Antigen-driven colonic inflammation is associated with development of dysplasia in primary sclerosing cholangitis

Bana Jabri (bjabri@bsd.uchicago.edu)
Chicago

Dustin Shaw
University of Chicago

Raúl Aguirre-Gamboa
University of Chicago
https://orcid.org/0000-0003-2505-3574

Marcos Viera
University of Chicago

Saideep Gona
University of Chicago

Nicholas DiNardi
University of Chicago

Anni Wang
University of Chicago

Anne Dumaine
University of Chicago

Jody Gederloos
University of Groningen

Zachary Earley
University of Chicago

Katherine Meckel
University of Chicago

Cezary Ciszewski
University of Chicago

Anabella Castillo
Icahn School of Medicine

Kelly Monroe
Washington University in St Louis

Joana Torres
https://orcid.org/0000-0003-2895-5821

Shailja Shah
Icahn School of Medicine

Jean-Frederic Colombel
Icahn Medical School of Medicine at Mount Sinai  https://orcid.org/0000-0001-6472-249X

Steven Itzkowitz
Icahn School of Medicine

Rodney Newberry
Washington University School of Medicine

Russell Cohen
University of Chicago

David Rubin
University of Chicago

Sarah Cobey
University of Chicago  https://orcid.org/0000-0001-5298-8979

Iris H. Jonkers
UMCG  https://orcid.org/0000-0003-2304-7939

Christopher Weber
University of Chicago

Joel Pekow
University of Chicago

Patrick Wilson
Weill Cornell Medicine  https://orcid.org/0000-0002-3537-1245

Luis Barreiro
University of Chicago

Article

Keywords:

Posted Date: June 1st, 2022

DOI: https://doi.org/10.21203/rs.3.rs-1697242/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License
Antigen-driven colonic inflammation is associated with development of dysplasia in primary sclerosing cholangitis
Abstract

Primary sclerosing cholangitis (PSC) is an autoimmune-like disease of the bile ducts that co-occurs with inflammatory bowel disease (IBD) in almost 90% of cases. Colorectal cancer is a major complication of patients with both PSC and IBD, and these patients are at a much greater risk compared to patients with IBD without concomitant PSC. Combining flow cytometry, bulk and single cell transcriptomics, and T- and B-cell receptor repertoire analysis of right colon tissue from PSC, IBD, and healthy controls we identified a unique adaptive inflammatory transcriptional signature associated with greater risk and shorter time to dysplasia specifically in patients with PSC. This inflammatory signature is characterized by antigen-driven IL-17A+ Foxp3+ CD4 T-cell and IgG-secreting B-cell responses. These data suggest that the mechanisms of PSC and IBD dysplasia are distinct and provide molecular insights to guide prevention of colorectal cancer in PSC.

Introduction

Primary sclerosing cholangitis (PSC) is a chronic, cholestatic, autoimmune-like liver disease with a strong HLA association\(^1\) that is characterized by liver fibrosis\(^2\) and concomitant inflammatory bowel disease (IBD)\(^3\). Colorectal neoplasia (CRN) is a major complication of patients with PSC and IBD\(^4\), with a 50% 25-year cumulative risk for CRN, which is five times greater than what is observed in IBD without PSC\(^5\). With such a high burden of CRN, there is a major clinical need to identify the risk factors that predispose patients with PSC to develop CRN. Furthermore, identifying the causes of CRN can aid in its prevention in this highly vulnerable population.
Although duration and severity of inflammation in IBD are known to correlate with CRN development\textsuperscript{6,7}, the precise mechanisms underlying these observations remain unknown. Inflammation is thought to impact cancer development at many stages, including initiation (introduction of mutations into proliferating cells) and promotion (preferential expansion of mutated cells via external proliferation signals)\textsuperscript{8}. In IBD without PSC, it is believed that reactive oxygen species (ROS) introduce DNA mutations in colonic epithelial cells which can then preferentially expand with additional proliferation signals\textsuperscript{9,10}. If this is the case, then IBD inflammation is relevant to the initiation of CRN but not necessarily promotion. Whether this is also the mechanism for CRN in PSC has not been investigated. However, the significantly higher risk of CRN in PSC, the limited genetic overlap between IBD and PSC\textsuperscript{11}, and the distinct clinical presentation of colitis between the two\textsuperscript{12,13} collectively suggest that the pathogenesis of CRN in PSC and IBD are distinct. We further hypothesize that the role of colonic inflammation in PSC, specifically with regards to promotion of CRN, is different than in IBD.

To determine whether the mechanisms promoting CRN differ in PSC versus IBD, and whether a PSC-specific signature of risk for CRN can be identified, we transcriptionally and cellulary profiled 71 patients with PSC (93\% of which also had a diagnosis of IBD, all of which will be referred to as ‘PSC’), 110 patients with IBD and no PSC (IBD), and 56 healthy control (HC) patients, including patients with and without active dysplasia (an early stage of CRN). Our analysis included a broad, unbiased tissue transcriptional profiling combined with flow cytometry analysis (Fig. 1A). In addition, given the strong HLA association with PSC but not IBD, we performed single-
cell transcriptomics of T-cells and B-cells with T- and B-cell receptor analysis to
evaluate the hypothesis that T- and B-cell antigen-driven responses contribute to the
development of CRN in patients with PSC. We focused on the right colon because
inflammation and dysplasia are most common in that location for patients with PSC\textsuperscript{14,15}.

Our study found that the nature of inflammation and the mechanisms promoting
dysplasia are distinct between PSC and IBD, and that PSC inflammation may be
antigen-driven.

\textbf{Results}

\textbf{PSC and IBD show markedly different inflammatory signatures}

To characterize the ways in which the intestinal environments of PSC and IBD
differ, we performed RNA sequencing (RNAseq) on colon tissue from a subset of
patients in our cohort that had no history of dysplasia in any segment of the colon
(clinical and demographic data from this cohort is summarized in Extended Data Table
1). This included samples from 65 patients with PSC, 103 patients with IBD, and 48 HC
patients with no history of dysplasia. The biopsies were all taken from the same location
(10cm distal to the ileocecal valve; right colon) to avoid any bias related to regional
immune and microbial differences across the colon\textsuperscript{16}. We analyzed samples from the
right colon because nearly all patients with PSC have a history of inflammation in the
right colon\textsuperscript{14}, and dysplasia is most common in the right colon of PSC\textsuperscript{15}. Although colitis
in IBD is not always right-sided\textsuperscript{17}, we controlled for the variability in distribution of
inflammation by enrolling patients with IBD only if they had a documented history of
right-sided inflammation.
An unsupervised clustering analysis using the 3,000 most hypervariable genes across diagnoses identified four distinct clusters of patients (Fig. 1B). Two of these clusters, “Uninflamed 1 and 2” (U1 and U2), were histologically and transcriptionally uninflamed, as defined by a pathologist or the Inflammatory Response Gene Set Enrichment Score, respectively, (Fig. 1C and D; and Extended Data Fig. 1A and B) and were therefore combined (collectively referred to as “U”) in subsequent analyses. Two smaller clusters of inflamed patients were identified and labeled “Inflamed 1 and 2” (I1 and I2), with I2 being more inflamed than I1 (Extended Data Fig. 1A and B). Genes significantly upregulated in I2 as compared to I1 (n=7,734, 51% of all genes tested at False Discovery Rate (FDR) <5%) were strongly enriched among gene ontology terms related to the regulation of both T- and B- cell activation (Fig. 1E), suggesting that I2 is associated to an adaptive inflammatory signature.

The distribution of diagnoses was markedly different across transcriptional clusters (Fig 1F). Nearly all HC patients fell in the U cluster, whereas there was a significant enrichment of patients with IBD, and to an even greater extent patients with PSC, in the I1 and I2 clusters. The I2 cluster had the greatest difference in proportion of PSC and IBD: 27% of patients with PSC fell in this cluster versus only 7% among patients with IBD (Chi-square p= 2.0x10^{-6}). Since there were both patients with IBD and patients with PSC in cluster I2 (albeit at different frequencies) we investigated whether there were any features unique to PSC I2 as compared to IBD I2. We observed immune pathways to be enriched in PSC I2 (Fig. 1G), including pathways related to T-cell activation and response to bacterial molecules. Therefore, although both PSC and IBD I2 demonstrated inflammation, the natures of these inflammations were transcriptionally
distinct. Of note, some of the genes belonging to the pathways enriched in PSC I2 were previously identified to be associated with PSC by genome-wide association studies (Fig. 1G, e.g., IDO1, SOCS1). These findings provide credence to the long-standing hypothesis that the nature of PSC and IBD inflammation are different—a hypothesis that was based on clinical observations demonstrating differences in patterns of inflammation and other phenotypic differences. Additionally, the features unique to PSC inflammation might also provide clues into potential mechanisms of dysplasia in PSC.

**PSC dysplasia is strongly associated with a unique inflammatory transcriptional signature**

Next, we investigated whether the I2 PSC signature was in any way related to the development of dysplasia. To do so, we performed RNA-seq data on colon biopsies collected from non-dysplastic mucosa in patients with right-sided dysplasia detected at the time of sampling (clinical and demographic data from this cohort are summarized in Extended Data Table 2). This included 6 patients with PSC and dysplasia, 7 patients with IBD and dysplasia, and 8 control patients with dysplasia (sporadic dysplasia). Because we have not specifically sampled dysplastic tissue, the data collected reflects the tissue environment in which the dysplasia developed rather than the transcriptional signature of a dysplastic lesion.

Using the cluster signatures generated from patients with no history of dysplasia we built a classification model using a regularized logistic regression (elastic net, see methods section), which was then deployed to predict cluster assignment of patients
with right-sided dysplasia. Out of sample test set (n=53), validation showed that our prediction model has perfect accuracy in ascribing cluster I2 (AUC=1). Strikingly, 83% of patients with PSC and right-sided dysplasia were assigned to cluster I2. In marked contrast, only 0% of control patients with right-sided sporadic dysplasia and 14% of patients with IBD and right-sided dysplasia were classified into I2 (Fig. 2A). Importantly, amongst patients with PSC classified as I2 we found no significant differences in gene expression between patients with and without right-sided dysplasia (Extended Data Fig. 2) suggesting that the I2 PSC signature is not impacted by the presence of dysplasia and may instead reflect an immunological and transcriptional state promoting the development dysplasia.

Consistent with the strong overlap between PSC dysplasia and the I2 transcriptional signature, we observed significantly higher inflammation levels, both histologically (Fig. 2B) and transcriptionally (Fig. 2C), in the tissue environment where PSC dysplasia developed versus those environments of IBD dysplasia or sporadic dysplasia. Strikingly, no genes were differentially expressed between IBD dysplasia and sporadic dysplasia-associated tissue (Fig. 2D), suggesting that the tissue transcriptional environments of patients with IBD and dysplasia versus patients with sporadic dysplasia are largely indistinguishable at the time which dysplasia is discovered. In contrast, 15% and 36% of all genes tested (n=15,146) were differentially expressed when contrasting the PSC dysplasia against that of IBD dysplasia and sporadic dysplasia, respectively (Fig. 2D). Taken together, these results suggest that when comparing CRN pathogenesis in IBD and PSC, inflammation is relevant at different stages. In IBD, because ROS are thought to induce mutations, but the tissue environment at the time of
dysplasia is uninflamed and indistinguishable from sporadic dysplasia, we believe that inflammation contributes to the initiation, but not the promotion of CRN in IBD. In contrast, PSC dysplasia is nearly always found in the context of active inflammation, suggesting that inflammation does play a role in the promotion of CRN in PSC. Whether or not PSC inflammation contributes to initiation of CRN remains to be determined.

The colonic immune landscape of PSC shows signs of antigen drive

Given the enrichment of CD4 T-cell activation in PSC I2 (Fig. 1G) we isolated right colon lamina propria CD4 T-cells and measured the expression of canonical markers associated with activation (interleukin-17 (IL-17), interferon gamma (IFNγ), tumor necrosis factor alpha (TNFα)) and regulation (forkhead box P3 (Foxp3)) by flow cytometry. Although we did not see quantitative increases in lamina propria CD4 T-cells that expressed single markers (Extended Data Fig. 3A-D), we found a significant increase in IL-17A⁺ Foxp3⁺ double positive (DP) CD4 T-cells in patients with PSC classified as I2, relative to patients with PSC classified as U (Fig. 3A) or patients with IBD classified as I2 (p=0.024). These results were particularly interesting given the previous implication of IL-17A⁺ Foxp3⁺ CD4 T-cells in the development of CRN. The IL-17A⁺ Foxp3⁺ DP T-cells in PSC I2 had lower CD4 surface expression than their IL-17A⁺ and Foxp3⁺ single positive (SP) counterparts (Fig. 3B), suggesting that IL-17A⁺ Foxp3⁺ DP cells were more activated or chronically stimulated. There was no increase in IL-17A⁺ Foxp3⁺, Foxp3⁺ IL-17⁺, IFNγ⁺ Foxp3⁺, or TNFα⁺ Foxp3⁺ CD4 T-cells, nor an increase in IFNγ⁺ FoxP3⁺ or TNFα⁺ FoxP3⁺ double positive (DP) cells in PSC I2 as compared to IBD I2 (Extended Data Fig. 3E-J).
We hypothesize that IL-17A+ Foxp3+ CD4 T-cells are likely to play a key role in promoting the unique dysplastic program seen in patients with PSC. To address this hypothesis, we assessed the transcriptional program of IL-17A+ Foxp3+ DP CD4 T-cells by single cell RNAseq (scRNAseq) on freshly isolated right colon lamina propria CD4 T-cells from patients with PSC (n=5 I2, 6 I1, and 4 U). By calibrating the threshold of transcriptional detection of cells co-expressing IL17A and FOXP3 transcript using our flow cytometry data (Extended Data Fig. 4A-D), we were able to identify both DP and SP cells by scRNAseq (Fig. 3C). We performed differential expression analysis between the DP and each of the SP populations. This analysis demonstrated that IL17A+ FOXP3+ DP CD4 T-cells were transcriptionally distinct from either IL17A+ or FOXP3+ CD4 SP cells (Fig. 3D). Of note, GZMM\textsuperscript{21} and IL-32\textsuperscript{22}, genes previously implicated in development of dysplasia, were both significantly increased in IL17A+ FOXP3+ DP CD4 T-cells, as compared to either FOXP3+ or IL17A+ SP cells (adjusted p-value < 0.1, Extended Data Fig. 5). More specifically, genes upregulated in IL17A+ FOXP3+ DP CD4 T-cells were enriched for biological functions related to immune activation (Fig. 3E), and significantly enriched for a pathogenic Th17 signature\textsuperscript{23} when compared to IL17A+ SP cells (GSEA p<0.01) (Fig. 3F). Collectively, these results suggest a pathogenic role for IL17A+ FOXP3+ DP CD4 T-cells in the promotion of CRN, perhaps via secretion of IL-17A in conjunction with other cytokines.

Finally, to assess whether we could identify signs of antigen-drive in the IL-17A+ Foxp3+ DP CD4 T-cells, we determined whether there was a TCR motif enriched in the non-germline encoded complementarity-determining region 3 (CDR3) of the IL17A+ FOXP3+ DP T-cell subset. While we did not find any preferential V, D, or J gene usage
in either the TCRβ or TCRα chains (Extended Data Fig. 6A-E), we identified an enrichment for the “Leucine-Alanine (LA)” amino acid (AA) motif (Fig. 3G). “LA” is a germline encoded motif within one of the open reading frames (ORF) of TRBD2. Although this motif is encoded within the germline TRBD2 sequence, it exists in only one of the possible ORFs of TRBD2. Thus, the use of this motif and the specific ORF usage suggest antigen-driven selection of the TCR in the FOXP3⁺ DP T-cell subset. Additionally, a comparison of SP and DP cells using TRBD2 demonstrated a specific enrichment of the “LA” AA motif in DP T-cells (Extended Data Fig. 6F), suggesting a preferential selection for this ORF amongst DP cells. Finally, we analyzed the V, D, and J usage amongst cells containing the “LA” motif (Extended Data Fig. 7A-D) and found that the Vα gene usage of DP cells containing the “LA” motif were distinct from DP cells without the LA motif (Extended Data Fig. 7C), further suggesting that these DP “LA”-containing cells display a distinct TCR.

Strong genetic HLA Class II association in complex immune disorders implies a pathogenic role for antigen-specific T- and B-cell responses\(^{18}\). In contrast to IBD, and similar to celiac disease, PSC is associated with HLA Class II\(^{1}\). PSC is specifically associated with the ancient AH8.1 ancestral (HLA-A*01:01-C*07:01-B*08:01-DRB3*01:01-DRB1*03:01-DQA1*05:01-DQB1*02:01 haplotype) and the HLA-DRB1*13:01-DQA1*01:03-DQB1*06:03 haplotype\(^ {24}\). The AH8.1 haplotype was observed in all patients with PSC who showed LA-containing DP cell expansions (Extended Data Table 3)- consistent with the hypothesis that an antigen of unknown origin is presented on a specific HLA-II molecule associated with these haplotypes, driving the expansion of “LA”-containing IL17A⁺ FOXP3⁺ DP CD4 T-cells.
As we found a unique, pathogenic-like T-cell population enriched in PSC I2, we wanted to probe for a potential B-cell response as well. Initial tissue RNAseq showed that immunoglobulin (Ig) transcripts were among the most strongly upregulated genes in I2 (Extended Data Fig. 8). Given that plasma cells are the predominant B-cell subset of the intestinal lamina propria\textsuperscript{25} and that they express Ig transcript to a much greater degree than any other B-cell subset, we focused our analysis on plasma cells. We found that PSC I2 plasma cells were nearly 100% surface CD19\textsuperscript{+} (Fig. 4A) and larger than plasma cells in PSC U (Extended Data Fig. 9A), suggesting that the plasma cells observed in inflamed patients are recently arrived, active antibody-secreting cells\textsuperscript{26}. We observed an ordinal increase across clusters of the proportion of plasma cells secreting IgG in both IBD and PSC (Fig. 4B). However, the proportion of plasma cells secreting IgG in PSC I2 was significantly greater than in IBD I2 (p = 0.016). A corresponding decrease in proportion of IgA- and IgM-secreting plasma cells is observed ordinally across clusters (Extended Data Fig. 9B and C) as well. PSC colitis is therefore uniquely characterized by an increased proportion of IgG-secreting plasma cells, which is not seen to the same degree in IBD colitis, even IBD I2.

We performed scRNAseq of total plasma cells derived from patients with PSC across clusters (n=4 I2, 3 I1, and 6 U) and determined clonal pools of cells based on similarities across heavy chain sequences. We focused our analysis on the largest clone in each subject, under the assumption that the largest clone is the one that is most likely chronically activated. The three largest clones in our dataset were found in patients with inflammation (Extended Data Fig. 9D), and were predominantly IgG in I2, and IgA in U and I1 (Extended Data Fig. 9E). We observed that there was a greater
mean amino acid divergence from inferred germline within complementarity-determining region 3 (CDR3) of the largest clones of I2 as compared to U (Fig. 4C). Additionally, the CDR3 of the I2 top clones were more diverse as computed by the mean pairwise amino acid divergence within the clone (Fig. 4D). These differences were not present when analyzing the entire length of the heavy chain (Extended Data Fig. 9F and G), meaning that the CDR3 specifically was more heavily mutated and diverse in the top I2 than the top U clones. A dendrogram of the sequences within the largest clone found in an I2 patient demonstrates lop-sided branching patterns characteristic of selection (Fig. 4E). Collectively, these data suggest that the clonal IgG plasma cells in I2 PSC are strongly antigen-driven. The signs of antigen drive in the plasma cells corroborates the preferential enrichment of a TCR motif amongst the pathogenic DP cells, further suggesting that PSC inflammation and dysplasia is antigen-driven.

**PSC I2 inflammation increases the risk of developing dysplasia**

If the I2 inflammatory signature triggers the development of dysplasia in PSC, we would expect that patients with PSC classified as I2 will have an increased risk of developing dysplasia over time as compared to patients with PSC that aren’t I2. To test that hypothesis, we classified patients with IBD and patients with PSC as I2 or non-I2 (I1 or U), according to their transcriptional cluster. For patients that were sampled at multiple time points, we classified them as I2 if at any point they had an I2 signature, otherwise, they were classified as non-I2. Therefore, we classified patients based on whether they have ever experienced I2 inflammation. Rather than looking at the impact of duration and severity of inflammation on the development of CRN (which have
already been associated with CRN in IBD), we investigated whether experiencing a
unique form of inflammation (I2) was associated with a shorter time to dysplasia. To do
so, we retrospectively calculated the time (in years) from the diagnosis of intestinal
colitis to either the first incidence of right-sided dysplasia or to the last recorded
colonoscopy. 64 patients with PSC were included in this analysis, of which 10 (16%)
developed right-sided dysplasia in our timeframe of observation (median 15.5 years).
127 patients with IBD were included, of which 17 (13%) developed right-sided dysplasia
(median 13.8 years). Of the patients that developed dysplasia, 6 patients with PSC
(60%) and 0 patients with IBD (0%) were classified as I2. By plotting the Kaplan-Meier
estimated probability of right-sided dysplasia stratified by I2 and non-I2, we found that
patients with PSC classified as I2 had a greater risk of developing dysplasia over time
than non-I2 patients with PSC (Fig. 5A, right, \( p = 0.05 \)). We did not find, however, any
difference in risk of dysplasia between I2 patients with IBD and non-I2 patients with IBD
(Fig. 5A, left). This suggests that the I2 signature is associated with a greater risk for
right-sided dysplasia in PSC but not IBD. We additionally tested whether right colon I2
status was associated with a greater risk for the development of dysplasia outside of the
right colon. Of the 64 patients with PSC and 127 patients with IBD in this analysis, 10
patients with PSC (16%) and 23 patients with IBD (18%) developed dysplasia outside of
the right colon, and 5 patients with PSC (50%) and 0 patients with IBD (0%) of those
patients were classified as I2, respectively. We found that I2 was not associated with an
increased risk of non-right-sided dysplasia in either PSC or IBD (Fig. 5B), suggesting
that I2 inflammation is associated with a greater risk of dysplasia specifically in the
region in which it is observed.
Discussion

The major aim of our study was to perform an unbiased analysis of the transcriptional profile and adaptive immune response at the site where dysplasia develops in patients with PSC. A second goal was to generate hypotheses on potential mechanisms underlying the high risk of CRN in PSC and determine whether we could identify a transcriptional signature that could predict risk of dysplasia in patients with PSC. A major strength of the design of our study is that we combined tissue RNA seq with flow cytometry, scRNAseq and BCR and TCR repertoire analysis. Furthermore, we controlled for regional variability in factors such as bacterial load and composition, immune subsets, and epithelial cell function across the large intestine by restricting our analysis to the right colon. Finally, we included key patient control groups such as patients with and without right-sided dysplasia, and patients with IBD that had a history of right-sided inflammation to match the predominant site of inflammation in PSC patients.

In its entirety, our study reveals that inflammation plays a role in the promotion of PSC dysplasia but not IBD dysplasia, and that the inflammatory transcriptional I2 signature may be a clinical predictor for the development of dysplasia in patients with PSC. Importantly, we observed the PSC dysplasia-associated signature in patients with PSC with no history of dysplasia, suggesting that this inflammation is not a response to dysplasia, but rather precedes it. Overall, the I2 signature may help identify patients with PSC that need to be more closely monitored for dysplasia and may require more aggressive therapies. A prospective study in which patients are classified as I2 or non-
I2 and followed for right-sided dysplasia outcomes is warranted and would validate our results. In the spirit of developing clinically useful tools, we propose the use of the I2 PSC classifier model consisting of 81 genes (Extended Data Fig. 10) as a surveillance tool to identify patients with PSC that are at higher risk of developing CRN.

Furthermore, our study suggests a role for adaptive immune responses in the development of dysplasia in patients with PSC. Indeed, the I2 signature is characterized by a clonally expanded Foxp3+ IL17A+ DP T cell and IgG-secreting B-cell immune responses. The involvement of Foxp3+ IL17A+ DP T cells in colitis-associated cancers was previously suggested\(^1\). Here, our study suggest that the DP are significantly increased in PSC compared to IBD, and that Foxp3+ IL17A+ DP T cells may play a distinct role in the progression of PSC dysplasia. The finding that the DP cells have an activated and pathogenic Th17 phenotype, suggest that they may be driving dysplasia because of the cytokines and factors that they produce (such as GZMM and IL32, which are both upregulated when compared to single positive T-cells (adjusted p-value < 0.1), Extended Data Fig. 5). In addition, the uniquely expanded and mutated IgG plasma cells might also contribute to the development of dysplasia, by promoting the expansion of pathogenic T cells. Indeed, in HLA-associated diseases such as celiac disease, B-cell and T-cell crosstalk has been implicated in the amplification of the pathogenic tissue destruction\(^30,31\).

These findings, in combination with the strong HLA-II association, suggest that specific antigens may be driving inflammatory adaptive immune responses that promote the proliferation of epithelial cell and CRN. If this hypothesis is correct, interventions that target this adaptive immune response, including targeting B-cells, or removing the
driving antigens could prevent or dramatically reduce the risk of CRN in PSC. Though
the antigens remain to be identified, some studies have pointed to bacteria as a
potential causative agent in PSC\textsuperscript{32} and in CRN\textsuperscript{33}. In support of this, small clinical studies
on PSC cohorts have shown improvements in liver function tests and inflammation in
response to antibiotic treatment\textsuperscript{34–36}. Our study has generated resources, specifically Ig
and TCRs from patients with PSC, that can be used to screen potential bacterial
antigens. If a particular target (or set of targets) is found, it would be possible to formally
test the relationship of a specific taxum or taxa with dysplasia.

Finally, although the relationship between the intestinal and liver pathologies in
PSC are unclear, our results provide a framework with which to investigate the immune-
mediated fibrosis of the bile ducts. It is possible that the same mechanisms that cause
intestinal inflammation are at play in the bile ducts, though such investigation remains
outstanding. Therefore, further investigation of the potential mechanisms of CRN
provided in this study could not only lead to interventions that lead to decreased rates of
CRN in PSC, but also decreased rates of liver pathologies.
Materials and Methods

Patient enrollment and ethics

All activities related to enrollment of patients, collection of samples, and sample analysis were approved by the University of Chicago Institutional Review Board (IRB) and performed under IRB protocols 15573A and 13-1080.

Adults scheduled for a standard of care colonoscopy at the University of Chicago Medicine (UCM) were screened for diagnosis and eligibility criteria for enrollment in the study on a weekly basis. Any patient that passed exclusion criteria was eligible for enrollment. Exclusion criteria included: patients with chronic infectious diseases such as human immunodeficiency virus (HIV) or hepatitis C (HCV); active, untreated *Clostridia difficile* infection; active infection with severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2); intravenous or illicit drug use such as cocaine, heroin, non-prescription methamphetamines; active use of blood thinners; severe comorbid diseases; patients on active cancer treatment; and patients who are pregnant. Approaching prospective patients was at the discretion of their treating physician and was not done in cases that would put patients at any increased risk, regardless of reason. Patients were approached the day of their procedure and informed, written consent was obtained before any samples were acquired.

Classification of patients into diagnosis groups

Patients enrolled in the study were categorized as either PSC, IBD, or healthy (no diagnosis of PSC or IBD) control patients. IBD and PSC patients were further sub-
classified by IBD type: Crohn’s disease, ulcerative colitis (UC), or indeterminant colitis.

A total of 77 PSC (23 PSC-Crohn’s, 48 PSC-UC, 1 PSC-indeterminant colitis, and 5 PSC patients without a diagnosis of IBD), 129 IBD (80 Crohn’s and 49 UC), and 63 healthy control patients were enrolled over the course of the study. Categorization of each patient into diagnosis groups was done after careful review of the patients’ medical health records to ensure the validity of each diagnosis. Should a patient’s diagnosis change over the course of the study (e.g. the subtype of IBD was re-diagnosed as UC, when previously Crohn’s), the most recent diagnosis was used for all time points at which the patient was assessed. Patients were classified as PSC if records of a diagnosis of PSC could be found in the patients’ chart along with supporting liver imaging and liver function tests consistent with the PSC diagnosis. A liver biopsy was not necessary to confirm a PSC diagnosis as consistent with current practices.

As nearly all patients with PSC have right-sided colitis, we only enrolled IBD patients who had an explicit history of right-sided colitis documented in their medical records. Any IBD patient without documented right-sided colitis were excluded from the study. Any patients without a diagnosis of PSC or IBD that were receiving screening colonoscopies for preventative cancer screening or diagnostic abnormalities such as diarrhea, were considered healthy patients. Any healthy controls consented to the study that were determined to have signs of endoscopic or histologic inflammation were retrospectively excluded from the study.

A diagnosis of dysplasia was determined by evaluation of the histological reports for each colonoscopy. If the pathologist’s report stated evidence of adenoma, low-grade dysplasia, high-grade dysplasia, or adenocarcinoma the patient was classified as
having dysplasia. If the pathologist reported indefinite dysplasia or were unable to
determine whether an abnormal lesion represented actual dysplasia or reactive
changes due to inflammation, the patient was classified as indefinite for dysplasia. If no
signs of *bona fide* or indefinite dysplasia were identified, the patient was classified as
non-dysplastic. Sporadic dysplasia was defined as the presence of dysplasia (typically
an adenoma) in the healthy control patients. A total of 6 PSC, 7 IBD, and 8 sporadic
dysplasia patients were included in our analysis.

**Collection of patient clinical and demographic data**

We searched all patients’ available electronic medical records for relevant
clinical and demographic information. The demographic information collected included
date of birth, sex, race, and ethnicity. We also recorded (when applicable) date of initial
IBD and PSC diagnosis, date of first incidence of dysplasia, and date of liver transplant.
For each procedure, we recorded (when applicable) the date of the procedure;
endoscopically and histologically scored inflammation in the right colon; location, stage,
and nature of dysplasia; endoscopically and histologically scored inflammation at the
site of dysplasia; and all IBD- or PSC-related medications currently taken by the
patients, including immunosuppressants, biologics, antibiotics, steroids, and ursodiol.

Endoscopically-scored inflammation was based on the clinician’s
evaluation of inflammation during the colonoscopy using the Mayo Endoscopic
Subscore system. The following scale was used:

0 = no diagnostic abnormality or quiescent inflammation

1 = mild inflammation
2 = moderate inflammation

3 = severe inflammation.

Histologically-scored inflammation was based on the pathologist’s evaluation of the inflammation based on the histologic criteria for grading of disease activity at University of Chicago Medicine. The criteria are the following:

0 = no diagnostic abnormality

1 = quiescent = features of chronicity (crypt distortion/shortening/drop-out, basal plasmacytosis, pyloric or Paneth cell metaplasia) in the absence of mild/moderate/severe activity

2 = mild = neutrophils present in the epithelium

3 = moderate = neutrophils present in crypt lumen forming crypt abscess

4 = severe = erosion or ulceration of epithelium

Collection of tissue specimens

During the colonoscopy, the endoscopist collected 8-10 tissue biopsies using 2.8mm or 3.2mm forceps at 10cm distal to the ileocecal valve. One of these biopsies was placed immediately into RNAprotect (Qiagen) and the remaining biopsies were placed into RPMI 1640 (Fisher Scientific). Samples were immediately transported on ice to the laboratory for processing, according to the protocols below.

Tissue biopsy RNAseq

The tissue biopsy in RNAprotect was stored at 4°C for 48-72 hours, after which the RNAprotect was removed, and the biopsy stored at -80°C until tissue
processing. Whole tissue biopsies stored at -80°C were thawed on ice and transferred
to Starstedt tubes (Fisher Scientific) containing 350µL RLT Plus (Qiagen) supplemented
with 1% 2-mercaptoethanol (Fisher Scientific) and equal quantities of 1.0mm and
0.5mm zirconium oxide beads (Next Advance; one scoop each). Biopsies were bead
beat 3 times for 1 minute at a setting of nine on a Bullet Blender 24 (Next Advance),
with one minute of cooling on ice between each beating. Lysates were processed using
the AllPrep DNA/RNA/miRNA Universal Kit (Qiagen). 500ng of purified RNA was used
as input in the TruSeq Stranded mRNA Library Prep kit (Illumina) to generate sample
libraries according to manufacturer’s specifications. Libraries were multiplexed and
sequenced at a depth of 20 million reads per sample (50 base pairs, single read) on a
HiSeq4000 sequencer.

Lamina propria lymphocyte isolation
Colonic epithelial cells and lymphocytes were isolated via mechanical disruption
and enzymatic digestion. Briefly, colonic biopsies were twice shaken at 250 revolutions
per minute (rpm) for 30 minutes at 37°C in 7mL RPMI 1640 (Fisher Scientific)
supplemented with 1% dialyzed fetal bovine serum (Biowest), 2mM EDTA (Corning),
and 1.5 mM MgCl₂ (Thermo Fisher Scientific). Cells were filtered through a 40µM filter
(Fisher Scientific), centrifuged, and pooled for subsequent analysis. This fraction was
considered the epithelial fraction. Subsequently, the remaining tissue was digested in
two sequential shakes at 250rpm at 37°C for 30 minutes in 15mL RPMI 1640
supplemented with 20% fetal bovine serum and 1mg/mL collagenase type IV, from
Clostridium histolyticum (Sigma-Aldrich). After each digestion, the solution was filtered,
centrifuged, and then combined for downstream experimentation. This fraction was
considered the lamina propria fraction.

**Surface flow cytometry and fluorescence activated cell sorting (FACS)**

The following directly conjugated antibodies were used to identify cell surface
markers (clone and manufacturer in parenthesis): CD45 (HI30; BD Biosciences), Ep-
CAM (9C4; BioLegend), CD3 (UCHT1; BioLegend), TCR α/β (IP26; BioLegend), CD4
(SK3; BD Biosciences), CD8 (RPA-T8; BD Biosciences), CD19 (HIB19; BD
Biosciences), CD27 (O323; BioLegend), and CD38 (HIT2; BioLegend). Cells were
stained for 15 min on ice using LIVE/DEAD Fixable Aqua or LIVE/DEAD Fixable Near-IR
(1:50, Thermo Fisher Scientific) diluted in PBS (Fisher Scientific), washed with PBS
supplemented 2% FBS, and subsequently stained in an antibody cocktail for 25 min at
4°C. Cells were washed with PBS/2%FBS, resuspended into PBS/2%FBS, and
subsequently run on a BD FACS Aria Fusion Flow Cytometer to sort purify the
populations of interest. CD4 T-cells (CD45⁺ EpCAM⁻negative > LIVE/DEAD⁻negative > FSC vs
SSC > singlets > CD3⁺ CD19⁻negative > CD4⁺ CD8⁻negative) and plasma cells (CD45⁺
EpCAM⁻negative > LIVE/DEAD⁻negative > FSC vs SSC > singlets > CD3⁻negative > CD38⁺
CD27⁺) from the remaining lamina propria fraction were sorted into 600 µL of RPMI 1640
supplemented with 10% FBS and 1% penicillin/streptomycin (Thermo Fisher Scientific)
for downstream experimentation including 10x Genomics sequencing and ELISpot. All
flow cytometry data were analyzed using FlowJo software version 10.7.2 (Tree Star).

**Enzyme-linked immune absorbent spot assay (ELISpot)**
A fraction of the purified plasma cells from FACS were used in this assay.

Preceding the isolation of the plasma cells, three rows of a flat-bottom 96-well polystyrene plates (Thermo Fisher Scientific) were coated with polyclonal goat-anti human IgA, IgG, and IgM antibodies (KPL) at a concentration of 5µg/mL, diluted in PBS (100µL/well). Plates were incubated at 4°C for a minimum of 24 hours. The day of the plasma cell isolation, the coated plates were washed three times with PBS/0.05% Tween-20 (BioRad) and then three times with PBS. Coated wells were then blocked with RPMI 1640 supplemented with 10% FBS and 1% penicillin/streptomycin at 37°C for a minimum of two hours. After FACS sorting, an equal number of plasma cells were serially diluted down the three rows of the plate at a 1:2 dilution and left to incubate at 37°C overnight. After the incubation, the cells were removed from the plate, and the wells were washed three times with PBS/0.05% Tween-20 and then three times with PBS. Each row then was incubated with Biotin-conjugated polyclonal goat anti-human IgA, IgG, or IgM (Southern Biotech) at a concentration of 1µg/mL at room temperature in the dark, for two hours. Subsequently, wells were washed three times with PBS/0.05% Tween-20, three times with PBS, and incubated in Streptavidin-Alkaline Phosphotase (Southern Biotech) at a dilution of 1:500 for two hours at room temperature in the dark. The wells were then washed three times in each PBS/0.05% Tween-20 and PBS, and the substrate NBT/BCIP (Thermo Scientific) was applied until individual spots were visible (less than 5 minutes total), after which the reaction was halted using room temperature tap water. Plates were left to dry upside-down in the dark, after which images were captured using a CTL ImmunoSpot Analyzer (ImmunoSpot) and spots were quantified manually in ImageJ (FIJI).
Phorbol myristate acetate/ionomycin stimulation assay

A portion of the lamina propria fraction was suspended in RPMI 1640 medium supplemented with 10% FBS, 1% penicillin/streptomycin, 1pg/mL phorbol myristate acetate (Sigma-Aldrich), 1.5 ng/mL ionomycin calcium salt (Sigma-Aldrich), 0.15% GolgiPlug (BD Bioscience), and 0.3% GolgiStop (BD Bioscience) in a total volume of 500µL in a polystyrene flat-bottom 24-well plate (Thermo Fisher Scientific). Cells were incubated at 37°C for three hours after which they were washed twice with ice cold RPMI 1640 medium supplemented with 10% FBS and 1% penicillin/streptomycin. Cells were stained for viability subsequently surface markers as for FACS, after which cells were fixed and permeabilized in a 1:4 solution of Fixation/Permeabilization Concentrate and Fixation/Diluent (eBioscience) for one hour at 4°C in the dark. Cells were washed twice with a 1:10 dilution of Permeabilization Buffer Solution (eBioscience) in nuclease-free water (Fisher Scientific), and subsequently stained for intracellular markers for one hour at room temperature in the dark. The following directly conjugated antibodies were used to identify intracellular markers (clone and manufacturer in parenthesis): CD45 (HI30; BD Biosciences), TCR α/β (IP26; BioLegend), CD4 (SK3; BD Biosciences), CD8 (RPA-T8; BD Biosciences), CD27 (O323; BioLegend), IFNγ (4S.B3; eBioscience), TNFα (MAb11; BioLegend), IL-17A (BL168; BioLegend), and Foxp3 (PCH101; Invitrogen). Cells were subsequently passed on either a BD LSRFortessa Flow Cytometer or a Cytek Aurora Flow Cytometer. All flow cytometry data were analyzed using FlowJo software version 10.7.2 (Tree Star).
**Single-cell RNAseq**

Cells were centrifuged and resuspended to a final concentration in RPMI 1640 medium supplemented with 10% FBS and 1% penicillin/streptomycin, and the suspensions were loaded into a Chromium Controller (10x Genomics, Inc.) under conditions to generate an anticipated yield of 1,000-10,000, depending on yield of cells from the tissue. Single-cell 5' RNA-seq libraries, as well as V(D)J libraries, were generated for each sample according to the manufacturer's instructions (Chromium Single Cell 5' Library Construction Kit V1 Chemistry, Single Cell V(D)J Enrichment Kit for Human T-cells, and Single Cell V(D)J Enrichment Kit for Human B-cells, all from 10x Genomics Inc). 5' libraries were sequenced to a minimum depth of 50,000 reads per cell for 5' gene expression libraries, or 5,000 reads per cell for V(D)J libraries, on an Illumina NovaSEQ6000.

**Bulk RNAseq analysis**

All bulk RNAseq samples were processed using a standard workflow based on the GENPIPES framework. Specifically, the “stringtie” type “rnaseq” pipeline was used. Reads were first trimmed using Trimmomatic software. Trimmed reads were aligned to the GRCh38 human reference genome using the STAR aligner following a two-pass mapping protocol. Alignments were then sorted and filtered for duplicates using Picard(sort, markduplicates) (“Picard Toolkit.” 2019. Broad Institute, GitHub Repository. http://broadinstitute.github.io/picard/). Gene-level read counts for downstream processing were calculated from spliced alignments using HTseq count.
Dimensionality reduction and clustering in non-dysplastic samples

In order to evaluate whether a unique signature was detectable in the colon biopsies of IBD or PSC patients with no history of dysplasia we searched for disease-specific transcriptional profiles through an unsupervised clustering strategy. To do so, we selected only samples with no history of dysplasia (healthy n=48, IBD n =103, PSC n=65). The normalized (log2 CPMs) expression matrix was then corrected for batch effect, and we selected the top 3,000 most variable genes by modeling the mean-variance relationship using the FindVariableFeatures, from the Seurat package (v4.0.0.0). Next, we calculated the principal components (PCs) by sample, for which we selected the first 40 PCs, as they explain at least 70% of the complete variance. These 40 PCs were then used as a distance matrix to perform hierarchical clustering from which we selected 4 biologically relevant clusters: U1, U2, I1 and I2. All statistical analyses involving dimensionality reduction and clustering were performed using R (v4.0.3, (“The R Project for Statistical Computing”)).

Differential expression and gene set enrichment analysis

Counts derived from the alignment were filtered for lowly expressed transcripts (median > 5). Furthermore, we included only protein coding genes and TCR and IG receptors, resulting in a total of 15,146 genes. Upon this set of genes we detected differentially expressed genes either across diagnosis or cluster by fitting a linear model to the log2 count per million reads (CPM) using the limma package (v3.46.0). In every contrast we included as covariates sex, age, and batch of sequencing.
We performed gene set enrichment analysis (GSEA) using the gseaGO function from the clusterProfiler (v.3.0.4) package over the log2 fold changes in expression between contrasts, cluster I2 vs I1.

To detect gene ontologies enriched in defined sets of genes, such as I2 PSC genes (I2 PSC versus I2 IBD contrast, p.adjusted < 0.05, log2 FC > 0). We performed over enrichment analysis using the enrichGO function from the clusterProfiler (v.3.0.4).

**Prediction of cluster assignment in dysplastic samples**

To assign a cluster (U1, U2, I1 and I2) to dysplastic samples, we constructed a classifier using an elastic net (eNet) model, which is a regularized regression approach. To do so, we decrease the potential noise within cluster assignment errors, by calculating cluster silhouette for each sample, and select only samples with a positive silhouette score. Defining a set of core cluster samples, we used the core cluster samples to detect differential expressed genes between U2, I1, and I2 clusters, and used all DEG (p.adjust < 0.05) in at least one contrast as the initial set of features to construct the eNet model. Next, we partitioned the cohort of core samples into a training set of 70% of all samples and a test set with the rest. To select the penalization score for eNet we used a 10x cross validation within the training cohort. The resulting classification model to predict I2 cluster adscription consisted of 81 genes with non-zero coefficients. The I2 model predicted with 100% accuracy in the out of sample test cohort (AUC=1).

**Transcriptional analysis of CD4 T-cells**
Fastq files were processed into gene count matrices using Cellranger v3.1.0 and the GRCh38 transcriptome downloadable from the Cellranger website. Analysis centered on the Seurat framework. An initial filtration step involved the removal of plasma cells from some samples. In addition, cells with mitochondrial read percentage greater than 50% were removed from further analysis. Finally, we dropped samples PSC28D and PSC40D entirely due to their very low T cell counts. Datasets were integrated using the SCTransform protocol. Specifically, SC-Transform was run on each sample while regressing mitochondrial read percentage as a covariate. Integration was performed using 20,000 genes followed by dimensionality reductions runPCA (utilizing 20 principle components for all dependent analysis), and runUMAP. After dimensionality reduction, unsupervised clustering was performed using FindNeighbors and FindClusters (resolution of 1). To define T-cell subpopulations, we employed a calibration strategy using corresponding flow cytometry data as a reference (Extended Data Fig. 5).

To perform differential gene expression analysis between subpopulations, we used a pseudobulking strategy. First, genes were filtered to have LogCPM > 0.01. Next, scran factor normalization was performed using the computeSumFactors function from the “scran” R package. Cells with size factors between 0.125 and 8 were preserved. Pseudobulk means were then calculated from the log counts as the per gene mean within each pseudobulk grouping. Pseudobulk means were used as input into an Limma-voom differential testing pipeline similar to those employed in bulk. Variance stabilization was performed using the Limma-Voom function voomWithQualityWeights.
and model fitting was performed using the Limma-Voom functions lmfit and eBayes. Resulting differential expression statistics were extracted using the topTable function.

**Repertoire analysis of CD4 T-cells**

Binary base call output from sequencing were put through the Cellranger mkfastq pipeline to generate fastq files, which were subsequently put through Cellranger vdj to generate full-length TCR sequences ([https://support.10xgenomics.com/single-cell-vdj/software/pipelines/latest/using/vdj](https://support.10xgenomics.com/single-cell-vdj/software/pipelines/latest/using/vdj)). Full-length TCR sequences were processed using IMGT/HiV-QUEST to identify productive sequences, determine V, D, and J gene usage, and identify the CDR3. Non-productive sequences, and sequences with the same cellular barcode were filtered from the analysis. TCRs were matched to gene expression profiles by barcodes and all subsequent analyses were performed by cell type. CDR3 amino acid sequences were trimmed from both ends to the left-most and right-most amino acid with a mutation within its codon (silent or missense). Trimmed amino acids from IL17A⁺ FOXP3⁺ double positive CD4 T-cells were queried for potential motifs using Sensitive, Thorough, Rapid, Enriched Motif Elicitation (STREME) web-based software, using IL17A and FOXP3 single positive CDR3s as a control. The proportion of cells containing the motif were then calculated.

**Repertoire analysis of plasma cells**

Binary base call output from sequencing were run through the Cellranger mkfastq pipeline to generate fastq files, which were subsequently run through Cellranger vdj to generate full-length Ig sequences ([https://support.10xgenomics.com/single-cell-](https://support.10xgenomics.com/single-cell-))
Full-length Ig sequences were processed using IMGT/HiV-QUEST to identify productive sequences, determine V, D, and J gene usage, and identify the CDR3. Non-productive sequences, and sequences with the same cellular barcode were filtered from the analysis. Partis v0.15.0 with default settings was used to simultaneously identify sets of sequences descended from the same naïve B cell and determine the sequence and the germline immunoglobulin genes used by each clone’s naïve ancestor. IgPhyML v1.1.0. was used to build clones’ phylogenetic trees by jointly optimizing tree topology and the parameters of a codon substitution model that incorporates variation in the mutability of nucleotide motifs in immunoglobulin genes. We manually verified that all the heavy chains within the top clones used the same light chain. Those that did not were removed from the clone, and the clonal size was re-adjusted. Custom code (available at https://github.com/cobeylab/psc_repertoire) was used for subsequent computational analyses.

For the entire sequence and separately for CDR3, the average amino acid divergence was computed between each sequence and the inferred naïve ancestor (to estimate average divergence from the clone’s ancestor) and for all pairs of sequences in a clone (to estimate standing diversity within clones at the time they were sampled). These analyses were conducted for the top clone in each dataset, including multiple clones in case of ties.

Patient genotyping and HLA imputation

DNA of patients was genotyped using the Illumina Infinium global screening array v1.0, with accompanying manifest file A5. Per patient, 200 ng of DNA was used for
hybridization and visualization was performed using the Illumina iScan. Results were exported using Genomestudio. Genotype calling was performed using opticall (version 0.8.1), all samples reported a call rate > 98%, and genotypes with a call rate below 95% were removed. Furthermore, rare variants (minor allele frequency < 0.01) and that do not follow Hardy Weinberg equilibrium (p < 0.0001) were removed from further analysis. Genotypes where then prepared for imputation following the guidelines and toolbox from the Michigan imputation server to be matched to the genome assembly GRCh37/hg19. Genotypes from chromosome 6 where then used to impute the HLA region using the four-digits multi ethnic HLA reference panel version 1. We then used the imputed four-digits HLA annotations to infer the HLA haplotypes for each patient.

**Time to dysplasia Kaplan-Meier analysis**

The medical records of each IBD and PSC patient was probed to determine the date of diagnosis of colitis, last date of follow-up at University of Chicago Medicine, and the date of first incidence of right-sided or non-right-sided dysplasia (if applicable). Right-sided dysplasia was dysplasia occurring in the cecum, ascending colon, or hepatic flexure. Non-right-sided dysplasia was considered dysplasia occurring in the transverse colon, splenic flexure, descending colon, sigmoid colon, or rectum. Right-sided dysplasia and non-right-sided dysplasia were considered independent events. We calculated the time from colitis diagnosis to right-sided dysplasia for each individual patient with a history of right-sided dysplasia or to the most recent colonoscopy at University of Chicago Medicine for patients with no documented right-sided dysplasia. We stratified samples into two groups: I2 group and non-I2. We defined I2 as any
samples for which an I2 inflammatory profile was ever detected in any of their visits, including visits after or during the first diagnosis of right-sided dysplasia. We then evaluated the difference in times to develop right-sided dysplasia from their first colitis related diagnosis using the Kaplan–Meier estimator using the survminer package (v0.4.8, https://CRAN.R-project.org/package=survminer), for both PSC patients and IBD patients. The same process was then repeated with non-right-sided dysplasia as the outcome.


9. Shah, S. C. & Itzkowitz, S. H. Colorectal Cancer in Inflammatory Bowel Disease:


27. Horns, F., Vollmers, C., Dekker, C. L. & Quake, S. R. Signatures of selection in the human antibody repertoire: Selective sweeps, competing subclones, and


Fig. 1: A subset of PSC patients with no history of dysplasia show a unique and highly inflamed transcriptional profile. A, Graphical representation of the methodology of this study. To control for biological differences across the span of the colon, we collected 8-10 tissue biopsies only from the right colon. Our patient cohort included PSC patients, IBD patients with a history of right-sided colitis, and control patients with no history of IBD or PSC. Patients were retrospectively determined to have right-sided dysplasia. From one biopsy, we isolated RNA for whole tissue bulk RNA sequencing. The remainder of the biopsies were mechanically and enzymatically disrupted and to isolate lamina propria CD4+ T-cells and plasma cells for analysis via flow cytometry, single cell RNAseq, and T-cell receptor and B-cell receptor analysis. B-D, Uniform manifold and approximation and projection (UMAP) plot using right colon tissue samples from subjects with no history of dysplasia at the time of sample collection. Samples are annotated by transcriptionally-determined cluster (B), histologically-scored inflammation (C), or Inflammatory Response Gene Set Enrichment Score (D). E, Dot plot showing the top 30 most significantly enriched (GSEA test q value < 0.1x10^-9) between I2 vs I1 clusters, color represents enrichment score and direction of effect. F, Distribution of subjects across clusters amongst subject without dysplasia, statistical significance determined by Chi-squared test. Red corresponds to cluster “I2”, purple to cluster “I1”, and blue to cluster “U” (combined U1 and U2). G, Alluvial plot connecting top enriched gene ontologies (biological processes) with the significantly up-regulated genes in I2-PSC vs I2-IBD, connections are colored by fold-change.
Fig. 2. The colonic dysplasia landscape of PSC is enriched in I2 signature and differs from that of IBD. A, Distribution of subjects with right colon dysplasia across clusters. Statistical significance determined by Chi-squared test. B, Histologically-scored inflammation at the site of dysplasia within the right colon using the University of Chicago Medicine histological criteria for grading of disease activity. 0 = no diagnostic abnormality, 1 = quiescent/minimally active, 2 = mild, 3 = moderate, 4 = severe. Right colon includes cecum, ascending colon, and hepatic flexure. Significance determined by Wilcoxon test (**p < 0.01, ***p < 0.001) for not significant (p > 0.05). C, single sample gene set enrichment analysis (ssGSEA) inflammatory response score calculated from the transcriptome of the right colon tissue biopsy. Significance determined by Wilcoxon test (**p < 0.01). D, Bar graph quantifying the percentage of differentially expressed genes in each comparison (proportion of genes up-regulated in PSC in purple, genes up-regulated in sporadic dysplasia in green, and genes up-regulated in IBD in orange).
Fig. 3: PSC inflammation is characterized by IL-17A+ FoxP3+ CD4 T-cells enriched for TCRs containing “LA.”

A, Proportion of right colon lamina propria CD4 T-cells positive for IL-17A and FoxP3 by flow cytometry after 3 hour stimulation with phorbol myristate acetate and ionomycin (n=56 samples). Significance determined by Wilcoxon test (“*” for p<0.05, “**” for p<0.01, “ns” for not significant (p>0.05)). Each symbol represents an individual patient (circles denote patients without dysplasia at the time of sampling, “x” denotes patients with dysplasia at the time of sampling.

B, Mean fluorescence intensity (MFI) of surface CD4 of cells from I2 PSC patients. Significance determined by Wilcoxon matched-pairs signed rank test (n=6; “*” for p<0.05).

C, UMAP of single-cell sequenced CD4 T-cells from PSC subjects, annotated by transcriptionally-determined cell type (n=15 patients, 25,942 cells).

D, Log 2 fold change (FC) of genes comparing IL-17A+ FoxP3+ to FoxP3+ CD4 T-cells (x-axis) or IL-17A+ (y-axis) amongst I2 PSC subjects. Each gene represented as a point. Genes uniquely differentially expressed in IL-17A+ FoxP3+ versus FoxP3+ highlighted in blue, genes uniquely differentially expressed in IL-17A+ FoxP3+ and IL-17A+ highlighted in orange, and genes differentially expressed in both comparisons highlighted in pink. Genes not differentially expressed in either comparison shown as open gray circles. Highest FC genes labeled on graph.

E, Most significantly enriched gene sets using genes upregulated in IL-17A+ FoxP3+ versus either IL-17A+ or FoxP3+ CD4 T-cells. GSEA p-value is shown.

F, Enrichment for a pathogenic IL-17 signature using genes differentially expressed in IL-17A+ FoxP3+ versus IL-17A+ CD4 cells. GSEA p-value is shown.

G, Proportion of CD4 subset containing amino acid motif “LA” in the TCR beta chain by cell type amongst I2 PSC patients. Gray lines denote paired values from the same patients. SP = “single positive”, DP = “double positive” i.e. IL-17A+ FoxP3+. Significance determined by Wilcoxon test.
Fig. 4: PSC inflammation is characterized by an influx of IgG plasma cells and plasma cells show signs consistent with antigen drive. A, Proportion of right colon plasma cells positive for surface CD19 by flow cytometry (n=81 samples). B, Proportion of IgG-secreting plasma cells amongst total right colon plasma cells as determined by ELISPOT (n=67 samples). C, Mean amino acid divergence from inferred germline within CDR3 of largest clones identified in each patient (n=13 patients). D, Mean pairwise amino acid divergence within CDR3 of largest clones identified in each patient. E, Dendrogram of heavy chain sequences within top clone of I2 patient (n=110 sequences). This clone demonstrates a “lop-sided” branching pattern, consistent with non-random mutation accumulation and antigen drive. Origin point represents inferred germline sequence. Scale bar represents codon substitutions per codon. (A-D) Significance determined by Wilcoxon test (*** for p<0.05, **** for p<0.01, ***** for p<0.001, ****** for p<0.0001, “ns” for not significant (p>0.05)). Each symbol represents an individual patient (open circles denote patients without dysplasia at the time of sampling, “x” denotes patients with dysplasia at the time of sampling, open squares denote patients indefinite for dysplasia at the time of sampling).
**Fig. 5:** I2 status is associated with a greater risk and shorter time to dysplasia in PSC but not IBD. 

**A,** Kaplan-Meier-estimated curves for risk of right-sided dysplasia over time. Patients were classified as I2 (in red) or non-I2 (I1 or U, in black), according to their transcriptional cluster. Gray dashed line marks for 0.5 probability of event (dysplasia diagnosis after colitis). For patients that were sampled at multiple timepoints, they were classified as I2 if at any point they had an I2 signature, otherwise they were classified as non-I2. Time (years) was calculated from the diagnosis of intestinal colitis to either the first incidence of right-sided colitis or to the last colonoscopy recorded colonoscopy. Statistical outliers for time of follow-up were removed from the analysis before calculating the Kaplan-Meier estimates. Subjects are subset by diagnosis: IBD (left) or PSC (right).

**B,** Kaplan-Meier curves as in A, but for the estimated risk of non-right-sided dysplasia over time. Patients were again classified as I2 or non-I2 based on their right colon tissue transcriptional profile. Time (years) was calculated from the diagnosis of intestinal colitis to either the first incidence of non-right-sided colitis or to the last colonoscopy recorded colonoscopy. Statistical outliers for time of follow-up were removed from the analysis before calculating the Kaplan-Meier estimates. Subjects are subset by diagnosis: IBD (left) or PSC (right).
**Extended Data Fig. 1: Clusters I1 and I2 are inflamed, with I2 being the most inflamed.**

**A,** Histologically-scored inflammation in the right colon of patients without a history of dysplasia, separated by transcriptionally-determined cluster. 0 = no diagnostic abnormality, 1 = quiescent inflammation, 2 = mild inflammation, 3 = moderate inflammation, 4 = severe inflammation. Significance determined by Wilcoxon test ("****" for p<0.0001). **B,** Single sample gene set analysis (ssGSEA) enrichment score for the Inflammatory Response gene set (HALLMARK_INFLAMMATORY_RESPONSE, MS932, Molecular Signatures Database v7.5.1) calculated from the right colon tissue transcriptome of patients without a history of dysplasia, separated by transcriptionally-determined cluster. Significance determined by Wilcoxon test ("****" for p<0.001, "*****" for p<0.0001).
Extended Data Fig. 2: PSC I2 patients with dysplasia are indistinguishable transcriptionally from PSC I2 patients without dysplasia. Volcano plot summarizing the differentially expressed gene analysis of PSC I2 subjects with right-sided dysplasia versus PSC I2 subjects with no history of dysplasia. 0 genes passed the threshold of significance (red dashed line, adjusted p-value > 0.05), suggesting that the transcriptional profile of PSC I2 subjects is identical whether or not the patient has right-sided dysplasia.
Extended Data Fig. 3: Cytokines secreted by CD4 T-cells across transcriptional clusters. A-D, Proportion of right colon lamina propria CD4 T-cells expressing IL-17A (A), IFNγ (B), TNFα (C), or Foxp3 (D) after 3 hours of stimulation with PMA/ionomycin. E, Proportion of right colon lamina propria CD4 T-cells that are IL-17A Foxp3− after 3 hours of stimulation with PMA/ionomycin. F, Proportion of right colon lamina propria CD4 T-cells that are Foxp3− IL-17A after 3 hours of stimulation with PMA/ionomycin. G-H, Proportion of right colon lamina propria cells that are Foxp3− and IFNγ− (G) or TNFα− (H) after 3 hours of stimulation with PMA/ionomycin. I-J, Proportion of right colon lamina propria cells that are Foxp3+ and IFNγ− (I) or TNFα− (J) after 3 hours of stimulation with PMA/ionomycin. A-J, Significance determined by Wilcoxon test (*** for p<0.05, **** for p<0.01, “ns” for not significant (p>0.05)).
Extended Data Fig. 4: Transcriptional identification of the IL17A+ FOXP3+ CD4 T-cells. A, Correlation of proportion of IL-17A+ FoxP3+ cells by flow cytometry versus scRNAseq at each quantile cutoff value used to identify positive (IL17A+ FOXP3+) cells. B, Normalized sum of differences in proportions between proportion of IL-17A+ FoxP3+ cells by flow cytometry and scRNAseq at each quantile cutoff value used to identify positive (IL17A+ FOXP3+) cells. C, Correlation of proportion of IL-17A+ FoxP3+ cells by flow cytometry versus scRNAseq at the quantile cutoff value used in Fig. 3 (0.94). Significance and correlation determined by two-sided Pearson correlation test. D, Proportion of each transcriptionally-determined cell type within total CD4 cells by patient.
Extended Data Fig. 5: Cytokine and cytotoxic granule expression of IL17A FOXP3 double positive CD4 T-cells as compared to IL17A single positive, FOXP3 single positive, and IL17A negative FOXP3 negative cells. Log 2 fold change of cytokine expression of double positive cells as compared to IL17A single positive (orange), FOXP3 single positive (blue), or IL17A FOXP3 double negative (gray). Filled circles represent genes significantly changed at p<0.1, and open circles represent genes that are not significantly changed p>0.1.
Extended Data Fig. 6: V(D)J usage by cell type in I2 PSC. A-E, TRBV (A), TRBD (B), TRBJ (C), TRAV (D) and TRAJ (E) gene usage by cell type amongst CD4 T-cells from I2 PSC patients. F, Proportion of cells containing amino acid motif “LA” in the TCR beta chain by cell type amongst I2 PSC patients using TRBD2. Gray lines denote paired values from the same patients. SP = “single positive”, DP = “double positive” i.e. IL17A+ FOXP3+. (*** for p<0.05). A-F, Datapoints and box plot color denotes cell type (pink = IL17A+ FOXP3+ DP, orange = IL17A+ SP, blue = FOXP3+ SP, gray = negative for IL17A and FOXP3).
Extended Data Fig. 7: V(D)J gene usage amongst IL17A+ FOXP3+ CD4 T-cells containing “LA” motif. A-D, TRBV (A), TRBJ (B), TRAV (C), and TRAJ (D) gene usage amongst IL17A+ FOXP3+ CD4 T-cells stratified by whether the Beta chain contains the “LA” amino acid motif. Significance determined by Chi-squared test.
Extended Data Fig. 8: Immunoglobulin genes are amongst the most differentially regulated genes in I2 versus U. Volcano plot of the negative log base 10 adjusted p-value versus log base 2 fold change of the genes differentially expressed in the whole tissue biopsies of I2 versus U patients. Closed circles denote genes coding for immunoglobulin constant region; heavy chain V, D, or J segments; or light chain V or J segments.
Extended Data Fig. 9: Features of the top plasma cell clones in PSC patients. A, Mean forward scatter (FSC) of right colon plasma cells across clusters as determined by flow cytometry. B, Proportion of IgA-secreting plasma cells amongst total right colon plasma cells as determined by ELISpot. C, Proportion of IgM-secreting plasma cells amongst total right colon plasma cells as determined by ELISpot. D, Proportion of the total repertoire made up by the top clone within each subject. E, Proportion of plasma cells of each isotype by clone. F, Mean amino acid divergence from inferred germline across entire heavy chain sequence of largest clones identified in each patient. G, Mean pairwise amino acid divergence across entire heavy chain sequence of largest clones identified in each patient. (A-D, F-G) Each symbol represents an individual patient (open circles denote patients without dysplasia at the time of sampling, “x” denote patients with dysplasia at the time of sampling, open squares denote patients indefinite for dysplasia at the time of sampling). Significance determined by Wilcoxon test (*** for p<0.005, **** for p<0.001, “ns” for not significant (p>0.05)).
Extended Data Fig. 10: Genes of the I2 PSC classifier model. Heatmap of expression of the 81 core I2 PSC genes (rows) amongst PSC patients (columns). Top annotation are patient characteristics: cluster, sex, inflammatory response (IR) ssGSEA score, and age. Annotated on the right are the genes present in wound healing related GO (Biological Processes) enriched with the 81 genes from the I2 signature.
Extended Data Table 1: Clinical and demographic information for patients with no history of dysplasia analyzed in Figure 1.

<table>
<thead>
<tr>
<th>Variable</th>
<th>HC, N = 48&lt;sup&gt;†&lt;/sup&gt;</th>
<th>IBD, N = 103&lt;sup&gt;†&lt;/sup&gt;</th>
<th>PSC, N = 65&lt;sup&gt;†&lt;/sup&gt;</th>
<th>p-value&lt;sup&gt;2&lt;/sup&gt;</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Demographic and Clinical Information</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sex</td>
<td></td>
<td></td>
<td></td>
<td>0.053</td>
</tr>
<tr>
<td>Female</td>
<td>27 (56%)</td>
<td>39 (38%)</td>
<td>23 (35%)</td>
<td></td>
</tr>
<tr>
<td>Male</td>
<td>21 (44%)</td>
<td>64 (62%)</td>
<td>42 (65%)</td>
<td></td>
</tr>
<tr>
<td>Race</td>
<td></td>
<td></td>
<td></td>
<td>0.070</td>
</tr>
<tr>
<td>Asian</td>
<td>2 (4.2%)</td>
<td>5 (4.9%)</td>
<td>5 (7.7%)</td>
<td></td>
</tr>
<tr>
<td>Black</td>
<td>16 (33%)</td>
<td>16 (16%)</td>
<td>16 (25%)</td>
<td></td>
</tr>
<tr>
<td>White</td>
<td>29 (60%)</td>
<td>82 (80%)</td>
<td>44 (68%)</td>
<td></td>
</tr>
<tr>
<td>Unknown</td>
<td>1 (2.1%)</td>
<td>0 (0%)</td>
<td>0 (0%)</td>
<td></td>
</tr>
<tr>
<td>Ethnicity</td>
<td></td>
<td></td>
<td></td>
<td>0.7</td>
</tr>
<tr>
<td>Hispanic/Latino</td>
<td>0 (0%)</td>
<td>2 (1.9%)</td>
<td>2 (3.1%)</td>
<td></td>
</tr>
<tr>
<td>Not Hispanic/Latino</td>
<td>48 (100%)</td>
<td>101 (98%)</td>
<td>63 (97%)</td>
<td></td>
</tr>
<tr>
<td>Age at procedure</td>
<td>52 (48, 55)</td>
<td>37 (28, 49)</td>
<td>34 (25, 46)</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Type of IBD</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>No IBD</td>
<td>48 (100%)</td>
<td>0 (0%)</td>
<td>5 (7.7%)</td>
<td></td>
</tr>
<tr>
<td>CD</td>
<td>0 (0%)</td>
<td>71 (69%)</td>
<td>16 (25%)</td>
<td></td>
</tr>
<tr>
<td>UC</td>
<td>0 (0%)</td>
<td>32 (31%)</td>
<td>43 (66%)</td>
<td></td>
</tr>
<tr>
<td>IC</td>
<td>0 (0%)</td>
<td>0 (0%)</td>
<td>1 (1.5%)</td>
<td></td>
</tr>
<tr>
<td>Age at IBD diagnosis</td>
<td>NA (NA, NA)</td>
<td>22 (16, 32)</td>
<td>22 (17, 30)</td>
<td>0.8</td>
</tr>
<tr>
<td>Age at PSC diagnosis</td>
<td>NA (NA, NA)</td>
<td>NA (NA, NA)</td>
<td>27 (19, 35)</td>
<td></td>
</tr>
<tr>
<td>Duration of IBD at procedure</td>
<td>NA (NA, NA)</td>
<td>12 (7, 19)</td>
<td>8 (4, 14)</td>
<td>0.009</td>
</tr>
<tr>
<td>Duration of PSC at procedure</td>
<td>NA (NA, NA)</td>
<td>NA (NA, NA)</td>
<td>6 (2, 11)</td>
<td></td>
</tr>
<tr>
<td>History of liver transplant</td>
<td>0 (0%)</td>
<td>0 (0%)</td>
<td>11 (17%)</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Medications</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5-aminosalicylic acid</td>
<td>0 (0%)</td>
<td>35 (34%)</td>
<td>30 (46%)</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>anti_IL12/23 mAb</td>
<td>0 (0%)</td>
<td>2 (1.9%)</td>
<td>1 (1.5%)</td>
<td>&gt;0.9</td>
</tr>
<tr>
<td>anti-intergrin mAb</td>
<td>0 (0%)</td>
<td>10 (9.7%)</td>
<td>7 (11%)</td>
<td>0.037</td>
</tr>
<tr>
<td>anti-TNFa mAb</td>
<td>0 (0%)</td>
<td>36 (35%)</td>
<td>10 (15%)</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Methotrexate</td>
<td>0 (0%)</td>
<td>6 (5.8%)</td>
<td>1 (1.5%)</td>
<td>0.2</td>
</tr>
<tr>
<td>Ursodiol</td>
<td>0 (0%)</td>
<td>0 (0%)</td>
<td>27 (42%)</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>JAK inhibitor</td>
<td>0 (0%)</td>
<td>3 (2.9%)</td>
<td>2 (3.1%)</td>
<td>0.6</td>
</tr>
<tr>
<td>Purine synthesis inhibitor</td>
<td>0 (0%)</td>
<td>36 (35%)</td>
<td>12 (18%)</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Steroids</td>
<td>0 (0%)</td>
<td>13 (13%)</td>
<td>9 (14%)</td>
<td>0.012</td>
</tr>
</tbody>
</table>

<sup>†</sup> n (%); Median (IQR)

<sup>2</sup> Pearson’s Chi-squared test; Fisher’s exact test; Kruskal-Wallis rank sum test
Extended Data Table 2: Clinical and demographic information for patients with active right-sided dysplasia analyzed in Figure 2.

<table>
<thead>
<tr>
<th>Variable</th>
<th>Sporadic, N = 8</th>
<th>IBD, N = 7</th>
<th>PSC, N = 6</th>
<th>p-value $^2$</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Demographic and Clinical Information</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sex</td>
<td></td>
<td></td>
<td></td>
<td>0.6</td>
</tr>
<tr>
<td>Female</td>
<td>6 (75%)</td>
<td>4 (57%)</td>
<td>3 (50%)</td>
<td></td>
</tr>
<tr>
<td>Male</td>
<td>2 (25%)</td>
<td>3 (43%)</td>
<td>3 (50%)</td>
<td></td>
</tr>
<tr>
<td>Race</td>
<td></td>
<td></td>
<td></td>
<td>0.053</td>
</tr>
<tr>
<td>Asian</td>
<td>0 (0%)</td>
<td>0 (0%)</td>
<td>0 (0%)</td>
<td></td>
</tr>
<tr>
<td>Black</td>
<td>3 (38%)</td>
<td>0 (0%)</td>
<td>0 (0%)</td>
<td></td>
</tr>
<tr>
<td>White</td>
<td>4 (50%)</td>
<td>7 (100%)</td>
<td>6 (100%)</td>
<td></td>
</tr>
<tr>
<td>Unknown</td>
<td>1 (12%)</td>
<td>0 (0%)</td>
<td>0 (0%)</td>
<td></td>
</tr>
<tr>
<td>Age at procedure</td>
<td>58 (53, 65)</td>
<td>56 (42, 61)</td>
<td>36 (31, 42)</td>
<td>0.10</td>
</tr>
<tr>
<td>Type of IBD</td>
<td></td>
<td></td>
<td></td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>No IBD</td>
<td>8 (100%)</td>
<td>0 (0%)</td>
<td>0 (0%)</td>
<td></td>
</tr>
<tr>
<td>CD</td>
<td>0 (0%)</td>
<td>3 (43%)</td>
<td>3 (50%)</td>
<td></td>
</tr>
<tr>
<td>UC</td>
<td>0 (0%)</td>
<td>3 (43%)</td>
<td>3 (50%)</td>
<td></td>
</tr>
<tr>
<td>IC</td>
<td>0 (0%)</td>
<td>1 (14%)</td>
<td>0 (0%)</td>
<td></td>
</tr>
<tr>
<td>Age at IBD diagnosis</td>
<td>NA (NA, NA)</td>
<td>41 (28, 46)</td>
<td>25 (14, 35)</td>
<td>0.2</td>
</tr>
<tr>
<td>Age at PSC diagnosis</td>
<td>NA (NA, NA)</td>
<td>NA (NA, NA)</td>
<td>33 (18, 39)</td>
<td>&gt;0.9</td>
</tr>
<tr>
<td>Duration of IBD at procedure</td>
<td>NA (NA, NA)</td>
<td>10 (4, 24)</td>
<td>15 (7, 17)</td>
<td>&gt;0.9</td>
</tr>
<tr>
<td>Duration of PSC at procedure</td>
<td>NA (NA, NA)</td>
<td>NA (NA, NA)</td>
<td>8.1 (2.9, 15.4)</td>
<td></td>
</tr>
<tr>
<td>Age at first diagnosis of right colon dysplasia</td>
<td>58 (53, 65)</td>
<td>47 (42, 59)</td>
<td>36 (31, 42)</td>
<td>0.074</td>
</tr>
<tr>
<td>Duration of IBD at first right colon dysplasia</td>
<td>NA (NA, NA)</td>
<td>5 (2, 22)</td>
<td>15 (6, 17)</td>
<td>0.6</td>
</tr>
<tr>
<td>Duration of PSC at first right colon dysplasia</td>
<td>NA (NA, NA)</td>
<td>NA (NA, NA)</td>
<td>7.6 (2.7, 15.3)</td>
<td></td>
</tr>
<tr>
<td>History of liver transplant</td>
<td>0 (0%)</td>
<td>0 (0%)</td>
<td>1 (17%)</td>
<td>0.3</td>
</tr>
<tr>
<td>Medications</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5-aminosalicylic acid</td>
<td>0 (0%)</td>
<td>2 (29%)</td>
<td>1 (17%)</td>
<td>0.3</td>
</tr>
<tr>
<td>anti_IL12/23 mAb</td>
<td>0 (0%)</td>
<td>0 (0%)</td>
<td>1 (17%)</td>
<td>0.3</td>
</tr>
<tr>
<td>anti-intergrin mAb</td>
<td>0 (0%)</td>
<td>1 (14%)</td>
<td>0 (0%)</td>
<td>0.6</td>
</tr>
<tr>
<td>anti-TNFa mAb</td>
<td>0 (0%)</td>
<td>1 (14%)</td>
<td>3 (50%)</td>
<td>0.043</td>
</tr>
<tr>
<td>Ursodiol</td>
<td>0 (0%)</td>
<td>0 (0%)</td>
<td>1 (17%)</td>
<td>0.3</td>
</tr>
<tr>
<td>JAK inhibitor</td>
<td>0 (0%)</td>
<td>1 (14%)</td>
<td>0 (0%)</td>
<td>0.6</td>
</tr>
<tr>
<td>Purine synthesis inhibitor</td>
<td>0 (0%)</td>
<td>1 (14%)</td>
<td>3 (50%)</td>
<td>0.043</td>
</tr>
<tr>
<td>Steroids</td>
<td>0 (0%)</td>
<td>1 (14%)</td>
<td>2 (33%)</td>
<td>0.2</td>
</tr>
</tbody>
</table>

$^1$ n (%); Median (IQR)

$^2$ Fisher’s exact test; Kruskal-Wallis rank sum test
## Extended Data Table 3: Patients imputed HLA haplotypes and percentage of IL17A+ FOXP3+ DP CD4 T−cells with LA motif

<table>
<thead>
<tr>
<th>Sample ID</th>
<th>Cluster</th>
<th>HLA</th>
<th>n DP T-cells</th>
<th>Proportion of DP T-cells with LA</th>
</tr>
</thead>
<tbody>
<tr>
<td>PSC50B</td>
<td>I2</td>
<td>AH8.1</td>
<td>91</td>
<td>0.14</td>
</tr>
<tr>
<td>PSC70D</td>
<td>I2</td>
<td>AH8.1</td>
<td>174</td>
<td>0.12</td>
</tr>
<tr>
<td>PSC84B</td>
<td>I2</td>
<td>AH8.1</td>
<td>123</td>
<td>0.11</td>
</tr>
<tr>
<td>PSC94</td>
<td>I2</td>
<td>AH8.1</td>
<td>56</td>
<td>0.11</td>
</tr>
</tbody>
</table>