease.

Keywords: Microbiome, Discrete Wavelet Transform, Metabolite, Human Microbiome Project 2, Inflammatory bowel disease, Single-cell gene signature, Deconvolution, Cell type, Multi-omics, Mendelian Randomization, Causality

Highlights

Our application of discrete wavelet transform to deconvoluted cell proportion data sheds light on additional correlations between cell types and metabolites, resulting in fruitful correlation networks and an abundance of significant correlations among the wavelet components. Moreover, our Mendelian Randomization tests support the causal impact of ursodeoxycholic acid, a secondary bile acid, and RUNX1 gene expression, responsible for hematopoietic stem cell development, on IBD.

Analyzing the IBD datasets from the Human Microbiome Project 2 (HMP2) to analyze microbial correlation to IBD, our research found new metabolites potentially linked to IBD such as lithocholic acid. Our research also analyzed genetic correlations with IBD, discovering new potential genetic biomarkers for IBD such as the NOMO2 gene. Our analysis of the interplay between microbial and genetic material found that this multi-omics linkage to IBD achieved a correlation of over 90%. Our research also focused on identifying cell types corresponding with the metabolites and genes found previously. By identifying these cell types such as intestinal lymphoid cells and enterocytes, this discovery can lead to targeted, effective drug development to combat IBD.

Acknowledgments

We would like to acknowledge the contributions of each member.

Stephanie Chang performed pairwise correlations between observed y variables and normalized abundance x variables calculated based on Pearson correlations. She employed a mean-equality t-test based on any abundance variable between two samples: a sample of IBD patients and a control sample of non-IBD participants. She also tested the transcriptomes and metabolites, intersecting each subset of marker genes with the gene signatures that coincide using a hypergeometric distribution, and produced the associated figures. In addition, she applied state-of-the-art visualization techniques t-SNE and UMAP to further visualize the data. She also performed the literature review and generated the associated analyses. She authored these corresponding sections in this paper.

Adeethyia Shankar performed deconvolution on the host transcriptomics data and then subsequently applied discrete wavelet transform. In addition, he carried out the cell type-metabolite pairwise correlations, creating the corresponding heatmaps and networks. Finally, Adeethyia performed the four Mendelian Randomization tests, including both the tables and figures from the result. He wrote these corresponding sections in this paper.

We would like to thank Dr. Yongzhong Zhao for suggesting our research topic and for guiding us throughout our entire research. This research would not have been possible without Dr. Zhao's support. We would also like to thank our research mentor, Dr. Xiaodi Wang, for advising us throughout our research. Moreover, we would like to thank Tong Liu for his assistance not only in R but also in the mathematical transformations that we performed in this research.

Commitments on Academic Honesty and Integrity

We hereby declare that we

1. are fully committed to the principle of honesty, integrity and fair play throughout the competition.
2. actually perform the research work ourselves and thus truly understand the content of the work.
3. observe the common standard of academic integrity adopted by most journals and degree theses.
4. have declared all the assistance and contribution we have received from any personnel, agency, institution, etc. for the research work.
5. undertake to avoid getting in touch with assessment panel members in a way that may lead to direct or indirect conflict of interest.
6. undertake to avoid any interaction with assessment panel members that would undermine the neutrality of the panel member and fairness of the assessment process.
7. observe the safety regulations of the laboratory(ies) where we conduct the experiment(s), if applicable.
8. observe all rules and regulations of the competition.
9. agree that the decision of YHSA is final in all matters related to the competition.

We understand and agree that failure to honour the above commitments may lead to disqualification from the competition and/or removal of reward, if applicable; that any unethical deeds, if found, will be disclosed to the school principal of team member(s) and relevant parties if deemed necessary; and that the decision of YHSA is final and no appeal will be accepted.

(Signatures of full team below)

x Adeethyia Shankar
Name of team member: Adeethyia Shankar

x Stephanie Chang
Name of team member: Stephanie Chang

x
Name of team member:

x
Name of supervising teacher: Xiaodi Wang

Declaration of Academic Integrity

The participating team declares that the paper submitted is comprised of original research and results obtained under the guidance of the instructor. To the team's best knowledge, the paper does not contain research results, published or not, from a person who is not a team member, except for the content listed in the references and the acknowledgment. If there is any misinformation, we are willing to take all the related responsibilities.

Names of team members Adeethyia Shankar Stephanie Chang

Signatures of team members Adeethyia Shankar Stephanie Chang

Name of the instructor

Signature of the instructor Xiaodi Wang

Date August 20, 2023

Table of Contents

1 Introduction

More than 3 million people suffer from Inflammatory Bowel Disease (IBD) in the United States [5]. Incidences of IBD have increased rapidly over the past several decades, and it has become a global health challenge. IBD is a severe autoimmune disorder and most prevalent in two subtypes, namely Crohn's disease (CD) and ulcerative colitis (UC). CD patients suffer from chronic inflammation of the gastrointestinal tract, while UC is often localized to the descending colon.

Recent metagenomic studies have advanced our understanding of the microbial ecosystem in different diseases significantly [22,30]. It has been shown that dysbiosis of the microbiome is associated with certain diseases, including colorectal cancer, diabetes, and IBD [34]. These conditions result from a complex interplay among host genetic, microbial, and environmental factors. The advent of genomic technologies of high-throughput analyses (including metagenomics, transcriptomics, metabolomics, proteomics, and viromics) has afforded us new tools for a better understanding of the pathogenesis and development of diseases towards precision medicine.

IBD is among the most closely studied diseases caused by multiple factors involving host genetics, the environment, and microbes. The past decade has witnessed substantially increased attention to IBD based on the host-microbiome interaction approaches, but much of the etiology remained largely unknown.

In this study, we employed the metadata, transcriptomics, and metabolomics datasets from the iHMP IBD dataset. Our research focuses on an intensive search for the microbial contributors to host phenotypes via an extensive study design (see **Figure 1**).

Our association analysis of transcriptomic gene expressions with host IBD phenotypes based on a few targeted transcriptomes from discoveries made by prior studies suggests weak associations. We further performed association analysis on all transcriptomes (untargeted) and successfully identified many genetic biomarkers with very strong statistical associations with the host IBD phenotypes. Moreover, NOMO2 and some retinitis pigmentosa genes have been found strongly associated with IBD traits. We also conducted metabolomic association analysis with the host IBD phenotypes based on both targeted and untargeted metabolites. As a result, our research identifies many metabolites with very strong statistical associations with host IBD phenotypes that are unknown in existing literature, such as C18n_QI7515, C18n_QI2042, C18n_QI382, C18n_QI7515, and C18n_QI2042.

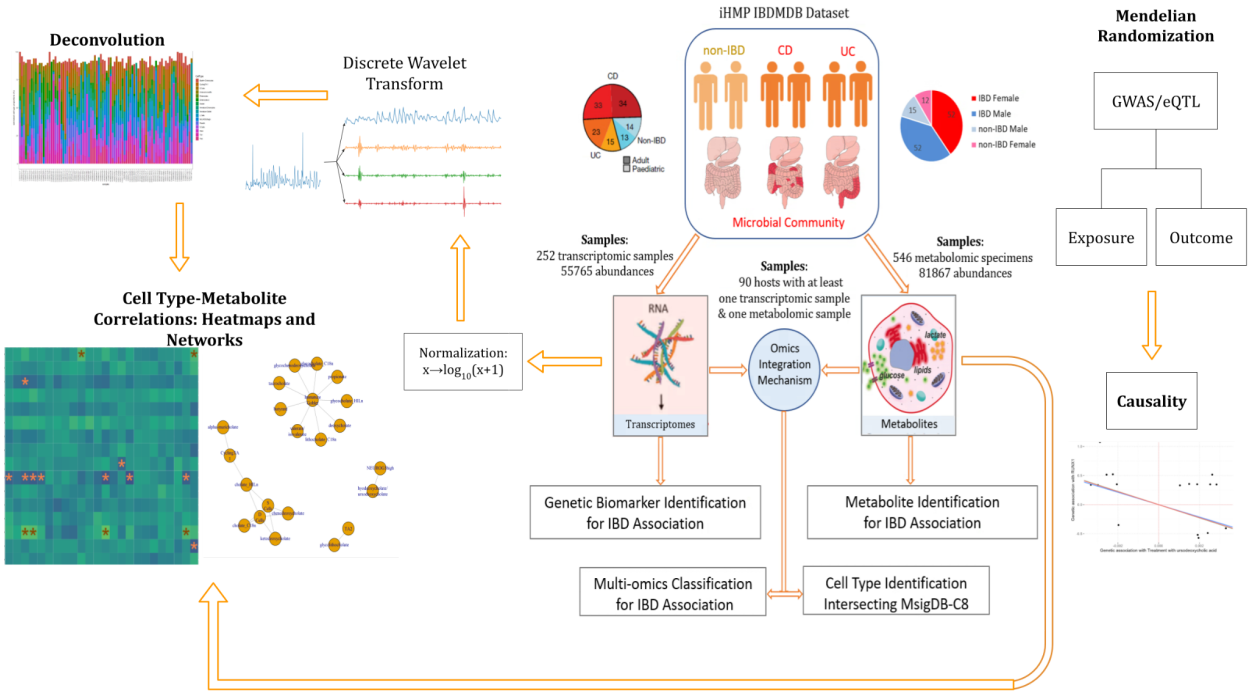


Figure 1. Workflow of study schematics on how microbiome multi-omics signature is associated with IBD connected to our pipeline of deconvoluting transcriptomic data, utilizing discrete wavelet transform, and illustrating cell type-metabolite correlations in the form of heatmaps and networks.

Host-microbial interactions in disease require an understanding of the complex data interplay of the host and microbiome. Based on a multi-omic view on the microbiome, we conducted a large-scale correlation study between transcriptomes and metabolites. Significant statistical correlations were found between transcriptomic gene expressions and host metabolites for the participants in the dataset. Strong cross-omics correlations suggest a multi-omic understanding of IBD is warranted but challenging. We found that when integrating both important information from transcriptomic and metabolomic abundances, a much stronger association between host multi-omics and IBD phenotypes is obtained than any single-omics being used alone. The correct identification of IBD phenotypes increased to more than 90% based on the aggregated two-dimensional indicators only.

To aid the visualization of microbiome data, discrete wavelet transform (DWT) has enabled multiview clustering of the data as shown in previous research [28]. Similar to the variety of microbiome correlations achieved in that paper, DWT now provides us with better performance via different angles to view cell type-metabolite correlations.

Finally, using the metabolites and transcriptomes selected above, we pinpointed these gene signatures in their corresponding cell types. This analysis plays an important role and provides useful information for biomedical applications, such as disease diagnosis and translational medicine research and development.

2 Materials and Methods

The IBDMDB dataset from the Integrative Human Microbiome Project (iHMP) was downloaded from https://ibdmdb.org [22]. In the metadata, there are numerous subject traits including the week number at which the samples were taken, race, sex, occupation, education level, and age of diagnosis. There were 131 participants taking part in the study, of which 104 were diagnosed with IBD and 27 were included for sample controls as non-IBD participants. The number of male and female participants is well-balanced.

On average, each participant took approximately two samples of the host transcriptomics test, yielding 252 samples of Host transcriptomics. 55765 transcriptomic measures were reported for each sample. Similarly, each participant took approximately 4 samples of metabolomic tests, yielding 546 samples of 81867 compound measures. We summarize the dataset in **Table 1**.

Table 1. Microbiome data summary

Data Type	Count	IBD	non-IBD	# of Abundances	Abundance Description
Metadata Participant	131	104	27	490	week_num, race, sex, age at diagnosis....
Female	64	52	12		
Male	67	52	15		
Multi-omics joint set*	90	68	22		
Host transcriptomics	252	201	51	55765	5S_rRNA, 7SK,..., yR211F11.2
Metabolomics	546	411	135	81867	C18n_QI06, C18n_QI07, … ,ILp_QI25068

* participants with minimum one host transcriptomic sample and metabolomic sample.

In the raw relative abundance data, the scales for host transcriptomes vary drastically from one gene to another. For each gene, it is also true that longitudinal and cross-sectional abundance can fluctuate in a wild range. And the same holds true for the scale for metabolites. In order to perform a meaningful correlation test, we first rank each abundance based on the dense ranking method to ensure samples of the same raw scores receive the same ranked score. We then divide the ranked scores by the maximum rank score for each abundance. In this way, every abundance is normalized between [0,1].

In this study, pairwise correlations between observed *y* variables and normalized abundance *x* variables were simply calculated based on Pearson correlations. We further employ a mean-equality t-test based on any abundance variable between two samples: a sample of IBD patients and a control sample of non-IBD participants.

We also tested the transcriptomes and metabolites using the C8 set in MsigDB. The C8 set includes 704 subsets for cell type signature genes, and each subset corresponds to a specific type of cell. For each set of marker genes found in one metabolite, we intersect it with every subset and check the number of gene signatures that coincide. A test based on hyper-geometric distribution is conducted to check whether the intersection is significant.

In addition, we perform an array of analyses using R [27]. We run deconvolution on preprocessed host transcriptomics data using the granulator R package [26], obtaining estimates for cell type proportions. Following deconvolution, we utilize DWT, compute Pearson correlations with BH adjustment, and obtain p-values using t-tests. We illustrate these correlations with heatmaps and networks, created using the ggplot2 and igraph R packages respectively [4,35]. Moreover, we perform mendelian randomization using the MendelianRandomization R package and create a causality network graph using the ggdag R package [1,25].

2.1 Preprocessing

To begin, we split the host transcriptomics data by biopsy location, namely the ileum and colon, into separate files. For the ileum, we include data from specifically the ileum, excluding the terminal ileum. For the colon, we include data from the ascending colon, descending colon, sigmoid, and transverse colon, as well as the cecum. After this separation, we convert each value to z-scores by row. This step is to ensure that the data is in the same mathematical space as the gene signature matrix that we use for deconvolution, which we obtained from Source Data Fig. 5 of Hickey et al. [15]. To convert the negative values to nonnegative in both the host transcriptomics data and the gene signature matrix, we subtract the minimum value in each column, followed by a log base 10 transformation of $x \rightarrow log_{10}(x + 1)$.

2.2 Deconvolution

Deconvolution allows us to obtain estimates of the cell type proportions present in a sample from the transcriptomic data. For the deconvolution, we used only a subset of the gene signature matrix containing cells from the same location as the ileum and colon host transcriptomics data, respectively. From the granulator package, we utilized two deconvolution methods: non-negative least squares (NNLS) and quadratic programming with non-negativity and sum-to-one constraint (QPROGWC). NNLS is a type of optimization problem where coefficients must always be nonnegative. This constraint, also present in QPROGWC, offers the advantage of preventing negative cell type proportions, which are biologically impossible. We treated the compositional cell type proportion estimates with another $x \rightarrow log_{10}(x + 1)$ log transform and then performed DWT. As DWT introduces negative values to the data, we again subtract the column minimum from each column.

2.3 Discrete Wavelet Transform

DWTs are state-of-art tools for signal processing. DWTs decompose any signal into different frequency components and capture both frequency and location information [6]. An M-band DWT uses a filter bank with M filters, where M2, to decompose an n-dimensional signal into M frequency components: one low-frequency component, and M-1 high-frequency components. In this research, we used filter banks from the 4-band 2 regular DWT.

4-band 2 regular DWT filter bank:

$[-0.06737,\ 0.09420,\ 0.4058,\ 0.5674,\ 0.5674,\ 0.4058,\ 0.09420,\ -0.06737]$

$[-0.09420,\ 0.06737,\ 0.5674,\ 0.4058,\ -0.4058,\ -0.5674,\ -0.06737,\ 0.09420]$

$[-0.09420,\ -0.06737,\ 0.5674,\ -0.4058,\ -0.4058,\ 0.5674,\ -0.06737,\ -0.09420]$

$[-0.06737,\ -0.09420,\ 0.4058,\ -0.5674,\ 0.5674,\ -0.4058,\ 0.09420,\ 0.06737]$

Below is an example of a 4-band 2 regular 12 x 12 DWT matrix *T* constructed with low pass filter α and high pass filters β, γ, δ.

$$T = \begin{bmatrix} \alpha_1 & \alpha_2 & \alpha_3 & \alpha_4 & \alpha_5 & \alpha_6 & \alpha_7 & \alpha_8 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & \alpha_1 & \alpha_2 & \alpha_3 & \alpha_4 & \alpha_5 & \alpha_6 & \alpha_7 & \alpha_8 \\ \alpha_5 & \alpha_6 & \alpha_7 & \alpha_8 & 0 & 0 & 0 & 0 & \alpha_1 & \alpha_2 & \alpha_3 & \alpha_4 \\ \beta_1 & \beta_2 & \beta_3 & \beta_4 & \beta_5 & \beta_6 & \beta_7 & \beta_8 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & \beta_1 & \beta_2 & \beta_3 & \beta_4 & \beta_5 & \beta_6 & \beta_7 & \beta_8 \\ \beta_5 & \beta_6 & \beta_7 & \beta_8 & 0 & 0 & 0 & 0 & \beta_1 & \beta_2 & \beta_3 & \beta_4 \\ \gamma_1 & \gamma_2 & \gamma_3 & \gamma_4 & \gamma_5 & \gamma_6 & \gamma_7 & \gamma_8 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & \gamma_1 & \gamma_2 & \gamma_3 & \gamma_4 & \gamma_5 & \gamma_6 & \gamma_7 & \gamma_8 \\ \gamma_5 & \gamma_6 & \gamma_7 & \gamma_8 & 0 & 0 & 0 & 0 & \gamma_1 & \gamma_2 & \gamma_3 & \gamma_4 \\ \delta_1 & \delta_2 & \delta_3 & \delta_4 & \delta_5 & \delta_6 & \delta_7 & \delta_8 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & \delta_1 & \delta_2 & \delta_3 & \delta_4 & \delta_5 & \delta_6 & \delta_7 & \delta_8 \\ \delta_5 & \delta_6 & \delta_7 & \delta_8 & 0 & 0 & 0 & 0 & \delta_1 & \delta_2 & \delta_3 & \delta_4 \end{bmatrix}$$

The DWT matrix *T* is an orthogonal matrix, meaning that T^t is the inverse of *T*. To perform DWT, we multiply each data sample S by the DWT matrix *T*. Then *S* is transformed into the Wavelet Domain, and can be expressed as:

$$TS = \begin{bmatrix} a_1 \\ d_1 \\ \vdots \\ d_{M-1} \end{bmatrix} \triangleq \tilde{S}$$

$$a_1 = \begin{bmatrix} a_{11} \\ \vdots \\ a_{1k} \end{bmatrix}, d_1 = \begin{bmatrix} d_{11} \\ \vdots \\ d_{1k} \end{bmatrix}, \ldots, d_{M-1} = \begin{bmatrix} d_{M1} \\ \vdots \\ d_{Mk} \end{bmatrix},$$

where $k = n/M$ and n is the dimension of S.

Let $C_1, C_2, \ldots, C_{Mk}$ be the column vectors of T^t. Then $\{C_1, C_2, \ldots, C_{Mk}\}$ forms an orthonormal basis of $\mathbb{R}^{Mk}$. Define

$$A_1 = a_{11}C_1 + \cdots + a_{1k}C_k$$

$$D_i = d_{i1}C_{i*k+1} + \cdots + d_{i,(i+1)k}C_{(i+1)*k}$$

for $i = 1, \ldots, M - 1$. Then A_1 is corresponding to a_1 and D_i is corresponding to d_i for $i = 1, \ldots, M - 1$. If we let

$$V = span\{C_1, \ldots, C_k\}$$

$$W_i = span\{C_{i*k+1}, \ldots, C_{(i+1)*k}\}, \text{ for } i = 1, \ldots, M - 1$$

then V, $W_1, \ldots, W_{M-1}$ are orthogonal to each other and therefore $\mathbb{R}^{Mk}$ is the direct sum of V, $W_1, \ldots,$ and W_{M-1}, that is

$$\mathbb{R}^{Mk} = V \oplus W_1 \oplus W_2 \oplus \cdots \oplus W_{M-1}$$

So, for $S \in \mathbb{R}^{Mk}$, S can be written uniquely as

$$S = A_1 + D_1 + \cdots + D_{M-1}$$

where $A_1 = Proj_V S$ = the orthogonal projection of S onto V, $D_i = Proj_{W_i} S$ = orthogonal projection of S onto W_i for $i = i, \ldots, M - 1$.

Indeed,

$$T^t \tilde{S} = A_1 + D_1 + \cdots + D_{M-1} = S$$

where A_1 is an approximation or low-frequency component of S, and D_1 through D_{M-1} are the details or high-frequency components of S. We apply four-band two-regular DWT to our deconvolution cell type proportions to obtain five different components: No DWT, A1, D1, D2, and D3.

3 Transcriptomics

Existing transcriptomic studies of IBD have discovered many important transcripts involved in the process of IBD pathogenesis. **Table 2** exemplifies some of the important previous findings on this subject.

Table 2. Genes and IBD studies

Reference	Year	Gene	Findings
Yang, et al. [37]	2021	IRF5	IRF5 regulates Th1 and Th17 immune responses and cytokine production and is a possible marker for managing IBD.
Bruns, et al. [2]	2009	ABCB1	ABCB1 gene may be responsible for the poor response of IBD patients to glucocorticoid treatment.
Yilmaz, et al. [38]	2018	PTPN2	PTPN2 modulates intestinal microbiota composition and it interacts with microbiota directly and affects disease severity in IBD patients.
Glocker, et al. [10]	2009	IL10R	Genes IL10RA and IL10RB form a heterotetramer to make up the interleukin-10 receptor and were found involved in hyper-inflammatory immune responses in the intestine.
Duerr, et al. [7]	2006	IL23R	IL23R gene encodes a subunit of the receptor for the proinflammatory cytokine interleukin-23 in IBD.
Horowitz, et al. [16]	2021	NOD2	Recessive inheritance of NOD2 alleles is a mechanistic driver of early-onset CD, likely due to loss of NOD2 protein function.
Chun, et al. [3]	2019	ATG16L1	ATG16L1 contributes to dysbiosis and immune infiltration prior to IBD disease symptoms.
Mehto, et al. [24]	2019	IRGM	IRGM suppresses the pro-inflammatory responses and slows down the activation of inflammasomes and protects from gut inflammation.
Goyette, et al. [11]	2015	HLA	Multiple HLA alleles, especially HLA-DRB1, are implicated for noteworthy differences in IBD.

Most of these studies have involved the identification of specific genetic associations with IBD and achieved a better understanding of the pathways in which genetic factors influence IBD.

3.1 Targeted transcriptome analysis

Our dataset from 252 samples of host transcriptomes from 131 hosts has a rich 55765 transcriptomic abundances for each sample and allowed us to test the findings on the targeted genes in the existing literature.

The two-sample mean-equality t-test—testing whether the sample means are the same between 201 samples from IBD hosts and 51 non-IBD hosts—was our primary statistical test on the degree of association of transcriptomics with IBD traits.

We further constructed a multi-mode (more continuous states compared with binary means) test by breaking the 252 samples into 10 groups (each group comprising around 25 samples) based on ascending ranking of any normalized gene abundance, then we calculated the correlation between ranking (1 to 10) with a ratio of samples from IBD to total sample size within each decile. **Table 3** reports the targeted genes and their p-values of statistical tests for association with IBD host traits.

Table 3. Targeted genes' association with IBD statistical tests

Targeted Gene	Two sample mean-equality test		Decile correlation t-test	
	t-statistic	p-value	Correlation	p-value
IRF5	-1.39	0.1700	-0.4312	0.2134
ABCB1	-5.10*	$1.55 * 10^{-6}$	-0.8193*	0.0037
PTPN2	1.57	0.1200	0.6242	0.0538
IL10RA	1.15	0.2540	0.3748	0.2860
IL10RB	-0.34	0.7330	-0.0391	0.9145
IL23R	1.66	0.1010	0.3402	0.3361
NOD2	3.44*	0.0009	0.6605*	0.0376
ATG16GL1	-1.32	0.1920	-0.5029	0.1385
IRGM	-0.56	0.5770	0.0551	0.8797
HLA-DRB1	1.65	0.1030	0.6495	0.0421

* statistically significant at p-value 5%

Our data analysis confirmed that only ABCB1 and NOD2 survive our statistical test in differentiating IBD and non-IBD samples while other genes lack statistical power in their association with IBD traits.

3.2 Host Transcriptome-wide Association Study of IBD

We also conducted the statistical tests in **Section 3.1** on all 55765 transcriptomes to determine if they are each correlated with IBD traits. There are 9% and 3% of transcriptomes that meet the 0.1% p-value significant cutoff for the two different statistical tests, respectively. This implies that there are many transcriptomes that may be associated with the host IBD traits compared with a standard normally distributed population. **Figure 2.A** depicts a Q-Q plot of observed $-\log_{10}$(p-value) against expected (assuming uniform p-value) $-\log_{10}$(p-value).

Based on the two sample mean-equality t-test, we found 594 transcriptomes with p-values exceeding the $5*10^{-8}$ significance threshold and we focused on AQP9, HCAR3, MMP10, and CXCR1 as distinct pattern breakout candidates with a strong association with host IBD traits (p-value exceeding $5*10^{-22}$ significance level). Our findings confirmed the other discoveries. Yu et al. identified AQP9 as an auxiliary diagnostic indicator for CD and enhanced insight into the immune cells underlying CD [39]. Wnorowski et al. found that HCAR3 is upregulated in IBD and gets suppressed after treatment with Infliximab [36]. Koller et al. reported that MMP10 improves colitis conditions and promotes anti-inflammation-associated colonic dysplasia development [20]. Gijsbers et al. discovered that CXCR1-binding chemokines in IBD are linked to down-regulated IL-8 and CXCL8 production by leukocytes in CD [9].

Based on the ten deciles' correlation test, we also identified CXCL5, NOMO2, RP11-143K11.1, and RP11-457M11.5 genes as distinct pattern breakout candidates with a strong association with host IBD traits (p-value exceeding $5*10^{-8}$ significance level). Singh et al. found a significant upregulated CXCL5 in IBD patients, as well as other chemokines [29]. Previous research has examined the role of NOMO2 and RP11 gene families in relation to IBD and has not reported their functions on IBD [40].

In **Figure 2.B**, the top chart illustrates that the IBD sample concentration in each decile correlates poorly with the ranking of deciles formed based on ranked values of the targeted genes. as suggested by their weak statistical powers in **Table 3**.

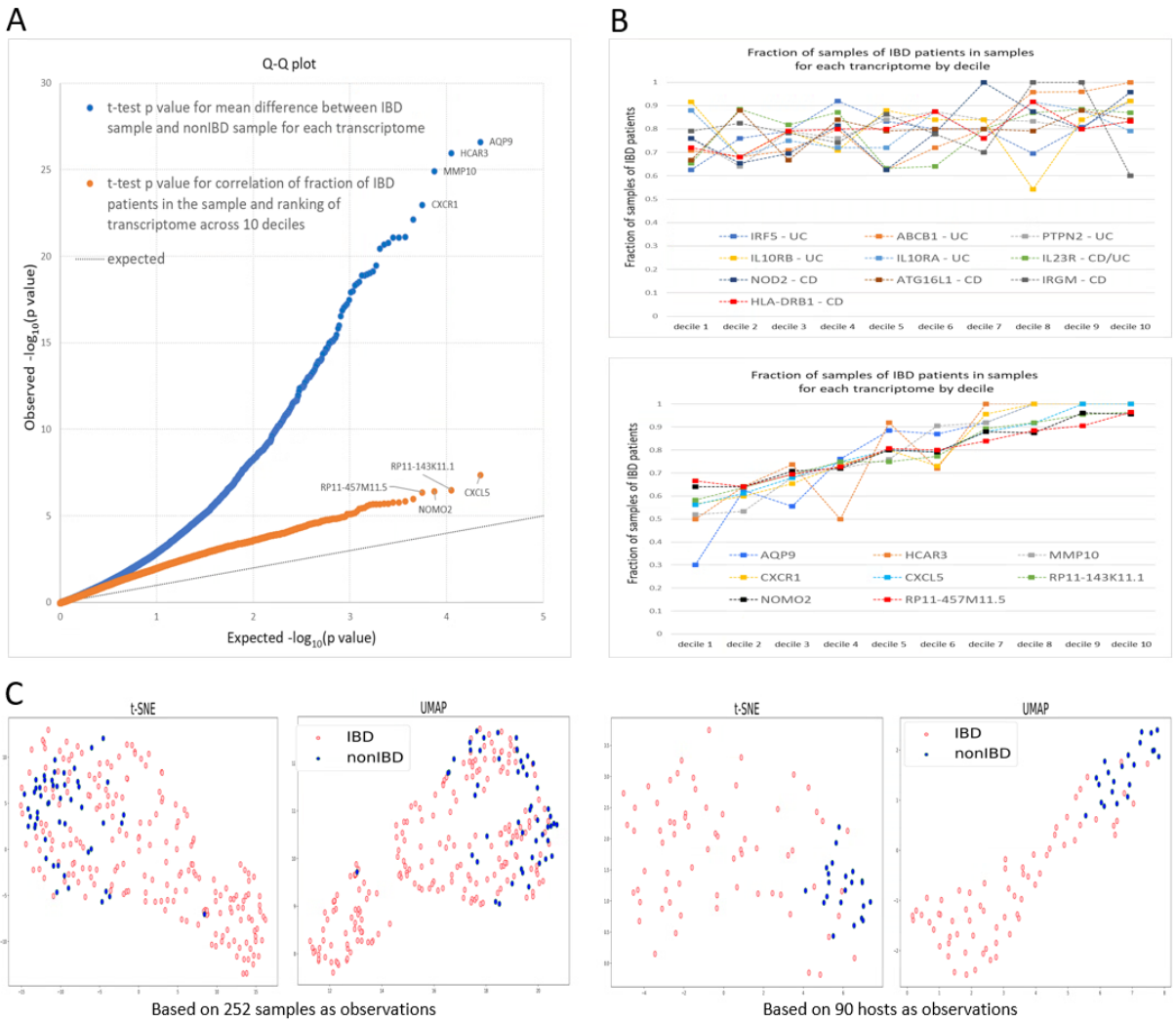


Figure 2. Host Transcriptomic analysis using IBD patient samples and non-IBD control samples.

(A) Q-Q plot of observed and expected $-\log_{10}$(p value) of two tests:
1). A blue dot represents a student t-test for mean-equality of any normalized transcriptomic abundance $x \in \{x_i \mid i=1,\ldots,55765\}$ where 201 samples of x were collected from IBD patients and 51 samples from non-IBD participants;
2). An orange dot represents a student t-test for correlation of ranking of x in ten deciles (each decile has around 25 samples ranked based on values of x) with fraction of IBD samples in each decile;

(B) When the targeted genes based on findings of existing literatures are selected as potential attributing genes to IBD, it shows these transcriptomes are weakly associated with the fraction of IBD samples in each of the ten deciles in the top chart;
When the untargeted genes from statistical break-outs are selected as potential attributing genes to IBD, it shows these transcriptomes are significantly associated with the fraction of IBD samples in each of the ten deciles in the bottom chart.

(C) 100 most significant transcriptomes from A.1 with mean-equality t-test p-value < $5*10^{-8}$ are selected as features for t-SNE and UMAP analysis. Left two plots are t-SNE and UMAP 2-dim plots using samples of the selected transcriptomes.
For each host, the average score of normalized transcriptomic abundances scores from all her/his samples as her/his scores is used to convert the dataset from 252 samples into 90 hosts-based samples. Based on the same top 100 transcriptomes of mean-equality t-test significances, two corresponding t-SNE and UMAP plots are drawn on the right as a comparison to 2 plots on the left.

The bottom chart displays a very strong increasing relationship between the abundance decile and the fraction of IBD samples in each of the ten deciles using the top untargeted genes with cross-sectional correlation p-value < 0.2%. (We reversed the ranking order if the correlation is

negative.) These untargeted genes explain 100% of IBD samples when the abundance level is strong (high deciles), and only 50% when the abundance level is weak (low deciles).

This poor differentiating ability at low abundance levels may be due to a biased sampling in the study since the baseline case for IBD sample concentration is 80%, making it statistically difficult to single out non-IBD samples. Another plausible explanation is offered by Lloyd-Price, that the transcriptomic gene expression of IBD hosts is highly variable longitudinally since IBD patients experienced both inflammatory states and inactive states [22]. When samples were collected during inactive states for IBD hosts, their abundance measures could be less distinguishable from non-IBD samples.

Evidence in **Figure 2.A** suggests that a very high number of genes could be collectively associated with IBD traits. We employed t-SNE and UMAP to reduce feature high dimensionality to only 2 dimensions for visual inspection as classification techniques are based only on large feature similarities (not on outcomes).

We then selected the 100 most significant transcriptomic abundances based on the two sample mean-equality t-test with a p-value less than $5*10^{-8}$ as features for t-SNE and UMAP. In the two left plots of **Figure 2.C**, it appears to be that most of the IBD samples can be reasonably separated from non-IBD control samples. However, the non-IBD samples tend to be embedded with IBD samples, similar to the case in **Figure 2.B**.

One potential way to navigate the issue of longitudinal instabilities of IBD is to pool the longitudinal samples from each host by averaging the samples. Lloyd-Price found that inter-individual variations accounted for the majority of variance in all measurement types and suggested that the potentially key connections between host microbes and disease phenotypes may be better discovered between individuals with and without IBD.

Using the simple averages to pool the samples for any host let us focus on the inter-individual differences. It also helps to smooth out any longitudinal variations of IBD conditions and acts as a denoising mechanism to the dataset. In the two right plots of **Figure 2.C**, most of the IBD hosts continue to be reasonably separable from non-IBD hosts. However, the 22 non-IBD hosts are still mixed with approximately 10 IBD hosts, albeit a big improvement already from decile cross-sectional correlation in lower **Figure 2.B**. With visual inspection comparing the two sets of t-SNE and UMAP plots on samples without pooling as well as with pooling, the pooling approach offers a clearer differentiation between IBD traits.

4 Metabolomics

Metabolites in gut microbiomes play a key role in mediating how the gut microbiota interacts with the host. Many existing studies have focused on and implicated specific classes of metabolites such as bile acids and short-chain fatty acids in IBD pathogenesis. **Table 4** lists some of their findings.

Table 4. Metabolomics and IBD studies

Reference	Year	Metabolite	Findings
Ward, et al. [33]	2017	HILn_QI82 C18n_QI48 lithocholate	Lithocholate acid (LCA), a secondary bile acid, exerts anti-inflammatory actions in the colon. LCA is hepatotoxic in many animal species including nonhuman primates.
Zhao, et al. [40]	2016	C18n_QI50 deoxycholate	High levels of fecal deoxycholic acid (DCA) may be a dangerous signal for colonic inflammation of IBD.
Medina, et al. [23]	2021	HILn_QI28433 propionate	Propionate kinase is the enzyme that catalyzes the anaerobic breakdown of L-threonine to propionate. It is responsible for propionate production in the gut and is depleted in patients with CD.
Gasaly, et al. [12]	2021	HILn-QI36 butyrate	Butyrate decreases inflammatory processes by stabilizing the gut barrier function. IBD patients exhibit a lower abundance of butyrate-producing content.

4.1 Targeted metabolomics analysis

Our dataset contains 546 samples of host metabolites from 131 hosts and it reports 81867 metabolomic abundances per sample. To check if our data analysis on the targeted genes agrees with the findings in existing literature, a two-sample mean-equality t-test between 411 samples from IBD hosts and 135 non-IBD hosts, was our primary statistical test on the degree of association of metabolites with IBD traits. We further constructed a 10 means (more continuous states compared with 2 means) test by breaking the 546 samples into 10 groups (each group comprising around 55 samples) based on ascending ranking of any normalized metabolomic abundance. Then, we calculated the correlation between ranking (1 to 10) of the ratio of samples from IBD to the total sample size within each decile. It appears in **Table 5** that among the few targeted metabolites, only lithocholate and propionate may effectively differentiate IBD and non-IBD samples, while other metabolites lack the statistical power to be found significant with sample IBD traits.

Table 5. Targeted metabolites' association with IBD statistical tests

targeted compound	two sample mean-equality t-test		decile correlation t-test	
	t-statistics	p-value	correlation coef.	p-value
HILn_QI82 lithocholate	3.42*	0.0007	-0.5513	0.0986
C18n_QI48 lithocholate	-1.94	0.0539	-0.3426	0.3325
C18n_QI50 deoxycholate	-1.79	0.0745	-0.3463	0.3269
HILn_QI36 butyrate	-1.88	0.0615	0.5334	0.1123
HILn_QI28433 propionate	-2.25*	0.0253	-0.6447*	0.0442

* statistically significant at 5%

4.2 Microbiome Metabolome-wide Study of IBD

We also conducted the same statistical tests in **Section 4.1** on all 81867 metabolomic abundances to see if they are individually correlated with IBD traits. There are 32% and 10% of metabolites that meet the 0.1% p-value significant cutoff for two statistical tests described in **Section 4.1**, respectively, implying that there is a large number of metabolites that may be strongly associated with the host IBD traits compared with a standard normally distributed population.

Figure 3.A depicts a Q-Q plot of the observed $-\log_{10}$(p-value) against the expected $-\log_{10}$(p-value) (assuming uniform p-value). Based on the two-sample mean-equality t-test, we found 5475 metabolites with p-values exceeding the $5*10^{-8}$ significance threshold. We singled out C18n_QI382, C18n_QI7515, and C18n_QI2042 as distinct pattern breakout candidates with a strong association with host IBD traits (p-value exceeding 10^{-43} significance level). Based on the ten deciles' correlation test, we also identified HILp_QI1898, C18n_QI522, and HILp_QI12317 as major pattern breakout candidates with a strong association with host IBD traits (p-value exceeding $5*10^{-8}$ significance level).

Among the top six candidates, only one metabolite has been discussed in existing studies. Huang et al. recently found biomarkers that are significantly different among hosts with IBD and without IBD are basically unknown metabolites including C18n_QI382 explaining the greatest differences in their A.I. application [17]. However, C18n_QI7515, C18n_QI2042, C18n_QI382, C18n_QI7515, and C18n_QI2042 do not yet exist in research studies. Further efforts may be warranted for better compound annotations and understanding of these compounds in future research.

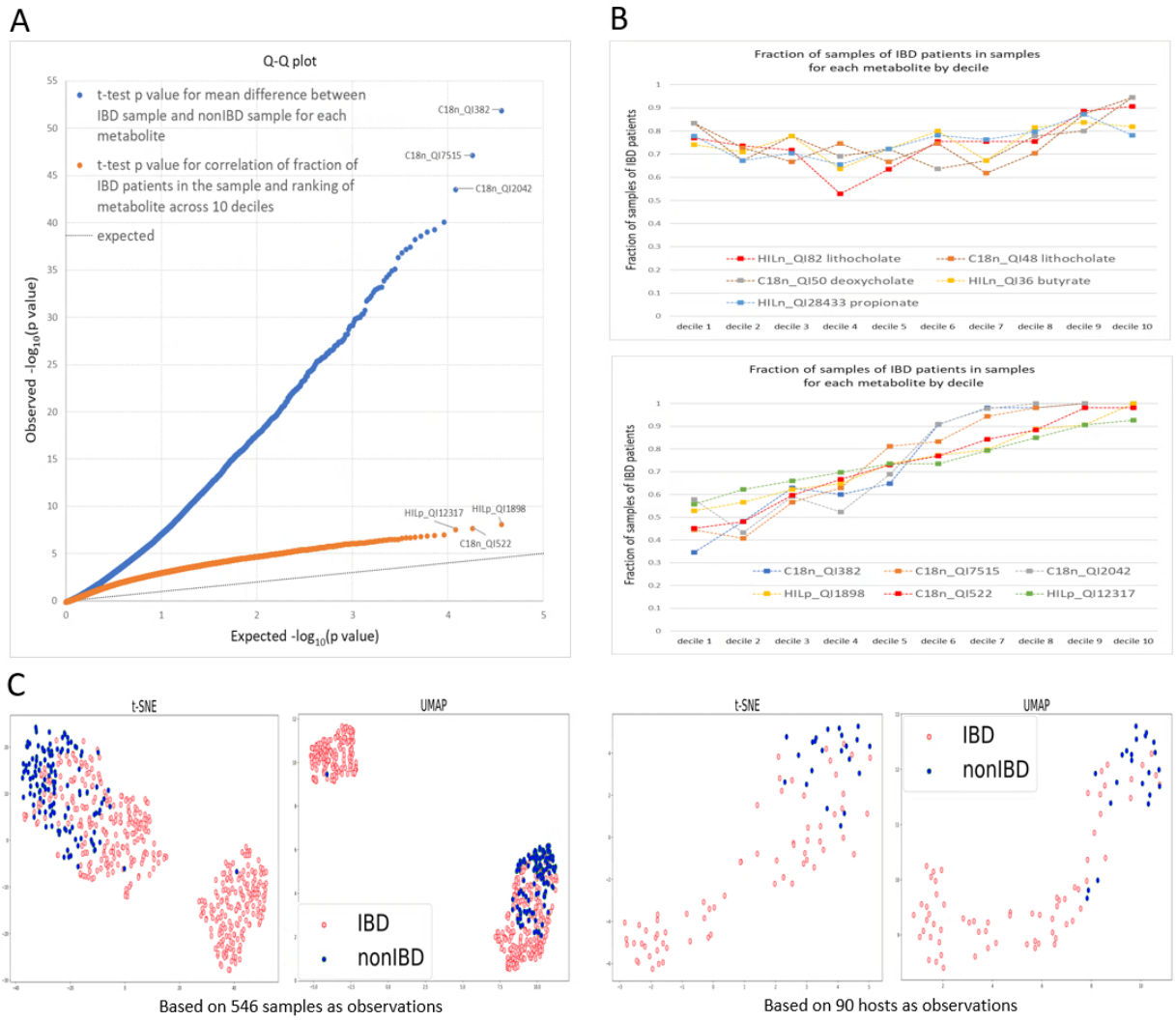


Figure 3. Metabolomic analysis using IBD patient samples and non-IBD control samples.

(A) Q-Q plot of observed and expected $-\log_{10}$(p value) of two tests:
1). A blue dot represents a student t-test for mean-equality of any normalized metabolomic abundance $x \in \{x_i \mid i=1,\ldots,81867\}$ where 411 samples of x were collected from IBD patients and 135 samples from non-IBD participants;
2). An orange dot represents a student t-test for correlation of ranking of x in ten deciles (each decile has around 55 samples ranked based on values of x) with fraction of IBD samples in each decile;

(B) When the targeted metabolites based on finding of existing literatures are examined as potential attributing metabolites to IBD, it shows these metabolites are very weakly related to the fraction of IBD samples in each of the ten deciles in the top chart;
When the untargeted metabolites from statistical break-outs are selected as potential attributing metabolites to IBD, it shows these metabolites are significantly associated with the fraction of IBD samples in each of the ten deciles in the bottom chart.

(C) 100 most significant metabolites from A.1 with mean-equality t-test p-value $< 5*10^{-8}$ are selected as a features for t-SNE and UMAP analysis. Left two plots are t-SNE and UMAP 2-dim plots using samples of the selected metabolites.
For each host, the average score of normalized metabolomic abundance scores from all her/his samples as her/his scores is used to convert the dataset from 546 samples into 90 hosts-based samples. Based on the same top 100 metabolites of mean-equality t-test significances, two corresponding t-SNE and UMAP plots are drawn on the right as a comparison to 2 plots on the left.

It appears that the ranked values of the targeted metabolite correlate poorly with the IBD sample concentration in each decile (**Figure 3.B**) and alongside a general weak statistical power (**Table 5**).

The bottom chart displays a very strong increasing relationship between the abundance decile and the fraction of IBD samples in each of the ten deciles (we reversed the ranking order if the correlation is negative) using the untargeted metabolites with cross-sectional correlation p-values < 0.02%. The untargeted metabolites explain 100% of IBD samples when the abundance level is strong (high deciles), and only 40% when the abundance level is weak (low deciles), again probably due to some samples of IBD hosts collected during non-acute states. This evidence is consistent with our findings from the t-SNE and UMAP plots in **Figure 2.B**. Despite these relatively unknown compounds, they seem generally capable of differentiating host IBD traits better than the untargeted transcriptomes do.

We found the top 100 most significant metabolites with a strong association with host IBD traits with a minimum two-sample mean-equality t-test p-value being $5*10^{-8}$ on 546 metabolomic samples and visualized with t-SNE and UMAP (**Figure 3.C**).

Portions of IBD samples can be reasonably separated from non-IBD control samples while the non-IBD samples still tend to be embedded with IBD samples. When pooling samples to 90 hosts based on each host's average metabolomic abundances over time, the t-SNE and UMAP do present a clearer separation of IBD outcome than the un-pooling approach as the two plots in the right panel compared with the left panel. However, the non-IBD are still embedded with the IBD hosts, much like the top 100 untargeted transcriptomes (**Figure 2.C)**.

4.3 Cell type-Metabolomics Correlation Analysis

Moreover, we performed pairwise correlations with BH adjustment between the cell type proportions obtained from deconvolution and the metabolomics data. For these correlations, we focused on 18 bile acids and three short-chain fatty acids. The three short-chain fatty acids we looked at are butyrate, propionate, and valerate/isovalerate, which we abbreviate as val/isoval. We list the 18 bile acids below in **Table 6.**

Table 6. Bile acids and abbreviations

Metabolite	Abbreviation
lithocholate	LCA*
chenodeoxycholate	CDCA
deoxycholate	DCA
hyodeoxycholate/ursodeoxycholate	HDCA/UDCA
ketodeoxycholate	keto-DCA
alpha-muricholate	α-MCA
cholate	CA*
glycolithocholate	GLCA
glycochenodeoxycholate	GCDCA
glycodeoxycholate	GDCA
glycoursodeoxycholate	GUDCA
glycocholate	GCA*
taurolithocholate	TLCA
taurochenodeoxycholate	TCDCA
taurodeoxycholate	TDCA
taurohyodeoxycholate/tauroursodeoxycholate	THDCA/TUDCA
tauro-alpha-muricholate/tauro-beta-muricholate	TαMCA/TβMCA
taurocholate	TCA

*Collected with both the C18-neg and HILIC-neg methods, duplicate metabolites are differentiated by suffix (i.e., _C18n or _HILn)

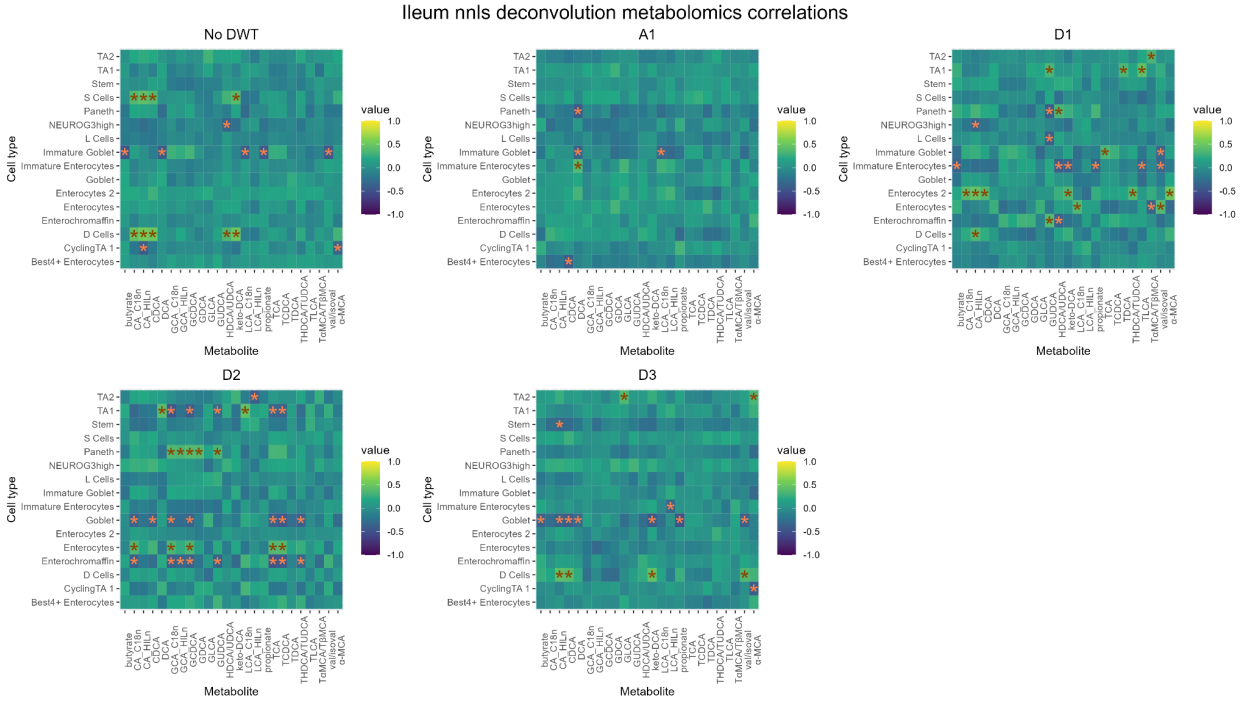
These bile acids consist of primary and secondary bile acids, and although some metabolites were collected using both the C18-neg and HILIC-neg methods, we differentiated between them by suffixing the metabolite abbreviation in **Figures 4, 5, 6, and 7**.

Using the results of these pairwise correlations, we present correlation heatmaps, starring significant correlations with p-value less than 0.05. Following the heatmaps, we created a network connecting cell types and metabolites with significant correlations for each the ileum and the colon.

The size of each vertex is based on $log_2(d + 1)$, where d is the degree of the vertex. To minimize overlapping in the networks, we abbreviate “Immature” as “Imm.”

4.3.1 Ileum

A



B

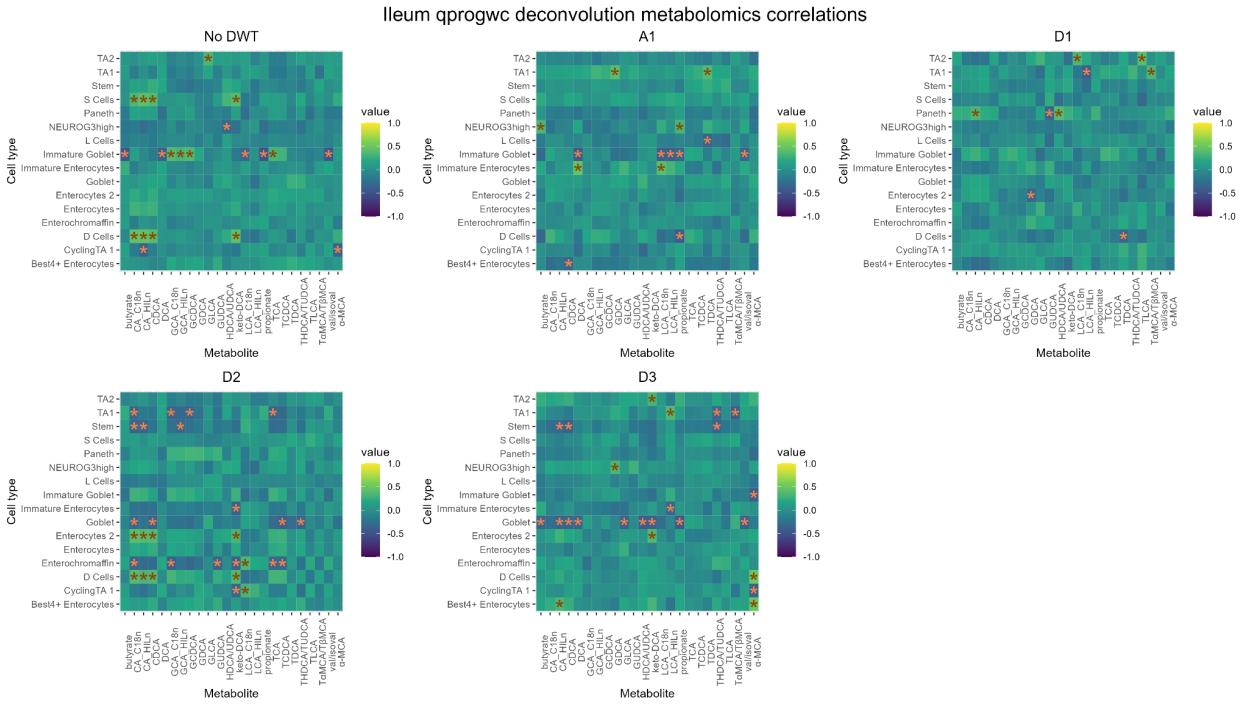


Figure 4. Ten heatmaps of correlations between ileum cell types and metabolites: No DWT, A1, D1, D2, and D3 wavelet components for both NNLS and QPROGWC deconvolution results. Significant correlations with $p < 0.05$ are starred, with dark orange stars for positive correlations and light orange stars for negative correlations.

In **Figure 4.A**, the D1 and D2 wavelet components present vastly more significant correlations than viewed in No DWT. For example, in D2, the TA1, Paneth, Goblet, Enterocytes, and Enterochromaffin cells enjoy numerous correlations with the metabolites. This trend extends into **Figure 4.B** with many keto-DCA correlations in D2. Across all ten heatmaps, propionate and valerate/isovalerate have barely a couple of significant correlations.

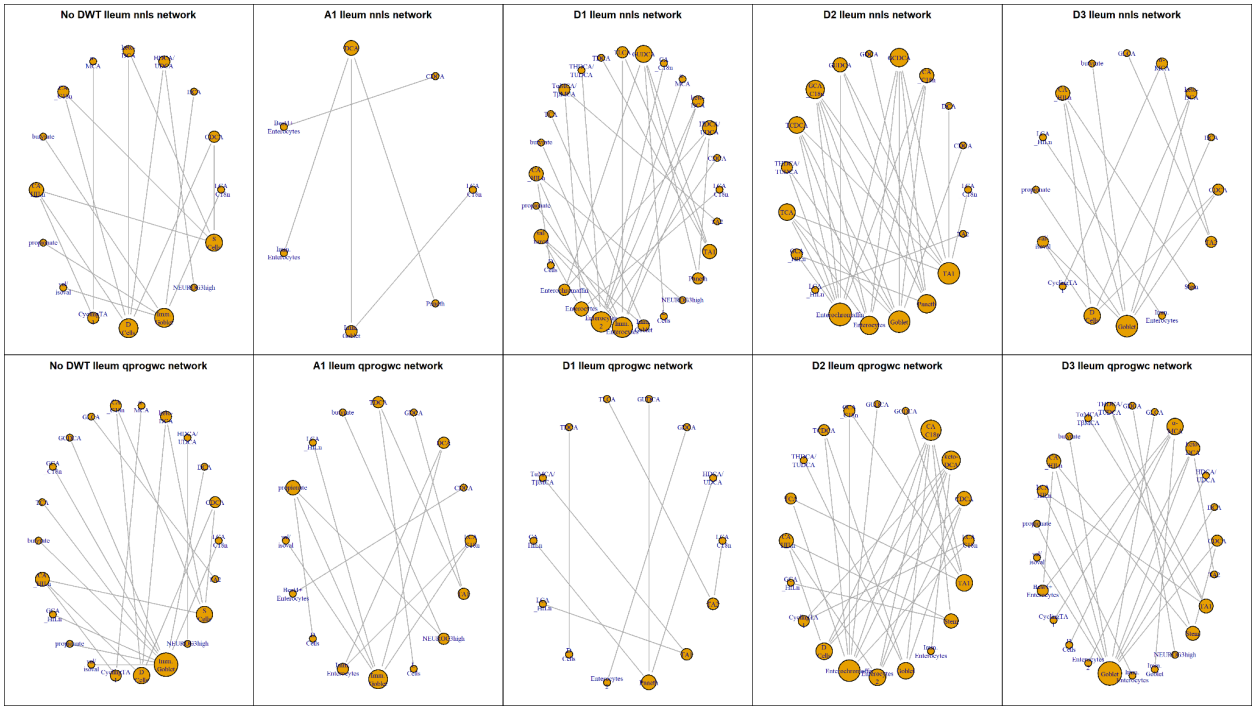


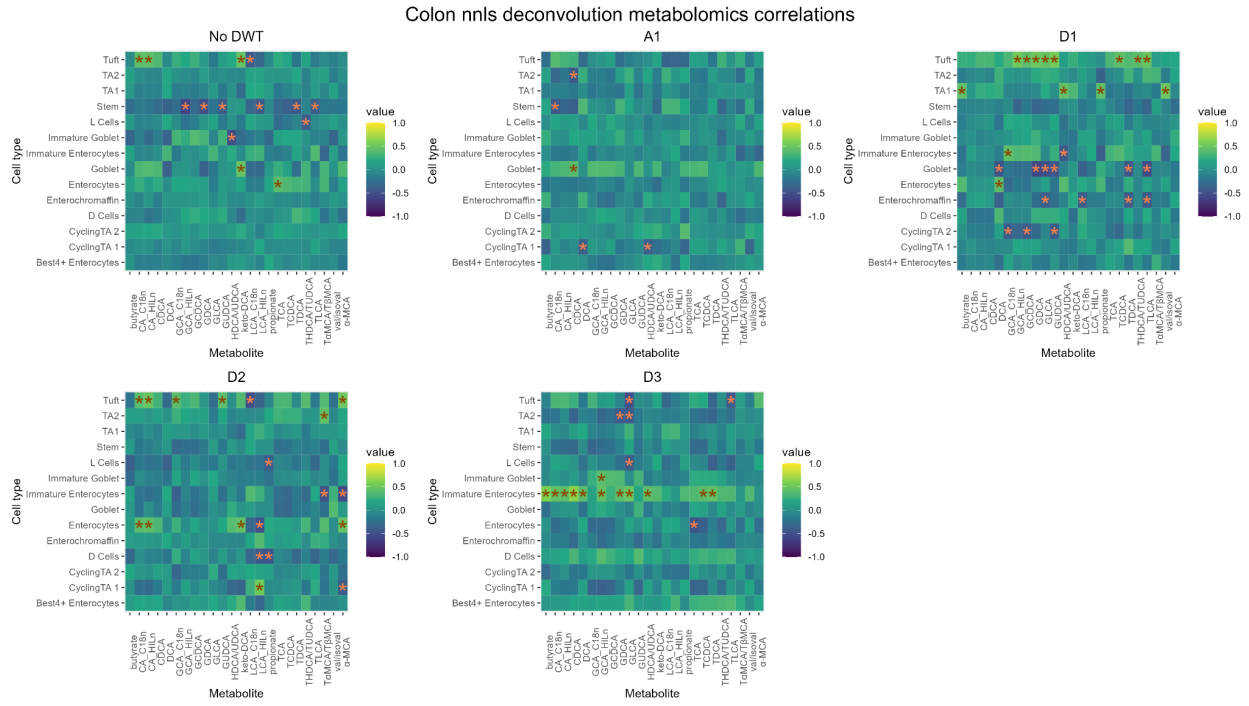
Figure 5. Ten networks of correlations between ileum cell types and metabolites: No DWT, A1, D1, D2, and D3 wavelet components for both NNLS and QPROGWC deconvolution results. Connections represent significant correlations with p-value less than 0.05. Nodes with more significant correlations are logarithmically larger.

In the **Figure 5** networks, the wavelet detail networks are visibly livelier than the No DWT networks alone, on both the NNLS and QPROGWC deconvolution cell type proportions. The A1 ileum NNLS network is also very sparse. While there is much heterogeneity among the networks, the Immature Goblet and Goblet cells consistently have many significant correlations with metabolites as evidenced by their large node size.

4.3.2 Colon

Here we present similar figures as in the ileum section above.

A



B

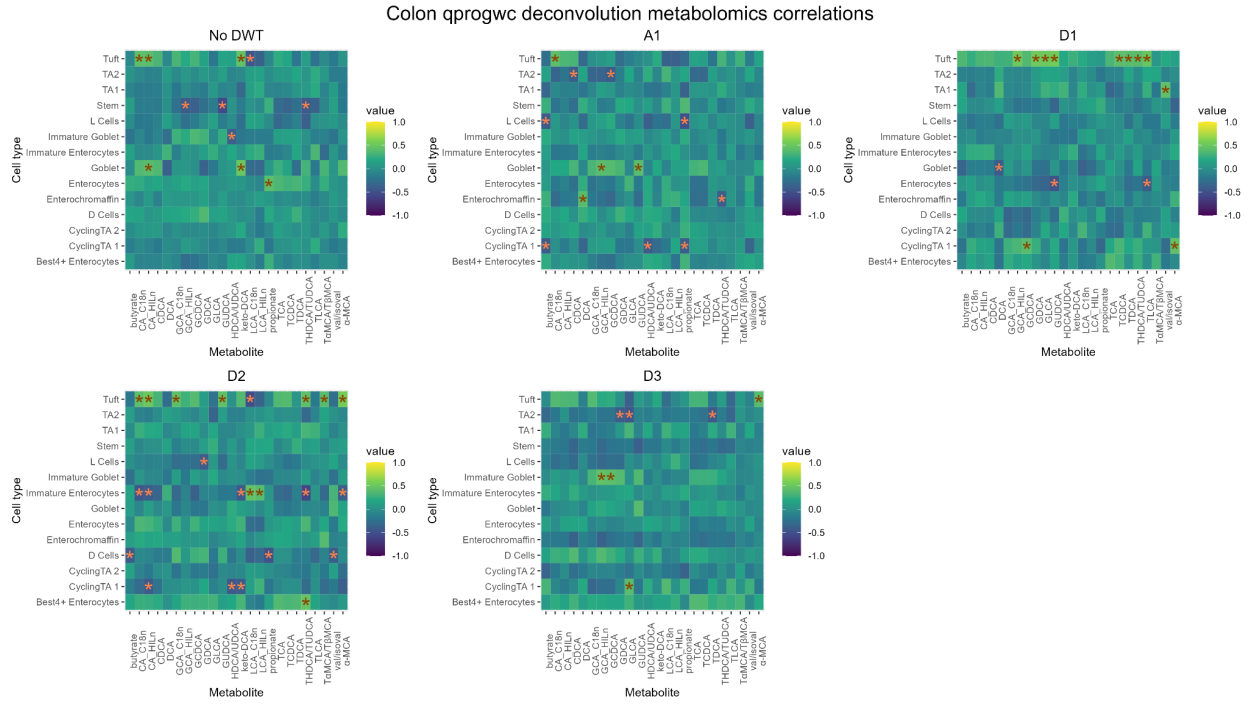


Figure 6. Ten heatmaps of correlations between colon cell types and metabolites: No DWT, A1, D1, D2, and D3 wavelet components for both nnls and QPROGWC deconvolution results. Significant correlations with p-value less than 0.05 are starred, with dark orange stars for positive correlations and light orange stars for negative correlations.

Although the additional views of the data offered by DWT do not present as many correlations as in the ileum data, there are some noticeable differences. In particular, tuft cells in both **Figure 6.A** and **Figure 6.B** present various significant correlations with the metabolites outside of those observed in the No DWT heatmap. Also, Immature Enterocytes in D3 of colon NNLS, Goblet cells in D1 of colon NNLS, and Immature Enterocytes in D2 of colon QPROGWC have many significant correlations.

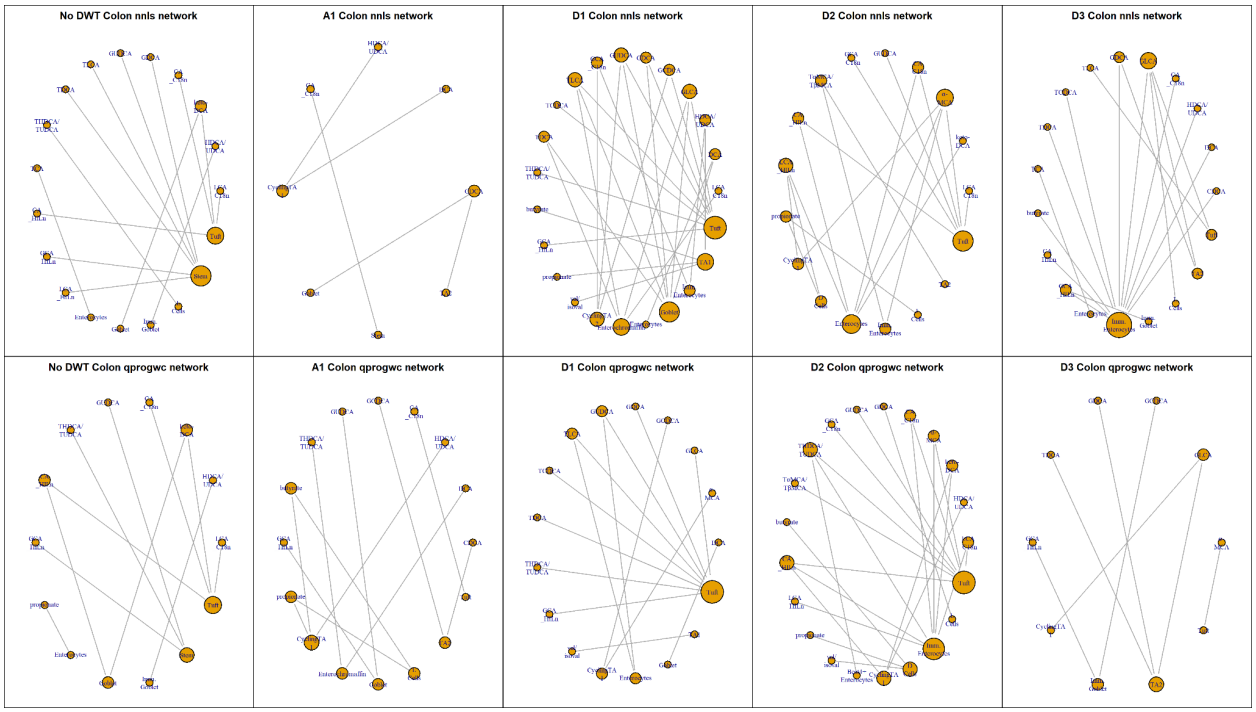


Figure 7. Ten networks of correlations between colon cell types and metabolites: No DWT, A1, D1, D2, and D3 wavelet components for both NNLS and QPROGWC deconvolution results. Connections represent significant correlations with p-value less than 0.05. Nodes with more significant correlations are logarithmically larger.

Similar to the ileum networks, the colon networks in **Figure 7** have busier detail networks, with the exception of the D3 colon QPROGWC network. Tuft cells are consistently observed to have numerous significant correlations with metabolites, but we especially observe many correlations with Immature Enterocytes outside of the No DWT network.

5 Integrated multi-omic analysis

Our transcriptomic and metabolomic analyses have provided a good understanding of the linkages and pathway to IBD at the microbial level. We ruled out a simple, universal ‘‘magic bullet’’ to diagnose IBD based on any single-omics attributes.

Systems-level views of the genetic immune response and the interaction between genetic factors and microbial metabolic patterns to disease susceptibility have just begun in recent years. Further integration of these-omic approaches will enhance and expand our knowledge about IBD and improve our ability to diagnose and treat various stages of IBD.

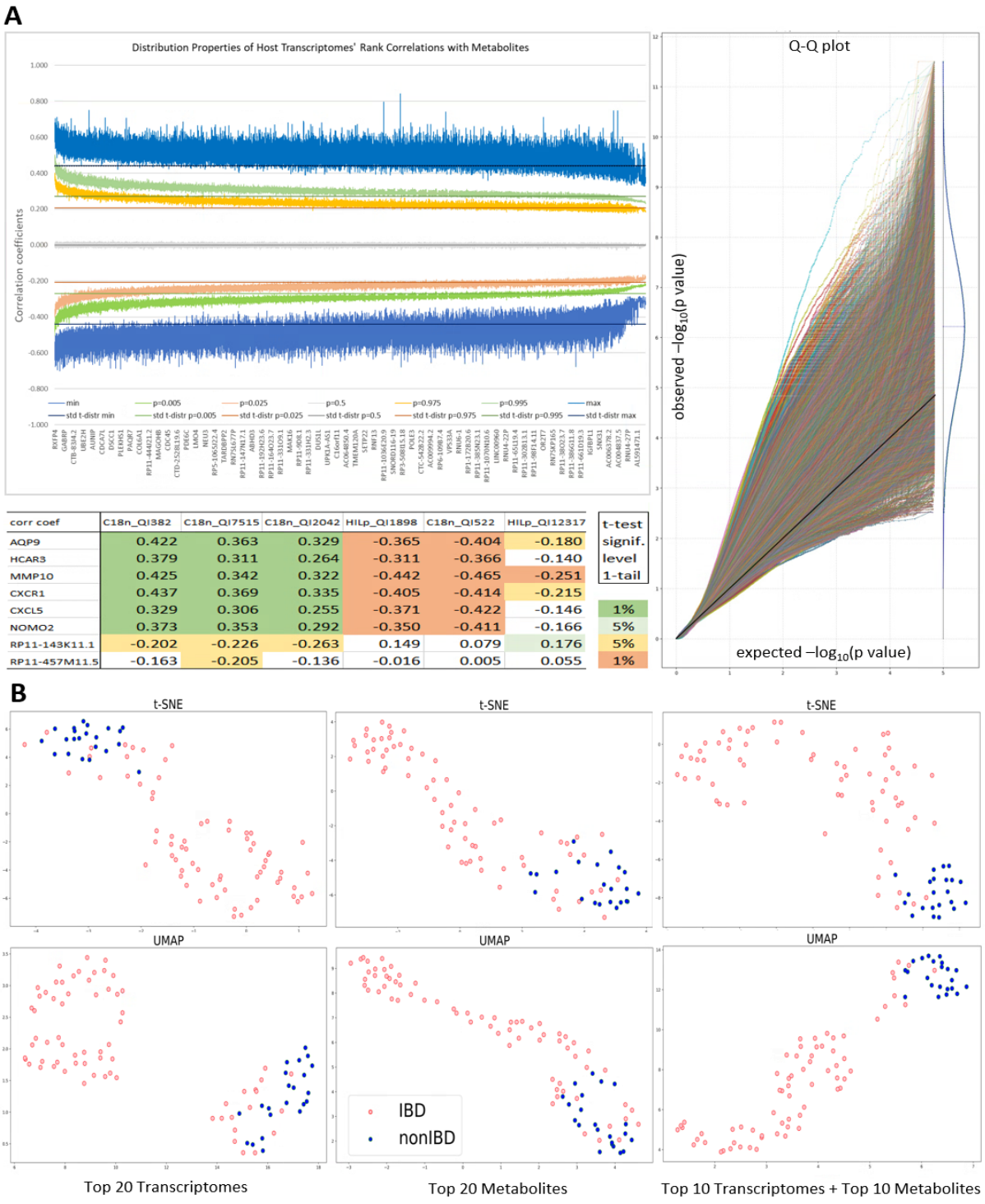
However, it is very difficult to disentangle the interactive effects or individual contributions of transcriptomes and metabolites on IBD due to their significant correlations. We first calculated their correlations using the averaged normalized omics abundances across 90 hosts. In the top left chart of **Figure 8.A,** the horizontal axis displays all 55765 transcriptomes, and the vertical axis displays distributional properties of pairwise correlation coefficients of each transcriptome with 81867 metabolites based on a total of 4.5 billion correlation calculations. The order of transcriptomes is arranged according to the descending range of correlation coefficients of two tails at a significant level of 1%. Compared with the standard t-distribution cumulative probability density levels of 0.5%, 2.5%, 97.5%, and 99.5% (the max and min levels are averaged numbers from bootstrapped samples based on random simulated runs of the standard t-distribution with matched sample sizes) with 88 degrees of freedom, the observed distributions of correlation coefficients appear to be far more significant at any threshold, supporting a multi-correlated notion between the two omics. In the right chart of **Figure 8.A**, the same correlation coefficients are presented in a Q-Q plot of observed $-\log_{10}$(p-value) against expected $-\log_{10}$(p-value) (assuming uniform p-value) where each line represents each transcriptome’s correlation with 81867 metabolites. In total, it also supports the multi-correlated notion between the two omics.

The table at the bottom left of **Figure 8.A** shows a total of 48 correlation coefficients between the 8 untargeted transcriptomes and 6 untargeted metabolites, of which 38 are correlation coefficients statistically significant at either 1% or 5% levels.

This evidence suggests that transcriptomes and metabolites may have reciprocal effects to each other in the ecosystem of hosts. Taking them together into consideration in a systematic way could aggregate host information better instead of looking at each piece of information in isolation. Again, we used t-SNE and UMAP as the initial steps for our integrative investigation. Both t-SNE and UMAP are unsupervised learning techniques for aggregating information to achieve dimension

reductions. It is to our comfort that they do not take the host phenotypes as input to avoid data overfitting(host phenotypes were superimposed on the same plots after the analysis).

In **Figure 8.B**, we chose only the top 20 transcriptomes (shown in the left panel), top 20 metabolites (shown in the middle panel), or top 10 transcriptomes + top 10 metabolites (shown in the right panel) as features for t-SNE and UMAP classification analysis over 90 hosts. Upon visual inspection, using 20 transcriptomes or 20 metabolites display similar differentiating powers between the reduced 2 dimension factors and host IBD traits when compared with using the top 200 features in **Figure 2.C** and **Figure 3.C**. However, combining the top 10 transcriptomes and top 10 metabolites together as features used by t-SNE and UMAP results in a significantly improved differentiation between hosts with and without IBD, leaving very few misclassified for non-IBD hosts, a big challenge when we use single omic for the same analysis. Upon visual inspection, only 2~3 hosts on the borderline could be misclassified, yielding a phenotype identification accuracy rate of ~96% for IBD hosts and ~90% for non-IBD hosts.



corr coef	C18n_QI382	C18n_QI7515	C18n_QI2042	HILp_QI1898	C18n_QI522	HILp_QI12317	t-test signif. level 1-tail
AQP9	0.422	0.363	0.329	-0.365	-0.404	-0.180	
HCAR3	0.379	0.311	0.264	-0.311	-0.366	-0.140	
MMP10	0.425	0.342	0.322	-0.442	-0.465	-0.251	
CXCR1	0.437	0.369	0.335	-0.405	-0.414	-0.215	
CXCL5	0.329	0.306	0.255	-0.371	-0.422	-0.146	1%
NOMO2	0.373	0.353	0.292	-0.350	-0.411	-0.166	5%
RP11-143K11.1	-0.202	-0.226	-0.263	0.149	0.079	0.176	5%
RP11-457M11.5	-0.163	-0.205	-0.136	-0.016	0.005	0.055	1%

Figure 8. Host Transcriptomes and Metabolites Interaction.

(A) The top left chart shows distributional properties of the pairwise correlation coefficients by each of 55765 transcriptomes. For any given transcriptome, one may find its quantile statistics for its correlations with all 81867 metabolites at quantile of min, 0.5%, 2.5%, 97.5%, 99.5%, and max compared with the standard t-distribution of corresponding quantile where the max and min levels are averaged numbers from bootstrapped samples based on randomly simulated runs of the standard t-distribution with matched sample sizes). The right chart displays the same correlation coefficients in a Q-Q plot of observed against expected -$\log_{10}$(p-value) where each line represents each transcriptome's correlation with 81867 metabolites. The table at the bottom reports the correlation coefficients between 8 untargeted transcriptomes and 6 untargeted metabolites.

(B) Three sets of t-SNE and UMAP on 90 hosts' average normalized abundances are plotted here: the left panel is based on top 20 transcriptomes, the middle panel is based on top 20 metabolites, and the right panel is based on top 10 transcriptomes combined with top 10 metabolites,

6 Cell type identification from metabolites and transcriptomes

To gain a better understanding of the mechanisms underlying abnormal pathological stats, it is important to classify and identify cells. Cell type identification allows us to find markers for specific cell types, which can broaden our understanding of environmental effects on microbiome biology. Smillie et al. found that many UC risk genes are cell type-specific and suggested a limited set of cell types and pathways, limiting functions for specific risk genes across genome-wide association region studies [30].

To explore IBD and map risk variants to specific cell types, we took advantage of our findings in the previous procedures on the significant linkages of some targeted transcriptomes with their untargeted metabolites and tried to identify cell types associated with the selected metabolites and transcriptomes. We analyzed two metabolites, namely HILn_QI82 and C18n_QI48, and the results are shown in **Figure 9**. For each metabolite, we selected the top 50 significant associated transcriptomes, and more than half of those can match the marker genes with MsigDB cell type (C8) dataset. Detailed marker genes and corresponding negative log p-values are shown in Fig. 5(B)-(C), with 32 marker genes for HILn_QI82 and 29 for C18n_QI48. These markers all have a p-value less than 10^{-4} from the t-test and are suitable for identifying the cell types.

We matched the transcriptomes and metabolites using the C8 set in MsigDB. The C8 set includes 704 subsets for cell type signature genes, and each subset corresponds to a specific type of cell. For each set of marker genes found in one metabolite, we intersect it with every subset of C8 and check the number of gene signatures that coincide. A test based on hyper-geometric distribution is conducted to check whether the intersection is significant in MsigDB.

Next, we conducted a hyper-geometric distribution test on the intersection between the marker genes found above and each subset in MsigDB C8. Under significant level 0.01, we found 17 cell types in HILn_QI82 while 31 in C18n_QI48, as in **Figure 9 (D)-(E)**. For each significant subset, it can match 2∼4 gene markers on average, and the most significant ones, such as Busslinger gastric mature pit cells and Busslinger duodenal late immature enterocytes, can match 5∼8 markers. Note that 11 subsets coincide with each other in two metabolites. Actually, these two metabolites both belong to lithocholate, and it is expected that common cell types be found in these two metabolites.

Figure 9. Matching cell types from metabolites and transcriptomes.

(**A**) The cell types identified from two metabolites HILn_QI82 (in green) and C18n_QI48 (in blue). (**B**) The top 50 significant transcriptomes and corresponding marker genes found from HILn_QI82 while (**C**) is from C18n_QI48. In (**B**) and (**C**), those transcriptomes matched with non-marker genes are omitted. (**D**) The cell types recognized from the marker genes in HILn_QI82, ranked by p-value in increasing order, while (**E**) is from C18n_QI48.

Note: the corresponding p-values (FDR-adjusted q-values in brackets) are in the negative $\log_{10}$ scale.

7 Mendelian Randomization

Mendelian randomization (MR) is a method used to infer causality between an exposure and an outcome. In this research, we conducted four MR tests using data from PhenoScanner, which included genome-wide association study (GWAS) data from the UK Biobank and expression quantitative trait loci (eQTL) data from the GTEx consortium [13,14,19,31,32]. Specifically, we tested in exposure vs outcome order, Treatment with UDCA vs IBD, RUNX1 vs IBD, Treatment with UDCA vs RUNX1, and RUNX1 vs Treatment with UDCA, where RUNX1 is a gene that impacts the development of hematopoietic stem cells [8]. For our MR tests, we used the simple median, weighted median, and penalized weighted median methods. We display our results from each test in **Table 7**, where we include the estimate, standard error, a 95% confidence interval for the estimate, and the negative log base 10 of the p-value.

Table 7. MR results from four tests

A. Treatment with UDCA vs IBD	Estimate	Std Error	95% CI		$-\log_{10}$(p-value)
Simple median	-42.8571	5.12789	-52.9076	-32.8067	16.19398
Weighted median	-42.7634	5.060178	-52.6811	-32.8456	16.53922
Penalized weighted median	-45.3362	5.523473	-56.162	-34.5104	15.64772
B. RUNX1 vs IBD					
Simple median	-0.00515	0.001303	-0.00771	-0.0026	4.117055
Weighted median	-0.00158	0.001334	-0.00419	0.001039	0.624157
Penalized weighted median	-0.00517	0.001336	-0.00779	-0.00255	3.961649
C. Treatment with UDCA vs RUNX1					
Simple median	-113.025	17.7662	-147.846	-78.2034	9.700193
Weighted median	-106.232	15.76359	-137.128	-75.3357	10.79748
Penalized weighted median	-112.215	19.27224	-149.987	-74.4417	8.237031
D. RUNX1 vs Treatment with UDCA					
Simple median	-0.00458	0.000623	-0.00581	-0.00336	12.71715
Weighted median	-0.00221	0.00041	-0.00302	-0.00141	7.17893
Penalized weighted median	-0.00203	0.00037	-0.00276	-0.00131	7.427514

Results from four MR tests, with p-values presented in negative log base 10.

In each of our tests, all of our obtained p-values were less than 0.05, with many values less than 10^{-10}, providing a putative causal relationship of RUNX1 gene expression and IBD, UDCA and

IBD, and a bidirectional relationship between RUNX1 and UDCA. We also include plots of our MR results in **Figure 10**.

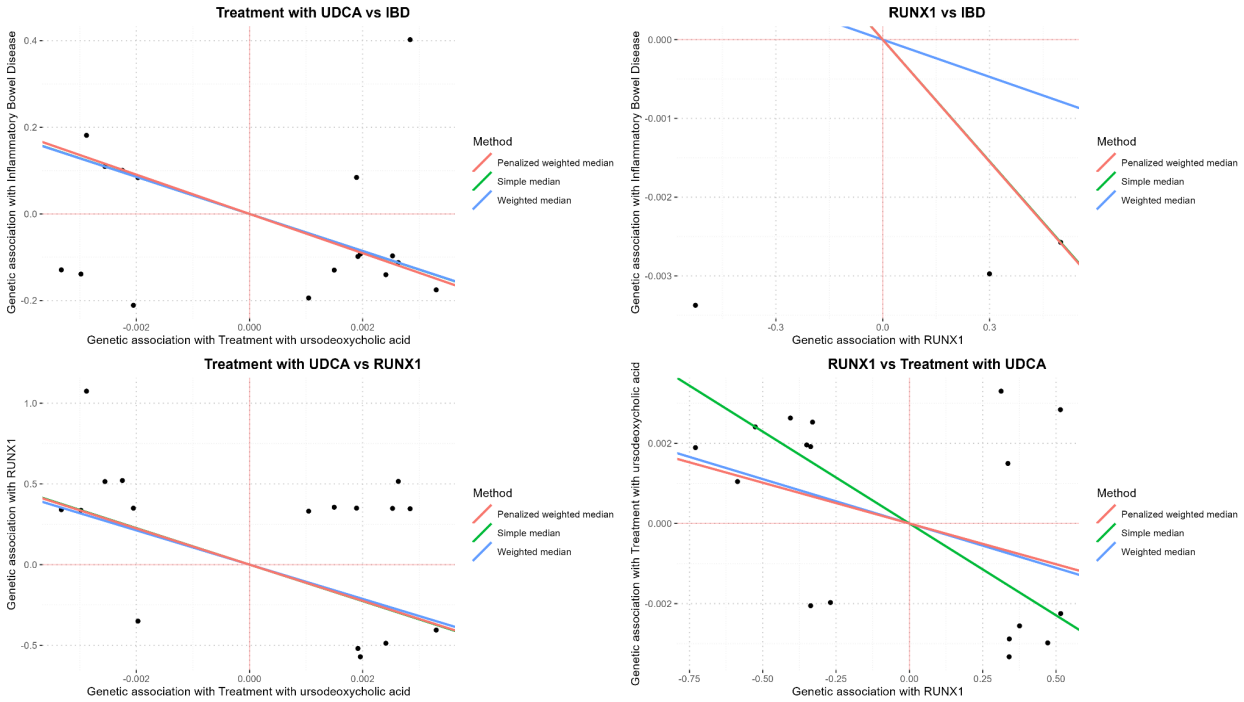


Figure 10. Four different mendelian randomization estimation results each using penalized weighted median, simple median, and weighted median methods.

The three methods we utilized are fairly consistent, nearly overlapping in the Treatment with UDCA vs IBD and Treatment with UDCA vs RUNX1 tests, corroborating our low p-values. We also illustrate these relationships with the causality network in **Figure 11**.

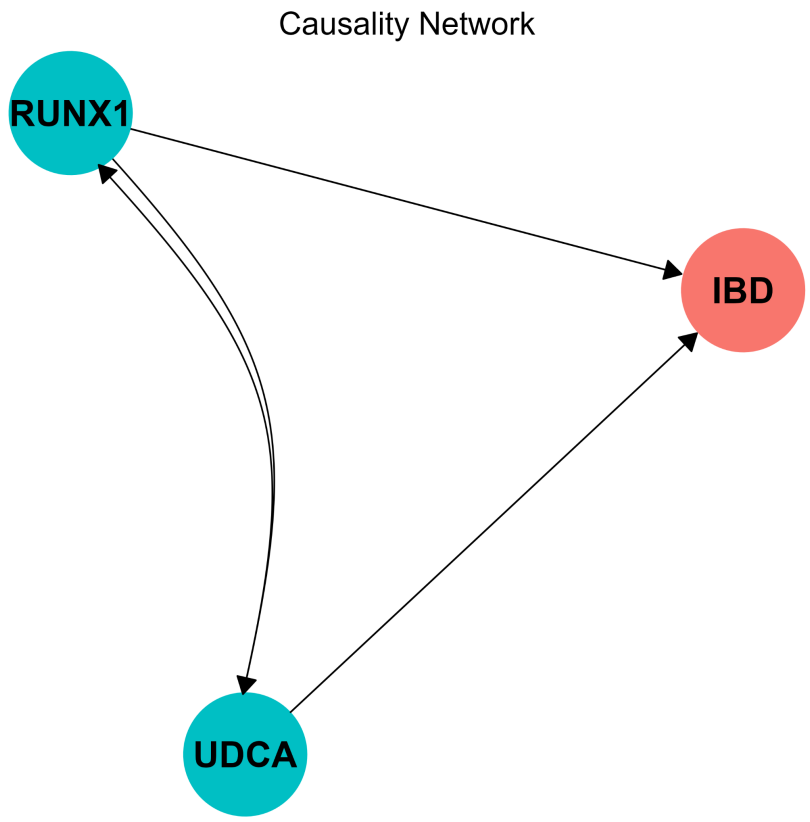


Figure 11. Causality network with directed causality between RUNX1 gene expression and IBD, as well as between UDCA and IBD. In addition, there is bidirectional causality between RUNX1 and UDCA.

RUNX1 and UDCA are connected with each other via bidirectional causality in **Figure 11**, with both being connected to IBD via directed causality.

8 Discussion

In summary, cutting-edge methodologies have been utilized to combine multi-omics datasets for data analysis and integrate multi-facet host-microbiome information to identify associations of the microbiome, metabolome, and host responses in IBD. Our analysis extensively covers all transcriptomes and metabolites for their statistical correlation and causality with IBD. It put forward a number of potential transcriptomes and metabolites for future studies to define their functional relevance in the complex etiology of IBD. Pooling samples of multi-omics data yielded a better association of host genetic and microbial biomarkers with host IBD phenotypes than using sample-based observations. With 10 transcriptomic features combined with 10 metabolic features based on t-SNE and UMAP classification schema, we were able to classify the host phenotypes on a reduced 2-dimensional space correctly for 87 hosts (out of 90 participants), a much better

association than using single-omics of 20 transcriptomes or 20 metabolites alone. Identifications of microbial statistical association with host IBD phenotypes are vital to the efficient searches for potential contributors with mechanisms underlying their interactions. In the future, combining the top features from significant multi-omics into a single model could be useful for building an improved classifier for IBD and treatment response based on patients' microbiomes.

Another significant contribution of this research is to identify the gene signature and corresponding cell types by testing the intersection of the metabolites and transcriptomes using the MsigDB C8. Our research has provided a new framework that enables us to study complex multi-omics in connection with host IBD traits and aims to benefit disease diagnosis and treatment, as well as drug development.

In addition, our correlation heatmaps and networks demonstrate an increased variety of correlations between cell types and metabolites, consistent with our previous research [28], and our MR causality analysis exemplifies the power of DWT in our framework.

While our research works toward providing a large-scale microbiome-driven host trait system that pushes the boundaries in exploring host trait correlation with microbiomes, we do acknowledge the various limitations of our findings. First, the sample size of our cohort is not big enough to generate very reliable statistical inferences. Larger studies with sufficient power to identify disease-specific associations are needed. Future continuations of this research may also include building structure to uncover meaningful information from extremely high dimensional factors' Riemann manifold by applying topological mapping methods. Secondly, we cannot predict disease events before their occurrence and cannot determine whether these associated multi-omic features of hosts are a consequence or cause of IBD. It is crucial that microbial molecular functions are identified in prospective studies in order to fully establish the mechanisms and causality of microbial influence in IBD. Last but not least, the inflammatory markers identified in this study were useful in distinguishing between hosts with or without IBD but not well enough to set apart CD from UC among IBD patients. The coming years in biomedical research may generate new techniques and data that, through experimental biology, enable a deeper understanding of diseases and their correlation to the human microbiome.

References

[1] Barrett, M. (2023). ggdag: Analyze and Create Elegant Directed Acyclic Graphs. https://CRAN.R-project.org/package=ggdag
[2] Bruns, T., Stallmach, A. (2009) Drug monitoring in inflammatory bowel disease: helpful or dispensable? Review Digestive Diseases. 27(3):394-403.
[3] Chun, E., Michaud, M. et al. (2019) The Crohn's disease polymorphism, ATG16L1 T300A, alters the gut microbiota and enhances the local Th1/Th17 response Elife 8:e39982.
[4] Csardi, G., & Nepusz, T. (2006). The igraph software package for complex network research. InterJournal, Complex Systems, 1695.
[5] Dahlhamer JM, Zammitti EP etal. (2015) Prevalence of Inflammatory Bowel Disease Among Adults Aged ≥18 Years — United States, 2015. MMWR Morb Mortal Wkly Rep 65:1166–1169
[6] Daubechies, I. (1992). Ten lectures on wavelets. Society for Industrial and Applied Mathematics. https://doi.org/10.1137/1.9781611970104
[7] Duerr, RH. et al. (2006) A Genome-Wide Association Study Identifies IL23R as an Inflammatory Bowel Disease Gene, Science. 314(5804): 1461–1463.
[8] Friedman, A. D. (2009). Cell cycle and developmental control of hematopoiesis by Runx1. Journal of Cellular Physiology, 219(3), 520–524. https://doi.org/10.1002/jcp.21738
[9] Gijsbers, K., Assche, G.V. et al. (2004) CXCR1-binding chemokines in inflammatory bowel diseases: down-regulated IL-8/CXCL8 production by leukocytes in Crohn's disease and selective GCP-2/CXCL6 expression in inflamed intestinal tissue. European Journal of Immunology 34(7): 1992-2000.
[10] Glocker, E.O. et al. (2009) Inflammatory Bowel Disease and Mutations Affecting the Interleukin-10 Receptor, New England J Med. 361(21): 2033–2045.
[11] Goyette, P., Boucher, G., Mallon, D. et al. (2015) High-density mapping of the MHC identifies a shared role for HLA-DRB1*01:03 in inflammatory bowel diseases and heterozygous advantage in ulcerative colitis. Nature Genetics 47, 172–179.
[12] Gasaly, N., Hermoso, M. A., & Gotteland, M. (2021) Butyrate and the Fine-Tuning of Colonic Homeostasis: Implication for Inflammatory Bowel Diseases. International journal of molecular sciences, 22(6), 3061.
[13] Ge, T., Chen, C.-Y., Neale, B. M., Sabuncu, M. R., & Smoller, J. W. (2017). Phenome-wide heritability analysis of the UK Biobank. PLOS Genetics, 13(4), e1006711. https://doi.org/10.1371/journal.pgen.1006711
[14] GTEx Consortium. (2015). Human genomics. The Genotype-Tissue Expression (GTEx) pilot analysis: Multitissue gene regulation in humans. Science (New York, N.Y.), 348(6235), 648–660. https://doi.org/10.1126/science.1262110
[15] Hickey, J. W., Becker, W. R., Nevins, S. A., Horning, A., Perez, A. E., Zhu, C., Zhu, B., Wei, B., Chiu, R., Chen, D. C., Cotter, D. L., Esplin, E. D., Weimer, A. K., Caraccio, C., Venkataraaman, V., Schürch, C. M., Black, S., Brbić, M., Cao, K., … Snyder, M. (2023). Organization of the human intestine at single-cell resolution. Nature, 619(7970), Article 7970. https://doi.org/10.1038/s41586-023-05915-x
[16] Horowitz, J.E., Warner, N., Staples, J. et al. (2021) Mutation spectrum of NOD2 reveals recessive inheritance as a main driver of Early Onset Crohn's Disease. Nature Sci Rep 11, 5595.
[17] Huang, Q., Zhang, X., and Hu, Z. (2021) Application of Artificial Intelligence Modeling Technology Based on Multi-Omics in Noninvasive Diagnosis of Inflammatory Bowel Disease. J. of Inflamm Res.14: 1933–1943.
[18] Joehanes, R., Zhang, X., Huan, T., Yao, C., Ying, S.-X., Nguyen, Q. T., Demirkale, C. Y., Feolo, M. L., Sharopova, N. R., Sturcke, A., Schäffer, A. A., Heard-Costa, N., Chen, H., Liu, P.-C., Wang, R., Woodhouse, K. A., Tanriverdi, K., Freedman, J. E., Raghavachari, N., … Munson, P. J. (2017). Integrated genome-wide analysis of expression quantitative trait loci aids interpretation of genomic association studies. Genome Biology, 18(1), 16. https://doi.org/10.1186/s13059-016-1142-6
[19] Kamat, M. A., Blackshaw, J. A., Young, R., Surendran, P., Burgess, S., Danesh, J., Butterworth, A. S., & Staley, J. R. (2019). PhenoScanner V2: An expanded tool for searching human genotype-phenotype associations. Bioinformatics (Oxford, England), 35(22), 4851–4853. https://doi.org/10.1093/bioinformatics/btz469
[20] Koller, F., Dozier, E.A. et al. (2012) Lack of MMP10 exacerbates experimental colitis and promotes development of inflammation-associated colonic dysplasia. Nature Laboratory Investigation volume 92, 1749–1759.
[21] Lavelle, A., Sokol, H. (2020) Gut microbiota-derived metabolites as key actors in inflammatory bowel disease. Nature Rev Gastroenterol Hepatol 17, 223–237.

[22] Lloyd-Price, J., Arze, C., Ananthakrishnan, A.N. et al. (2019) Multi-omics of the gut microbial ecosystem in inflammatory bowel diseases. Nature 569, 655–662.
[23] Medina, J. M., Fernández-López, R. et al. (2017) Propionate Fermentative Genes of the Gut Microbiome Decrease in Inflammatory Bowel Disease. Journal of clinical medicine, 10(10).
[24] Mehto, S., Jena, K.K. et al (2019) The Crohn's Disease Risk Factor IRGM Limits NLRP3 Inflammasome Activation by Impeding Its Assembly and by Mediating Its Selective Autophagy Molecular. Cell 73, 429–445.
[25] Olena Yavorska & James Staley. (2023). MendelianRandomization: Mendelian Randomization Package. https://CRAN.R-project.org/package=MendelianRandomization
[26] Pfister, S., Kuettel, V., & Ferrero, E. (2023). granulator: Rapid benchmarking of methods for *in silico* deconvolution of bulk RNA-seq data (1.8.0) [R]. Bioconductor version: Release (3.17). https://doi.org/10.18129/B9.bioc.granulator
[27] R: The R Project for Statistical Computing. (n.d.). Retrieved August 18, 2023, from https://www.r-project.org/
[28] Shankar, A., Chang, S., Zhao, Y., Wang, X., & Liu, T. (2022). Wavelet-Based Microbiome Correlations of Host Traits. Proceedings of the 2022 6th International Conference on Computational Biology and Bioinformatics, 13–19. https://doi.org/10.1145/3589437.3589440
[29] Singh,U.P., Singh, N,P. et al. (2016) Chemokine and cytokine levels in inflammatory bowel disease patients. Cytokine. 77: 44–49.
[30] Smillie, C.S., Biton, S. et al., (2019) Intra- and Inter-cellular Rewiring of the Human Colon during Ulcerative Coliti. Cell 178, 714–730.
[31] Staley, J. R., Blackshaw, J., Kamat, M. A., Ellis, S., Surendran, P., Sun, B. B., Paul, D. S., Freitag, D., Burgess, S., Danesh, J., Young, R., & Butterworth, A. S. (2016). PhenoScanner: A database of human genotype-phenotype associations. Bioinformatics (Oxford, England), 32(20), 3207–3209. https://doi.org/10.1093/bioinformatics/btw373
[32] Sudlow, C., Gallacher, J., Allen, N., Beral, V., Burton, P., Danesh, J., Downey, P., Elliott, P., Green, J., Landray, M., Liu, B., Matthews, P., Ong, G., Pell, J., Silman, A., Young, A., Sprosen, T., Peakman, T., & Collins, R. (2015). UK Biobank: An Open Access Resource for Identifying the Causes of a Wide Range of Complex Diseases of Middle and Old Age. PLOS Medicine, 12(3), e1001779. https://doi.org/10.1371/journal.pmed.1001779
[33] Ward, J., Lajczak, N. K. et al. (2017) Ursodeoxycholic acid and lithocholic acid exert anti-inflammatory actions in the colon. American journal of physiology. Gastrointestinal and liver physiology, 312(6), G550–G558.
[34] Westermann, A.J., Vogel, J. (2021) Cross-species RNA-seq for deciphering host–microbe interactions. Nat Rev Genet 22, 361–378.
[35] Wickham, H. (2016). ggplot2: Elegant Graphics for Data Analysis. Springer-Verlag New York. https://ggplot2.tidyverse.org
[36] Wnorowski, A., Wnorowska, S. et al. (2021) Alterations in Kynurenine and NAD+ Salvage Pathways during the Successful Treatment of Inflammatory Bowel Disease Suggest HCAR3 and NNMT as Potential Drug Targets. International Journal of Molecular Sciences 22(24):13497.
[37] Yang, Y. et al. (2021) RF5 Acts as a Potential Therapeutic Marker in Inflammatory Bowel Diseases. Inflammatory Bowel Diseases, 27(3), 407–417.
[38] Yilmaz, B. et al. (2018) The presence of genetic risk variants within PTPN2 and PTPN22 is associated with intestinal microbiota alterations in Swiss IBD cohort patients PLoS One. 13(7): e0199664. Published online.
[39] Yu, B, Yin, Y. et al. (2021) Diagnostic and Predictive Value of Immune-Related Genes in Crohn's Disease. Frontiers in Immunology 12: 643036.
[40] Zhao, S., Gong, Z. et al. (2016) Deoxycholic Acid Triggers NLRP3 Inflammasome Activation and Aggravates DSS-Induced Colitis in Mice. Frontiers in immunology, 7, 536.