THE UNIVERSITY OF CHICAGO

A UNIQUE ADAPTIVE INFLAMMATORY SIGNATURE IS LINKED TO DYSPLASIA IN PRIMARY SCLEROSING CHOLANGITIS

A DISSERTATION SUBMITTED TO

THE FACULTY OF THE DIVISION OF THE BIOLOGICAL SCIENCES

AND THE PRITZKER SCHOOL OF MEDICINE

IN CANDIDACY FOR THE DEGREE OF

DOCTOR OF PHILOSOPHY

INTERDISCIPLINARY SCIENTIST TRAINING PROGRAM: IMMUNOLOGY

BY

DUSTIN GARLABAN JAMES SHAW

CHICAGO, ILLINOIS

DECEMBER 2021

To all those that helped me along my way.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGEMENTS

In more ways than one, this thesis was a product of the work and efforts of so many people, and I am forever thankful to all those who contributed and supported me. To pursue such a project, and to even be a graduate student, is a privilege not afforded to many and I am grateful beyond words for the people and circumstances that have enabled it.

I would like to thank Bana Jabri for taking a chance on me and encouraging me to take this path, as I would have never thought to do so otherwise. Thank you for all that you have taught me in the ways of critical thinking and approaches to human studies. Thank you also for your patience in those times in which I struggled and making sure that I kept moving, even when I felt like giving up.

I would like to thank Patrick Wilson for his mentorship and sharing his excitement and innovative spirit with me. You allowed me to become someone unafraid of trying new methods and experimental techniques, which ultimately solidified the success of this project. The environment you created in your lab made it a joy to work in and made me feel empowered to do my very best work.

I am indebted to the people who helped move this project forward. An enormous thank you to Nick DiNardi, Saideep Gona, and Anni Wang who did so much work day-to-day to keep this project running. I very much enjoyed working with you and wish you the best of luck in your future careers in science and medicine. I must thank all our collaborators, but especially Raúl Aguirre-Gamboa, Jacob Barlow, and Marcos Viera for their enormous contributions to this project. Raúl completely transformed the transcriptional data in a way that I am still impressed with to this day. Jacob gave us our first potential look into the source of the antigen with his very

own innovative techniques. Marcos's analysis of the immunoglobulin sequences was central to the demonstration of antigen drive and has given us the tools we need to help us identify the antigen. It was a joy to work with all three of you. Thank you all for being so committed to this project with me. I found our interactions so engaging and makes me excited for a career in science full of collaborations.

I also want to thank the members of both labs, especially Elaine, Jordan, Kishan, Wiola, and Zach from Bana's lab for keeping me grounded and for all the fun times and laughs we had together. From Patrick's lab, I would like to especially thank Angela, Carole, Chris, and Haley for making feel welcome and for all the time we spent together. I would like to thank Toufic for always looking out for me and giving me advice as a young graduate student; I very much enjoyed going to WWE Monday Night Raw all those times. I need to also thank Cezary for teaching me everything I know about flow cytometry, and for guiding me and always helping me out when I needed it.

I would like to give an enormous thanks to the friends who supported me and made the past years such a fun period of my life. To Anthony, Christine, Claire, Cody, David, Desmond, Jeff, Jeremy, Mike, Molly, Remington, Tess, and Van: I am grateful to be surrounded by you and for all the laughs and adventures that we had together. Thanks for listening to me vent and sticking with me during low periods. I could not imagine that anyone could have better friends.

Finally, I need to thank my family. I come from two extended families with very different relationships to education. For one family, education was not a given and they had to fight to obtain it. To them, education was precious because it meant access to a good job and then indirectly safety and stability. I am grateful for the lessons I learned from this family's experiences, as it has taught me the value and necessity of education, and to never take it for

granted. My work ethic comes from them. For my other family, education was something bigger than a diploma. It was an exciting, intellectual pursuit that surpassed its practical aspects. This family taught me how enjoyable and fulfilling an education can be, and how this can drive you to new heights. My inspiration comes from them. Getting to this point is a direct result of all the support, guidance, energy, resources, lessons, and time given to me from my immediate and extended families.

Dougie, thanks for all that we shared together over the years. Though we fought, I know that you are someone I can always depend on no matter what. I admire your maturity, kindness, thoughtfulness, level-headedness, creativity, and your ability to always get things done with just a second to spare. I am looking forward to your thesis defense in the next year or two, and hope that we can work together in the future.

Dad, thank you for all those times that you spent helping me with math homework, even when I went off to college. You always spent as much time as necessary to not only help me finish the assignment, but to make sure that I fully understood the concepts and how to solve the problem. I appreciate that you always took a genuine interest in my work, and that you were always someone that I could discuss it with.

Maman, thank you for all those late nights that you spent awake with me, proof-reading my essays and helping me glue my science fair poster together. Thank you for always being someone I could go to, no matter what. I'm sorry that I forgot to bring my ID to the SATs and thank you for driving me to the make-up test date. Also, don't blame yourself for the C+ that I got in social studies in 7th grade- I shouldn't have left it to the night before.

The unconditional love and support that the three of you have given me is more than I could have ever asked for and has made me the person that I am today.

ABSTRACT

Primary sclerosing cholangitis (PSC) is an immune-mediated, cholestatic liver disease characterized by progressive fibrosis of the liver bile ducts. PSC patients have significantly greater rates of liver failure, portal hypertension, and cholangiocarcinoma (CCA) than the general population, and the median survival time from diagnosis is only 12 years without a liver transplant. Aside from transplant, there are no therapies that can prevent or cure the liver pathology. Nearly all PSC patients have a secondary diagnosis of inflammatory bowel disease (IBD), however the presentation of PSC colitis differs from that which is observed in IBD alone. Additionally, PSC patients have significantly increased risk for colorectal cancer (CRC) that is far greater than the already increased risk of CRC in IBD. Some hypothesize that PSC colitis is distinct from IBD though no differences have been formally identified. Others believe that PSC is antigen driven, due to the strong association with the human leukocyte antigen (HLA) locus by genome-wide association studies (GWAS). In this thesis we provide formal evidence that PSC colitis is distinct from IBD, and is characterized by the presence of lamina propria immunoglobulin G (IgG)-producing plasma cells and interleukin (IL)-17A+ forkhead box P3 (Foxp3)+ CD4 T-cells. We demonstrate that both cell types show signs of selection, consistent with the antigen drive hypothesis. 16S sequencing identified a handful of bacterial taxa enriched specifically in PSC colitis, suggesting that a driving antigen could be bacterial in origin. Finally, we show that PSC patients with this unique colitis are at greater risk for dysplasia than PSC patients without it. Our data suggests that an antigen in PSC drives inflammation and subsequently CRC, representing a hypothetical therapeutic target for the prevention of colonic inflammation and CRC, and potentially liver inflammation as well.

INTRODUCTION

Adaptive immunity and tolerance

Very broadly, the function of our immune system is to maximize our survival and ability to reproduce by recognizing and managing any foreign- or self-agent that would do us harm. Though colloquially thought of as a defense against pathogens, our immune system plays vital roles in homeostasis, wound healing, and defense against cancer. We have evolved two immune compartments termed the innate and adaptive immune system. Adaptive immune systems, as they are in humans, can be found in all jawed vertebrates[1] whereas virtually all living organisms possess some aspect of innate immunity[2], highlighting the importance of immunity for the survival of individuals and species. The innate immune system consists of physical and chemical barriers as well as a set of specialized immune cells that recognize evolutionarily conserved molecular patterns that signal foreign invasion or damage to our tissues[3,4]. Neutrophils, macrophages, eosinophils, and other specialized cells of the innate immune system are quickly activated in response to pathogen or damage signals that bind to their pattern recognition receptors (PRRs) and are the first line of defense against dangerous agents. However, innate immune cells can only respond to agents that present in the context of the limited set of molecules recognized by their PRRs.

In contrast, T-cells and B-cell, the adaptive immune cells collectively named lymphocytes, can recognize an essentially limitless number of amino acid-based molecules due to the near infinite possible conformations of their receptors. The diversity of specificities afforded to T-cell receptors (TCRs) and B-cell receptors (BCRs), also known as immunoglobulins (Igs), is achieved by genetic recombination of these receptors. Early in its

development, a T-cell will randomly splice and merge one of 52 possible variable genes, with one of two diversity genes, and one of 13 joining genes to create a DNA segment coding for half of the TCR called the beta chain. The other half of the TCR, the alpha chain, also will randomly splice one of 70-80 V genes with one of 61 J genes. This process is referred to as V(D)J recombination and allows for approximately 5.8 million possible unique TCRs. B-cells also undergo V(D)J recombination using 65, 27, and 6 possible V, D, and J genes for its heavy chain, and 70 and 9 possible V and J genes for its light chain resulting in 3.4 million possible combinations of Ig. Additionally, TCRs and Ig undergo junctional diversification due to the improper joining of V, D, and J segments, resulting in an estimated 10^{18} and 10^{14} possible unique TCRs and Igs respectively[5].

The near infinite possible conformations of TCRs and Igs are the foundation by which lymphocytes pools are poised to respond to nearly any potential pathogenic antigen. In 1976, Frank Burnet postulated that from this pool of randomly generated receptors that there is by chance at least one that can recognize an assaulting antigen, which will lead to the activation and division of the antigen-specific lymphocyte to a great enough degree to allow for the clearance of the antigen[6]. This theory, called clonal selection, is now validated by countless scientific studies and has inspired many to investigate the TCR and Ig repertoires in the context of many infections and diseases. In certain contexts of antigen exposure such as bacterial or viral infection, there are observable “clonal expansions” of specific TCRs and Ig in the tissues in which the antigen is present[7–10]. Analysis of the TCR and Ig repertoire of individuals with infections or immune-mediated diseases can potentially be a very powerful tool in understanding how the adaptive immune system is responding to these assaults on homeostasis. Some labs, including ours, have

even develop technique to clone and express TCRs[11] and Igs[12] for functional studies to determine specificity and explore factors that contribute to antigen affinity.

With such diversity is appreciated that randomly generated TCRs can recognize tissue antigens as well. It is critical to an organism's survival that T-cells and B-cells only respond to dangerous antigens and not to our non-pathogenic self-antigens, which is achieved to a great deal of success through control mechanisms called central and peripheral tolerance. In the thymus, developing T-cells are presented an array of self-peptides loaded into major histocompatibility complex (MHC; also known as HLA in humans) molecules. The fate of the T-cell depends on the strength of the interaction of its TCR with the MHC-peptide complex[13,14]. T-cells with too high affinity will die by apoptosis, while those that do not signal either too strongly or too weakly will be allowed to mature and pass into the periphery. In cases with the TCR affinity is strong, but not strong enough to induce apoptosis, the T-cell can develop into a regulatory T-cell (Treg), which are able to suppress immune responses. Both of these mechanisms of central tolerance are essential for avoiding inflammatory responses against one's tissue[15,16]. Autoreactive T-cells that escape central tolerance mechanisms can be deleted[17], anergized[18], or converted to a Treg in the periphery[19]. Similar mechanisms of tolerance exists in B-cells, with the added ability to edit their receptor during selection in order to decrease affinity for self and avoid deletion[20]. Tolerization is an ongoing process and is not perfect, and a number of diseases are associated with defects in or loss of tolerance[21–23].

The intestine is an immunological battleground that perfectly exemplifies the difficult task taken on by the immune system. The human gastrointestinal tract is a region in which host tissue is exposed to a broad variety of ingested materials, secreted molecules, and over 100 trillion bacteria[24]. Within this diverse antigen environment, the intestinal immune

system must allow passage of beneficial nutrients and metabolites, ward off pathogens, and keep commensal bacteria in check to avoid opportunistic infections. Barriers such as mucus and tight junctions between sheets of epithelial cells prevent bacteria and viruses from invading into the more sterile inner tissues[25]. Immunoglobulin A (IgA), a non-inflammatory isotype of Ig, as well as antimicrobial peptides are constantly secreted into the lumen of the gut in order to maintain and ward off bacteria[26,27]. Dendritic cells (DCs) constantly sample antigens proximal to the epithelial surface of the lumen and migrate to organized lymphoid structures to present to T-cells[28]. In the absence of activating pathogen or damage signals, DCs will tolerize T-cells to the presented antigens, whereas DCs activated by such signals will relay activation signals to T-cells and induce an adaptive response. Occasionally, as is thought to be the case in celiac disease (CeD) for example[23], when DCs present otherwise innocuous peptides in the context of pathogen or damage signals, an inappropriate adaptive immune response is mounted against a non-dangerous antigen, resulting in tissue destruction.

Immune-mediated inflammatory disorders

Diseases in which tissue is chronically damaged by inflammatory pathways are called immune-mediated inflammatory disorders (IMIDs). These include CeD, type 1 diabetes (T1D), systemic lupus erythematosus (SLE), and IBD. Collectively, IMID has an incidence of approximately 5-7% in Western society, though incidence varies geographically[29]. IMIDs can affect any tissues and contribute significantly to morbidity, mortality, and reduced quality of life. Though therapies exist for many IMIDs, much remains to be understood about the mechanisms and causes of these diseases. One of the best characterized and understood IMIDs is CeD, in

which a subset of genetically susceptible individuals mount a T-cell mediated inflammatory response against gluten peptides[30].

A physician named Willem Dicke first identified that a component of wheat was the causative agent in CeD in the 1940s[31]. Since then, countless studies have filled in the many details about the complex pathogenesis of this disease. CeD is driven by the ingestion of gluten in a subset of genetically vulnerable individuals. CeD almost exclusively occurs in individuals of HLA-DQ2 or HLA-DQ8 haplotypes[32], because proline-rich gluten peptides are more strongly bound by the HLA molecules coded by this haplotype than other haplotypes[33]. Gluten peptides loaded on HLA can then present to T-cells resulting in an antigen-dependent expansion of T-cells expressing specific TCRs[34,35], leading to tissue destruction. In addition to the T-cell response, there is also a coordinated B-cell response which is also restricted in its Ig usage[36]. Importantly, the epitopes recognized by gluten-specific T- and B-cells are largely overlapping[37,38]. For this reason, although the exact function of B-cells in CeD pathogenesis is not known, it is hypothesized that they potentiate inflammation by presenting recognizable gluten epitopes to T-cells. Though there are many advances in understanding the mechanisms of tissue destruction in CeD, for many other IMIDs, we have little to know understanding of their pathogenesis. Many IMIDs have a strong association with HLA[39], so it is believed that there might also be a similar antigen drive as seen in CeD, though none have yet to be identified.

The roles of inflammation and infection in the development of cancer

It is well appreciated that cancer arises from sites of chronic infection and inflammation[40]. As early as 1863, German physician Rudolf Virchow speculated that cancer occurred in inflamed tissues, from inflammatory factors that promoted cell proliferation[41].

Induction of cell proliferation by the immune system is critical mechanism of homeostatic processes such as wound healing and clearing of infections. For example, T-cells that infiltrate mechanically damaged tissues produce several growth factors that induces proliferation of surrounding healthy cells as a way of healing the wound[42]. In the context of certain enteric infections such as *Listeria monocytogenes*, our immune system promotes epithelial cell proliferation and turnover via secretion of IL-11 and IL-22 in order to dispose of infected cells and clear the pathogen[43]. Though immune factors are critical for survival and homeostasis, they can also unintentionally induce tumor growth and promotion. Epidermal growth factor (EGF), a peptide secreted by macrophages and essential to wound healing[44], can also induce the proliferation of tumor cells[45]. Tumor cells can even secrete their own factors, such as colony stimulating factor-1 (CSF-1), that induce EGF and create a positive feedback loop favoring tumor proliferation[46].

Pro-proliferative signals alone are not sufficient to induce and promote cancer, however. Though cytokines and other factors can promote proliferation and progression of tumor cells, they themselves cannot convert a healthy cell into a neoplastic cell. Rather, the initial generation of a cancerous cell is dependent on the accumulation of somatic mutations that disrupt the function of tumor suppressors or oncogenes. It was first recognized in 1941 that cancers arose from “subthreshold neoplastic states”, or seemingly normal cells which have accumulated mutations which would allow them to uncontrollably divide in response to the appropriate external proliferation signals[47,48]. There is a critical role for the events that ‘initiate’ neoplasia, i.e. events that cause somatic mutations, and the events that ‘promote’ tumorigenesis such as the secretion of inflammatory factors[40]. Initiation is irreversible, and though not necessarily pathogenic, results in a set of cells that are poised to become cancerous, given the right

conditions. Theoretically, such cells could remain dormant for many years, until conditions that promote their proliferation occur. Promotion alone also cannot produce tumors, and rather creates the milieu in which somatically mutate cells can divide and metastasize. Without initiation, no amount of cytokine can create a tumor. However, without promotion, a mutated cell cannot develop into a tumor either. Initiation and promotion are likely non-binary states, and there probably exists a spectrum of potency to each. One could even imagine situations in which their potencies could compensate for each other in the process of tumorigenesis. For example, it could be possible that highly mutated cells would require less inflammatory signals to develop into a tumor, due to a sufficient number of defects in intrinsic cellular regulation. Conversely, it might also be possible that a cell baring relatively few somatic mutations could become a tumor in response to extremely intense proliferation signals. To what degree tumorigenesis depends on the degree of initiation and promotion might differ by cancer type, the individual, or other factors. Additionally, the potency of initiation and promotion could depend on the number of mutations and genes affected or the nature (i.e. cellular and cytokine composition) of the inflammatory milieu respectively. Regardless, the study of the relationship between initiation and promotion in tumorigenesis is a high-yield pursuit, with implications for the prevention and treatment of cancer.

Infectious disease is also highly associated with the development of cancer, and approximately 15% of the worldwide incidence of cancer can be attributed to infections[49]. Viruses and bacteria increase the risk of cancer by contributing to both initiation and promotion events. First, chronic inflammation associated with persistent infections can introduce somatic mutations by the production of reactive oxygen species (ROS) that directly damage DNA[50,51]. This chronic inflammation can also create a pro-proliferative environment for cancers as

described in the paragraphs above. Secondly, certain viruses such as hepatitis B virus (HBV), the virus associated with hepatocellular carcinoma (HCC), can integrate its genome into our chromosomes, disrupting endogenous gene function of critical tumor suppressors, and even introduce virally-encoded oncogenes[52]. Human papilloma virus (HPV), in the context of cervical cancer development, will integrate its E6 and E7 genes into epithelial cell genomes, which allows for the stable and persistent expression of these viral proteins that bind and interfere with tumor suppressive capacity of p53 and retinoblastoma, respectively[53,54].

As infections contribute to both the initiation and promotion of certain tumors, they are prime targets in the prevention of cancer. Prevention of initial infections and clearance of chronic infections could theoretically reduce the number of initiation and promotion events that would lead to cancer. However, the role of viruses and bacteria in the generation of initiation mutations also means that clearing an infection will not entirely reduce the risk of cancer associated with the pathogen. In fact, after initial initiation events, clearance of certain pathogens can exacerbate tumorigenesis. For example, one study determined that after integration of the HPV genome into our chromosomes, that elimination of non-integrated, replicating virus is a critical step in the progression of aberrant epithelial cells into overt carcinoma[55]. Therefore, when it comes to intervention in pathogen-associated and inflammatory cancers, timing is key, and there exists a point of no return after which clearance of the infection will not stop tumorigenesis and may even exacerbate it.

Primary Sclerosing Cholangitis

PSC is a heterogenous, progressive liver disease characterized by inflammation and fibrosis of the intra- and extra-hepatic biliary ducts[56]. Progression of PSC leads to wall

thickening and narrowing or complete blockage of the bile ducts, resulting in the loss of smaller intrahepatic ducts as well as dilation of the ducts proximal to the stricture[57]. Clinical manifestations of PSC include jaundice, pruritus, hepatosplenomegaly, fatigue, and weight loss[58,59]. Patients with PSC often have elevated serum alkaline phosphatase (AP) levels, which serves as a prognostic marker for clinical outcomes[60]. Cholangiography of affected livers will often show a characteristic "beading" of the bile ducts corresponding to the strictures and proximal dilations[61].

Historically, visualization of strictures via endoscopic retrograde cholangiopancreatography (ERCP) was the standard in diagnosis of PSC, however this is now reserved for the evaluation of major strictures, interventional procedures, collection of tissue specimens, or diagnosis of early stage PSC, before strictures can be visualized non-invasively[62]. Usually, a diagnosis of PSC is rendered in the context of chronic cholestatic liver test abnormalities (including elevated AP) and cholangiographic evidence of multifocal bile duct strictures, after ruling out other liver abnormalities including primary biliary cirrhosis (PBC) and IgG4-associated autoimmune hepatitis (IgG4 AIH). Though a liver biopsy is not necessary for diagnosis of PSC, one can occasionally be taken if there are suspicions of PSC that only affects the small ducts or to rule out other liver disorders[63]. These biopsies, if deterministic, will often show inflammation and occasionally "onion skin" fibrosis which is pathognomic to PSC[64].

The liver abnormalities in PSC lead to cirrhosis, portal hypertension, and liver failure[65]. CCA is observed at incredibly higher rates in PSC patients than in the general population. Patients with PSC have a 398-fold increased risk of developing CCA as compared to the general population[66]. Additionally, one study surveying gallbladders removed from PSC patients found that 37% of removed gallbladders had dysplasia and 14% had overt carcinoma[67]. Though

estimated to have an incidence rate of only 0.77 per 100,000 individuals in North America[68], PSC is the fourth leading indication for liver transplant[56], and approximately 40% of transplanted individuals will re-experience symptoms of PSC[69]. Without a liver transplant, however, median survival from time of diagnosis is estimated at about 12 years[58]. There are currently no therapies to cure, reverse, or alleviate the liver pathologies associated with PSC.

The unknown etiology of PSC complicates the efforts in developing therapies for this disease. Risk for PSC is greater in relatives of individuals with PSC as compared to relatives of unaffected individuals[70], suggesting a genetic component to PSC. GWAS point to a number of genetic loci potentially implicated in PSC[71,72], the most strongly associated of which being the chromosomal locus coding for HLA. The degree of association with the HLA locus is to a similar degree observed in other classic autoimmune diseases[39,73], suggesting that HLA could play a role in PSC much like it does in CeD. A few studies have linked *HLA-DRB1* to PSC[74,75], however this HLA molecule has yet to be definitively implicated. Outside of the HLA locus, genes within 22 genetic loci have been identified in PSC[71–73,76,77], all of which are shared with other autoimmune diseases[39]. These genes include: ATXN2, BACH2, BCL2L11, CCDC88B, CCL20, CD226, CD28, CLEC16A, CTLA4, GPR35, FOXP1, HDAC7, IL2, IL21, IL2RA, MMEL1, MST1, NFKB1, PRKD2, PSMG1, RFX4, RIC8B, SH2B3, SIK2, SOCS1, STRN4, TCF4, TNFRSF14, and UBASH3A.

Genetic loci only explain about 10% of disease susceptibility[39], however. It is therefore hypothesized that there is a significant environmental contribution to PSC, though none have been formally identified. A prevailing hypothesis in the field is that PSC could be caused by a microbe. In support of this, a number of studies have observed differences in the fecal microbial communities of PSC subjects versus healthy controls or versus other closely related disorders[78–

[81]. Additionally, though low in sample size, a few clinical studies provide preliminary evidence that both vancomycin and metronidazole can improve liver function tests and PSC-associated inflammation[82–84]. One study even found that *Klebsiella pneumonia*, a bacterium isolated from the feces of patients with PSC, can induce intestinal barrier permeability that allows bacteria to translocate to the liver, inducing a hepatic inflammatory response[85]. *K. pneumonia*, thus, is one potential therapeutic target in PSC, though it remains to be seen whether other bacteria found in the feces of PSC patients could also reproduce the effects observed by this bacterium. There are likely other bacteria that can replicate these pathologies, especially since vancomycin and metronidazole, neither of which directly affect *K. pneumonia*, improved the symptoms of PSC patients.

Intestinal inflammation in PSC

The impact of the gut microbiota on disease progression has been a subject of great interest, especially given the intestinal pathologies associated with PSC. In addition to the liver pathologies that burden these patients, nearly all individuals with PSC have a concurrent diagnosis of IBD[86]. IBD is a group of inflammatory diseases that affects the intestines, mainly classified as either ulcerative colitis (UC) or Crohn's disease (CD). UC is characterized by continuous span of inflammation of the mucosa of the colon, starting rectally, though it can progress to the entire extent of the colon. Symptoms include constipation, diarrhea (with or without blood and/or mucus), bowel movement urgency, abdominal pain, fever, malaise and weight loss[87]. Complications of UC include severe bleeding, toxic megacolon, and perforation of the colon. CD, in contrast, consists of non-continuous, deeper inflammation that can affect any part of the gastrointestinal tract, leading to strictures, fistulas, and abscesses[88]. The inflammation

observed in both forms of IBD are non-specific, though features such as granulomas and fissures can be used to distinguish between them[89]. Additional diagnostics such as imaging and endoscopy can be used to distinguish the two types, though differentiation between the two can be difficult when a patient presents with exclusively colonic inflammation[90]. In contrast to either UC or CD, the IBD observed in the context of PSC is most often pan-colitic, with inflammation most severe in the ascending colon, terminal ileum involvement, and sparing of the rectum[91,92]. PSC-IBD can also present as inflammation affecting only the ascending, without visible inflammation distally[93,94]. PSC patients are also younger at the age of IBD diagnosis than patients without PSC[92].

Due to the near ubiquitous presence of IBD in patients with PSC, many studies have investigated for genetic links between the two diseases. GWASs demonstrate a much weaker association with the HLA locus in IBD than in PSC, and only about 5% of all risk loci identified in IBD are also found in PSC[39]. The genome wide genetic correlation (r_G) between PSC and UC is 0.29, and r_G between PSC and CD is only 0.04. In contrast, r_G between UC and CD is 0.56[72], suggesting that UC and CD are more genetically similar to each other than PSC to either form of IBD. These genetic analyses underscore the differences in clinical observations between IBD and PSC-IBD, suggesting that PSC-colitis is in fact its own unique form of inflammation.

How the liver pathologies in PSC related to the intestinal manifestations is unknown, but it is hypothesized that both pathologies are related and may have a common origin. The intestine and liver are anatomically interconnected, allowing for travel of metabolites, nutrients, immune cells, bile acids, and other factors between the two. The liver delivers bile into the duodenum via the common bile duct, and liver-derived produces travel through the small intestine and reach the colon. Blood draining from the large intestine converge into the portal vein which feeds into the

capillaries of the liver. By this route, nutrients, metabolites, translocated bacteria, and bacterial products can reach the liver tissue. Some propose that pathogenic T-cells activated in the colon can travel to and seed liver tissue, thus causing the liver pathologies[95]. Others believe that bacteria or bacterial byproducts leak from the gut to the liver causing an inflammatory response, based on data showing that colonic bacterial overgrowth can cause liver disease[96].

Despite the clinical and genetic differences between the intestinal inflammation, there are no differences in the treatment of IBD with or without PSC. The medical management of IBD in either case is dependent on the severity and extent of the inflammation, as determined by colonoscopy. For mild to moderate colitis that is distally restricted, rectal application of a 5-aminosalicylic acid derivative (5-ASA), a nonsteroidal anti-inflammatory drug (NSAID) clinically proven to induce remission, is recommended[97]. If colitis extends more proximally, if the inflammation becomes severe, or if the patient does not respond to rectal 5-ASA, then they can receive oral 5-ASA, steroids, or biologics including anti-tumor necrosis factor (TNF) agents, anti-integrin antibody, anti-interleukin(IL)-12/23 antibody, or janus kinase (JAK) inhibitors[98–100]. Though each therapy acts via different mechanisms, the goal of each is to induce sustained remission of the intestinal inflammation and reduce occurrence and severity of symptoms. In the case of severe disease refractory to medications, or with complications such as toxic megacolon, hemorrhage, or perforations, a partial or complete resection of the colon or affected portion of small intestine is indicated[101,102].

Another indication for surgical resection is the presence of CRC or precursory dysplasia[101,102]. It is well appreciated that IBD is associated with increased risk for colorectal cancer[103,104], with both UC and CD having very comparable frequencies of and time to the development of CRC[105]. Though CRC can occur in individuals without IBD (referred to as

sporadic CRC), the risk is much greater in those with IBD[106]. CRC is a slowly progressing, epithelial-derived cancer that generally begins as dysplasia, which is defined as an abnormal epithelium confined to the basement membrane[107]. Dysplasia, though itself benign, has the potential to progress to CRC, and is the best marker for risk of malignant CRC in IBD[108]. For this reason, IBD patients are recommended to undergo dysplasia surveillance colonoscopies every one to three years[109], so that a partial or complete colectomy can be performed before the dysplasia progresses to malignant CRC.

Within cohorts of IBD patients, the major risk factors associated with the development of CRC include duration of disease[110], extent of disease[111], and severity of inflammation[112]. Another major factor that increases the risk of CRC within IBD cohorts is a concomitant diagnosis of PSC. In fact, patients with PSC-IBD are at a shocking five times greater risk for CRC than IBD patients without PSC[113]. One study reported that the absolute cumulative risk of CRC at 10, 20, and 25 year post diagnosis of colitis to be 2%, 5%, and 10% for patients with IBD, whereas it is 9%, 31%, and 50% for patients with PSC and IBD[114]. Due to the dramatic increase in risk for CRC, it is recommended that patient with IBD-PSC receive a screening colonoscopy on a yearly basis[115]. Consistent with the observations that PSC-IBD almost always involves the right colon, CRC is most often right-sided in PSC as opposed to IBD where CRC is more often left-sided[116,117]. Beyond the identification of PSC as a risk factor for IBD, it has never been formally demonstrated that the risk factors that predict dysplasia in IBD would apply to a cohort of PSC patients. Additionally, no study to date has looked for clinical and demographic risk factors unique to the development of CRC in PSC.

As the risk factors for CRC in IBD are linked to the presence, duration, and severity of inflammation, many hypothesize that inflammation is directly involved in the pathogenesis of

CRC. Though no formal mechanisms by which inflammation leads to CRC have been published, many believe that chronic inflammation damages DNA via ROS, resulting in mutations that confer loss or gains of function in important genes within epithelial cells[118]. Many studies have investigated the molecular mechanisms of CRC and found that genomic instability is at the basis of neoplasia. Chromosomal instability (CIN) accounts for about 85% of all CRC and microsatellite instability (MSI) for about 15% in both sporadic and IBD-associated CRC[119]. CINs are genomic alterations that include structural aberrations such as point mutations, insertions, deletions, chromosomal rearrangements, or gains or losses of entire chromosomes[120]. Microsatellites are short repeats of nucleotides often located near coding regions in DNA[121]. In MSI, microsatellites become hypermutable due to defects in DNA mismatch repair, leading to increased risk for genetic mutations[122]. Generally, the genes affected in CIN and MSI are the same in sporadic and IBD-associated CRC, however the order in which mutations arise differs[118]. For example, loss of APC function mutations are found earlier in the development of sporadic than in IBD-associated CRC, while p53 mutations are found earlier in IBD-associated than sporadic CRC.

The morphology and presentation of the two CRCs also differ. Sporadic CRC starts as a raised, well-defined polyp whereas IBD-associated CRC most often begins as a flat lesion, which can be difficult to see without specialized visualization methods[123,124]. IBD patients are also more likely to have multifocal CRC[125]. IBD demonstrates a "premalignant field effect" in which mutations found in CRC tumors can be observed in non-cancerous portions of the colon[126]. This likely suggests that inflammation introduces mutations in otherwise healthy tissue, creating regions at greater risk for developing into discernable dysplasia. Counterintuitively, given the dependence of CRC on inflammation in IBD, one recent study published that there is a reduction

in the density of CD3+, CD8+, and Foxp3+ immune cells in the stroma directly surrounding IBD-associated tumors, as compared to sporadic tumors[127]. However, better survival was reported in both groups for individuals with high frequencies of cells expressing these markers, potentially meaning that these cells have an anti-tumor function. Nonetheless, it would be interesting to understand why there would be a reduction of these cells in IBD patients. As inflammation can have both pro- and anti-tumor effects, it is likely dysplasia outcomes rely on multiple factors of inflammation.

Thesis aims

Whether PSC dysplasia develops by the same mechanisms as IBD dysplasia is completely unknown, and no studies have formally investigated either risk factors or mechanism of CRC in PSC specifically. The morbidity and mortality in this patient population is incredibly high, and given the lack effective medications and interventions, further investigation is critical. Because of the differences in presentation of inflammation and rates of CRC, we sought to investigate whether there were any features of the nature of PSC inflammation as compared to IBD inflammation that underlay the dramatically increased rates of CRC in PSC. We hypothesize that features unique to PSC colitis are directly related to the increased rates of CRC, and that the mechanisms by which CRC develop are different than in either IBD or sporadic CRC. Given the strong association with HLA, what we know about the impact of HLA in CeD, and the fact that PSC has features consistent with other autoimmune disorders, we hypothesize (as others have before us) that PSC intestinal and liver inflammation are antigen driven, and that a careful analysis of the adaptive immune repertoire would show signs consistent with antigen drive. We took a broad, systematic, and unbiased approach to this investigation, incorporating

clinical, transcriptional, and ex vivo cellular data to address our hypothesis. We present our work in the following chapters. We hope that that these data will be one steppingstone towards future investigations that will discover the exact mechanisms of PSC, identification of a driving antigen, and ultimately lead to therapies or even a cure.

MATERIALS AND METHODS

Enrollment of study subjects

All activities related to enrollment of subjects, collection of samples, and sample analysis were approved by the University of Chicago Institutional Review Board (IRB) and performed under IRB protocols 15773A and 13-1080.

On a weekly basis, 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. Any subject that passed our exclusion criteria was eligible for enrollment. Our exclusion criteria included: individuals 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 hard drug use such as cocaine, heroin, non-prescription methamphetamines; active use of blood thinners; severe comorbid diseases; individuals on active cancer treatment; and individuals who are pregnant. Approaching prospective subjects 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. Subjects that passed the inclusion criteria were approached the day of their procedure by a dedicated set of clinical coordinators and research technologists, who would get written consent of those individuals willing to be enrolled in our study.

Classification of subjects into diagnosis groups

Individuals enrolled in the study were categorized as either PSC, IBD, or control subjects. Categorization of each subject into diagnosis groups was done after careful review of the subjects' medical health records to ensure the validity of each diagnosis. Subjects 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. PSC subjects were further categorized as PSC-no IBD, PSC-UC, PSC-CD, or PSC-indeterminant colitis (PSC-IC) depending on the respective IBD diagnosis found in the individuals' chart. IBD subjects were also classified as UC, CD, or IC based on the details surrounding their IBD diagnosis. Importantly, as nearly all individuals with PSC and IBD have right sided colitis, we only enrolled IBD subjects who had an explicit history of right sided colitis. Any IBD subject without documented right sided colitis were excluded from the study. Any individuals without a diagnosis of PSC or IBD that were receiving screening colonoscopies for cancer screening or diagnostic abnormalities such as persistent diarrhea, were considered control subjects. Individuals who did not have any confirmed sporadic dysplasia on colonoscopy were classified as healthy controls, and those with a sporadic adenoma were classified as sporadic dysplasia controls. Any controls consented to the study that were determined to have signs of endoscopic or histologic inflammation were retrospectively excluded from the study.

It is not uncommon for subjects' IBD diagnosis to change over time. We therefore assigned subjects to an IBD subtype by the clinicians' diagnosis after evaluating all findings from the day of the procedure. Any individual without a diagnosis of PSC, that subsequently received a diagnosis of PSC or had strong suspicions for PSC, were excluded from the analysis. Individuals with an initial diagnosis of PSC that were later confirmed to not have PSC were re-classified as IBD without PSC.

A diagnosis of dysplasia was determined by evaluation of the histological reports associated with the colonoscopy. If the pathologist's report mentioned clear signs of adenoma,

low-grade dysplasia, high-grade dysplasia, or carcinoma the subject 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 subject was classified as indefinite for dysplasia. If no signs of *bona fide* or indefinite dysplasia were identified, the subject was classified as non-dysplastic.

Collection of patient clinical and demographic data

We searched all subjects' 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 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-related medications currently taken by the patients, including immunosuppressants, biologics, antibiotics, and steroids.

Collection of tissue specimens

During the colonoscopy, upon reaching the extent of the colon, the endoscopist would collect 8-10 tissue biopsies using 2.8mm or 3.2mm forceps. One of these biopsies was placed immediately into RNAprotect (Qiagen) and the remaining biopsies were placed into RPMI 1640 (Fisher Scientific). The samples were subsequently transported on ice back to the laboratory for processing. The tissue biopsy in RNAprotect was stored for 48-72 hours at 4C, after which the

solution was removed, and the tissue stored at -80C until needed. Tissue biopsies in RPMI were immediately processed upon return to the lab.

Tissue biopsy RNAseq

Whole tissue biopsies stored at -80C were thawed on ice and transferred to Starstedt tubes (Fisher Scientific) containing 350uL RLT Plus (Qiagen) supplemented with 1% 2-mercaptoethanol (Fisher Scientific) and equal quantities of 1.0mm and 0.5mm zirconium oxide beads (Next Advance). Biopsies were bead beat 3 times for 1 minute at a setting of 9 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 (50bp SR) on a HiSeq4000.

Lymphocyte isolation

Colonic epithelial cells and lymphocytes were isolated via mechanical disruption and enzymatic digestion. Briefly, colonic biopsies were twice shaken at 250rpm for 30min at 37C 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 40uM 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 37C for 30min in 15mL RPMI 1640 supplemented with 20% fetal bovine serum and 1mg/mL collagenase (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 15min on ice using LIVE/DEAD Fixable Aqua or LIVE/DEAD Fixable Near-IR (Thermo Fisher Scientific) diluted in PBS (Fisher Scientific), washed with PBS supplemented 2% FBS, and subsequently stained in an antibody mastermix for 25min at 4C. Cells were washed with PBS/2%FBS, resuspended into PBS/2%FBS, and subsequently run on a BD FACSAria Fusion Flow Cytometer to sort purify the populations of interest. Up to 10,000 CD4 T-cells (CD45+ EpCAMneg > LIVE/DEADneg > FSC vs SSC > singlets > CD3+ CD19neg > CD4+ CD8neg) from the laminal propria fraction, and 10,000 epithelial cells (EpCAM+ CD45neg > LIVE/DEADneg > FSC vs SSC > singlets > EpCAMhi CD44neg) were sorted directly into Starstedt tubes (Fisher Scientific) containing 350uL RLT Plus Buffer (Qiagen) supplemented with 1% 2-mercaptoethanol (Fisher Scientific). These samples were vortexed for 30sec and stored at -80C until RNA isolation. CD4 T-cells and plasma cells (CD45+ EpCAMneg > LIVE/DEADneg > FSC vs SSC > singlets > CD3neg > CD38+ CD27+) from the remaining lamina propria fraction were sorted into 600uL 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 were coated with polyclonal goat-anti human IgA, IgG, and IgM antibodies (KPL) at a concentration of 5ug/mL, diluted in PBS. Plates were left at 4C 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 37C 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 37C 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 1ug/mL at room temperature, in the dark, for 2 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 2 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, 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 500uL in a polystyrene flat-bottom 24-well plate (Thermo Fisher Scientific). Cells were incubated at 37C for 3 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 live dead and subsequently surface markers as above, after which cells were fixed and permeabilized in a 1:4 solution of Fixation/Permeabilization Concentrate and Fixation/Diluent (eBioscience) for 1 hour at 4C. 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 1 hour at room temperature. 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.

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 (provided by Saideep Gona)

All bulk RNAseq samples were processed using a standard workflow based on the GENPIPES framework[128]. Specifically, the "stringtie" type "rnaseq" pipeline was used. Reads were first trimmed using Trimmomatic software[129]. Trimmed reads were aligned to the GRCh38 human reference genome using the STAR aligner[130] 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[131].

Tissue differential expression and gene set enrichment analysis (provided by Raúl Aguirre-Gamboa)

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)[132]. 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)[133] package over the log2 fold changes in expression between sporadic, IBD, and PSC dysplasia. We then manually annotated the top 100 most significantly enriched ontologies in 5 categories (Table 2). These top 100 were then visualized in an enrichment map using the ggraph package (v2.0.4, (https://CRAN.R-project.org/package=ggraph)), reflecting the sizes of the gene sets (ontologies) and the sharedness of genes among them through the edges of the graph.

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.

Dimensionality reduction and clustering in non-dysplastic samples (provided by Raúl Aguirre-Gamboa)

In order to evaluate whether the immune PSC dysplastic signature was detectable in the colon biopsies of IBD or PSC patients with no history of dysplasia in an unbiased approach, we

decided to detect disease transcriptional profiles through an unsupervised clustering strategy. To do so, we selected only samples with no history of dysplasia (healthy n=48, IBD n =100, PSC n=59). The normalized (log2 CPMs) expression matrix was then corrected for batch effect, and 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)[134]. 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 40PCs 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")).

Prediction of cluster assignment in dysplastic samples (provided by Raúl Aguirre-Gamboa)

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. To select the penalisation score for eNet we used a 10x cross validation within the non dysplastic cohort using 60% of samples for training and 40% for testing.

Repertoire analysis of plasma cells (portions provided by Marcos Viera)

Binary base call output from sequencing were put through the Cellranger mkfastq pipline to generate fastq files, which were subsequently put through Cellranger vdj to generate full-length Ig sequences (https://support.10xgenomics.com/single-cell-vdj/software/pipelines/latest/using/vdj). 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[135]. Non-productive sequences, and sequences with the same cellular barcode were filtered from the analysis. Partis v0.15.0[136,137] 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[138,139]. 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.

Transcriptional analysis of CD4 T-cells (portions provided by Saideep Gona)

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[134]. 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[140]. 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 de- pendent 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 (Figure 9).

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 computeSum- Factors function from the "scran" R package[141]. 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 EdgeR, Limma-voom[142,143] 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 pipline 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). 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+ CD4 T-cells were queried for potential motifs using Sensitive, Thorough, Rapid, Enriched Motif Elicitation (STREME) web-based software[144], using IL17A and FOXP3 single positive CDR3s as a control. The proportion of cells containing the motif were then calculated.

Disease diagnosis to dysplasia outcome (provided by Raúl Aguirre-Gamboa)

To evaluate whether IBD or PSC samples that have ever acquired an I2 inflammatory profile in the gut, had an increased risk of developing right side dysplasia, we compared the probability of developing dysplasia from the data of colitis diagnosis (date of IBD diagnosis for PSC diagnosed samples). We first calculated the time from disease diagnosis to right side dysplasia for each individual patient with a history of right-sided dysplasia as events, or to the latest colonoscopy procedure in their medical records as the latest time point with a negative dysplasia diagnosis. Next, we stratified samples into two groups: I2 group and others. We

defined I2 as any samples for which an I2 inflammatory profile was ever detected in any of their visits. We then evaluated the difference in times to develop 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)). The same process was then repeated with non-right-sided dysplasia as the outcome.

16S sequencing (adapted from Barlow et al. Nature communications 2020)

A portion of DNA extracted from the whole tissue biopsies was dedicated to 16S sequencing. We used a digital polymerase chain reaction (dPCR)-based method to calculate absolute abundance measures of bacteria taxa as described previously[145]. Briefly, total concentration of 16S ribosomal RNA was quantified using universal primers to 16S and the QX200 droplet dPCR system (Bio-Rad). DNA then amplified and libraries generated using barcoded universal primers against the variable 4 region of 16S in triplicate, pooled, and quantified for subsequent sequencing. Samples were sequenced on the Illumina MiSeq platform using 300bp paired-end sequencing conditions.

Processing of all sequencing data was performed using QIIME 2 2019.1. Raw sequence data were demultiplexed and quality filtered using the q2-demux plugin followed by denoising with DADA2.

Whole Exome Sequencing

Reads were first trimmed using Trimmomatic software[129]. Trimmed reads were aligned to the GRCh38 human reference genome using the Burrow-Wheeler Aligner (BWA; https://github.com/lh3/bwa), and a panel of normal (PoN) was generated from all the reads from

all the samples. Variants were called using Mutect2[146], with the PoN and GnomAD (v3.1.1; https://gnomad.broadinstitute.org/) population germine information as additional inputs. Variants were filtered for those likely to be somatic and then functionally annotated with Funcotator (https://github.com/broadinstitute/gatk/).

RESULTS

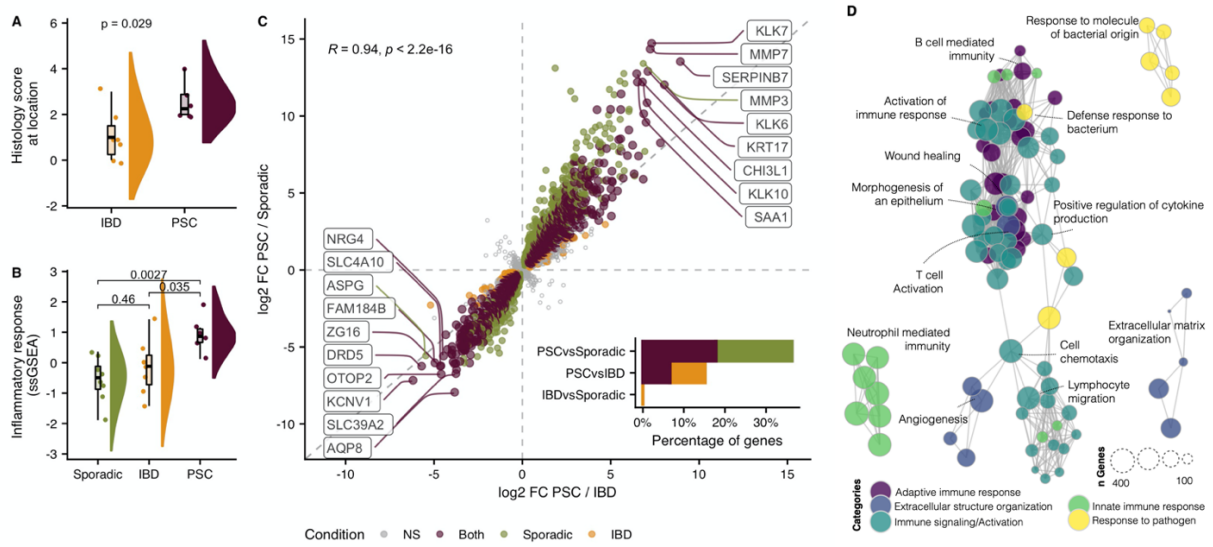
Introduction

We investigated whether there were biological factors present in PSC but absent in IBD that potentially underpin the differences in inflammation patterns and increased risk of CRC in PSC. To maximize our chances of finding relevant differences, we began with an unbiased transcriptional sequencing of colonic tissue from patients with PSC, as well as patients with IBD and healthy controls. There are many regional differences in factors such as bacterial load and composition[147], immune subsets[148], and epithelial cell function and identity[149] across the large intestine, so we restricted our analysis to the right colon as a way of controlling for these regional factors. We selected the right colon, as PSC-colitis is most common and active in the right colon, and dysplasia is most often right-sided. Though IBD patients do not always have right-sided inflammation, we only analyzed IBD patients with a documented history of right-sided colitis, to ensure that these IBD controls had the potential to present with right-sided colitis, even if they weren't actively inflamed at the time of sampling.

The environment of PSC dysplasia differs from that of IBD or sporadic dysplasia

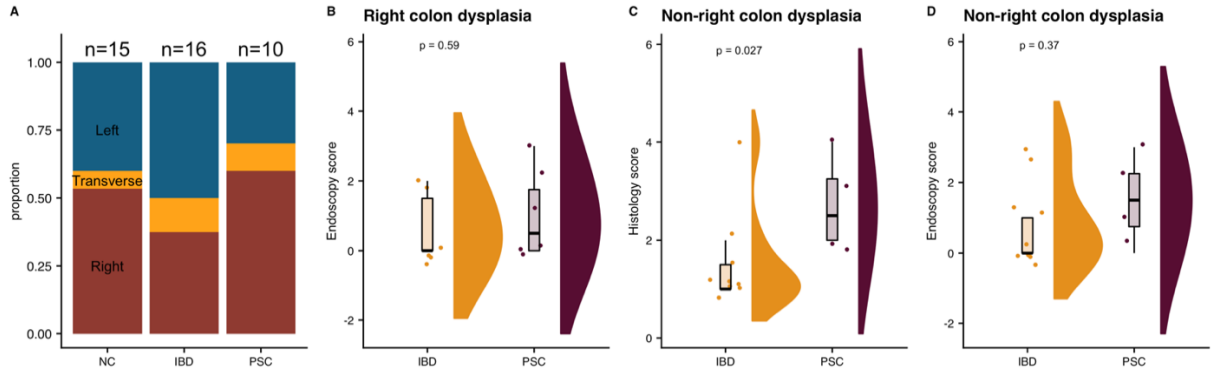
As we are particularly interested in understanding how the development of CRC in PSC differs from IBD, we began by analyzing patients with dysplasia at the time of sampling. Consistent with previously reported literature, we observed that PSC dysplasia was most often right-sided (hepatic flexure to ileo-cecal valve), whereas IBD dysplasia was more common distal to the hepatic flexure (transverse colon to rectum) (Figure 2A). As severity of inflammation is a risk factor for CRC, we quantified the severity of inflammation in the same segment of colon in which the dysplasia was identified. Amongst patients with right-sided dysplasia, we observed

Figure 1: The colonic dysplasia landscape of PSC differs from that of IBD.



a, Histologically scored inflammation at the site of dysplasia within the right colon. 0 = no diagnostic abnormality, 1 = quiescent/minimally active, 2 = mild, 3 = moderate, 4 = severe. Right colon consisted of the cecum, ascending colon, and hepatic flexure. **b,** single sample gene set enrichment analysis (ssGSEA) inflammatory response score calculated from the transcriptome of the right colon non-dysplastic tissue biopsy. **c,** Log 2 fold change (FC) of genes comparing PSC-dysplasia and IBD-dysplasia (x-axis) versus PSC-dysplasia and sporadic dysplasia (y-axis). Genes uniquely differentially expressed in PSC-dysplasia versus sporadic highlighted in green, genes uniquely differentially expressed in PSC-dysplasia versus IBD-dysplasia highlighted in yellow, and genes differentially expressed in both comparisons highlighted in purple. Inset bar graph quantifies the percentage of differentially expressed genes in each comparison (proportion of genes upregulated in PSC in purple, genes upregulated in sporadic dysplasia in green, and genes upregulated in IBD in yellow). **d,** Enrichment map of the 100 most significantly enriched gene sets upregulated in PSC-dysplasia versus IBD-dysplasia and sporadic dysplasia. Gene sets annotated by pathway and representative pathways per theme are labeled. Size of the circle represents size of the enriched gene set.

Figure 2: Histology score, but not endoscopy score, is greater in any area of dysplasia in PSC but not IBD.



a, Location of dysplasia as a proportion of total subjects with dysplasia at the time of collection. If a subject had multi-focal dysplasia, the most proximal dysplasia is noted. Right-sided dysplasia colored maroon, transverse dysplasia colored yellow, and left-sided dysplasia colored blue. **b,** Endoscopically scored inflammation at the site of dysplasia within the right colon. 0 = no diagnostic abnormality/quiescent, 1 = mild, 2 = moderate, 3 = severe. **c,** Histologically scored inflammation at the location of dysplasia, as done in Fig. 1a, for all dysplastic regions outside of the right colon (distal to hepatic flexure). **d,** Endoscopically scored inflammation at the location of dysplasia, as done in b, for all dysplastic regions outside of the right colon. (b-d) Significance determined by Wilcoxon test.

that PSC dysplasia always arose in the context of histologically active inflammation, whereas IBD dysplasia was most often found in areas of quiescent disease or areas without any discernable signs of IBD (Figure 1A). We did not observe any differences in endoscopically determined inflammation, however (Figure 2B).

We performed RNAseq on non-dysplastic tissue from the right colon of PSC, IBD, and non-IBD (sporadic) patients with dysplasia in the right colon. The clinical and demographic features of the patients included in this analysis are summarized in Table 1. We imputed the single sample inflammatory response gene set enrichment score (ssGSEA; HALLMARK_INFLAMMATORY_RESPONSE gene set from the Molecular Signature Database), for each sample. Consistent with what was observed histologically, we found that PSC dysplasia had a higher inflammatory response than IBD dysplasia, and that IBD dysplasia had the same levels of inflammation as non-colitis-associated sporadic dysplasia (Figure 1B). This suggests that PSC dysplasia is associated with active inflammation whereas IBD dysplasia, despite having severity and duration of inflammation as risk factors for development, has the same level of inflammation as a patient who developed dysplasia independently of IBD-associated colitis. In fact, there were almost no genes differentially expressed in IBD-dysplasia as compared to sporadic dysplasia (Figure 1C, inset bar graph). Though there were fewer genes differentially expressed in PSC dysplasia versus IBD dysplasia than in PSC dysplasia versus sporadic dysplasia, the direction and fold-change of gene expression in both comparisons were highly correlated (Figure 1C). The genes upregulated in PSC with respect to both comparisons were significantly enriched for many immune processes, as well as processes related to defense against pathogens and extracellular structural organization (Figure 1D, Table 2). Therefore, PSC dysplasia, but not IBD or sporadic dysplasia, is associated with an immune signature and

Table 1: Clinical and demographic information for patients in Figure 1.

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

[1] n (%); Median (IQR)

[2] Fisher's exact test; Kruskal-Wallis rank sum test

Table 2: Top 100 most significantly enriched gene sets in PSC dysplasia

ID	Description	Category	Set size	Normalized enrichment score	Adjusted p-value
GO:0002449	lymphocyte mediated immunity	Adaptive immune response	320	2.120084348	5.01E-09
GO:0051251	positive regulation of lymphocyte activation	Adaptive immune response	323	2.114437086	5.01E-09
GO:0051249	regulation of lymphocyte activation	Adaptive immune response	458	2.029728356	5.01E-09
GO:0030098	lymphocyte differentiation	Adaptive immune response	324	1.969046105	5.01E-09
GO:0046651	lymphocyte proliferation	Adaptive immune response	238	1.935121811	5.01E-09
GO:0002819	regulation of adaptive immune response	Adaptive immune response	156	2.004129521	5.40E-08
GO:0002920	regulation of humoral immune response	Adaptive immune response	107	2.280250319	5.01E-09
GO:0002455	humoral immune response mediated by circulating immunoglobulin	Adaptive immune response	132	2.210111491	5.01E-09
GO:0006959	humoral immune response	Adaptive immune response	262	2.281148622	5.01E-09
GO:0050853	B cell receptor signaling pathway	Adaptive immune response	119	2.148358573	5.01E-09
GO:0019724	B cell mediated immunity	Adaptive immune response	203	2.145308814	5.01E-09
GO:0016064	immunoglobulin mediated immune response	Adaptive immune response	200	2.144070478	5.01E-09
GO:0002460	adaptive immune response based on somatic recombination of immune receptors built from immunoglobulin superfamily domains	Adaptive immune response	332	2.143227007	5.01E-09
GO:0050864	regulation of B cell activation	Adaptive immune response	172	2.053178086	5.01E-09

Table 2 continued.

GO:0042113	B cell activation	Adaptive immune response	282	2.03362906	5.01E-09
GO:0050871	positive regulation of B cell activation	Adaptive immune response	130	2.055369145	3.07E-08
GO:0002377	immunoglobulin production	Adaptive immune response	177	1.956491267	3.48E-08
GO:0050870	positive regulation of T cell activation	Adaptive immune response	200	2.069570217	5.01E-09
GO:0050863	regulation of T cell activation	Adaptive immune response	301	1.949216443	5.01E-09
GO:0042110	T cell activation	Adaptive immune response	430	1.916694154	5.01E-09
GO:0030217	T cell differentiation	Adaptive immune response	233	1.886199039	3.31E-08
GO:0050867	positive regulation of cell activation	Cellular process	374	2.121873227	5.01E-09
GO:0006909	phagocytosis	Cellular process	341	2.006062163	5.01E-09
GO:0007204	positive regulation of cytosolic calcium ion concentration	Cellular process	254	1.908060426	5.01E-09
GO:0043410	positive regulation of MAPK cascade	Cellular process	466	1.71643973	5.01E-09
GO:0051480	regulation of cytosolic calcium ion concentration	Cellular process	278	1.842885218	2.32E-08
GO:0033108	mitochondrial respiratory chain complex assembly	Cellular process	101	-2.227186387	6.10E-08
GO:0006119	oxidative phosphorylation	Cellular process	138	-2.073903665	6.22E-08
GO:0032963	collagen metabolic process	Extracellular structure organization	90	2.303609806	5.01E-09
GO:0030198	extracellular matrix organization	Extracellular structure organization	344	2.200900499	5.01E-09
GO:0043062	extracellular structure organization	Extracellular structure organization	344	2.200900499	5.01E-09
GO:0045765	regulation of angiogenesis	Extracellular structure organization	277	1.962264808	5.01E-09
GO:0001525	angiogenesis	Extracellular structure organization	471	1.949945494	5.01E-09

Table 2 continued.

GO:1901342	regulation of vasculature development	Extracellular structure organization	308	1.919372122	5.01E-09
GO:0042060	wound healing	Extracellular structure organization	436	1.851246659	5.01E-09
GO:0030574	collagen catabolic process	Extracellular structure organization	39	2.222579091	1.91E-08
GO:0045766	positive regulation of angiogenesis	Extracellular structure organization	154	2.003135862	4.77E-08
GO:0022617	extracellular matrix disassembly	Extracellular structure organization	70	2.142816854	6.22E-08
GO:0002009	morphogenesis of an epithelium	Extracellular structure organization	480	1.655025341	1.14E-07
GO:0002696	positive regulation of leukocyte activation	Immune signaling/activation	362	2.105537828	5.01E-09
GO:0002697	regulation of immune effector process	Immune signaling/activation	412	2.063739083	5.01E-09
GO:0002703	regulation of leukocyte mediated immunity	Immune signaling/activation	194	1.970874615	5.01E-09
GO:0050727	regulation of inflammatory response	Immune signaling/activation	320	2.018820013	5.01E-09
GO:0002253	activation of immune response	Immune signaling/activation	497	2.038786394	5.01E-09
GO:0002699	positive regulation of immune effector process	Immune signaling/activation	200	1.96926	5.01E-09
GO:0050851	antigen receptor-mediated signaling pathway	Immune signaling/activation	296	1.987717781	5.01E-09
GO:0002429	immune response-activating cell surface receptor signaling pathway	Immune signaling/activation	431	1.996092285	5.01E-09
GO:0002757	immune response-activating signal transduction	Immune signaling/activation	431	1.996092285	5.01E-09

Table 2 continued.

GO:0070661	leukocyte proliferation	Immune signaling/activation	263	1.946128722	5.01E-09
GO:1902105	regulation of leukocyte differentiation	Immune signaling/activation	257	1.940558781	5.01E-09
GO:0002764	immune response-regulating signaling pathway	Immune signaling/activation	464	1.972859551	5.01E-09
GO:0002768	immune response-regulating cell surface receptor signaling pathway	Immune signaling/activation	461	1.971995537	5.01E-09
GO:0002440	production of molecular mediator of immune response	Immune signaling/activation	263	1.924734625	5.01E-09
GO:0002521	leukocyte differentiation	Immune signaling/activation	474	1.948067454	5.01E-09
GO:0002683	negative regulation of immune system process	Immune signaling/activation	394	1.781502053	5.01E-09
GO:1903706	regulation of hemopoiesis	Immune signaling/activation	437	1.711194557	3.62E-08
GO:0071621	granulocyte chemotaxis	Immune signaling/activation	114	2.251038966	5.01E-09
GO:0097530	granulocyte migration	Immune signaling/activation	133	2.224517285	5.01E-09
GO:0030595	leukocyte chemotaxis	Immune signaling/activation	202	2.173401676	5.01E-09
GO:0097529	myeloid leukocyte migration	Immune signaling/activation	193	2.143076625	5.01E-09
GO:1903039	positive regulation of leukocyte cell-cell adhesion	Immune signaling/activation	219	2.117996513	5.01E-09
GO:0060326	cell chemotaxis	Immune signaling/activation	263	2.087082061	5.01E-09
GO:0022409	positive regulation of cell-cell adhesion	Immune signaling/activation	256	2.067205914	5.01E-09
GO:0050900	leukocyte migration	Immune signaling/activation	457	2.109907627	5.01E-09
GO:0007159	leukocyte cell-cell adhesion	Immune signaling/activation	325	2.026583642	5.01E-09
GO:1903037	regulation of leukocyte cell-cell adhesion	Immune signaling/activation	292	1.973915207	5.01E-09

Table 2 continued.

GO:0045785	positive regulation of cell adhesion	Immune signaling/activation	395	1.988302228	5.01E-09
GO:0030335	positive regulation of cell migration	Immune signaling/activation	485	1.941476303	5.01E-09
GO:0022407	regulation of cell-cell adhesion	Immune signaling/activation	386	1.868366801	5.01E-09
GO:0002685	regulation of leukocyte migration	Immune signaling/activation	192	1.933847399	1.53E-08
GO:0002548	monocyte chemotaxis	Immune signaling/activation	57	2.17374667	7.38E-08
GO:0072676	lymphocyte migration	Immune signaling/activation	106	2.089692001	7.63E-08
GO:0002687	positive regulation of leukocyte migration	Immune signaling/activation	131	2.017863731	1.17E-07
GO:0070098	chemokine-mediated signaling pathway	Immune signaling/activation	80	2.32712843	5.01E-09
GO:1990868	response to chemokine	Immune signaling/activation	89	2.216763381	5.01E-09
GO:1990869	cellular response to chemokine	Immune signaling/activation	89	2.216763381	5.01E-09
GO:0001819	positive regulation of cytokine production	Immune signaling/activation	389	2.039303919	5.01E-09
GO:0034341	response to interferon-gamma	Immune signaling/activation	181	1.973677727	9.52E-09
GO:0034612	response to tumor necrosis factor	Immune signaling/activation	285	1.820238949	4.69E-08
GO:0030449	regulation of complement activation	Innate immune response	92	2.182792365	5.01E-09
GO:0006956	complement activation	Innate immune response	141	2.281783738	5.01E-09
GO:0006958	complement activation, classical pathway	Innate immune response	118	2.212561611	5.01E-09
GO:0032943	mononuclear cell proliferation	Innate immune response	240	1.936133271	5.01E-09
GO:0002275	myeloid cell activation involved in immune response	Innate immune response	499	1.785196562	5.01E-09
GO:0043299	leukocyte degranulation	Innate immune response	488	1.781983183	5.01E-09

Table 2 continued.

GO:0036230	granulocyte activation	Innate immune response	464	1.766070807	5.01E-09
GO:0030593	neutrophil chemotaxis	Innate immune response	94	2.284151499	5.01E-09
GO:1990266	neutrophil migration	Innate immune response	110	2.2732604	5.01E-09
GO:0042119	neutrophil activation	Innate immune response	457	1.779640769	5.01E-09
GO:0002283	neutrophil activation involved in immune response	Innate immune response	446	1.767850129	5.01E-09
GO:0002446	neutrophil mediated immunity	Innate immune response	455	1.763015774	5.01E-09
GO:0043312	neutrophil degranulation	Innate immune response	443	1.761789789	5.01E-09
GO:0071222	cellular response to lipopolysaccharide	Response to pathogen	171	2.124179959	5.01E-09
GO:0071219	cellular response to molecule of bacterial origin	Response to pathogen	179	2.105390059	5.01E-09
GO:0032496	response to lipopolysaccharide	Response to pathogen	279	2.084222476	5.01E-09
GO:0002237	response to molecule of bacterial origin	Response to pathogen	293	2.075103827	5.01E-09
GO:0071216	cellular response to biotic stimulus	Response to pathogen	203	2.037928974	5.01E-09
GO:0042742	defense response to bacterium	Response to pathogen	216	1.924042202	5.01E-09
GO:0031349	positive regulation of defense response	Response to pathogen	323	1.923376686	5.01E-09
GO:0032103	positive regulation of response to external stimulus	Response to pathogen	446	1.918197595	5.01E-09

inflammation, and IBD dysplasia indistinguishable from sporadic dysplasia in these regards. Importantly, the increase in histologically scored inflammation in PSC was observed for dysplasia distal to the hepatic flexure (Figure 2C), suggesting these results are applicable to dysplasia occurring anywhere within the colon.

PSC colitis is unique and distinct from IBD colitis

The immune signature we observe associated with PSC dysplasia could either be a response to the dysplasia or could precede the dysplasia. If the former, the inflammation does not likely have a causative link in the development of dysplasia. If this inflammation precedes dysplasia, it is more likely to be associated with the development of the dysplasia. To determine whether the PSC dysplasia inflammatory signature could be found in advance of the dysplasia, we analyzed the right colon tissue of healthy control, IBD, and PSC subjects with no history of dysplasia and no diagnosis of dysplasia at the time of sampling. Clinical and demographic information for the patient included in this analysis are summarized in Table 3. Overall, we observed that histologically scored and endoscopically scored inflammation was higher in PSC patients than in IBD (Figure 3A and 4A). However, there was no significant difference in transcriptionally scored inflammation, and both IBD and PSC patients were more inflamed than healthy controls (Figure 3B). This highlights a disparity in how severity of inflammation is quantified histologically and transcriptionally. There are many methods by which pathologists can evaluate inflammation, each with its own criteria, and no consistent methods is used across all hospitals[150]. At UCM, pathologists use a four-tiered grading system, based on the presence of neutrophils in different compartments of the colonic tissue to determine the severity of inflammation (Table 4). Our analysis of inflammation in PSC dysplasia revealed an increase in

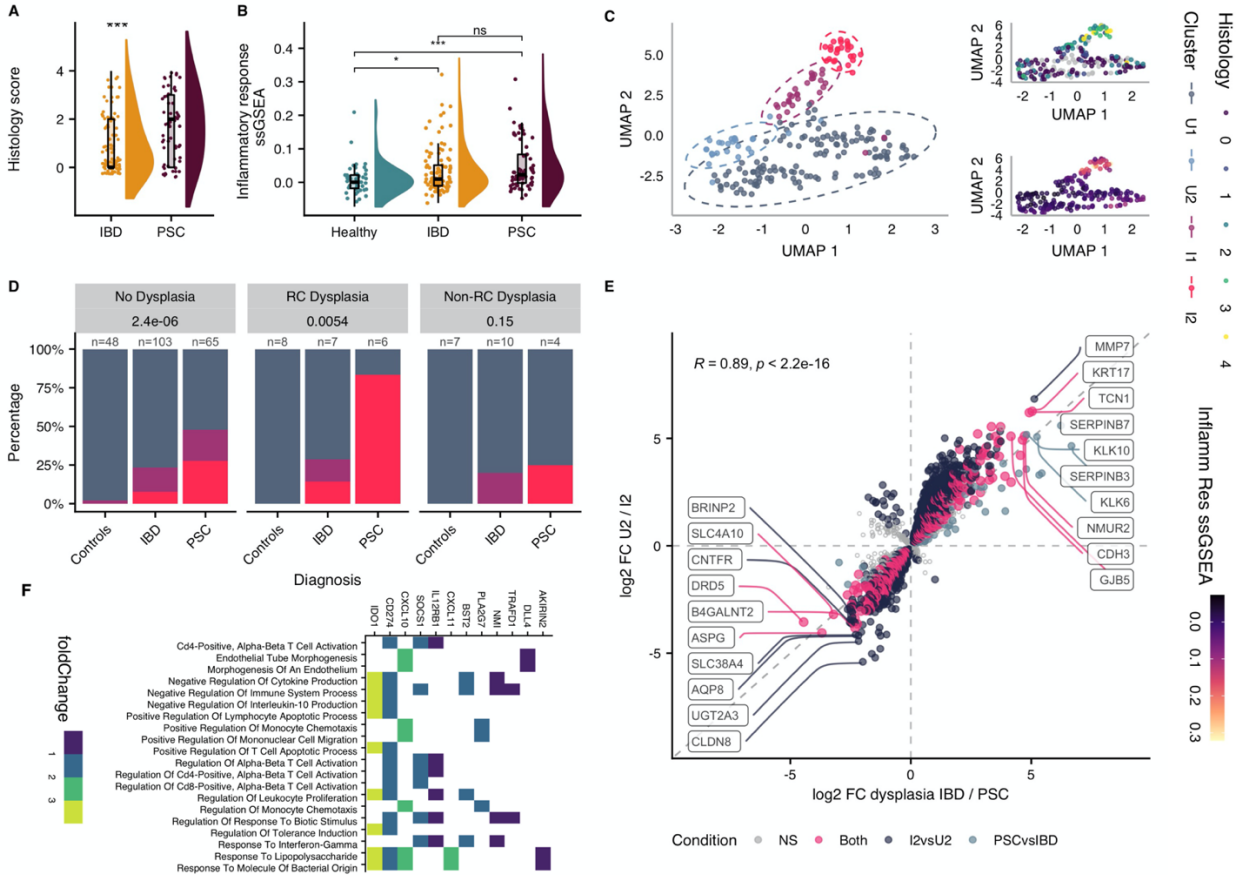
Table 3: Clinical and demographic information for patients in Figure 3.

Variable	NC, N = 46[1]	IBD, N = 105[1]	PSC, N = 65[1]	p-value[2]
Demographic and Clinical Information				
Sex				0.056
Female	26 (57%)	40 (38%)	23 (35%)	
Male	20 (43%)	65 (62%)	42 (65%)	
Race				0.070
Asian	2 (4.3%)	5 (4.8%)	5 (7.7%)	
Black	15 (33%)	16 (15%)	16 (25%)	
White	28 (61%)	84 (80%)	44 (68%)	
Unknown	1 (2.2%)	0 (0%)	0 (0%)	
Ethnicity				0.7
Hispanic/Latino	0 (0%)	2 (1.9%)	2 (3.1%)	
Not Hispanic/Latino	46 (100%)	103 (98%)	63 (97%)	
Age at procedure	52 (50, 55)	37 (28, 49)	34 (25, 46)	<0.001
Type of IBD				
No IBD	46 (100%)	0 (0%)	5 (7.7%)	
CD	0 (0%)	71 (68%)	16 (25%)	
UC	0 (0%)	34 (32%)	43 (66%)	
IC	0 (0%)	0 (0%)	1 (1.5%)	
Age at IBD diagnosis	NA (NA, NA)	22 (17, 32)	22 (17, 30)	>0.9
Age at PSC diagnosis	NA (NA, NA)	NA (NA, NA)	27 (19, 35)	
Duration of IBD at procedure	NA (NA, NA)	12 (7, 18)	8 (4, 14)	0.012
Duration of PSC at procedure	NA (NA, NA)	NA (NA, NA)	6 (2, 11)	
History of liver transplant	0 (0%)	0 (0%)	11 (17%)	<0.001
Medications				
5-aminosalycilic acid	0 (0%)	35 (33%)	30 (46%)	<0.001
anti_IL12/23 mAb	0 (0%)	2 (1.9%)	1 (1.5%)	>0.9
anti-intergrin mAb	0 (0%)	10 (9.5%)	7 (11%)	0.046
anti-TNFa mAb	0 (0%)	37 (35%)	10 (15%)	<0.001
Methotrexate	0 (0%)	6 (5.7%)	1 (1.5%)	0.2
Ursodiol	0 (0%)	0 (0%)	27 (42%)	<0.001
JAK inhibitory	0 (0%)	3 (2.9%)	2 (3.1%)	0.7
Purine synthesis inhibitor	0 (0%)	38 (36%)	12 (18%)	<0.001
Steroids	0 (0%)	13 (12%)	9 (14%)	0.013

[1] n (%); Median (IQR)

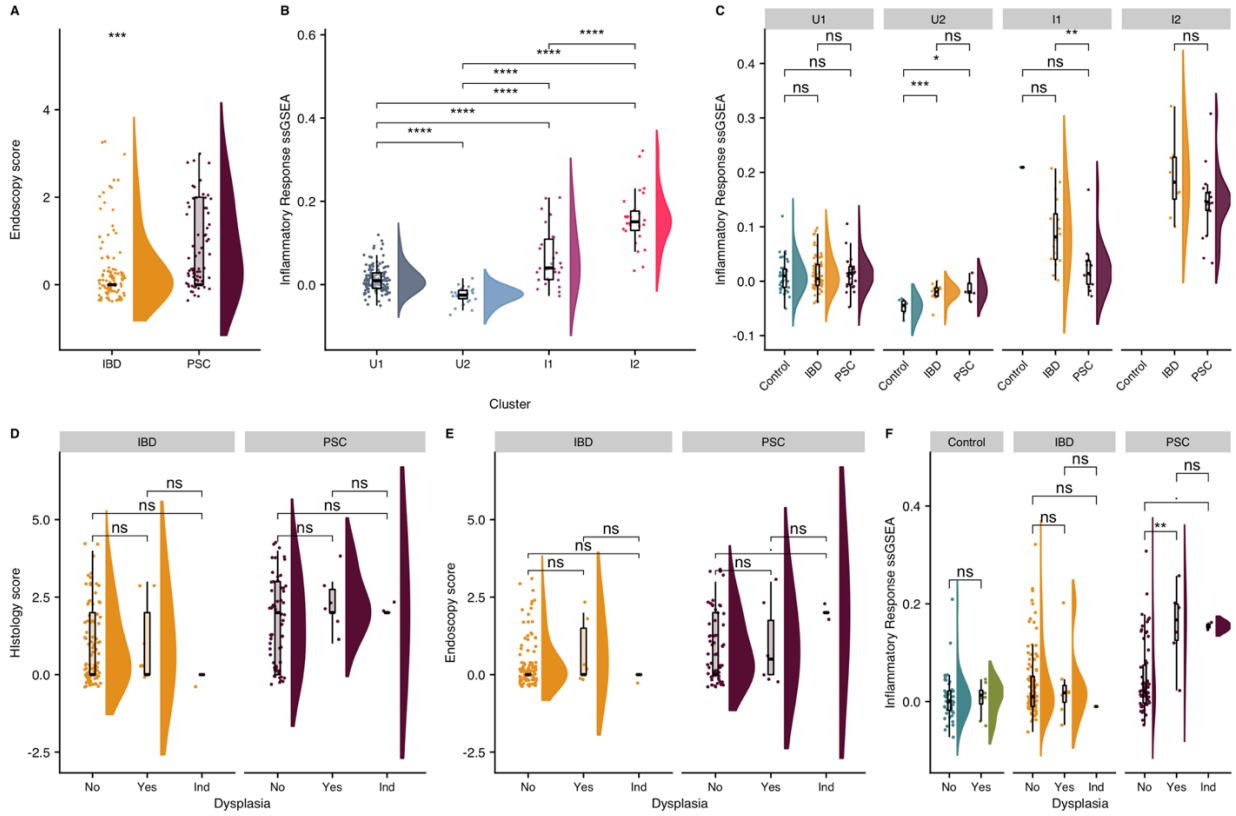
[2] Pearson's Chi-squared test; Fisher's exact test; Kruskal-Wallis rank sum test

Figure 3: A subset of PSC patients without dysplasia share a similar transcriptional profile to PSC patients with dysplasia.



a, Histologically scored inflammation in the right colon of patients with no history of dysplasia. Scoring performed as in Fig. 1a. **b,** Inflammatory response ssGSEA calculated as in Fig. 1b. in the right colon of patients with no history of dysplasia. **c,** 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 (left), histologically scored inflammation (top right), or inflammatory ssGSEA (bottom right). **d,** Distribution of subjects across clusters amongst subject without dysplasia (left), with right-sided dysplasia (middle), and with non-right sided dysplasia (right). Statistical significance determined by Chi-squared test. **e,** Log 2 fold change (FC) of genes comparing I2 subjects and U subjects (y-axis) versus PSC-dysplasia versus IBD/sporadic dysplasia (x-axis). Genes uniquely differentially expressed in I2 vs U highlighted in dark blue, genes uniquely differentially expressed in PSC-dysplasia versus IBD/sporadic dysplasia highlighted in light blue, and genes differentially expressed in both comparisons highlighted in pink. **f,** 20 most significantly upregulated gene sets in PSC-I2 versus IBD-I2, ordered by fold-change. Most significantly associated genes with each pathway listed on x-axis.

Figure 4: Inflammation across diseases, dysplasia, and clusters.



a, Endoscopically scored right colon inflammation of patients without a history of dysplasia, as done in Extended Data Fig. 1b. **b,** Inflammatory response ssGSEA score across clusters. **c,** Inflammatory response ssGSEA between diagnoses, within clusters. **d,** Histologically scored right colon inflammation of patients with or without right colon dysplasia at the time of sampling. **e,** Endoscopically scored right colon inflammation of patients with or without right colon dysplasia at the time of sampling. **f,** Inflammatory response ssGSEA of patients with or without right colon dysplasia at the time of sampling. (a-f) Significance determined by Wilcoxon test (“.” for p<0.1, “*” for p<0.05, “**” for p<0.01, “***” for p<0.001, “****” for p<0.0001, , “ns” for not significant (p>0.05)). (d-f) Ind = Indefinite for dysplasia.

Table 4: Histologic criteria for grading of disease activity at UCM (provided by Dr. Christopher Weber)

Activity Grading	Histologic Criteria
Quiescent	Features of chronicity (crypt distortion/shortening/drop-out, basal plasmacytosis, pyloric or Paneth cell metaplasia) in the absence of mild/moderate/severe activity
Mild	Neutrophils present in epithelium
Moderate	Neutrophils present in crypt lumen forming crypt abscess
Severe	Erosion or ulceration of epithelium

several adaptive and innate immune pathways, and not just neutrophils (Figure 1D). To remain as unbiased as possible and to account for as many factors of inflammation as possible, we will give more weight to transcriptionally determined inflammation. Therefore, though PSC patients were histologically and endoscopically more inflamed than the IBD cohort, there was no difference in the transcriptionally determined inflammation, and are therefore comparable in distribution of inflammation.

Using the 3,000 most hypervariable genes across diagnoses, we were able to identify four transcriptionally distinct cluster of subjects (Figure 3C, left). Two of these clusters, uninflamed 1 and 2 (U1 and U2) were histologically and transcriptionally uninflamed (Figure 3C, right and 4B) and were therefore combined in subsequent analyses (collectively referred to as U). Two smaller clusters of inflamed subjects were identified and labeled inflamed 1 and 2 (I1 and I2), with I2 being more inflamed than I1 (Figure 4B). The distribution of diagnoses was not equivalent across transcriptional clusters (Figure 3D, left). Nearly all healthy controls fell in the U cluster, whereas there was a significant enrichment of IBD patients, and to an even greater extent PSC patients in I1 and I2. Nearly 25% of IBD patients were I1 or I2, and nearly 50% of PSC patients were I1 or I2. Across all clusters, PSC patients were either equally inflamed or less inflamed than their IBD counterparts (Figure 4C), suggesting that comparison of disease groups within clusters is fair, and that any inflammatory feature unique to PSC within a transcriptional cluster is not due to greater inflammation.

Using core gene for each cluster, we predicted the transcriptional cluster identity of PSC, IBD, and control patients with dysplasia at the time of sampling. We observed that nearly all cases of right-sided PSC dysplasia were assigned I2, whereas the majority of right-sided IBD and sporadic dysplasia were assigned U (Figure 3D, center). This observation is consistent with the

fact that PSC dysplasia is inflamed while IBD and sporadic dysplasia are uninflamed. PSC-dypsplasia patients were also more inflamed than the non-dysplastic PSC population (Figure 4F), confirming that inflammation is not a general feature of all PSC patients, and that PSC dysplasia is highly associated with strong inflammation. Notably, the majority of non-right-sided sporadic, IBD, and PSC dysplasia were assigned cluster U (Figure 3D, right). This suggests that inflammation is important specifically within the region that dysplasia develops, and that inflammation in one segment of the colon is irrelevant to the development of dysplasia in other segments.

Given that nearly all PSC dysplasia subjects were categorized as I2 and nearly all IBD dysplasia as U, we wanted to directly compare the overlap between these two sets. Plotting the fold-change and direction of gene expression in PSC dysplasia and IBD dysplasia versus the fold-change and directions of the gene expression in I2 versus U revealed an extremely strong correlation between the two signatures (Figure 3E). The I2 gene signature is likely the same signature as the PSC-dysplasia signature, meaning that about 25% of PSC patients without a history of dysplasia share the same signature as a dysplastic PSC patient. The I2 signature could therefore be implicated in the development of PSC dysplasia, or at the very least marks PSC patients at risk for dysplasia.

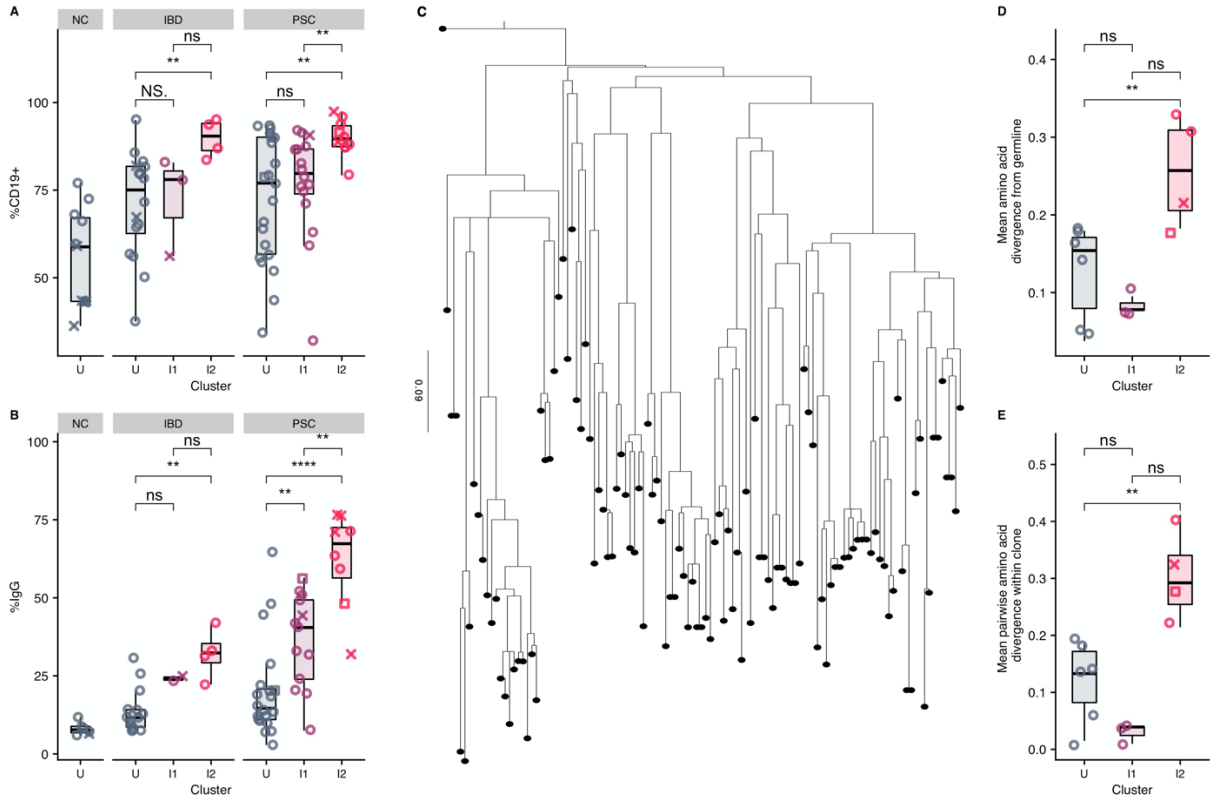
Since a subset of IBD patients were also classified as I2, we wanted to investigate whether there were any features unique to PSC I2 as compared to IBD I2. Since PSC dysplasia specifically is preferentially enriched in I2, the features unique to PSC I2 as compared to IBD I2 are likely the most relevant to the development of PSC dysplasia. Surprisingly, though PSC I2 and IBD I2 subjects are equally inflamed transcriptionally, we observed several immune pathways to be specifically enriched in PSC I2 (Figure 3F). We noted that several pathways

related to T-cell activation and response to bacterial molecules were enriched. Interestingly, some of the genes most correlated with the enriched pathways were previously identified as associated with PSC by GWAS (IDO1, SOCS1)[39].

PSC inflammation is characterized by antigen driven IgG plasma cells

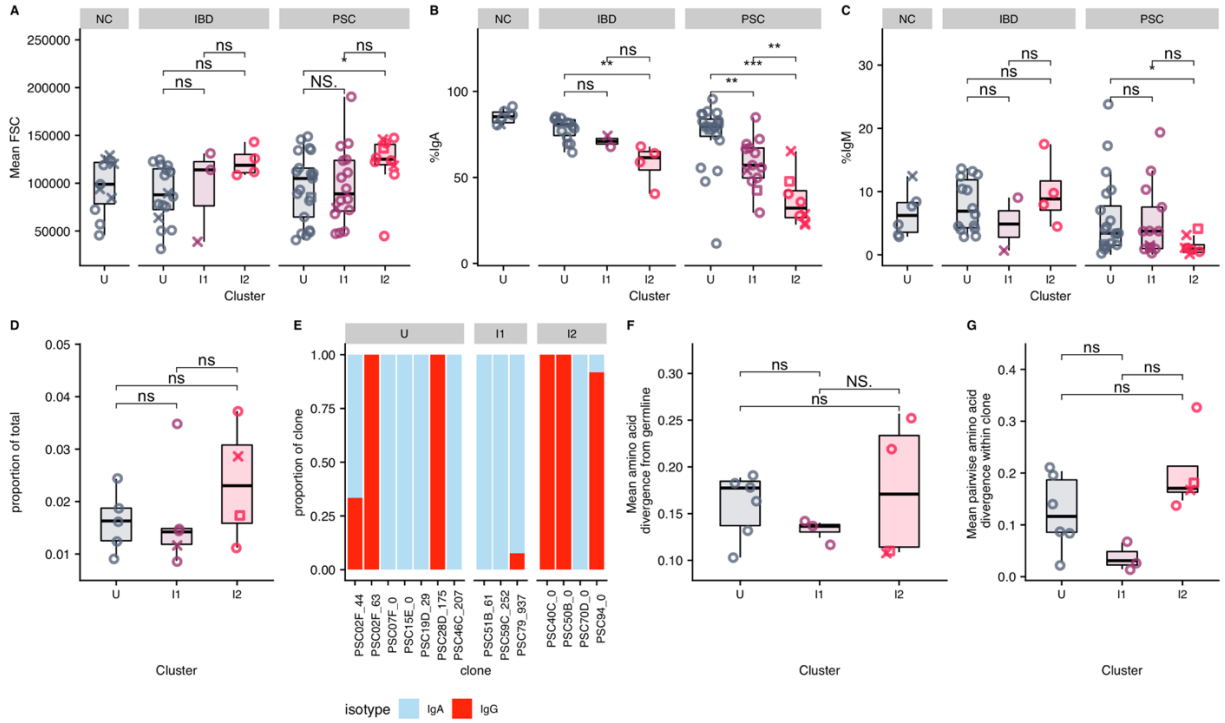
We wanted to further characterize the nature of I2 inflammation and identify additional differences between PSC I2 and IBD I2. Given the strong association of HLA in PSC and the T-cell signature observed uniquely in PSC I2, we paid special attention to differences in the adaptive immune compartment. We extracted plasma cells from the right colon of patients across clusters to determine if there was any visible B-cell phenotype. We found that plasma cells from I2 PSC and IBD patients were nearly 100% surface CD19+ (Figure 5A), suggesting that the plasma cells observed in these patients are recently arrived, active antibody secreting cells[151]. These plasma cells in PSC I2, and likely also IBD I2, were larger (Figure 6A) than plasma cells in U subjects as well. In a healthy colon, most plasma cells are of the IgA isotype[152], though the proportion of plasma cells secreting pro-inflammatory IgG increases in active colitis[153]. We observed an ordinal increase across clusters of the proportion of plasma cells secreting IgG in both IBD and PSC (Figure 5B). However, the proportion of plasma cells secreting IgG in PSC I2 was greater than in IBD I2 ($p = 0.016$). In our dataset, and consistent with previous reports, the percent of plasma cells secreting IgG in IBD I2 patients never exceeded 50%, whereas the majority of PSC I2 subjects had more than 50% of their plasma cells secreting IgG. Upwards of 75% of all plasma cells in some PSC I2 subjects were of the IgG isotype. A corresponding decrease in IgA- and IgM-secreting plasma cells is observed ordinally across clusters (Figure

Figure 5: PSC inflammation is characterized by an influx of IgG plasma cells and plasma cells show signs consistent with antigen drive.



a, Frequency of right colon plasma cells positive for surface CD19 by flow cytometry. **b,** Frequency of IgG-secreting plasma cells amongst total right colon plasma cells as determined by ELISpot. **c,** Representative dendrogram of heavy chain sequences within top clone of I2 patient. 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. **d,** Mean amino acid divergence from inferred germline within CDR3 of largest clones identified in each patient. **e,** Mean pairwise amino acid divergence within CDR3 of largest clones identified in each patient. (a, b, d, e) 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" denote patients with dysplasia at the time of sampling, open squares denote patients indefinite for dysplasia at the time of sampling).

Figure 6: 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,** Frequency of IgA-secreting plasma cells amongst total right colon plasma cells as determined by ELISpot. **c,** Frequency of IgA-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 $p<0.05$, "**" for $p<0.01$, "***" for $p<0.001$, , "ns" for not significant ($p>0.05$)).

6B,C). PSC colitis is therefore uniquely characterized by an influx of IgG-dominant plasma cells, which is not seen to the same degree in IBD colitis, even when equally inflamed.

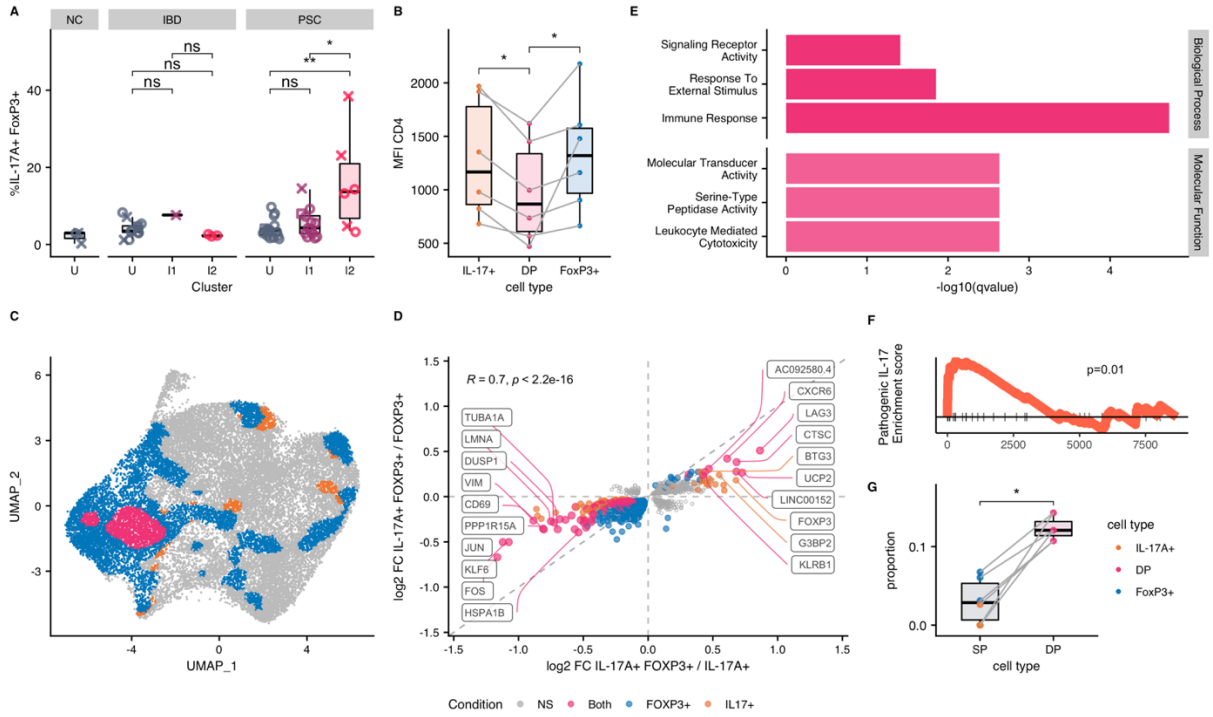
As we had observed a unique plasma cell phenotype in PSC, we want to see if these plasma cells showed signs of antigen drive, given the strong association of PSC with HLA. We performed single cell sequencing of total plasma cells derived from PSC patients across clusters and determined clonal pools of cells based on similarities across heavy chain sequences. Given that PSC inflammation is characterized by an active turnover of plasma cells, we focused our analysis on the largest clone in each subject, under the assumption that the largest clone is the most likely to be a response to antigen drive. We observed that the three largest clones in our dataset were found in inflamed PSC patients (Figure 6D), and that the clones were predominantly IgG in I2 subjects, and IgA in U and I1 (Figure 6E). A representative dendrogram of the sequences within the largest clone of an I2 PSC patient demonstrates lop-sided branching patterns characteristic of non-random selection of mutations (Figure 5C). We observed, that within complementarity-determining region 3 (CDR3) of the largest clones, that there was a greater mean amino acid divergence from inferred germline in I2 subjects as compared to U (Figure 5D). Additionally, the CDR3 of the I2 top clones were more diverse as computed by the mean pairwise amino acid divergence within the clone (Figure 5E). These differences were not present when analyzing the entire length of the heavy chain (Figure 6F,G), meaning that the CDR3 specifically was more heavily mutated and diverse in the top clones of I2 than the top clones of U. Collectively, these are all signs that the IgG plasma cells in PSC I2 are expanded in response to antigen drive.

PSC inflammation is characterized by antigen driven IL-17+ Foxp3+ CD4 T-cells

We next proceeded to search for a corresponding subset of T-cell unique to PSC I2 patients. We did not find any difference in the proportion of lamina propria CD4 T-cells positive for either interferon gamma (IFNγ), IL-17, TNFα, or Foxp3 across clusters within PSC (Figure 8A,B,C,D). Previous work demonstrated an increase in pathogenic IL-17A+ Foxp3+ CD4 T-cells in the cancerous lesions of IBD patients[154], so we search specifically for the presence of these cells as well. To our surprise, we found an increase in the percentage of CD4 T-cells expressing IL-17A and Foxp3 in PSC I2 as compared to PSC U (Figure 7A) or IBD I2 ($p = 0.024$). As controls, we did not see an increase of either IFNγ+ Foxp3+ or TNFα+ Foxp3+ cells in PSC I2 as compared to IBD I2 (Figure 8E,F). There was also no increase in singularly positive IL-17A or Foxp3 CD4 T-cells in PSC I2 (Figure 8G,H). Given the previous association with the development of colitis-associated CRC, and that these cells were uniquely present in the subset of PSC enriched for dysplasia, we turned our focus to the IL-17A+ Foxp3+ CD4 T-cells. We found that these double positive (DP) cells had lower surface expression of CD4 than their IL-17A or Foxp3 single positive (SP) counterparts (Figure 7B), suggesting that the DP cells were more activated or chronically stimulated[155].

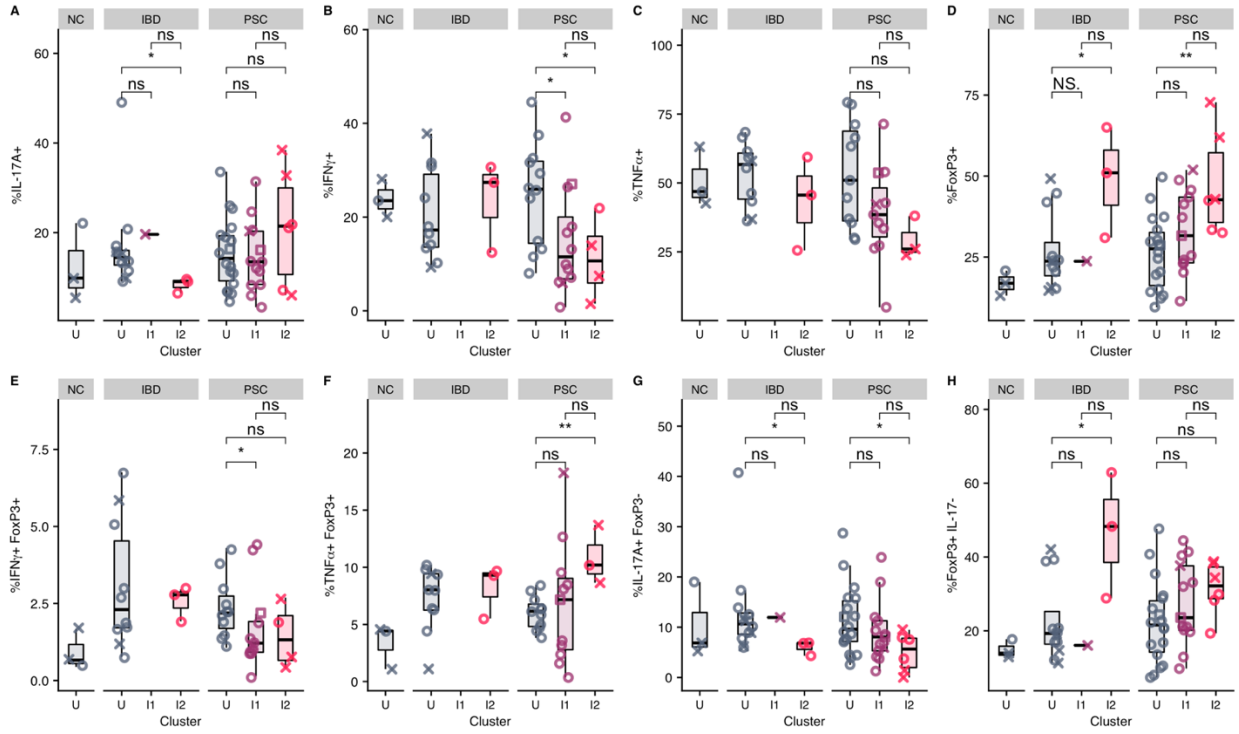
Our analysis of these cells to this point was restricted to flow cytometry after stimulation and fixation. To better describe these cells *ex vivo* we performed single cell RNAseq on freshly isolated CD4 T-cells from PSC patients. By calibrating the threshold of transcriptional detection of cells co-expressing IL17A and FOXP3 transcript using our flow cytometry data (Figure 9A-D), we successfully identified both DP and SP cells by single cell RNAseq (Figure 7C). We compared the transcriptional signature of DP cells to IL17A and FOXP3 SP cells and found a strong correlation between the two (Figure 9C), meaning that DP cells are distinct from either IL17A or FOXP3 SP cells. Using the genes increased in DP cells as compared to both IL17A or

Figure 7: PSC inflammation is characterized by IL-17A+ Foxp3+ CD4 T-cells enriched for TCRs containing "LA."



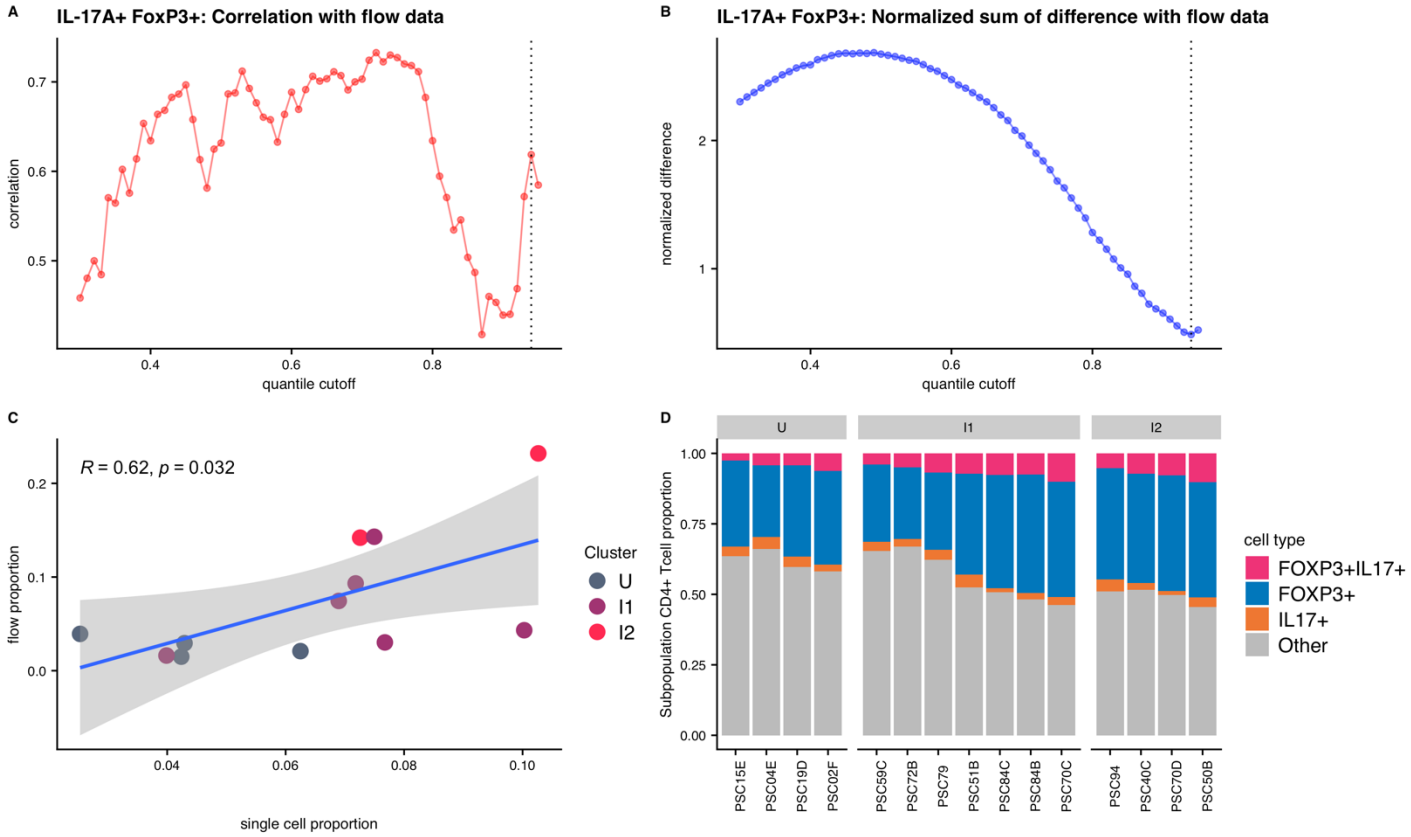
a, Percent of right colon lamina propria CD4 T-cells positive for IL-17A and Foxp3 by flow cytometry after 3 hour stimulation with phorbol myristate acetate and ionomycin. **b,** Mean fluorescence intensity (MFI) of surface CD4 of cells from I2 PSC patients. Significance determined by Wilcoxon matched-pairs signed rank test. **c,** UMAP of single-cell sequenced CD4 T-cells from PSC subjects, annotated by transcriptionally-determined cell type. **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. **f,** Most significantly enriched gene sets using genes upregulated in IL-17A+ Foxp3+ CD4 T-cells versus either IL-17A+ or Foxp3+ CD4 T-cells. **g,** Enrichment for a pathogenic IL-17 signature using genes differentially expressed in IL-17A+ Foxp3+ CD4 versus IL-17A+ CD4 cells. GSEA p-value is shown. **g,** Proportion of cells containing amino acid motif "LA" in the TCR beta chain by cell type amongst I2 PSC patients. (b,d) 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. ("*" for $p<0.05$, , "ns" for not significant ($p>0.05$)).

Figure 8: Cytokines secreted by CD4 T-cells across transcriptional clusters



Frequency of right colon lamina propria CD4 T-cells expressing IL-17A (**a**), IFNγ (**b**), TNFα (**c**), or Foxp3 (**d**). Frequency of right colon lamina propria CD4 T-cells co-expressing IFNγ and Foxp3 (**e**) or TNFα and Foxp3 (**f**). Frequency of right colon lamina propria CD4 T-cells that are IL-17+ Foxp3- (**g**) or Foxp3+ IL-17- (**h**). (a-h) Cells were assessed for cytokine and transcription factors by flow cytometry after 3 hour stimulation with phorbol myristate acetate and ionomycin. Significance determined by Wilcoxon test ("*" for $p<0.05$, "**" for $p<0.01$, "ns" for not significant ($p>0.05$)).

Figure 9: 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. 4 (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.

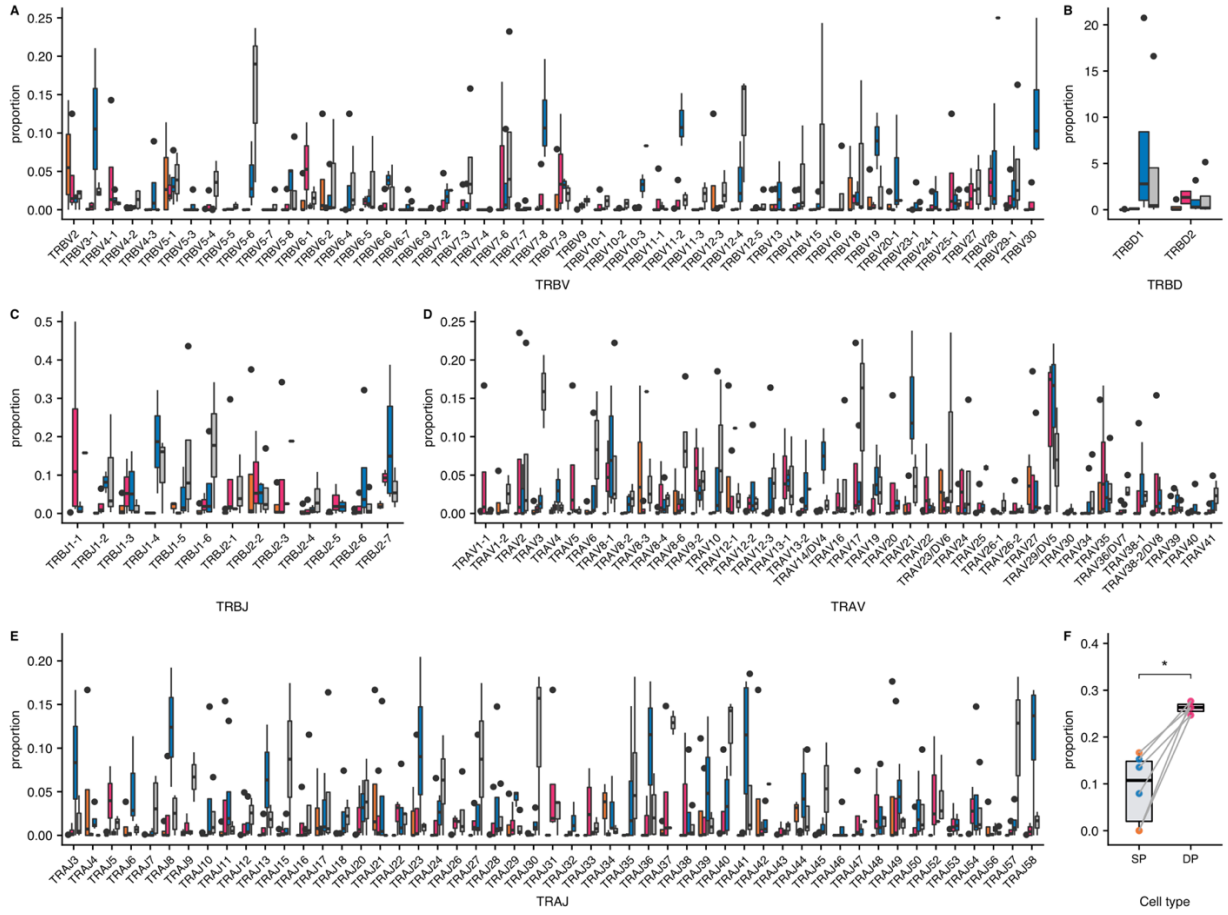
FOXP3 SP cells, we found an enrichment for activation modules (Figure 7E), further emphasizing that DP cells are highly activated. We also found that DP cells were enriched for a pathogenic Th17 signature[156], as compared to IL17A SP cells suggesting that DP cells might be contributing to the inflammation (Figure 7F).

We wanted to determine whether the DP cells showed signs of antigen drive much like the IgG plasma cells. We performed TCR repertoire analysis of the CD4 T-cells from I2 PSC patients, stratifying the cells by identity as DP or IL17A/FOXP3 SP cells. We did not find any preferential V, D, or J gene usage in either the TCRβ or TCRα chains (Figure 10A-E). We proceed to analyze the non-germline encoded region of the TCRβ chain of each cell for potential amino acid motif enrichments. Using the STREME web-based software, we identified a candidate amino acid motif "LA." We found this "LA" amino acid motif in a greater proportion of DP cells than in SP cells (Figure 7G). LA is a germline encoded motif within one of the open reading frames (ORF) of TRBD2. However, even when analyzing only the cells using TRBD2 we found an enrichment of the "LA" motif in DP cells (Figure 10F), suggesting a preferential selection for this ORF amongst DP cells. A summary table of the V, D, and J genes used by "LA" containing cells is listed in Tables 5 and 6. We analyzed the V, D, and J usage amongst cells containing the "LA" motif and found that the Vα gene usage of DP cells containing the "LA" motif were distinct from DP cells without the LA motif (Figure 11C), further suggesting that these DP "LA"-containing cells are distinct within the TCR repertoire.

PSC inflammation is associated with greater risk for dysplasia

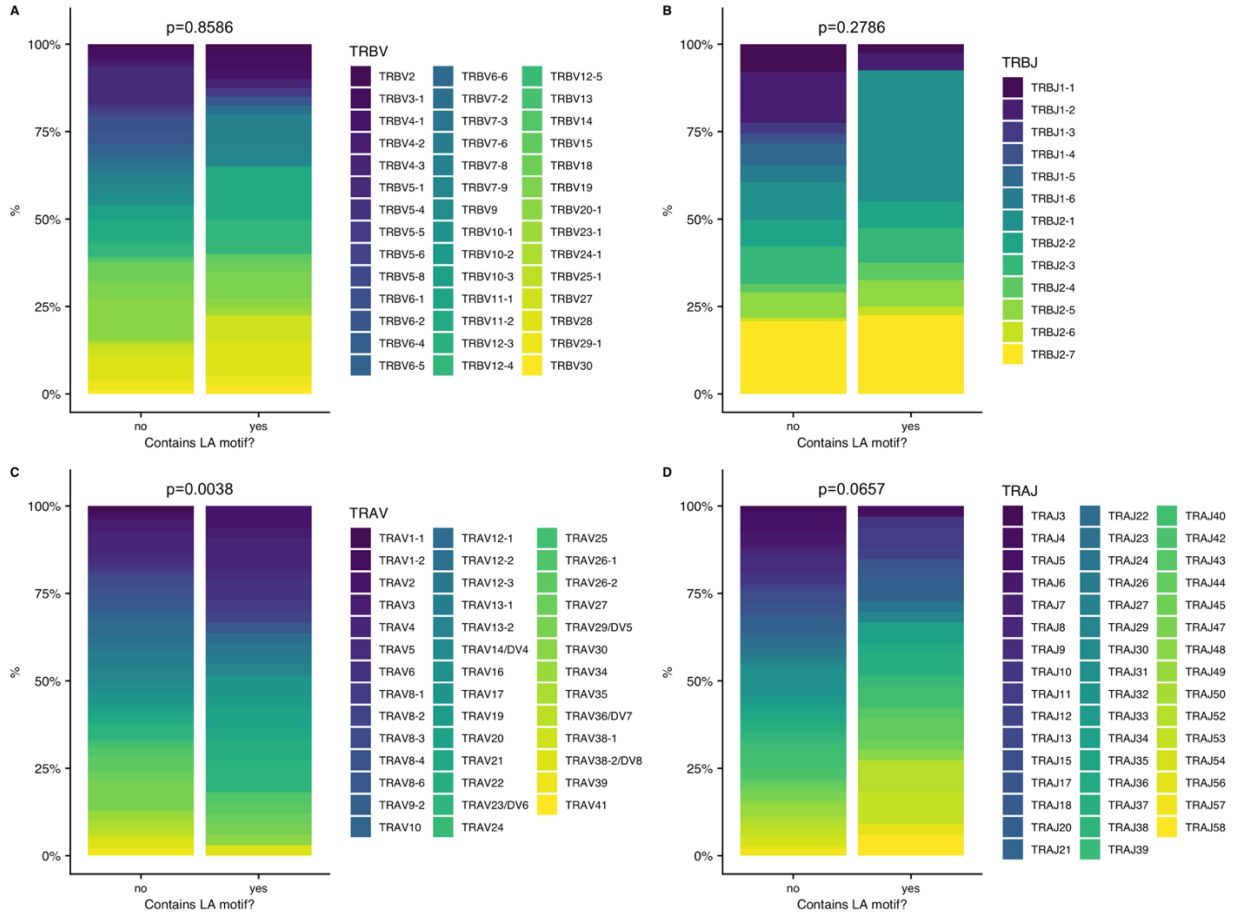
Given the association of I2 inflammation with PSC dysplasia, we tested whether I2 status could be used as a marker for risk of dysplasia. We stratified patients as I2 if at any point these

Figure 10: V(D)J usage by cell type in I2 PSC



TRBV (**a**), TRBD (**b**), and TRBJ (**c**) gene usage by cell type amongst CD4 T-cells from I2 PSC patients. 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. IL-17A+ Foxp3+. ("*" for $p<0.05$).

Figure 11: V(D)J gene usage amongst IL17A+ FOXP3+ CD4 T-cells containing "LA" motif



TRBV (**a**), TRBJ (**b**), TRAV (**c**), and TRAJ (**d**) gene usage amongst IL-17+ Foxp3+ CD4 T-cells stratified by whether the Beta chain contains the "LA" amino acid motif. (a-d) Significance determined by Chi-squared test.

Table 5: Amino acid and V, D, and J gene usage of TRB chain of cells containing "LA" motif

subject	barcode	CDR3 amino acid sequence	TRBV	TRBJ	TRBD
PSC50B	AATCCAGTCTACCTGC	CASSLPFLAGVPYEQYF	TRBV7-9*01 F	TRBJ2-7*01 F	TRBD2*02 F
PSC50B	ACTGAACAGGCTCTTA	CASSHVLAGGPTSYNEQFF	TRBV23-1*01 ORF	TRBJ2-1*01 F	TRBD2*02 F
PSC50B	AGGTCCGTCCCTAACC	CASSRLAGGGQYF	TRBV3-1*01 F	TRBJ2-4*01 F	TRBD2*02 F
PSC50B	CACACAAAGGACTGGT	CASGGSLANTEAFF	TRBV7-9*03 F	TRBJ1-1*01 F	TRBD2*02 F
PSC50B	CAGCCGAGTCTTGCGG	CASNKRLASGANVLTF	TRBV19*01 F	TRBJ2-6*01 F	TRBD2*01 F
PSC50B	CGGACACCACCCTATC	CASSPGLAGVGKHEQFF	TRBV12-4*01 F	TRBJ2-1*01 F	TRBD2*02 F
PSC50B	CTTTGCGGTATATCCG	CSVEGSGLAQYF	TRBV29-1*01 F	TRBJ2-3*01 F	TRBD2*01 F
PSC50B	GATGCTAAGGGTTTCT	CASSPGLSGGAGRGLASGVPGEQYF	TRBV18*01 F	TRBJ2-7*01 F	TRBD2*01 F
PSC50B	GTCGGGTGTTACGTCA	CASSEGVGLANTDTQYF	TRBV2*01 F	TRBJ2-3*01 F	TRBD2*01 F
PSC50B	TCTCATATCACCCTCA	CASSLLASGTNEQFF	TRBV7-8*01 F	TRBJ2-1*01 F	TRBD2*02 F
PSC50B	TTCTCAAGTAGCTCCG	CASRYRYLAGEETQYF	TRBV28*01 F	TRBJ2-5*01 F	TRBD2*02 F
PSC50B	TTGTAGGCACGTGAGA	CATSTRLAANEQFF	TRBV15*02 F	TRBJ2-1*01 F	TRBD2*01 F
PSC50B	TTTGTCAGTCGAACAG	CASSRGGLAGPLQYF	TRBV7-8*01 F	TRBJ2-4*01 F	TRBD2*02 F
PSC70D	AACTGGTAGGAGCGTT	CASDRNGKLAGGNYEQFF	TRBV19*01 F	TRBJ2-1*01 F	TRBD2*02 F
PSC70D	AAGGCAGGTAGCACGA	CACGLANSYEQYF	TRBV30*01 F	TRBJ2-7*01 F	TRBD2*01 F
PSC70D	ACACCCTAGCCAACAG	CASSSLAGGNYEQYF	TRBV11-2*01 F	TRBJ2-7*01 F	TRBD2*02 F
PSC70D	AGCCTAAAGAGTACCG	CASSFGLARTGELFF	TRBV28*01 F	TRBJ2-2*01 F	TRBD2*01 F
PSC70D	AGCTCTCTCGAGAGCA	CASSQVIGLAGGGGEQFF	TRBV4-1*01 F	TRBJ2-1*01 F	TRBD2*01 F
PSC70D	AGCTTGAGTCGCGGTT	CASSSLAGGPVGEQYF	TRBV11-2*01 F	TRBJ2-7*01 F	TRBD2*02 F
PSC70D	AGTAGTCCACCACCAG	CASSALAGARNEQFF	TRBV5-5*01 F	TRBJ2-1*01 F	TRBD2*02 F
PSC70D	CAAGGCCCACCGATAT	CASSTGLAGGQGTQYF	TRBV11-2*01 F	TRBJ2-3*01 F	TRBD2*02 F
PSC70D	CATCAAGGTTGGACCC	CASSLVPGLAETQYF	TRBV11-2*01 F	TRBJ2-5*01 F	TRBD2*01 F
PSC70D	CCTACACAGAGGACGG	CASSPGLAGTYNEQFF	TRBV7-8*01 F	TRBJ2-1*01 F	TRBD2*02 F
PSC70D	CGTCACTAGGTAAACT	CASSFKLAGNTDTQYF	TRBV27*01 F	TRBJ2-3*01 F	TRBD2*02 F
PSC70D	CGTGTCTCATATACGC	CASSNRGLAGGYNEQFF	TRBV27*01 F	TRBJ2-1*01 F	TRBD2*02 F
PSC70D	GATGAAAGTGACCAAG	CASSFVGLAGVEQFF	TRBV11-2*01 F	TRBJ2-1*01 F	TRBD2*02 F
PSC70D	GGATGTTAGGCGACAT	CASSFGLASEQFF	TRBV28*01 F	TRBJ2-1*01 F	TRBD2*01 F
PSC70D	GTACTTTTCCAAACAC	CASRLAGGPNSGNEQFF	TRBV7-2*01 F	TRBJ2-1*01 F	TRBD2*02 F
PSC70D	TAAGCGTTCCTTAATC	CASSQGLAGTYEQYF	TRBV4-3*01 F	TRBJ2-7*01 F	TRBD2*02 F
PSC70D	TACTTGTGTGCCTGGT	CASSQGRGGRLAGELFF	TRBV3-1*01 F	TRBJ2-2*01 F	TRBD2*01 F
PSC70D	TGACGGCTCATCGATG	CASSLILAESNYGYTF	TRBV28*01 F	TRBJ1-2*01 F	TRBD1*01 F
PSC70D	TGTGTTTCACTCTGTC	CASRILAQGGKTQYF	TRBV6-2*01 F	TRBJ2-5*01 F	TRBD1*01 F
PSC70D	TTCTACAGTACCCAAT	CSARAPRLAGVRYEQYF	TRBV20-1*01 F	TRBJ2-7*01 F	TRBD2*02 F
PSC70D	TTCTCCTCAGACAAAT	CASSSLAGASYEQYF	TRBV12-4*01 F	TRBJ2-7*01 F	TRBD2*01 F
PSC94	ATAGACCGTTCCCGAG	CASSLAFGLYGYTF	TRBV27*01 F	TRBJ1-2*01 F	TRBD1*01 F
PSC94	CGTTCTGTCAGCTCTC	CASNGLAGGRFNEQFF	TRBV19*01 F	TRBJ2-1*01 F	TRBD2*01 F

Table 5 continued.

PSC94	CTAGAGTAGGCTAGAC	CASTLGLAGADEQFF	TRBV12-4*01 F	TRBJ2-1*01 F	TRBD2*01 F
PSC94	CTCCTAGAGTACGCGA	CASRGLAGNTGELFF	TRBV11-2*01 F	TRBJ2-2*01 F	TRBD2*01 F
PSC94	GCATACAGTCTTCAAG	CASTLGLAGADEQFF	TRBV12-4*01 F	TRBJ2-1*01 F	TRBD2*01 F
PSC94	GTAGTCATCTGGTGTA	CASSFGLANGGGSSYEQYF	TRBV7-8*02 F	TRBJ2-7*01 F	TRBD2*01 F

Table 6: V and J gene usage of TRA chain of cells containing "LA" motif

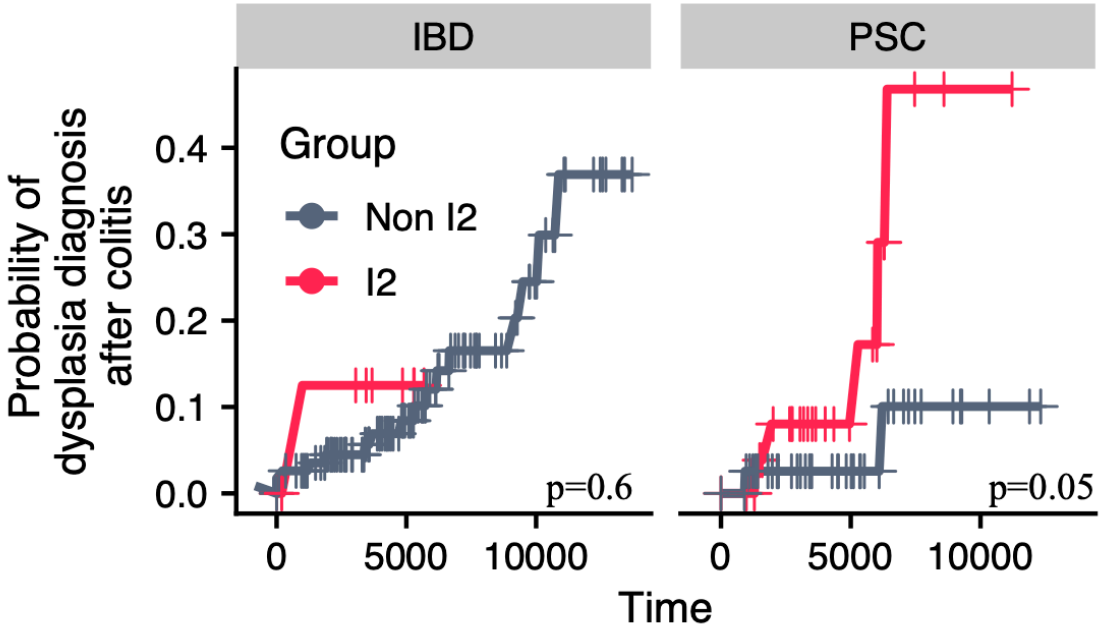
subject	barcode	TRAV	TRAJ
PSC50B	AATCCAGTCTACCTGC	TRAV22*01 F	TRAJ52*01 F
PSC50B	ACTGAACAGGCTCTTA	TRAV4*01 F	TRAJ44*01 F
PSC50B	AGGTCCGTCCCTAACC	TRAV23/DV6*03 (F)	TRAJ58*01 ORF
PSC50B	CACACAAAGGACTGGT	TRAV8-1*01 F	TRAJ37*01 F
PSC50B	CGGACACCACCCTATC	TRAV26-1*01 F	TRAJ29*01 F
PSC50B	CTTTGCGGTATATCCG	TRAV5*01 F	TRAJ17*01 F
PSC50B	GATGCTAAGGGTTTCT	TRAV23/DV6*03 (F)	TRAJ24*02 F
PSC50B	GATGCTAAGGGTTTCT	TRAV19*01 F	TRAJ48*01 F
PSC50B	GTCGGGTGTTACGTCA	TRAV3*01 F	TRAJ34*01 F
PSC50B	TCTCATATCACCCTCA	TRAV2*01 F	TRAJ40*01 F
PSC50B	TTCTCAAGTAGCTCCG	TRAV6*02 (F)	TRAJ10*01 F
PSC50B	TTTGTCAGTCGAACAG	TRAV4*01 F	TRAJ4*01 F
PSC70D	AACTGGTAGGAGCGTT	TRAV29/DV5*01 F	TRAJ20*01 F
PSC70D	AGCCTAAAGAGTACCG	TRAV23/DV6*01 F	TRAJ20*01 F
PSC70D	AGCTCTCTCGAGAGCA	TRAV8-2*01 F	TRAJ12*01 F
PSC70D	AGCTTGAGTCGCGGTT	TRAV12-1*01 F	TRAJ18*01 F
PSC70D	AGCTTGAGTCGCGGTT	TRAV8-6*02 F	TRAJ56*01 F
PSC70D	CAAGGCCCACCGATAT	TRAV13-1*01 F	TRAJ11*01 F
PSC70D	CCTACACAGAGGACGG	TRAV22*01 F	TRAJ34*01 F
PSC70D	CGTCACTAGGTAAACT	TRAV13-2*01 F	TRAJ53*01 F
PSC70D	CGTGTCTCATATACGC	TRAV20*02 F	TRAJ37*01 F
PSC70D	CGTGTCTCATATACGC	TRAV30*01 F	TRAJ40*01 F
PSC70D	GGATGTTAGGCGACAT	TRAV4*01 F	TRAJ39*01 F
PSC70D	TACTTGTGTGCCTGGT	TRAV6*02 (F)	TRAJ36*01 F
PSC70D	TGTGTTTCACTCTGTC	TRAV17*01 F	TRAJ52*01 F
PSC70D	TTCTACAGTACCCAAT	TRAV21*01 F	TRAJ53*01 F
PSC70D	TTCTCCTCAGACAAAT	TRAV20*02 F	TRAJ45*01 F
PSC70D	TTCTCCTCAGACAAAT	TRAV26-2*01 F	TRAJ44*01 F
PSC94	ATAGACCGTTCCCGAG	TRAV17*01 F	TRAJ58*01 ORF
PSC94	CGTTCTGTCAGCTCTC	TRAV27*01 F	TRAJ52*01 F
PSC94	CTAGAGTAGGCTAGAC	TRAV38-2/DV8*01 F	TRAJ43*01 F
PSC94	CTCCTAGAGTACGCGA	TRAV12-3*01 F	TRAJ53*01 F
PSC94	GTAGTCATCTGGTGTA	TRAV2*01 F	TRAJ11*01 F

subjects were identified as being I2, and measured time from diagnosis of colitis to the first incidence of right-sided dysplasia. We found that I2 status amongst PSC patients, but not IBD patients, was associated with a greater risk over time for the development of right-sided dysplasia (Figure 12). I2 status was not associated with a greater risk of developing dysplasia outside of the right colon (Figure 13), emphasizing that dysplasia risk is only increased in the region for which I2 inflammation is observed. This poises I2 status as a predictor of dysplasia risk in PSC and could potentially use clinically to assess risk of dysplasia amongst PSC patients.

PSC inflammation is associated with an expansion of CRC-associated bacterial taxa

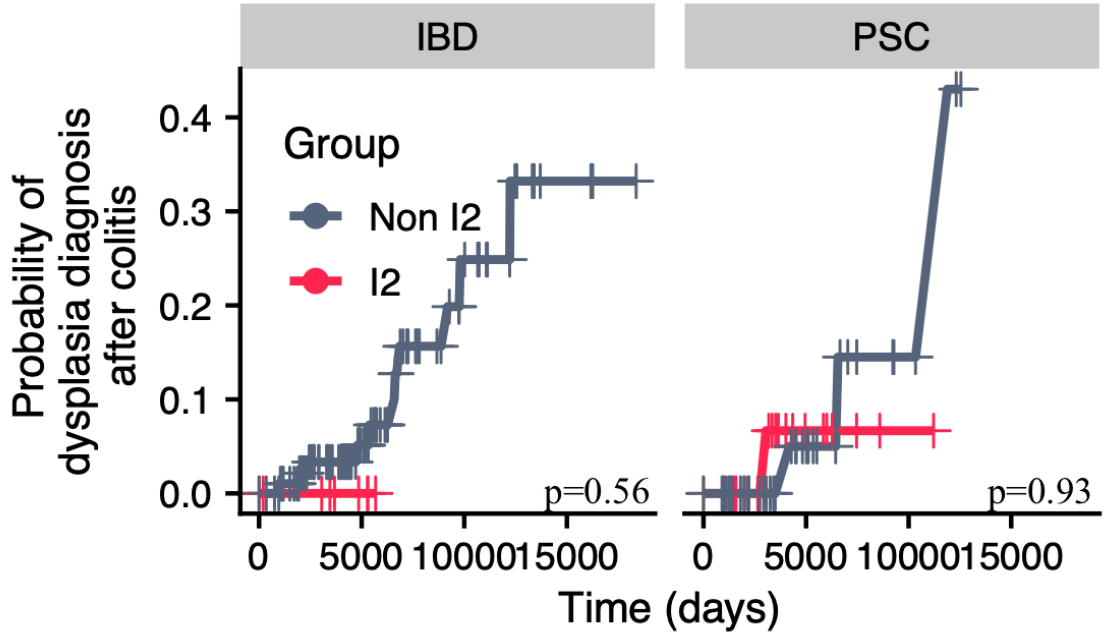
To further explore the possibility of antigen drive, we performed 16S sequencing on the DNA extracted from the tissue biopsies. We chose to explore the possibility of an antigen of bacterial origin due to previous work showing alterations in the PSC microbiome and the clinical trials showing preliminary beneficial effects of antibiotics on liver function and colitis. 16S sequencing demonstrated no difference in total bacterial load or Shannon diversity index (SDI) between IBD and PSC (Figure 14 A,B), nor across transcriptional clusters (Figure 14 C,D). We probed for bacterial taxa enriched in PSC-I2 versus PSC-U and IBD-I2 versus IBD-U, with the idea that these comparisons would identify taxa that are enriched in PSC inflammation and IBD inflammation specifically. We identified a handful of taxa to be specifically enriched in PSC inflammation as compared to IBD inflammation (Figure 14 E). Notably, three out of five of these taxa were previously implicated in the development of CRC[157,158]. Any one, or combination of, these taxa could represent a source of antigen driving the expansion of IgG plasma cells or DP CD4 T-cells. Though further exploration is necessary to determine whether this is true, we are equipped with representative Ig and TCRs to formally test this hypothesis.

Figure 12: Status as I2 is associated with a greater risk and shorter time to dysplasia in PSC but not IBD.



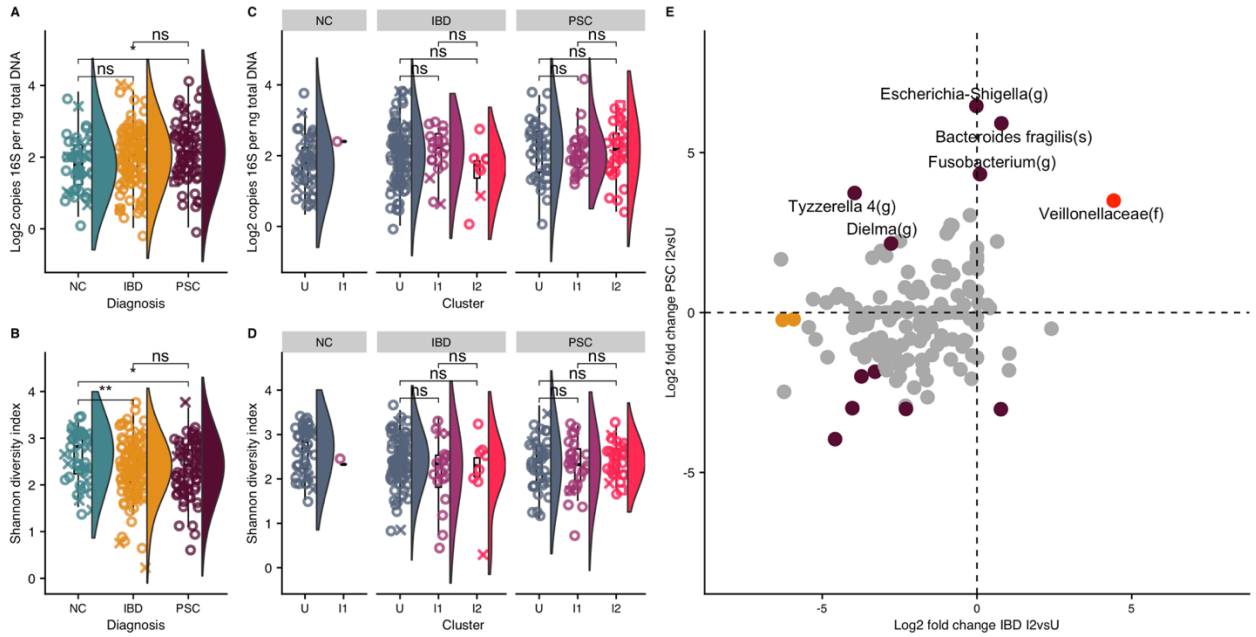
Kaplan-Meier-estimated curves for risk of right-sided dysplasia over time, with subjects stratified as either I2 at any timepoint or not based on transcriptional profiling. Subjects are subset by diagnosis: IBD (right) or PSC (left).

Figure 13: Status as I2 in the right colon is not associated with risk for non-right-sided dysplasia.



Kaplan-Meier-estimated curves for risk of non-right-sided dysplasia over time, with subjects stratified as either I2 at any timepoint or not based on transcriptional profiling. Subjects are subset by diagnosis: IBD (right) or PSC (left).

Figure 14: CRC-associated bacteria are enriched in PSC inflammation.



a, Log2 copies of 16S DNA per ng of total input DNA across diagnosis groups. **b,** Shannon diversity index of samples across diagnosis groups. **c,** Log2 copies of 16S DNA per ng total input DNA across clusters. **d,** Shannon diversity index of samples across diagnosis groups. **e,** Log2 fold change of absolute abundance of bacteria taxa in PSC I2 vs U (y-axis) and IBD I2 vs U (x-axis). Each point represents a bacterial taxeme. Taxa significantly different (non-adjusted p-value <0.5) in PSC I2 vs U highlighted in purple, in IBD I2 vs U highlighted in orange, and taxa significantly different in both comparisons in red. Taxa with a >2 log2 fold change in abundance in PSC I2 versus U are annotated by name and taxonomic level(f = family, g = genus, s = species). (a-d) Significance determined by Wilcoxon test ("*" for p<0.05, "**" for p<0.01, "ns" for not significant (p>0.05)).

Results Summary and Conclusion

The major aim of our study was to uncover immunological factors that explain the difference in presentation of colitis and increased risk of CRC in PSC. By directly comparing the environments in which dysplasia occurred, we determined that PSC dysplasia is highly associated with inflammation as determined histologically and transcriptionally. The milieu of IBD dysplasia, on the other hand, is indistinguishable from sporadic dysplasia, despite all the evidence suggesting that they arise by different mechanisms. Though our transcriptional analysis was limited to right-sided dysplasia, the histologically scored inflammation of non-right sided dysplasia followed the same trends as in the right colon, suggesting that these results are applicable to dysplasia arising anywhere in the colon. Given these data, we conclude that PSC dysplasia, but not IBD or sporadic dysplasia, is dependent on active inflammation.

We observed the PSC dysplasia-associated inflammatory signature in a subset of PSC patients with no history of dysplasia. This inflammation is therefore not a response to dysplasia, and likely represent a driving force or risk factor for the development of dysplasia in PSC. We demonstrated that PSC patients who at any point had this inflammatory signature were at higher risk for dysplasia than PSC patients without this signature. This signature could be a clinical indicator of PSC patients at greatest risk for dysplasia and could be implemented to improve screening and management of CRC risk in this already vulnerable population. As this inflammatory signature is only associated with risk for dysplasia in the segment of the colon in which it occurs, clinicians could also use the presence of this inflammation as an indicator of the segment of colon in which dysplasia is most likely to develop. This added information could allow for more careful CRC screenings, as well as a better utilization of time and resources.

The dysplasia-associated inflammation in PSC is distinct from IBD inflammation in several ways. Overall, the transcriptional signature of PSC inflammation is enriched for T-cell activation, response to pathogen, and other immune signatures. These enrichments are observed despite actively inflamed PSC patients having a comparable level of inflammation to actively inflamed IBD patients. We observed both an influx of recently arrived IgG plasma cells and an increase in IL-17A+ Foxp3+ CD4 T-cells in PSC inflammation. The largest plasma cell clones in inflamed PSC patients had highly mutated and highly diverse CDR3, consistent with signs of antigen drive. IL-17A+ Foxp3+ CD4 T-cells in inflamed PSC patients were enriched for an amino acid motif within their CDR3, suggesting selection and antigen drive in these cells as well. To our knowledge, this is the first formal evidence of what has been long hypothesized to be the case in PSC: that inflammation is antigen driven.

DISCUSSION

Biological and clinical implications in the understanding and management of PSC

Despite known differences in presentation of inflammation and CRC outcomes, little to know evidence existed on the nature of PSC inflammation in comparison to IBD inflammation. In this thesis we detailed, for the first time, differences in the natures of the inflammations and showed that these differences were directly related to clinical outcomes. First, we demonstrated that PSC dysplasia is always found in the context of active inflammation, whereas IBD dysplasia is not. This suggests that IBD and PSC inflammation are relevant in different ways to the development of dysplasia. In the initiation and promotion theory of cancer development, cancer develops as a result of cells that accumulate genomic alterations that are then preferentially expanded in response to a proliferation signal. In this model genomic alterations are the irreversible initiation event, and the proliferation signal acts as the promotion event. IBD, which is not often found at the time of dysplasia, cannot therefore promote cancer, as it would have to be providing active signal for the mutated cells to preferentially divide. Since duration and severity of colitis are factors that predict dysplasia in IBD, perhaps IBD inflammation is more relevant as an inducer of mutations than as a proliferation signal. PSC on the other hand, is always found at the instance of dysplasia. Therefore, it is likely that PSC inflammation can act as a promoter of cancer. Perhaps it is in fact a very strong promoter of cancer. We cannot exclude the possibility that PSC inflammation can induce genomic alterations, and thus is relevant as an initiator of cancer, however its role as a promoter of cancer might at least partially explain why PSC patients are at a greater risk for the development of cancer than IBD. Additionally, it would be interesting to further investigate why it is that PSC dysplasia is always inflamed whereas IBD

inflammation is not. Could this be due to a dependence of PSC dysplasia on inflammation due to the nature of the genomic alternations? Or could it be that PSC inflammation is such a strong promoter of cancer that it can very easily induce dysplasia with fewer genomic alterations?

We also demonstrated that the nature of PSC inflammation is unique and distinct from IBD inflammation and is characterized by an increased frequency of IgG plasma cells and IL-17A+ Foxp3+ CD4 T-cells. This is one of the first, if not the first, study that formally identified unique cellular features increased in PSC inflammation as compared to IBD inflammation specifically. We were also able to implicate this unique inflammation with the risk for dysplasia, suggesting that these cell types might be relevant to the development of dysplasia. In the model of pathogenesis for CeD, CD4 T-cells control a Th1 immune response resulting in villous atrophy. We see the IL-17+ Foxp3+ CD4 T-cells as a potential source of cytokines driving the development of dysplasia. IL-17A has previously been implicated in epithelial cell proliferation in conjunction with other cytokines[43]. IL-17A+ Foxp3+ could therefore secrete another cytokine or set of cytokines that have yet to be identified which are also inducing dysplasia. As the IL-17A+ Foxp3+ CD4 T-cells demonstrate a pathogenic-like signature, we believe them to be directly related to the development of dysplasia and inflammation, though the exact mechanisms must still be worked out. The influx of IgG plasma cells might also play an important role in the pathogenesis of inflammation and dysplasia. In CeD, it is thought that B-cells potentiate the T-cell response by presenting antigen to T-cells on HLA class II, thus activating the T-cells to further cause inflammation. This could also be the case in PSC, where we see both a unique B-cell phenotype found in conjunction with the T-cell phenotype.

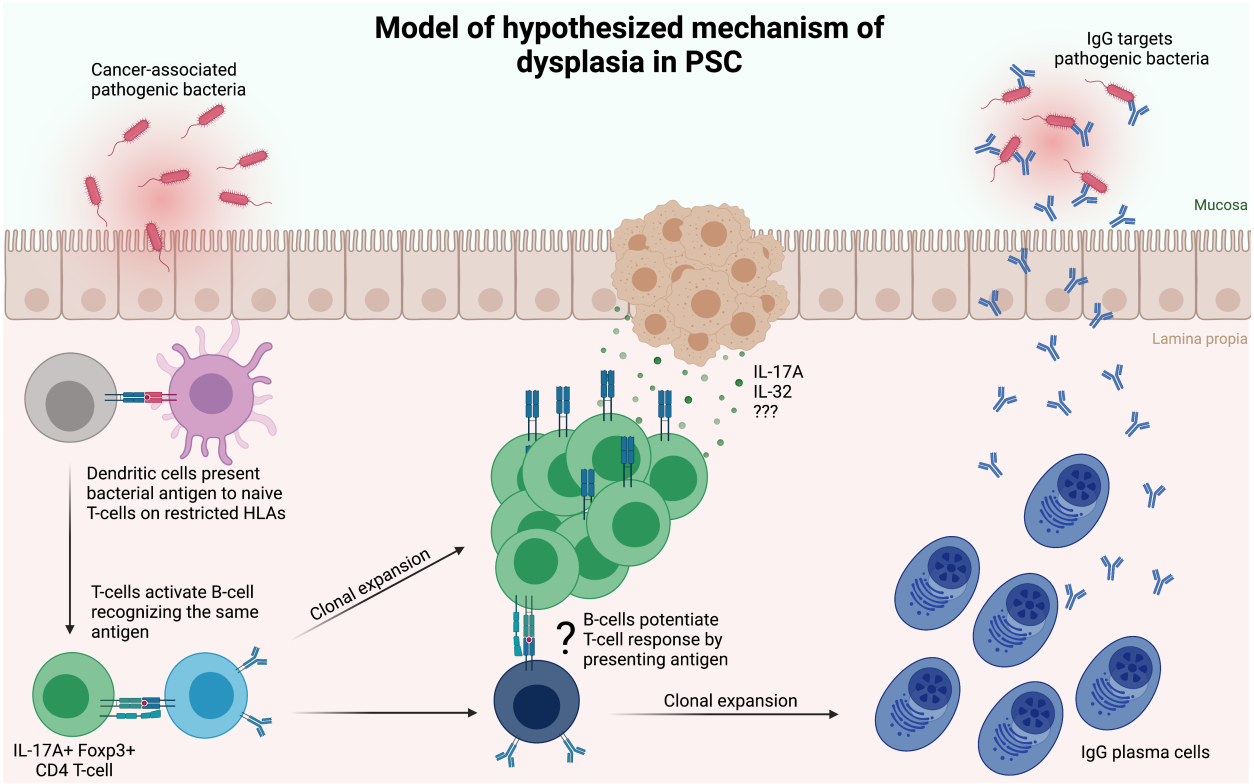
Both the plasma cells and CD4 T-cells show signs of antigen drive. T-cell antigen drive is found in the form of an enriched amino acid motif within the CDR3. B-cell antigen drive was

found in the form of a heavily mutated, highly diverse CDR3 within the plasma cells of I2 subjects. This is the first evidence that has ever been provided that PSC is in fact antigen driven, as previously hypothesized. If PSC is in fact antigen driven, the implications for the prevention of the disease are massive. If the antigen can be eliminated from the gastrointestinal system of these patients, then we can dramatically reduce the burden of inflammation and dysplasia in these patients. However, this antigen could be from many sources including bacteria, viruses, fungus, diet, or self. We hypothesize that there could be a single or small number of antigens from a bacterial source that drive B- and T-cell expansions, and that these expansions, or the antigen itself, drive dysplasia. There is mounting evidence that bacterial-derived molecules can induce DNA damage[159], serving as an initiator of tumorigenesis. Perhaps the source of the antigen, or the antigen itself is such a carcinogen. Removal of the antigen from the intestinal environment should eliminate or at least greatly reduce the increased risk for CRC in PSC patients. If the antigen is of bacterial origin, this might represent a unique situation in which CRC is prevent by antibiotic. If not bacterial, identification of the antigen would still facilitate the design or application of preventative therapies.

We propose an antigen driven model of immune-mediated dysplasia that is in line with all the findings in this thesis (Figure 15). This is the first comprehensive mode of PSC that fully encapsulates the broad hypothesized mechanisms of dysplasia. Though there are many questions left to be answered, we hope that this model provides some important concepts and inspires future directions that will lead to a full understanding of the mechanisms of dysplasia in PSC.

Even without finding an antigen, this work contributed to our understanding of PSC inflammation. These findings are clinically relevant, as we have demonstrated that I2 inflammation is associated with active dysplasia and a greater risk for dysplasia in PSC. We

FIGURE 15: Model of hypothesized mechanism of dysplasia in PSC.



Hypothesized mechanisms of colonic dysplasia in PSC: A pathogenic bacteria colonizes the epithelium of genetically susceptible PSC patients. Dendritic cells present bacterially derived peptides to T-cells on restricted HLAs. Those T-cells are converted to IL-17A-secreting Foxp3 T-cells that then provide T-cell help to naïve B-cells recognizing the same antigens. B-cells potentiate the T-cell response by providing antigen to the already clonally expanded IL-17A+ Foxp3+ CD4 T-cells. These CD4 T-cells produce IL-17A along with other cytokines that provide a proliferation signal to epithelial cells. IgG plasma cells secrete antibody that targets the same bacterial antigens presented by dendritic cells.

demonstrated that histologically scored inflammation is not always concordant with transcriptionally determined inflammation. We must therefore reevaluate how inflammation is assessed in PSC patients, as neutrophil compartmentalization does not capture the full essence of PSC inflammation. We would suggest assessing the presence of IgG plasma cells as well as DP cells as additional markers of inflammation and risk factors for dysplasia. Or the presence of I2 inflammation itself could be a clinical predictor of which patients are likely to develop dysplasia. More practically, I2 inflammation could be reduced to a handful of signature genes for which expression could be measured by quantitative polymerase chain reaction (qPCR) from tissue biopsies. Those patients with a significant score for I2 inflammation by qPCR could then be monitored more closely for signs of dysplasia. Additionally, assessing I2 inflammation across the colon could indicate where each PSC patients is likely to develop dysplasia.

Inflammation and cancer

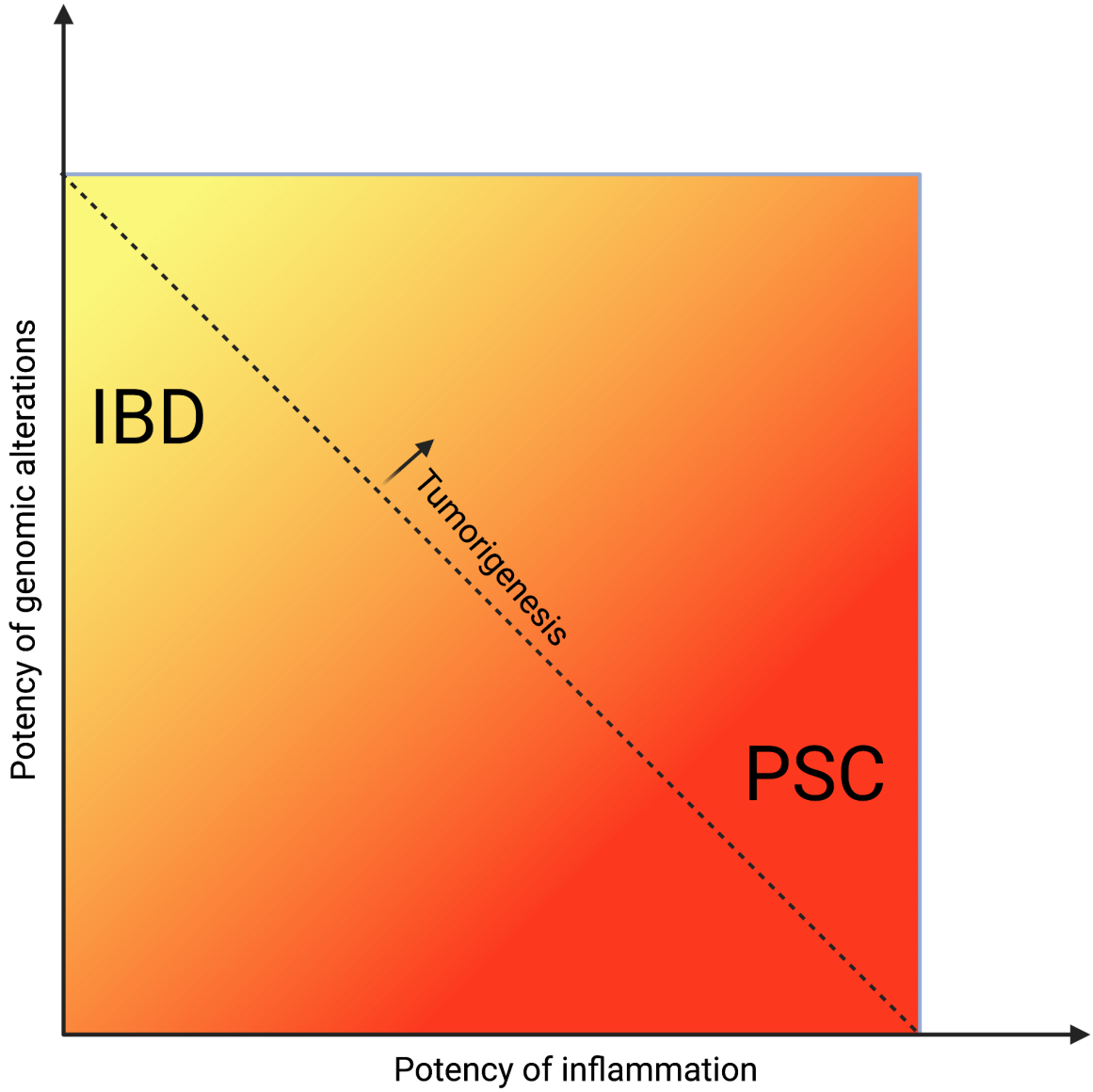
Since PSC colitis confers a several-fold greater risk of CRC than IBD colitis, not all inflammations are equivalently capable of inducing cancer. Somehow PSC inflammation is more potent than IBD inflammation, either in frequency, severity, or nature. Our evidence suggests that the severity of inflammation does not differ between IBD and PSC, but rather that the nature of the inflammation is different. PSC inflammation is therefore likely more carcinogenic by nature. This could have to do with the composition or quantity of cytokine or other secreted factors associated with PSC inflammation.

Inflammation can be relevant to the induction or promotion of tumors, which begs the question as to which of these events PSC inflammation contributes. We observed that PSC dysplasia, but not IBD dysplasia, was characterized by active inflammation. Additionally, PSC

dysplasia patients were more inflamed than PSC patients without dysplasia. Based on this data, we hypothesized that PSC dysplasia is dependent on active inflammation, whereas IBD dysplasia is not. We further hypothesize that IBD dysplasia does not depend on active inflammation because it is characterized by a greater number or potency of mutations than PSC dysplasia. In contrast, we hypothesize that PSC inflammation is a more potent inducer of proliferation, thus dysplasia would require fewer baseline mutations to proliferate. We theorize that there exists a spectrum of potency of mutations and inflammation, and that the summation of their effects (after passing a certain threshold) result in tumors. In our theory, lowly mutated cells can become tumors in the context of excessive inflammation, whereas highly mutated cells would require little to no external signal to proliferate uncontrollably. We believe that IBD dysplasia exists on the potent mutations/weak inflammation end of the spectrum, and PSC dysplasia is on the weak mutations/potent inflammation end (Figure 16).

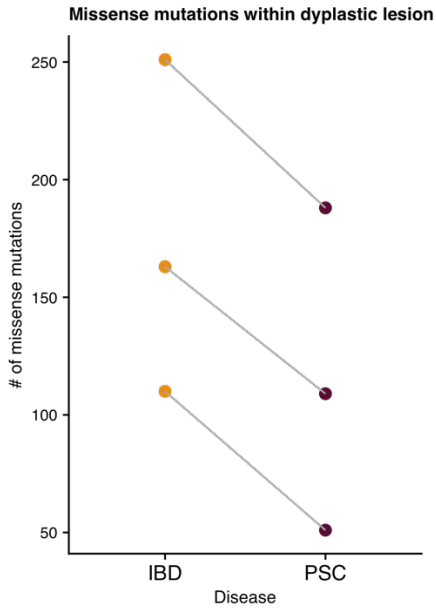
We wanted to test this more formally, by looking at the mutational landscape of dysplasia in IBD and PSC. If our hypothesis is correct, there should be fewer mutations in PSC dysplastic lesions than IBD dysplastic lesions. We had access to DNA from three sets of matched IBD and PSC dysplasia. These samples were matched by age, sex, race, duration of disease, and the stage, type, and location of the dysplasia. We performed whole exome sequencing on these samples and quantified the total number of missense mutations in each sample. In line with our hypothesis, we observed that PSC dysplasia had fewer missense mutations than the paired IBD dysplasia (Figure 17). We wish to expand this dataset to a greater number of matched samples to draw stronger conclusions about the mutational burden in PSC. Previous work demonstrates that IBD dysplasia has similar mutations as sporadic dysplasia. However, no formal study has compared either of these dysplasia to PSC. As our preliminary data suggests, there might be

Figure 16: Potency of mutations and inflammation in IBD and PSC dysplasia development.



The x-axis represents the theoretical potency of inflammation in inducing proliferation of mutated cells. The y-axis represents the theoretical potency of genomic alterations in pro-cancerous cells. The dashed line represents an arbitrary threshold for the development of dysplasia. We believe PSC patients to have more potent inflammation and less mutations, while IBD patients have less potent inflammation but more potent mutations.

Figure 17: PSC dysplasia has fewer missense mutations than IBD dysplasia.



Total number of missense mutations from dysplastic lesion DNA from IBD and PSC patients. Gray lines represent matched pairs. Samples were matched by age, sex, race, duration of disease, and the stage, type, and location of the dysplasia.

fewer mutations in PSC. Additionally, it is possible that the nature of the mutations themselves are different and confer differential proliferative capacity *in vivo.*

This concept has broader implications on the role of inflammation in carcinogenesis. It raises the question as to whether certain inflammations are better inducers of cancer while others are better propagators. What factors would make an inflammation potent in each regard, and could we identify common factors across inflammation? Is there a point after which sufficient mutations cause proliferation completely independently of inflammation? And importantly, how can we control these inflammatory factors to prevent the initiation and proliferation of cancer?

Future directions

Our results provide the first formal evidence that PSC inflammation is antigen driven, and that this inflammation, distinct from IBD colitis, is strongly associated with dysplasia. This opens the possibility of many future studies which can address additional details about PSC inflammation and dysplasia, address prospective questions, and independently validate our results. We outline below the future directions that we believe are most exciting and important:

Identification of driving antigen

The influx of highly mutated and diverse IgG plasma cells and increase in motif-containing, activated, pathogenic-like DP CD4 T-cells strongly suggests antigen drive to PSC inflammation. It would be highly valuable to identify the driving antigen, as this is a prime target for therapeutic intervention to prevent inflammation and dysplasia. Our 16S data identified several bacteria uniquely associated with PSC inflammation which are possibly the source of driving antigen. We do not eliminate the possibility that other bacterial taxa are involved in

pathogenesis, so a broader approach to the role of bacteria in pathogenesis, while paying special attention to the previously identified taxa, is necessary. Prospectively, we would like to preserve tissue biopsies under special conditions that maintain bacteria and mucus layers so that we can perform bacterial immunohistochemistry and fluorescence *in situ* hybridization (FISH). These assays would allow us to analyze the compartmental distribution and composition of bacteria within the tissue. We could determine whether I2 PSC patients have any bacteria adhering to the epithelium or invading into the tissue, as these are potentially the bacteria inducing inflammation.

We would also like to collect fresh mucosal scrapings or lumenal contents from PSC patients to identify endogenously IgG-coated bacterial populations. Identification of these populations by flow and then enrichment via FACS or magnetic bead-based methods would allow for subsequent DNA sequencing of these bacteria. The taxa found in the IgG-coated fraction in PSC, but not the IgA-coated fraction, or IgG-coated fraction in IBD represent additional potential sources of antigen drive. We of course expect that at least a subset of taxa identified by 16S sequencing to be coated by IgG.

The full-length Ig and TCR sequences identified by our single cell sequences are additional useful tools for identifying potential antigens. From the paired heavy and light chain sequences we can generate an unlimited amount of monoclonal antibody (mAb) from the expanded IgG clones of PSC I2 patients. We can test these mAbs for reactivity to *ex vivo* or cultured bacterial cell lysates by enzyme-linked immunosorbent assay (ELISA). Then, specific proteins of interest can be isolated using immunoprecipitation. We can also clone the “LA”-containing DP TCRs into reporter T-cell lines to test whether lysates or antigens that are bound by the mAbs also activate these TCRs. Though we have evidence that the source of antigen is

potentially bacterial, we cannot eliminate the possibility of a viral, dietary, or self-antigen. Fortunately, these mAbs and reporter T-cell can be also used to test a broader array of antigens.

Further characterization of DP CD4 T-cells

Our single cell sequencing of DP cells determined that these cells are activated and have a pro-pathogenic signature, potentially implicating them in inflammation and dysplasia pathogenesis. Due to limitations in single cell sequencing, we have not been able to fully characterize these cells. Bulk sorting and sequencing of these cells would provide a much better idea of their functional properties. We are currently unable to bulk sequence these cells because IL-17A and Foxp3 are intracellular markers that require fixation and permeabilization of the cell to visualize by flow cytometry. Fixation degrades RNA, preventing high-quality RNAseq. Further flow cytometric characterization of these cells is necessary to identify a suitable surface protein that can serve as a proxy to distinguish these cells from IL-17A and Foxp3 SP cells. LAG3, a gene associated with TCR activation, codes for a surface protein that could potentially serve as such a proxy. Assessing the overlap between protein expression of Lag3 and IL-17a+ Foxp3+ is a simple experiment that can be done by flow cytometry. Additionally, we can test additional candidate surface markers based on the literature and a deeper probing of the single cell transcriptional data.

Mechanisms of dysplasia and the dependence on active inflammation

We have not provided any evidence beyond association that DP T-cells or IgG plasma cells play a role in the development of dysplasia. We are inspired by the work in CeD, which clearly point to T-cells as mediators of tissue destruction, and potentially B-cells as potentiators

of tissue destruction. Given the similarities between CeD and PSC with respect to HLA association and antigen drive, we hypothesize that lymphocytes may be involved in the development of dysplasia in PSC. If we can successfully sequence bulk DP cells, we would specifically investigate expression of different cytokines to identify mediators of a pro-proliferative or pro-tumorigenic signal. IL-17A itself has already been implicated in epithelial cell proliferation. There are likely other factors involved in the development of dysplasia, however, given that there is no difference in total IL-17A+ cells across transcriptional clusters within PSC. In fact, there is a greater frequency of IL-17A+ Foxp3- cells in PSC U than PSC I2 (Figure 8H). We would therefore test whether IL-17A in conjunction with other cytokines could produce epithelial cell proliferation or dysplastic morphology. Intestinal organoid cultures could be treated with various cytokines found to be increased in DP cells, and then assessed for proliferation markers such as Ki-67 and p53.

We are particularly interested in further investigating the hypothesis that PSC dysplasia depends on active inflammation unlike IBD dysplasia. As we hypothesize that this has to do with the relative potency of the accumulate mutations and inflammatory milieu, we would need to directly prove the relevance of both factors. First, we would like to expand our whole exome sequencing analysis to a much larger of matched PSC and IBD dysplasia samples. With only three sets of samples, it is impossible to draw any conclusions about mutational abundance or about differences in genes mutated. In expanding this dataset, we can draw stronger conclusions about the mutational burden in PSC with respect to IBD, and potentially also identify driver mutations unique to PSC.

We could also compare the proliferative ability of organoids derived from dysplastic lesions. We could test various inflammatory conditions to determine whether IBD and PSC

epithelial cells have intrinsic differences in their ability to respond to cytokine or other proliferation signals. We could test *in vitro* the potency of the PSC and IBD inflammation by treating control organoids with supernatants from short-term biopsy cultures and measuring proliferative response. This assay in particular would be useful in determining which cytokines are mechanistically relevant to dysplasia in PSC.

Relationship between intestinal and liver pathologies.

How the liver and intestinal pathologies in PSC relate to each other is unknown. The liver of PSC patients also succumbs to inflammation and cancer, so the obvious question is whether the nature of the liver inflammation is similar to intestinal inflammation. Liver biopsies are invasive, and not routinely taken for research purposes, making it difficult to assess inflammation in the same way that we did for the colon. Immunohistochemistry and FISH on tissue slides cut from archived standard of care biopsies would allow for the quantification of IgG plasma cells and DP CD4 T-cells. However, correlation of these results with colon transcriptional cluster would not be possible as these are not likely subjects we have enrolled in our study. Additionally, Ig and TCR analysis would pose a challenge without being able to sequence freshly isolated cells. Nonetheless, investigation of these cell types in the liver is warranted.

Approaches to the study of complex human diseases

Human tissue-based studies are challenging as they do not offer the same set of optimized conditions as in animal models. Many animal models allow for complete control of genetic and environment factors, can introduce experimental interventions *ad libidum*, and have a near

limitless supply of diseased and control study subjects. Human studies on the other hand, are mostly constrained to descriptive observations, are vulnerable to the immense heterogeneity of genetic and environmental influences and are limited in the possible readouts due to the rarity of a disease or amount of tissue available for study. While these factors hinder what can be achieved in human tissue-based research, this field still brings one critical advantage over animal models: that all observations, if done correctly, are true and directly relevant to the disease which is being studied.

Animal models often rely on several assumptions about the pathogenesis of the human diseases that they model or can only mimic the symptoms of the disease. For monogenic diseases or diseases for which the pathogenesis is very well understood, animal models are powerful tools in the investigation of details of pathogenesis and in the search of potential therapies. However, these models are less effective for the study of complex disease for which the pathogenesis is unknown. Using these models may not even reveal any true details about the disease itself and studying them will only provide more information about the mechanisms within that model. One proposed mouse model of PSC is the *Mdr2* knockout mouse, which develops spontaneous cholestatic liver injury similar to PSC[160]. However, this model does not also have colitis, so other groups have proposed treating *Mdr2* knockout mice with dextran sulfate sodium (DSS) to induce colitis, in order to have both the intestinal and hepatic features of PSC[161]. While this model might replicate some of the pathologies, it is very likely that PSC colitis is not caused by DSS, and no genetic studies have ever found associations between PSC and MDR2. Therefore, continued, thorough investigation using human tissue is essential if there is ever to be an accurate, comprehensive mouse model of PSC.

The study of complex human diseases is difficult due to the technical difficulty in acquiring the necessary samples, as well as the immense heterogeneity that exists within populations with the disease of interest. However, thoughtful design and careful execution can maximize the likelihood for success of a study, despite the technical and biological obstacles. In fact, the natural heterogeneity that exists within a population being studies is an opportunity to investigate how multiple factors influence disease progression and can lead to key details about the pathogenesis of a disease. As most complex human diseases are caused by the intricate interaction of numerous host and environmental factors, embracing the complexity, and leveraging the heterogeneity is critical. PSC is one such complex disease. In our study of PSC, we have tried numerous approaches and techniques to varying degrees of success. We have learned a lot about strategies that were effective and spent a lot of effort and time on strategies that were ultimately unsuccessful. We describe below our experience with different approaches to studying PSC with the hopes that these can be more generally applied to other studies of complex diseases.

Practical considerations and study design

From a practical standpoint, accessibility to samples and patient data are essential for any study on human disease: there is no possible project without data. However, even when access to patient data and samples is possible, there are certain environments that enable successful projects. As an academic medical center, UCM facilitates close interactions between the clinical and research faculty. We have been fortunate to have access to a near limitless supply of intestinal tissue, which allows us to perform and optimize all our experimental read-outs. Additionally, as a quaternary care center for gastrointestinal diseases, we are fortunate to acquire

tissue samples from a large cohort of patients with PSC from across the region, though PSC is very rare in the general population.

The design of human studies is key and begins with the selection of the disease to study. Heterogenous diseases, where symptoms and outcomes are highly variable, are difficult to study as there could be multiple mechanisms of disease. SLE, for example, is thought by some to be a collection of many different diseases, each with potentially different mechanisms of pathogenesis. Without correctly sub-setting these subjects by the relevant variables, it is unlikely that experimental readouts within comparison groups converge in any meaningful way. Uncontrolled extraneous factors could mask and effects of the comparison groups. UC and CD are genetically heterogenous, with a lot of variability in disease presentation, which may contribute to why no definitive cause of these diseases has been identified. PSC on the other hand is strongly associated with HLA by GWAS, suggesting that PSC patients are genetically similar within their HLA, and that one or a small set of HLAs are directly relevant to disease. In this way, genetic factors, especially with respect to immune function, are better controlled. Perhaps if researchers were to subset IBD and other IMID patients by HLA, there might be clearer immunological readouts.

Another key design feature of our study that led us to our findings was restricting our analysis to a specific segment of the colon. We restricted our analysis to the right colon because inflammation and dysplasia were most often observed in the right colon. In doing so, we controlled for regional variability in physiology, immune function, and bacterial composition across the colon. Additionally, in selecting the right side specifically, we maximized our chances of observing active inflammation and dysplasia in our population.

The selection of appropriate controls is also critical in study design. We paid special attention to only enroll IBD patients with a history of right-sided disease, as essentially all PSC patients have right-sided colitis. This study, along with others, highlight the importance of regional inflammation in the development of CRC. Had we not selected IBD controls with a history of right-sided inflammation, we may have inadvertently compared PSC patients to tissue that was essentially entirely healthy, and without the potential to develop colitis-associated dysplasia. We observed differences in inflammation and dysplasia across comparably inflamed populations, allowing us to draw specific conclusions about PSC inflammation, as opposed to inflammation in general.

Leveraging of clinical data and integration with biological readouts

The clinical relevance and importance of a translation study depends on how we leverage the clinical data and integrate it with the biological readouts. Our analysis was centered around the development of dysplasia, so collection of data such as dates of first dysplasia, location of dysplasia, histologically and endoscopically scored inflammation at the dysplastic lesion was critical. Important details to collect are not always obvious at the beginning of the study, and it is important to revisit patient data and re-analyze biological readouts periodically. For example, the link between active inflammation and dysplasia was not obvious to us at the onset of the study. We originally noticed that I2 PSC patients were getting colectomies at a higher rate than the non-I2 PSC patients, and only through careful examination of the patient records did we notice that the indication for colectomy was dysplasia. Even after uncovering the relationship between active inflammation and dysplasia, we initially had not sub-setted PSC patients based on the location of the dysplasia. Only in revisiting the charts once again, and carefully recording the

location in which dysplasia developed were we able to separate out PSC patients with right-sided and non-right-sided dysplasia. This distinction turned out to be key because only right-sided dysplasia patients are enriched for I2, and I2 is only a risk factor for right-sided dysplasia.

It is tempting to rely on pre-determined clinical factors when performing comparisons. We could have based our entire analysis based off histologically and endoscopically scored inflammation, for example. While doing so could have been useful in progressing our work, it is built on certain assumptions that may not be relevant to pathogenesis. At UCM, inflammation is determined by the presence of neutrophils in different compartments of the colonic mucosa. This method of course is clinically sound in the assessment of colitis activity but does not capture the entire nature of the inflammation. As we have shown, PSC inflammation is characterized by a high proportion of IgG plasma cells and IL-17A+ Foxp3+ CD4 T-cells, none of which are assessed in the clinical scoring of disease activity. Instead, by scoring inflammation transcriptionally and assigning groups based on their transcriptional profile, we were more unbiased in our comparisons, which ultimately led to our discovery of these relevant cell types. Histologically and endoscopically, there is no difference in the levels of inflammation between PSC patients with and without dysplasia. Had we depended only on those clinical scores, we would not have been able to conclude anything about the importance of inflammation in PSC dysplasia. On the other hand, the transcriptionally determined inflammation was significantly increased in PSC dysplasia as compared to those without dysplasia, and the distribution of subjects across clusters was also different by dysplasia status.

These results are seemingly incongruous though not incompatible, as they measure different parts of inflammation. As inflammation is complex in its composition and we were not exactly sure what to look for, it was important for us to use the most unbiased measure of

inflammation available to us instead of wholly relying on the clinical standards. This allowed us to come to the correct conclusions and raises questions regarding how inflammation is evaluated clinically. Perhaps a new system is necessary for the appropriate scoring of PSC inflammation. Or, perhaps clinicians and pathologists should consider a more unbiased and universal transcriptional system to evaluate inflammation- one that has a greater likely hood of capturing the relevant immunological components.

Selection of appropriate transcriptional read-outs

Only a limited amount of tissue can be collected from each colonoscopy, so careful allocation of the tissue for the appropriate assays can maximize the value of these precious samples. The selection of the best assays is not obvious, and we learned the effectiveness of different techniques through trial and error. In general, we found transcriptional analyses to be the most powerful because they allow for the most unbiased analysis of the sample. Whole tissue RNAseq was especially effective, because it served as a survey of the cell composition and activity of each tissue sample. Tissue RNAseq provided the information we needed to focus our investigation on the most relevant factors. One tissue biopsy provides more than enough RNA to perform RNAseq, and the DNA that is extracted at the same time can be used for genotyping, WES, and 16S sequencing. RNA and DNA extraction is therefore a very efficient use of a tissue sample. However, RNAseq results are not immediate, require the batching of many samples together, and can be very costly. Our dependence on RNAseq to determine the transcriptional cluster of each sample created a bottleneck in our analysis, requiring that we wait extended periods of time before being able to analyze large batches of our samples. Nonetheless, tissue

RNAseq was extremely effective at the onset of the study because it directed us to the correct hypotheses.

Bulk RNAseq was effective for surveying the tissue, however we found it ineffective when looking at adaptive cells. During our investigation, we attempted to analyze CD4 T-cells in two ways: RNAseq on bulk sorted CD4 T-cells and single cell RNAseq of the same population. We found that bulk sequencing did not provide fine enough resolution to identify changes in the composition of the CD4 T-cells. Single cell sequencing, on the other hand, allowed us to identify subsets of CD4 cells based on expression of cytokines and transcription factors. Single cell sequencing also allows the pairing of transcriptome with full-length TCR sequences. This is what allowed us to identify the "LA" motif enrichment in the DP cells and would not have been possible with bulk RNAseq.

Generalized approach to studying complex diseases

Were we to begin another investigation of a complex human disease, we would proceed according to the following general steps:

1. Thoroughly read available clinical and translational literature available on the disease.
2. Extensively consult clinicians and pathologists on their approaches to managing and diagnosis the disease and get a thorough understanding of the information that they collect.
3. Set up a framework of patient clinical data by probing medical records. Validate known findings in this local cohort, and explore other potential clinical factors associated with outcomes.

4. Perform a high-quality tissue RNAseq, controlling for factors such as method of sampling, location of sampling, and extent of disease. Allow for variability in potentially biologically relevant factors such as severity of disease, medications, age, and sex. Set up the appropriate control samples.
5. Analyze the transcriptome without pre-determined comparison groups. Observe the variability in transcriptional signatures, then try to explain the variance by integrating the appropriate clinical data.
6. Generate hypotheses and test them with multi-parameter assays such as immunohistochemistry, flow cytometry, and single cell RNAseq which allow for the testing of specific hypotheses while also collecting additional information that might be relevant.
7. Upon narrowing down particular cell types of interest, perform in-depth sequencing of this cell type to then generate information about the function of these cells in disease pathogenesis.

REFERENCES

1. Cooper, M. D. & Alder, M. N. The Evolution of Adaptive Immune Systems. *Cell* **124**, 815–822 (2006).

2. Buchmann, K. Evolution of Innate Immunity: Clues from Invertebrates via Fish to Mammals. *Front. Immunol.* **0**, 459 (2014).

3. Janeway, C. A. Approaching the Asymptote? Evolution and Revolution in Immunology. *Cold Spring Harb. Symp. Quant. Biol.* **54**, 1–13 (1989).

4. P, M. The danger model: a renewed sense of self. *Science* **296**, 301–305 (2002).

5. Charles A Janeway, J., Travers, P., Walport, M. & Shlomchik, M. J. *Immunobiology: The Immune System in Health and Disease*. (Garland Science, 2001).

6. Burnet, F. M. A modification of jerne's theory of antibody production using the concept of clonal selection. *CA. Cancer J. Clin.* **26**, 119–121 (1976).

7. Charles, E. D. *et al.* Clonal expansion of immunoglobulin M+CD27+ B cells in HCV-associated mixed cryoglobulinemia. *Blood* **111**, 1344–1356 (2008).

8. Yang, Y. *et al.* Focused specificity of intestinal T H 17 cells towards commensal bacterial antigens. *Nat. 2014 5107503* **510**, 152–156 (2014).

9. Nielsen, S. C. A. *et al.* Human B cell clonal expansion and convergent antibody responses to SARS CoV-2. *bioRxiv* (2020) doi:10.1101/2020.07.08.194456.

10. Minervina, A. A. *et al.* Longitudinal high-throughput tcr repertoire profiling reveals the dynamics of t-cell memory formation after mild covid-19 infection. *Elife* **10**, 1–17 (2021).

11. Aarnoudse, C. A., Krüse, M., Konopitzky, R., Brouwenstijn, N. & Schrier, P. I. TCR reconstitution in Jurkat reporter cells facilitates the identification of novel tumor antigens by cDNA expression cloning. *Int. J. Cancer* **99**, 7–13 (2002).

12. Wrammert, J. *et al.* Rapid cloning of high-affinity human monoclonal antibodies against influenza virus. *Nat. 2008 4537195* **453**, 667–671 (2008).

13. Hogquist, K. A., Baldwin, T. A. & Jameson, S. C. Central tolerance: learning self-control in the thymus. *Nat. Rev. Immunol. 2005 510* **5**, 772–782 (2005).

14. Moran, A. E. & Hogquist, K. A. T-cell receptor affinity in thymic development. *Immunology* **135**, 261–267 (2012).

15. Ramsdell, F. & Ziegler, S. F. FOXP3 and scurfy: how it all began. *Nat. Rev. Immunol. 2014 145* **14**, 343–349 (2014).

16. Anderson, M. S. *et al.* Projection of an Immunological Self Shadow Within the Thymus by the Aire Protein. *Science (80-.).* **298**, 1395–1401 (2002).

17. Chen, Y. *et al.* Peripheral deletion of antigen-reactive T cells in oral tolerance. *Nat. 1995 3766536* **376**, 177–180 (1995).

18. Lechler, R., Chai, J.-G., Marelli-Berg, F. & Lombardi, G. The contributions of T-cell anergy to peripheral T-cell tolerance. *Immunology* **103**, 262 (2001).

19. Sakaguchi, S., Wing, K. & Yamaguchi, T. Dynamics of peripheral tolerance and immune regulation mediated by Treg. *Eur. J. Immunol.* **39**, 2331–2336 (2009).

20. Nemazee, D. Mechanisms of central tolerance for B cells. *Nat. Rev. Immunol. 2017 175* **17**, 281–294 (2017).

21. Bennett, C. L. *et al.* The immune dysregulation, polyendocrinopathy, enteropathy, X-linked syndrome (IPEX) is caused by mutations of FOXP3. *Nat. Genet. 2001 271* **27**, 20–21 (2001).

22. Yurasov, S. *et al.* Defective B cell tolerance checkpoints in systemic lupus erythematosus. *J. Exp. Med.* **201**, 703–711 (2005).

23. Bouziat, R. *et al.* Reovirus infection triggers inflammatory responses to dietary antigens and development of celiac disease. *Science (80-.).* **356**, 44–50 (2017).

24. Thursby, E. & Juge, N. Introduction to the human gut microbiota. *Biochem. J.* **474**, 1823 (2017).

25. Capaldo, C. T., Powell, D. N. & Kalman, D. Layered defense: how mucus and tight junctions seal the intestinal barrier. *J. Mol. Med. 2017 959* **95**, 927–934 (2017).

26. Gutzeit, C., Magri, G. & Cerutti, A. Intestinal IgA production and its role in host-microbe interaction. *Immunol. Rev.* **260**, 76–85 (2014).

27. Muniz, L. R., Knosp, C. & Yeretssian, G. Intestinal antimicrobial peptides during homeostasis, infection, and disease. *Front. Immunol.* **0**, 310 (2012).

28. Mann, E. R. & Li, X. Intestinal antigen-presenting cells in mucosal immune homeostasis: Crosstalk between dendritic cells, macrophages and B-cells. *World J. Gastroenterol.* **20**, 9653 (2014).

29. Kuek, A., Hazleman, B. L. & Östör, A. J. K. Immune-mediated inflammatory diseases (IMIDs) and biologic therapy: a medical revolution. *Postgrad. Med. J.* **83**, 251 (2007).

30. Jabri, B. & Sollid, L. M. Mechanisms of Disease: immunopathogenesis of celiac disease. *Nat. Clin. Pract. Gastroenterol. Hepatol. 2006 39* **3**, 516–525 (2006).

31. DICKE, W. K., WEIJERS, H. A. & KAMER, J. H. v. D. Coeliac Disease The Presence in Wheat of a Factor Having a Deleterious Effect in Cases of Coeliac Disease. *Acta Pædiatrica* **42**, 34–42 (1953).

32. Tollefsen, S. *et al.* HLA-DQ2 and -DQ8 signatures of gluten T cell epitopes in celiac disease. *J. Clin. Invest.* **116**, 2226–2236 (2006).

33. Kim, C.-Y., Quarsten, H., Bergseng, E., Khosla, C. & Sollid, L. M. Structural basis for HLA-DQ2-mediated presentation of gluten epitopes in celiac disease. *Proc. Natl. Acad. Sci.* **101**, 4175–4179 (2004).

34. Qiao, S.-W. *et al.* Posttranslational Modification of Gluten Shapes TCR Usage in Celiac Disease. *J. Immunol.* **187**, 3064–3071 (2011).

35. Qiao, S.-W., Christophersen, A., Lundin, K. E. A. & Sollid, L. M. Biased usage and preferred pairing of α- and β-chains of TCRs specific for an immunodominant gluten epitope in coeliac disease. *Int. Immunol.* **26**, 13–19 (2014).

36. Steinsbø, Ø. *et al.* Restricted VH/VL usage and limited mutations in gluten-specific IgA of coeliac disease lesion plasma cells. *Nat. Commun. 2014 51* **5**, 1–12 (2014).

37. Osman, A. A. *et al.* B cell epitopes of gliadin. *Clin. Exp. Immunol.* **121**, 248–254 (2000).

38. Bateman, E. A. L. *et al.* IgA antibodies of coeliac disease patients recognise a dominant T cell epitope of A-gliadin. *Gut* **53**, 1274–1278 (2004).

39. Jiang, X. & Karlsen, T. H. Genetics of primary sclerosing cholangitis and pathophysiological implications. *Nat. Rev. Gastroenterol. Hepatol. 2017 145* **14**, 279–295 (2017).

40. Coussens, L. M. & Werb, Z. Inflammation and cancer. *Nat. 2002 4206917* **420**, 860–867 (2002).

41. Virchow, R. Die krankhaften Geschwülste. *Berlin: August Hirschwald* (1863).

42. Landén, N. X., Li, D. & Ståhle, M. Transition from inflammation to proliferation: a critical step during wound healing. *Cell. Mol. Life Sci. 2016 7320* **73**, 3861–3885 (2016).

43. Disson, O. *et al.* Peyer's patch myeloid cells infection by Listeria signals through gp38+ stromal cells and locks intestinal villus invasion. *J. Exp. Med.* **215**, 2936–2954 (2018).

44. Schultz, G., Clark, W. & Rotatori, D. S. EGF and TGF-α in wound healing and repair. *J. Cell. Biochem.* **45**, 346–352 (1991).

45. Haque, A. S. M. R. *et al.* CD206 + tumor-associated macrophages promote proliferation and invasion in oral squamous cell carcinoma via EGF production. *Sci. Reports 2019 91* **9**,

1–10 (2019).

46. Goswami, S. *et al.* Macrophages Promote the Invasion of Breast Carcinoma Cells via a Colony-Stimulating Factor-1/Epidermal Growth Factor Paracrine Loop. *Cancer Res.* **65**, 5278–5283 (2005).

47. Rous, P. & Kidd, J. G. CONDITIONAL NEOPLASMS AND SUBTHRESHOLD NEOPLASTIC STATES A STUDY OF THE TAR TUMORS OF RABBITS. *J. Exp. Med.* **73**, 365–390 (1941).

48. MacKenzie, I. & Rous, P. THE EXPERIMENTAL DISCLOSURE OF LATENT NEOPLASTIC CHANGES IN TARRED SKIN. *J. Exp. Med.* **73**, 391–416 (1941).

49. Pisani, P., Parkin, D. M., Muñoz, N. & Ferlay, J. Cancer and infection: estimates of the attributable fraction in 1990. *Cancer Epidemiol. Prev. Biomarkers* **6**, (1997).

50. Hagen, T. M. *et al.* Extensive oxidative DNA damage in hepatocytes of transgenic mice with chronic active hepatitis destined to develop hepatocellular carcinoma. *Proc. Natl. Acad. Sci.* **91**, 12808–12812 (1994).

51. TJ, M., BD, P. & LA, L. Mutagenic spectrum resulting from DNA damage by oxygen radicals. *Biochemistry* **30**, 207–213 (1991).

52. Sung, W.-K. *et al.* Genome-wide survey of recurrent HBV integration in hepatocellular carcinoma. *Nat. Genet. 2012 447* **44**, 765–769 (2012).

53. Werness, B., Levine, A. & Howley, P. Association of human papillomavirus types 16 and 18 E6 proteins with p53. *Science (80-.).* **248**, 76–79 (1990).

54. Dyson, N., Howley, P., Munger, K. & Harlow, E. The human papilloma virus-16 E7 oncoprotein is able to bind to the retinoblastoma gene product. *Science (80-.).* **243**, 934–937 (1989).

55. Mine, K. L. *et al.* Gene network reconstruction reveals cell cycle and antiviral genes as major drivers of cervical cancer. *Nat. Commun. 2013 41* **4**, 1–11 (2013).

56. Lee, Y.-M. & Kaplan, M. M. Primary Sclerosing Cholangitis. *N. Engl. J. Med.* **332**, 924–933 (2009).

57. Teefey, S. A., Baron, R. L., Rohrmann, C. A., Shuman, W. P. & Freeny, P. C. Sclerosing cholangitis: CT findings. *https://doi.org/10.1148/radiology.169.3.3055028* **169**, 635–639 (1988).

58. Wiesner, R. H. *et al.* Primary sclerosing cholangitis: Natural history, prognostic factors and survival analysis. *Hepatology* **10**, 430–436 (1989).

59. Farrant, J. M. *et al.* Natural history and prognostic variables in primary sclerosing cholangitis. *Gastroenterology* **100**, 1710–1717 (1991).

60. de Vries, E. M. G. *et al.* Alkaline phosphatase at diagnosis of primary sclerosing cholangitis and 1 year later: evaluation of prognostic value. *Liver Int.* **36**, 1867–1875 (2016).

61. MacCarty, R. L., LaRusso, N. F., Wiesner, R. H. & Ludwig, J. Primary sclerosing cholangitis: findings on cholangiography and pancreatography. *https://doi.org/10.1148/radiology.149.1.6412283* **149**, 39–44 (1983).

62. Fung, B. M. & Tabibian, J. H. Biliary endoscopy in the management of primary sclerosing cholangitis and its complications. *Liver Res.* **3**, 106 (2019).

63. Lindor, K. D., Kowdley, K. V. & Harrison, M. E. ACG clinical guideline: Primary sclerosing cholangitis. *Am. J. Gastroenterol.* **110**, 646–659 (2015).

64. Portmann, B. & Zen, Y. Inflammatory disease of the bile ducts–cholangiopathies: liver biopsy challenge and clinicopathological correlation. *Histopathology* **60**, 236–248 (2012).

65. Porayko, M. K. *et al.* Patients with asymptomatic primary sclerosing cholangitis frequently have progressive disease. *Gastroenterology* **98**, 1594–1602 (1990).

66. Boonstra, K. *et al.* Population-based epidemiology, malignancy risk, and outcome of primary sclerosing cholangitis. *Hepatology* **58**, 2045–2055 (2013).

67. Lewis, J. T., Talwalkar, J. A., Rosen, C. B., Smyrk, T. C. & Abraham, S. C. Prevalence and risk factors for gallbladder neoplasia in patients with primary sclerosing cholangitis: Evidence for a metaplasia-dysplasia-carcinoma sequence. *Am. J. Surg. Pathol.* **31**, 907–913 (2007).

68. Molodecky, N. A. *et al.* Incidence of primary sclerosing cholangitis: A systematic review and meta-analysis. *Hepatology* **53**, 1590–1599 (2011).

69. Ueda, Y. *et al.* Long-term Prognosis and Recurrence of Primary Sclerosing Cholangitis After Liver Transplantation: A Single-Center Experience. *Transplant. Direct* **3**, e334 (2017).

70. Bergquist, A. *et al.* Increased Risk of Primary Sclerosing Cholangitis and Ulcerative Colitis in First-Degree Relatives of Patients With Primary Sclerosing Cholangitis. *Clin. Gastroenterol. Hepatol.* **6**, 939–943 (2008).

71. Liu, J. Z. *et al.* Dense genotyping of immune-related disease regions identifies nine new risk loci for primary sclerosing cholangitis. *Nat. Genet. 2013 456* **45**, 670–675 (2013).

72. Ji, S.-G. *et al.* Genome-wide association study of primary sclerosing cholangitis identifies

new risk loci and quantifies the genetic relationship with inflammatory bowel disease. *Nat. Genet.* **49**, 269 (2017).

73. Ellinghaus, D. *et al.* Analysis of five chronic inflammatory diseases identifies 27 new associations and highlights disease-specific patterns at shared loci. *Nat. Genet. 2016 485* **48**, 510–518 (2016).

74. Bowlus, C. L., Li, C.-S., Karlsen, T. H., Lie, B. A. & Selmi, C. Primary sclerosing cholangitis in genetically diverse populations listed for liver transplantation: Unique clinical and human leukocyte antigen associations. *Liver Transplant.* **16**, 1324–1330 (2010).

75. Hov, J. R. *et al.* Electrostatic modifications of the human leukocyte antigen-DR P9 peptide-binding pocket and susceptibility to primary sclerosing cholangitis. *Hepatology* **53**, 1967–1976 (2011).

76. Melum, E. *et al.* Genome-wide association analysis in primary sclerosing cholangitis identifies two non-HLA susceptibility loci. *Nat. Genet. 2010 431* **43**, 17–19 (2010).

77. Folseraas, T. *et al.* Extended analysis of a genome-wide association study in primary sclerosing cholangitis detects multiple novel risk loci. *J. Hepatol.* **57**, 366–375 (2012).

78. Rühlemann, M. *et al.* Consistent alterations in faecal microbiomes of patients with primary sclerosing cholangitis independent of associated colitis. *Aliment. Pharmacol. Ther.* **50**, 580 (2019).

79. Sabino, J. *et al.* Primary sclerosing cholangitis is characterised by intestinal dysbiosis independent from IBD. *Gut* **65**, 1681–1689 (2016).

80. Vieira-Silva, S. *et al.* Quantitative microbiome profiling disentangles inflammation- and bile duct obstruction-associated microbiota alterations across PSC/IBD diagnoses. *Nat. Microbiol. 2019 411* **4**, 1826–1831 (2019).

81. Kummen, M. *et al.* The gut microbial profile in patients with primary sclerosing cholangitis is distinct from patients with ulcerative colitis without biliary disease and healthy controls. *Gut* **66**, 611–619 (2017).

82. De Chambrun, G. P. *et al.* Oral vancomycin induces sustained deep remission in adult patients with ulcerative colitis and primary sclerosing cholangitis. *Eur. J. Gastroenterol. Hepatol.* **30**, 1247–1252 (2018).

83. Davies, Y. K. *et al.* Long-term treatment of primary sclerosing cholangitis in children with oral vancomycin: An immunomodulating antibiotic. *J. Pediatr. Gastroenterol. Nutr.* **47**, 61–67 (2008).

84. Tabibian, J. H. *et al.* Randomised clinical trial: vancomycin or metronidazole in patients

with primary sclerosing cholangitis - a pilot study. *Aliment. Pharmacol. Ther.* **37**, 604–612 (2013).

85. Nakamoto, N. *et al.* Gut pathobionts underlie intestinal barrier dysfunction and liver T helper 17 cell immune response in primary sclerosing cholangitis. *Nat. Microbiol. 2019 43* **4**, 492–503 (2019).

86. Fausa, O., Schrumpf, E. & Elgjo, K. Relationship of Inflammatory Bowel Disease and Primary Sclerosing Cholangit is. *Semin. Liver Dis.* **1**, (1991).

87. Danese, S. & Fiocchi, C. Ulcerative Colitis. *http://dx.doi.org/10.1056/NEJMra1102942* **365**, 1713–1725 (2011).

88. Chang, J. T. Pathophysiology of Inflammatory Bowel Diseases. *https://doi.org/10.1056/NEJMra2002697* **383**, 2652–2664 (2020).

89. Geboes, K. What histologic features best differentiate Crohn's disease from ulcerative colitis? *Inflamm. Bowel Dis.* **14**, S168–S169 (2008).

90. Nikolaus, S. & Schreiber, S. Diagnostics of Inflammatory Bowel Disease. *Gastroenterology* **133**, 1670–1689 (2007).

91. Loftus, E. V *et al.* PSC-IBD: a unique form of inflammatory bowel disease associated with primary sclerosing cholangitis. *Gut* **54**, 91–96 (2005).

92. Joo, M. *et al.* Pathologic features of ulcerative colitis in patients with primary sclerosing cholangitis: A case-control study. *Am. J. Surg. Pathol.* **33**, 854–862 (2009).

93. Sano, H. *et al.* Clinical characteristics of inflammatory bowel disease associated with primary sclerosing cholangitis. *J. Hepatobiliary. Pancreat. Sci.* **18**, 154–161 (2011).

94. KK, J. *et al.* Inflammatory bowel disease in patients with primary sclerosing cholangitis: clinical characterization in liver transplanted and nontransplanted patients. *Inflamm. Bowel Dis.* **18**, 536–545 (2012).

95. Adams, D. H. & Eksteen, B. Aberrant homing of mucosal T cells and extra-intestinal manifestations of inflammatory bowel disease. *Nat. Rev. Immunol. 2006 63* **6**, 244–251 (2006).

96. Lichtman, S. N., Keku, J., Clark, R. L., Schwab, J. H. & Sartor, R. B. Biliary tract disease in rats with experimental small bowel bacterial overgrowth. *Hepatology* **13**, 766–772 (1991).

97. JK, M. *et al.* Rectal 5-aminosalicylic acid for induction of remission in ulcerative colitis. *Cochrane database Syst. Rev.* (2010) doi:10.1002/14651858.CD004115.PUB2.

98. Ko, C. W. *et al.* AGA Clinical Practice Guidelines on the Management of Mild-to-Moderate Ulcerative Colitis. *Gastroenterology* **156**, 748–764 (2019).

99. Feuerstein, J. D. *et al.* AGA Clinical Practice Guidelines on the Management of Moderate to Severe Ulcerative Colitis. *Gastroenterology* **158**, 1450–1461 (2020).

100. Feuerstein, J. D. *et al.* AGA Clinical Practice Guidelines on the Medical Management of Moderate to Severe Luminal and Perianal Fistulizing Crohn's Disease. *Gastroenterology* **160**, 2496–2508 (2021).

101. Cima, R. R. & Pemberton, J. H. Medical and Surgical Management of Chronic Ulcerative Colitis. *Arch. Surg.* **140**, 300–310 (2005).

102. Strong, S. A. Surgical management of Crohn's disease. *Surg. Treat. Evidence-Based Probl.* (2001).

103. Jess, T., Rungoe, C. & Peyrin-Biroulet, L. Risk of Colorectal Cancer in Patients With Ulcerative Colitis: A Meta-analysis of Population-Based Cohort Studies. *Clin. Gastroenterol. Hepatol.* **10**, 639–645 (2012).

104. Ekbom, A., Adami, H. O., Helmick, C. & Zack, M. Increased risk of large-bowel cancer in Crohn's disease with colonic involvement. *Lancet* **336**, 357–359 (1990).

105. Choi, P. M. & Zelig, M. P. Similarity of colorectal cancer in Crohn's disease and ulcerative colitis: implications for carcinogenesis and prevention. *Gut* **35**, 950–954 (1994).

106. Olén, O. *et al.* Colorectal cancer in ulcerative colitis: a Scandinavian population-based cohort study. *Lancet* **395**, 123–131 (2020).

107. Farraye, F. A., Odze, R. D., Eaden, J. & Itzkowitz, S. H. AGA Technical Review on the Diagnosis and Management of Colorectal Neoplasia in Inflammatory Bowel Disease. *Gastroenterology* **138**, 746-774.e4 (2010).

108. Shanahan, F., Weinstein, W. M. & Bernstein, C. N. Are we telling patients the truth about surveillance colonoscopy in ulcerative colitis? *Lancet* **343**, 71–74 (1994).

109. Clarke, W. T. & Feuerstein, J. D. Colorectal cancer surveillance in inflammatory bowel disease: Practice guidelines and recent developments. *World J. Gastroenterol.* **25**, 4148 (2019).

110. Eaden, J. A., Abrams, K. R. & Mayberry, J. F. The risk of colorectal cancer in ulcerative colitis: a meta-analysis. *Gut* **48**, 526–535 (2001).

111. A, E., C, H., M, Z. & HO, A. Ulcerative colitis and colorectal cancer. A population-based study. *N. Engl. J. Med.* **323**, 7–8 (1990).

112. Rutter, M. *et al.* Severity of inflammation is a risk factor for colorectal neoplasia in ulcerative colitis. *Gastroenterology* **126**, 451–459 (2004).

113. Soetikno, R. M., Lin, O. S., Heidenreich, P. A., Young, H. S. & Blackstone, M. O. Increased risk of colorectal neoplasia in patients with primary sclerosing cholangitis and ulcerative colitis: A meta-analysis. *Gastrointest. Endosc.* **56**, 48–54 (2002).

114. Broomé, U., Löfberg, R., Veress, B. & Eriksson, L. S. Primary sclerosing cholangitis and ulcerative colitis: Evidence for increased neoplastic potential. *Hepatology* **22**, 1404–1408 (1995).

115. Chapman, R. *et al.* Diagnosis and management of primary sclerosing cholangitis. *Hepatology* **51**, 660–678 (2010).

116. K, S., L, R., A, B., WD, C. & BA, L. The risk for cancer or dysplasia in ulcerative colitis patients with primary sclerosing cholangitis. *Am. J. Gastroenterol.* **94**, 1643–1649 (1999).

117. MM, C. *et al.* More right-sided IBD-associated colorectal cancer in patients with primary sclerosing cholangitis. *Inflamm. Bowel Dis.* **15**, 1331–1336 (2009).

118. Beaugerie, L. & Itzkowitz, S. H. Cancers Complicating Inflammatory Bowel Disease. *http://dx.doi.org/10.1056/NEJMra1403718* **372**, 1441–1452 (2015).

119. Itzkowitz, S. H. Molecular Biology of Dysplasia and Cancer in Inflammatory Bowel Disease. *Gastroenterol. Clin.* **35**, 553–571 (2006).

120. Sansregret, L., Vanhaesebroeck, B. & Swanton, C. Determinants and clinical implications of chromosomal instability in cancer. *Nat. Rev. Clin. Oncol. 2018 153* **15**, 139–150 (2018).

121. Li, K., Luo, H., Huang, L., Luo, H. & Zhu, X. Microsatellite instability: a review of what the oncologist should know. *Cancer Cell Int. 2020 201* **20**, 1–13 (2020).

122. Lower, S. S., McGurk, M. P., Clark, A. G. & Barbash, D. A. Satellite DNA evolution: old ideas, new approaches. *Curr. Opin. Genet. Dev.* **49**, 70–78 (2018).

123. Rhodes, J. M. & Campbell, B. J. Inflammation and colorectal cancer: IBD-associated and sporadic cancer compared. *Trends Mol. Med.* **8**, 10–16 (2002).

124. Kiesslich, R. *et al.* Methylene blue-aided chromoendoscopy for the detection of intraepithelial neoplasia and colon cancer in ulcerative colitis. *Gastroenterology* **124**, 880–888 (2003).

125. Han, Y. D. *et al.* Prognosis of ulcerative colitis colorectal cancer vs. sporadic colorectal cancer: propensity score matching analysis. *BMC Surg. 2017 171* **17**, 1–6 (2017).

126. Baker, K. T., Salk, J. J., Brentnall, T. A. & Risques, R. A. Precancer in ulcerative colitis: the role of the field effect and itsclinical implications. *Carcinogenesis* **39**, 11 (2018).

127. Soh, J. S. *et al.* Immunoprofiling of Colitis-associated and Sporadic Colorectal Cancer and its Clinical Significance. *Sci. Reports 2019 91* **9**, 1–10 (2019).

128. M, B. *et al.* GenPipes: an open-source framework for distributed and scalable genomic analyses. *Gigascience* **8**, (2019).

129. AM, B., M, L. & B, U. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* **30**, 2114–2120 (2014).

130. A, D. *et al.* STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**, 15–21 (2013).

131. S, A., PT, P. & W, H. HTSeq--a Python framework to work with high-throughput sequencing data. *Bioinformatics* **31**, 166–169 (2015).

132. Ritchie, M. E. *et al.* limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res.* **43**, e47–e47 (2015).

133. Wu, T. *et al.* clusterProfiler 4.0: A universal enrichment tool for interpreting omics data. *Innov.* **2**, 100141 (2021).

134. Hao, Y. *et al.* Integrated analysis of multimodal single-cell data. *Cell* **184**, 3573-3587.e29 (2021).

135. Brochet, X., Lefranc, M.-P. & Giudicelli, V. IMGT/V-QUEST: the highly customized and integrated system for IG and TR standardized V-J and V-D-J sequence analysis. *Nucleic Acids Res.* **36**, W503–W508 (2008).

136. Ralph, D. K. & IV, F. A. M. Consistency of VDJ Rearrangement and Substitution Parameters Enables Accurate B Cell Receptor Sequence Annotation. *PLOS Comput. Biol.* **12**, e1004409 (2016).

137. Ralph, D. K. & IV, F. A. M. Likelihood-Based Inference of B Cell Clonal Families. *PLOS Comput. Biol.* **12**, e1005086 (2016).

138. Hoehn, K. B., Lunter, G. & Pybus, O. G. A Phylogenetic Codon Substitution Model for Antibody Lineages. *Genetics* **206**, 417–427 (2017).

139. Hoehn, K. B. *et al.* Repertoire-wide phylogenetic models of B cell molecular evolution reveal evolutionary signatures of aging and vaccination. *Proc. Natl. Acad. Sci.* **116**, 22664–22672 (2019).

140. Hafemeister, C. & Satija, R. Normalization and variance stabilization of single-cell RNA-seq data using regularized negative binomial regression. *Genome Biol. 2019 201* **20**, 1–15

(2019).

141. Lun, A. T. L., McCarthy, D. J. & Marioni, J. C. A step-by-step workflow for low-level analysis of single-cell RNA-seq data with Bioconductor. *F1000Research 2016 52122* **5**, 2122 (2016).

142. Robinson, M. D., McCarthy, D. J. & Smyth, G. K. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* **26**, 139–140 (2010).

143. Law, C. W., Chen, Y., Shi, W. & Smyth, G. K. voom: precision weights unlock linear model analysis tools for RNA-seq read counts. *Genome Biol. 2014 152* **15**, 1–17 (2014).

144. Bailey, T. L. STREME: accurate and versatile sequence motif discovery. *Bioinformatics* (2021) doi:10.1093/BIOINFORMATICS/BTAB203.

145. Barlow, J. T., Bogatyrev, S. R. & Ismagilov, R. F. A quantitative sequencing framework for absolute abundance measurements of mucosal and lumenal microbial communities. *Nat. Commun. 2020 111* **11**, 1–13 (2020).

146. Benjamin, D. *et al.* Calling Somatic SNVs and Indels with Mutect2. *bioRxiv* 861054 (2019) doi:10.1101/861054.

147. Donaldson, G. P., Lee, S. M. & Mazmanian, S. K. Gut biogeography of the bacterial microbiota. *Nat. Rev. Microbiol. 2015 141* **14**, 20–32 (2015).

148. James, K. R. *et al.* Distinct microbial and immune niches of the human colon. *Nat. Immunol. 2020 213* **21**, 343–353 (2020).

149. Calderó, J. *et al.* Regional distribution of glycoconjugates in normal, transitional and neoplastic human colonic mucosa. *Virchows Arch. A 1989 4154* **415**, 347–356 (1989).

150. Bressenot, A. M. *et al.* Review article: the histological assessment of disease activity in ulcerative colitis. *Aliment. Pharmacol. Ther.* **42**, 957–967 (2015).

151. Landsverk, O. J. B. *et al.* Antibody-secreting plasma cells persist for decades in human intestine. *J. Exp. Med.* **214**, 309 (2017).

152. Spencer, J. & Sollid, L. M. The human intestinal B-cell response. *Mucosal Immunol. 2016 95* **9**, 1113–1124 (2016).

153. Kett, K., Rognum, T. O. & Brandtzaeg, P. Mucosal subclass distribution of immunoglobulin G-producing cells is different in ulcerative colitis and Crohn's disease of the colon. *Gastroenterology* **93**, 919–924 (1987).

154. Keerthivasan, S. *et al.* -Catenin Promotes Colitis and Colon Cancer Through Imprinting

of Proinflammatory Properties in T Cells. *Sci. Transl. Med.* **6**, 225ra28-225ra28 (2014).

155. Grishkan, I. V., Ntranos, A., Calabresi, P. A. & Gocke, A. R. Helper T cells down-regulate CD4 expression upon chronic stimulation giving rise to double-negative T cells. *Cell. Immunol.* **284**, 68 (2013).

156. Lee, Y. *et al.* Induction and molecular signature of pathogenic TH17 cells. *Nat. Immunol. 2012 1310* **13**, 991–999 (2012).

157. Dejea, C. M. *et al.* Patients with familial adenomatous polyposis harbor colonic biofilms containing tumorigenic bacteria. *Science (80-.).* **359**, 592–597 (2018).

158. Kostic, A. D. *et al.* Fusobacterium nucleatum Potentiates Intestinal Tumorigenesis and Modulates the Tumor-Immune Microenvironment. *Cell Host Microbe* **14**, 207–215 (2013).

159. Dziubańska-Kusibab, P. J. *et al.* Colibactin DNA-damage signature indicates mutational impact in colorectal cancer. *Nat. Med. 2020 267* **26**, 1063–1069 (2020).

160. Smit, J. J. M. *et al.* Homozygous disruption of the murine MDR2 P-glycoprotein gene leads to a complete absence of phospholipid from bile and to liver disease. *Cell* **75**, 451–462 (1993).

161. Battista, K. D. *et al.* A Novel Murine Model of Primary Sclerosing Cholangitis Associated Inflammatory Bowel Disease. *FASEB J.* **31**, 469.1-469.1.