Chapter 1

Introduction

One in 10,000 people in Western countries lives with Primary Sclerosing Cholangitis (PSC), an immune-mediated, inflammatory disease of the bile ducts. PSC is a rare disease, which confers risk of serious disease sequelae including hepatobiliary malignancy and progression to end-stage liver failure, where the only treatment option is liver transplantation. Inflammatory bowel disease (IBD) is highly co-morbid, present in up to 80% of patients with PSC. Patients with PSC and IBD also have a high risk of developing colorectal cancer. The absence of effective medical therapy for PSC reflects our current limited understanding of disease aetiology and pathogenesis. Over the past decade several genome-wide association studies (GWAS) have investigated the genetic architecture of PSC, identifying genetic variants associated with disease susceptibility. Whilst it was anticipated that these findings would translate into further biological understanding of PSC pathogenesis and the identification of potential therapeutic drug targets, progress has been slow. This thesis will explore the biological significance of genetic risk variants associated with PSC susceptibility and the genes and cell types they perturb, with the aim of identifying potential future therapeutic targets. This introductory chapter will provide an overview of our current understanding of the genetic and biological architecture of PSC, and the associated challenges for the functional follow-up of genetic risk loci associated with rare complex diseases, such as PSC.

1.1 What is Primary Sclerosing Cholangitis?

PSC is chronic progressive fibro-obliterative disease of the intra- and extra-hepatic bile ducts of the liver. It is characterised by recurrent biliary inflammation leading to a progressive, diffuse, multi-focal, biliary stricturing and fibrosis. Eventually this can progress to complete obliteration of small bile ducts and resultant cholestasis. Common symptoms can range from fatigue to the sequelae of cholestasis such as jaundice, pruritus and recurrent cholangitis. Progressive fibrosis and cirrhosis of the hepatic parenchyma can

result in end-stage liver failure, with up to 20% of patients requiring liver transplantation within 10 years of diagnosis [1, 2]. High transplant rates are further precipitated by the lack of any effective medical treatments that can attenuate or halt the progression of this debilitating disease. Even liver transplantation itself does not always offer a cure, with more than 20% of patients experiencing recurrent disease in their transplant graft [3]. In addition, inflammation-associated biliary dysplasia results in a greatly increased risk of cholangiocarcinoma and gallbladder cancer and thus PSC confers a 15% lifetime risk of developing hepatobiliary or colorectal malignancy [4, 5]. Therefore, although it is a rare disease, PSC places a disproportionately high burden on gastroenterology, oncological and transplant services, remaining the 5th most common indication for liver transplantation across the UK and Europe [6, 7].

PSC is strongly associated with inflammatory bowel disease (IBD), most commonly ulcerative colitis (UC), which co-exists in 60-80% of PSC patients. The clinical phenotype of PSC-associated IBD (PSC-IBD) is distinct from that of lone IBD, more commonly affecting the right side of the colon with rectal sparing [8–10]. Despite its milder inflammatory phenotype, PSC-associated IBD carries a significantly higher risk of colonic malignancy, which is ten-fold that of the general population [4, 11]. Furthermore, the lesser common PSC-associated with Crohns disease, has been associated with a lower risk of liver transplant, death and malignancy compared to PSC associated with UC [12]. There are several other clinical subtypes of PSC which have been demonstrated to confer different prognoses compared to 'classical' PSC. The first is small-duct PSC, defined by the absence of cholangiographic evidence of PSC in the presence PSC affecting the small bile ducts on histological examination. Patients with small-duct PSC demonstrate improved survival (6% versus 34%) and lower risk of cholangiocarcinoma (0% versus 11%) [13]. Conversely, the second clinical subtype, PSC with an elevated IgG4 concentration, has been associated with an increased risk of progression to cirrhosis; 50% versus 12% of those without [14, 15].

Despite trials of multiple therapeutic agents in PSC, none have proven successful and there is absence of any effective medical therapies which can either cure or attenuate disease progression in PSC. Perhaps the most widely trialed therapeutic agent in PSC is Ursodeoxycholic acid (UDCA), given its proven efficacy in the treatment of other cholestatic diseases such as PBC. UDCA is postulated to have two mechanisms of actions; reducing hydrophobicity of bile and a direct effect on adaptive immunity by inhibiting dendritic cell response [16]. Since 1990, at least twelve trials, nine of which were randomised and placebo-controlled, have studied the effect of varying doses of UDCA on liver biochemistry [17]. Whilst most observed an improvement in liver biochemistry, there was no demonstrable effect on time to transplantation or liver-related death. These trials are notable due to their small numbers of patients in each study arm, and their short duration compared to the natural history of PSC. Despite these findings, UDCA remains widely prescribed for PSC.

Various immunosuppressive drugs have been trialled in PSC, including placebo-controlled trials of ciclosporin and methotrexate, and uncontrolled trials of steroids, azathioprine and tacrolimus [17]. Whilst again these trials remained limited by sample size and duration, they still failed to identify any effect on hepatic cholangiography, transplant or survival. Moreover, there has been limited enthusiasm for further trialling of immunosuppressive agents, given that the commonality of PSC-IBD means that many patients with PSC are taking immunosuppressants at the time of PSC diagnosis and progression.

1.2 PSC is a complex disease

Complex diseases result from a complex interplay between genetic and environmental factors, most of which remain unidentified. Monogenic diseases results from a rare variant that exerts a large, usually qualitative effect upon a single gene resulting in a disease phenotype. In contrast, the genetic component of a complex disease is driven by multiple variants with modest to small effect sizes, acting in a predominantly additive fashion [18, 19]. Familial studies support both an environmental and genetic aetiology for PSC, with familial and geographic clustering of cases, particularly in Northern Europe where prevalence estimates are as high as 16/100,000 [20]. In comparison, prevalence estimates in Asia are as low as 4/100,000 [21, 22]. Whilst one of the best means of quantifying the genetic and environmental influences on complex disease is by the comparison of disease concordance in monozygotic versus dizygotic twins, there are currently no published twin studies in PSC. Familial clustering of disease provides another means to estimate the level of genetic influence on complex disease. The relative risk ratio of a sibling (λs) is the risk of disease development in the siblings of an affected individual and is calculated as the prevalence of a complex disease among siblings divided by the prevalence of the disease in the population at large. Disease prevalence in the first-degree relatives of PSC-affected patients is significantly increased compared to that of the general population with a λs of approximately 100 [23, 24]. Despite being a rare disease with observed familial clustering, PSC does not display a classical Mendelian inheritance pattern, and is considered to be a complex disease driven by multiple dynamic gene-environment and gene-gene interactions, acting in concert to cause the PSC phenotype. Environmental factors that have been implicated include a protective effect from coffee consumption (OR=0.52, p=0.006) and smoking (OR=0.33, $p<0.001$) on the development of PSC [25].

1.3 Genome-wide association studies

Genome-wide association studies (GWAS) are the standard study design for testing the association of genetic variants throughout the genome with the presence of a complex

trait. Disease-associated variants are those for which one allele occurs significantly more frequently in cases compared with controls. These variants mark regions of the genome associated with the trait and are called 'risk loci'. The GWAS design was facilitated by an important biological observation; linkage disequilibrium (LD), the non-random association of alleles between nearby genetic variants [26]. Patterns of LD between nearby genetic variants allow the capture of most of the common variation within the human genome by assaying just a small subset of single-nucleotide polymorphisms (SNPs) [27]. Approximately five million common SNPs with a minor allele frequency (MAF) <0.5, could be well-tagged (r^2>0.8) using a subset of just 500,000 SNPs, in East Asian and European populations [28]. Because LD patterns vary by population, the International Hapmap Project was set up to map the patterns of LD across several populations, providing the foundation for GWAS [29]. Thus, one could perform a GWAS by genotyping just a small subset of variants across the genome followed by the imputation of non-genotyped variants using the LD structure from reference panels, dramatically reducing the costs of genotyping and increasing the economic scalability of GWAS. To account for the hundreds of thousands of genetic variants tested in a GWAS, the genome-wide significance threshold is set, by convention, at $p<5\times10^{-8}$ to account for multiple testing and to avoid type I (false-positive) statistical errors [30]. In order to achieve sufficient statistical power, GWAS therefore requires thousands of cases and controls. The amassing of increasingly large sample sizes to improve statistical power was facilitated by a second biological observation. The results from several early GWAS of immune-mediated diseases (IMDs) led to the observation that many genetic associations were shared across multiple IMDs [31, 32]. This facilitated the development of the Immunochip, a targeted genotyping array with dense coverage across approximately 130,000 SNPs within 186 known risk loci, from twelve immune-mediated diseases. Similarly, genetic architecture was shared across many metabolic disorders resulting in the development of the Metobochip, which was designed for studying metabolic and cardiovascular disease [33]. These chips provided a cost-effective means of identifying common and rare variants associated with complex traits, at a fraction of the cost of a GWAS chip. This allowed the genotyping of increasingly large samples sizes, although at the cost of being unable to identify rare variants and variants outside of the predefined regions included in the chip [34]. 'Common' variants, those occurring at a frequency of >5% in the general population, typically have low to moderate effect upon complex traits, with odds ratios (OR) of up to 1.4. 'Rare' variants (those with a MAF<1%) with larger effect sizes are less likely to be represented by genotyping chips and reference panels, due to the fact that variants with large effect on disease risk are likely to be kept at low frequency due to negative selection pressure [35, 36]. The identification of rare variants associated with complex traits through GWAS has been made possible over the past decade by improvement in LD reference panels, an exponential reduction

in the cost of GWAS and the development of collaborative research consortia for the meta-analysis of GWAS data from increasingly large sample sizes.

PSC research has derived significant benefit from the genetics revolution. To date, at least six studies have examined the effects of genetic variation on PSC susceptibility [37–42]. Although PSC was not included in the original design of the Immunochip, in 2013 an Immunochip study of 3,789 PSC cases of European ancestry and 25,079 population controls identified twelve genome-wide significant associations outside the human leukocyte antigen (HLA) complex, nine of which were new [40]. This was further improved in 2017, when the largest PSC GWAS to date, including 4,796 cases and 19,955 population controls, identified fifteen regions of the genome associated with PSC susceptibility, four of which were new. To date, we have identified a total of 23 regions of the genome associated with susceptibility to PSC.

1.4 Genetic associations within the human leucocyte antigen

In keeping with findings in many other IMDs, the strongest genetic associations with PSC have been observed within the highly polymorphic HLA region, supporting an important role for the adaptive immune system in the pathogenesis of PSC. The HLA gene complex is an ∼7 million base pair (bp) region of DNA on the short arm of chromosome 6, encoding more than 250 genes, of which approximately a third relate to immune cell function. The HLA plays an essential role in the tuning of the adaptive immune system, encoding cell-surface proteins responsible for the regulation and presentation of foreign antigens to T-cells. Many IMDs have been associated with particular HLA SNPs or haplotypes, suggesting the involvement of disease-specific antigens. However, for many diseases, including PSC, the causative antigens remain unidentified [43]. Despite the evidence supporting a strong effect of the HLA on PSC susceptibility, dissecting this region into the underlying specific genes that confer disease risk is challenging due to several characteristics of the HLA region. These include high levels of variation and strong patterns of LD, extending up to several thousand kilobases, in addition to the presence of multiple genes of potential relevance with roles in antigen presentation and immune-regulation within this particularly gene-dense region [26].

Class I HLAs present intracellular foreign antigens to CD8+ T-cells, inducing CD8+ mediated cytotoxicity. The strongest observed effect on PSC susceptibility is with the class I HLA haplotype, HLA-B*08:10 [40]. Liu *et al* identified rs4143332 as the top associated SNP with PSC risk, which was in almost perfect LD (r^2=0.996) with HLA-B*08:01. Step-wise conditional analysis containing both SNP and HLA allele genotypes identified rs4143332 as SNP most associated with PSC risk, both within and outside of the HLA

(OR=3.00, p=3.7×10^{-246}). Other associations have also been observed adjacent to class II HLA haplotypes HLA-DRB1*03:01, 04:01, 07:01 and 13:01, [44, 45]. Class II HLAs present extracellular antigens to CD4+ T-helper cells which stimulate antibody-producing B-cells. Liu *et al* identified a complex HLA class II association signal determined by HLA-DQA1*01:03 and SNPs rs532098, rs1794282 and rs9263964, which along with rs4143332 (tagging HLA-B*08:01), explained most of the HLA association signal in PSC.

1.5 Genetic associations outside of the HLA

Genetic associations outside of the HLA further support the role of immune dysregulation in PSC pathogenesis, as the majority occur within, or close to genes involved in immune-cell function. Notably, several of the risk loci with the greatest effect on PSC susceptibility may affect genes involved in T-effector and T-regulatory cell signalling pathways, including *CD28*, *MST1*, *IL2* and *IL2RA* (Figure 1.1). Furthermore, polymorphisms in these genomic regions are also associated with a number of other IMDs including type I Diabetes (T1DM), rheumatoid arthritis (RhA), multiple sclerosis (MS) and systemic lupus erythematosus (SLE) [46–49]. Disease risk loci outside of the HLA provide an important anchor for unravelling some of the potential pathogenic mechanisms involved in PSC. Evidence supporting the potential pathogenic contribution of some of these risk loci to PSC disease biology is discussed below.

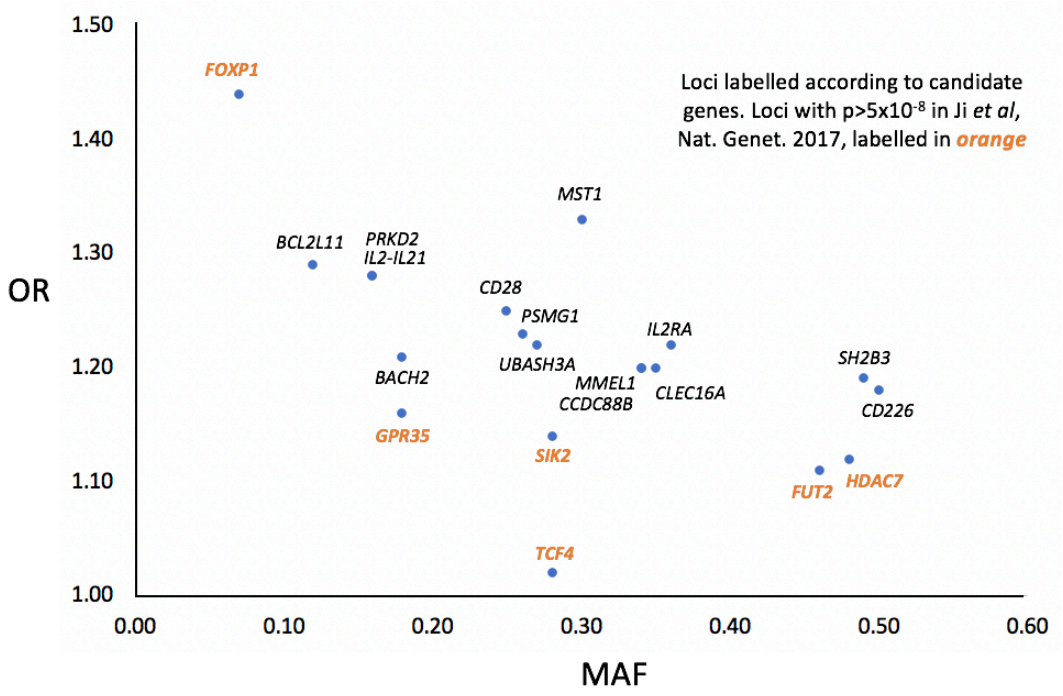

Figure 1.1: Twenty of the twenty-two non-HLA PSC risk loci plotted according to their effect size (OR) and MAF in Ji *et al*'s GWAS data [42].

1.5.1 PSC risk loci in coding regions of the genome

Of the 22 PSC risk loci outside of the HLA region, only four have lead SNPs within coding regions of the genome. The genetic risk locus on chromosome 3 has one of the strongest effects on PSC risk (OR=1.33, CI=1.26-1.40). The lead SNP for this signal, rs3197999, is a missense variant in *MST-1*. *MST-1* encodes macrophage stimulating protein (MSP), which is expressed primarily in the liver by biliary epithelial cells. It functions as part of Hippo pathways that regulate tumour suppression, and deletion of *MST-1* in hepatocytes results in excessive proliferation and hepatomegaly [50]. *MST-1* is known to have a role in cellular immunity, modulating integrin- and selectin-mediated lymphocyte migration and chemotaxis in lymphoid tissues [51]. Moreover, autosomal recessive *MST-1* deficiency is an identified cause of combined immunodeficiency [52]. Taken together, this suggests that *MST-1* may play an important role in homing of lymphocytes between the gut and the liver, supporting one of the most common hypotheses of disease pathogenesis in PSC, the 'gut-homing T-cell' hypothesis [53]. Associations with the *MST-1* region have also been reported in UC and CD, and a study of the lead variant, rs3197999, suggests that the risk allele is associated with a gain of function and increased stimulatory effect of MSP on chemotaxis and proliferation in a monocyte THP-1 cell line [54]. Whilst it is more common for risk variants with the largest effect sizes to occur within protein coding

regions, it is important to note that LD in this region extends over a large number of theoretically relevant genes [55]. Indeed, an IBD fine-mapping study of the *MST-1* locus identified a credible set of 437 SNPs explaining >95% of the variation within this region, any one of which could be the true causal variant [56].

Another PSC risk locus in a coding region is on chromosome 2 within *GPR35* (G-protein-coupled receptor 35). The lead variant in this region, rs3749171, is a missense variant, located in the 3' exon of *GPR35*. Structural modelling suggests that the residue affected by this threonine to methionine substitution is found in the third trans-membrane helix and is predicted to effect the efficiency of signalling through the GPR35 receptor [39]. GPR35 is expressed in high levels in the gastrointestinal tract, predominantly by intestinal crypt enterocytes and sub-populations of immune cells [57]. It functions as a receptor for kynurenic acid, an intermediate in the tryptophan metabolic pathway, which is found at high concentrations in bile and intestinal fluid, and increases during inflammation [58, 59]. Furthermore, variation in this gene is also associated with IBD risk [60] and increased levels of plasma kynurenic acid have been reported in patients with IBD [61]. GPR35 has also been shown to promote the activity of Na/K-ATPase, with the PSC-associated lead variant, rs3749171, inducing a more pronounced increase in Na/K-ATPase activity, enhancing glycolysis and proliferation in intestinal epithelial cells [62]. Whilst genetic associations with the *GPR35* region have been robustly replicated across multiple IBD GWAS, associations with PSC have not been consistently reported across all PSC GWAS [41]. This includes the most recent and largest PSC GWAS to date, where associations with this region failed to reach genome-wide significance [42].

1.5.2 PSC risk loci in non-coding regions of the genome

The vast majority of risk loci associated with IMDs, of which PSC is no exception, occur in non-coding regions of the genome and are presumed to exert regulatory effects upon nearby genes. In the absence of a proven association been gene and locus, candidate genes have been historically assigned to non-coding risk loci according to a combination of their genomic proximity to the lead variant and existing knowledge of a genes biological function. The non-coding PSC risk locus on chromosome 2 occurs 3' downstream of *CD28* (OR=1.25, 95% CI=1.19-1.32) and has been implicated by genetic association with several other IMDs, including MS and RhA [47, 63]. Due to its role in T-cell signalling, *CD28* has been highlighted as a candidate gene for this locus. This gene encodes a co-stimulatory protein on T-cells necessary for activation and proliferation. Co-stimulation through CD28 and the T-cell receptor (TCR) induces the production of multiple interleukins. These include IL-2, a cytokine with a dual role in both the activation of the inflammatory immune response via T-effector cells, and suppression of the inflammatory immune response via T-regulatory cells. Interestingly, in PSC a greater proportion of CD4+ and CD8+ liver-infiltrating

T-cells are CD28- in comparison with controls without liver disease (30.3% vs 2.5% for CD4+ and 68.5% vs 31.9% in CD8+) as well as controls with other forms of liver disease including primary biliary cirrhosis (PBC) and non-alcoholic steatohepatitis (NASH) [64]. These CD28- cells are induced by TNFα and infiltrate the peri-biliary region where they secrete pro-inflammatory cytokines resulting in apoptosis of biliary epithelial cells.

Interleukin-2 receptor alpha (IL2RA), also known as CD25, is constitutively expressed by T-regulatory cells (T-regs). It binds IL-2 to promote the survival and proliferation of T-regs, thus promoting an anti-inflammatory and immune-suppressive response. Both the *IL-2* (OR 1.33, 95% CI=1.26-1.40) and *IL2RA* (OR=1.22, 95% CI=1.16-1.28) genes have been implicated in PSC by genetic associations in non-coding regions on chromosomes 4 and 10 respectively. *IL2RA* knock-out mice develop a phenotype similar to PSC with the spontaneous development of T-cell mediated biliary inflammation and colitis [65]. However evidence is not just restricted to mice and activated liver-derived T-lymphocytes of PSC patients demonstrate reduced expression of the IL-2 receptor and impaired proliferative response and functional capacity in comparison with patients with PBC, autoimmune hepatitis (AIH) or healthy controls [66]. Furthermore, a link between homozygosity for polymorphisms in the *IL2RA* gene and reduced numbers of FOXP3+ T-regs has been demonstrated in the peripheral blood of patients with PSC [67]. Collectively, the *CD28*, and *IL-2* and *IL2RA* risk loci may support an important role for defects in T-regulatory pathways in the pathogenesis of PSC.

Non-coding genetic associations within the introns of *SIK2, HDAC7* and *PRKD2* (on chromosomes 11, 12 and 19 respectively) highlight the potential pathogenic importance of T-cell selection in PSC pathogenesis [40]. Negative selection of immature T-cells within the thymus is essential for the development and maintenance of tolerogenic response, the disruption of which facilitates the development of IMDs. SIK2 (salt-inducible kinase 2) regulates the expression of both IL-10 in macrophages, and leukocyte transcription factor, Nur77 [68, 69]. Following engagement of the thymocyte TCR, PRKD2 (serine-threonine protein kinase D2) phosphorylates HDAC7, leading to loss of HDAC7-mediated repression of Nur77 (regulated by SIK2) [70]. This results in nuclear exclusion of HDAC7 and loss of HDAC7's regulatory functions, ultimately resulting in apoptosis and negative selection of immature T-cells [71]. Notably, *HDAC7* has also been implicated by genetic association, with IBD [72], although genetic associations in both the *HDAC7* and *SIK2* regions fell short of genome-wide significance in the most recent and most well-powered PSC GWAS [42].

1.5.3 Current genetic understanding of PSC subtypes

The association between PSC and IBD provides an important opportunity for further understanding of the genetics of PSC. The increased commonality of IBD means that

GWAS of IBD dwarf those of PSC, both in terms of number and sample sizes [60, 73]. Consequently, there have been 240 regions of the genome associated with risk of IBD, compared to just twenty-three in PSC. Unfortunately, the absence of phenotype data identifying those IBD cases with concomitant PSC has limited our insights into PSC genetic risk from published IBD GWAS.

Despite the presence of coexistent IBD in up to 80% of PSC patients, most of the HLA associations with PSC are distinct from those with IBD. The exception is HLA-DRB1*15:01 which is associated with increased risk of PSC, increased risk of UC, and decreased risk of Crohn's disease (CD) [74]. In the largest PSC GWAS to date, Ji *et al* performed Bayesian tests of colocalisation between IBD and PSC GWAS summary statistics to identify fourteen non-HLA loci with strong evidence of shared causal variants between PSC and IBD [42]. Six of the fourteen non-HLA loci associated with both PSC and IBD displayed strong evidence of a shared causal variant with UC, CD or both (*MST1, IL21, HDAC7, SH2B3, CD226* and *PSMG1*), demonstrating an important degree of shared genetic variation between both diseases (Figure 1.2). However, four of the same fourteen loci demonstrated strong evidence that the causal variant was independent from that in UC and CD (*IL2RA, CCDC88B, CLEC16A* and *PRKD2*). This demonstrates that even between highly co-morbid diseases, significant associations in the same genomic regions will not always share the same causal variant.

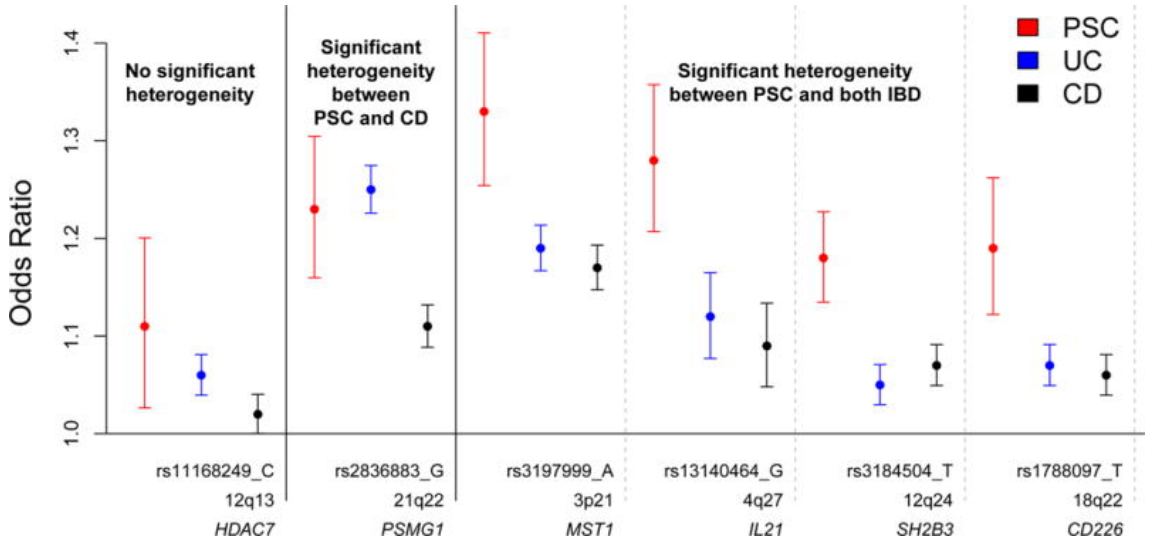

Figure 1.2: Figure taken from Ji *et al* demonstrating odds ratios (and their 95% confidence intervals) for PSC, UC and CD across the 6 PSC associated SNPs demonstrating strong evidence for a shared causal variant (maximum posterior probability >0.8) [42].

Genetic studies of the PSC sub-phenotypes, small-duct PSC and PSC with raised IgG4, have been significantly impeded by the low prevalence of PSC and the lack of power resulting from the further subdivision of already small cohorts of patients with PSC. As a result, genetic studies of PSC sub-phenotypes have, to date, focused solely on associations within the HLA. HLA-DRB1*15:01 is present at increased frequency in patients with PSC and high IgG4 levels, the sub-phenotype of PSC reportedly associated with increased risk of progression to cirrhosis [14, 75]. Interestingly, as mentioned above, this haplotype is also associated with increased risk of UC, which in turn has been associated with increased severity of PSC. Patients with small-duct PSC without concomitant IBD, the sub-phenotype which confers an improved survival and lower risk of cholangiocarcinoma, also demonstrate distinct HLA associations [13]. Small-duct PSC without IBD is associated only with HLA-DRB1*13:01 and is otherwise distinctly different from large-duct PSC with IBD in terms of its HLA associations [76]. This may support the hypothesis that small-duct PSC without IBD is a distinct clinical entity from large-duct PSC. However, these results must be interpreted with caution as this study analysed genotype data for just four classical HLA loci in only 87 small-duct and 485 large-duct PSC patients compared with 1117 controls. As both sample sizes and depth of phenotype data in PSC research increases, future studies will be able to further delineate the distinct and overlapping genetic architecture of PSC sub-phenotypes.

1.6 Current hypotheses of disease pathogenesis in PSC

In order to identify potential proteins and biological pathways for therapeutic target, there is an urgent need for a greater understanding of the causal biology underlying PSC. There are currently three main working hypotheses of PSC pathogenesis; the 'gut-homing T-cell', 'toxic bile' and 'leaky gut' hypotheses. Genetic support for these hypotheses is one means of establishing whether the underlying biological observations on which they are based, represent disease causation, or the effects of an established disease process. The three main hypotheses of PSC pathogenesis and existing genetic support for these hypotheses are discussed below.

1.6.1 The 'gut-homing T-cell' hypothesis

The PSC 'gut-homing T-cell hypothesis' is the hypothesis that memory T-cells, originally activated by inflammation within the gut are recruited to the liver where they cause the inflammation observed in PSC [53]. PSC is histologically characterised by T-cell rich portal infiltration with peri-ductal inflammation, portal fibrosis and progressive loss of

the bile ducts, known as ductopenia. Between 50-70% of patients with PSC also have concomitant IBD, although the observed course of hepatobiliary inflammation is notably independent from that of the colon. Approximately 75% of the blood supply to the liver originates from the intestine via the portal vein, thus creating an anatomical connection between the liver and gut. The portal vein drains into the hepatic sinusoids which are lined by fenestrated epithelia with Kupffner cells (a specialised liver-resident macrophage that phagocytoses pathogens or antigens from the portal blood). First proposed by Grant *et al* in 2001, the 'gut-homing T-cell hypothesis' conjectures that memory T-cells, originally activated by inflammation within the gut and expressing gut-specific ligands CCR9+ and $\alpha 4\beta 7+$, are recruited to the liver due to aberrant inflammation-induced expression of their receptors MAdCAM-1 and CCL25 [53]. High levels of MAdCAM-1 (mucosal addressin cell adhesion molecule) and CCL25 (chemokine C-C motif ligand 25) are usually restricted to the mucosal vessel endothelia of the gut and small intestine, respectively. In health, lymphocytes expressing the MAdCAM-1 receptor, CCR9, are found almost exclusively in the small intestine, with <10% of T-cells being CCR9+ in normal colon [77]. In active colitis however, their numbers increase, with approximately 90% of CD4+ and 30% of CD8+ tissue-infiltrating effector T-cells being CCR9+ [78]. Furthermore, in active colitis, intestinal CCL25, is up-regulated and levels correlate with mucosal TNFα expression and endoscopic measures of disease severity [78]. In support of the gut-homing T-cell hypothesis, MAdCAM-1 is found to be aberrantly expressed on the portal vein endothelium and CCL25 on the liver sinusoidal endothelium of patients with PSC [79]. Furthermore, in PSC it has been observed that 20% of liver-infiltrating lymphocytes express the respective MAdCAM-1 and CCL25 receptors, CCR9 and $\alpha 4\beta 7$ [80]. The majority of these CCR9+ T-lymphocytes are CD45RA+ CCR7+CD11a(high) and secrete IFN-γ in keeping with an effector memory phenotype. After recruitment to the liver, Grant *et al* proposed that CCR9+ and $\alpha 4\beta 7+$ gut-derived lymphocytes are likely to use other chemokines such as CXCL12 and CXCR6 to localise to biliary epithelium where they mediate targeted inflammation of the bile ducts.

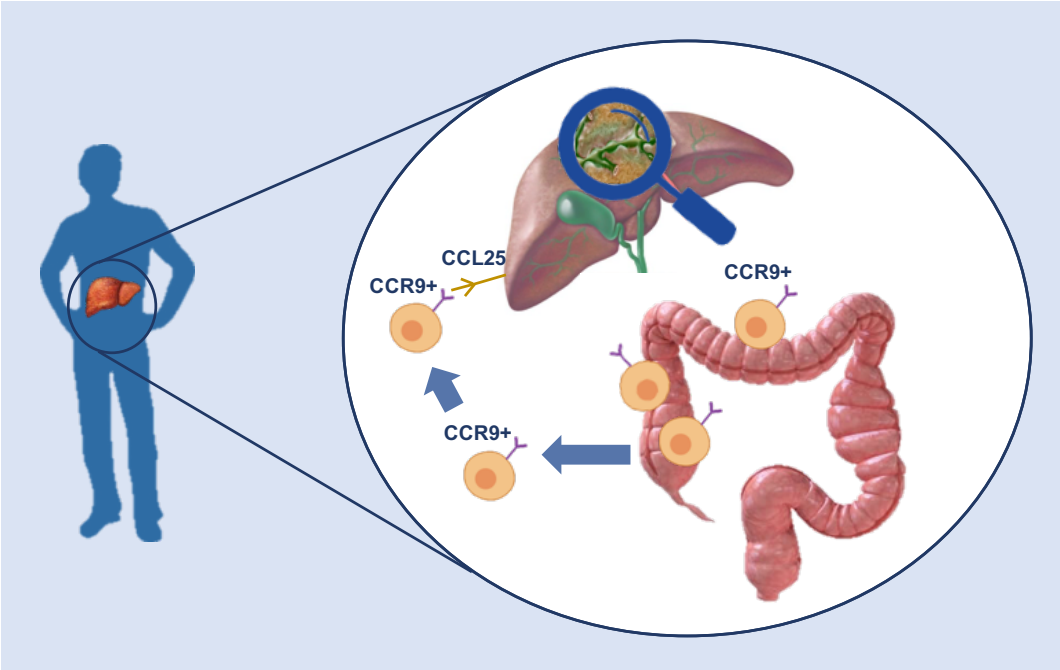

Figure 1.3: The 'gut-homing' T-cell hypothesis of PSC pathogenesis.

The α4β7 dimer (co-expressed alongside CCR9 in gut-homing T-cells) is an integrin complex expressed on T-cells, normally restricted to the gut. The importance of the integrin α4 gene has recently been confirmed by an IBD GWAS study, which has shown that the IBD risk increasing variant also increases expression of integrin α4 in stimulated monocytes [81, 82]. In IBD, this pathway is already the target of successful therapeutic blockade by Vedolizumab, a monoclonal antibody to the α4β7 integrin, which inhibits T-cell trafficking to the gut mucosa [83, 84]. Genetic studies in PSC have not yet detected any association with the integrin α4 gene, although are likely to be underpowered to do so, given that samples sizes in excess of 25,000 were required to detect the association with IBD. However disappointingly, clinical trials in patients with PSC and IBD have consistently observed no improvement of liver biochemistry with Vedolizumab treatment [85, 86].

Associations with several HLA and non-HLA PSC risk loci in close proximity to genes involved in T-cell biology such as *IL2RA* and *IL2/IL21*, supports a role for aberrant T-cell activation in PSC pathogenesis [42]. Furthermore, a study using high-throughput sequencing of TCRβ repertoires found significantly higher sharing of TCRβ repertoires in the gut and liver of PSC-IBD patients compared to paired normal gut and liver tissue, suggesting a common clonal origin between gut- and liver-derived memory T-cells of PSC-IBD patients. This finding is likely to result from reaction to a common antigen [87]. Therefore gut-homing T-cells may have an important pathogenic role in PSC.

1.6.2 The 'toxic bile' hypothesis

PSC belongs to the group of cholestatic liver diseases in which bile acid accumulation, or cholestasis, causes inflammation, apoptosis and necrosis of cells within the surrounding hepatic parenchyma. Whilst most hypotheses of PSC pathogenesis cite stricturing biliary inflammation as the cause of bile acid accumulation, the 'toxic bile' hypothesis conversely proposes that abnormal composition of bile itself mediates bile-duct injury and resultant cholestasis [88]. This hypothesis was based upon observations of the MDR2$^{-/-}$ knock-out mouse, a mouse deficient in a bile acid canalicular transporter, which is similar to the human MDR3/ABCB4 transporter that mediates biliary excretion of phospholipids. Phospholipids, excreted into the bile cannaliculi, combine with bile acids and cholesterol to form mixed micelles, which protect the biliary epithelium against the detergent properties of bile acids [89]. However due to their inability to secrete phospholipids into bile, MDR2$^{-/-}$ knockout mice spontaneously develop bile-duct injury with macroscopic and microscopic features closely resembling human PSC [90]. Notably, however, there have been no associations yet identified between genetic variants in *ABCB4* or other genes involved in the bile acid pathway with PSC risk. A second mechanism for protecting the apical surface of hepatocytes and cholangiocytes exists in the form of the 'HCO3- umbrella', which protects against attack from apolar hydrophobic bile acids. Decreased biliary HCO3- secretion can result in bile acid toxicity and thus damage to hepatocytes and cholangiocytes [88]. Early GWAS studies reported potential genetic associations with *GPBAR1* (G-protein-coupled bile acid receptor 1), which encodes a receptor involved in HC03- regulation [91], however this region has consistently fallen short of genome-wide significance in subsequent larger studies [40, 41].

Based upon the 'toxic bile' hypothesis, subsequent trials of therapeutic agents known to modify the bile acid composition have yielded mixed results. Ursodeoxycholic acid (UDCA), is a hydrophilic dihydroxy bile acid, very effective in the treatment of sister biliary condition, PBC. However trials in PSC have proved disappointing, with meta-analyses confirming no benefit on liver transplant rates, liver-related death or hepatic decompensation and only a small improvement in serum liver function tests with standard doses. Moreover, at high doses there was an increased risk of progression to hepatic decompensation and liver transplantation [92], attributed to the production and accumulation of hepatotoxic bile acids, such as lithocholic acid [93]. *Nor*UDCA, a C_{23} homologue of UDCA with a side chain shortened by one methylene group, is secreted into bile in an unconjugated, glucoronidated form and metabolised to non-hepatotoxic *nor*-lithocholate [94, 95]. It is known have anti-fibrotic properties, with a phase II trial in PSC reporting a significant improvement in serum ALP (alkaline phosphatase), a common surrogate measure of PSC disease activity [2, 96]. Furthermore trials of Obeticholic acid, an FXR agonist which down-regulates cytochrome P450, limiting bile salt production, has been recently approved for treatment

in PBC, with the results of phase II trials in PSC awaited [97]. Overall, whilst the evidence supports that toxic bile acid accumulation expedites biliary inflammation, both genetic and clinical studies provide minimal support for this hypothesis as the underlying causal process in PSC.

1.6.3 The 'leaky gut' hypothesis

The 'leaky gut' hypothesis conjectures that disruption of colonic permeability leads to microbial infection of bile, activating cholangiocytes and subsequently leading to hepatic inflammation and fibrosis [98]. In health, colonic pathogens and commensals remain confined to the colon due to the presence of mesenteric lymph nodes. These act as sites for the induction of tolerance to food proteins and protection against live commensal intestinal bacteria, penetrating the systemic immune system [99]. In the presence of intestinal inflammation, such as in IBD, the inner mucus layer of the intestinal mucosal barrier demonstrates increased permeability allowing interaction between the intestinal microbiota and the normally inaccessible surface epithelium [100]. Further disruption of the tight junctions connecting these epithelial cells allows translocation of bacteria across the mucosal barrier, where it enters the portal circulation [101]. This is supported by several observations. Firstly, the more frequent finding of translocated bacterial products in the explanted livers of patients with PSC, compared to other liver disorders [102]. Secondly, the transient improvement of serum ALP following treatment with metronidazole, an antibiotic which alters intestinal bacterial composition [103]. Thirdly, colectomy performed prior to liver transplantation is associated with a significantly decreased risk of recurrent PSC, post-transplantation [104].

The microbiome has a recognised role in the immune-pathogenesis of colonic inflammation in IBD, via the induction of T-regulatory cells and down-regulation of pro-inflammatory and up-regulation of anti-inflammatory cytokines [105]. In CD, intestinal microbial dysbiosis has been shown to be characterised by reduced microbial richness with an increase in mucus-degrading *Ruminococcus gnavus* [106] and a decrease in *Faecalibacterium prausnitzii, Bifidobacterium adolescentis* and *Dialister invisus* species [107]. In contrast, intestinal microbial richness in UC remains normal, but with a reduction in levels of butyrate-producing bacterial species *Roseburia hominis* and *F. prausnitzi*, a short-chain fatty acid with known anti-inflammatory properties [108]. Surprisingly, the few existing studies in PSC suggest that PSC demonstrates an intestinal microbial dysbiosis signature, independent from both UC and CD. PSC has been shown to be characterised by decreased microbiota diversity, and over-representation of *Lactobacillus, Fusobacterium* and *Enterococcus* genera, with one taxonomic unit belonging to the *Enterococcus* genus associated with increased levels of serum ALP [109]. More recently, several studies have confirmed a link between genetic variation and the gut microbiome, identifying genetic variants with effects upon

gut microbial composition in healthy individuals [110, 111]. In IBD, risk alleles within *NOD2*, have been associated with increased relative abundance of *Enterobacteriaceae* [112], with evidence that the increased susceptibility to ileal CD, conferred by these genetic variants is partially mediated by the microbiome itself. Furthermore variants within *FUT2* that have been associated (although not consistently replicated) with PSC risk, are also associated with changes in the commensal phyla in affected PSC patients [38]. These changes are characterised by reduced *Proteobacteria* and elevated *Firmicutes*. *FUT2* encodes galactoside 2-alpha-L-fucosyltransferase-2 and variants within the gene result in altered recognition and binding of various pathogens to FUT2 carbohydrate receptors on the mucosal surface. Overall, whilst intestinal microbial dysbiosis and translocation across a leaky gut barrier might be a consequence of disease pathogenesis, the evidence supporting a causal role is minimal.

Importantly, both the 'gut-homing T-cell' and 'leaky gut' hypotheses assume an inflamed colon as a key component of the disease model. They therefore cannot explain the presence of PSC in patients without IBD. However, whilst only 60-80% of PSC patients have diagnosed concomitant IBD, the milder IBD phenotype which often displays only microscopic levels of inflammation, may mean that actual rates of IBD in PSC are much higher [8–10]. Of the three hypotheses of PSC pathogenesis, the 'gut-homing T-cell' hypothesis is still widely considered the most biologically plausible causal mechanism on account of the supporting experimental and (albeit limited) genetic evidence.

1.7 Challenges in deciphering PSC risk loci

Our DNA, laid down at conception, provides a unique anchor for improving our understanding of the underlying causal biology of disease. This is because the genetic variants associated with disease susceptibility perturb genes and biological pathways that contribute to disease causality, and allow us to differentiate between cause and consequence of disease. The twenty-three genetic risk loci associated with PSC offer the potential for huge insight into the causal biology of this disease, if only we can robustly identify the true causal variants driving these loci and the genes they perturb.

When trying to extract disease-relevant biological insights from genetic risk loci there are two major hurdles. Firstly, identifying the causal variants driving the signals within each locus can be challenging due to patterns of LD or correlation between nearby genetic variants. The GWAS design uses this to its advantage, utilising several hundred thousand 'tagging' SNPs to capture a large proportion of the common variation in the human genome to powerfully identify loci associated with disease. Resultantly, the most strongly associated SNP identified by GWAS is likely to be in high LD with many other SNPs, any of which may be the causal SNP [113]. Identifying the causal variant within each

PSC locus is important for the design of follow-up studies investigating the underlying function of that variant. Statistical approaches aimed at identifying the most likely causal variant within risk loci are known as fine-mapping methods, and have been successful in resolving the causal variant for many IMD-associated risk loci. For example, fine-mapping of seventy-six RhA and T1DM loci defined credible sets containing five or fewer causal variants at five RhA and ten T1DM loci [114]. Furthermore, fine-mapping of 94 of the 240 known IBD risk loci resolved eighteen associations down to a single causal variant with >95% certainty, and twenty-seven associations to a single variant with >50% certainty [56]. Interestingly, of these forty-five variants, thirteen were found to be significantly enriched for protein-coding changes, three caused direct disruption of transcription-factor binding sites and ten were tissue-specific epigenetic marks in specific immune cells.

The second hurdle in understanding the functional importance of genetic risk loci is identifying the genes they affect. The vast majority of genetic variants associated with IMDs are located within non-coding regions of the genome, a complicating factor when considering their functional evaluation. Indeed, of the 22 known PSC risk loci outside of the HLA, only four have lead SNPs within coding regions of the genome. In the search to unravel the function of non-coding risk variants, it is now understood that many exert their influence via gene regulatory mechanisms and exert a quantitative effect upon gene expression. This is supported by the finding that up to 93% of GWAS risk loci occur in regulatory regions of the genome [115]. As such, variation in gene expression is an important component of the genetics of complex disease.

Understanding the epigenetic regulation of gene expression is already assisting the translation of genetic associations to disease mechanisms. This includes identifying genetic variants that alter gene expression either directly through a regulatory element, or indirectly by DNA methylation and chromatin accessibility. Defining the epigenetic changes that regulate genes associated with disease can improve both our ability to predict disease risk and our understanding of the underlying pathogenesis. Some of these epigenetic regulatory mechanisms are discussed below.

1.7.1 Expression quantitative trait loci

Expression quantitative trait loci (eQTLs) are genomic loci in which the abundance of a gene transcript is directly modified by a genetic polymorphism, usually within a regulatory element. Similar to any complex trait, the abundance of a gene transcript is a quantitative trait that can be measured [116]. In recent years eQTL mapping methods have been developed which test for association between genetic polymorphisms and transcript abundance, by simultaneously assaying gene expression and genetic variation on a genome-wide scale, in a large number of individuals. Importantly, variants associated with complex traits are more likely to be eQTLs than MAF-matched variants from GWAS analyses

chosen at random, confirming the importance of examining eQTLs in the functional study of genetic risk loci associated with complex diseases [117–119].

Variants that are eQTLs can act either in *cis* (within 1 megabase (Mb) of a gene transcription start site (TSS)), or *trans* (at least 5Mb up- or down-stream of the TSS), to directly alter gene expression. *Cis*-eQTLs tend to have greater effect sizes in comparison to their *trans*-eQTL counterparts, and thus modest sample sizes in the order of tens-to hundreds are sufficient for the detection of *cis*-eQTLs [120–122]. *Cis*-eQTL are often located close to the TSS of genes, with eQTL effect sizes generally tending to increase as the distance to TSS decreases [123]. In addition to altering transcription factor binding sites, *cis*-eQTL also tend to overlap other active regulatory elements such as activating DNAse-I hypersensitive sites that affect chromatin accessibility [124], whilst being depleted for repressive regulatory elements such as CTCF binding sites [125]. Many are also located within gene introns, however perhaps surprisingly, do not always affect the expression of that particular gene. For example, non-coding intronic variants within the *FTO* gene and associated with susceptibility to type-2 diabetes mellitus (T2DM), affect the gene expression levels of *IRX3*, a gene located several megabases away [126]. Measuring of *trans*-eQTLs requires much larger sample sizes than *cis*-eQTLs in order to generate enough power to detect their smaller effect sizes and correct for the greater number of tests required to measure the effects of each variant on all genes [118, 127]. Consequently, comparatively fewer *trans*-eQTLs have been reported within the literature. However, when observed, their presence can identify entire networks of gene pathways causally involved in disease pathogenesis. For example, a *trans*-eQTL analysis identified a SNP in the *IRF7* locus associated with T1DM susceptibility that exhibited *trans*-regulatory effects on an interferon regulatory factor 7 (IRF7)-driven inflammatory network enriched for viral response genes [128].

Multiple recent studies have demonstrated that genetic effects on gene expression can differ significantly between cell types and environments [129, 130]. Indeed, an eQTL may only be active in one particular cell type or state of activation [131–133]. Therefore in order to fully understand the functional mechanisms underlying GWAS risk loci it is important to examine the right cell type, in the right state of activation, at the right time. Identifying the relevant cell-type or stimulated state in which an eQTL is active remains challenging, and several studies have sought to address this through the mapping of eQTLs across several cell types challenged with multiple stimuli [130]. Interestingly, in a study combining RNA-seq with ATAC-seq (Assay for Transposase-Accessible Chromatin using sequencing), the majority of stimulus-specific eQTLs with a detectable effect upon chromatin accessibility also altered chromatin accessibility in unstimulated (naïve) cells [134]. Therefore in order to unravel the biological significance of disease-associated risk loci, it may also be important to examine other epigenetic markers including chromatin

accessibility and DNA methylation.

1.7.2 Histone modification

Histone modification marks are a common means of exploring the genetic determinants of chromatin conformation. To form chromosomal structures, chromatin is tightly packaged into an array of nucleosomes, each consisting of 147 bp of DNA. These wrap around alkaline proteins known as histones, which are arranged in octamers (H3, H4, H2A and H2B) and separated by linker DNA. The terminal tails of these histone octomers are subject to many forms of postranslational modification including methylation, acetylation, phosphorylation, and ubiquitination, which imparts functionality to nucleosomes both in the compaction of chromatin and in gene regulation [135]. Specific combinations of histone modifications provide landmarks for gene regulatory proteins. Commonly studied histone marks include H3K4me1, H3K4me3 and H3K27ac. H3K4me1 describes the mono-methylation of the fourth lysine from the N-terminal of the H3 protein and marks enhancer and promoter elements. H3K4me3 (trimethylation at lysine 4 on histone H3) marks the 5' region of active genes and is commonly associated with the activation of transcription. H3K27ac (acetylation at lysine 27 on histone 3) is found at both proximal and distal regions of TSSs [136]. The development of ATAC-seq and ChIP-Seq (chromatin immunoprecipitation followed by sequencing) technology, has enabled the genome-wide profiling of DNA-binding proteins and histone modifications.

1.7.3 DNA methylation

DNA methylation also plays an important role in the regulation of transcription, and is a potential candidate for exploring the functional importance of non-coding disease-associated risk variants. DNA methylation describes the addition of a methyl group to the 5' position of a cytosine residue that is 5' to a guanosine, commonly annotated as a 'CpG' site [137]. These methyl groups project into the groove of DNA, reversibly altering the biophysical properties of DNA to facilitate or prevent the binding of proteins [138]. CpG pairing generally occurs at a lower than expected frequency throughout the genome, with the exception of some particular CpG rich regions called 'CpG islands'. About half of CpG islands are associated with the promoter regions of genes [139], whilst the other half are located within genes or intergenic regions, often marking TSSs [140]. In general, DNA methylation is associated with gene repression with an inverse relationship between the extent of DNA methylation and expression levels of proximal genes [141, 142]. One mechanism via which cytosine methylation may lead to transcriptional silencing is via DNA methyltransferases interacting with transcription factors leading to site-specific methylation of promoter regions, influencing the assembly of transcriptional machinery

[143]. Methyl-binding proteins may exert influences on gene expression through a second functional domain which represses the transcription or recruiting of co-repressors or histone deactetylases which in turn affect chromatin modelling [144]. The combination of genome-wide SNP genotyping, CpG DNA methylation assays and RNA sequencing can be used to identify SNPs that influence DNA methylation (methQTLs) as well as down-stream gene expression [145]. Indeed, it has already been demonstrated that disease-associated variants have widespread effects on DNA methylation in *trans*, reflecting differential occupancy of *trans* binding sites by *cis*-regulated transcription factors [146].

1.8 Translating genetic risk loci into biological drug targets

GWAS have identified many genetic risk loci associated with susceptibility to complex disease. Nevertheless, the value of these genetic associations in the development of biological drug targets has been doubted due to the modest to small effect sizes of the vast majority of these risk loci. However, the impact of variant and gene discovery on the development of therapeutics has been greater than initially anticipated. Abatacept, is a drug highly successful in the treatment of RhA that targets the protein product of *CTLA4* [147]. However, at the *CTLA4* risk locus, the RhA risk increasing allele has an OR of just 1.1 [148]. Similarly, in IBD, Vedolizumab is a monoconal antibody that targets components of the α4β7 dimer, encoded by *ITGA4* and *ITGB7* [84]. At the *ITGA4* IBD locus, the IBD risk increasing allele also has an OR of just 1.1 [60]. Furthermore, Ustekinumab, used for the induction and maintenance of remission in refractory CD, is a monoclonal antibody that targets IL12B [149]. At the *IL12B* CD risk locus, the risk increasing allele has an OR of just 1.2 [150]. When trying to understand why a drug targeting a gene for which the lead variant of the risk locus has only a small to modest effect size, it is important to consider the allelic series. The presence of multiple causal variants within that gene, with the same direction of effect may generate a genotype–phenotype dose–response curve, explaining more than the effect of the individual causal variant [151]. Significantly, Nelson *et al* have demonstrated that drug mechanisms with genetic support are twice as likely to succeed from phase I trials to approval, than those without [152]. Therefore, it is anticipated that by expanding our understanding of the true causal variants and genes implicated by genetic risk loci, we will be more able to identify putative therapeutic targets for PSC.

1.9 Outline of this thesis

In this introductory chapter I have given an outline of the current knowledge of the genetic architecture of PSC. In addition, I have described the current hypotheses of PSC pathogenesis and the benefits and challenges of deciphering the genes and biological pathways impacted by PSC risk loci. The aim of this thesis is to build upon our current knowledge of the mechanisms by which genetic risk loci associated with PSC might result in the disease phenotype. This thesis aims to define the genetic variants, genes and cell types perturbed by each of the PSC risk loci, in an effort to bring us closer to drug target discovery.

In chapter 2, I describe the fine-mapping of each PSC risk locus. I use Bayesian fine-mapping methods to define a single causal variant or small set of credible causal variants with >95% posterior probability of causality. I describe two PSC risk loci which are fine-mapped to single variant resolution with >95% certainty, and a further three loci resolved to a credible causal variant with >50% certainty. In order to define the mechanisms via which non-coding variants impact upon PSC risk, I analyse all PSC credible causal variants for enrichment of known regulatory elements in PSC-relevant cell-types and tissues. Thus, I identify individuals credible causal variants which overlap enhancer or promoter elements in cell-types and tissues relevant to PSC.

In chapter 3, using colocalisation I aim to identify PSC risk loci which directly influence gene expression (i.e. are eQTLs) or indirectly influence gene regulation via DNA methylation or chromatin accessibility. I perform Bayesian colocalisation between PSC risk loci and functional QTLs measured in relevant cell-types and tissues. For each of three PSC risk loci, I find evidence of colocalisation with an eQTL for a single gene across multiple tissues. By fine-mapping these loci in the colocalising functional QTL traits, I further refine the credible sets for two PSC risk loci. Thus through a combination of colocalisation and fine-mapping, for each of these three risk loci I identify the dysregulated gene, a set of relevant cell-types and tissues in which the eQTL is active, a single or small set of credible causal variants, a direction of effect upon gene expression and the functional mechanism via which the causal variant perturbs the quantitative expression of that gene.

In chapter 4, I describe the generation and analysis of eQTL maps measured in the cell-types of most potential relevance to PSC. I develop eQTL maps in six PSC-specific T-cell subsets, including the rare CCR9+ gut-homing T-cells, isolated from the peripheral blood of patients with PSC and IBD. I perform differential gene expression according to disease phenotype with *DESeq2*. I map eQTLs in all six individual cell-types using *QTLtools* and identify eQTLs that are cell-type specific and those that are shared across multiple cell-types using *mashR*. Finally, I conduct colocalisation of eQTLs with PSC and IBD risk loci, identifying two genes that are causal in the pathogenesis of PSC, and three genes that are causal in the pathogenesis of IBD.

In Chapter 5, I discuss the major findings and discoveries from the previous chapters. I propose relevant further work, which could build upon and further the findings of this thesis. Finally, I conclude by discussing the future direction of genetic research in PSC.