OPEN ACCESS

ORIGINAL ARTICLE

Molecular classification of Crohn's disease reveals two clinically relevant subtypes

Matthew Weiser,[1,2] Jeremy M Simon,[1] Bharati Kochar,[2,3] Adelaide Tovar,[3,4] Jennifer W Israel,[1] Adam Robinson,[3] Gregory R Gipson,[3] Matthew S Schaner,[3] Hans H Herfarth,[3] R Balfour Sartor,[3] Dermot P B McGovern,[5] Reza Rahbar,[6] Timothy S Sadiq,[6] Mark J Koruda,[6] Terrence S Furey,[1,2,7] Shehzad Z Sheikh[1,2,3,4]

► Additional material is published online only. To view please visit the journal online (http://dx.doi.org/10.1136/gutjnl-2016-312518).

For numbered affiliations see end of article.

Correspondence to
Dr Shehzad Z Sheikh, Departments of Medicine and Genetics, University of North Carolina at Chapel Hill, 7340B Medical Biomolecular Research Building, Chapel Hill, NC 27517, USA; sheisx@med.unc.edu

MW, JMS, TSF and SZS contributed equally.

Received 27 June 2016
Revised 09 September 2016
Accepted 18 September 2016
Published Online First 13 October 2016

ABSTRACT

Objective The clinical presentation and course of Crohn's disease (CD) is highly variable. We sought to better understand the cellular and molecular mechanisms that guide this heterogeneity, and characterise the cellular processes associated with disease phenotypes.

Design We examined both gene expression and gene regulation (chromatin accessibility) in non-inflamed colon tissue from a cohort of adult patients with CD and control patients. To support the generality of our findings, we analysed previously published expression data from a large cohort of treatment-naïve paediatric CD and control ileum.

Results We found that adult patients with CD clearly segregated into two classes based on colon tissue gene expression—one that largely resembled the normal colon and one where certain genes showed expression patterns normally specific to the ileum. These classes were supported by changes in gene regulatory profiles observed at the level of chromatin accessibility, reflective of a fundamental shift in underlying molecular phenotypes. Furthermore, gene expression from the ilea of a treatment-naïve cohort of paediatric patients with CD could be similarly subdivided into colon-like and ileum-like classes. Finally, expression patterns within these CD subclasses highlight large-scale differences in the immune response and aspects of cellular metabolism, and were associated with multiple clinical phenotypes describing disease behaviour, including rectal disease and need for colectomy.

Conclusions Our results strongly suggest that these molecular signatures define two clinically relevant forms of CD irrespective of tissue sampling location, patient age or treatment status.

Significance of this study

What is already known on this subject?

- Crohn's disease (CD) is a clinically heterogeneous disease with variable presentation and progression.
- Current therapies for patients with CD are largely based on subjective clinical classifications resulting in an inconsistent response. Unlike other complex diseases, notably certain cancers, there is currently no way to determine a personalised approach to therapy.
- Genetic association studies suggest that underlying molecular or genetic factors contribute to disease heterogeneity, but precise mechanisms are poorly understood.

What are the new findings?

- Gene expression and chromatin accessibility can subdivide CD into at least two distinct molecular subclasses that associate with specific disease phenotypes.
- These subclasses exhibit stark expression differences in cellular metabolism and immune pathways.
- Similar molecular phenotypes are also observed in treatment-naïve paediatric patients with CD, thus defining true disease subclasses independent of treatment history and patient age.

How might it impact on clinical practice in the foreseeable future?

- Our findings thus provide evidence of a molecular basis for the observed heterogeneity in CD. They also offer a unique opportunity to begin exploring customised treatment options guided by disease subclass providing novel therapeutic avenues and predictors of disease outcomes.

INTRODUCTION

Crohn's disease (CD) is a chronic heterogeneous inflammatory disorder with distinct patterns of clinical behaviour. CD may present or evolve with time into a more complex phenotype with patients developing strictures, fistulae and/or abscesses, and many patients experience highly variable response to therapies. Genetic associations[1 2] and a recently defined lipid metabolism-based gene expression signature predictive of disease involvement[3] suggest that molecular or genetic factors are associated with and may contribute to disease heterogeneity, but precise mechanisms are poorly understood. Molecular subtypes defined by gene expression that impact clinical phenotypes have also been documented in other complex diseases, especially cancers.[4–6] Whether adult CD can be similarly separated into two or more subgroups and whether

To cite: Weiser M, Simon JM, Kochar B, *et al*. *Gut* 2018;**67**:36–42.

these molecular classes can explain disease phenotypes remains largely unknown.

Understanding how genes are regulated provides complementary information to gene expression. We and others have studied gene regulation by focusing on accessible chromatin, which allows transcriptional regulators to bind to the otherwise highly condensed nuclear genome. Chromatin accessibility has been associated with promoters, enhancers, silencers and insulators,[7] and changes as cellular identity is established through differentiation and development[8 9] or in response to cellular stresses. Chromatin profiling provides a mechanism as to why expression is changing, and whether observed changes may be transient or persistent. We have shown that chromatin accessibility can differentiate disease subtypes[10] and helps to describe genetic and environmental contributors to disease.[11] Therefore, we sought to determine whether multiple clinically distinct subclasses of adult CD exist by examining both gene expression, using RNA-seq, and chromatin accessibility, using formaldehyde-assisted isolation of regulatory elements (FAIRE-seq),[12] in unaffected colon mucosa from patients with CD and non-IBD.

RESULTS

Whole genome interrogation of the colonic transcriptional and chromatin landscape reveals two distinct molecular classes in CD

To determine, in an unbiased manner, whether gene expression levels separated samples into distinct molecular groups, we performed a principal components analysis (PCA) using gene expression profiles from a combined set of 21 patients with CD and 11 patients with non-IBD. A striking clustering pattern emerged, whereby the individuals with CD were divided into two distinct, expression-based subclasses, one of which clustered with the non-IBD controls (figure 1A). To specifically interrogate these two CD subclasses, we identified genes differentially expressed between these two groups of patients with CD (849 genes at False Discovery Rate (FDR)<0.05; figure 1B, see online supplementary table S1 for top 20 differentially expressed genes in each CD subclass). Surprisingly, when looking at the top 25 differentially expressed genes regardless of direction, most had tissue-specific expression patterns that discriminated colon from the small intestine (ileum), including *NXPE4*, *CWH43* and *CA2* (colon-specific) as well as *RBP2*, *TM6SF2*, *APOB*, *MTTP*, *CREB3L3* and *CPS1* (ileum-specific).[13] CD samples similar to non-IBD controls (figure 1A) exhibited abundant expression of the above colon-specific genes, whereas the other CD subclass showed expression patterns more consistent with ileum despite being sampled from the colon. To explore this more globally, we compared all these differentially expressed genes with 947 genes with known significant differential expression between colon and ileum (figure 1C).[14] We found that 34% of the genes more highly expressed in the colon-like CD samples were indeed markers of normal colon, and 44% of ileum marker genes were more highly expressed in the ileum-like CD samples ($p<1\times10^{-95}$, hypergeometric test). To validate these expression differences, we performed reverse transcription-quantitative PCR on 18 CD samples of unaffected colon mucosa (9 colon-like, 9 ileum-like) using *CEACAM7* and *APOA1* as a proxy for colon-like and ileum-like expression patterns (see online supplementary figure S1A). In agreement with the RNA-seq data, *CEACAM7* was significantly more abundant in colon-like CD samples (p=0.017, one-sided t-test), whereas *APOA1* was significantly more abundant in ileum-like CD samples (p=0.020, one-sided t-test).

To determine whether these expression changes represented a fundamental shift in the functional cellular identity of these tissues, we investigated chromatin accessibility by performing FAIRE-seq[12] on the same samples from both CD subclasses. Supporting a fundamental shift in underlying molecular phenotypes, we identified 3339 300-bp regions with significantly differential chromatin accessibility between colon-like and ileum-like CD samples (figure 1D; p<0.05, two-sided t-test), hereafter referred to as differentially accessible regions (DARs). These DARs could be divided into two classes based on greater accessibility in colon-like or ileum-like CD samples, and further, an unsupervised PCA of FAIRE-seq data nearly separated ileum-like from colon-like CD subclasses (see online supplementary figure S1C). Subclass-specific changes in the chromatin landscape corresponded strongly to differences in nearby (within 50 kb) gene expression (figure 1E,F; p≤0.01, permutation). Additionally, both colon-specific and ileum-specific DARs exhibited a significant enrichment for CD genome-wide association study (GWAS) loci[15] compared with what was expected due to random chance (colon-specific p=0.018, ileum-specific p=0.006; permutation), suggesting that changes in chromatin accessibility occur at disease-relevant regions of the genome.

We next sought to annotate these DARs based on tissue-specific gene regulatory information. Post-translational modifications on histone proteins serve to compartmentalise the genome and demarcate putative function of regulatory elements.[16] Using Chromatin Immunoprecipitation (ChIP-seq) data from the Roadmap Epigenomics Project,[17] we assessed the enrichment of six histone modifications reflective of underlying regulatory activity (active: H3K4me1, H3K4me3, H3K27ac, H3K36me3; repressive: H3K27me3, H3K9me3) around colon-specific and ileum-specific DARs. We found that colon-specific DARs were demarcated by H3K27ac and H3K4me1 modifications present in colon but not ileum (see online supplementary figure S1B), suggesting these DARs function as active regulatory regions only in the normal colon. In contrast, ileum-specific DARs demonstrated positive H3K27ac and H3K4me1 enrichment found only in normal small intestine, despite these samples originating from colon tissue. These suggest that regulatory activity in DARs contribute to the colon-like and ileum-like expression levels. To confirm regulatory activity, we cloned three DARs (two with colon-specific and one with ileum-specific chromatin accessibility) into luciferase vectors upstream of a minimal promoter in both orientations using THP-1 monocytes (see online supplementary figure S1D). Relative to empty vector controls, two DARs (associated with *SATB2-AS1* and *DEPDC7*) exhibited a significant increase (p<0.01, one-sided t-test) in luciferase activity in both orientations, strongly suggestive of enhancer function. The third DAR (associated with *SLC16A9*) also enhanced luciferase activity significantly ($p=8.9\times10^{-5}$, one-sided t-test), however, only in the reverse orientation.

Together, these data support the existence of two molecularly distinct subclasses of CD. Furthermore, chromatin accessibility data suggest these subclasses exist due to stable molecular transformations of the genomic architecture in colon tissue cells, and not transient differences due to external cellular signalling.

Whole genome RNA-seq analysis reveals colon-like and ileum-like subclasses in treatment-naïve paediatric patients with CD

Gene expression profiles in adult patients with CD may vary due to treatment history. Therefore, we sought to determine whether treatment-naïve paediatric patients with CD also segregated into similar molecular classes. We performed PCA on

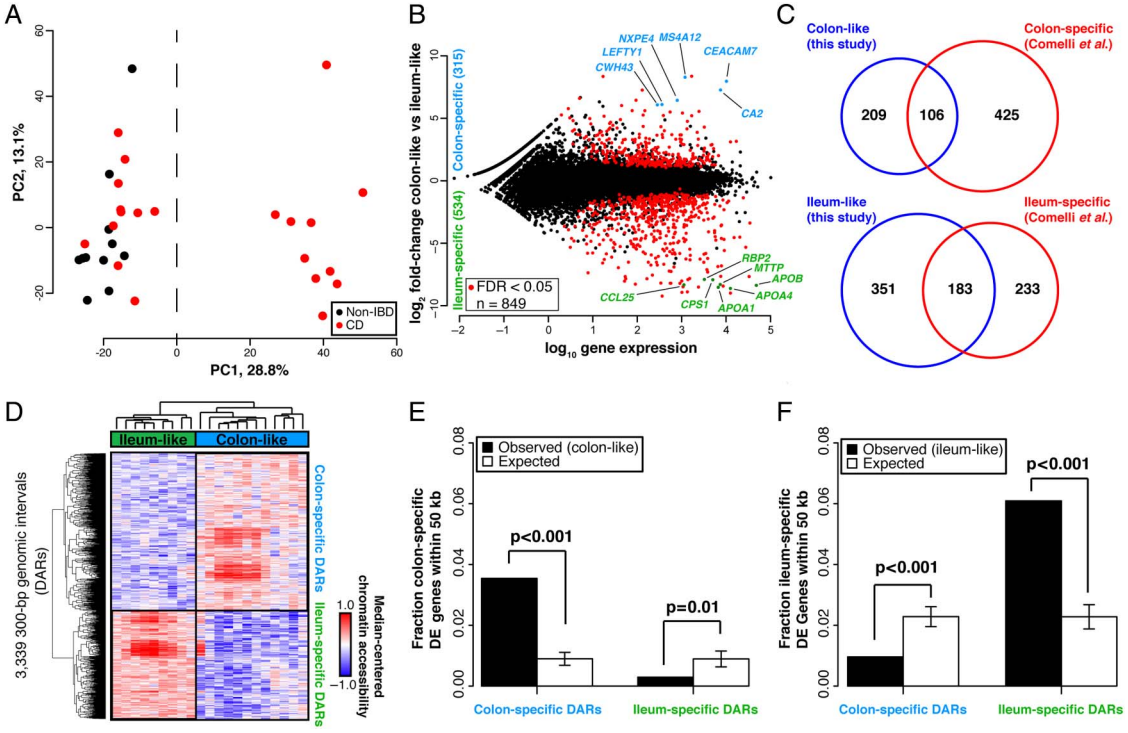

Figure 1 Two distinct molecular subtypes in adult Crohn's disease (CD). (A) Principal components analysis (PCA) analysis of RNA-seq data from colon tissue from 21 patients with CD and 11 patients with non-IBD shows two distinct clusters. (B) Eight hundred and forty-nine genes are differentially expressed between the two CD subclasses (adjusted p<0.05, DEseq), defined as colon-like and ileum-like. Known markers of colon and ileum are highlighted. (C) Genes upregulated in colon-like (top) and ileum-like (bottom) CD subclasses overlap previously defined colon-specific and ileum-specific genes, respectively. (D) Differentially accessible regions (DARs) identified using formaldehyde-assisted isolation of regulatory elements (FAIRE-seq) (p<0.05, two-sided t-test, normalised read counts, 300 bp windows) show distinct profiles in colon-like and ileum-like CD samples. (E) Colon-like-associated DARs are enriched and ileum-like-associated DARs are depleted near genes upregulated in colon-like samples (p≤0.01, permutation test). (F) Ileum-like-associated DARs are enriched and colon-like-associated DARs are depleted near genes upregulated in ileum-like samples (p<0.001, permutation test).

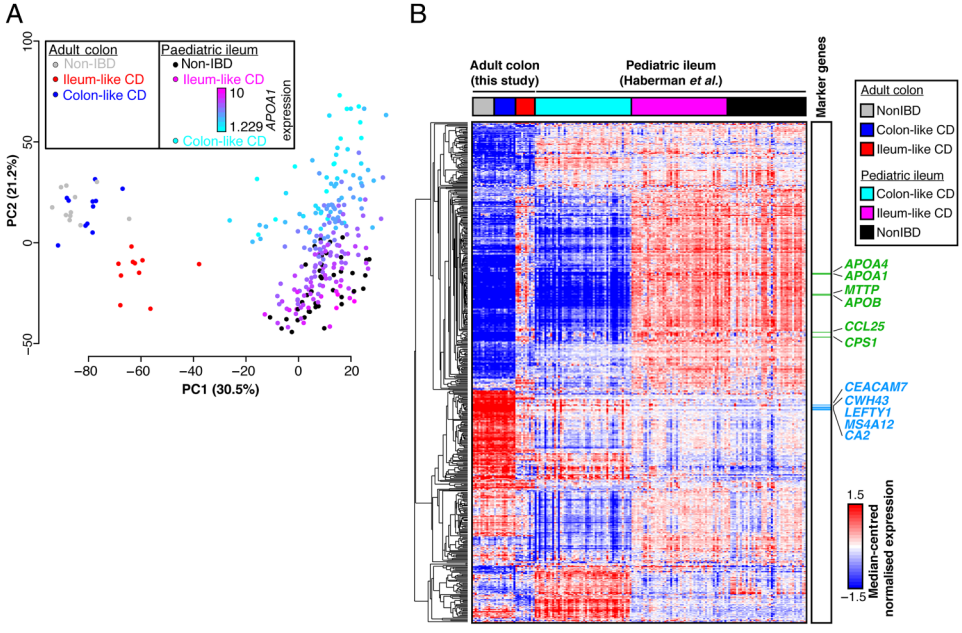

Figure 2 Treatment-naïve paediatric Crohn's disease (CD) samples show similar molecular subtypes. (A) Principal components analysis (PCA) analysis of combined RNA-seq data from adult colon tissue and paediatric ileum tissue from patients with CD and non-IBD shows separation of tissue types (PC1) and replicates ileum-like and colon-like clusters (PC2). Expression of *APOA1* (blue-pink, low-high) in paediatric samples aligns well with subclasses. (B) Hierarchical clustering of RNA-seq data using 500 colon-specific and ileum-specific genes show clusters of genes associated with ileum-like and colon-like samples across both the adult colon and paediatric ileum cohorts, as well as genes associated with tissue of origin.

previously published RNA-seq data from ileal biopsies in age-matched paediatric patients with CD (n=201) and non-IBD (n=40) generated within the Pediatric Risk Stratification Study.[3] Although a clustering as distinct as with the adult samples was not observed, non-IBD ileum samples clustered with some CD samples along the first principal component, whereas the other CD samples were separate (see online supplementary figure S2). To determine whether this pattern was related to the adult CD molecular subtypes, we performed PCA on combined adult colon and paediatric ileum expression data (figure 2A). Unsurprisingly, samples predominantly separated by tissue of origin (colon vs ileum; first principal component). However, a

separation indicative of two molecular subclasses was evident along the second principal component, and correlated nearly exactly with the first principal components in single cohort PCAs (see figure 1A online supplementary figure S2). Furthermore, paediatric CD samples fell on a spectrum highly correlated with expression of *APOA1*, a marker gene of the ileum and indicator of disease outcome in the paediatric cohort.[3] This pattern aligned well with the ileum-like (*APOA1*-high) and colon-like (*APOA1*-low) subclasses we identified in the adult CD colon.

To closely examine the relationship between the two CD subclasses across the adult and paediatric cohorts, we assessed gene expression patterns across the 500 most variably expressed known colon and ileum marker genes[14] using hierarchical clustering (figure 2B). To focus this analysis, we selected the 50 paediatric ileum samples each that were most colon-like and most ileum-like based on the PCA (figure 2A, second principal component). Many of the colon and ileum representative genes described above (e.g., *APOA1*, *CEACAM7*, *MTTP*, *LEFTY1* and *CA2*) exhibited highly consistent expression patterns across all samples in a defined molecular subclass, regardless of cohort. Interestingly, for these 500 genes, expression patterns were extremely consistent between colon-like CD and non-IBD colon samples, as well as between ileum-like CD and non-IBD ileum samples. Importantly, we note that a subset of genes differentiate all colon tissues from all ileum tissues indicating that tissue-of-origin-specific expression is not completely lost. Together, these data strongly suggest that the colon-like and ileum-like molecular signatures define two forms of CD present regardless of tissue sampling location, patient age or treatment status.

Metabolic and immune activation gene expression profiles characterise distinct molecular phenotypes in adult and paediatric CD

Gene expression differences between colon-like and ileum-like subclasses in the adult and paediatric CD cohorts went beyond the marker genes described above (figure 1B,C). To evaluate this more broadly, we computed and compared pathway-level expression patterns of both CD subclasses and non-IBD controls in both patient cohorts.[18 19] We then grouped significantly altered pathways based on similarity in both gene composition and direction of expression difference (see figure 3 online supplementary figures S3 and S4). First, numerous pathways related to interferon signalling, G-protein coupled receptor (GPCR) signalling, and antigen processing were significantly upregulated in patients with CD as a whole relative to non-IBD controls in both cohorts (see online supplementary figure S3: 'CD-enriched', red). Given that the adult patient cohort consisted of disease-unaffected tissue, and the paediatric cohort was treatment-naïve, this suggests a basal activation of the immune system in CD. In contrast, many pathways related to RNA processing, translation and transcription were downregulated in CD in the paediatric patients (see online supplementary figure S3: 'CD-depleted', blue). These results strongly corroborate studies in mice and humans linking overall defects in cellular protein processing in CD to immune activation and unabated inflammation.[20]

Next, we identified significantly altered pathways that described how the ileum-like and colon-like CD subclasses differed. Among the most pronounced effects in both cohorts reflected strong differences in metabolic activity including pathways involved in lipid metabolism and metabolism of foreign (xenobiotic) agents (figure 3A: 'ileum-like-enriched', red; figure 3B: 'colon-like-depleted', blue). Interestingly, energy production by way of the tricarboxylic acid (TCA) cycle was significantly affected in opposing ways: in adults, it was upregulated in the colon-like class and simultaneously downregulated in the ileum-like class in colon tissue (figure 3A: 'colon-like-enriched, ileum-like-depleted', pink); whereas in paediatric patients it was downregulated in the colon-like class and upregulated in the ileum-like class in ileal tissue (figure 3B: 'ileum-like-enriched, colon-like-depleted', purple). This suggests that energy production increases in patients where the subtype is more similar to the tissue of origin, and decreases when gene expression adopts patterns of the opposite tissue. In addition, several pathways related to GPCR signalling were upregulated in the ileum-like subclass in colon tissue (figure 3A: 'ileum-like-enriched', red) and upregulated in the colon-like subclass in ileum tissue (figure 3B: 'colon-like-enriched', green). GPCRs are highly expressed in monocytes and macrophages central to the development and progression of inflammation in CD, mainly through migration and accumulation within the inflamed tissues.[21]

Taken together, dysregulation of metabolic pathways may represent defining features of CD subtypes. Although dysregulation of lipid metabolism has been previously described in CD,[3] our data suggest these alterations may be specific only to patients within a certain subclass and dependent on the tissue being assayed. Furthermore, these data indicate that despite a striking similarity in expression of key ileum and colon marker genes (figure 2B), there are key differences in pathway-level expression patterns between adult patients with CD and paediatric patients with CD, such as the immune response (e.g., nucleotide-binding oligomerization domain (NOD) signalling, toll-like receptor (TLR) signalling signalling, interleukin signalling), which point towards clinically relevant phenotypes and characteristics of each subclass.

Molecular subclasses are associated with clinical outcomes

To understand the impact of molecular phenotype on clinical disease, we studied the clinical characteristics of treatment-naïve paediatric patients with respect to molecular subclassification. We again used the 50 ileum-like and 50 colon-like paediatric samples defined above. We found that significantly more patients with colon-like CD displayed both colon and ileum involvement ($p=0.0014$), had deep ulcers ($p=0.0002$) and showed macroscopic inflammation ($p=0.0156$), whereas more patients with ileum-like CD showed no inflammation ($p=0.0122$) and a strong skew in prevalence towards colon-only involvement ($p=0.0528$; table 1).

We next performed an in-depth retrospective chart review and a prospective follow-up of all adult patients with CD (see table 1 and online supplementary table S1) to see if a similar clinical impact of molecular subclass was observed. Patients with colon-like CD had greater rectal disease involvement ($p<0.01$) and were more likely to eventually need a colectomy ($p=0.01$). Patients with ileum-like CD tended to have ileal disease ($p=0.06$) and required the use of postoperative biological therapy ($p=0.02$). Although our sample size is small, these data suggest that molecular subtypes of CD can stratify patients into clinically distinct and relevant subgroups, and may prospectively identify those more likely to require intensive medical therapy.

DISCUSSION

We identified two distinct molecular phenotypes in colon tissue obtained from adult patients with CD. When we applied the same analytical approach to paediatric RNA-seq data from ileum, two molecular phenotypes also emerged, although the clinical consequences in both circumstances were relevant to the population of study and limited to the data collected for each population in terms of clinical phenotype. Although the adult patient sample size was limited, in only the colon-like CD

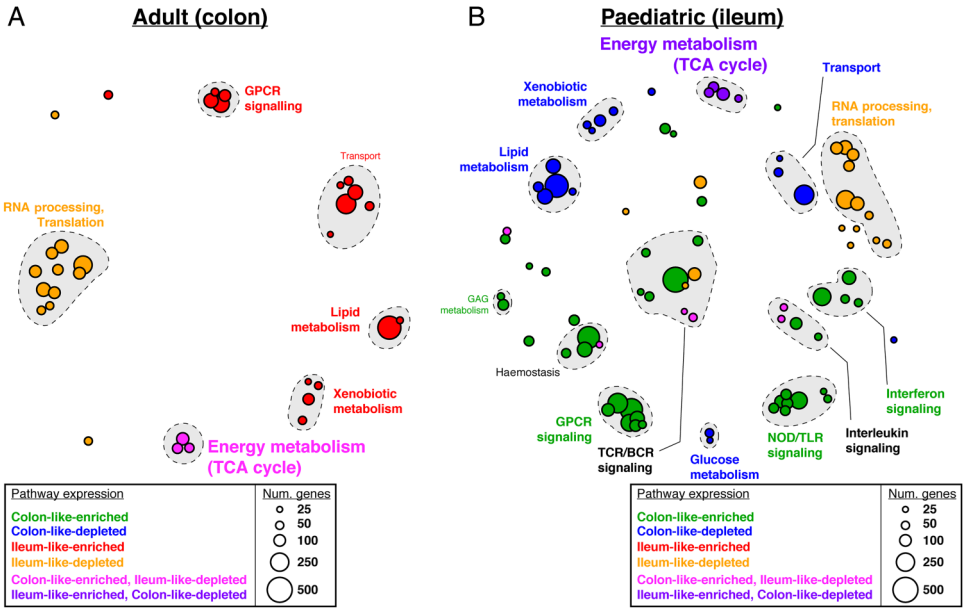

Figure 3 Pathways enriched for differentially expressed genes associated with Crohn's disease (CD) phenotypes. Pathway enrichments were determined using gene set association analysis (GSAA) (FDR<0.1, permutation test) for all pairwise comparisons between all CD, colon-like CD, ileum-like CD and non-IBD samples. Separate analyses in (A) adult colon and (B) paediatric ileum show similar and unique pathways associated with CD phenotypes. Each circle represents a Reactome-defined pathway with the size reflecting the number of genes in the pathway. Pathways were grouped based on similarity in gene membership, and labels describing multipathway clusters are shown. See online supplementary figures S3 and S4 for CD versus non-IBD comparisons and full list of pathways.

subclass were there patients with rectal disease, or that required a colectomy. Rectal CD is particularly difficult to manage, and although it may represent a unique CD phenotype, its underlying molecular mechanisms are unknown.[22–24] Paediatric patients with colon-like disease were more likely to have macroscopic inflammation, deep ulcers and involvement of both the ileum and colon. These data emphasise the need to continue and expand these studies over time to incorporate the evolving clinical phenotype in both adult and paediatric patients, and the need to study both tissues in the same patient. Only through such studies will we uncover whether the tissue of origin dictates potential clinical phenotype similarly.

The ileum-like CD subclass was primarily characterised by an upregulation of pathways involved in lipid and xenobiotic metabolism. The mechanistic link between these clinical phenotypes and the lipid metabolic signature defined here is yet to be explored; however, the implication that lipid metabolism may be involved is interesting for several reasons. First, lipid metabolism and altered levels of certain lipids have been previously associated with IBD (reviewed in [25]), and may be linked to inflammation state and immune signalling (reviewed in [26]), in addition to diet and intestinal microbiota composition.[27] Lipid levels may also be a cause or result of increases in intestinal oxidative stress, which has also been shown to be elevated in patients with CD (reviewed in [28]). Notably, two of the top differentially expressed genes between CD subclasses were in fact *GSTA1* and *GSTA2*, key mediators of the oxidative stress response.[29] Furthermore, diet, cholesterol and microbiota composition have each been studied either themselves as potential therapies or how their levels change as a result of surgery or biologic use (reviewed in [25]).

The diversity in cellular state demonstrated at both the transcriptional and chromatin levels has important potential therapeutic implications. Our results suggest that additional testing of patients with CD for particular molecular signatures, especially metabolic pathways, may determine potential therapeutic subgroups. This approach has already increased our understanding of human cancer biology. Patient subclassification in various forms of cancer, most notably invasive breast tumours,[4 5] has led to numerous associations with clinical outcome and helped to shape future treatment strategies. There is growing consensus that subtypes exist in CD as well, each with its own presentation, genetic makeup and prognosis. As a first step, molecular stratifications of archived patient tissue and serum from major clinical trials could be performed in the context of response to biological and microbial therapies for CD.[30]

Our current studies did not allow for the investigation of both intestinal regions from the same individual, but they provide significant motivation for future exploration of these molecular subclasses longitudinally in larger cohorts of matched colon and ileum tissue of the same patient. New larger studies should allow for a more complete understanding of the molecular effects of host-environment interactions on disease and its utility in guiding clinical decisions. In addition, our sample numbers here are too limited to study the effects of genetic variation on identified changes in chromatin accessibility and gene expression. Our chromatin accessibility data strongly suggested that changes occurred at genomic loci previously associated with disease through genetic variation. Future studies with larger sample sizes will help identify specific relationships between genetic disease predisposition, regulatory activity, gene expression levels and clinical phenotypes and to better characterise individual-level disease subphenotypes within this very heterogeneous disease and tissue. Genomic studies on composite cell types (e.g., immune and epithelial cells) may also become necessary to study cell-specific mechanisms driving phenotype-specific disease pathogenesis.

MATERIALS AND METHODS

For more detailed information, see the online supplementary methods. All processed sequencing data are available in Gene Expression Omnibus (GEO) under accession GSE85499 with

Table 1 Clinical phenotypes associated with Crohn's disease (CD) subclasses

	Paediatric (ileum)			Adult (colon)		
Phenotype	**Ileum-like (n=50)**	**Colon-like (n=50)**	**p Value**	**Ileum-like (n=10)**	**Colon-like (n=11)**	**p Value**
Location						
Ileum-only	10	10	1	2	0	0.21
Colon-only	21	11	0.0528	1	6	0.06
Ileum+colon	19	29	**0.0014**	7	5	0.39
Patient characteristics						
Mean age	12.1	12.4	0.6193	35.1	35.3	0.98
Male	28	29	1	3	3	1
Female	22	21	1	7	8	1
Smoker	NA	NA	NA	5	4	0.67
Inflammation						
Macroscopic	29	41	**0.0156**	NA	NA	NA
Microscopic	8	6	0.7742	NA	NA	NA
None	13	3	**0.0122**	NA	NA	NA
Phenotypes and involvement						
Deep ulcers	10	29	**0.0002**	NA	NA	NA
Perianal	NA	NA	NA	0	4	0.09
Sigmoid	NA	NA	NA	2	7	0.08
Rectal	NA	NA	NA	0	9	**<0.01**
Ileal disease	29	39	0.0528	9	5	0.06
Inflammatory	NA	NA	NA	1	4	0.31
Stricturing	NA	NA	NA	7	4	0.2
Penetrating	NA	NA	NA	5	4	0.67
Preoperative treatment history						
Steroids	NA	NA	NA	8	10	0.59
5-ASA	NA	NA	NA	8	10	0.59
Immunomodulation	NA	NA	NA	8	8	1
Anti-TNF	NA	NA	NA	7	10	0.31
Non-anti-TNF biologic	NA	NA	NA	0	2	0.48
Postoperative outcome						
Postoperative disease recurrence	NA	NA	NA	=2/8	=0/10	0.18
Postoperative biologic use	NA	NA	NA	=6/9	=1/11	**0.02**
Colectomy	NA	NA	NA	0	6	**0.01**
Second resection	NA	NA	NA	3	6	0.39
Median time to first resection (years)	NA	NA	NA	3	5	0.91
Median time from first to second resection (years, if applicable)	NA	NA	NA	3.5	2	0.33

Paediatric and adult patient phenotypes, segregated by colon-like and ileum-like classifications, were compared using Fisher's exact test (for categorical data) or two-sided t-test (for continuous data). Significant associations (p<0.05) are bolded. See online supplementary tables S2 and S3 for data on individual patients.
5-ASA, 5-aminosalicylic acid; TNF, tumour necrosis factor; NA, not available.

raw sequence data available through dbGaP. Full data tables are also posted at http://fureylab.web.unc.edu/datasets/crohns-disease-molecular-subtypes/.

Statistics

Differential gene expression was detected using DEseq with an adjusted p value threshold of 0.05. Differential chromatin accessibility was detected using a two-sided t-test with p value threshold of 0.05. Pathway enrichments were determined using GSAA (FDR<0.1, permutation test) for all pairwise comparisons between all CD, colon-like CD, ileum-like CD and non-IBD samples based on differential gene expression p values calculated using two-sided t-tests. Clinical phenotype associations were tested with two-sided t-test or Fisher's exact test.

Study approval

All procedures were approved under the University of North Carolina Institutional Review Board (IRB) protocols 10-0355, 14-2445 and 11-0359.

Author affiliations
[1]Department of Genetics, University of North Carolina at Chapel Hill, Chapel Hill, North Carolina, USA
[2]Curriculum in Bioinformatics and Computational Biology, University of North Carolina at Chapel Hill, Chapel Hill, North Carolina, USA
[3]Center for Gastrointestinal Biology and Disease, University of North Carolina at Chapel Hill, Chapel Hill, North Carolina, USA
[4]Curriculum in Genetics and Molecular Biology, University of North Carolina at Chapel Hill, Chapel Hill, North Carolina, USA
[5]F. Widjaja Foundation Inflammatory Bowel and Immunobiology Research Institute, Cedars-Sinai Medical Center, Los Angeles, California, USA
[6]Department of Surgery, University of North Carolina at Chapel Hill, Chapel Hill, North Carolina, USA
[7]Department of Biology, University of North Carolina at Chapel Hill, Chapel Hill, North Carolina, USA

Acknowledgements We thank B. Aronow and L. Denson for assistance with paediatric patient data (CCFA RISK cohort). All sequencing experiments were performed at the UNC High-Throughput Sequencing Facility (HTSF).

Contributors MW, JMS, BK, AT, AR, GRG, MSS and DPBM acquired data. MW, JMS and JWI analysed and interpreted data. MW and JMS prepared figures, drafted and revised the manuscript. RBS, HHH, RR, TSS and MJK provided help with tissue acquisition and patient phenotyping. SZS and TSF designed and supervised the

study, acquired, analysed and interpreted the data, drafted and revised the manuscript and obtained funding. SZS conceptualised the study and acted as study sponsor. All authors uphold the integrity of the work, approved the manuscript in its entirety and are accountable for all aspects of the work.

Funding National Institute of Environmental Health Sciences (R01-ES024983), National Institute of Diabetes and Digestive and Kidney Diseases (P01-DK046763, P30-DK034987, R01-DK094779, T32-DK007634 and U01-DK062413), American Gastroenterological Association Research Scholar Award (SZS), Broad Medical Research Program, Crohn's and Colitis Foundation of America's Career Development Award (SZS) and Microbiome Consortium, UNC Team Translational Science Award, and Helmsley Trust SHARE 2, Project 3.

Competing interests None declared.

Ethics approval University of North Carolina IRB.

Provenance and peer review Not commissioned; externally peer reviewed.

REFERENCES

1 Cleynen I, Boucher G, Jostins L, *et al*. Inherited determinants of Crohn's disease and ulcerative colitis phenotypes: a genetic association study. *Lancet* 2016;387:156–67.
2 Jostins L, Ripke S, Weersma RK, *et al*. Host-microbe interactions have shaped the genetic architecture of inflammatory bowel disease. *Nature* 2012;491:119–24.
3 Haberman Y, Tickle TL, Dexheimer PJ, *et al*. Pediatric Crohn disease patients exhibit specific ileal transcriptome and microbiome signature. *J Clin Invest* 2014;124:3617–33.
4 Perou CM, Sørlie T, Eisen MB, *et al*. Molecular portraits of human breast tumours. *Nature* 2000;406:747–52.
5 Cancer Genome Atlas Network. Comprehensive molecular portraits of human breast tumours. *Nature* 2012;490:61–70.
6 Brannon AR, Reddy A, Seiler M, *et al*. Molecular stratification of clear cell renal cell carcinoma by consensus clustering reveals distinct subtypes and survival patterns. *Genes Cancer* 2010;1:152–63.
7 Thurman RE, Rynes E, Humbert R, *et al*. The accessible chromatin landscape of the human genome. *Nature* 2012;489:75–82.
8 Dixon JR, Jung I, Selvaraj S, *et al*. Chromatin architecture reorganization during stem cell differentiation. *Nature* 2015;518:331–6.
9 McKay DJ, Lieb JD. A common set of DNA regulatory elements shapes Drosophila appendages. *Dev Cell* 2013;27:306–18.
10 Simon JM, Hacker KE, Singh D, *et al*. Variation in chromatin accessibility in human kidney cancer links H3K36 methyltransferase loss with widespread RNA processing defects. *Genome Res* 2014;24:241–50.
11 Simon JM, Davis JP, Lee SE, *et al*. Alterations to chromatin in intestinal macrophages link IL-10 deficiency to inappropriate inflammatory responses. *Eur J Immunol* 2016;46:1912–25.
12 Simon JM, Giresi PG, Davis IJ, *et al*. Using formaldehyde-assisted isolation of regulatory elements (FAIRE) to isolate active regulatory DNA. *Nat Protoc* 2012;7:256–67.
13 Uhlén M, Fagerberg L, Hallström BM, *et al*. Proteomics. Tissue-based map of the human proteome. *Science* 2015;347:1260419.
14 Comelli EM, Lariani S, Zwahlen MC, *et al*. Biomarkers of human gastrointestinal tract regions. *Mamm Genome* 2009;20:516–27.
15 Welter D, MacArthur J, Morales J, *et al*. The NHGRI GWAS Catalog, a curated resource of SNP-trait associations. *Nucleic Acids Res* 2014;42:D1001–6.
16 Ernst J, Kellis M. Discovery and characterization of chromatin states for systematic annotation of the human genome. *Nat Biotechnol* 2010;28:817–25.
17 Roadmap Epigenomics C, Kundaje A, Meuleman W, *et al*. Integrative analysis of 111 reference human epigenomes. *Nature* 2015;518:317–30.
18 Xiong Q, Ancona N, Hauser ER, *et al*. Integrating genetic and gene expression evidence into genome-wide association analysis of gene sets. *Genome Res* 2012;22:386–97.
19 Xiong Q, Mukherjee S, Furey TS. GSAASeqSP: a toolset for gene set association analysis of RNA-Seq data. *Sci Rep* 2014;4:6347.
20 Garg AD, Kaczmarek A, Krysko O, *et al*. ER stress-induced inflammation: does it aid or impede disease progression? *Trends Mol Med* 2012;18:589–98.
21 Smith PD, Ochsenbauer-Jambor C, Smythies LE. Intestinal macrophages: unique effector cells of the innate immune system. *Immunol Rev* 2005;206:149–59.
22 Singh S, Ding NS, Mathis KL, *et al*. Systematic review with meta-analysis: faecal diversion for management of perianal Crohn's disease. *Aliment Pharmacol Ther* 2015;42:783–92.
23 Bailey EH, Glasgow SC. Challenges in the medical and surgical management of chronic inflammatory bowel disease. *Surg Clin North Am* 2015;95:1233–44, vii.
24 Harb WJ. Crohn's disease of the colon, rectum, and anus. *Surg Clin North Am* 2015;95:1195–210, vi.
25 Agouridis AP, Elisaf M, Milionis HJ. An overview of lipid abnormalities in patients with inflammatory bowel disease. *Ann Gastroenterol* 2011;24:181–7.
26 Shores DR, Binion DG, Freeman BA, *et al*. New insights into the role of fatty acids in the pathogenesis and resolution of inflammatory bowel disease. *Inflamm Bowel Dis* 2011;17:2192–204.
27 den Besten G, van Eunen K, Groen AK, *et al*. The role of short-chain fatty acids in the interplay between diet, gut microbiota, and host energy metabolism. *J Lipid Res* 2013;54:2325–40.
28 Rezaie A, Parker RD, Abdollahi M. Oxidative stress and pathogenesis of inflammatory bowel disease: an epiphenomenon or the cause? *Dig Dis Sci* 2007;52:2015–21.
29 Gorrini C, Harris IS, Mak TW. Modulation of oxidative stress as an anticancer strategy. *Nat Rev Drug Discov* 2013;12:931–47.
30 Sands BE, Anderson FH, Bernstein CN, *et al*. Infliximab maintenance therapy for fistulizing Crohn's disease. *N Engl J Med* 2004;350:876–85.