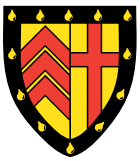
UNIVERSITY OF CAMBRIDGE

Primary sclerosing cholangitis: from genetic risk to disease biology

Elizabeth Claire Goode



Clare College

This dissertation is submitted for the degree of Doctor of Philosophy, May 2020

Declaration

This dissertation is the result of my own work and includes nothing which is the outcome of work done in collaboration except as declared in the Preface and specified in the text. It is not substantially the same as any that I have submitted, or am concurrently submitting, for a degree or diploma or other qualification at the University of Cambridge or any other University or similar institution except as declared in the Preface and specified in the text. I further state that no substantial part of my dissertation has already been submitted, or is being concurrently submitted, for any such degree, diploma or other qualification at the University of Cambridge or any other University or similar institution except as declared in the Preface and specified in the text. This dissertation does not exceed the prescribed limit of 60 000 words.

Elizabeth Claire Goode
May, 2020

Abstract

Primary sclerosing cholangitis: from genetic risk to disease biology
Elizabeth Claire Goode

One in 10,000 people in the Western world lives with Primary Sclerosing Cholangitis (PSC), an immune-mediated, inflammatory disease of the bile ducts that is highly co-morbid with inflammatory bowel disease (IBD). PSC confers risk of serious disease sequelae including hepatobiliary malignancy and progression to end-stage liver failure, for which the only treatment option is liver transplantation. The absence of effective medical therapies for PSC reflects our current limited understanding of the disease's aetiology and pathogenesis.

Our DNA, laid down at conception, gives us an unrivalled opportunity to understand 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. Twenty-two regions of the genome, outside of the HLA, have been associated with PSC risk. These loci offer the potential for huge insight into the causal biology of PSC, if only we can robustly identify the true causal variants driving these loci and the genes they perturb. However, this is complicated by several scientific challenges. Firstly, the majority of disease-associated risk loci occur within non-coding regions of the genome. Secondly, patterns of correlation between variants within a risk locus means that the true causal variant driving the signal could be any of those highly correlated with the variant with the smallest p-value.

In this thesis, I present analyses aimed at identifying the true genes and causal variants underlying each of the twenty-two PSC risk loci. Many non-coding risk variants associated with complex disease exert a quantitative affect upon gene expression i.e. are expression quantitative trait loci (eQTLs). Colocalisation assesses the evidence that a single shared causal variant is responsible for driving PSC risk and gene expression via an eQTL. In order to assign dysregulated genes to PSC risk loci, I perform colocalisation with eQTLs mapped in multiple cell-types and tissues mechanistically relevant to PSC. Because PSC is rare, eQTLs have not previously been mapped in all cell-types most relevant to this disease. In addition, I therefore map eQTLs in six peripheral blood T-cell subsets (including the rare CCR9+ gut-homing T-cells) from ∼80 patients with PSC and IBD. With colocalisation, I assign causal genes to five PSC risk loci, and assign other epigenetic regulatory features

including methylation or histone modification, to six risk loci. Statistical fine-mapping of each risk locus in both the GWAS and eQTL data enables me to resolve three PSC risk loci to a single causal variant and nine loci to 95% credible sets containing ten or fewer variants.

The results presented in this thesis identify three genes (*PRKD2, ETS2* and *UBASH3A*), causal in the pathogenesis of PSC, which are currently the target of existing or experimental therapeutic agents. Firstly, reduced expression of *PRKD2* causes excessive cell-autonomous T-follicular helper cell development and B-cell activation, and is associated with increased risk of PSC. Several studies are investigating the therapeutic effects of increasing the kinase activity of PRKD2. *ETS2* is involved in the induction of pro-inflammatory cytokine release from macrophages and IL-2 regulation in Th to Th0 transition. ETS2 inhibitors are currently the subject of early therapeutic trials. Finally, *UBASH3A* attenuates the NF-$\varkappa$B/I-KKβ pathway, an inflammatory pathway that is already targeted by proteasome inhibitors and acetylsalicylic acid, both of which could be potentially therapeutic in PSC.

PSC is a debilitating disease with serious disease sequelae, for which new therapeutic options are urgently needed. In this thesis, I elucidate multiple genes with a causal role in PSC pathogenesis, several of which are potential candidates for future therapeutic target.

Acknowledgements

As I began to write these final, but perhaps most important, few paragraphs I realise how fortunate I have been to be assisted by so many wonderful individuals along the road to completing this thesis. There a great number of people to whom I owe thanks and would like to make special mention to just some of them below.

In May of 2016, I began the work presented in this thesis under the supervision of Dr Carl Anderson. I wish to express my sincere gratitude to Carl for welcoming me to his research group and for his support and guidance in the years since. It is only with his enthusiasm, encouragement, scientific brilliance and humour that the work in this thesis has been made possible. I would also like to express my sincere appreciation to my secondary supervisors, Professor Nicole Soranzo, Dr Tim Raine and Dr Simon Rushbrook who have provided additional scientific clarity, wisdom and support. I would like to pay special regards to Simon for introducing me to this field of research back in 2012, and his mentorship ever since.

There are three talented scientists, whose assistance has been integral to the work presented in thesis and all of whom have shown me extensive kindness and patience. Firstly, I would like to thank Dr Loukas Moutsianas for teaching a naïve medic to code and for his guidance with fine-mapping and colocalisation. Secondly, I am grateful to Dr Laura Fachal who guided me through the functional annotation analyses and was always there to answer every genetics question, big or small. Thirdly, I would like to express my gratitude to Dr Nikos Panousis, whose direction was invaluable in the RNA-seq and eQTL mapping component of this thesis. To all three of you, I am indebted.

For the past four years I have been fortunate to be supported by a fellowship from the University of Cambridge Wellcome Trust Clinical PhD program. The Wellcome Sanger Institute has been an incredible source of support throughout my fellowship and a wonderful place to spend the last few years. I would also like to thank each and every member of the Anderson Team at the Wellcome Sanger Institute. This team has evolved and expanded during my time at Sanger and I have been lucky to spend time with a group of such kind, talented and good-humoured individuals. I would like to pay special thanks to Dr Rebecca McIntyre who has been a source of personal support and who also performed all of the DNA extraction from my study samples. I would also like to thank Mr

Ben Bai for his guidance with mashR *('Nobody really understands how mashR works...')*, Dr Alex Sazonovs for his help with Latex *('It's us against Latex!')*, Dr Carla Jones for her patient explanations of T-cell biology and Mr Sigurgeir Olafsson, Alex and Ben for helping solve my innumerable coding issues!

I would also like to extend my gratitude to Dr Norihito Kawasaki and Dr Alex Wittmann at the Quadram Institute, for their guidance in the preparation of samples for my T-cell eQTL study, alongside the many patients from the Norfolk and Norwich University Hospital who donated their time (and blood!) to this study. In addition I would like to thank the patient support charity *PSC Support*, who generously funded the sample acquisition for the T-cell study.

To my esteemed examiners, Dr Richard Sandford and Professor Heather Cordell, I would like to express my gratitude for their invaluable comments and for a thoroughly enjoyable viva (albeit via Zoom)!

Undoubtedly, my proudest achievement during my PhD has been the arrival of my daughter, Alexandra. Thank you for being a source of such joy, for reminding me of the most important things in life, and for sleeping well so that Mummy could write her thesis! I would also like to express my sincere gratitude to Dr Pat Tate whose unwavering positivity encouraged me to put confidence in myself, and has kept me (mostly) sane. My final and most special debt of gratitude goes to my wonderful husband, Grégory for his great love and support. Je tiens à exprimer ma profonde gratitude pour ton amour et ton soutien et pour m'avoir aidée à prendre confiance en moi. To him, I am wholeheartedly grateful.

Finally, to my parents and the many family, friends and colleagues who I have not named individually, but who have also given me their support. I hope that along the way I have shown you my appreciation, and to you all I give a heart-felt thank you.

Contents

List of Figures

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.

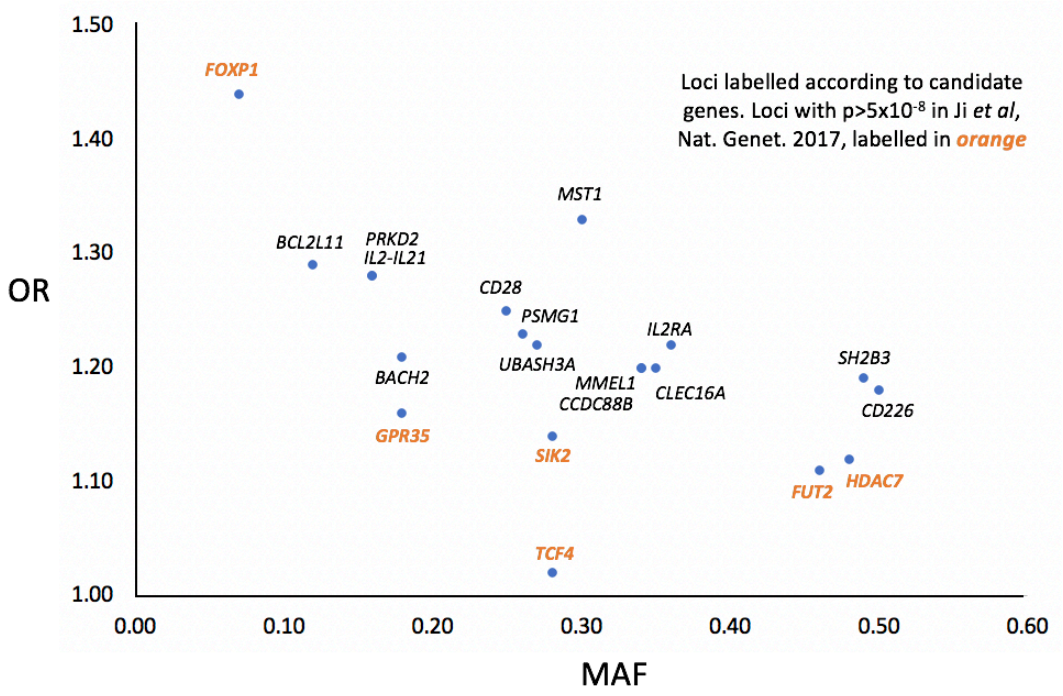


Figure 1.1: Twenty of the twenty-two non-HLA PSC risk loci plotted according to their effect size (OR) and MAF in Ji *et al*'s GWAS data [42].

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.

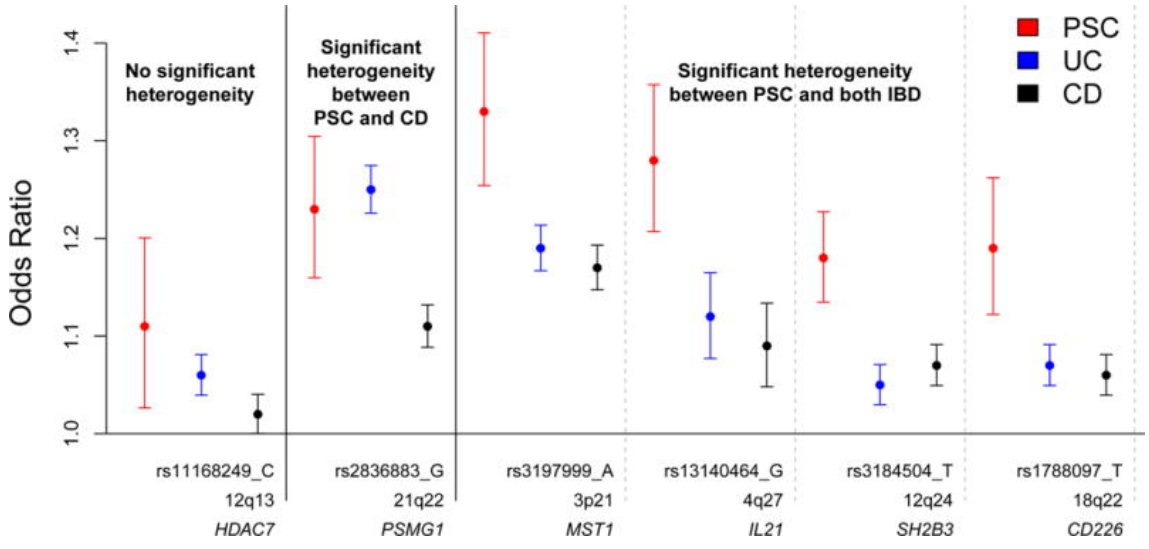


Figure 1.2: Figure taken from Ji *et al* demonstrating odds ratios (and their 95% confidence intervals) for PSC, UC and CD across the 6 PSC associated SNPs demonstrating strong evidence for a shared causal variant (maximum posterior probability >0.8) [42].

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.

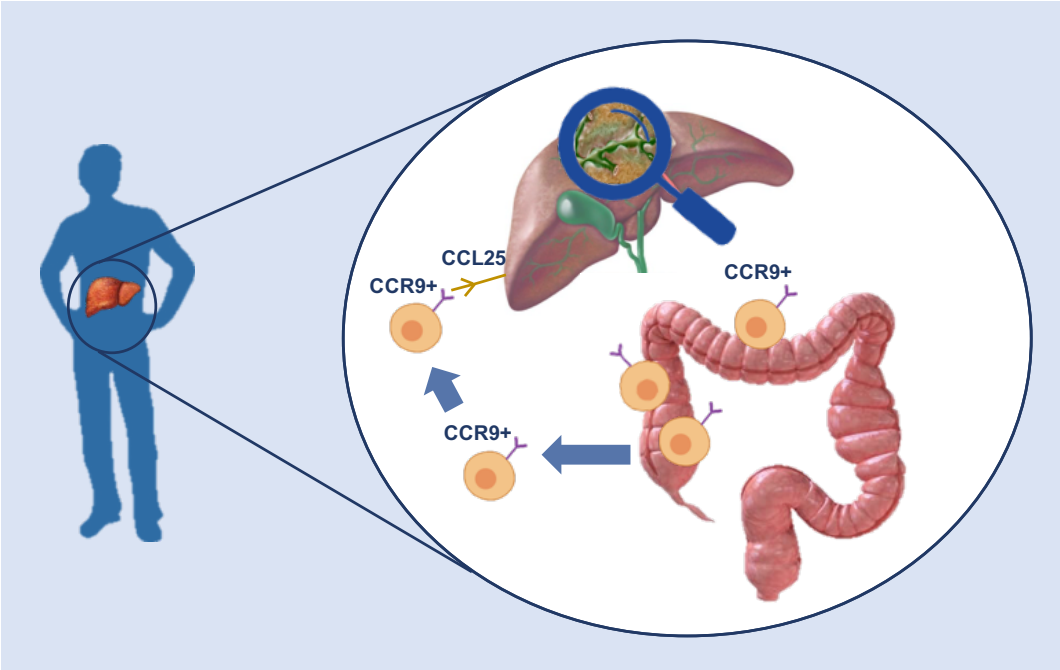


Figure 1.3: The 'gut-homing' T-cell hypothesis of PSC pathogenesis.

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.

Chapter 2

Fine-mapping of disease-associated risk loci in Primary Sclerosing Cholangitis

2.1 Introduction

GWAS have identified many thousands of regions of the genome associated with immune-mediated disease (IMD), which have been replicated both within and between diseases. One biological phenomenon facilitating the success of GWAS is that of linkage disequilibrium (LD), the non-random association of alleles between nearby genetic variants. LD blocks can consist of anywhere between two and thousands of highly correlated single-nucleotide polymorphisms (SNPs). The GWAS design uses LD to its advantage, utilising several hundred thousand 'tagging' SNPs to capture a large proportion of the common variation in the human genome, to accurately identify loci associated with disease. The most strongly associated variant with the smallest p-value within the disease-associated locus is referred to as the 'lead variant'. However, one important shortcoming of the GWAS design is that the lead variant identified by GWAS is likely to be in high LD with many other SNPs, resulting in a statistical association of approximately equivalent strength across all of the variants in high LD (defined as $r^2>0.8$), any one of which could be the true causal variant. Distinguishing the causal variant from the often hundreds of variants in LD is not possible from GWAS alone. Whilst conditional analysis is one means of identifying the number of independent association signals within a region, it cannot ascribe an individual probabilistic measure of causality for each individual variant within a locus. Identifying the true causal variant within each risk locus is an essential step in translating genetic associations into biological functions that explain disease processes. Knowledge of the precise location of the causal variant within, for example, a gene intron, exon, splice junction, promoter or enhancer region, may provide important clues about the mechanism via which the variant

exerts its effect on disease risk. With the development of CRISPR/Cas9 gene editing technology, which allows the introduction of targeted mutations within cellular DNA, knowledge of the causal variant can now greatly facilitate the design of functional assays investigating the mechanistic impact of disease-associated variants. Furthermore, it allows recall-by-genotype experiments, studying individuals with and without a particular causal variant.

Statistical approaches aimed at identifying the most likely causal variant within GWAS risk loci are known as fine-mapping methods. Fine-mapping aims to define a single variant or credible set of variants which contain the true causal variant with a high probability. Several conditions are important when conducting fine-mapping studies. Firstly, whilst GWAS requires only one variant in LD with the true causal variant to detect a signal for disease association, fine-mapping requires that all common SNPs within a region are genotyped or well-imputed [113]. This is because an important fine-mapping assumption is that the true causal variant is included within the data. Imputation using reference panels such as UK10K or the 1000 Genomes Project, allows the incorporation of variants that were not included within the original genotyping array [153, 154]. This aims to satisfy the assumption that when estimating the relative evidence for each variant being causal, the true causal variant is present within the analysis [113]. Secondly, fine-mapping utilises subtle differences in the strength of association between tightly correlated variants to infer causality. It is therefore especially sensitive to data quality and stringent quality control is essential to remove genotyping errors and batch effects. Large sample sizes are also necessary to achieve sufficient power to differentiate between SNPs in high LD. As shown in Figure 2.1, taken from Huang, Fang, Jostins *et al* [56], power to identify the causal variant in a correlated pair increases with the significance of the association, and therefore with sample size and effect size. Initially, the generation of larger sample sizes was achieved by the use of cheaper custom genotyping arrays, such as the Immunochip, at the expense of analyses restricted to only those regions of the genome previously associated with risk of those IMDs included within the Immunochip, of which PSC was not included. However more recently, the reducing cost of sequencing has allowed GWAS on an unprecedented scale, with the meta-analysis of pooled data through collaborative consortia.

The standard approach for refining association signals is via conditional analysis, a step-wise, iterative approach which conditions on the SNP with the lowest p-value for association and continues to add SNPs until no additional SNP reaches the p-value threshold, usually set at 5×10^{-8}. This process provides information about the number of complementary signals within a locus, but cannot assign an probabilistic measure of causality to each individual variant within the locus. Furthermore, p-values are not necessarily comparable between studies as they are heavily influenced by characteristics of the individual study design such as power and locus-specific factors such as minor

allele frequency (MAF) and effect size [113]. Currently, the most common approaches to fine-mapping therefore employ Bayesian methods in which the evidence for the association of each variant is tested using an approximate Bayes factor (ABF). These are then used to calculate the posterior probability (PP) for each variant being causal within a region. Each PP describes the ratio for that variant being causal, versus all other variants in the region, and thus Bayesian PPs are more comparable between variants of the same or different studies. Furthermore, the Bayesian approach enables the weighting of evidence for a particular variant being causal according to prior knowledge, known as the prior probability. Typical fine-mapping approaches in complex diseases aim to define the number of independent signals within a risk locus, and to identify a 'credible set' of variants, in which the sum of the PPs is >0.95, and thus the credible set is >95% likely to contain the true causal variant.

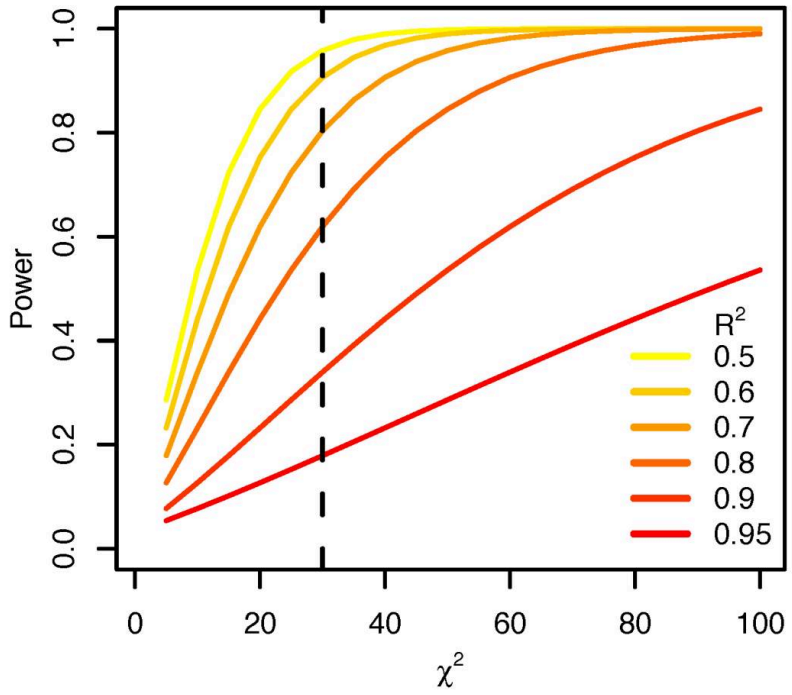


Figure 2.1: Power (y axis) to identify the causal variant in a correlated pair increases with the significance of the association (x axis), and therefore with sample size and effect size (vertical dashed line shows genome-wide significance level). Figure taken from Huang, Fang, Jostins *et al* [56].

Fine-mapping approaches have been applied to several IMDs to date and have been successful in resolving some disease risk loci down to single or small set of credible causal variants. For example Westra *et al* fine-mapped 76 rheumatoid arthritis (RhA) and type 1 diabetes (T1DM) risk loci, defining credible sets of ≤ 5 causal variants at 5 RhA and 10

T1DM loci [114]. IBD remains the most extensively fine-mapped IMD to date. Huang, Fang and Jostins *et al* fine-mapped 94 of the 240 known IBD risk loci and resolved 18 associations down to a single causal variant with >95% certainty, and 27 associations to a single variant with >50% certainty [56]. Importantly, of these 45 variants, 13 were found to be significantly enriched for protein-coding changes, 3 caused direct disruption of transcription-factor binding sites and 10 were tissue-specific epigenetic marks in specific immune cells. In addition de Lange and Moutsianas *et al* resolved an additional 7 IBD loci to a single credible variant with >50% PP of being causal [60]. To date, there have been no fine-mapping studies of the genetic risk loci associated with PSC.

2.2 Chapter overview

Twenty-three regions of the genome have been associated with susceptibility to PSC. The first step in translating these genetic associations into biological understanding of disease mechanisms is to accurately define those variants which are responsible for driving each risk locus. In this chapter I describe the first fine-mapping analysis of genetic risk loci associated with PSC susceptibility. I apply Bayesian fine-mapping approaches to PSC risk loci, with the aim of resolving each locus to a single causal variant or a small set of credible causal variants. To identify those credible variants in non-coding loci which overlap known functional regions of the genome, I perform annotation of the fine-mapped variants to define their functional effects.

2.3 Methods

2.3.1 Fine-mapping

There are multiple computational software programs which employ Bayesian approaches to fine-mapping. In order to develop a fine-mapping analysis pipeline that could be easily applied to data-sets for which full genotype data might not be available, I aimed to use a method of fine-mapping which could be applied to summary statistic data. At the time of conducting this study, several methods for fine-mapping using summary statistics and a SNP correlation matrix, were available. These included *Paintor* [155], *Caviar* [156], *CaviarBF* [157] and *FINEMAP* [158]. Of these four methods, the first three implement an exhaustive search through all possible causal SNP configurations and therefore become computationally slow when considering more than one independent causal variant within each region. I therefore opted to use *FINEMAP* v1.3, a computational software program for the fine-mapping of complex traits [159]. The *FINEMAP* model is made up of four components; the likelihood function, priors, likelihood evaluation and

search algorithm. It differs from the other three methods in that it employs a Bayesian approach to explore a set of the most likely causal configurations of variants within each region via a shotgun stochastic search algorithm [160]. By focusing the analysis on those variants with a non-negligible causal probability rather than searching through all possible causal configurations, *FINEMAP* avoids becoming computationally slow or intractable when considering several causal variants in a data-set with many thousands of variants within each region.

The first step of Bayesian fine-mapping requires a set of summary statistics and risk loci. I used summary statistics from the largest available PSC GWAS dataset, published by Ji *et al* [42]. Ji *et al* identified 15 PSC risk loci outside of the HLA, associated with PSC susceptibility. In addition, a further seven risk loci have been previously reported as associated with PSC from other studies (*CCL20, CPR35, NFKB1, SIK2, HDAC7, RFX4, TCF4*), however these did not reach genome-wide significance in Ji *et al's* study. I therefore excluded these seven loci, focusing fine-mapping efforts on the 15 significant PSC risk loci (shown in Table 2.1). *FINEMAP* assumes that each region to be fine-mapped includes at least one causal SNP, and that all causal SNPs are included within the data (either directly genotyped or imputed). Genotyping of the PSC GWAS data-set had been previously conducted on three different genotyping arrays; the Illumina Omni 2.5-8 and Omni 2.5-4 and the Affymetrix Affy 6 with imputation using a combined reference panel of the 1000 Genomes Phase 1 integrated version 1 and the UK10K cohort [153]. Quality control had been previously conducted by Ji *et al* using strict standards for genetic association analysis [42]. For the purposes of fine-mapping, I defined each of the 15 PSC risk loci as 1Mb regions centred upon the lead GWAS SNP. GWAS summary statistics required for the analysis were the variant RSID, the chromosome and base pair (bp) position of each SNP, all reported according to Ensembl build 37, the major and minor alleles along with the MAF, the estimated effect size (β) and standard error (SE) of the effect size.

The second step of Bayesian fine-mapping requires the calculation of an LD matrix with the estimation of LD between variants within each risk locus using Pearson correlations. Recent studies support the use of original genotype data, where available, for the calculation of LD structure over the use of reference panels [159]. This is not only because the LD matrix will then match exactly the study population, but because the size of the reference panel for calculation of the LD matrix needs to scale with the GWAS sample to maintain optimal fine-mapping performance. Fine-mapping with smaller reference panels (e.g. 100 individuals) misleadingly results in smaller credible sets, with much lower coverage over variants than the larger reference panels or original genotype data. Benner *et al* demonstrated that a reference panel of 1,000 individuals is sufficient when summary statistics originate from a GWAS with 5-10,000 individuals. For this reason, I used full original genotype data from Ji *et al* [42] to calculate the SNP correlation matrix, using

computational software program *LD Store* v1.1 [159]. Importantly, this ensured that the ancestry of the LD cohort matched exactly the ancestry of the GWAS cohort. For any two variants, LD information was only extracted if absolute Pearson correlation was above zero. One of the drawbacks of fine-mapping methods based on summary statistics is that they are more sensitive to the choice of data-set used to calculate the LD matrix. I therefore also conducted fine-mapping analyses using LD structure derived from the UK10K project reference panel, to assess any differences dependent upon the choice of LD matrix.

FINEMAP assumes that each SNP is causal with prior probability of 1 / number of SNPs in the genomic region. I left prior probabilities for the number of independent causal SNPs in the genomic region unspecified, however repeated the analyses with iterations assuming between one and five independent causal variants per region. For each of the 15 risk loci, the analysis output included model-averaged posterior summaries for each SNP, posterior summaries for each causal configuration, posterior summaries for the number of independent signals per region and the 95% credible sets for each causal signal conditional on other causal signals in the genomic region. To declare a locus fine-mapped to single causal variant, I defined that the PP of causality for that single variant had to be $\geq 95\%$. Evidence for additional independent signals within a risk locus was taken as a PP of >50% in support of two or more independent signals within a risk locus. In order to check the fine-mapping assumption that each locus contained a true causal variant and all potential causal variants had been included within the analysis, I searched the UK10K and 1000 Genomes reference panels for any SNPs in moderate or high LD (defined as an r^2>0.5 and >0.8 respectively) with the most probable fine-mapped variant, noting any that were not included within the analysis.

2.3.2 Functional annotation

The majority of common disease-associated variants are located within non-coding regions of the genome. These non-coding variants are thought to overlap functional DNA elements involved in gene regulation such as transcription factor binding sites, open chromatin or gene enhancer regions [117]. Functional annotation complements statistical fine-mapping methods by providing independent information about the likely biological function of each variant. In recent years functional annotation profiles have been developed across many hundreds of tissue and cell types collated into databases such as the Encyclopedia of DNA Elements (ENCODE) database [161].

I aimed to further define the function of those non-coding variants included within the credible sets from the above fine-mapping analysis, by assessing which credible causal variants overlapped functional regions of the genome. Several existing fine-mapping approaches have integrated functional annotation as a means of prioritising causal variants. *Paintor* is one example of such an approach which integrates association strength with

functional genomic annotation data to improve the accuracy in selecting credible causal variants for functional validation [155]. Genomic Annotation Shifter *(GoShifter)* is a statistical approach that tests for enrichment of functional annotations overlapping a disease-associated variant, as a means to prioritising variants for further functional follow-up [162]. *GoShifter* identifies all variants in high LD (defined as an $r^2>0.8$) with the lead GWAS variant, and the median size (in bp) of the tested annotation feature, X. *GoShifter* defines the 'locus' as the region between the two furthest SNPs linked with the lead variant, plus twice the median size of X. *GoShifter* identifies the proportion of loci in which at least one SNP overlaps X, and compares this to a null distribution of iterations, generated by repeated random shifting of the site of X within the locus. The p-value is computed as the proportion of iterations for which the number of overlapping loci is equal to or greater than that for the tested SNPs. *GoShifter* then uses stratified enrichment analysis to asses the significance of an overlap with X, independent of overlap with any other colocalising annotation, Y. This involves separating the locus into two fragments- that which overlaps Y, and that which does not, Y_0. X is then shifted separately within Y and Y_0 to generate the null distribution, and the significance of the observed overlap assessed by the proportion of loci in which any SNPs overlaps annotation X in Y or Y_0. The delta overlap describes the difference between the observed proportion of loci overlapping X and the mean proportion of loci overlapping X under the null derived by shifting and provides a measure of the effect size of the observed enrichment. In the absence of enrichment, the observed overlap will be close to the mean overlap of the null, and delta-overlap will be close to 0, whereas stronger enrichment corresponds with larger delta overlap. Finally, to identify loci in which the overlap between a SNP and an annotation is particularly informative and thus should be higher priority for further functional evaluation, the 'overlap score' is calculated. The overlap score describes the probability that each locus overlaps an annotation by chance, and is only computed for loci that overlap the annotation in the observed data. Loci with better (lower) overlap scores suggest significant enrichment and are therefore proposed to be higher priority for functional evaluation of causal variants.

To identify those non-coding credible variants that overlapped gene regulatory features, I used a modified version of *GoShifter*, with modifications similar to those implemented in a published study by Ulirsch *et al* [163]. These modifications included substitution of the high-LD variants with the credible set variants from my fine-mapping analysis. Therefore, the 'locus' provided to *GoShifter* was defined as the region between the two furthest credible SNPs linked with the variant with the highest PP of causality from fine-mapping, plus twice X (the median size in bp of the tested annotation). In the first stage of the analysis *GoShifter* therefore takes all PSC credible causal variants across all non-coding PSC loci and tests for regulatory features enriched across all PSC credible causal variants.

In the second stage, it tests for overlap of each locus with those enriched features from the first stage of the analysis, to prioritise credible causal variants based upon those with the lowest overlap score.

I used the ENCODE v4 database [161] for all annotations, including promoters, enhancers, histone acetylation marks and DNase-I hypersensitive sites for 28 whole tissue and immune cell sub-types, relevant to PSC. I defined relevant PSC tissues as any immune cell type and any tissue from an organ system affected by PSC. Instead of using a set of high LD variants, for each independent signal within each locus, I input the 95% credible set of variants defined from my fine-mapping study and performed 20,000 local shift iterations per annotation. I calculated delta-overlap scores to measure the enrichment of overlap between annotations and credible variants. I adjusted the enrichment p-values for multiple testing using the Benjamin Hochberg FDR correction at 5% [164]. I chose this less stringent form of multiple testing correction as some annotations are not independent (for example enhancer marks are made up from a combination of annotations) and therefore a more lenient method than the Bonferonni correction is required. I calculated overlap scores for those loci that overlapped annotations. Lower scores suggest significant enrichment and higher priority as causal variants. Although *GoShifter* does not define a overlap score threshold to interpret significance, I prioritised variants from credible sets based upon the variant or variants with the lowest feature overlap score per PSC risk locus.

For ease of reference, SNPs are referred to according to their RSID. The 15 PSC risk loci are numbered 1 to 15, and referred to according to their PSC risk region number, chromosome and bp position of the lead GWAS SNP for the region, according to Ensembl build 37 (see Table 2.1). Where a region is referred to according to a nearby candidate gene, it is important to note that candidate genes are assigned according to locality and biological plausibility and do not necessarily describe a proven causal association between variant and gene, unless specifically stated otherwise.

2.4 Results

I conducted fine-mapping of the fifteen PSC risk loci in the GWAS dataset published by Ji *et al* [42]. Genotyping, quality control (QC) and imputation had been previously conducted on the genotype data of 24,751 individuals of European ancestry, including 4,796 PSC cases 19,955 controls, as previously described by Ji *et al.* Fine-mapping of the PSC risk loci identified nineteen independent signals across fifteen risk loci (Figure 2.2). Evidence supported just one causal signal within eleven of the fifteen regions with >50% certainty. For seven of those regions the PP supporting one independent causal signal was >70%. In the remaining four regions the evidence supported the presence of two independent signals with >50% certainty. For each signal detected, variants were sorted

by the PP of causality and added to the credible set of associated variants until the sum of their PPs exceeded 95%. These credible sets ranged in size from one to sixty-two variants (Table 2.1). For all loci, review of the 1000 Genomes and UK10K data-sets revealed that there were no SNPs in high LD (r^2>0.8) with the most probable causal variants, missing from the analysis. Two of the fifteen risk loci resolved to a single causal variant with >95% certainty and three loci to larger credible sets where one credible variant was assigned >50% PP of causality. Of these five variants, one was enriched for a significant protein-coding change, one caused direct disruption to a splice site and three overlapped tissue-specific epigenetic marks in PSC-relevant tissue- and cell-types (Table 2.2).

Sensitivity analysis using a different LD matrix derived from UK10K reference cohort demonstrated that when considering one independent causal variant, choice of LD matrix did not affect the most probable SNP or PP of causality according to *FINEMAP*. As expected, when considering more than one independent causal variant in each region, the results of fine-mapping were more sensitive to the choice of LD matrix. For five of the fifteen loci, the most probable fine-mapped causal variant using LD from GWAS and UK10K remained the same. For seven of the fifteen loci the credible sets derived from both analyses were identical, with a slight redistribution of the PP of causality to a neighbouring, highly-correlated variant. Fine-mapping attempts in the remaining three loci resulted in credible sets containing more than forty credible causal variants with complex patterns of LD in all three regions, which was not improved with a different choice of LD matrix.

GoShifter identified significant enrichment of credible causal variants from all 19 independent signals with promoter and enhancers annotations in all five tested immune-cell types (B-cell, CD14+ monocytes, macrophage, peripheral blood mononuclear cells and T-regulatory cells) and eleven gastro-intestinal tissues (colonic mucosa, duodenal mucosa, gastroesophageal sphincter, large intestine, liver, duodenal muscle, rectal smooth muscle, rectal mucosa, Sigmoid colon, small intestine and transverse colon). It is important to note that *GoShifter* is applied only to variants within non-coding regions and therefore PSC regions 4 and 10 (loci within coding regions) and PSC region 15 (a splice site region), were not included in this analysis.

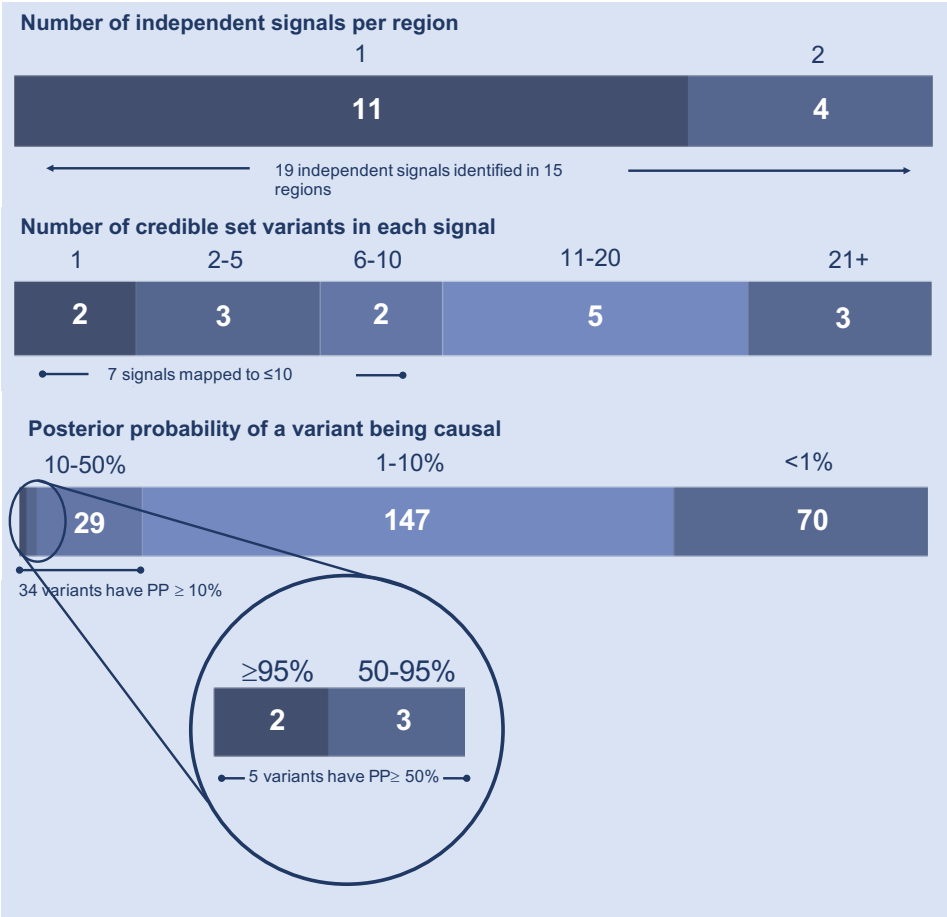


Figure 2.2: Summary of fine-mapping the PSC risk loci.

Table 2.1: Fine-mapping of PSC risk loci

			Region			Fine-mapping			
Region	**Signal**	**Chr**	**Candidate Gene**	**Lead GWAS SNP**	**Position (b37)**	**SNP PP_{max}**	**Position (b37)**	**PP Causal**	**Credible set size**
1	1	1	*MMEL1*	rs3748816	2526746	rs61763697	2810791	0.07	62
2	1	2	*BCL2L11*	rs72837826	111933001	rs72837826	111933001	0.18	12
3	1	2	*CD28*	rs7426056	204612058	rs5837875	204647878	0.19	6
	2					rs231799	204707417	0.17	
4	1	3	*MST1*	rs3197999	49721532	rs11716895	49762779	0.11	13
	2					rs13083791	49721798	0.07	
5	1	3	*FOXP1*	rs80060485	71153890	rs80060485	71153890	0.99	1
	2					rs36023390	71523093	0.14	
6	1	4	*IL2-IL21*	rs13140464	123499745	rs13119723	123218313	0.09	50
7	1	6	*BACH2*	rs56258221	91030441	rs7750271	91036225	0.20	12
8	1	10	*IL2RA*	rs4147359	6108439	rs4147359	6108439	0.46	5
9	1	11	*CCDC88B*	rs663743	64107735	rs35247680	63884747	0.61	2
	2					rs663743	64107735	0.41	
10	1	12	*SH2B3*	rs3184504	111884608	rs3184504	111884608	0.99	1
11	1	16	*CLEC16A*	rs725613	11169683	rs725613	11169683	0.16	12
12	1	18	*CD226*	rs1788097	67543688	rs1610555	67543147	0.08	44
13	1	19	*PRKD2*	rs313839	47221557	rs313839	47221557	0.23	14
14	1	21	*ETS2*	rs2836883	40466744	rs4817988	40468838	0.58	10
15	1	21	*UBASH3A*	rs1893592	43855067	rs1893592	43855067	0.62	5

PP; posterior probability of causality

Table 2.2: PSC risk loci overlapping gene regulatory features

Region	Signal	Chr	Cand. gene	FINEMAP SNP	FM PP	GoShifter SNP	FM PP	Overlaps promoter in these tissue	Overlaps enhancer in these tissues
1	**1**	1	MMEL1	rs61763697	0.07	rs60733400	0.02	SI,RM,L,CM,MD, CD14,SC,PBMC,TC,DM,LI,MR,SC	SC,SI,L,GOS,BC,MR,TC,CM,CD14, PBMC,LI,DM, MD
2	**1**	2	BCL2L11	rs72837826	0.18	rs72836345	0.18	SI,RM,L,CM,CD14,SC,PBMC,TC,DM,LI,MR,M	SC,GOS,BC,RM,TC,CM,CD14,PBMC
3	**1**	2	CD28	rs5837875	0.19	rs5837875	0.19	RM,CM,PBMC,L,DM,MR, RM,CD14,M,TR	SI,SC,CD14, PBMC,RM,GOS,CM,TR
	2			rs231779	0.16	rs231779	0.16	RM,CM,PBMC,L,DM,MR, RM,CD14,M,TR	SI,SC,CD14, PBMC,RM,GOS,CM,TR
5	**1**	3	FOXP1	rs80060485	0.99	rs80060485	0.99	CD14,MR,RM,BC,TR,M	SC,GOS,MR,RM,LI,BC,TR
	2			rs36023390	0.14	rs36023390	0.14	CD14,MR,RM,BC,TR	SC,GOS,MR,RM,LI,BC,TR,CM
6	**1**	4	IL2-IL21	rs13119723	0.09	rs67963613	0.01	SI,RM,L,CM,DM,SC,PBMC,CD14,TC, DM,LI,MR,RM,M,TR	SC,SI,L,GOS,BC,RM,MR,TC,CM,CD14,PBMC,L,MD,DM,TR
7	**1**	6	BACH2	rs7750271	0.20	rs7750271	0.20	SI,RM,L,CM,MD,SC,PBMC,CD14,TC, DM,LI,MR,RM,M,BC	SC,SI,L,GOS,BC,RM,MR,TC,CM,PBMC,CD14
8	**1**	10	IL2RA	rs4147359	0.46	rs4147359	0.46	CD14,M,L,DM,CD14,CM,SC,RM,TR,LI,CM,BC,SI,MR,MD	SC,SI,L,GOS,BC, RM,TC,CM, CD14,PBMC,MD, L,DM, CM,TRs,LI,BC,SI,MR
9	**1**	11	CCDC88B	rs35247680	0.61	rs35247680	0.61	SI,RM,L,CM, MD, CD14,SC,PBMC, TC,DM,LI,MR,M	L,SC,GOS,BC,RM, TC,CM,CD14,PBMC,LI ,SI,DM
	2			rs663743	0.41	rs663743	0.41	SI,RM,L,CM, MD, CD14,SC,PBMC, TC,DM,LI,MR,M,L	L,SC,GOS,BC,RM, TC,CM,CD14,PBMC,LI ,SI,DM,TC
11	**1**	16	CLEC16A	rs725613	0.16	rs113344842	0.02	SC,SI,L,GOS,BC,MR,TC,CM,CD14,RM,MD,DM,CM, PBMC,RM,TRs,LI	CM,CD14,PBMC,MR,RM,M,L,DM,CD14,CM,PBMC,SC,BC, RM,SI,TR,LI,CM,MR,MD
12	**1**	18	CD226	rs1610555	0.08	rs4891781	0.03	PBMC,SC,TR,MD,SI,PBMC	LI,RM
13	**1**	19	PRKD2	rs313839	0.23	rs313839	0.23	SI,RM,L,CM,MD,CD14,SC,PBMC,TC,DM, LI,MR,RM,M	SC,SI,L,GOS,BC,RM,MR,TC,CM,CD14,PBMC,LI,MD,DM
14	**1**	21	ETS2	rs4817988	0.58	rs2836883	0.05	CM,RM,CD14,PBMC,M,SI	SC,SI,L,GOS,RM,TC,CM, CD14,PBMC,MD,DM,BC

BC; B-cell, CD14; CD14+ monocyte, CM; Colonic mucosa, DM; Duodenal mucosa, GOS; gastroesophageal sphincter, LI; Large intestine, L; liver, M; macrophage, MD; Duodenal muscle.
MR; Rectal smooth muscle, PBMC; Peripheral blood mononuclear cell, RM; Rectal mucosa, SC; Sigmoid colon, SI; small intestine, TC; Transverse colon, TR; T-regulatory cell

2.4.1 Loci mapped to a single causal variant

Two of the fifteen PSC risk loci mapped to a single causal variants with $\geq 95\%$ PP of causality. The first single variant credible set was in PSC region 5 (Chr3:71153890), where the GWAS lead SNP, rs80060485 at position 71153890, was predicted to be causal with a PP of 99%. *FINEMAP* strongly supported the presence of a second independent signal within this region with 83% certainty. Signal 2 could not be well fine-mapped with 14% PP of causality for the most probable causal variant, rs36023390 at position 71523093 (Figure 2.3a). The presence of two independent causal variants was supported by the finding that the causal configuration with the highest PP contained both rs80060485 and rs36023390 and that these two SNPs were not correlated ($r^2=0$). *GoShifter* identified that the credible causal variant for signal 1, rs80060485, overlapped promoter and enhancer marks in three immune cell types and ten gastrointestinal tissue types (see Table 2.2).

The fine-mapped causal variant, rs80060485, occurs within an intron of *FOXP1* (forkhead box P1), a transcription factor with an important role in B- and T-cell differentiation. CD4+ T-follicular helper (T-FH) cells, are a specialised T-cell subset found in germinal centres, which interact with B-cells, inducing antibody formation and response. Foxp1 is a negative regulator of T-FH cell differentiation, directly and negatively regulating IL-21 production [165]. Foxp1-deficient CD4+ T cells preferentially differentiate into CD4+ T-FH cells, resulting in substantially enhanced germinal centre and antibody responses. T-FH cells can also be found in the periphery where they are characterised by the expression of chemokine receptor type 5 (CXCR5) and the inhibitory receptor, programmed death 1 (PD-1). Circulating T-FH cells lacking the chemokine (C-C motif) receptor 7 (CCR7), closely resemble lymphoid tissue-derived T-FH cells, that are pathogenic in autoimmunity [166]. Interestingly, the frequency of potentially pathogenic CCR7lowCXCR5+PD-1+CD4+ T-FH cells is increased in patients with PSC, compared to healthy donors [167], suggesting Foxp1 and T-FH cells may have an important role in PSC pathogenesis. Whilst it is not yet clear whether or how the expression of *FOXP1* is affected by the intronic rs80060485 variant, this analysis demonstrates that this variant overlaps several important markers for active enhancers, suggesting a mechanism via which this variant may exert a quantitative effect upon the expression of *FOXP1* or several other genes within the region.

The second locus fine-mapped to a single causal variant with $\geq 95\%$ PP of causality was PSC region 10 (Chr12:111884608). Fine-mapping confirmed that rs3184504, the lead GWAS SNP, was the most probable causal variant with 99% certainty (Figure 2.3b). The rs3184504 SNP, is a multi-allelic missense variant which is positioned within exon 3 of the *SH2B3* (Scr homology 2 adaptor protein 3) gene. The rs3184504*A and rs3184504*C alleles code for a basic polar arginine and the rs3184504*G allele codes for a polar glycine at position 262 in the pleckstrin homology domain of the SH2B3 protein. The minor allele

for this locus, present at a frequency of 15%, is the PSC risk increasing rs3184504*T allele, which codes for a non-polar tryptophan at this position. Analysis of the functional effect of this missense mutation using Ensembl's variant effect predictor (VEP) assigned the rs3184504*C>T SNP a PHRED-like scaled CADD score of 11.08, where a score of ≥10 indicates polymorphisms predicted to be within the 10% most deleterious substitutions in the human genome [168].

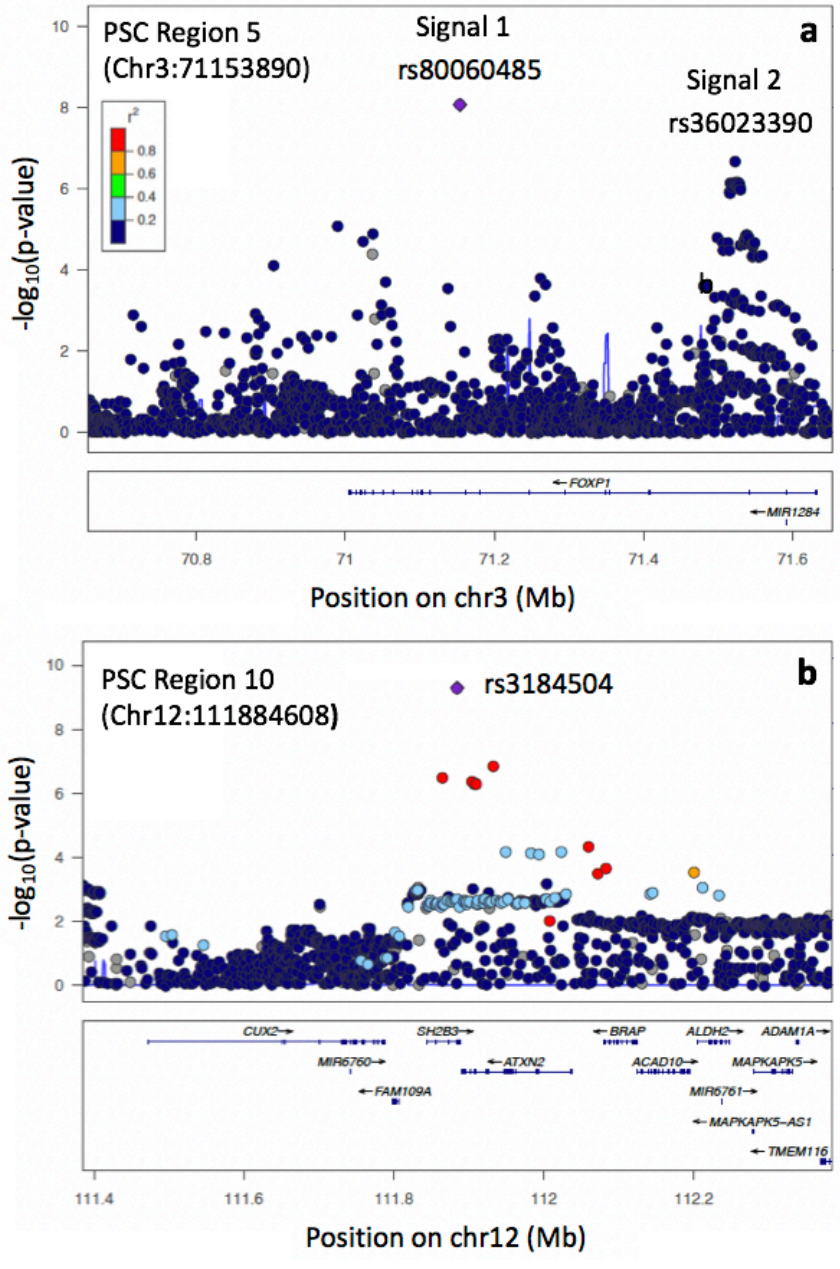


Figure 2.3: Regional association plots for PSC risk loci mapped to single variants.

SH2B3 is an interesting gene in the pathogenesis of PSC, as it is a negative regulator of T-cell activation, TNF production, and Janus kinase (JAK) 2 and 3 signalling. It

is known to encode the T-cell adapter protein LNK, which regulates T-cell receptor-, growth factor- and cytokine receptor-mediated signalling [169]. The *SH2B3* locus is a shared risk locus with several other IMDs and rs3184504 remains the lead SNP in GWAS of coeliac disease (CeD), rheumatoid arthritis (RhA), type 1 diabetes mellitus (T1DM) and autoimmune hepatitis (AIH) [169–171]. Fine-mapping of this risk locus in RhA predicted that rs3184504 was the most probable causal variant for this locus with 76% PP of causality [114]. Expression quantitative trait loci (eQTL) studies have shown that rs3184504 is associated with increased expression of genes involved in IFNγ production [172]. Furthermore, functional investigation of this locus has shown that peripheral blood mononuclear cells isolated from individuals homozygous for the rs3184504*A allele, which increases risk of RhA and T1DM, display increased production of pro-inflammatory cytokines in response to bacterial stimuli compared to individuals homozygous for the non-risk G allele [173]. The same study also suggested that the SH2B3 protein has an inhibiting function on the MDP-NOD2-RIP2 pathway, which responds to bacterial ligands, with disease-associated alleles causing diminished inhibitory activity of SH2B3. Unfortunately, they did not include analysis of individuals homozygous for the minor rs3184504*T allele, which not only increases the risk of PSC, but also of AIH [174], suggesting the resultant Arg262Trp amino acid substitution may contribute to an aberrant immune- and inflammatory-response targeted at the hepato-biliary system.

2.4.2 Variants with a greater than 50% posterior probability of causality

Three signals mapped to credible sets containing more than one variant, where one variant within each credible set had >50% PP of causality. The first was within PSC region 9 (Chr11:64107735), where rs35247680 at position 63884747 was predicted to be causal with 61% PP. This SNP is a non-coding variant within an intron of *MACROD1*. *GoShifter* demonstrated that this variant overlapped promoter marks enriched in four immune cell-types and nine gastrointestinal tissues and overlapped enhancer marks in four immune cell-types and eleven gastrointestinal tissues (Table 2.2), thereby suggesting several mechanisms via which this credible causal variant may regulate expression of nearby genes. There was evidence to support a second independent signal within this region with 68% certainty (Figure 2.4a). The most probable causal variant for signal 2 was the previously reported lead GWAS SNP for this locus, rs663743 at position 64107735. Independence of these two signals was supported by the fact that these variants were not highly correlated with one another (r^2=0.02). The rs663743 SNP is non-coding and within the 5' untranslated region overlapping a promoter region for *CCDC88B* (coiled-coil domain containing 88B). However with 41% PP of causality attributed to rs663743, signal

2 could not be considered fine-mapped.

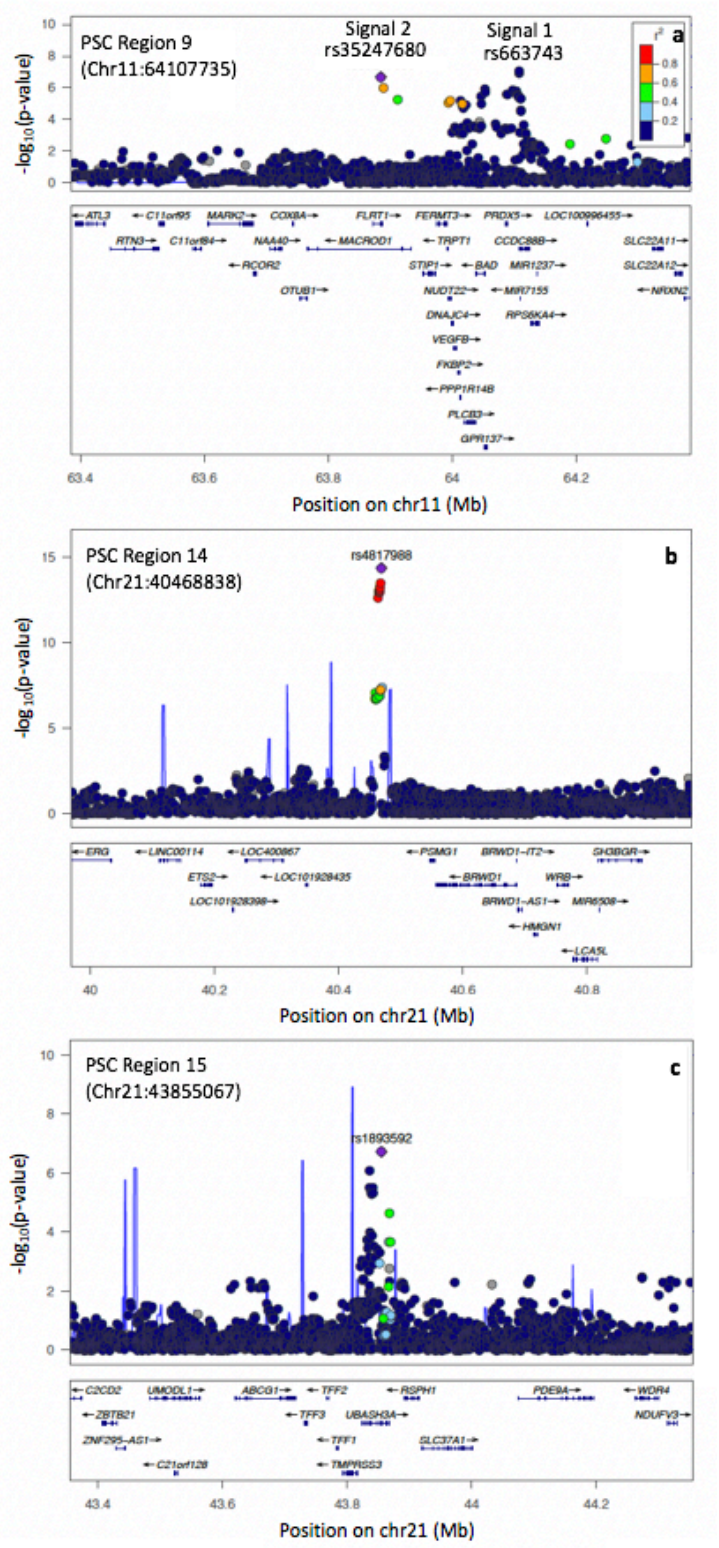


Figure 2.4: Regional association plots for PSC risk loci mapped to casual variants with >50% posterior probability of causality.

The second region mapped to a credible set containing a variant with >50% PP of causality was in PSC region 14 (Chr21:40466744). The lead GWAS variant for this region,

rs4817988, is highly correlated (r^2>0.8) with more than ten other variants of genome-wide significance that lie in close proximity, all with similar measures of association (Figure 2.4b). Fine-mapping of this region identified a 95% credible set containing ten of these high-LD variants with the most probable causal variant for this region, rs4817988 at position 40468838 with 58% PP of causality. The next most probable causal variants were rs2836884 at position 40467643 (8% PP) and rs2836883 at position 40466744 (5% PP). This locus is located 5' upstream of *PSMG1*, which is the commonly quoted candidate gene for this region, based upon its genomic locality and evidence from the study of paediatric IBD colon where levels of PSMG1 were increase compared to healthy colon [175]. Importantly, although rs4817988 is located in a non-coding region, it overlaps a CTCF transcription factor binding site and has been correlation with the expression of *ETS2 (v-ets avian erythroblastosis virus E26 oncogene homolog 2)* in the GTEx eQTL analysis of whole blood [176]. *GoShifter* prioritised rs2836883, followed closely by rs4817988 with the lowest overlap scores and found these variants to overlap promoter marks in three immune cell types and three gastrointestinal tissue types (Table 2.2). This same locus has also been associated with IBD and has been the subject of an IBD fine-mapping study [56]. Huang, Fang, Jostins *et al* resolved the *ETS2* locus to a credible set of 10 variants, with a 39% PP attributed to the most probable variant, rs9977672. In PSC, I mapped this region to a 10 variant credible set, 8 of which overlapped with the IBD credible set for this region, however prioritising a different variant, rs4817988 at position 40468838, as causal with 58% certainty. The two non-overlapping variants within the IBD credible set, one of which is the most probable IBD fine-mapped variant, are both present within the PSC data-set. It is likely that for this region, the same variant is causal is both PSC and IBD, although further analysis is required to validate this hypothesis.

The third locus fine-mapped to a credible set containing a variant with >50% PP of causality was PSC region 15 (Chr21:43855067). Ji *et al* reported an association between the Chr21:43855067 locus and PSC risk, driven by lead SNP rs1893592, and proposed *UBASH3A* as the most likely gene affected by this risk locus on the basis that this SNP was an eQTL of *UBASH3A* in one B-cell only [129] and two whole blood analyses [118, 177] (Figure 2.4c). Fine-mapping of this region confirmed that rs1893592 at position 43855067, which is located three bases downstream of the 10^{th} exon of *UBASH3A* within the splice consensus sequence, was the most probable causal variant in this region with 62% PP of causality. The PSC risk reducing rs1893592*C allele, disrupts the conserved 5' splice donor sequence at this position, and is predicted to cause partial retention of the downstream intron and possible non-stop mediated decay [178]. I fine-mapped this locus to a credible set containing just four additional variants, each located within intronic regions of *UBASH3A*, but with a low individual probability of causality; rs11203203 (14%), rs3788013 (9%), rs9974339 (6%) and rs876498 (6%), and all in low LD (r^2<0.6)

with the most probable SNP, rs1893592. Interestingly, this locus has also been reported as associated with T1DM, where a fine-mapping study has identified the second most probably PSC SNP, rs11203203 at Chr21:43836186, as the most probable causal variant for this locus in T1DM with 39% PP of causality [114]. In this T1DM fine-mapping study, the 95% credible set contained four variants, of which only rs11203203 is contained with both the PSC and T1DM credible sets. Notably, a review of the summary statistics from both data-sets showed that SNPs from both credible sets were considered within both the PSC and T1DM fine-mapping analyses. Whilst it is possible that different SNPs within this same locus may precipitate different IMDs, it is more likely, where the credible sets overlap, that it is the same causal variant responsible for both IMDs. Further work to colocalise the signals in PSC and T1DM at this locus would be helpful to establish a shared causal variant.

2.4.3 Variants with a greater than 20% posterior probability of causality

Fine-mapping of two loci resulted in credible sets containing at least one causal variant with >20% PP of causality. Although these loci could not be considered fine-mapped, a large credible set with >20% PP attributed to one SNP could, in combination with functional annotation, provide useful information about the potential causal variants within a locus. The first locus containing at least one causal variant with >20% PP of causality was PSC region 8 (Chr10:6108139 region). The lead GWAS SNP for this locus, rs4147359 (Chr10:6108139), located upstream of *IL2RA*, was predicted to be the most probable causal variant for this region with 46% PP of causality. The 95% credible set included four other variants, two intergenic and two intronic variants (Figure 2.5a). These variants were found to be enriched for overlapping regulatory regions in PSC-relevant tissues. *GoShifter* identified variants within this locus as potentials for functional follow up with one of the lowest overlaps scores across all credible variants from all non-coding loci observed for rs4147359, also the most probable causal variant from fine-mapping of this locus. This variant is located within an intergenic region and overlaps a marker of active transcription, H3K36me3. This suggests that the mechanism via which rs4147359 may increase PSC risk is through modulation of an active transcription histone acetylation mark, although the downstream gene and direction of effect cannot be identified from either of these analyses. *FINEMAP* could not distinguish whether there were one or two independent signals within the region, with equal evidence for both, although the most probable causal configuration contained just one single variant, rs4147359. Interestingly, this locus has been previously fine-mapped in a study of individual and combined summary statistics for T1DM and RhA [114]. In the combined T1DM and RhA data, this locus

was fine-mapped to a credible set containing 3 SNPs; rs706778 (89% PP), rs7072793 (4% PP) and rs7096384 (3% PP). Reassuringly, two of these T1DM/RhA credible set variants (rs706778 and rs7072793) were also included within the PSC credible set, with 12% and 9% PP of causality respectively. Both PSC and T1DM/RhA GWAS datasets included all SNPs within the credible sets for both of these IMDs. Given that the T1DM/RhA fine-mapping study included data from 11,475 RhA cases and 9,334 T1DM cases, compared to the 4,796 PSC cases analysed in this study, the T1DM/RhA fine-mapping study was better powered to fine-map individual rick loci. Therefore, where fine-mapping of risk loci within one data-set is inconclusive, the sharing of genetic architecture between IMDs means that other fine-mapping studies of the same locus can be informative.

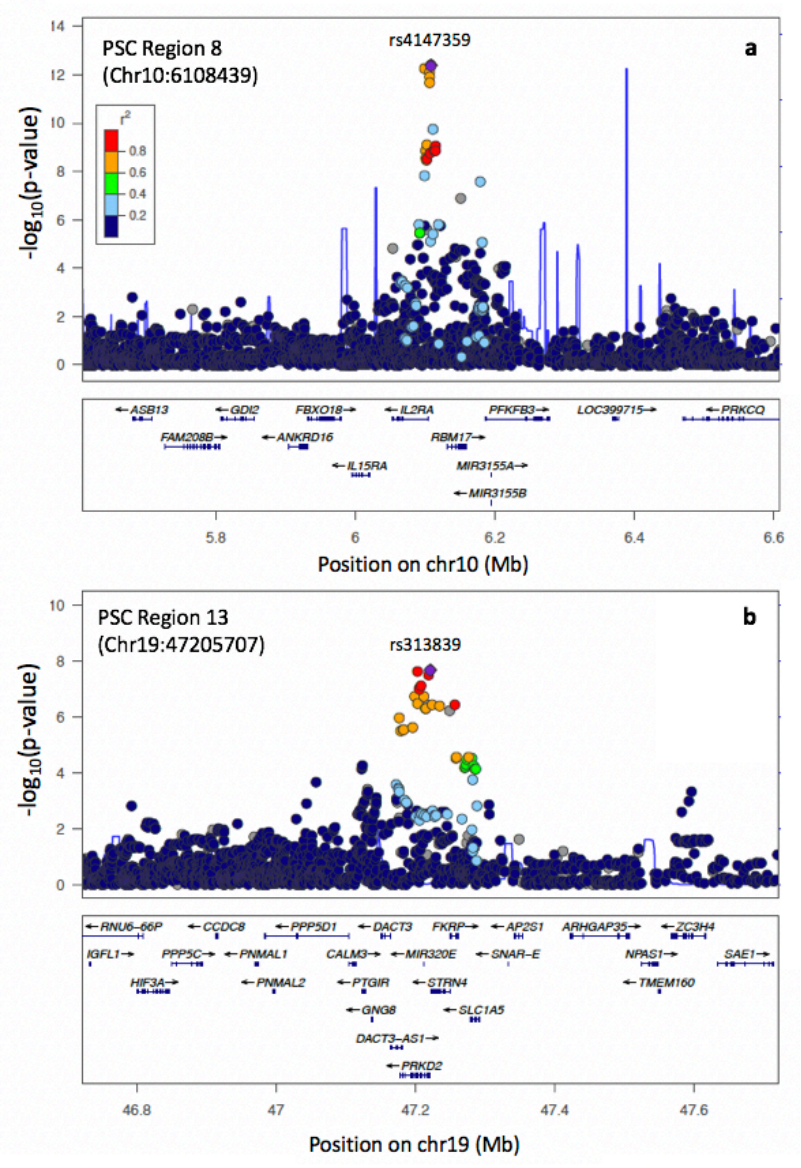


Figure 2.5: Regional association plots for PSC risk loci mapped to casual variants with >20% posterior probability of causality.

The second locus containing at least one causal variant with >20% PP of causality was PSC region 13 (Chr19:47205707). This locus was fine-mapped to a credible set containing fourteen variants, with the most probable causal variant for the region being rs313839 at position 47221557 with 23% PP of causality (Figure 2.5b). The rs313839 variant lies within an intron region that has been associated with binding of many transcription factors though chromatin immunoprecipitation (ChIP)-sequencing studies and overlaps a promoter region for a gene called *PRKD2*. Notably, rs313839 was also prioritised by *GoShifter* as it overlaps transcription activation marks, H3K4me3 and H3K27ac, in CD14+ monocytes. This finding is in keeping with the known role of *PRKD2* in monocyte migration and adhesion [179], [180]. Analysis of the major and minor allele sequences using *PROMO*, which identifies putative transcription factor binding sites [181], showed that rs313839 C>G (where rs313839*G is the PSC risk increasing allele) resulted in loss of binding motifs for transcription factors LEF-1 TCF1A, TCF-4E, DEF:GLO:SQUA, TCF-3 and ADR1, and gain of motifs for EIIE-A, VSF-1, V-MYB and MYB2. Although further investigation is required to identify the genes affected by these transcription factors, the results of this study suggest that the most probable causal variant, rs313839, modulates transcription factor binding, with evidence supporting *PRKD2* as a likely candidate gene for this locus.

2.4.4 Loci not well-resolved with fine-mapping

Several regions could not be fine-mapped to credible sets with the majority of the PP for causality attributed to one variant. However, four regions were resolved to relatively small credible sets. PSC region 3 (Chr2:204612058) was mapped to a six variant credible set, with the majority of the PP for causality (19%) attributed to rs5837875 at position 204647878, a variant located in an transcription factor binding site and associated with expression of *CD28* in one whole blood eQTL analysis [118]. PSC region 2 (Chr2:111933001), region 7 (Chr6:91030441) and region 11 (Chr16:11169683) were each resolved to credible sets of twelve variants, in which the respective most probable causal variants had PPs of 0.18, 0.20 and 0.16. Although PSC region 2 (Chr2:111933001) and region 11 (Chr16:11169683) have been reported as associated with IBD and PBC respectively, the Chr2:111933001 locus has not been considered in either of the published IBD fine-mapping studies [56], [60] and there have been no published fine-mapping studies in PBC, to date. PSC region 7 (Chr6:91030441) has been fine-mapped to a credible set of nine variants in RhA [114]. Whilst the most probable causal variants differed between PSC and RhA, all nine variants within the RhA credible set were contained within the PSC credible set.

Four PSC risk loci were not well resolved with fine-mapping, defined by large credible sets with the most probable causal variant assigned $\leq 10\%$ PP of causality. PSC region 4 (Chr3:49721532) *MST1* (Figure 2.6a) and region 6 (Chr4:123499745) *IL2-IL21* (Figure 2.6b) both contain many variants in tight LD with one another, extending over a wide

genomic region of >500Mb, all with very similar strengths of association. Both of these loci have been associated with IBD, however fine-mapping in IBD was unable to resolve the Chr3:49721532 *MST1* and Chr4:123499745 *Il2-Il21* loci, with credible sets containing 437 and 29 variants respectively, a likely consequence of the extended, complex patterns of LD observed within these regions. PSC region 1 (Chr1:2526746) (Figure 2.7a) and region 12 (Chr18:67543688) (Figure 2.7b) both contain many variants with similar strengths of association, all in tight LD with one another. Under such circumstances *FINEMAP*, which utilises subtle differences in strengths of association between tightly correlated variants, performs less well, and prioritisation of non-coding causal variants using functional annotation of genomic regions become more important to infer causality. The *GoShifter* overlap scores for these two loci were comparatively high, compared to other loci, suggesting *GoShifter* was unable to easily prioritise causal variants from these loci in comparison with other PSC risk loci. However, based upon the variant with the lowest overlap score relative to other credible variants within the same locus, for PSC region 1 (Chr1:2526746) *GoShifter* prioritised rs60733400 at position 2516781 with an overlap score of 0.17. This variant overlaps several regulatory features including H3K27ac, an active enhancer marker in CD14+ monocytes. For PSC region 12 (Chr18:67543688), *GoShifter* prioritised rs4891781 at position 67524646 with an overlap score of 0.14, as it overlapped an H3K9ac mark in peripheral blood mononuclear cells (PBMCs). This variant is in tight LD with the most probable causal variant from fine-mapping, and lies only 10Kb upstream. However in both cases the overlap scores were relatively high (>0.14) suggesting that in comparison to the other PSC risk loci, these loci should not be prioritised for further for functional follow-up. PSC region 12 (Chr18:67543688) has been associated with T1DM, however fine-mapping of locus in T1DM was no more successful, with a reported credible set containing 32 variants [114].

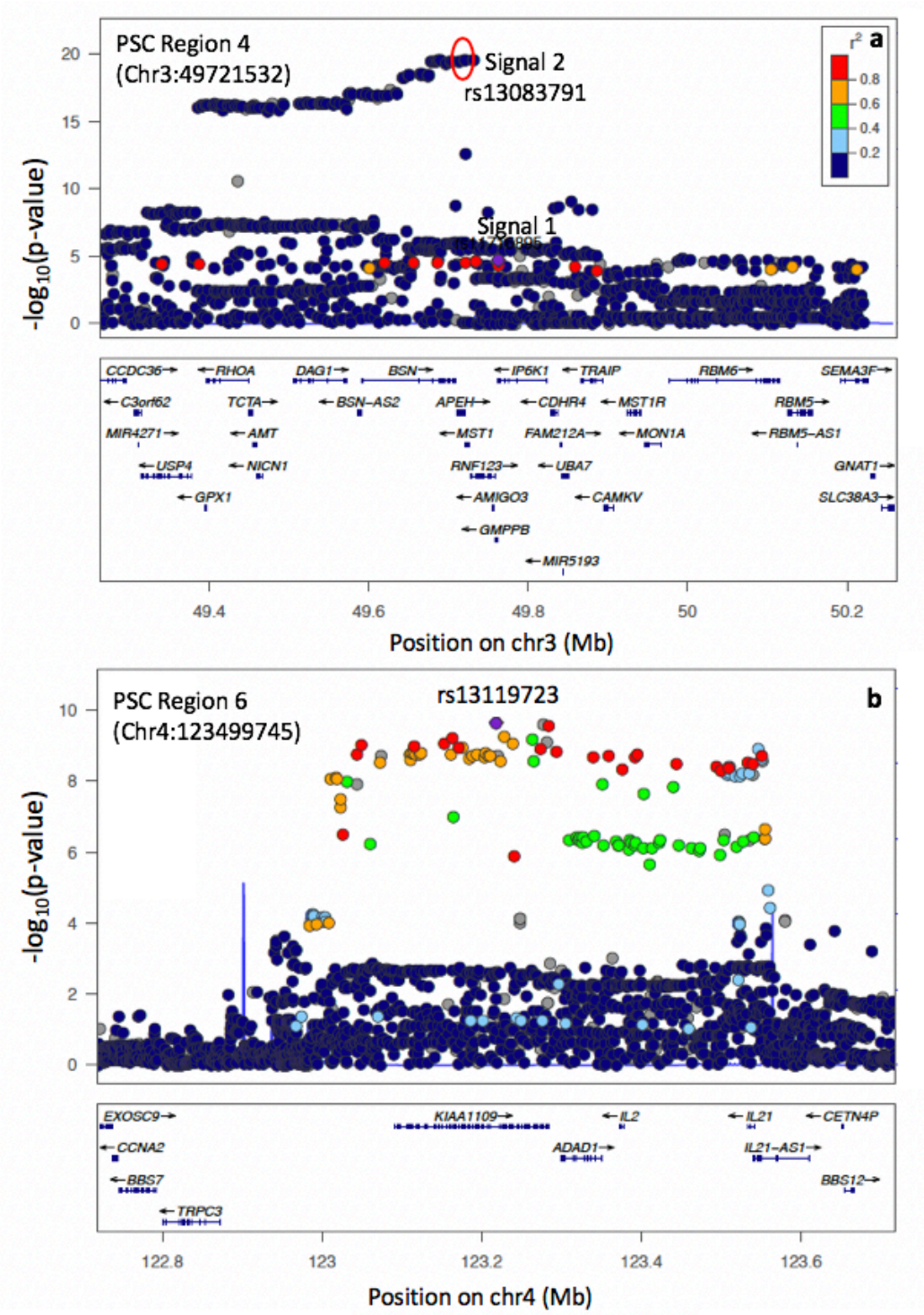


Figure 2.6: Regional association plots for PSC risk loci not well resolved with fine-mapping.

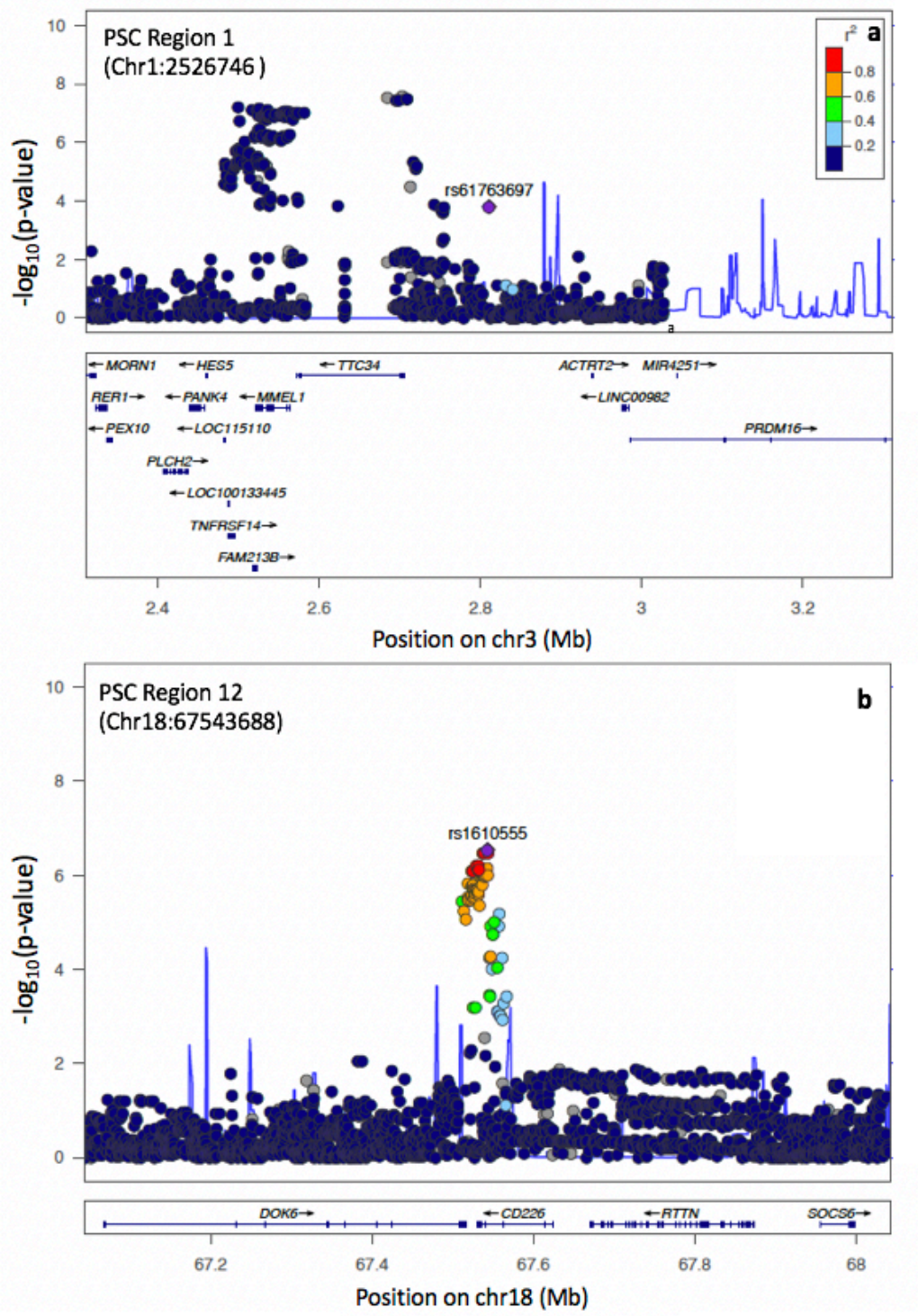


Figure 2.7: Regional association plots for PSC risk loci not well resolved with fine-mapping.

2.5 Discussion

In this chapter I perform the first fine-mapping analysis of risk loci associated with PSC. Using an established Bayesian fine-mapping method I was able to fine-map five of these risk

loci to a credible set containing a variant with >50% PP of causality, two of which were fine-mapped to a single causal variant with >95% PP of causality. Of these five variants, one is enriched for a significant protein-coding change, one is predicted to cause direct disruption to a splice site and three overlap tissue-specific epigenetic marks in PSC-relevant cell- and tissue-types. Stringent prior quality control and filtering of GWAS data means that for common variant associations, the results of this fine-mapping analysis are likely to be robust. This is supported by the finding that PSC credible sets are significantly enriched for variants that overlap enhancer or promoter regions in PSC-relevant tissues and cell-types. This analysis however, illustrates some of the challenges associated with fine-mapping. Only seven loci were mapped to credible sets containing ≤ 10 credible causal variants and for only 5 of these loci was is possible to identify a single causal variant as a promising candidate with >50% PP. For several loci, the presence of large credible sets with multiple plausible causal variants, each with low PPs of causality means that functional annotation is essential for prioritisation. The generation of precise annotation maps in disease-relevant tissues will therefore be crucial to our ability to further interpret these risk loci. Nevertheless, the identification of causal variants for even just a few loci remains a valuable outcome.

Precise fine-mapping should frequently point to the same variant in different diseases with shared risk loci. IBD remains the disease which shares the most genetic architecture with PSC. Of the 15 PSC risk loci fine-mapped within this study, it has been previously reported that five loci; Chr3:49721532 *(MST1)*, Chr4:123499745 *(Il2-Il21)*, Chr12:111884608 *(SH2B3)*, Chr18:67543688 *(CD226)* and Chr21:40466744 *(ETS2)*, demonstrate strong evidence for a shared causal variant with IBD [42]. Three of these loci, Chr3:49721532 *(MST1)*, Chr4:123499745 *(Il2-Il21)* and Chr21:40466744 *(ETS2)*, have been the subject of fine-mapping in IBD [56, 60]. Huang, Fang, Jostins *et al* resolved the *ETS2* locus to a credible set of ten variants, with a 39% PP attributed to the most probable variant, rs9977672. In PSC, I fine-mapped this region to a ten variant credible set, eight of which overlap with the IBD credible set for this region, however prioritising a different variant, rs4817988 at position 40468838, as causal with 58% certainty. The two non-overlapping variants within the IBD credible set, one of which is the most probable IBD fine-mapped variant, are both present within the PSC dataset. It is likely that for this region, the same variant is causal is both PSC and IBD, supporting a higher prior in any future fine-mapping studies and consideration of these two additional IBD credible variants in any future functional follow-up studies. Fine-mapping in IBD was however unable to resolve the Chr3:49721532 *MST1* and Chr4:123499745 *Il2-Il21* loci, with credible sets containing 437 and 29 variants respectively, a likely consequence of the extended, complex patterns of LD observed within these regions. This is an important negative finding, as for those regions not well resolved by fine-mapping in PSC, it has been suggested that the future

use of larger GWAS sample sizes will enable the statistical resolution of more risk loci, due to the ability to better distinguish subtle difference in LD between tightly correlated variants. Sample sizes in IBD GWAS dwarf those of PSC, with the numbers of subjects now approaching 60,000. However despite this, for these high LD regions, fine-mapping was no more successful in the larger samples sizes of IBD. One remedy to this might be to leverage LD from other ethnicities by undertaking GWAS in populations with different LD structure to improve fine-mapping resolution [113]. However, to date, GWAS in PSC have only included individuals of European ancestry, and the number of non-European individuals included in IBD GWAS is comparatively small.

Several of the PSC risk loci in this study have been reported as risk loci and fine-mapped in RhA and T1DM. Similar to IBD, this provides an important means of verifying the precision of these PSC fine-mapping results and where PSC fine-mapping has not resolved a locus to a small credible set, it provides the opportunity to review fine-mapping results from IMDs with larger sample sizes and thus greater power to differentiate between highly correlated SNPs. However, whilst we know there is significant sharing of risk loci between IMDs, to date, there have been no studies that determine shared risk loci between PSC and other IMDs, outside of IBD. In order to conclusively prove that a risk locus is shared between two traits, and thus the results from the fine-mapping of one trait are applicable to both traits, some form of statistical analysis is required. Colocalisation is a statistical means of assessing the probability that the signal observed in two traits e.g. a PSC risk locus and a T1DM risk locus, is driven by the same causal variant. Whilst colocalisation does not define which is the true causal variant for each colocalising trait, in combination with fine-mapping it provides a powerful means of determining shared genetic architecture and resolving causal variants. This is an analysis explored in the following chapter of this thesis.

An important step in using genetic risk loci to further our biological understanding of disease causation is to identify the genes impacted. When the fine-mapped variant falls within a coding region of the genome, this can be relatively straightforward. For example, fine-mapping of PSC region 10 (Chr12:111884608) identified rs3184504, a missense variant located in exon three of *SH2B3*, as the most probable causal variant with 99% certainty. *SH2B3* is an interesting gene in the pathogenesis of PSC, as it is a negative regulator of T-cell activation, TNF production, and Janus kinase (JAK) 2 and 3 signalling, with several studies exploring the allele-specific effects of this SNP on immune- and inflammatory-response and subsequent risk of IMD. However, the majority of genetic associations with IMD fall within non-coding genomic regions, and PSC is no exception. Of the 15 PC-risk loci fine-mapped in this study, only two loci were fine-mapped to variants within a coding region. For the remaining 13 loci, identifying the precise genes impacted by the non-coding variants, and the direction of effect on gene expression remains challenging. For example,

fine-mapping and annotation of the Chr19:47205707 *PRKD2* locus supported rs313839 at position 47221557 as the most probable causal variant. The PSC risk increasing allele at this position is predicted to cause direct disruption to a binding site for multiple transcription factors. However, identifying the gene affected by this mutation, and whether the PSC risk increasing allele results in increased or decreased expression of that gene is not possible from fine-mapping and annotation alone.

One means of identifying the genes affected by the many non-coding risk loci is via colocalisation analysis with variants that exert a quantitative effect upon gene expression (eQTL). Colocalisation can be performed between any two types of traits, for example two disease risk loci, or a disease risk locus and an eQTL locus. For example, the Chr19:47205707 *PRKD2* locus is also a T1DM locus, which has been shown to colocalise with an eQTL in monocytes [182]. It is likely that the gene affected by this same PSC locus is also *PRKD2*, however colocalisation is a statistical means of measuring the probability that this is true. Colocalisation with eQTLs allows us to identify the genes impacted by non-coding risk loci, in addition to identifying the direction of effect a particular disease risk allele has upon downstream gene expression. Following on from fine-mapping, an important next step to infer biological understanding from genetic risk loci in PSC is therefore to identify the genes impacted, by colocalisation with eQTLs mapped in relevant cell types, an analysis which is presented in the next chapter of this thesis.

Fine-mapping of genomic regions associated with disease risk is an important step in understanding the biological mechanisms via which risk variants exert their effect to cause disease. Through fine-mapping, we can filter the often many hundreds of potential causal variants within a locus to a single variant or set of variants responsible for the observed association. In many cases, functional annotation of these credible sets gives us insight into the mechanisms via which they alter gene expression and thus the biological pathways that may be important in disease causation. Colocalisation with eQTLs measured in disease-relevant cell-types and tissues is an important next step for identifying those genes, cell-types and biological pathways affected by disease risk loci, and will bring us one step closer to understanding the causal biology of PSC.

Chapter 3

Statistical colocalisation of Primary Sclerosing Cholangitis risk loci with functional quantitative trait loci

3.1 Introduction

The majority of genetic variants associated with complex disease risk are located within non-coding regions of the genome. In the quest to unravel the function of non-coding risk variants, our next challenge is to identify the precise genes upon which they impact. It is now understood that many non-coding risk variants exert their influence via epigenetic gene regulatory mechanisms and exert a quantitative rather than a qualitative effect upon gene expression. Variation in gene expression is therefore an important mechanism underlying susceptibility to complex diseases. Expression quantitative trait loci (eQTL) are genetic variants that exert a quantitative effect upon gene expression, i.e. the abundance of a gene transcript is directly modified by a genetic polymorphism, usually within a regulatory element. In recent years eQTL mapping methods have been developed, which test the association between genetic polymorphisms and transcript abundance by assaying gene expression and genetic variation on a genome-wide scale, in a large number of individuals. Similar to any complex trait, the abundance of a gene transcript is a quantitative trait that can, with a sufficient sample size, be mapped with considerable power [116]. Variants associated with complex diseases are demonstrably enriched for eQTLs [117]. Nicolae *et al* have shown that SNPs associated with complex traits are significantly more likely to be eQTLs than MAF–matched SNPs chosen from high-throughput GWAS platforms that are not associated with complex traits. Investigating eQTLs in the functional study of genetic risk loci associated with complex diseases such as PSC therefore remains a priority.

In order to further investigate the mechanism via which non-coding genetic variants drive risk of complex disease, one challenge has been the integration of complex trait

association data with eQTL data to measure the plausibility of a shared causal variant between the two traits. Over the past decade, several methods have been developed to try and address this challenge. One of the first methods of assessing whether two traits shared a causal variant was by crude visual comparison of the overlap between two signals. A number of computational tools were developed to facilitate the visual comparison of trait-associated and gene-expression data [183]. For example, a study exploring eQTL data for a particularly gene-dense region on chromosome 17q23 strongly associated with susceptibility to asthma [184], found by visual comparison, that the same asthma-associated variants also had highly significant effects on the expression of *ORMDL3* [185]. However, observation of visual overlap cannot prove a causal relationship between, for example, *ORMDL3* and asthma because the abundance of eQTLs throughout the human genome make the chance finding of an overlap highly likely [186]. Indeed, inference about shared causality between two traits requires a more robust statistical assessment of colocalisation.

Plagnol *et al* proposed a 'proportionality-testing' method which tests a null hypothesis of proportionality of regression coefficients for any set of SNPs across two traits, with the assumption that where there are multiple causal variants, these are shared between both signals [187]. However it has been subsequently demonstrated that this method is biased as a result of having to specify a subset of SNPs on which to base the analysis [188]. Moreover these, and other methods, reliant on individual level genotype data have become impractical with the development of collaborative consortia facilitating the meta-analysis of GWAS data from increasingly large sample sizes. In 2014, Giambartolomei *et al* published *Coloc*, a method to test for colocalisation between two pairs of traits, which overcomes many of these shortcomings by using a Bayesian model with single-SNP summary statistics [189]. *Coloc*, discussed further in the following Methods section, assesses the plausibility of a single shared causal variant driving two traits, requiring densely-genotyped or well-imputed summary statistics that have undergone stringent QC. *Coloc* bases its analysis upon all SNPs within a locus, assuming each SNP is *a priori* equally likely to affect the traits under analysis. Furthermore it estimates the posterior probability (PP) for five different hypotheses ranging from no shared genetic variation between two traits within a region (PP0), to shared genetic variation with the same causal variant driving each signal (PP4). *Coloc* can be applied to any two pairs of traits, including disease traits or functional (epigenetic) traits such as eQTL, histone acetylation marks (histQTL) and methylation marks (methQTL). *Coloc*, and other methods using a similar statistical approach have become the singular method of analysis for performing colocalisation between genetic traits.

Gene expression is the subject of both global and local regulatory variation, i.e. there are eQTLs which act across multiple tissues, in addition to tissue-specific regulatory variation [125]. Colocalisation between disease-associated risk loci and functional traits,

therefore requires careful consideration of the tissue, cell-type or activation state in which the functional trait has been measured. However, identifying the relevant cell-type or stimulated state in which an eQTL is active remains challenging, as demonstrated by several studies, which have sought to address this through the mapping of eQTLS across multiple cell types challenged with multiple stimuli [129, 130]. Importantly, it has been demonstrated that eQTLs are enriched for disease-associated variants in disease-relevant cell- or tissue-types [190, 191]. For example, a recent IBD GWAS and colocalisation study found that a chromosome 2 IBD risk locus co-localised with an eQTL that increased expression of integrin α4 in stimulated monocytes, an eQTL that was not active in unstimulated monocytes [60]. Furthermore, this pathway is already the target of successful therapeutic blockade in IBD, by Vedolizumab, a monoclonal antibody to the α4β7 integrin which inhibits T-cell trafficking to the gut mucosa [83, 84]. Therefore, in order to unravel the molecular basis of disease-specific risk loci, the evidence supports the preferential use of eQTLs measured in disease-relevant tissues for colocalisation. However, paucity of published eQTL data means that colocalisation with eQTLs in mechanistically-related tissue/cell types may be limited by data availability. One interesting finding of a study combining RNA-seq with ATAC-seq (Assay for Transposase-Accessible Chromatin using sequencing) data, found that the majority of stimulus-specific eQTLs with a detectable effect upon chromatin accessibility also altered chromatin accessibility in the unstimulated state [134]. On this basis, colocalisation with other functional QTL, for example chromatin accessibility, histone modification or DNA methylation, may indicate the presence of an eQTL in another (unstudied) stimulation state, in addition to revealing the epigenetic mechanism via which disease-associated risk variants may influence gene expression. Therefore in order to fully understand the functional mechanisms underlying GWAS association signals using colocalisation, it is important to examine the relevant cell type, in the right state of activation, at the right time.

3.2 Chapter overview

Colocalisation is one means of identifying the mechanistic impact of non-coding disease risk loci, by examining whether the same non-coding variant is responsible for regulation of gene expression (i.e. is an eQTL). In this chapter I perform colocalisation between PSC risk loci and functional QTL in multiple immune cell- and gastrointestinal tissue-types. Genetic variation is often shared between several immune mediated diseases (IMDs), implicating the same genes and biological pathways as causal mechanisms for autoimmunity. I therefore perform colocalisation between PSC risk loci and other IMDs to identify risk loci that are PSC-specific and those that are shared. Genetic variants tend to exert a greater effect upon gene expression than upon risk of complex disease. For those risk loci that colocalise with

other regulatory or functional QTL traits, I harness this increased power by fine-mapping the colocalising functional QTL data, in an effort to further refine the fine-mapping results presented in the previous chapter.

3.3 Methods

3.3.1 Colocalisation analysis

To test the plausibility of a single shared causal variant between each of the 22 PSC risk loci, and the same regions in multiple functional QTL and IMD GWAS data sets, I implemented Bayesian tests of colocalisation using R package *Coloc* [189]. Specifically, I used the *coloc.abf* function as it implements approximate Bayes Factor colocalisation methods which can be applied to per-SNP summary statistics. I used full summary statistics for the largest PSC GWAS [42] and the datasets outlined in Table 3.1. The 22 genomic loci for colocalisation were defined as 1Mb regions of interest centred on the most associated or 'lead' PSC GWAS SNP for each locus.

Coloc requires per-SNP summary level data for each of the two input traits. This must consist of all variants within the locus, including those variants that did and did not reach the predetermined threshold for genome-wide significance or false discovery rate (FDR). Colocalisation can be conducted using different combinations of input data for each trait to approximate Bayes factors, depending upon the data available. The first combination of input data includes per-SNP p-values and MAF, sample size and ratio of cases:controls (if using a case-control trait). The second combination includes per-SNP regression coefficients (β) and the variance of these regression coefficients (SE^2), in addition to sample size and ratio of cases:controls. Where available, I used regression coefficients and their variance in preference to p-values and MAFs to approximate Bayes factors, as the former combination is more accurate when using imputed data. Where data availability meant that p-values and MAF were used to approximate Bayes factors, I preferentially used the MAF derived from the same dataset under investigation. Where study-specific MAF data was not available, I used the MAF derived from the UK10K reference panel, as all data-sets included only individuals with European ancestry and thus this was the reference panel that best represented the study population. To interpret the direction of effect of an eQTL on gene expression in the context of the PSC risk allele, I matched eQTL and GWAS reference alleles for all loci. To minimise the chance of combining the wrong alleles, I discarded all A/T and C/G variants that had MAF>0.45.

In this Bayesian method of colocalisation, binary vectors representing a sequence of SNPs by whether each individual SNP is causal (1) or not (0) are paired, with each binary vector representing one trait, and pre-assigned to one of five hypotheses (H0, H1, H2, H3,

H4);

H0: No SNP is associated with either trait.
H1: A SNP is associated with trait 1 (PSC), but no SNP is associated to trait 2 (IMD or eQTL)
H2: A SNP is associated with trait 2 (IMD or eQTL), but no SNP is associated to trait 1 (PSC).
H3: Both traits are associated with genetic variation in the region, but this is driven by different causal variants.
H4: Both traits are associated with genetic variation in the region and share the same causal variant.

For each PSC risk locus tested, the probability of the data for each hypothesis is calculated and the aggregate support (probabilities) for each hypothesis combined with the prior probability, to obtain posterior probabilities for each hypothesis (PP0, PP1, PP2, PP3, PP4). The *Coloc* method uses approximate Bayes factors. Bayes factors are summary measures for the ranking of associations, similar to p-values and are defined as the ratio of the probability of the data under the null and alternative hypotheses [192]. Bayesian methods require the definition of prior probabilities for all five hypotheses. In line with recommendations made by the authors, for GWAS/eQTL analyses I set prior probabilities to 1×10^{-4} for individual trait associations and 1×10^{-6} for the probability of a SNP being associated with both QTL and PSC traits (denoted as the p^{12}). In a study of shared genetic variation between four IMDs (not including PSC or IBD) Fortune *et al* suggested that the selection of priors for colocalisation between two IMD traits should be set at a less stringent threshold between 1×10^{-5} and 1×10^{-6} for the prior probability of a SNP being associated with both traits (p^{12}) [193]. This is due to the expectation of more shared genetic variation between loci of IMDs. In their study, whilst the choice of p^{12} did not change which diseases were associated, the posterior odds for H3:H4 did vary with p^{12}. To inform the choice of priors for colocalisation between PSC and the other IMDs in this study, I tested how varying the prior may impact upon the results of colocalisation. I performed colocalisation between PSC and UC (the IMD expected to show the most genetic overlap with PSC), varying the p^{12} from 1×10^{-4} to 1×10^{-7} and examined the weights of the resulting PP3:PP4.

I performed colocalisation for each of the twenty-two PSC risk loci with the data-sets outlined in Table 3.1. I focused on loci for which the PP for the H4 hypothesis (PP4) was >80%, and subsequently refer to this as evidence of colocalisation when reporting results. I also noted regions for which the PP for the H3 hypothesis (PP3) was >80%, which suggests shared genetic variation between two traits, but a different causal variant driving each signal. Finally I noted regions for which PP4 did not reach the 80% threshold,

but where some of the PP had been attributed to PP0, PP1 or PP2, as this can, in the presence of a low PP3, indicate a loss of power to detect colocalisation.

Coloc makes a number of important assumptions. Firstly it assumes that the two traits undergoing colocalisation have been measured in two datasets of unrelated individuals. The method also assumes that the individuals in both datasets are of the same ethnicity and thus the MAF and LD structure are identical. Because the PSC GWAS data set is derived from individuals of European ancestry, only functional QTL and IMD GWAS data derived from European individuals could be included in this analysis. Resultantly, I excluded one large eQTL meta-analysis of whole blood from 32,000 individuals of many ethnicities [194]. A third *Coloc* assumption is that the true causal variant is included within each set of SNPs, requiring that the dataset for each trait is densely-genotyped or well-imputed. In situations where the true causal variant is not present within both datasets, this tends to result in a decrease in the resulting PP4 statistic. The final assumption of this method is that there is, at most, only one independent association for each trait within the region of interest. It is however not uncommon for genomic regions to contain more than one independent association signal. Indeed, fine-mapping of the PSC GWAS data from the previous chapter supported the presence of 19 independent signals across the 15 fine-mapped PSC risk loci. For those regions in which there is more than one independent signal, *Coloc* considers only the strongest of these distinct association signals.

3.3.2 Functional QTL data

Colocalisation of disease-associated risk loci with functional QTLs requires careful consideration of the choice of cell-type or tissue in which the functional QTL trait has been measured. Those tissues potentially relevant to PSC could be any whole-tissue or cell-type from the gastrointestinal or hepato-biliary systems, or any immune-cell type. To find published and un-published eQTL data for inclusion in my analysis, I performed a literature search of existing eQTL studies. From this, I gathered together 42 functional QTL data-sets covering five gastrointestinal whole tissues, six immune-cell types and five different functional traits including gene-expression (*cis*-eQTL), histone marks (histQTL), DNA methylation (methQTL) and splice site QTL (spliceQTL) data (Table 3.1). All data included for colocalisation in this analysis had been subject of prior QC conducted by the publishing authors.

Datasets used for colocalisation included functional QTL data from the Blueprint epigenome project phase 2 data release [195]. The Blueprint epigenome project is a large-scale research project which aims to generate at least 100 reference epigenomes for distinct haematopoietic cell-types in health and common autoimmune diseases (not including PSC or IBD). Blueprint have isolated CD14+CD16- monocytes, CD45+CD66b+CD16+ neutrophils and CD3+CD4+CD45RA+ naïve T-cells from the peripheral blood of between

Table 3.1: Characteristics of data-sets included in colocalisation analysis

data-set	Tissue type / GWAS	Trait	Condition	Sample size
GTEx v7	Liver	eQTL	unstimulated	153
	Transverse Colon	eQTL	unstimulated	246
	Sigmoid Colon	eQTL	unstimulated	203
	Terminal Ileum	eQTL	unstimulated	122
	Whole Blood	eQTL	unstimulated	369
	EBV-Transformed Lymphocytes	eQTL	unstimulated	117
Blueprint	Naïve T cells	eQTL	unstimulated	171
	(CD3+CD4+CD45RA+)	Methylation	unstimulated	133
		H3K4me1	unstimulated	104
		H3K27ac	unstimulated	142
		PSI	unstimulated	171
Blueprint	Neutrophils	eQTL	unstimulated	192
	(CD45+CD66b+CD16+)	Methylation	unstimulated	197
		H3K4me1	unstimulated	173
		H3K27ac	unstimulated	174
		PSI	unstimulated	192
Blueprint	Monocytes	eQTL	unstimulated	194
	(CD14+CD16-)	Methylation	unstimulated	196
		H3K4me1	unstimulated	172
		H3K27ac	unstimulated	162
		PSI	unstimulated	194
Glinos et al, unpub	T regulatory cells	eQTL	unstimulated	123
	(CD3+CD4+CD25highCD127-)	H3K4me3	unstimulated	73
		H3K27ac	unstimulated	91
		ATAC	unstimulated	88
Panousis et al, unpub	Macrophages	eQTL	CIL (6 and 24 hrs)	83
	(derived from iPS cells)	eQTL	Ctrl (6 and 24 hrs)	81
		eQTL	IFNB (6 and 24 hrs)	84
		eQTL	IFNG (6 and 24 hrs)	84
		eQTL	IL4 (6 and 24 hrs)	85
		eQTL	LIL10 (6 and 24 hrs)	75
		eQTL	MBP (6 and 24 hrs)	44
		eQTL	P3C (6 and 24 hrs)	86
		eQTL	PIC (6 and 24 hrs)	44
		eQTL	PIC (6 and 24 hrs)	45
		eQTL	Prec (Day 0 and 2)	42
		eQTL	R848 (6 and 24 hrs)	83
		eQTL	sLPS (6 and 24 hrs)	81
Kim-Hellmuth et al, 2017	Monocytes	eQTL	unstimulated	134
	(CD14+)	eQTL	LPS (90' and 6hrs)	134
		eQTL	RNA lipofectamine (90' and 6hrs)	134
		eQTL	MDP (90' and 6hrs)	134
Astle et al, 2016	Lymphocyte counts	GWAS		173,480
	Monocyte counts	GWAS		173,480
	Neutrophil Counts	GWAS		173,480
De Lange et al, 2017	Ulcerative colitis	GWAS		12,160
	Crohns Disease	GWAS		12,160
Cordell et al, 2015	Primary Biliary cirrhosis	GWAS		2,764
Bradfield et al, 2011	Type 1 Diabetes	GWAS		9,934
Trynka et al, 2011	Coeliac Disease	GWAS		12,041
Okada et al, 2012	Rheumatoid Arthritis	GWAS		29,880
Beecham et al, 2013	Multiple Sclerosis	GWAS		14,802
Bentham et al, 2015	Systemic Lupus Erythematosus	GWAS		7,219

100-200 healthy adults, followed by epigenomic analyses [196]. These include gene expression, CpG methylation, H3K4me1, a marker for active and poised enhancers, H3K27ac, a marker for active enhancers and active promoters and percentage splice index (PSI) which provides the inclusion level of each exon, indicating perturbation of a splice site. I also included published data from the Genotype-Tissue Expression (GTEx) Consortium v7 [197]. GTEx is an established data resource and tissue bank for the study of the relationship between genetic variation and gene expression in multiple human post-mortem tissues. Included within the GTEx database are whole tissue *cis*-eQTL maps for PSC-relevant tissues including liver, transverse and sigmoid colon, terminal ileum, whole blood and Epstein Barr Virus (EBV)-transformed lymphocytes (immortalised B-cells) isolated from between 100-400 individuals. To try and capture colocalisations with eQTLs only active in the stimulated state, I included published data from an eQTL study of CD14+ monocytes derived from 134 healthy individuals and stimulated with microbe-associated molecular patterns; lipopolysaccharide (LPS), RNA lipofectamine and muramyl dipeptide (MDP) [198]. Two sets of unpublished data, were also included for colocalisation. The first was an eQTL dataset measured in induced pluripotent stem cell (iPSC)-derived macrophages differentiated from the skin fibroblasts of up to 85 healthy donors, and exposed to 13 different states of stimulation. These included stimuli mimicking bacterial, viral and allergic response, and measured at 6 and 24 hour time-points (data kindly provided by Dr Nikolaus Panousis, Postdoctoral Fellow at the Wellcome Trust Sanger Institute). The second was data from an analysis of unstimulated T-regulatory cells (CD3+CD4+CD25highCD127-) derived from the peripheral blood of 70-125 healthy individuals and subject to RNAseq, ChIP-Seq and ATAC-seq (data kindly provided by Dr Daphne Glinos, former PhD student at the Wellcome Trust Sanger Institute). Finally, in order to identify PSC risk loci that colocalised with other IMDs, I downloaded summary statistics for the largest available GWAS study for each of eight IMDs from the GWAS catalogue (https://www.ebi.ac.uk/gwas/). These were UC and CD [60], Primary Biliary Cholangitis (PBC) [199], Type 1 Diabetes (TIDM) [200], Coeliac disease (CeD) [201], Rheumatoid arthritis (RhA) [148], multiple sclerosis (MS) [202] and systemic lupus erythematosus (SLE) [203]. I also conducted colocalisation between PSC risk loci and risk loci associated with lymphocyte, neutrophil and monocyte counts from a GWAS of human blood cell trait variation [204].

3.3.3 Fine-mapping of functional QTL loci

In the previous chapter I presented the results of fine-mapping the PSC risk loci. Fine-mapping is influenced by several factors including the sample size of the cohort, the effect size, the MAF and thus the strength of association of the variants within the locus. One of the challenges of studying a rare complex disease such as PSC is that amassing the GWAS samples sizes comparable to more common IMDs such as T1DM and IBD is not

feasible. Genetic variants tend to exert a greater effect upon gene expression than upon complex disease risk. Therefore, where colocalisation proves that a GWAS trait shares a causal variant with a functional trait, there will often be more power to fine-map within the functional QTL data and resolve the locus to a single causal variant, or small set of credible variants. With the aim of improving upon the fine-mapping of the PSC risk loci described in Chapter 2, I developed the following workflow pipeline (Figure 3.1). For each PSC risk locus I conducted fine-mapping in the PSC GWAS data (Chapter 2), followed by colocalisation with the multiple functional QTLs listed in Table 3.1. Where I observed a PSC-QTL colocalisation, I then fine-mapped the colocalising functional QTL data, using the same methods as described in Chapter 2. Fine-mapping requires an LD matrix, ideally calculated from the original genotype data rather than a reference panel [159], I therefore conducted fine-mapping in those functional QTL traits for which full genotype data was available for the calculation of SNP correlation matrices. Functional trait fine-mapping was therefore limited to the Blueprint data and Glinos *et al's* T-regulatory QTL data.

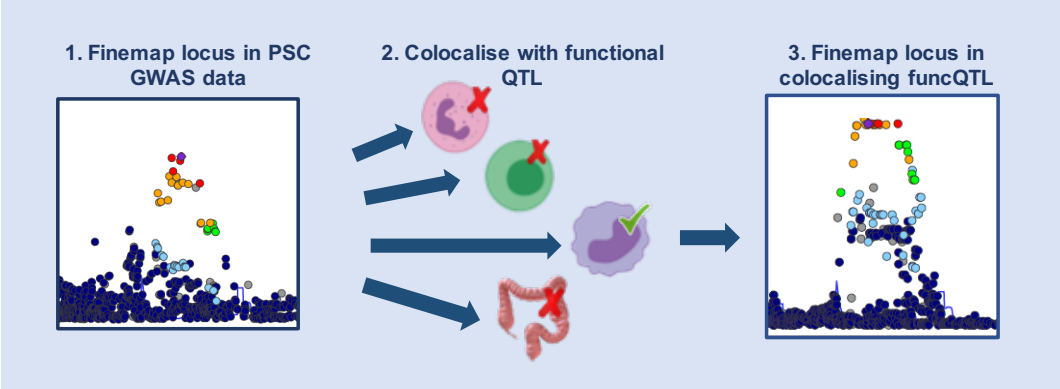


Figure 3.1: Schematic diagram of the GWAS fine-mapping - colocalisation - functional-trait fine-mapping pipeline to resolve the causal variants driving PSC risk loci, and the genes they perturb.

Throughout the analyses described in this chapter, all SNPs are referred to according to their RSID, and all base pair (bp) positions are reported according to Ensembl build 37. For ease of reference, all loci are referred to according to their chromosome and bp position (b37) of the most probable causal SNP from fine-mapping in Chapter 2 and where possible, the gene identified by colocalisation. Where a gene has not been identified by colocalisation with an eQTL, I use the GWAS candidate gene, stipulating where a causal association between a locus and a gene is proven and where it is not.

3.4 Results

I performed colocalisation analysis between the twenty-two non-HLA PSC risk loci and eight other IMDs (Table 3.2). To inform my choice of priors for this analysis, I first tested how varying the prior impacted upon the PP3 and PP4 weights. I performed colocalisation between PSC and UC, varying the p^{12} (prior probability for a SNP being associated with both traits) from 1×10^{-4} to 1×10^{-7}. For 7 PSC risk loci, Figure 3.2 demonstrates how varying the p^{12} changes the weights for PP3 and PP4. Fortune *et al* previously recommended a p^{12} threshold somewhere between 1×10^{-5} and 1×10^{-6}. Although the weights for PP3 and PP4 varied with a p^{12} of 1×10^{-5} and 1×10^{-6}, the results of colocalisation (number of loci for which PP4>80%) were the same. I therefore chose to retain the more stringent of the two p^{12} thresholds, which was set at 1×10^{-6} for all subsequent colocalisation analyses.

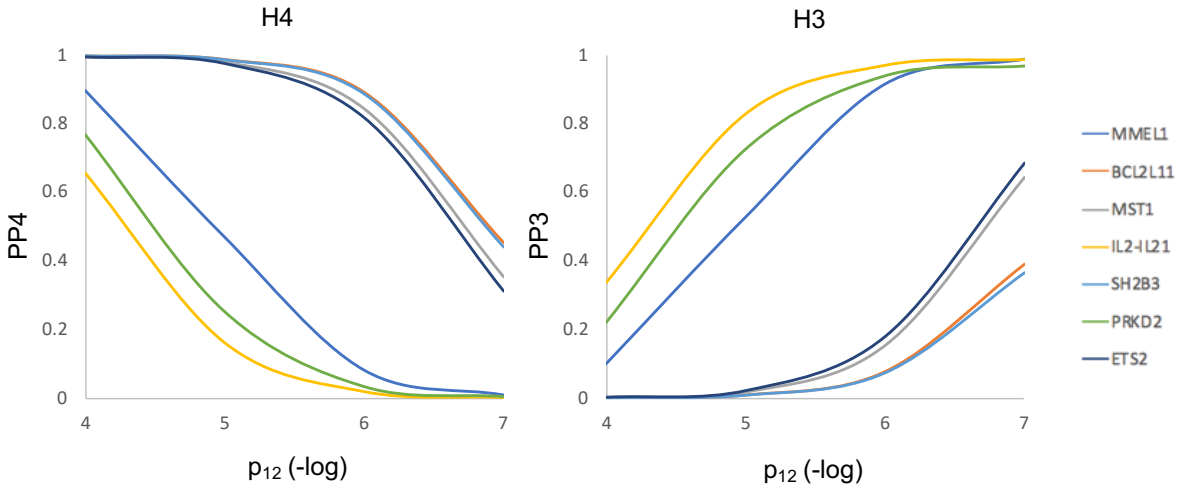


Figure 3.2: Colocalisation between seven PSC risk loci with UC and the evidence for PP4 and PP3 with varying p^{12}.

For those seven risk loci not reaching genome-wide significance in Ji *et al's* data, the results consistently supported no evidence of colocalisation and therefore the results for these seven loci are not subsequently shown. Supporting previous observations of shared genetic architecture between IMDs, eleven of the remaining fifteen PSC risk loci, colocalised with at least one other IMD with PP4>80%. I observed the largest number of colocalisations between PSC and UC, a finding that was expected due to the genetic overlap between PSC and IBD (particularly UC). Four loci colocalised between PSC and UC and two of these four were also shared with CD. Four loci also colocalised with loci for T1DM. There were several risk loci which could not be resolved to a single causal variant or small set of credible variants from Chapter 2's fine-mapping efforts. For these

Table 3.2: Colocalisation between PSC risk loci and immune-mediated diseases

Chr	Region	OR	p-value	UC H4	CD H4	PBC H4	T1DM H4	CeD H4	RhA H4	MS H4	SLE H4
1	*MMEL1*	1.20	5.12E-13	0.08	0.00	0.56	0.01	0.36	0.45	0.95	0.02
2	*BCL2L11*	1.29	2.18E-11	0.89	0.05	0.73	0.00		0.23	0.00	0.08
2	*CD28*	1.25	4.12E-16	0.06	0.01	0.00	0.00		0.00	0.00	0.07
3	*MST1*	1.33	5.25E-26	0.85	0.74	0.01	0.00	0.08	0.00	0.00	0.00
3	*FOXP1*	1.44	2.80E-15	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.00
4	*IL2-IL21*	1.28	8.25E-14	0.02	0.06	0.56	0.00		0.04	0.00	0.01
6	*BACH2*	1.21	1.09E-09	0.01	0.07	0.04	0.18	0.49	0.87	0.00	0.02
10	*IL2RA*	1.22	1.44E-16	0.05	0.00	0.00	0.00		0.95		0.00
11	*CCDC88B*	1.20	1.81E-13	0.00	0.56	0.03	0.82	0.00	0.29	0.00	0.04
12	*SH2B3*	1.18	3.86E-13	0.89	0.84	0.94	1.00	1.00	0.20	0.00	0.73
16	*CLEC16A*	1.20	5.22E-13	0.00	0.00	0.57	0.61		0.00	0.00	0.05
18	*CD226*	1.19	5.87E-12	0.00	0.03	0.88	0.76		0.02	0.00	0.25
19	*PRKD2*	1.28	2.12E-12	0.03	0.00	0.00	0.96	0.01	0.00		0.00
21	*ETS2*	1.23	3.40E-13	0.82	0.79	0.00	0.00	0.00	0.00		0.00
21	*UBASH3A*	1.22	2.42E-12	0.05	0.00	0.00	0.82	1.00	0.42	0.00	0.00

OR; odds ratio for lead GWAS SNP risk allele, p-value; for lead GWAS SNP
PP H4>0.8 highlighted in green, evidence for PP H3>0.8 highlighted in red

loci, in Chapter 2 I had examined fine-mapping studies of the same loci in other IMDs to define the most likely causal variants. Reassuringly, for all loci where this was the case, colocalisation supported a shared causal variant between PSC and the other IMD locus. For example, fine-mapping of PSC region 8 (Chr10:6108139) resolved this locus to a five variant credible set with a 46% PP of causality supporting rs4147359 as the most probable causal variant. Westra *et al* have previously fine-mapped this locus in RhA to a three variant credible set in which rs706778 has 89% PP of causality [114]. Here, I show that this same locus colocalised in PSC and RhA with 95% PP4 supporting a common causal variant between the two IMDs. Thus rs706778 is the most probable causal variant for both PSC and RhA.

When trying to identify the gene perturbed by a PSC risk locus, colocalisation with an eQTL is the most useful functional trait, as it enables us to identify not only the gene affected, but whether PSC risk is conferred by increased or decreased expression of that gene. The gene quantitatively affected by an eQTL, or the eQTL-gene pair is called an eGene. I conducted colocalisation analysis between the twenty-two PSC risk loci and 42 functional QTL datasets covering five gastrointestinal whole-tissue types, six immune cell-types and five different functional traits including gene-expression (eQTL), histone marks (histQTLs), DNA methylation (methQTLs) and splice site QTL (spliceQTLs) (Table 3.1). For those seven risk loci not reaching genome-wide significance in Ji *et al's* data, the results consistently supported no evidence of colocalisations with any functional traits, thus the results for these seven loci are not shown. I found colocalisations with eQTL for four of the remaining fifteen PSC risk loci. Of these four loci, three colocalised

with one eGene and one colocalised with two eGenes. Where a disease risk locus colocalises with an eQTL, further colocalisation of the same region with another functional QTL such as a histQTL or methQTL helps us to idenitfy the epigenetic mechanism via which that eQTL affects gene expression. For example an eQTL may decrease expression of gene X by impeding transcription factor binding, evidenced by colocalisation of the same locus with an eQTL of gene X and a H3K27ac mark (histQTL). Of the four loci that colocalised with one or more eGenes, I found evidence that all four also colocalised with another functional QTL; two with methQTLs, one with a histQTL and one with a spliceQTL.

Where colocalisation for a risk locus identifies the same single eGene in more than one cell-type or tissue, particularly those mechanistically related to PSC, this lends further weight to a causal role for this gene in disease pathogenesis. This was the case for three of the fifteen PSC risk loci; Chr19:47205707 *PRKD2*, Chr21:40466744 *ETS2* and Chr21:43855067 *UBASH3A*. For each of these three loci, I found colocalisations with one eGene across several cell-types and tissues. Followed by functional trait fine-mapping, for these three loci this allowed me to identify a perturbed gene, a direction of effect, a set of relevant cell-types, a single or small set of credible causal variants and the mechanism via which the causal variant potentially dysregulated the quantitative expression of that gene. The colocalisation and functional trait fine-mapping results for these three loci are discussed in more detail below. This is followed by the discussion of two other loci of interest; Chr12:11184608 *SH2B3* and Chr18:67543688 *CD226*.

Table 3.3: Colocalisation of PSC risk loci with functional QTLs

GWAS			Colocalisation				PSC GWAS			Functional Trait		
Chr	**Lead SNP**	**eGene**	**QTL Type**	**Tissue type**	**Tissue State**	**H4 PP**	**Risk allele**	**Beta**	**p-value**	**Risk allele**	**Beta**	**p-value**
3	rs3197999	n/a	methQTL cg06313718	Monocytes	Unst.	0.81	A	0.26	2.60E-13	A	-0.61	2.43E-08
		n/a	methQTL cg06313718	Neutrophils	Unst.	0.85				A	-1.20	2.37E-37
11	rs663743	*AP003774.1*	eQTL	Whole blood	Unst.	0.97	G	0.17	8.42E-08	G	0.82	1.45E-35
		AP003774.1	eQTL	EBV-transformed lymphocytes	Unst.	0.96				G	0.96	3.73E-18
		CCDC88B	eQTL	Monocytes	Unst.	0.85				G	0.57	7.73E-08
		n/a	methQTL cg07884764	Neutophils	Unst.	0.96				G	-0.71	4.29E-12
16	rs725613	n/a	methQTL cg04616529	CD4+ T cells	Unst.	0.96	T	0.20	5.50E-10	T	0.93	3.49E-18
		n/a	methQTL cg00121339	Neutrophils	Unst.	0.92				T	-0.85	4.58E-19
19	rs60652743	*PRKD2*	eQTL	Transverse Colon	Unst.	0.94	A	0.26	1.01E-07	A	-0.21	1.16E-06
		PRKD2	eQTL	Sigmoid Colon	Unst.	0.94				A	-0.39	8.14E-08
		PRKD2	eQTL	Monocytes	Unst.	0.94				A	-1.14	2.30E-25
		n/a	methQTL cg00838415	Monocytes	Unst.	0.94				A	-0.77	6.26E-11
		n/a	methQTL cg00838415	Neutrophils	Unst.	0.94				A	-0.74	4.15E-10
		n/a	methQTL cg08634012	Monocytes	Unst.	0.94				A	-0.97	2.17E-17
		n/a	methQTL cg08634012	Neutrophils	Unst.	0.95				A	-0.89	1.53E-14
21	rs2836883	*ETS2*	eQTL	Whole blood	Unst.	0.83	G	0.30	5.40E-14	G	0.15	2.50E-08
		ETS2	eQTL	Monocytes	Unst.	0.87				G	0.69	4.38E-11
		ETS2	eQTL	Macrophages	IL-4 at 6hrs	0.84				G	0.91	1.91E-09
		n/a	H3K7acQTL	Monocytes	Unst.	0.90				G	1.04	4.77E-23
		n/a	H3K7acQTL	Neutrophils	Unst.	0.92				G	0.62	2.16E-08
21	rs1893592	*UBASH3A*	eQTL	Whole blood	Unst.	0.99	A	0.20	1.90E-07	A	-0.21	8.07E-16
		UBASH3A	eQTL	Transverse Colon	Unst.	0.95				A	-0.20	9.44E-07
		UBASH3A	eQTL	CD4+ T cells	Unst.	0.99				A	-1.25	9.71E-39
		UBASH3A	eQTL	T regs	Unst.	1.00				A	-0.25	1.37E-13
		n/a	spliceQTL	CD4+ T cells	Unst.	0.99				A	1.29	6.19E-43

3.4.1 The *PRKD2* locus

The Chr19:47205707 risk locus colocalised with an eQTL for *PRKD2* in monocytes with 94% PP of causality. Notably, the PSC risk increasing allele was associated with decreased expression of *PRKD2*. This locus also colocalised with two CpG methylation sites (cg00838415 and cg08634012) in both monocytes and neutrophils, suggesting that the causal variant for this locus may exert its repressive effect upon gene expression via hypermethylation. Interestingly, although this region colocalised with an eQTL decreasing expression of *PRKD2* in transverse and sigmoid colonic tissue (PP4=94%) and the 1Mb region surrounding this PSC risk locus also contains a significant IBD risk locus, the evidence supported a different causal variant driving the IBD signal (PP3 for colocalisation with UC and CD of 94% and 97% respectively). However, co-localisation with other IMDs demonstrated that the causal variant for this region was shared between PSC and T1DM. Furthermore, in T1DM this locus has been reported as an eQTL for *PRKD2* in monocytes [182], a finding I was able to replicate by conducting colocalisation between T1DM and the Blueprint monocyte eQTL data (PP4=96%). Thus, these results support that PSC risk, T1DM risk and expression of *PRKD2* in monocytes are all likely driven by the same causal variant. The most probable causal variant was identified by fine-mapping this locus in the PSC GWAS data which resolved the region to a fourteen variant credible set with the majority of the PP attached to rs313839 (PP=23%), followed by rs112445263 (PP=20%) (see Chapter 2). This finding was replicated by fine-mapping the same region in the monocyte *PRKD2* eQTL data, resulting in an eight variant credible set led by rs313839 (PP=14%), and rs112445263 (PP=14%) two variants in high LD (r^2=0.98) with one another (Table 3.4 and Figure 3.3). The remaining PP was split evenly across a further 6 variants in high LD, all with r^2>0.8). Fine-mapping supported a second independent signal in the *PRKD2* eQTL data with 55% PP of causality. This was supported by the finding that the most probable causal configuration contained two uncorrelated SNPs; rs313839 and rs314675.

Confirming the fine-mapping assumption that all potential causal variants have been included in the analysis, a search of the 1000 Genomes and UK10K reference panels found there were no SNPs in high LD (r^2>0.8) with rs313839, missing from the eQTL data. The most probable credible variant for this locus, rs313839, lies within an intron. Colocalisation with two methQTLs suggests that this variant alters two CpG methylation sites in monocytes and neutrophils. Furthermore, rs313839*C>G (where rs313839*G is the PSC risk increasing allele) has also been associated with the disruption of many transcription factor binding motifs though ChIPseq studies, as previously discussed in Chapter 2. This suggests several plausible mechanisms via which rs313839*G may exert its repressive effect upon *PRKD2* expression and subsequent effect upon PSC risk. However, the location of other variants in the credible set within gene regulatory elements may

also be important in driving the observed molecular QTL trait. For example, rs402072 at Chr19:47219122, is the only variant that is in high LD with rs313839 and also lies within several hundred base pairs of the transcription start site (TSS) within the promoter region.

PRKD2 (Protein kinase D2) is a member of the serine/threonine protein kinase family and is known to be highly expressed in PSC-relevant tissues including whole blood, small intestine, colon and liver [176]. PRKD2 has known roles in monocyte migration and adhesion. In THP-1 cells (a widely used experimental model of monocytes) expression of a dominant-negative form of PRKD2 resulted in decreased monocyte migration in response to stimulus [179]. Knockdown of *PRKD2* was shown to reduce adhesion of THP-1 cells to endothelial cells in culture, whereas activation of PRKD2 through phosphorylation at Ser 744/748 was shown to increase adhesion to endothelial cells [180]. Monocytes and their macrophage progenitors play an important role in immune-regulation and tissue-repair. Therefore genetic variants that result in decreased expression of *PRKD2* may impair monocyte migration into tissues and subsequent tissue regeneration. *PRKD2* is however not only active in monocytes. The importance of *PRKD2* in T-cells has been demonstrated *in vivo* through T-cell-mediated immune responses in mice expressing PRKD2 variants that cannot be phosphorylated by protein kinase C [205]. While PRKD2 catalytic activity is not essential for the development of mature peripheral T- and B-lymphocytes [206], PRKD2-mutant mice show a striking reduction in the ability of the T-cell receptor (TCR) to induce production of pro-inflammatory cytokines such as interleukin 2 (IL-2) and interferon-γ (IFN-γ), which are important for optimal T-cell-dependent antibody responses [205]. In response to TCR stimulation in Jurkat cells (a model of peripheral T-cells), PRKD2 was activated and translocated from the cytoplasm to the nucleus, to allow IL-2 and IFN-γ promoter up-regulation [207]. Furthermore, in T-cell specific PRKD2-deficient mice, the generation of CD4+ thymocytes is abrogated. This defect is likely to be caused by attenuated TCR signalling during positive selection and incomplete CD4+ lineage specification. The role of PRKD2 in activated T-cells/thymocytes may explain the absence of an observed effect in the naïve CD4+ T-cells studied in my colocalisation analysis. This suggests that the generation of eQTL maps in other T-cell subsets, in different states of activation, may be useful in the further investigation of *PRKD2* in immune-mediated disease risk.

Table 3.4: Fine-mapping of PSC risk loci in functional QTL data

	GWAS Finemapping			Colocalisation			MolQTL Finemapping		
Chr	SNP	PP	CS	QTL type	Cell type	Gene	SNP	PP	CS
11	rs663743*	0.41	2	eQTL	Monocyte	*CCDC88B*	rs663743	0.03	245
19	rs313839	0.23	14	eQTL	Monocyte	*PRKD2*	rs112445263	0.14	8
21	rs4817988	0.58	10	eQTL	Monocyte	*ETS2*	rs4817987	0.07	47
				H3K27ac	Monocyte	N/A	rs2836878	0.13	11
21	rs1893592	0.61	5	eQTL	CD4+ T-cell	*UBASH3A*	rs1893592	1.00	1
				SpliceQTL	CD4+ T-cell	*UBASH3A*	rs1893592	1.00	1

*2nd signal in region, CS; Credible set size, PP; Maximum posterior probability

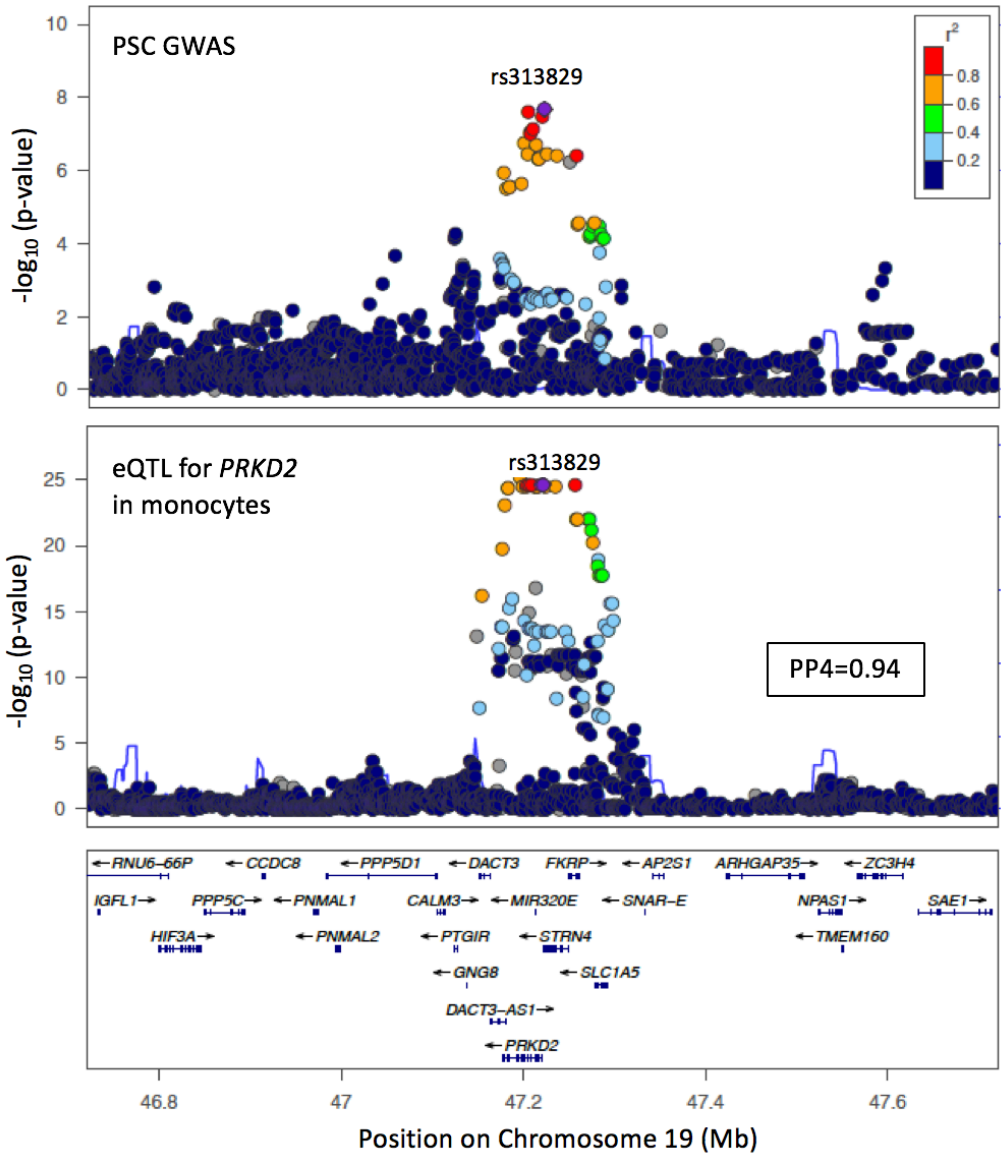


Figure 3.3: Chr19:47205707 regional association plots for most probable fine-mapped SNP, rs313839, in PSC GWAS data and colocalising eQTL data for *PRKD2* in monocytes.

3.4.2 The *ETS2* locus

The Chr21:40466744 locus colocalised with a eQTL for *ETS2* in three different tissues; whole blood, monocytes and IL-4 stimulated macrophages. GWAS studies in both IBD and PSC have consistently proposed that the most likely candidate gene for the Chr21:40466744 locus is *PSMG1* [42, 73]. This was based upon the paucity of genes in this region, and a study of colonic biopsies from paediatric-onset IBD patients, which demonstrated a 'modest' increase in the colonic expression of *PSMG1* in IBD cases compared to controls [175]. Indeed, *PSMG1* which encodes proteasome assembly chaperone 1, has a biologically plausible role in IBD, as part of the ubiquitin-protesosome system. The ubiquitin-protesosome system regulates the generation of peptide antigen presented to MHC class I [208] and TCRs, in addition to regulating co-stimulatory signaling [209]. However, the results of my analysis instead support *ETS2* as the gene dysregulated by this locus. In each tissue, the PSC risk increasing allele was associated with increased expression of *ETS2*. This locus also co-localised with a histQTL for H3K7ac, a marker associated with higher activation of transcription, in both unstimulated monocytes and neutrophils. This suggests that the mechanism by which the causal variant increases expression of *ETS2* in monocytes may, for example, be via increasing the affinity of transcription factor binding. Where a risk locus does not colocalise with an eGene in a particular cell-type, colocalisation with a functional QTL may suggest the presence of an eQTL in a different activation state [134]. It is therefore possible that this locus may be an eQTL of *ETS2*, if investigated in stimulated or activated neutrophils.

Colocalisation of this locus with an eQTL for *ETS2* in iPSC-derived macrophages, six hours following stimulation with IL-4, is particularly interesting as this is a stimulus that mimics the allergic response. Of note, there was no evidence for colocalisation with a macrophage eQTL in either the resting state or the multiple other stimulation states outlined in Table 3.1. This is particularly notable because the vast majority of eQTLs in these data are shared widely across stimulation states. Not only does this highlight the importance of studying cells in the correct state of activation on our ability to identify eQTLs, but also supports a role for *ETS2* in the autoimmune response. The *ETS2* locus also colocalised with a GWAS locus for neutrophil counts (PP4=83%), where the PSC risk increasing allele was associated with a reduction in neutrophil counts. This is biologically plausible given the role of *ETS2* in inducing expression of pro-inflammatory cytokines in macrophages, and the close interactions between macrophages and neutrophils in the inflammatory response.

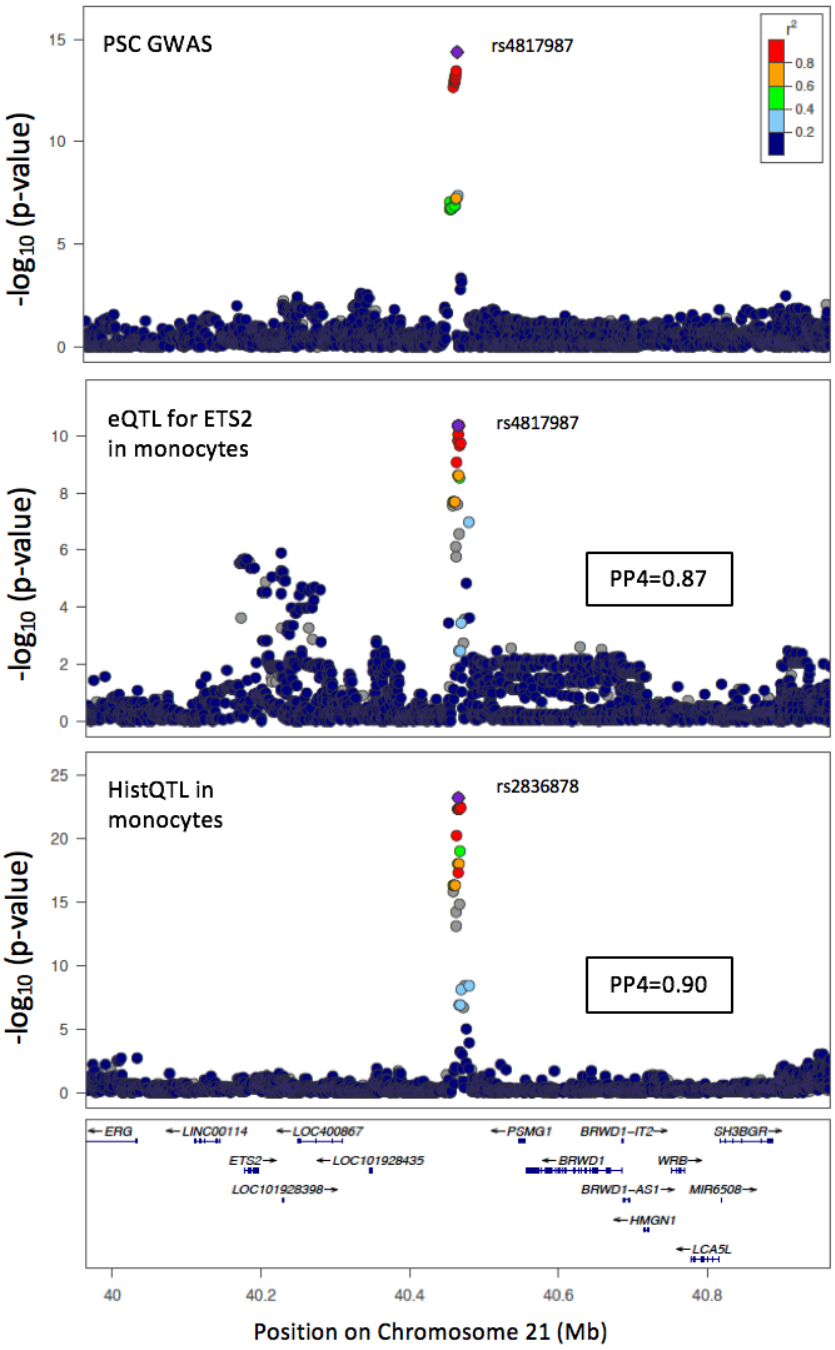


Figure 3.4: Chr21:40466744 regional association plot showing the most probable fine-mapped SNP for PSC GWAS (rs4817987) and colocalising eQTL data for *ETS2* in monocytes (fine-mapped to rs4817987) and for a H3K27ac histQTL in monocytes (fine-mapped to rs2836878).

The Chr21:40466744 locus colocalised with UC and CD (PP4 of 84% and 80% respectively), with no evidence supporting colocalisation with any other IMD (Table 3.2).

However, *ETS2* is a ubiquitously expressed transcription factor, with a well-defined role in macrophages. Cytokine-dependent phosphorylation of Ets-2 results in Ets-2 directly binding the promoters of matrix metaloproteinases *MMP-1* and *MMP-9*, and the induction of other pro-inflammatory target genes including *TNFa, IL-1β*, and chemokines, *CCL2/MCP-1* and *CCL3/MIP-1α* [210]. In mice with severe macrophage-induced pneumonitits, the prevention of Ets-2 phosphorylation on Thr(72) by the Ets-2(A72) mutant allele results in decreased tissue macrophage infiltration [211]. Thus, activated Ets-2 has an important role in the persistent inflammatory response. It is therefore biological plausible that increased expression of *ETS2* could contribute to driving the aberrant inflammatory response observed in PSC. Although I did not observe any colocalisation of this locus with functional QTLs in T-regulatory or CD4+ T-cells, *ETS2* also has a role in IL-2 regulation, the first cytokine produced when naïve T-helper (Th) cells are activated and differentiate into dividing pre-Th0 proliferating precursors. A study by Panagoulias *et al* has demonstrated that Ets-2 binds to the IL-2 promoter which allows transition of naïve Th cells to Th0 cells upon stimulation with antigen, and that Ets-2 silencing allows for constitutive IL-2 expression in unstimulated T-cells [212]. Indeed, they hypothesise that disturbance of this pathway could cause deranged Th cell plasticity and resultant autoimmune disease. Further analysis of eQTL maps in different T-cell subsets and activation states would be necessary to evaluate any effect of the *ETS2* risk locus on *ETS2* expression in T-cells.

Fine-mapping of the Chr21:40466744 *ETS2* locus within the functional QTL data did not prove useful in resolving the causal variant(s) driving this locus. Fine-mapping in the monocyte eQTL data resulted in a credible set of forty-seven variants compared to ten variants in the GWAS data fine-mapping (presented in Chapter 2). This larger credible set is partially attributable to the higher numbers of variants directly genotyped in the whole-genome sequenced eQTL data. Resultantly, the PP was more evenly split between a larger number of very highly correlated variants (Figure 3.4). Furthermore, the failure of eQTL fine-mapping to improve upon the GWAS fine-mapping is a consequence of the reduced strength of association between the lead variant in the eQTL signal compared to the GWAS signal, reducing the power to fine-map. Fine-mapping in the histQTL data resulted in a similar sized credible set of eleven variants compared to ten in the PSC GWAS data. Whilst all ten variants in the GWAS credible set overlapped with the histQTL credible set, the evidence supported rs2836878 as the most probable causal variant in the histQTL signal (PP=13%), compared to rs4817988 in the GWAS data (PP=58%).

3.4.3 The *UBASH3A* locus

Of all PSC risk loci, the Chr21:43855067 locus was the most extensively investigated prior to this study. This locus was already a known eQTL of *UBASH3A* from two whole-blood

and one B-cell only analysis [172, 213, 214] and a likely shared risk locus with CeD and RhA [148, 201]. I confirmed, with colocalisation, that this PSC locus shared a causal variant with CeD (PP4=100%), as well as T1DM (PP4=82%). However colocalisation resulted in almost equivocal evidence supporting a shared (PP4=42%) or different causal variant (PP3=54%) driving the signal in RhA.

Colocalisation confirmed that this locus is an eQTL of *UBASH3A* in T-regulatory cells (PP4=100%) and naïve CD4+ T cells (PP4=99%). In both T-cell types, the PSC risk increasing rs1893592*A allele, which is also the major allele at this locus, was associated with decreased expression of *UBASH3A*. Interestingly, although there was no evidence supporting shared genetic variation with UC or CD, the Chr21:43855067 rs1893592 locus also colocalised with a eQTL of *UBASH3A* in transverse colon tissue (PP4=95%), but not sigmoid colon, a pattern of colonic involvement reminiscent of the PSC-associated IBD phenotype. Fine-mapping of this locus in PSC GWAS and CD4+ T-cell eQTL data supported rs1893592 as the most probable causal variant. As a result of the higher strengths of association in the functional QTL data increasing power to fine-map the signal, fine-mapping in the eQTL data attributed 99% of the PP4 to rs1893592 compared to 61% in the GWAS data. The rs1893592 variant is thought to alter the conserved 5' splice donor sequence. The predicted consequence of the PSC protective rs1893592*C allele is to increase expression of the downstream intron, causing intron 10 to be retained in the *UBASH3A* mRNA [42, 215]. This was supported by the finding of a colocalisation with a spliceQTL in CD4+ T-cells (PP4=99%), which was also fine-mapped to the same causal variant, rs1893592, with 100% PP4 of causality.

UBASH3A has a described role in human T-cells where it has been shown to attenuate the NF-$\varkappa$B signalling pathway upon TCR stimulation, by specifically suppressing activation of the I$\varkappa$B kinase complex, through a ubiquitin-dependent mechanism [216]. In the T-cell eQTL data used for colocalisation in this analysis, the PSC protective rs1893592*C allele was associated with increased *UBASH3A* expression. It has been previously demonstrated in human primary CD4+ T cells that following TCR stimulation, the PSC-protective rs1893592*C allele is associated with a significant reduction in the overall mRNA levels of UBASH3A, but an increase in the proportion of a normally occurring, but low-abundant UBASH3A transcript that retains intron-9 sequences and cannot produce full-length UBASH3A protein [217]. The reduction in UBASH3A mRNA subsequently results in increased secretion of IL-2, a key cytokine in T-cell function and activation. This therefore provides important insights into how dysregulation of *UBASH3A* splicing and expression may be causal in the pathogenesis of PSC.

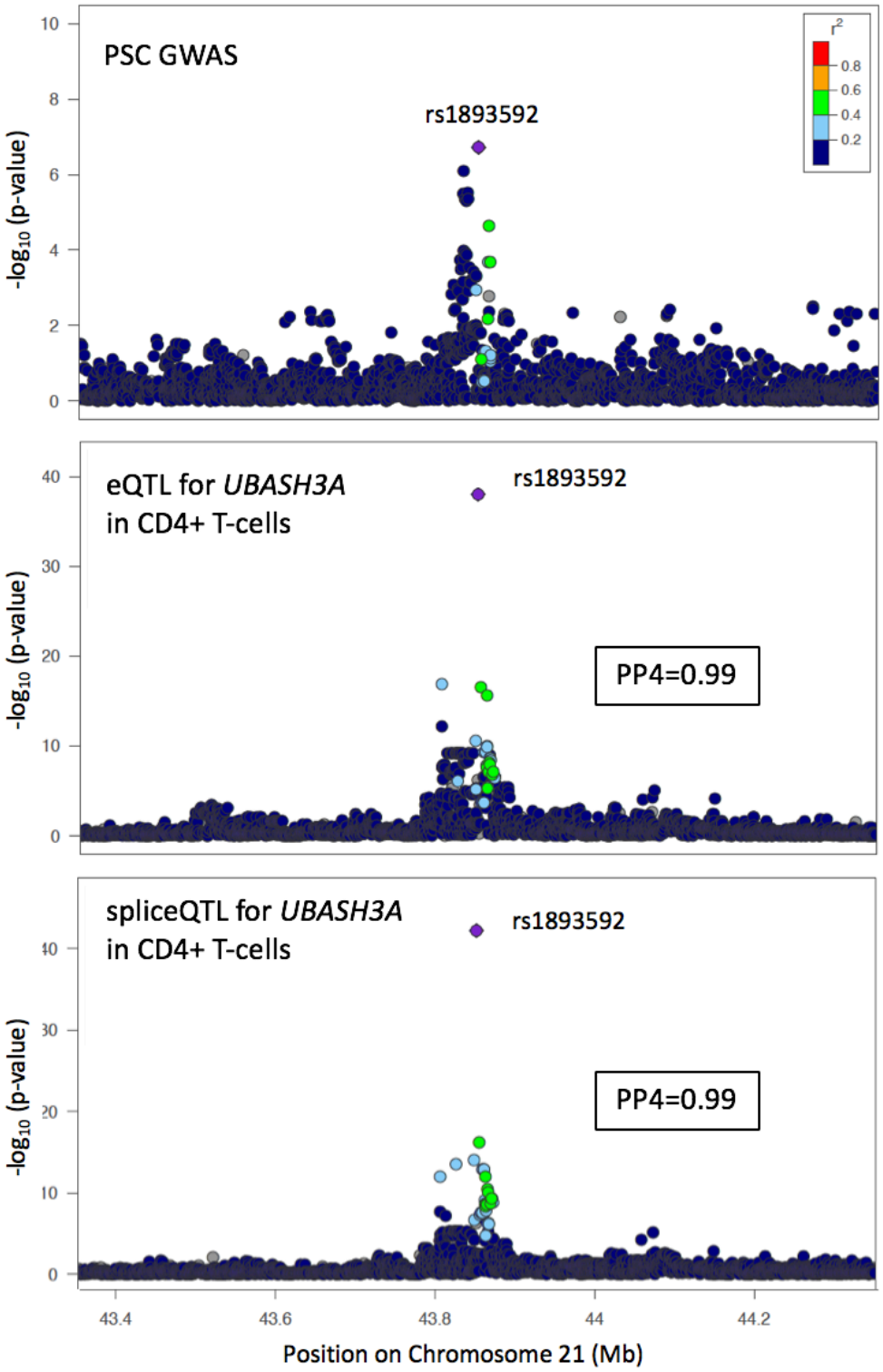


Figure 3.5: Chr21:43855067 regional association plots for fine-mapped SNP, rs1893593, in PSC GWAS and colocalising eQTL data for *UBASH3A* and spliceQTL data for *UBASH3A*.

3.4.4 The *SH2B3* locus

The Chr12:11184608 *SH2B3* PSC risk locus colocalised with five other IMDs; UC, CD, PBC, T1DM and CeD, supporting a single-shared causal variant driving all six diseases. Fine-mapping in the PSC GWAS data (Chapter 2), highlighted rs3184504, a missense variant within the 3rd exon of *SH2B3*, as the most probable causal variant (PP=99%). Whilst this locus has not been fine-mapped in UC, CD, PBC or CeD, a fine-mapping study of T1DM resolved this locus to a credible set including two variants; rs653178 (PP4=66%) and rs3184504 (PP4=33%) [114]. Notably, both of these variants were included within the PSC fine-mapping analysis.

SH2B3 is ubiquitously expressed across many cell and tissue types, with a role in the regulation of signalling pathways involved in cell migration, differentiation, inflammation and haematopoiesis [218]. It was therefore unsurprising that this locus colocalised with GWAS traits for leucocyte, monocyte and neutrophil counts. The PSC risk increasing rs3184504*T allele is associated with an increase in all myeloid and lymphoid cell counts, compared to the reference rs3184504*C allele. *SH2B3* is a negative regulator of T-cell activation, TNF production, and Janus kinase (JAK)-2 and -3 signalling. Another eQTL study has shown that the autoimmune hepatitis (AIH) rs3184504*A risk allele, which results in the same protein coding change as rs3184504*T, is associated with increased expression of genes involved in IFNγ production [172]. This suggests a mechanism via which this locus might contribute to increased immune cell counts and aberrant immune- and inflammatory-response.

3.4.5 The Chr18:67543688 locus

PBC is an immune-mediated inflammatory condition affecting the small bile ducts that is often considered a sister condition to PSC. Somewhat surprisingly, colocalisation of PSC risk loci with PBC identified only two loci for which there was evidence of a single shared causal variant between both traits. The first was the Chr12:11184608 *SH2B3* locus discussed above, and the second was the Chr18:67543688 locus. The Chr18:67543688 locus could not be well fine-mapped in the PSC GWAS data with a credible set containing 44 variants. Furthermore, this locus did not colocalise with any other IMD, other than PBC, in this analysis. It is therefore possible that the genes and pathways affected by this locus are perhaps the most likely candidates for bile-duct specific effects. This 1Mb region of the genome contains only four genes (see Figure 2.7, Chapter 2); *DOK6, CD226, RTTN* and *SOCS6*, however I did not find subsequent evidence for colocalisation with any functional QTLs or eQTLs to support a causal role for one of these four genes. Two of these four candidate genes, *CD226* and *SOCS6*, have important roles in relevant immune cell pathways. The first gene, *CD226*, is expressed on the surface of natural killer

cells, T-cell subsets, platelets and monocytes. CD226 mediates cellular adhesion to other cells expressing its ligands, CD112 and CD155. The second gene, *SOCS6* (suppressor of cytokine signalling 6), is a cytokine-inducible negative regulator of cytokine signaling. The third gene in this region, *DOK6*, is a less likely candidate for involvement in PSC or PBC pathogenesis as it is expressed mainly in the central nervous system where it is involved in the receptor tyrosine kinase signalling cascade [219]. Whilst little is known about the final candidate gene in this locus, *RTTN*, it encodes Rotatin, an intracellular protein thought to play a role in the maintenance of normal ciliary structure [220]. Cholangiocytes are ciliated cells which have a role in expediting bile flow, and disturbance of the normal structure or function of cholangiocyte cilia is likely to contribute to several cholangiopathies [221]. Thus *RTTN* is the only gene within the Chr18:67543688 PSC-PBC risk locus with a potential role in bile duct homeostasis, highlighting it as a potential candidate gene for further investigation.

3.5 Discussion

In this chapter I describe the first investigation of PSC risk loci using colocalisation with multiple traits including IMD risk, cell count indices, eQTLs and functional QTLs across a variety of PSC-relevant cell-types and tissues. By combining colocalisation to identify the genes impacted by PSC risk loci and the epigenetic mechanisms underlying the gene perturbation, with fine-mapping in the colocalising functional traits, I identify the genes, cell-types and causal variants affected by several PSC risk loci. For four of the fifteen PSC risk loci, this was successful in identifying the genes perturbed and for three of these five loci, it was successful in identifying a single causal variant, or small set of credible variants. Perhaps most notably, these analyses determine that the most probable causal variant driving the Chr19:47205707 PSC and T1DM risk locus, rs313839, results in a reduction of *PRKD2* expression in monocytes and colonic tissue, possibly mediated by hyper-methylation. Similarly, I have fine-mapped the shared PSC-IBD Chr21:40466744 risk locus to a set of ten credible variants, of which the true causal variant increases expression of *ETS2* by activating transcription in monocytes and macrophages subject to allergic stimulus. Thus the results of this study can guide further functional follow-up of these loci in terms of causal variants, direction of effect upon gene expression and relevant cell types in which the effects are mediated. Furthermore, they advocate the combination of colocalisation and functional trait fine-mapping as an alternative approach to resolving the causal variants driving complex trait loci in rare diseases, in which amassing the large sample sizes required to improve upon GWAS trait fine-mapping is unlikely to be feasible. However, for some non-coding PSC risk loci, this pipeline was not effective in determining either causal variants or genes. For example the Chr2:111933001 (*BCL211*),

Chr6:91030441 (*BACH2*), Chr10:6108439 (*IL2RA*) and Chr18:67543688 (*CD226*) loci did not colocalise with any functional QTL in the tissues and cell types included within this analysis. Each of these loci did, importantly, colocalise with at least one other IMD (Table 3.2), all of which are more common diseases with larger GWAS sample sizes, meaning that colocalisation and fine-mapping in those other colocalising IMD traits may be an alternative route to resolving these PSC risk loci.

This study focused on colocalisation with functional traits in cells and tissue types relevant to PSC. This was based upon previous studies demonstrating that some eQTLs are only active in particular cell types or activation states [130], and that eQTLs are enriched for disease-associated variants in disease-relevant tissue-types [190, 191]. The choice of disease-relevant tissues in this study was however limited by two factors. Firstly, designating a cell-type 'relevant' in a disease such as PSC, in which we have limited understanding of disease pathogenesis, is challenging. Secondly, the limited availability of functional QTL data with publicly accessible full summary statistics in these 'relevant' cell types further impairs this choice. However the results from this analysis serve to highlight the importance of conducting colocalisation with eQTLs measured in the relevant cell-types. For example, analysis of the Chr21:40466744 locus supported *ETS2* as the most likely gene perturbed by this risk locus, with colocalisations observed in monocytes and IL-4 stimulated macrophages. Whilst *ETS2* has a described role in the induction of pro-inflammatory cytokine release from macrophages, *ETS2* also has a role in IL-2 regulation in naïve Th transitioning to Th0 cells upon antigenic stimulation. Given this role in naïve Th cells, it is unsurprising that we did not find any colocalisation with eQTLs for *ETS2* in the available CD4+ or T-regulatory cell datasets. However, it is plausible that if examined in the right T-cell subtype or activation state, the Chr21:40466744 locus may also be an eQTL of *ETS2* in some T-cell subtypes. Similarly, investigation of the Chr19:47205707 risk locus found it colocalised with an eQTL for *PRKD2* in monocytes, a gene with a role in the adhesion of monocytes to endothelial cells. Whilst this gene also has a role in negative selection of T-cells, I did not find any colocalisation with a *PRKD2* eQTL in the available CD4+ T-cell and T-regulatory cell data. Furthermore, there are no published and publicly available eQTL datasets for T-cells in the activated or stimulated state, again introducing the possibility that the correct cell type has not been examined. Future work could focus upon conducting combined colocalisation and fine-mapping in functional QTL data from all available cell-types and tissues, with the added risk of introducing noise by examining traits across multiple tissue types and the difficulty of interpreting colocalisations with genes in tissues such as brain or muscle which are seemingly remote from PSC pathogenesis. Another solution would be to use the current hypotheses of disease pathogenesis in PSC to select those cell types of most potential mechanistic relevance to PSC and to build eQTL maps in those PSC-specific

cell types. This is an analysis presented in the following chapter.

Several properties of *Coloc* are likely to have influenced the results presented in this chapter. Firstly, Bayesian colocalisation analysis is strongly influenced by the choice of priors. Indeed, as the p^{12} threshold is increased (e.g. from 10^{-6} to 10^{-5}), there is more certainty that the data supports a shared causal variant between both traits. This can be especially important in regions where there are extended patterns of strong LD and thus uncertainty as to whether the data supports the H3 or H4 hypothesis, because it is in keeping with both scenarios. For these loci, the choice of prior becomes the determinant of the colocalisation. An example of this is the Chr4:123499745 locus near the candidate gene *IL2-IL21*, for which there was no evidence supporting colocalisation with any functional QTLs or IMDs at $p^{12}=10^{-6}$, but with evidence supporting shared genetic variation with several other IMDs driven by a different causal variant (PP3>80%) (Table 3.2). However, the evidence supporting colocalisation (PP4) increases as the p^{12} threshold is increased (Figure 3.2). Whilst this may favour a higher p^{12} for the detection of more colocalising IMD traits, it is known that variants associated with complex traits are more likely to be eQTLs than MAF-matched variants from GWAS analyses chosen at random, thus supporting the more stringent choice of priors used in this analysis [117, 222]. Secondly, *Coloc* makes the assumption that each risk locus contains only one independent signal. For those regions in which there were more than one independent signal, *Coloc* considers only the strongest of these distinct association signals. Where each of the association signals explains a similar proportion of the variance of the trait, the PP4 will drop and PP3 proportionately increase [189]. Fine-mapping of the PSC risk loci described in Chapter 2 supported the presence of two independent signals in four of the 15 loci. For those four PSC risk loci containing two independent signals, there was evidence for colocalisation with functional QTLs for only one of these four risk loci. This was the Chr11:64107735 locus, which colocalised with an eQTL for *CCDC88B* in monocytes and an eQTL for *AP003774.1* in whole blood and EBV-transformed lymphocytes. A future means of investigating these multi-signal loci is to include a step-wise conditional regression [223] to identify additional independent signals within a locus, and to perform colocalisation on the resultant conditional p-values, as a means to accounting for multiple independent signals [189].

Colocalisation with eQTLs, functional QTLs and other IMDs allows us to ascribe a gene, the direction of effect on gene expression associated with PSC risk, the epigenetic mechanism dysregulating that genes expression as well as the other IMDs impacted via the same gene and epigenetic mechanism. With the example of the Chr19:47205707 risk locus, colocalisation identified that the causal PSC risk increasing allele for this locus correlated with an eQTL reducing expression of *PRKD2*, via hypermethylation, and that the same causal variant also conferred risk of T1DM. In order however, to unequivocally prove that T1DM risk at this locus is also mediated by perturbation in *PRKD2* expression,

I needed to performed further colocalisation between T1DM and monocyte eQTL data. More recently, Giambartolomei *et al* have published methods to quantify the evidence supporting a common causal variant in a particular region across multiple traits from summary statistics [224]. This method, *Moloc*, was published in 2018 after the analysis presented in this chapter was largely complete. Similar to *Coloc*, *Moloc* uses a Bayesian framework to integrate GWAS and functional QTL data, with the same three assumptions pertaining to the inclusion of the true causal variant within the data, a maximum of one independent association per region and that samples are drawn from the same ethnic population and thus share LD structure. The future use of such a method would be advantageous in providing a quantification of evidence for a shared causal variant between all traits tested for one locus, avoiding the need for the multiple rounds of pair-wise colocalisation conducted in this analysis. Such an approach could also be useful in the fine-mapping of PSC risk loci. An important part of this analysis was to conduct fine-mapping of loci within functional traits, in an effort to identify the causal variant driving these colocalising traits. Whilst data availability meant that this approach could only be applied to four of the PSC risk loci, it was successful in improving fine-mapping resolution for two of these loci. An potentially fruitful future analysis might focus upon boosting power for fine-mapping by combining multiple colocalising datasets for a single locus into one meta-dataset using a model that allows for mixed effect sizes, followed by fine-mapping of the meta-dataset. Methods based upon similar approaches have been published by Wallace *et al* [225] and will form part of my future follow-up work, not presented in this thesis.

Using a combination of colocalisation and fine-mapping across multiple traits, I have been able to identify the genes, causal variants and epigenetic mechanisms implicated by five PSC risk loci. In addition, my work highlights some of the cell-types in which these aforementioned genes and mechanisms are especially relevant. However, several loci remain unresolved, and future work should focus upon using current knowledge of PSC pathogenesis to build eQTL maps in the most PSC-relevant cell types, followed by similar colocalisation and fine-mapping analyses. This analysis, presented in the following chapter is a means to further understanding the causal biology of PSC.

Chapter 4

T-cell expression quantitative trait loci maps in Primary sclerosing cholangitis

4.1 Introduction

Colocalisation of GWAS risk loci with eQTLs provides a powerful way to indentify the functional role of the numerous non-coding risk loci by assigning molecular function to them. As shown in the previous chapter, colocalisation using published eQTL datasets for a variety of immune cell types and tissues has enabled the identification the genes perturbed by six of the studied PSC risk loci. The failure to identify the genes underlying the remaining risk loci may result, in part, from the failure to identify genetic variants that regulate gene-expression in cell-types and states relevant to PSC.

Whilst many eQTLs are shared across multiple tissues, some remain highly specific to a particular cell type, tissue, environment or activation state [130]. One of the ongoing challenges is to identify the correct cell-type or tissue in which to map eQTLs for colocalisation with GWAS risk loci. Indeed, it has been shown that when trying to unravel the molecular basis of disease-specific risk loci, the choice of disease-relevant tissues supports the finding of eQTLs enriched for disease-associated variants [190, 191]. However, colocalisation analysis remains limited by the availability of published eQTL summary statistics. Furthermore, since PSC is a rare disease, there are currently no published eQTL studies of the cell types perhaps most relevant to PSC, in the environments most relevant to PSC. Therefore eQTL mapping in PSC-specific cell types, in PSC-specific environments, is of great scientific interest.

Identification of the cell types of most potential relevance to the causal pathogenesis of a disease relies upon existing knowledge of disease pathogenesis, which unfortunately, in PSC remains limited. As with many immune-mediated diseases (IMDs), T-regulatory cells

have been implicated in the pathogenesis of PSC, not least supported by the finding of two PSC risk loci near the *IL2RA* and *IL2/IL21* genes, the protein products of which are expressed or involved in pathways of T-regulatory cells. Histological observations provide further evidence of potentially relevant cell types. PSC is histologically characterised by a T-cell rich portal infiltration with peri-ductal inflammation, portal fibrosis and progressive loss of the bile ducts, known as ductopenia [226]. Moreover, evidence for potentially relevant cell types comes from the strong link with IBD, which is present in 50-70% of patients with PSC [23]. The liver and colon are anatomically linked with 75% of the blood supply to the liver originating from the intestine via the portal vein. In PSC, it has been shown that 20% of liver-infiltrating lymphocytes express gut-specific ligands CCR9 and α4β7. The 'gut-homing T-cell hypothesis' suggests that these CCR9+ memory T-cells are originally activated by inflammation within the gut and are recruited to the liver due to the observed aberrant inflammation-induced expression of their receptors MAdCAM-1 and CCL25 [53, 79]. In support of this, the vast majority of these CCR9+ liver-infiltrating T-lymphocytes in PSC 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 α4β7+ gut-derived lymphocytes are likely to use other chemokines such as CXCL12 and CXCR6 to localise to the biliary epithelium where they mediate targeted inflammation of the bile ducts. To date, no existing studies have mapped eQTLs in any of the aforementioned cell types. Therefore, some of the most potentially relevant cell types for the focus of future eQTL mapping efforts in PSC include the CD4+ and CD8+ effector-memory T-cells with the CCR9+ phenotype. Furthermore, one of the means of evaluating cells in the PSC-specific activated state, most representative of the active disease transcriptional phenotype, is to derive those cells from individuals with the active inflammatory condition.

4.2 Chapter Overview

Many studies have sought to map genetic variants associated with quantitative changes in gene expression in order to assign molecular function to non-coding disease risk loci via colocalisation. However eQTLs are known to be specific to both tissue type and activation state. Thus, one means of better understanding the genetic risk loci associated with susceptibility to PSC is to explore eQTL maps specific to the tissues and activation states of the disease. In this chapter, I describe the generation of eQTL maps in six peripheral blood T-cells subtypes, currently hypothesised to be important in the causal pathogenesis of PSC. These cells are derived from patients with active PSC and the highly co-morbid condition, UC. I describe the entire study process from patient recruitment to sample preparation and RNA sequencing analysis. I perform differential gene expression

analysis based on disease status. I map eQTLs for each cell type and identify those shared across several T-cell subtypes and those specific to an individual T-cell subtype. Finally, I perform colocalisation with genetic risk loci for PSC, IBD and other immune-mediated diseases (IMDs) in order to identify the genes perturbed by disease-associated risk loci.

4.3 Methods

4.3.1 Sample type and Patient recruitment

The PSC-specific cell-types chosen for analysis in this study were; T-regulatory cells (T-regs), non-activated memory T-cells (T-mems) and activated CD4+ and CD8+ effector-memory T-cells that are positive and negative for the gut-homing ligand, CCR9 (CD4+CCR9-, CD4+CCR9+, CD8+CCR9-, CD8+CCR9+) [53]. The surface marker phenotype of each cell subtype is shown in Table 4.1. I aimed to recruit a total of 80 patients for this study, based upon evidence that previous studies with similar numbers of individuals have identified eQTLs. For example, the GTEx Consortium pilot study of post mortem tissues was able to detect tissue-specific quantitative genetic traits for a median sample size of 105 for the 9 high-priority tissues [176]. Furthermore, the HapMap study of genetic variants underlying variation in gene expression detected an abundance of *cis*-regulatory variants in the human genome with a median sample size of just 40 individuals in each population group [120]. However, PSC is a rare disease with a prevalence of 1 in 10,000 and there are predicted to be just 7,000 patients living with PSC in the UK. Due to the rarity of PSC it is therefore difficult to recruit large numbers of PSC patients with a homogenous, active, disease phenotype. To address this difficulty, I aimed to recruit a total of 80 patients, 40 with PSC and concomitant UC and a further 40 with UC alone. Both PSC-UC and UC patients harbour increased numbers of CCR9+ effector-memory T-cells that have been activated in the inflamed colon [78, 80], and thus this combined cohort would facilitate a sample size large enough to detect eQTLs.

I recruited patients for this study from the Autoimmune liver disease clinic in the Department of Gastroenterology at the Norfolk and Norwich University Hospital. I was granted prior ethical approval for the study by the Norfolk and Norwich University Hospital Human Tissue Bank (reference number: 20122013-57 HT). Given that the ultimate aim of this study was to perform colocalisation with loci associated with risk of PSC in European populations, all patients were of white European ancestry. In order to minimise immune influences on the transcriptome, patients on steroids or biologic therapy, as well as those with previous cancer diagnoses, were excluded. In addition, given that one of the important cell types under investigation was the CCR9+ effector-memory T-cell activated within the inflamed colon, patients with previous colectomy were also excluded. Finally, all recruited

Table 4.1: Fluorochrome-labelled antibody panel defining six subtypes of T-cell by FACS

Cell type	Abbreviation	Antibody panel
T-regulatory cells	T-reg	CD3+CD4+CD25+CD45RO+CD127low
Memory T-cells (non-activated)	T-mem	CD3+CD4+CD45RO+CD25-
CD4+ CCR9- effector memory T-cells	CD4+CCR9-	CD3+CD4+CD62L-CD45RO+CD199-
CD8+ CCR9- effector memory T-cells	CD8+CCR9-	CD3+CD8+CD62L-CD45RO+CD199-
CD4+ CCR9+ effector memory T-cells	CD4+CCR9+	CD3+CD4+CD62L-CD45RO+CD199+
CD8+ CCR9+ effector memory T-cells	CD8+CCR9+	CD3+CD8+CD62L-CD45RO+CD199+

patients had a serum alkaline phosphatase raised above the reference range for the upper limit of normal, but no histological or radiological evidence of cirrhosis to ensure an active PSC transcriptome. A total of seventy-nine donors were recruited; forty-four with PSC and UC and thirty with lone UC. Five healthy controls (HC) for the pilot study set-up which were also included for analysis.

4.3.2 Sample preparation

I drew 50mls of peripheral blood from each donor, and processed this immediately at 4°C to prevent activation or degradation of cells. From whole blood, I separated peripheral blood mononuclear cells (PBMCs) over Ficoll and stained them with a fluorochrome labelled antibody panel designed to isolate the six T-cell subtypes, using three rounds of two-way sorting, as shown in Table 4.1. I sorted cells directly into chilled cell lysis buffer (Buffer RLT Plus, *Qiagen*) using a Sony SH800 fluorescent activated cell sorter (FACS). Samples were then immediately stored at -80°C. An example of the standard FACS gating strategy used is shown in Figure 4.2.

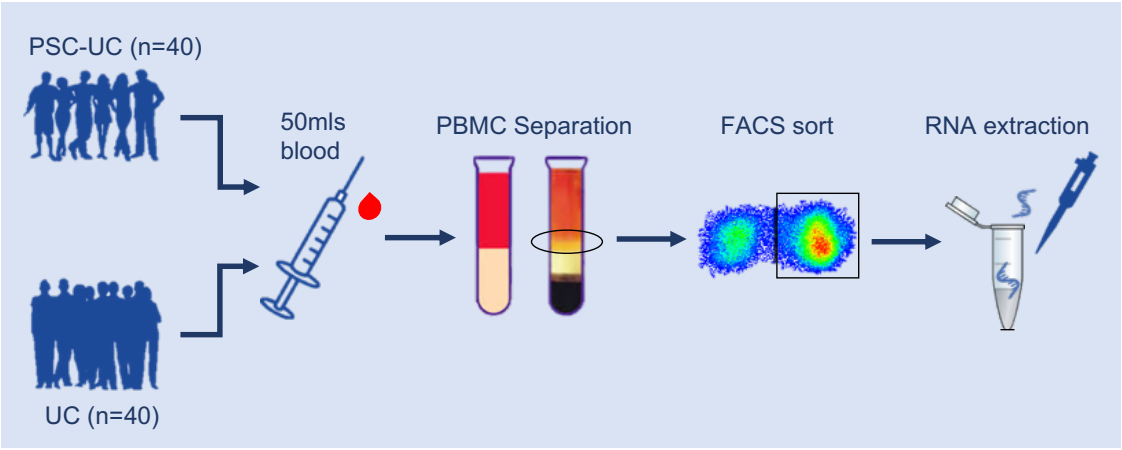


Figure 4.1: Sample preparation pipeline.

During the set-up phase, I verified a small subset of twelve samples (two of each cell type) to >95% purity by performing repeated FACS on already-sorted samples, using the

same gating strategy. To minimise cellular perturbation, I performed all cell sorting using a 100µm nozzle at low sorting pressures using chilled, preservative-free Hank's Balanced Salt Solution (HBSS). Maximum time from acquisition of the whole blood sample to freezing of lysed, FACS sorted, T-cell samples, was six hours. Technical failure of the cell-sorter calibration on two occasions resulted in the loss of all T-cell samples from three individuals (two with PSC-UC and one with lone UC). Therefore, in total 456 T-cell samples were isolated from 76 individuals; 42 with PSC-UC, 29 with UC and 5 healthy controls.

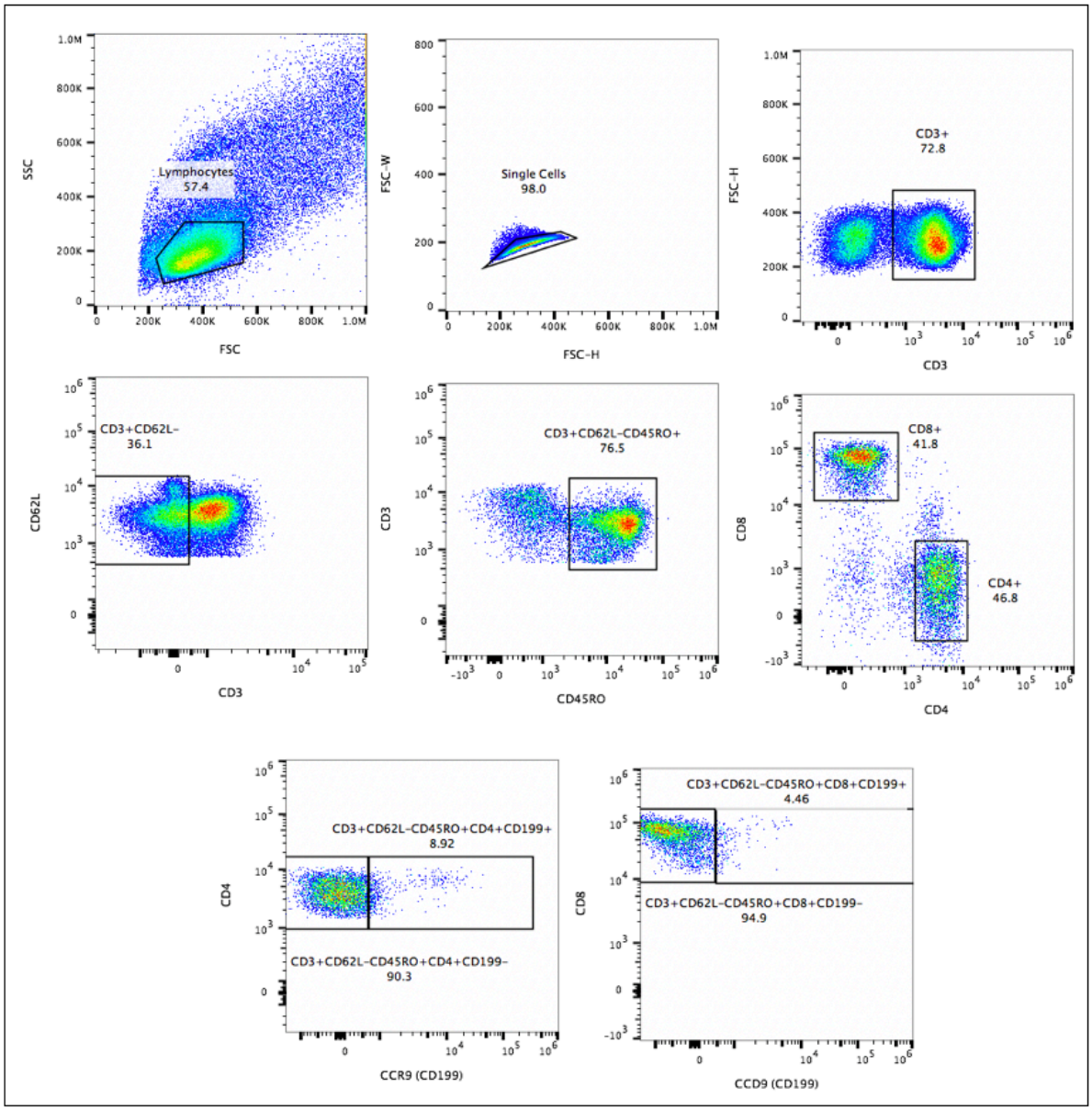


Figure 4.2: Gating strategy used for FACS separation of CD4+CCR9-, CD4+CCR9+, CD8+CCR9- and CD8+CCR9+ central effector T-cells from peripheral blood mononuclear cells.

4.3.3 RNA extraction, library preparation and sequencing

I sequenced six different cell-type samples from seventy-six donors giving a total of 456 libraries. I performed RNA extraction using the *Qiagen* RNAeasy Micro plus kit. I checked RNA concentration and quality on a 20% subset of samples (equally representative of all cell-types) using the Agilent 2100 Bioanalyser, confirming RNA integrity number (RIN)

of >8.0. All samples were then sent to the Wellcome Sanger Institute RNA Pipelines for library preparation and RNA sequencing. Library preparation was performed by Sanger Pipelines using NEBNext Ultra II Directional RNA kit, with a poly(A) pulldown using oligo d(T) beads. Samples were then sequenced using 75 base-pair, paired-end read sequencing, performed on the Illumina HiSeq 4000. Four plates, each containing 96 samples, were pooled at 96-plex and run over twelve lanes (eight samples sequenced per lane) and the fifth plate containing 76 samples was run at 76-plex across ten lanes (7.6 samples per lane). The expected number of reads per samples was ∼60 million reads.

4.3.4 Read alignment, counts and quality control

I aligned reads to the human genome and transcriptome, using *STAR* (Spliced Transcripts Alignment to a Reference) software [227] and the reference genome; Genome Reference Consortium Human Build 38 Release 29 (GRCh38.p12). This is a comprehensive reference transcriptome, which includes protein coding RNA, all known non-coding RNA, non-sense mediated decay transcripts, and both processed and unprocessed pseudogenes. The reference genome is however incomplete, particularly around centromeres, meaning that reads can be incorrectly mapped to other places within the genome (albeit with low quality) resulting in false positive calls. I therefore included decoy contigs, known true human genome sequence that is not included within the reference genome, to map reads that would otherwise map to other regions of the genome.

Read counts were assigned to genes using *FeatureCounts*, implemented in R [228]. For RNA samples, greater than 75% alignment of the total number of reads to the genome was considered successful [229]. Samples with less than 60% of reads aligned to the genome were immediately removed from the analysis, and those between 60-75% aligned initially retained, but ultimately excluded following further quality control (QC) steps described below. Across all samples, the mean proportion of the total reads mapping to exons was 0.79, with a median of 0.80. Samples with a proportion of exonic mapped reads less than 0.6 were also removed from the analysis. Following these preliminary QC steps, 6 T-cell samples were removed from the analysis (Figure 4.3).

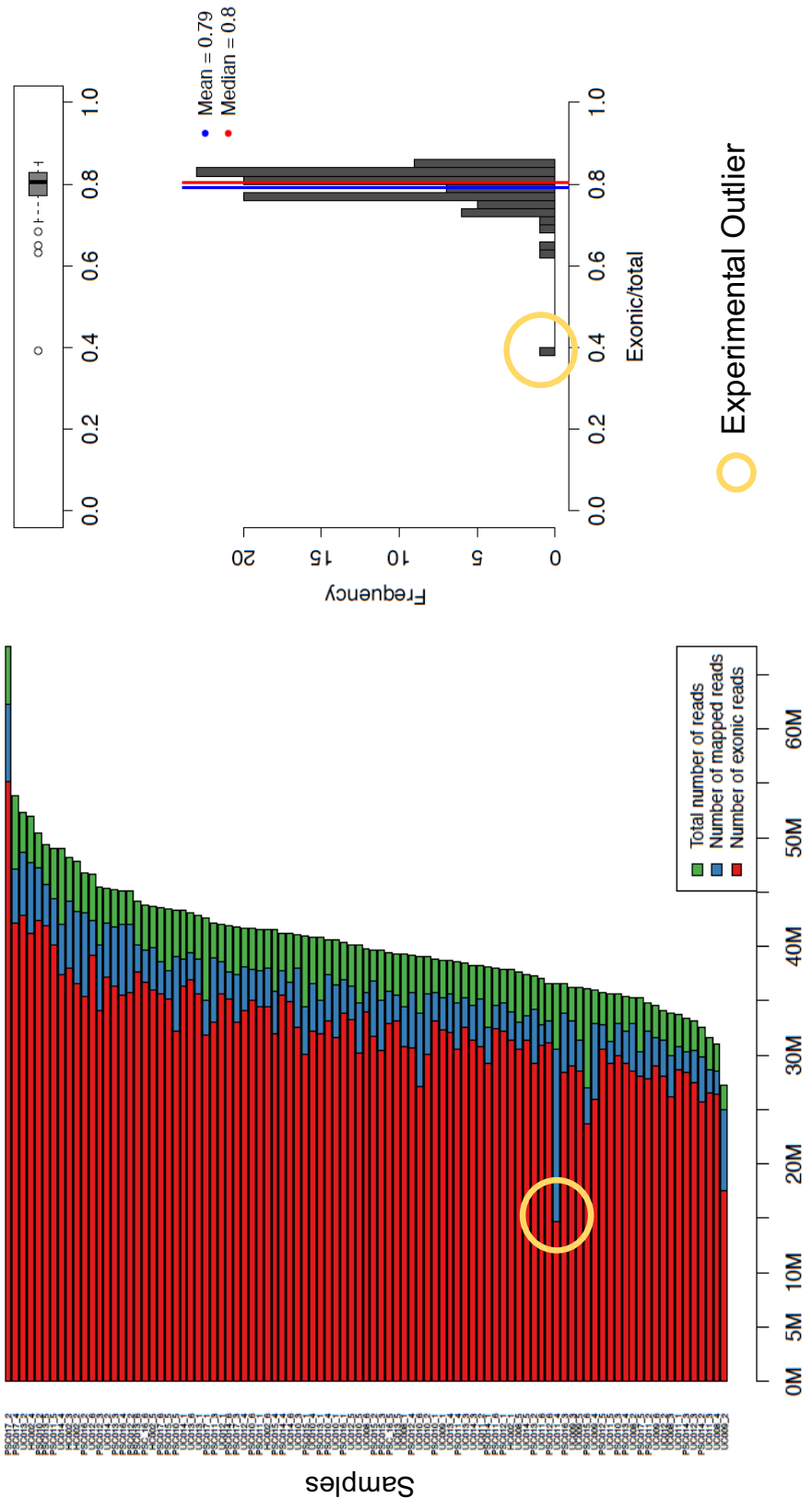


Figure 4.3: Proportion of reads mapped to exons for a subset of 96 of the total 456 samples, highlighting an experimental outlier which was subsequently excluded due to a low proportion of reads mapped to exons compared to the mean.

Duplicated genes within the pseudo-autosomal regions (PAR) were removed and normalisation performed by calculating transcripts per million (TPM). The number of reads mapping to a genes is affected by both sequencing depth (as each library has different sequencing depth) and gene length. TPM is a normalisation method that allows comparisons of genes across samples by normalising for both length of each gene and sequencing depth. Genes not expressed, or expressed at extremely low levels, defined as a sum of TPMs across all samples of <0.5, were removed. Because the presence of lowly expressed genes can decrease the sensitivity to detect differentially expressed genes, I performed a further filtering step, retaining only genes with a mean TPM of ≥ 1 in at least one disease condition.

In order to visualise samples that were experimental outliers, I performed principal component analysis (PCA) of the top 500 most variably expressed genes across all samples of all cell-types, implemented in *DESeq2* [230]. PCA uses linear combinations of gene expression values to define a new set of unrelated variables called principal components. Principal components (PCs) are orthogonal variables, where the PCs are ordered by the proportion of variation they explain in the data. This allows the description of a dataset and its variance by using a reduced number of variables, with the first two components describing the largest variability. The distances in the projection of the space defined by the principal components correlates with the similarities between the samples and thus the transcriptomes of different cell types. PC1 enabled CD4+ T-cells to be distinguished from CD8+ T-cells, explaining 52% of the variance (Figure 4.4). PC2 enabled samples from males and females to be distinguished (8% variance) and PC3 enabled CCR9+ and CCR9- cells to be distinguished (7% variance) (Figures 4.4 and 4.5). PCA was also used to identify experimental outliers, by performing PCA of the top 500 most variably expressed genes for all samples labelled according to sex, disease type and cell type. This process identified two outlying samples which did not cluster with the other samples of the same cell type (samples A and B shown in Figure 4.4), and therefore they were removed from the analysis. PCA also identified a further four outlying samples derived from two patients, which did not cluster with other samples of the expected sex (Figure 4.5). These four outlying samples were collected on the same day, and PCA confirmed that each sample clustered with the expected cell type and were therefore likely to be a direct swap or mislabelling of four samples between two patients. These samples were retained within the experiment for subsequent analysis using the *MBV* module of *QTLtools* [231] which matches genotype with transcriptome data (discussed later in Sample mismatch and amplification bias section).

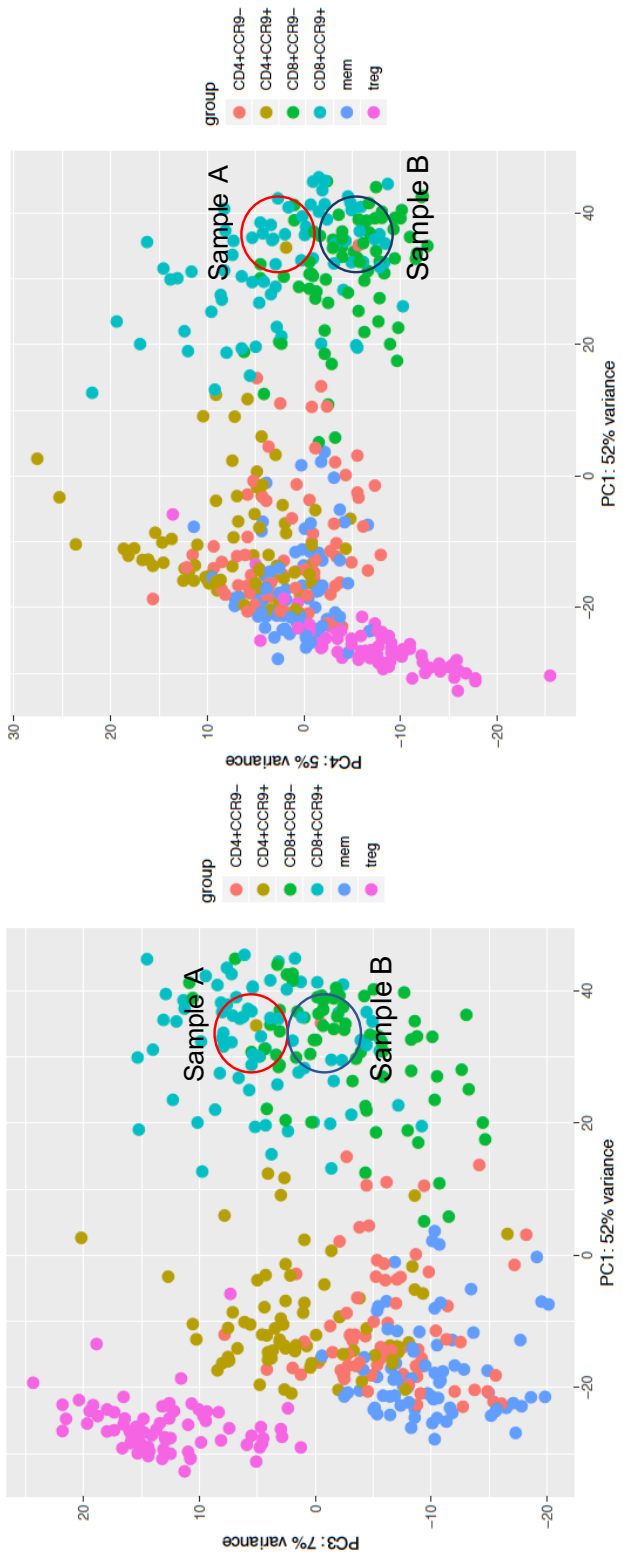


Figure 4.4: Principal component analysis of the top 500 most variably expressed genes, identifying two experimental outliers which did not cluster with their expected cell types.

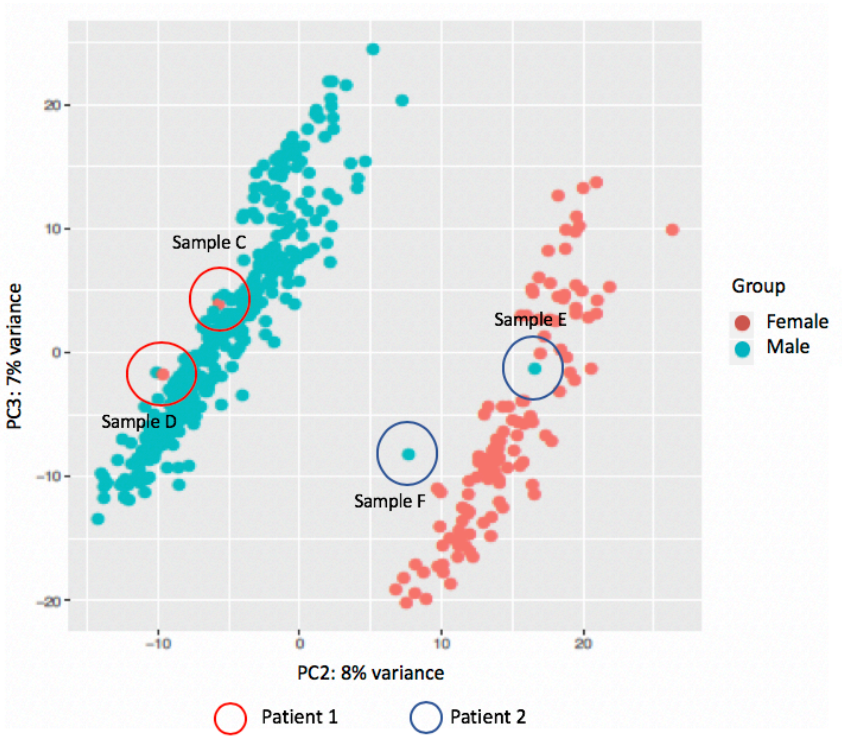


Figure 4.5: PCA analysis of the top 500 most variably expressed genes, identifying four experimental outliers from two patients.

To confirm that the gating strategy and FACS had successfully isolated the expected T-cell subtypes, I compared expression of known marker genes such as *CD4, CD8, CCR9* and *FOXP3* across all cell types. For this, I used the *PlotCounts* function implemented in *DESeq2* to visualise normalised counts of marker genes according to each cell type. This demonstrated good correlation between expected and observed marker gene expression for all cell types (Figure 4.6). The four CD4+ cell subtypes were shown to express high levels of *CD4* in comparison with the two CD8+ cell subtypes, which in turn expressed high levels of *CD8*. *FOXP3* is a transcription factor important in the development of T-regs. The T-regs in this study expressed high levels of *FOXP3*, compared to the other five cell types. *CCR9* expression was high in the two CCR9+ cell subtypes and the T-reg cell population and low in the T-mems, CD4+CCR9- and CD8+CCR9- cell types.

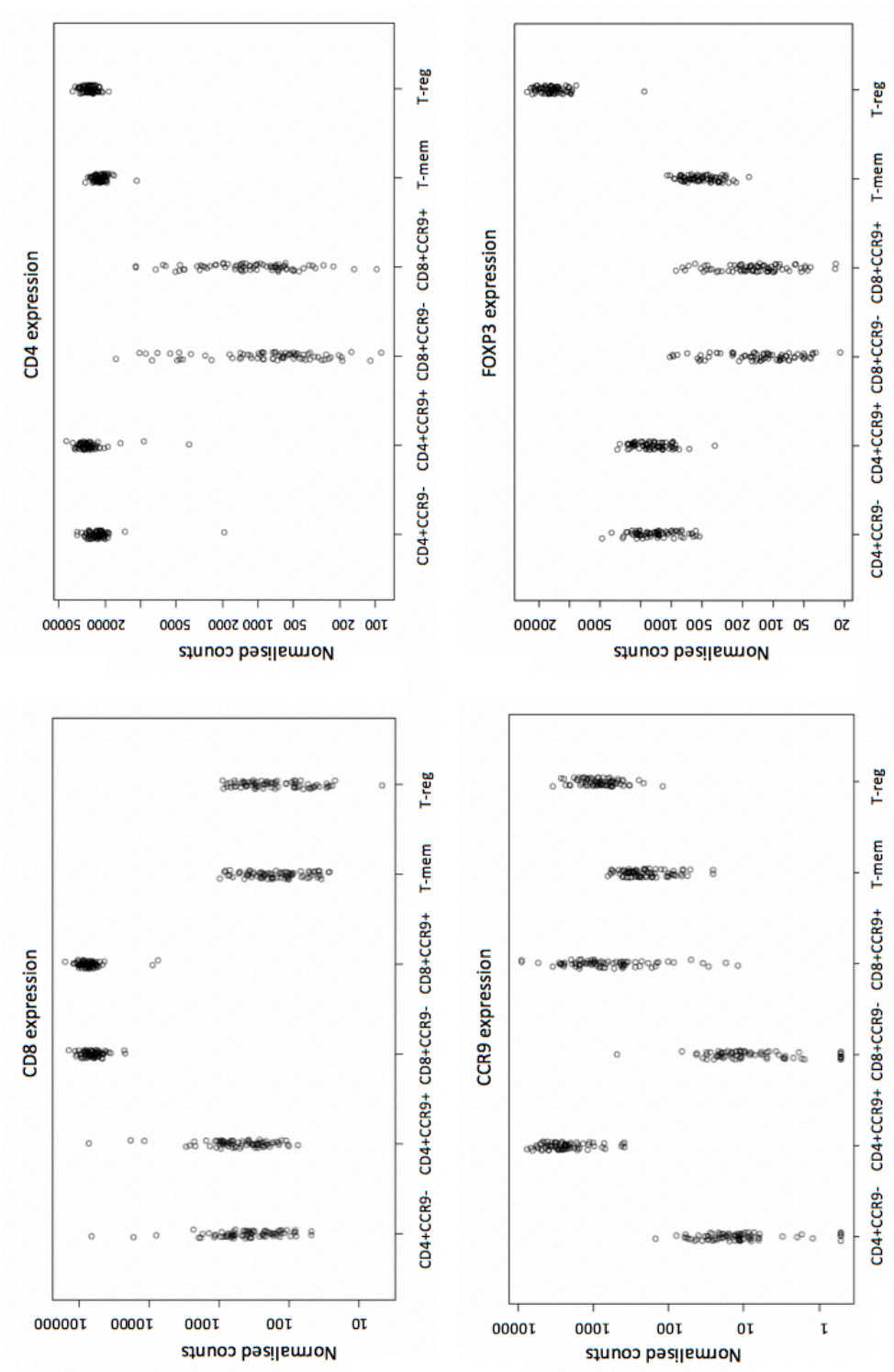


Figure 4.6: Expression of marker genes across all cell types.

4.3.5 Differential gene expression

As previously discussed, I recruited patients with both PSC-UC and lone UC for inclusion within this study, based upon evidence that the colonic inflammation in both PSC-UC

and UC patients results in increased numbers of CCR9+ effector-memory T-cells. Thus, I hypothesised that these cells would have a similar activated phenotype, with similar transcriptomic profiles in both disease groups. In order to prove that the cell types from the PSC-UC and UC groups had a similar transcriptomic profile, I performed differential gene expression (DGE) analysis between disease groups (PSC-UC, UC and HC) in each of the six T-cell subtypes.

For the analysis of differential gene expression I used *DESeq2* package version: 1.25.0. *DESeq2* is a tool for analysis of differential gene expression, using shrinkage estimation for dispersions and fold changes to improve stability and interpretability of estimates [230]. There are several similar methods available, including *edgeR* and *limma-voom*. I chose the *DESeq2* method over *limma-voom* because it offers a more stringent count normalisation method, based upon generalised linear modelling (GLM) or negative binomial modelling rather than linear modelling. This is especially important when dealing with very small sample sizes. For example, the HC sample group in my study contained only 5 individuals and *DESeq2* has been shown to have comparatively improved specificity and sensitivity as well as good control of false positive errors, even with small samples sizes [232]. In comparison with *DESeq2*, the *edgeR* method also uses a negative binomial distribution, with comparable specificity and sensitivity and I chose the former due to improved usability.

The input for *DESeq2* is the raw count matrix K (where 'count' refers to the number of sequencing reads unambiguously mapped to gene in a sample), including only those genes and samples taken forward following the aforementioned QC steps. Each row of the count matrix contains one gene *i*, and each column contains the number of counts for that gene in a sample *j*. *DESeq2* firstly normalises for sources of systemic variation between samples; library size and sequencing depth. This is important because not all samples have been sequenced to exactly the same depth and larger library sizes result in higher counts. It also normalises for two important sources of within-sample gene-specific effects. The first is related to gene length, because the total number of reads mapped to a given transcript is proportional to the expression level of the transcript multiplied by the length of the transcript [233]. The second is related to GC content which is heterogeneous across the genome and can affect the mapping of reads [234]. The method of normalisation used by *DESeq2* is called the 'median-of-ratios' method, which I have described in Figure 4.7. The output of this normalisation is a normalisation factor, S_{ij}, for each sample in the experiment [235]. *DESeq2* models the counts for K_{ij} as following a negative binomial distribution with mean μ_{ij} and dispersion α_i (dispersion estimation described more fully below). μ_{ij} is a quantity, q_{ij}, proportional to the concentration of cDNA fragments from the gene in the sample, scaled by the normalisation factor S for that sample;

$$\mu_{ij} = S_{ij} * q_{ij}$$

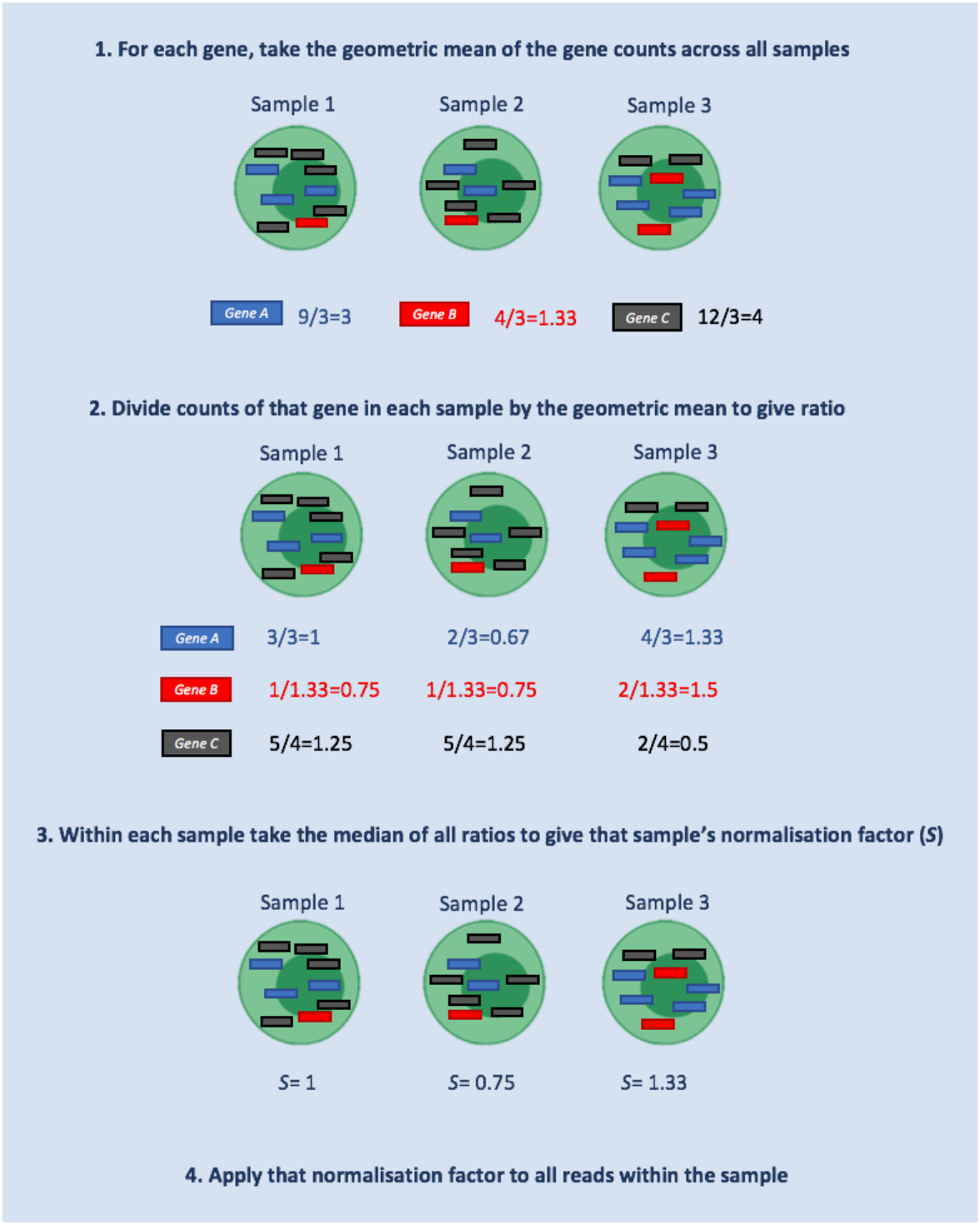


Figure 4.7: Schematic representation of the *DESeq2* method of normalisation.

To compare two groups (eg. PSC-UC versus UC), *DESeq2* fits a GLM with logarithmic link of the overall expression strength of a gene and the $\log_2$ fold change (LFC) between the two groups, as a combination of explanatory factors or covariates such as group, patient and sample;

$$log2q_{\mathrm{ij}} = a + (b * group) + (c * patient) + (d * sample) + e$$

where a is the intercept, b, c and d are parameters estimated from the data, and e is the error term. When comparing a gene's expression level between groups, *DESeq2* accounts for the within group variability of that gene's expression using dispersion estimation, α_{i} to model the variance of counts, Var K_{ij}.

$$VarK_{ij} = \mu_{\mathrm{ij}} + (\alpha_{\mathrm{i}} * \mu_{ij})$$

For the statistical inference of differential expression, it is important that estimation of the dispersion parameter, α_{i} is accurate. Because some RNAseq experiments, such as the HC group in my study, include only a few biological replicates, estimating the within group variability is difficult, especially because genes expressed at very low levels have much higher dispersion estimates. If used, these higher dispersion estimates would introduce noise and affect the accuracy of the differential expression analysis. To account for this *DESeq2* assumes that genes with a similar average expression have similar dispersion. It then estimates gene-wise dispersions (for each gene separately) using a maximum likelihood and shrinks dispersion estimates towards a fitted average dispersion curve, using an empirical Bayes approach. As sample size increases, the scale of shrinkage decreases.

When estimating log fold change (LFC), there is strong variance for genes expressed at very low levels. This is a result of working with count data, where even a small error in counting mapped reads causes a comparatively big change in LFC estimation for those genes expressed at very low levels. If unaccounted for, this would make the downstream estimation of effect sizes difficult to compare across the range of data. *DESeq2* deals with this by shrinking LFC estimates towards zero using an empirical Bayes method. This can be visualised on an MA plot, which shows the differences between measurements taken in samples, by transforming the data onto M (log ratio or LFC) and A (mean of normalised counts) scales, then plotting these values. Figure 4.8 shows two MA plots for all of the data in my differential gene expression (DGE) study, before and after shrinkage has been applied. This demonstrates that shrinkage is stronger when counts are low and dispersion is high, removing the problem of exaggerated LFCs for genes with low counts.

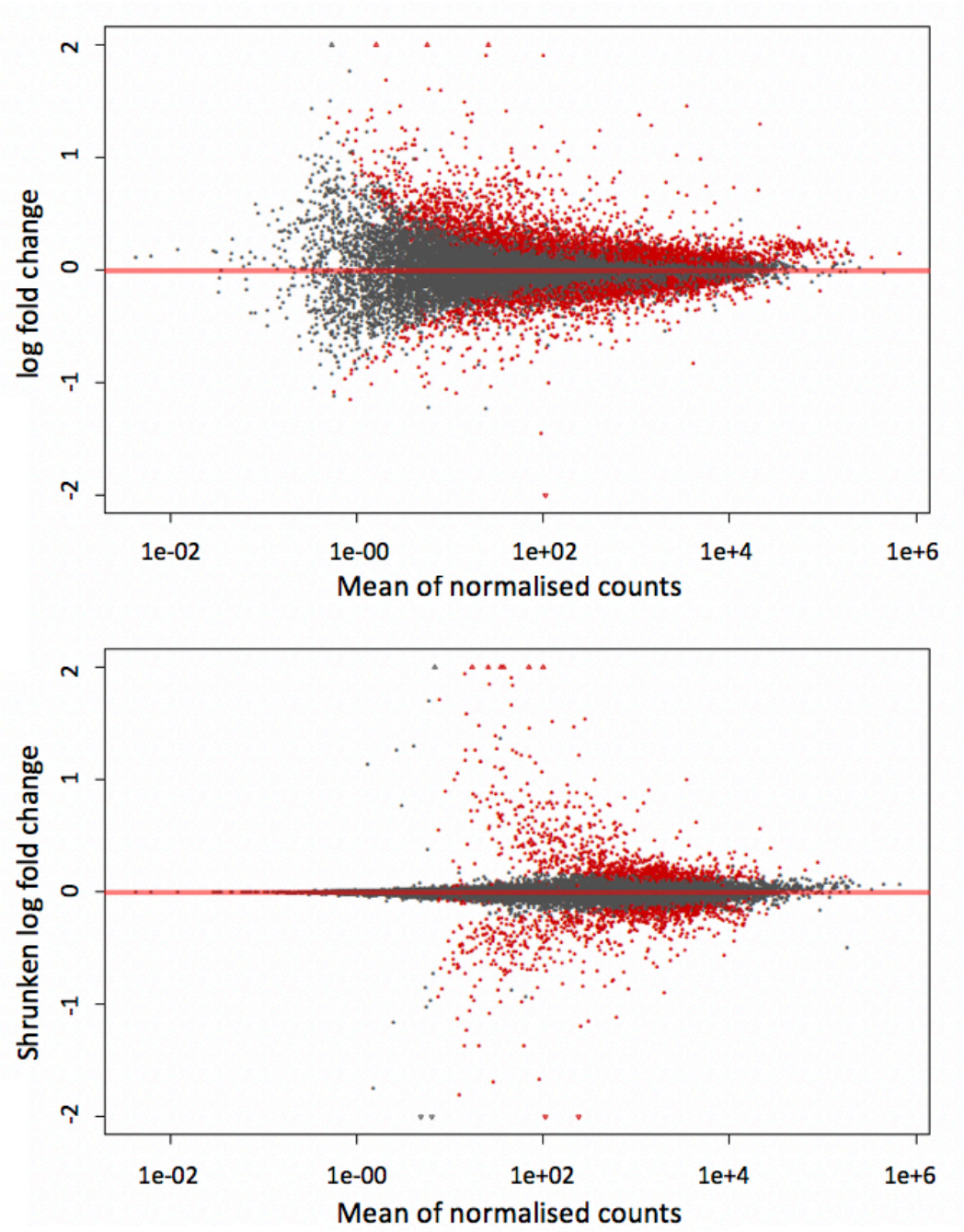


Figure 4.8: MA plots with and without shrinkage applied. Points are coloured red where the adjusted p-value is less than 0.05, and plotted as open triangles pointing either up or down if they fall outside of the window.

Having fit a GLM for each gene, the next stage is to test whether the coefficient for each model is significantly different from zero. *DESeq2* uses a Wald test for significance

where the shrunken estimate of LFC is divided by its standard error and the resulting Z statistic compared to a standard normal distribution with a resultant p-value. Because many thousands of genes are tested, it is possible to obtain some significant p-values just by chance (false positives), hence in the final stage of the analysis I corrected the p-values for multiple testing. I used the Benjamini-Hochberg (BH) correction method [164] to obtain adjusted p-values at a 5% false discovery rate (FDR).

I performed differential gene expression analysis between each of the three disease groups (PSC-UC, UC and HC), in a pair-wise fashion. I controlled for known covariates in the *DESeq2* model including patient age, sex, use of drugs including 5-aminosalicylates and azathioprine, and the sample sequencing run. I reported genes as differentially expressed if the adjusted p-value was <0.05. I performed gene ontology (GO) analysis of all DEGs in each group, using web-based GO platform, *g:Profiler* [236], to elucidate aspects of the underlying disease biology.

4.3.6 Genotype QC and imputation

Paired genotype and expression data is required for the mapping of eQTL. DNA samples from blood or saliva were available for 74 of the 76 patients. DNA extraction of all samples was performed by Dr Rebecca McIntyre, Senior Staff Scientist at the Wellcome Trust Sanger Institute. Extraction was performed using *Qiagen* DNeasy Blood and Tissue Kit and sequenced by the Wellcome Sanger Institute DNA pipelines, using the Illumina Omni2.5-8Exome BeadChip. I performed all QC on the raw genotype data, using the PLINK software version 1.9, following Anderson *et al*'s published standards for the QC of genotype data for genome-wide case-control association studies [237]. I considered all autosomal and chromosome X SNPs without insertions or deletions. The sequence of pre-imputation QC is shown in Figure 4.10 with further details on per-SNP and per-individual QC outlined below.

The removal of suboptimal SNPs is important for avoiding false-positive associations which reduce the ability to identify true associations correlated with disease risk. To remove individuals and SNPs with a particularly high error rate, but maximise the number of SNPs remaining within the study, I first removed individuals with a genotype call rate of $<95\%$ and SNPs with call rate of $<95\%$. SNPs with a very low frequency can be difficult to call using current genotype calling algorithms due to the small numbers of heterozygotes and homozygotes. Furthermore, power to detect association at rare variants is low, and thus I removed variants with a MAF <0.01.

Per-individual QC included the identification of individuals for whom information on sex was disconcordant between genotype and ascertained sex. This was done by calculating the homozygosity rate across all X chromosome SNPs for each individual within the sample and comparing this to the expected rate. Males are expected to have a homozygosity rate

around 1 (with some variation due to genotyping error), and females a homozygosity rate of around 0.2. This is because males have just one copy of the X-chromosome and thus cannot be heterozygous for any marker outside of the pseudo-autosomal Y chromosome region. There were no sex discrepancies between genotype and ascertained sex in my samples. In order to reduce the effect of population stratification, I next identified any individuals of ancestry divergent from the expected European ancestry. Excluding variants from regions of known high LD, I identified a pruned set of 62,805 independent variants from my dataset, all with an $r^2<0.2$ and MAF>0.01. I then identified the same subset of variants within the 1000 Genomes dataset. Using this pruned set of independent variants, I performed a PCA analysis of my individuals combined with the 1000 Genomes cohort. By plotting the first and second principal components of this combined dataset, I could visually identify that all of my samples were clustered with the known European individuals of the 1000 Genomes dataset (labelled 'PSC' in Figure 4.9). Notably, of all individuals passing QC and retained for downstream analysis, three were of Southern European/Iberian ethnicity, highlighted on Figure 4.9 and all remaining samples were of Northern European ethnicity. All samples from individuals of Northern and Southern European ethnicity were retained for further analysis.

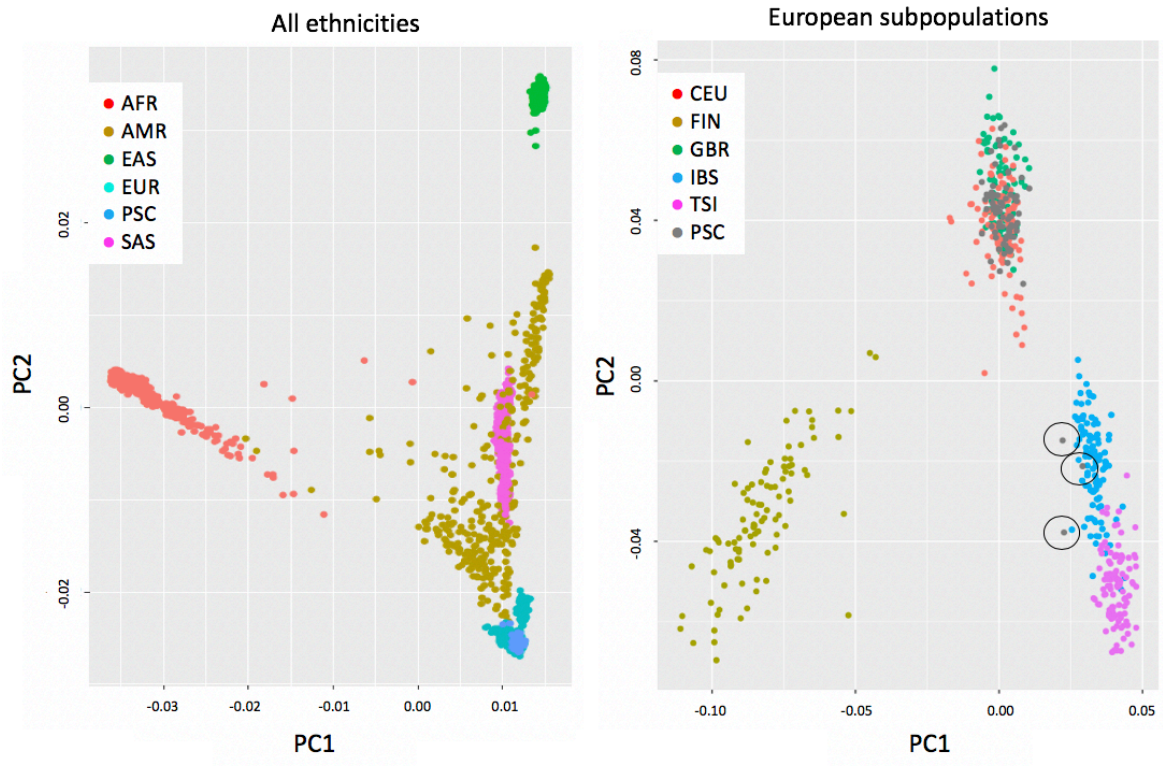


Figure 4.9: PCA of study samples compared to 1000 Genomes samples of known ethnicity using a pruned set of 62,805 independent variants with an $r^2<0.2$ and MAF>0.01.

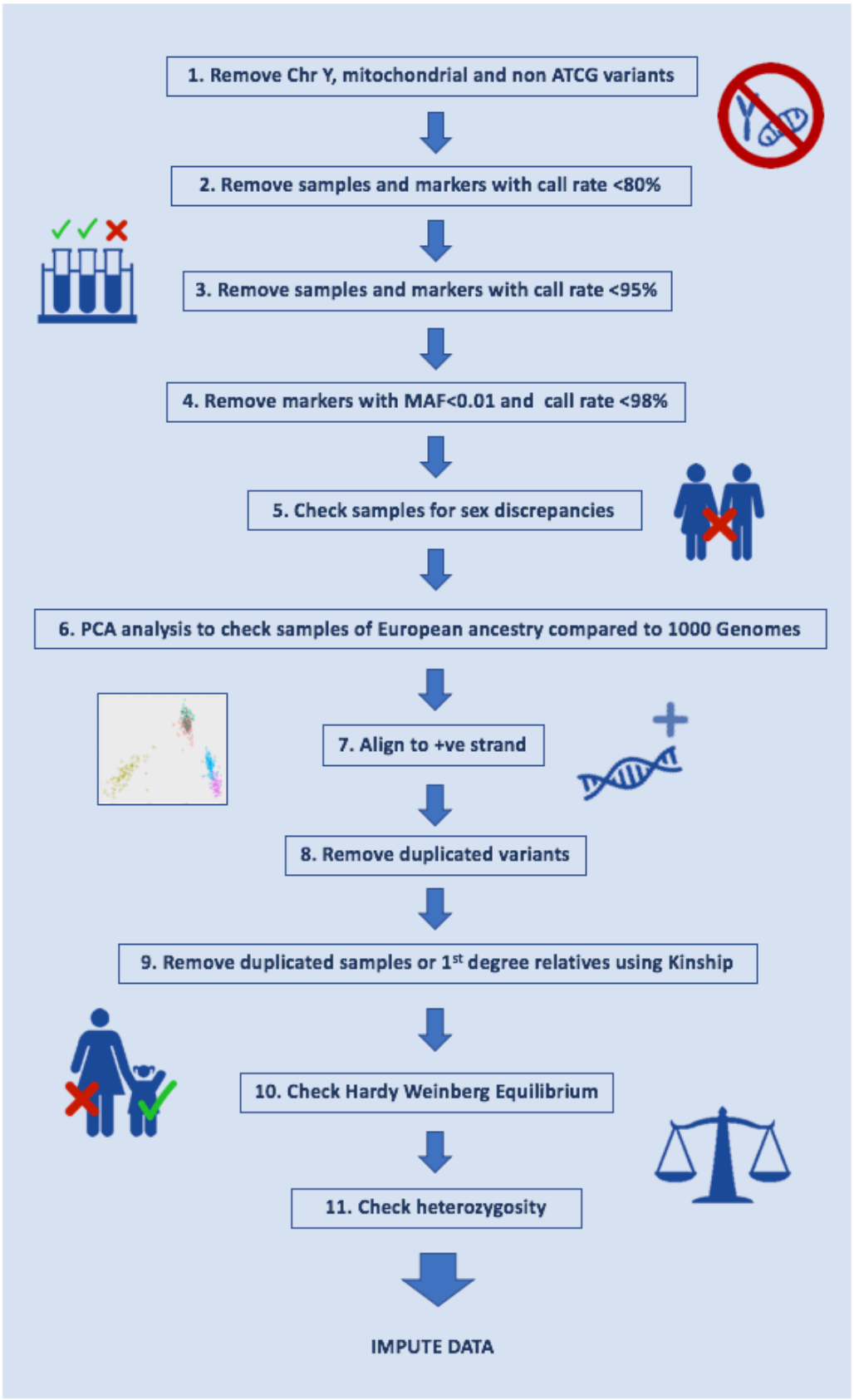


Figure 4.10: Outline of pre-imputation QC of genotype data.

The heterozygosity rate per individual can be used as a measure of DNA sample quality. Considering only autosomal chromosomes, I examined the distribution of the heterozygosity rate, excluding any samples with a heterozygosity rate more than two standard deviations from the mean. The mean heterozygosity rate was 0.274 and one sample fell outside the two standard deviations threshold resulting in its removal from the analysis. To avoid the bias of over-represented genotypes introduced by first-or second degree relatives, the next stage of per-individual QC was to identify any duplicated or related individual, to ensure the maximum relatedness between any pair of individual was less than second-degree relatives. I used KING software v2.2 and a set of 2,513,131 variants with MAF>0.01, and call rate >98% to infer close relatives based on the estimated kinship coefficients. I identified two first degree relatives (kinship coefficient range 0.177 to 0.354), one of which was removed from subsequent analysis.

SNPs with extensive deviation from Hardy-Weinberg Equilibrium (HWE) may indicate selection, occurring at loci associated with disease, but can often be indicative of genotype calling error. As part of the per-marker QC, I removed variants with a HWE p-value of $<1 \times 10^{-8}$. Following the above QC steps a dataset including 71 individuals and 1,590,593 variants remained, and were put forward for imputation. I imputed a further ∼5.5 million variants against the UK10K, 1000 Genomes phase 3 and Haplotype Reference Consortium reference panels, using the Wellcome Sanger Imputation and Phasing Service pipeline, IMPUTE2 [238]. IMPUTE2 provides an 'info' score related to the quality of the imputation for each SNP. Post-imputation QC consisted of removing any SNPs with a low info score <0.3. This threshold was decided by plotting an info score frequency curve and assigning the threshold at the inflexion point [239]. The final post-imputation QC step was to re-check the HWE as described above. The resultant post-imputation, post-QC dataset consisted of 7,027,506 SNPs. Mapping of eQTLs requires the addition of known covariates within the model, including principal components (PCs) from the genotype data. Therefore, using the final QC'd and imputed genotype dataset, I performed a PCA using the PLINK (v1.9) *PCA* function with the aforementioned pruned set of 62,805 independent variants from low LD regions. I retained the resulting genotype PCs for inclusion as covariates in the downstream eQTL analysis.

I processed all genotype and imputed data in ensembl build 37, but for further downstream processing performed a genome coordinates conversion or 'lift-over' to ensembl build 38 using CrossMap v0.3.5 which supports the conversion of variant call format (VCF) files between different genome assemblies [240].

4.3.7 eQTL mapping

I conducted all eQTL analysis and mapping using *QTLtools* v1.1 9, which provides a complete toolset for molecular QTL discovery and analysis [241]. The analysis outlined

below was performed using a normalised gene expression matrix, which had undergone prior QC (as described in the RNA sequencing and sample QC section above), and the previously QC'd and imputed genotype data (as described in the Genotype QC and Imputation section above).

4.3.7.1 Identifying sample mismatches and amplification bias

To ensure that the genotype and gene expression data for each individual in the study was a true match, I used the MBV (Match BAM to VCF) module of *QTLtools* [231]. MBV identifies sample mislabelling, cross-sample contamination and PCR amplification bias. The input files for MBV were the VCF file containing the genotype data for all 71 individuals within my study, and the BAM file for the mapped RNA reads for each individual at a time. For each SNP site in the VCF file, MBV aggregates the sequencing reads and discards those SNPs not reaching a minimal coverage parameter threshold. For each individual within the VCF file, it calculates the proportion of heterozygous and homozygous genotypes for which both alleles have been captured by the sequencing reads and reports the two concordance measures for each individual. Where both measures are close to 100% concordance, this describes a match between genotype and gene expression datasets. Where there is decreased heterozygous concordance with no change in homozygous concordance this is described as 'no match' between genotype and gene expression, but in fact represents a match but with amplification bias effect (Figure 4.11). Twenty-three percent of samples demonstrated heterozygosity concordance of less than 0.66 with no change in homozygous concordance. In order to account for the effect of such amplification bias, the fraction of heterozygosity concordance for each sample was taken forward as a covariate for inclusion in the eQTL analysis.

There were no instances of sample contamination within this dataset, which can be detected by a reduction in the fraction concordance at homozygous compared to heterozygous sites. I detected four cases of 'unexpected matches', two from the same male recruit and two from the same female recruit (Figure 4.12). These were the same four samples detected to be outliers on PCA according to sex, as previously described in the RNA sequencing and sample QC section above. This was the result of an accidental direct swap of two RNA sample labels (CD8+CCR9- and CD8+CCR9+ samples) from one male individual with two RNA sample labels for the same two cell types from one female individual. Following this stage of QC, these four samples could be re-assigned to the correct individual and therefore retained for eQTL mapping.

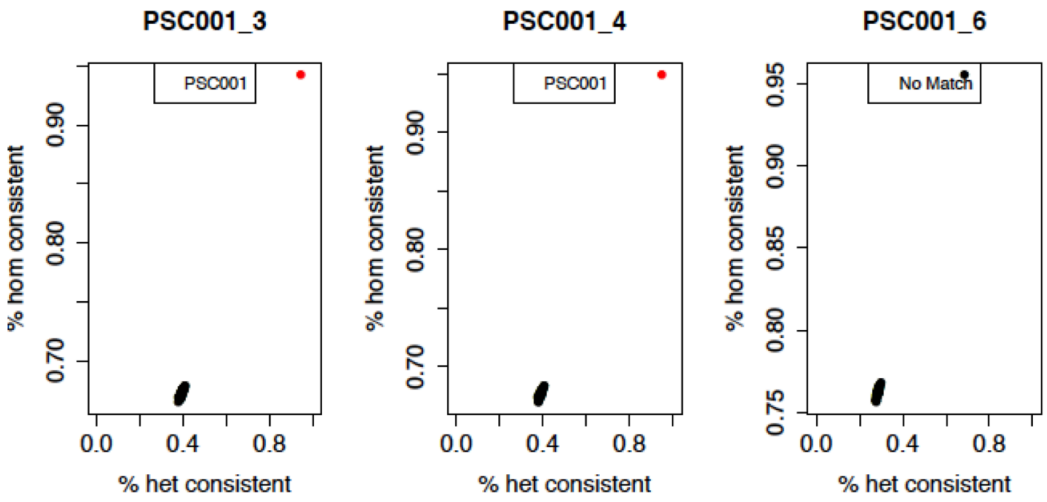


Figure 4.11: Concordance at heterozygous genotypes (x-axis) versus concordance at homozygous genotypes (y-axis), for each individual genotype sample (black dots). A match between genotype (box at top) and gene expression data (plot title) is coloured red (two left hand examples). A mismatch or amplification bias is coloured black (right hand example).

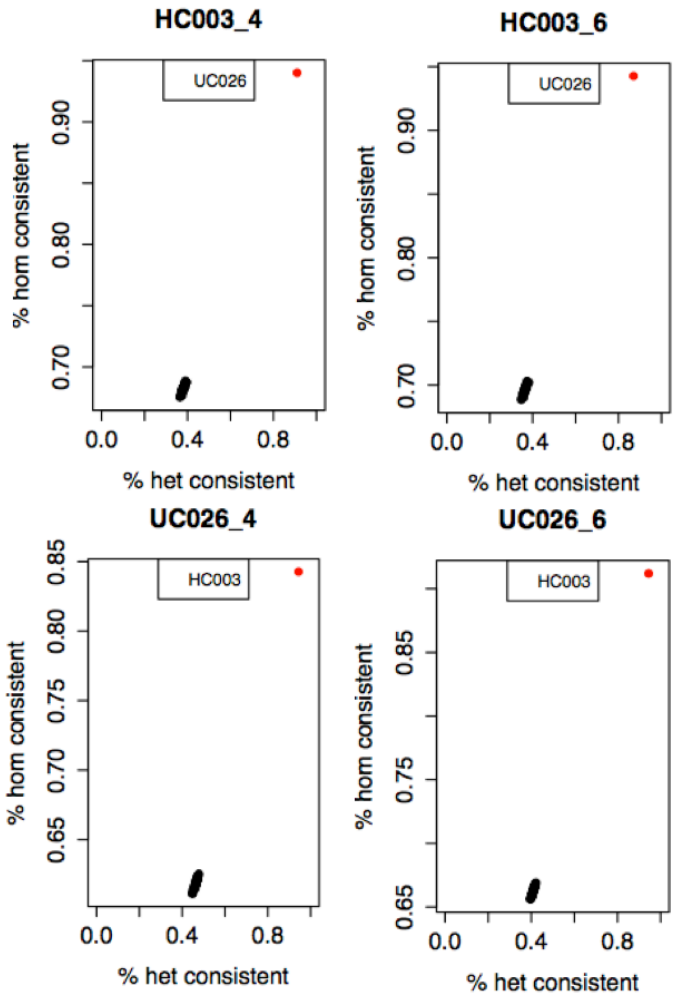


Figure 4.12: Concordance at heterozygous genotypes (X-axis) versus concordance at homozygous genotypes (Y-axis), for each individual genotype sample (black dots). An sample mismatch is shown by a match between a different genotype (in box at top) and gene expression data (plot title) in all four examples.

4.3.7.2 Identifying *cis*-eQTLs

For the identification and mapping of *cis*-eQTLs in each of my T-cell subsets, I used *QTLtools* [241]. Mapping eQTLs involves the testing of association between gene expression (phenotype of interest) and all the genetic variants within a window upstream and downstream of the transcription start site (TSS) of the gene, with millions of tests performed genome-wide. A linear regression model is fitted between the genotypes and gene expression, including multiple covariates to correct for batch and other effects, in order to find the best nominal associated variant per gene. Analysis of all gene-variant pairs requires millions of association tests, each producing a nominal p-value. Whilst adjustment of nominal p-values to correct for multiple testing and avoid false positives

must be performed, the presence of linkage disequilibrium (LD) means that the tests are not entirely independent, calling for a less stringent correction than the Bonferonni method. To deal with this issue *QTLtools* uses permutations to derive adjusted p-values per phenotype/gene. *QTLtools* uses a beta approximation permutation scheme based on Ongen *et al*'s *FastQTL* beta approximation permutation scheme, to correct for the testing of multiple variants per gene [242]. This scheme creates multiple permuted datasets by keeping the genotypes static (thus preserving correlation structure between variants) and permuting the gene expression data for every gene. For every permutation the best nominal association is retained to form a distribution of p-values expected under the null hypothesis of no association. Next, an adjusted p-value is calculated based on how likely it is that an observed association obtained in the nominal pass, originates from the null. *QTLtools* models this distribution of p-values expected under the null hypothesis of no associations, using a beta distribution. It approximates the tail of the null distribution to estimate adjusted p-values at any significance threshold, with no lower bounds.

The input files for *QTLtools* are the zipped and indexed VCF file (which had been previously QC'd and imputed as described above), the indexed and zipped gene expression BED files (reporting normalised expression in TPM and QC'd as previously described) and the covariate files. In my analysis I included age, sex, the first three genotype principle components (described in Genotype QC and imputation section above), fraction of heterozygosity concordance (described in Sample mismatch and amplification bias section) and a variable number of gene expression-derived principal components (PCs) as covariates. I calculated the gene expression PCs in the same way as the genotype PCs, using PLINK v1.9's *PCA* function.

To map eQTLs I ran *QTLtools* for each of the six T-cell datasets, using 1,000 permutations and a *cis*-window of 1Mb. To maximise the number of eQTL discoveries, I optimised the QTL mapping by performing multiple runs of the analysis including an increasing number of gene expression-derived principal components from zero to 50. To account for the thousands of genes tested genome-wide, I performed an FDR correction on the set of adjusted p-values obtained by the permutation analysis for every gene, using the R package, *qvalue* [243]. In contrast to the p-value, which measures significance in terms of the false positive rate, the q value is a measure of significance in terms of the false discovery rate. An FDR threshold of 5% therefore means that on average, 5% of the eQTLs called significant are truly null [243]. In order to find the optimal number of gene expression PCs required to detect the maximum number of eQTLs, I plotted the number of eQTLs against the number of expression PC's included within the linear regression model. For each cell type, I settled on the number of PC's that maximised the number of eQTLs, and included this number of PCs in the covariate model which was taken forward for subsequent analyses as described below [215, 241].

Further analysis of eQTL data, for example for meta-analysis or colocalisation, requires all the nominal associations (including those that do not reach statistical significance). To generate this data, I used the *QTLtools nominal pass* function, the same gene expression BED and genotype VCF files as described above and the covariate files containing the same number of gene expression PCs for detecting the maximum number of significant eQTLs.

4.3.8 Identifying shared and tissue-specific eQTL

Having mapped eQTLs for six individual cell types, an important question is to identify those eQTLs which are shared across cell types, and those that are cell-type specific. By allowing for the correlations of effect sizes among cell types using a form of meta-analysis, this can increase power by improving estimation of effect sizes and allow for more accurate comparison of effect sizes between tissues. Several statistical methods for analysing shared eQTL associations have been published which learn the patterns of eQTL sharing from the data using a hierarchical model [244–246]. However, each has its own limitations, for example the model by Flutre *et al* is limited by the assumption that correlations are non-negative and equal, such that it does not allow for genetic variants leading to an increased effect in one trait and a decrease in another [244]. Furthermore, Flutre *et al*'s methods provides flexibility at the cost of becoming computationally intractable when considering even moderate numbers of tissues or cell types and thus the authors sought to solve this by restricting effects to either a single cell type, or shared across all cell types. Another method published by Wei *et al* allows for all patterns of sharing, but is limited by the assumption that nonzero effects are uncorrelated among conditions, and thus focuses only on testing for significant effects and not on estimating effect sizes [245]. *MashR* (multivariate adaptive shrinkage in R) is a method that addresses these limitations, allowing for shared, condition-specific and arbitrary patterns of correlation among conditions, as well as providing measurements of significance and effect size estimates [246].

I used *mashR*, implemented in R, for further analysis of my eQTL data. The input data for *mashR* are the nominal pass of the individual cell-type analysis performed with *QTLtools* as described above. These include the effect size estimates (β's) and corresponding standard errors (SE) for all eQTL/Gene pairs in each cell type with no significance threshold. These measurements are the input for *mashR*'s two-step empirical Bayes procedure, which firstly learns the patterns of sparsity, sharing and correlations among effects from the individual cell-type results and secondly, combines these learned patterns to produce improved estimates of effect and their corresponding significance. For the first step, *mashR* requires a subset of 'strong' tests, corresponding to the strongest effects in the individual cell-type analysis. I identified this subset of 'strong' tests by taking the most significant eQTL per gene across all six cell types, from all significant eGenes

from the individual cell-type analysis. This produced a strong subset of 5,487 eGenes. Next, *mashR* requires a 'random' subset of all tests, which is an unbiased representation including null and non-null tests. I created a 'random' subset of 200,000 tests using the R function, *set.seed*, which is a reproducible random number generator. The random subset is used by *mashR* to estimate the correlation structure between tests, via a PCA-like approach and the strong subset is used to define the data-driven covariance matrices. The *mashR* model is then fitted to the random tests using both the data-driven covariance and *mashR*'s in-built canonical covariances. I then used the resultant *mashR* model to compute posterior summaries for all of my data. For each eQTL/Gene test in each of the six cell-types, the output includes the posterior β, SE, *lfsr* (local false sign rate, analogous to an FDR) and $\log_{10}$Bayes factor (a measure of the overall significance for a non-zero effect in any condition).

The final stage of the analysis is to call cell-type specific and shared eQTL from the *mashR* posterior summaries. From the posterior summaries for all of my data, I identified the subset of eQTL/Gene pairs significant in at least one cell type at *lfsr*<0.05. From this subset I extracted data for the most significant eQTL per gene, defined by the eQTL/gene pair with the smallest *lfsr* across any of the six cell types as described by Kim-Hellmuth *et al* in the analysis of cell-type specific eQTLs in the GTEx data [247].

4.3.9 Colocalisation

I performed colocalisation with the eQTL data derived for each individual cell type using the output data from *QTLtools'* nominal pass and permutation pass. I performed colocalisation at the fifteen PSC risk loci reported by Ji *et al* [42] with GWAS summary statistics from the same study using the same methods for colocalisation as previously described in Chapter 3. Where the PP4 for colocalisation of a PSC risk locus with a T-cell eQTL was >0.8 for at least one cell type, I explored whether this same locus also colocalised with the same eQTL in other cell types using the *mashR* eQTL data. I took the posterior results of *mashR* analysis for posterior standard deviation (standard error), lfsr (analogous to an FDR) and posterior mean (β) for each cell type, and performed colocalisation at those PSC risk loci, visualising the results on regional association plots. Finally, given that the majority of the study cohort were patients with UC and that genetic architecture is shared across many IMDs, I conducted colocalisation with other IMDs. I performed colocalisation with 240 IBD, 100 RhA and 45 T1DM risk loci, using their associated GWAS summary statistics [60, 148, 200] and nominal pass eQTL data for each T-cell subset (derived from the *QTLtools* individual cell-type eQTL analysis).

4.4 Results

4.4.1 Differential gene expression

I tested 20,547 genes for differential expression between each of the three disease groups (PSC-UC, UC and HC). Characteristics of the study cohort, according to disease group are shown in Table 4.2. I controlled for covariates including patient age, sex, use of 5-aminosalicylates or azathioprine and the sample sequencing run. The results of this analysis showed no significant differences is gene expression across all six T-cell subtypes in the PSC-UC group compared to the UC group (Table 4.3). Given that both groups share the UC phenotype, this finding is not unexpected. Furthermore, the results supported no significant changes in gene expression between both the PSC-UC and UC groups versus HC, in T-regs, CD4+CCR9-, CD4+CCR9+ and CD8+CCR9+ cells. Whilst there were a few DEG's (≤ 7) between the above comparator groups, genes are reported at a 5% FDR, therefore a false positive rate of 5% is expected, limiting any interpretation where such low number of DEGs are reported. Further visualisation of normalised counts for these few genes in each disease group confirmed that most genes reported as differentially expressed, were false positives.

Differential gene expression was observed between both PSC-UC and UC groups compared to HCs in two cell populations; T-memory and CD8+CCR9- T-cells. Using *gProfiler*, I performed GO analysis of all genes differentially expressed between these disease groups. GO term analysis of 367 DEGs in the T-memory cells of UC compared to HCs demonstrated enrichment of pathways involved in cellular metabolic activity ($p=1.1 \times 10^{-12}$) (Figure 4.13). The finding of a more metabolically active phenotype in the T-memory cells of patients with UC versus HCs may support a role for these cells in the disease pathogenesis. GO analysis was unable to find any more specific pathway enrichment based upon these DEGs. There were 101 DEGs between PSC-UC and HCs in T-memory cells. However GO analysis did not find any significant pathway enrichment, likely a result of the relatively low numbers of DEGs between these two groups.

The second cell type demonstrating significant numbers of DEGs in the PSC-UC and UC groups compared to the HC group, were the CD8+CCR9- T-cells. Here, 94 and 34 genes were differentially expressed in PSC-UC and UC groups compared to HCs respectively. GO analysis did not find any specific pathway enrichment for any of the DEGs, again, likely a result of the low numbers of DEGs. However the finding of a difference between the transcriptomes of CD8+CCR9- cells of PSC-UC and UC patients versus HCs, in the absence of any difference in the transcriptomes of CD4+CCR9- in the same groups, is interesting given existing evidence in IBD, of a CD8+ T-cell signature of immune-cell exhaustion, driving a more severe disease course in IBD [248]. Indeed, it has been reported that elevated expression of genes involved in antigen-dependent T-cell

Table 4.2: Characteristics of the study cohort according to disease group.

	PSC-UC n=42	UC n=29	HC n=5
Gender (% male)	78	69	60
Mean Age (Range)	50 (17-86)	52 (48-56)	44 (28-51)
UDCA use (%)	71	0	0
5-ASA use (%)	67	90	0
Azathioprine use (%)	57	21	0

responses, including IL-7 signaling and TCR ligation, specific to CD8+ T-cells and absent in CD4+ T-cells, can predict a more severe disease phenotype in IBD patients [248, 249]. Importantly, I found that several genes involved in TCR antigen recognition were up-regulated in CD8+CCR9- T-cells of PSC-UC groups versus HC, including *TRAV38-2DV8* (T cell receptor alpha variable) and *TRBV25-1* (T cell receptor beta variable), genes which encode the variable domain of T cell receptor (TCR) α and β chains respectively. In the CD8+CCR9- T-cells of PSC-UC versus HC, there was increased expression of *BTLA* (B- and T-lymphocyte attenuator), a gene induced during activation of T cells, and decreased expression of *IL-15*, a cytokine which prevents apoptosis and maintains memory T cells in the absence of antigen. Thus it appears that the CD8+CCR9- memory T-cells in PSC patients express genes consistent with a more active phenotype with reduced repression of apoptosis, compared to HCs.

Table 4.3: Comparison of differentially expressed genes for each of six T-cell subtypes according to disease group, reported at 5% FDR

Samples	Group 1	Group 2	Total no. of DEGs	No. up-regulated	No. down-regulated
T-reg	PSC-UC	UC	3	1	2
T-reg	PSC-UC	HC	5	3	2
T-reg	UC	HC	3	1	2
T-mem	PSC-UC	UC	0	0	0
T-mem	PSC-UC	HC	101	32	69
T-mem	UC	HC	367	143	224
CD4+CCR9-	PSC-UC	UC	1	0	1
CD4+CCR9-	PSC-UC	HC	7	1	6
CD4+CCR9-	UC	HC	4	1	3
CD8+CCR9-	PSC-UC	UC	0	0	0
CD8+CCR9-	PSC-UC	HC	94	47	47
CD8+CCR9-	UC	HC	33	27	6
CD4+CCR9+	PSC-UC	UC	1	1	0
CD4+CCR9+	PSC-UC	HC	1	0	1
CD4+CCR9+	UC	HC	2	1	1
CD8+CCR9+	PSC-UC	UC	2	2	0
CD8+CCR9+	PSC-UC	HC	6	2	4
CD8+CCR9+	UC	HC	3	0	3

Numbers of differentially expressed genes (DEGs) are for group 1 versus group 2

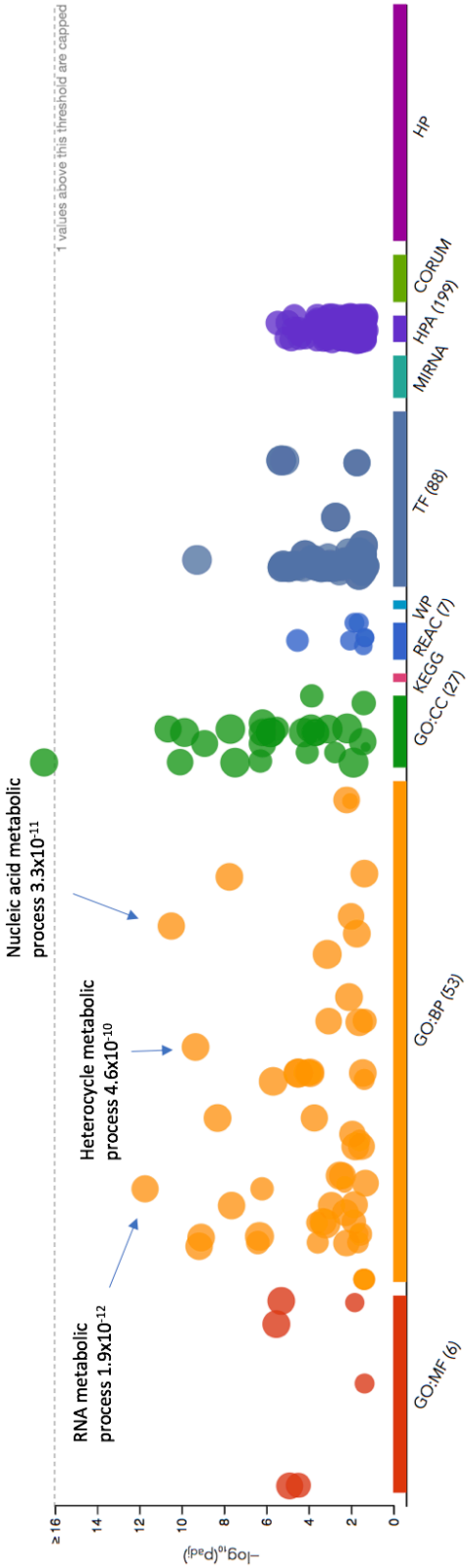


Figure 4.13: Gene ontology pathway analysis for DEGs in T-memory cells of UC compared to HC. Figure generated using g:profiler [236], 20/12/2019.

4.4.2 eQTL mapping

I mapped *cis*-eQTLs for six T-cell subtypes. For four of the T-cell subtypes (T-regulatory, T-memory, CD4+CCR9- and CD8+CCR9- T-cells), the optimal number of expression-derived PCs for detecting maximum number of significant eQTLs at 5% FDR was nine, for T-regs and CD8+CCR9- T-cells this number was eight (Figure 4.14). After extracting all significant eQTL/gene pairs, I detected a median of 1,337 eQTLs per cell type (5% FDR). The largest number of eQTLs (2,804) were detected in T-memory cells and the fewest (901) in CD8+CCR9+ cells (Figure 4.14). This is likely to reflect that T-memory cells were the most abundant cell type, and CCR9+ cells the least abundant, thus influencing the power to detect eQTLs for each cell type. Whilst data for the initial numbers of cells per sample was not available, the lesser-abundant CCR9+ cells underwent more amplification bias compared to the other cell types, as represented by the heterozygosity concordance rate was included within the covariate model. For each cell type, I plotted the position of each eQTL in relation to the gene transcription start sites (TSS), demonstrating that the majority of significant eQTLs were within 100,000 bp of the TSS (Figure 4.15). This is in keeping with the findings of several previous studies that most *cis*-eQTLs occur in close proximity to gene TSS [120, 250, 251].

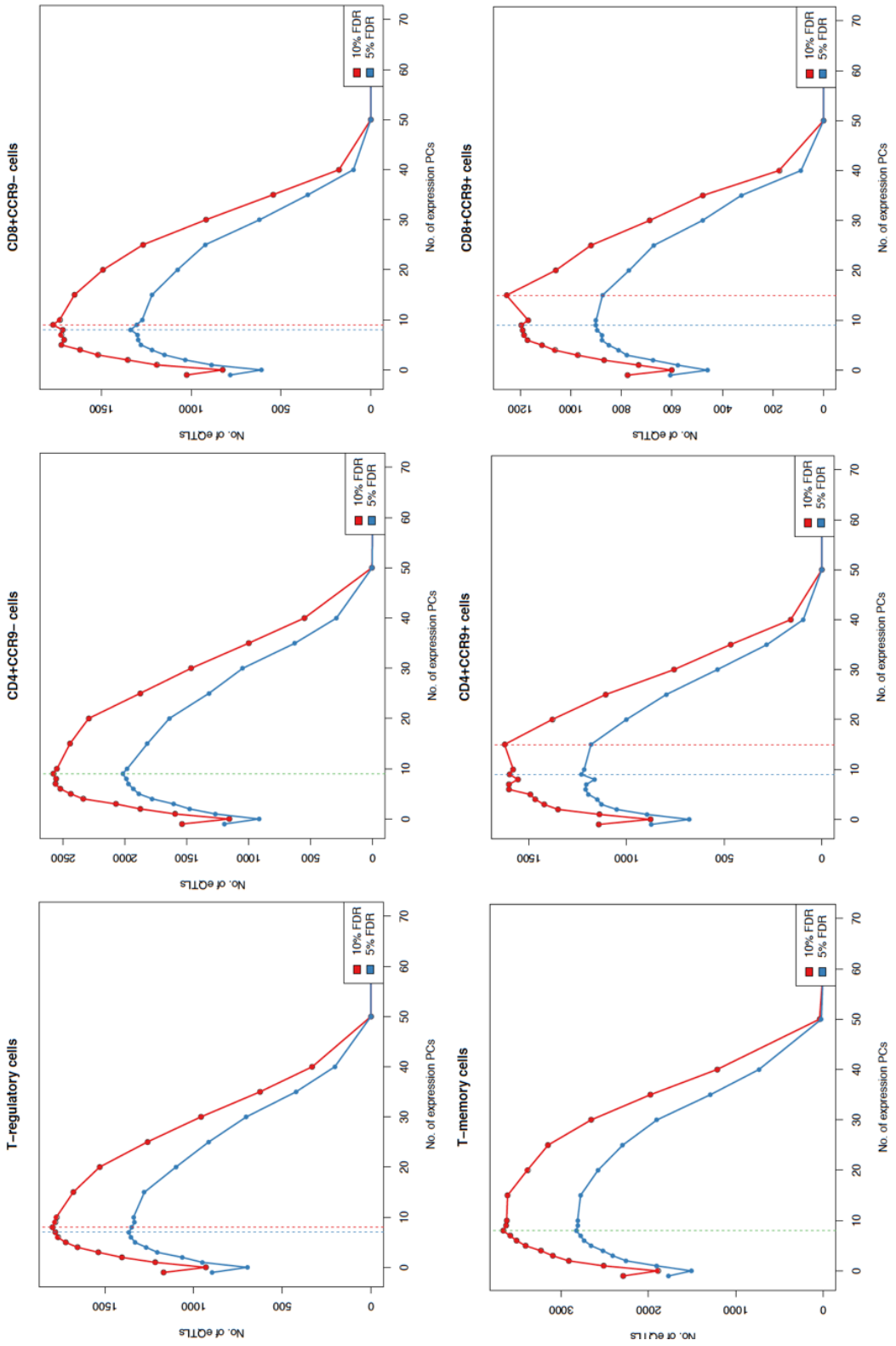


Figure 4.14: Number of significant eQTLs (y-axis) mapped for each individual cell type at 5% (blue line) and 10% FDR (red line), using covariate models with different numbers of gene-expression derived PCs from zero to fifty (x-axis).

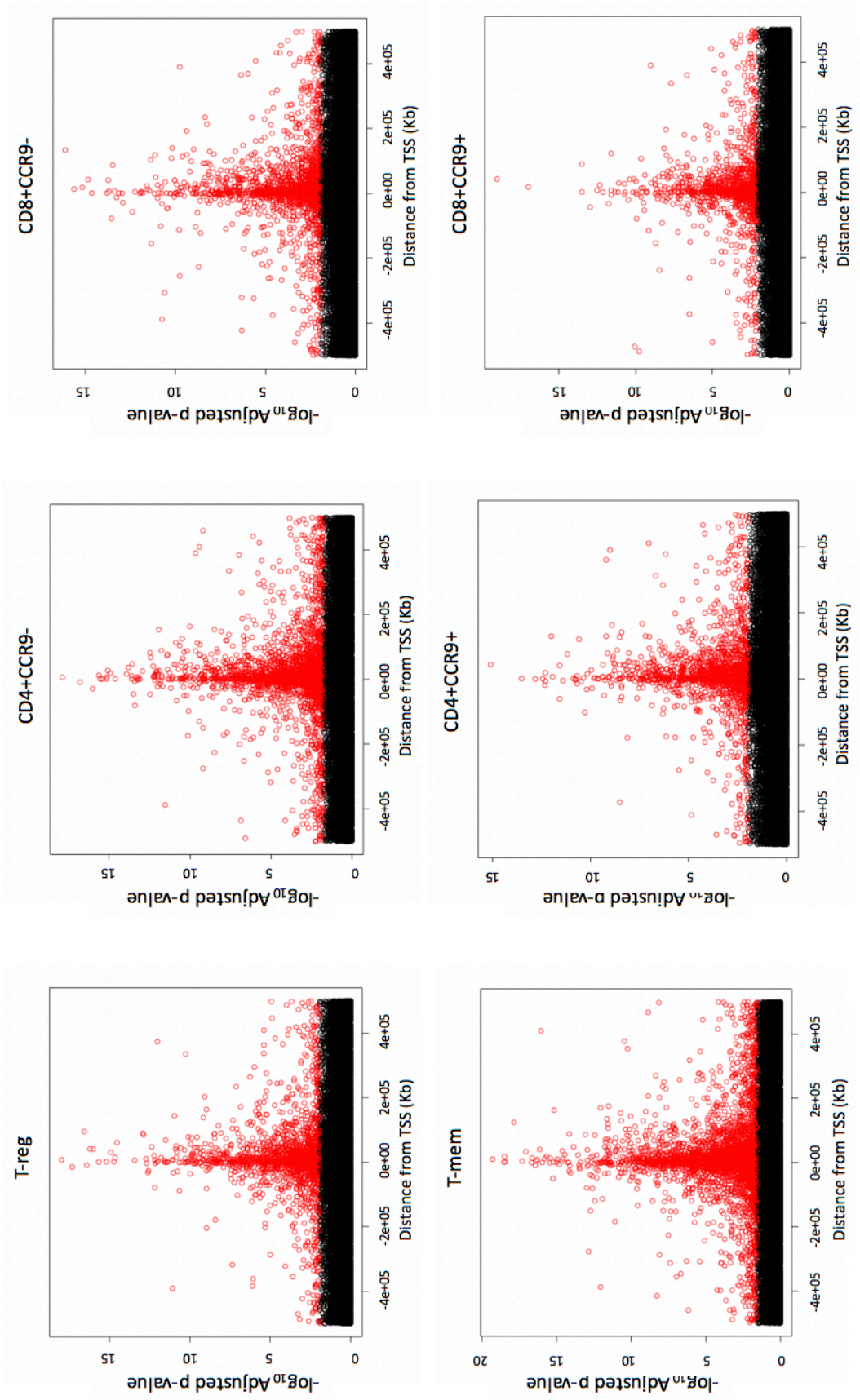


Figure 4.15: Distance from transcription start site (TSS) for each significant eQTL (coloured red for those less than 5% FDR) per cell type.

4.4.3 Shared and tissue-specific eQTLs

With *mashR*, I identified a set of 10,459 significant unique eGenes (5% FDR). This number was more than three times the sum of all significant, unique eGenes detected in the individual cell-type analysis, demonstrating the enormous boost in power provided by the aggregation of measurements across the six cell types to improve the estimates of the β/SE's. Of these 10,459 unique eGenes, 87% (9,176) were shared across all 6 cell types, 4.7% (489) were specific to a single cell type. The distribution of eQTL-sharing across the six cell types is shown in Figure 4.16. These data suggest that the vast majority of eQTLs are shared across all six T-cell subtypes, with very few cell-type specific eQTLs. This finding is not unexpected given that all six of these cell types are subsets of peripheral blood T-cells subject to similar disease conditions. GO analysis of the eGenes using g:profiler [236] did not highlight any gene sets or pathways enriched for cell-type specific or shared eQTLs.

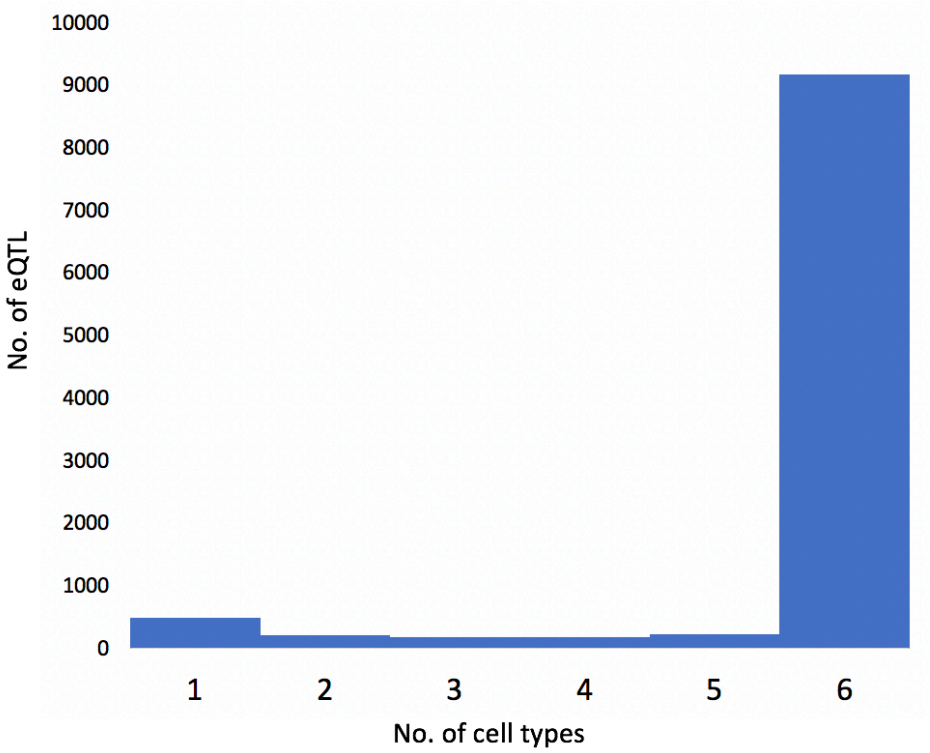


Figure 4.16: Number of cell-type specific and shared QTLs.

4.4.4 Colocalisation of disease-risk loci with eQTL

To identify eQTLs with a causative role in PSC pathogenesis, I performed colocalisation of the fifteen PSC risk loci reported by Ji *et al* [42], with the PSC GWAS summary statistics and eQTL data for each individual T-cell subtype. Two of the fifteen risk loci colocalised (PP4≥0.8) with eQTLs in one or more T-cell subtypes (Table 4.4).

Table 4.4: Colocalisation of PSC risk loci with eQTLs mapped in individual cell-types and eQTLs mapped with *mashR*

					Individual cell type analysis			*Mashr* analysis		
Chr	GWAS SNP	Condition	eGene	Cell type	PP4	eQTL Beta	eQTL p-value	PP4	eQTL post mean (beta)	eQTL lfsr (p-value)
11	rs663743	PSC	*AP003774.1*	T-reg	0.99	0.98	7.27E-06	0.99	1.00	1.90E-18
				T-mem	0.95	1.07	1.35E-07	0.95	0.82	1.71E-14
				CD4+CCR9-	0.95	0.96	1.83E-05	0.99	0.93	3.15E-15
				CD4+CCR9+	0.44	0.86	8.95E-04	0.72	0.81	3.41E-11
				CD8+CCR9-	0.70	0.88	4.38E-04	0.96	0.83	4.44E-12
				CD8+CCR9+	0.21	0.75	2.10E-03	0.00	0.74	5.36E-10
21	rs1893592	PSC	*UBASH3A*	T-reg	0.09	-1.32	1.06E-02	0.98	0.67	4.81E-08
				T-mem	0.91	0.93	4.83E-04	1.00	0.83	1.29E-11
				CD4+CCR9-	0.29	0.79	3.37E-02	0.99	0.77	4.42E-10
				CD4+CCR9+	0.47	0.88	6.69E-03	1.00	0.67	3.87E-10
				CD8+CCR9-	0.23	0.77	5.26E-02	0.99	0.68	9.10E-09
				CD8+CCR9+	0.02	0.82	4.71E-01	0.00	0.60	1.00E-06

Colocalisation of the Chromosome 21 rs1893592 PSC risk locus demonstrated that this locus was an eQTL of *UBASH3A* in T-memory cells. Whilst this is in keeping with my previous finding of colocalisation of this locus with an eQTL for *UBASH3A* in both T-regs and CD4+ naive T-cells in Chapter 3, there was no evidence from the individual cell type analysis to support colocalisation with this eQTL in the other T-cell subsets (all PP4<0.5) (Table 4.4). To identify if this GWAS risk locus was an eQTL of *UBASH3A* across all T-cell subsets, I conducted colocalisation with the eQTL data from the *mashR* analysis. Colocalisation with the *mashR* data supported that this risk locus colocalised with an eQTL for *UBASH3A* across five of the six cell types, including T-regs, T-mems, CD4+CCR9-, CD4+CCR9+ and CD8+CCR9- T-cells, with PP4≥ 0.98 (Figures 4.17 and 4.18). This supports the finding that the Chromosome 21 rs1893592 SNP is an eQTL of *UBASH3A* across most T-cell sub types. Furthermore, plotting of the *UBASH3A* eQTL at this SNP confirmed that the PSC risk increasing rs1893592*A allele reduced expression of *UBASH3A* across all T-cell subtypes (Figure 4.19), in keeping with my previous findings for this locus in Chapter 3.

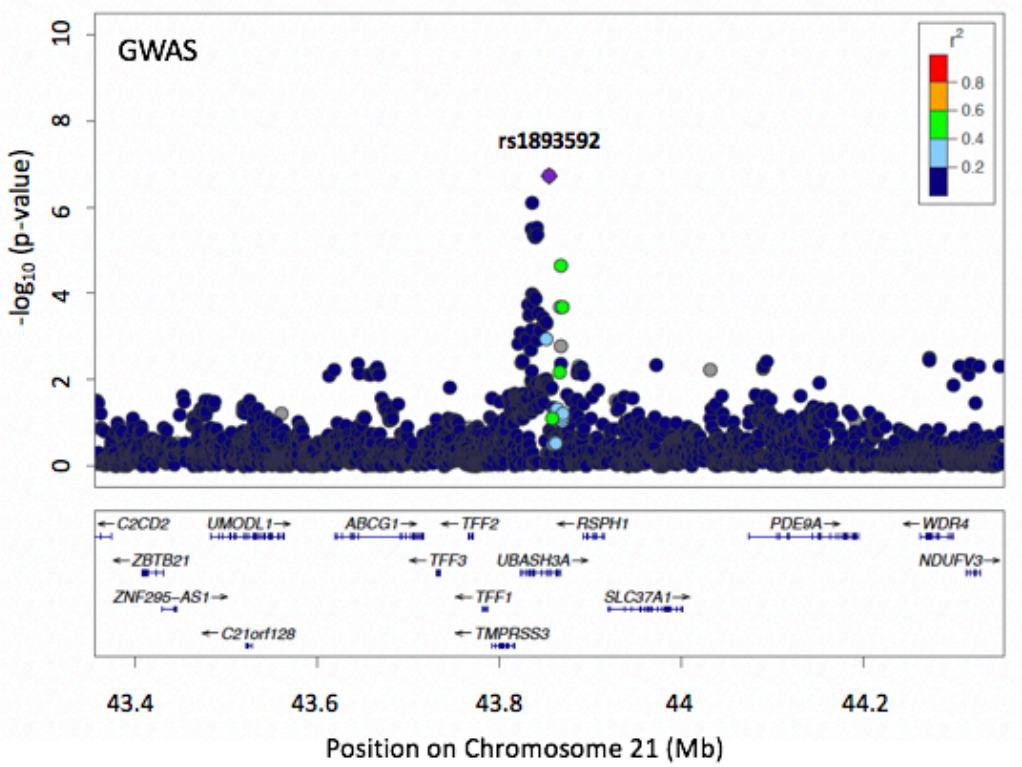


Figure 4.17: Regional association plot for the Chromosome 21 rs1893592 risk locus in PSC GWAS data.

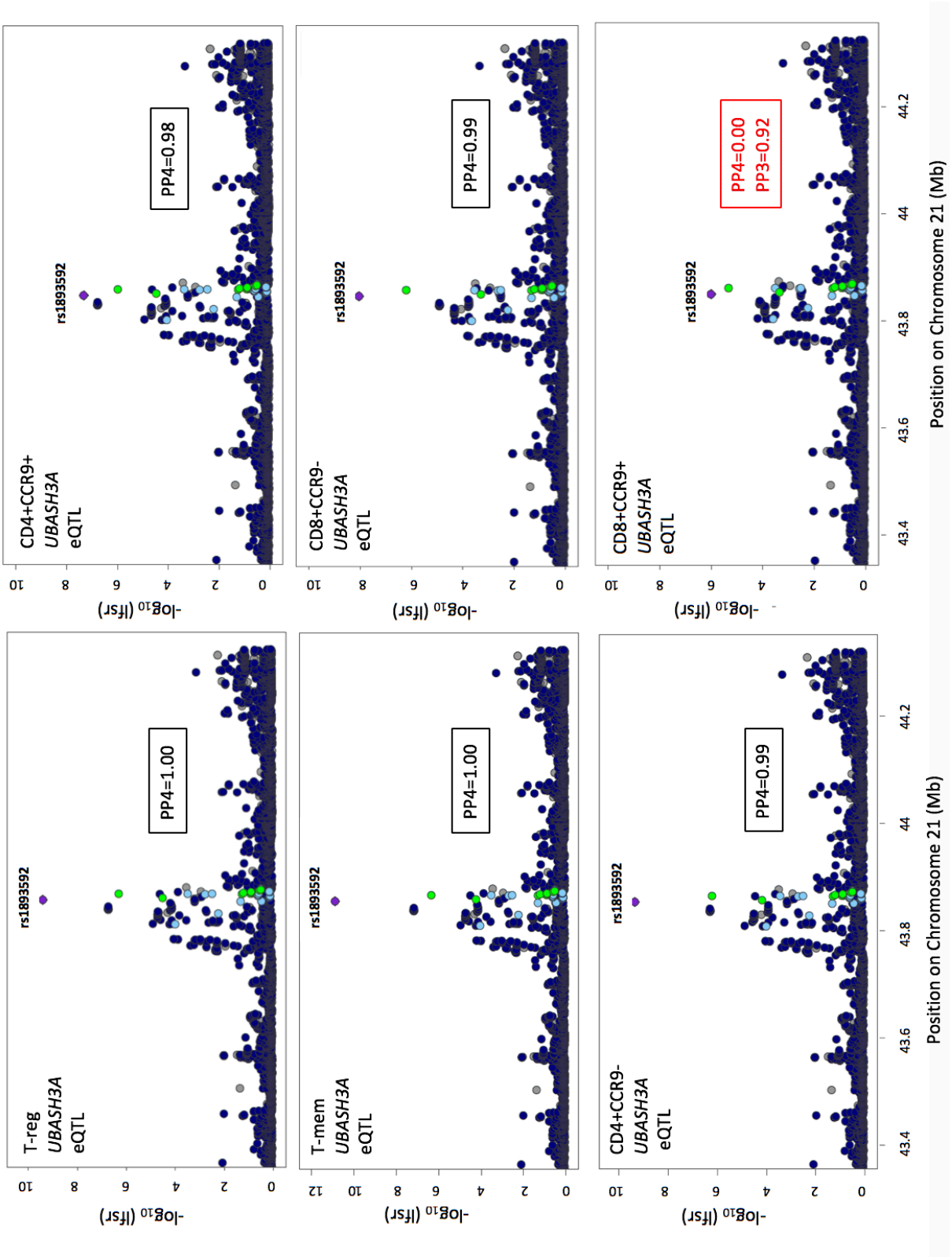


Figure 4.18: Regional association plots for colocalisation between PSC GWAS and eQTLs for *UBASH3A* in T-cells at Chromosome 21 rs1893592 risk locus, using *mashR* eQTL data.

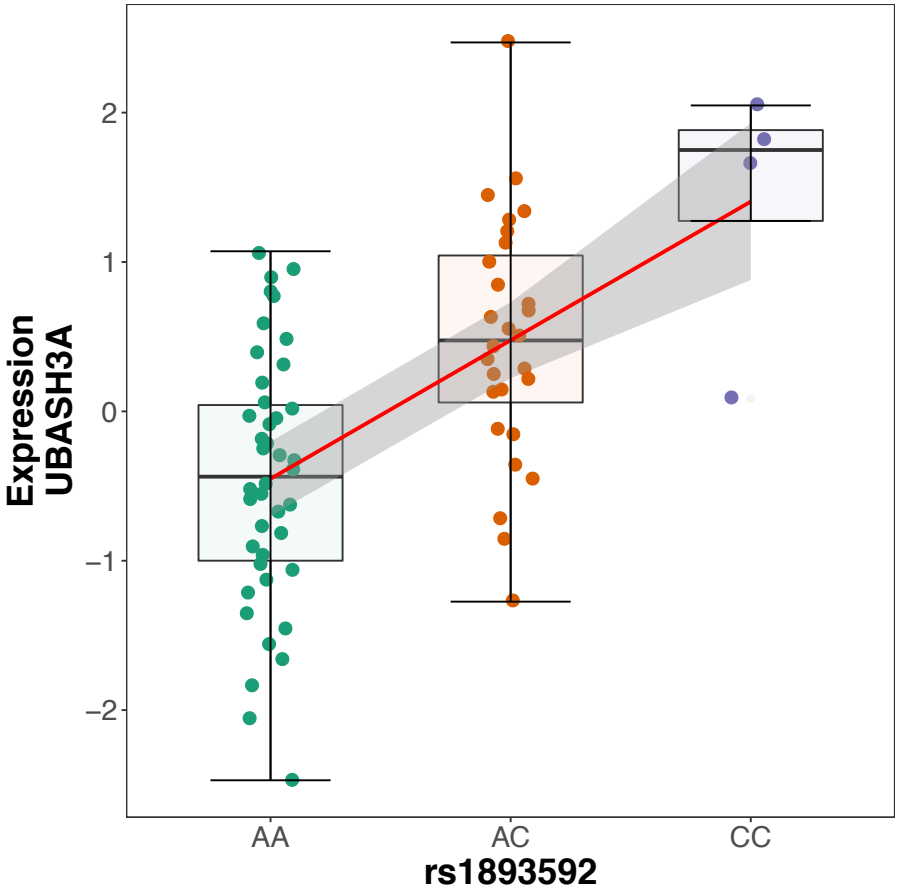


Figure 4.19: Expression of *UBASH3A* according to Chromosome 21 rs1893592 genotype in T-memory cells.

The second PSC risk locus which colocalised with an eQTL in one or more T-cell subsets was the Chromosome 11 rs663743 PSC risk locus. This locus colocalised with an eQTL for *AP003774.1* in three of the six T-cell subtypes; T-regs, T-mems and CD4+CCR9- T-cells with $\geq$ 95% PP (PP4) of causality (Figure 4.20). In addition, there was some evidence to support colocalisation of this locus with an eQTL for *AP003774.1* in CD8+CCR9- T-cells with PP4 of 0.70. Following *mashR* analysis, the strength of the association for this eQTL increased across all six cell types (Table 4.4). Subsequent colocalisation of this locus within the *mashR* eQTL data supported the finding that this PSC risk locus colocalised with an eQTL for *AP003774.1* in four of the six cell types including T-mem, T-reg, CD4+CCR9- and CD8+CCR9- T-cells (PP4 of $\geq$ 0.95) with some additional evidence (PP4=0.72) to support colocalisation with CD4+CCR9+ T-cells (4.21). Plotting of the *AP003774.1* eQTL at rs663743 confirmed that the PSC risk increasing rs663743*G allele reduced expression of *AP003774.1*, with a consistent direction of effect across all cell types (Figure 4.22).

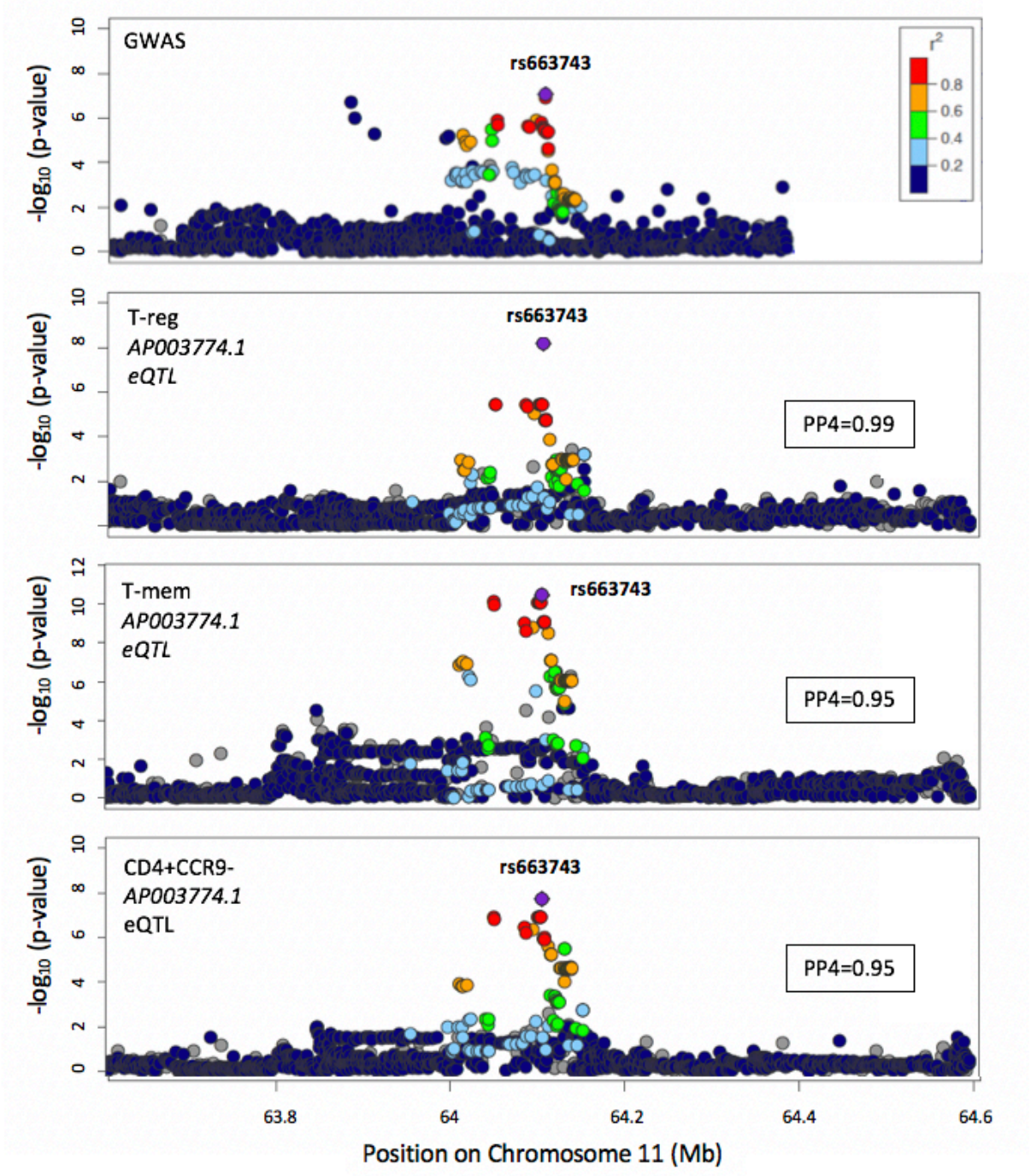


Figure 4.20: Colocalisation between PSC GWAS and *AP003774.1* eQTL data from the individual cell-type analysis, at the chromosome 11 rs663743 PSC risk locus.

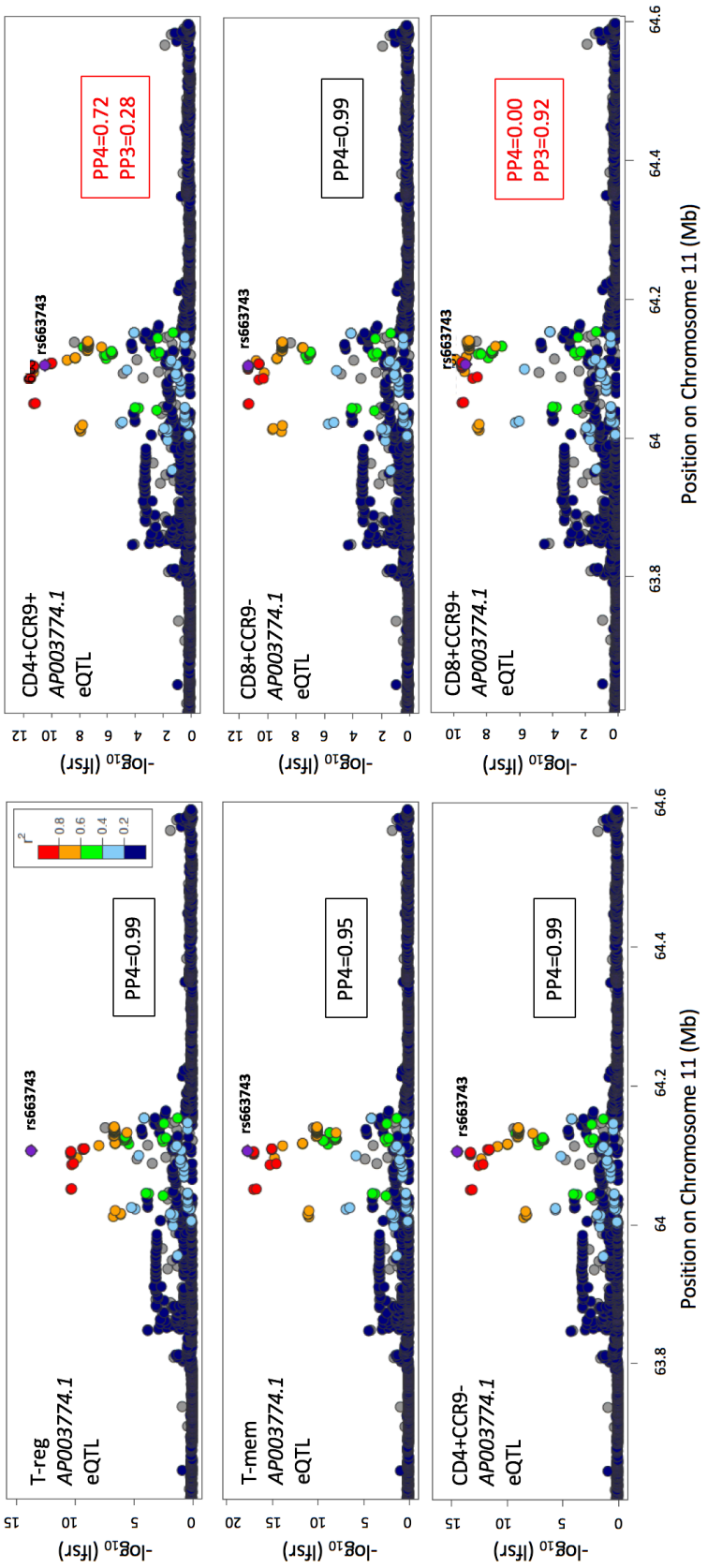


Figure 4.21: Colocalisation between PSC GWAS and*AP003774.1* eQTL data from the *mashR* analysis, at the chromosome 11 rs663743 PSC risk locus.

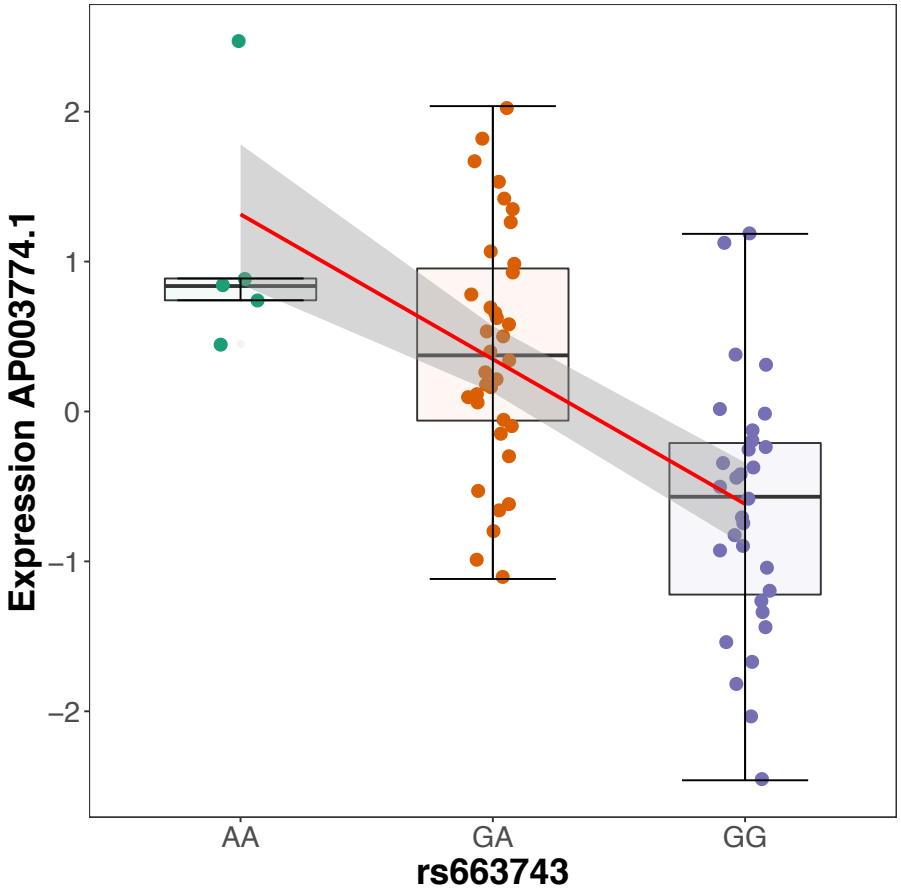


Figure 4.22: Expression of *AP003774.1* according to Chromosome 11 rs663743 genotype in T-regulatory cells.

AP003774.1 is a long non-coding RNA or lncRNA. LncRNA's are defined as transcripts with lengths exceeding 200 nucleotides that are not translated into protein. Whilst the function of the majority of lncRNAs are unknown, it has been shown that lncRNAs are themselves important regulators of gene expression, via interactions with transcription factors or epigenetic modifiers [252, 253]. LncRNAs thus provide a link between non-coding variants and protein-coding genes. Moreover, there is accumulating evidence that lncRNAs are important regulators of both immune cell differentiation and the innate and adaptive immune responses [254–256]. They have also been implicated in the pathogenesis of several IMDs, including (but not limited to) SLE, RhA, T1DM and MS [257–259]. Indeed, one study that mapped *cis*-eQTLs at 460 IMD-associated SNPs found that >10% affected the expression of a lncRNA [260]. Whilst little is known about *AP003774.1*, according to GTEx, this lncRNA is highly expressed in PSC-relevant tissues including colon, small intestine and whole blood (Figure 4.23) [176]. In addition, a search of the database for immune cell eQTL expression epigenomics (DICE) demonstrated that amongst immune cells, *AP003774.1* is most highly expressed in T-cells and NK cells, with lower expression in monocytes [261]. In Chapter 2, I demonstrated that this same region overlaps both

promoter and enhancer elements in multiple PSC-relevant tissues, suggesting plausible mechanisms via which this eQTL for *AP003774.1* may interact with epigenetic modifiers to regulate expression of other genes in the region. More specifically, this locus overlaps H3K27me3, a marker of an inactive or silenced regulatory region, in keeping with the PSC risk increasing allele reducing expression of *AP003774.1* (Figure 4.22). Interestingly, Ricano-Ponce *et al* demonstrated that expression of *AP003774.1* is also linked to another IMD, MS, where the lead GWAS SNP for the MS risk locus (rs694739 at Chr11:64097233, build 37) has been shown to decrease the expression level of *AP003774.1* in PBMCs [260]. Whilst this region has not been fine-mapped in MS, the MS lead SNP, rs694739, lies close to the fine-mapped SNP for this locus in PSC (rs663743 at Chr11:64107735) and both SNPs are in high LD with one another (r^2=0.74). In previous chapters I show that this same rs663743 risk locus in PSC colocalises with a monocyte eQTL for another gene, *CCDC88B*, which is not expressed in T-cells. It is therefore of particular note that Ricano-Ponce *et al* similarly observed that this same MS SNP also affected the expression of *CCDC88B* in PBMCs and that many SNPs associated with IMDs can affect the expression of more than one gene within a 500Kb region. It is therefore likely that this PSC risk locus functions as an eQTL for two different genes in two different cell types; *AP003774.1* in T-cells and *CCDC88B* in monocytes.

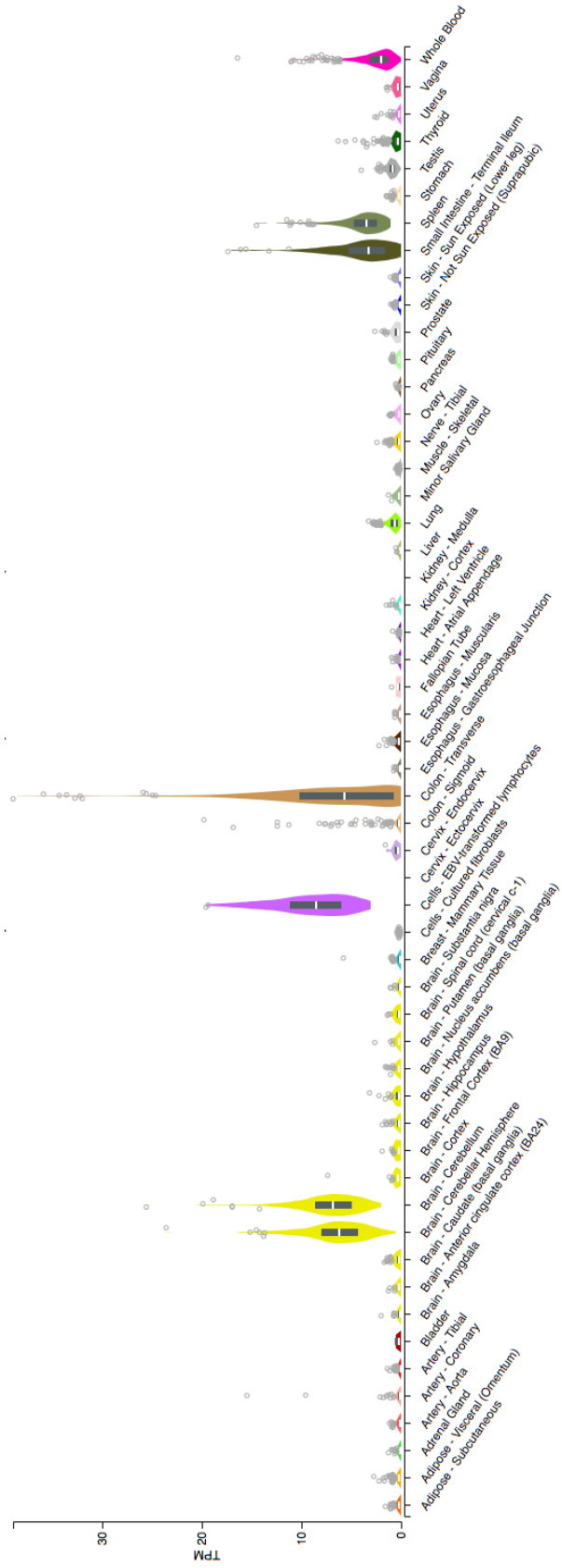


Figure 4.23: Expression of *AP003774.4* across multiple human tissues (figure generated by GTEx portal, 25/02/20 [176]).

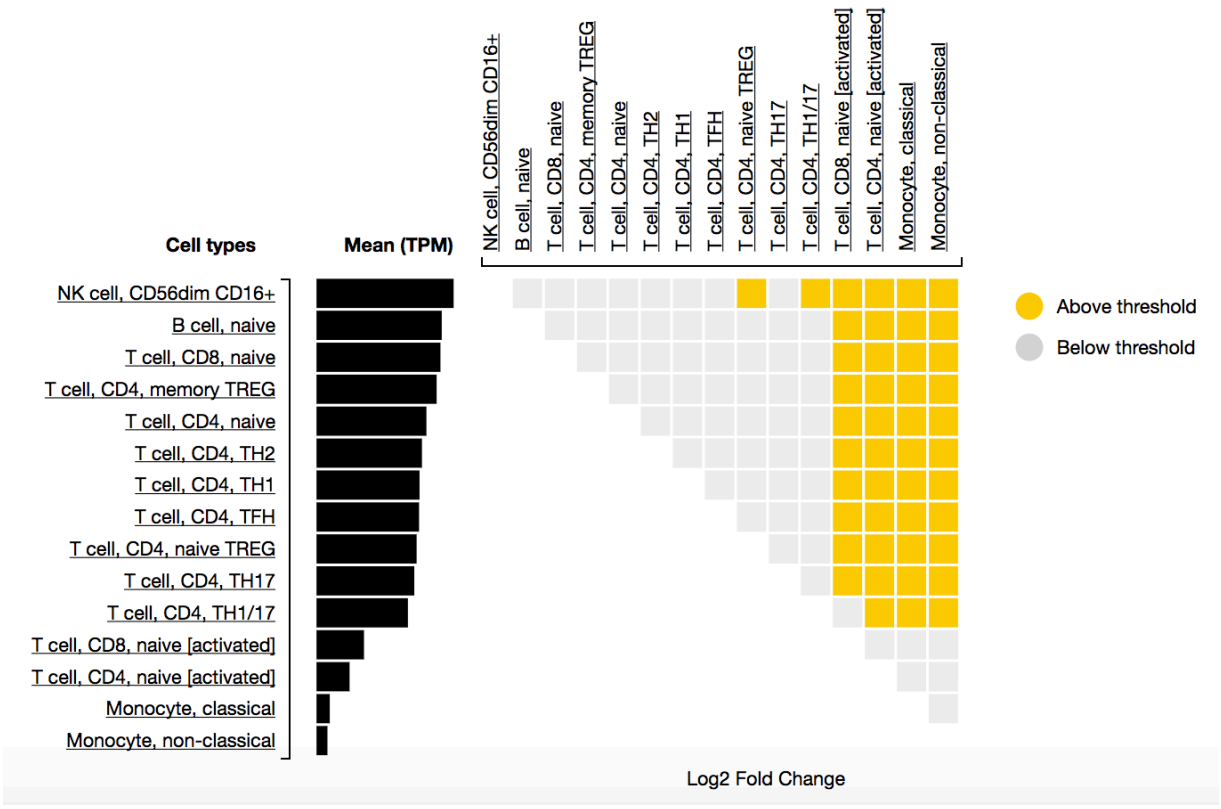


Figure 4.24: Expression of *AP003774.4* across multiple immune cell types (figure generated by the Database of immune cell eQTL expression [261], 26/02/2020).

Many genetic risk loci are known to be shared across multiple IMDs and similar eQTL studies have performed colocalisation of eQTLs with a range of IMDs. The majority of samples for this study were derived from patients with UC. In an effort to identify other IMD risk loci that function as eQTLs, I performed colocalisation of T-cell eQTLs with UC, CD and two other IMDs; RhA and T1DM. I identified ten IMD risk loci that colocalised with eQTLs for one or more genes. The results of colocalisation with all IMDs are shown in Table 4.5, however given that the focus of this thesis is PSC, only those IBD risk loci that colocalised with T-cell eQTLs are discussed further.

This analysis identified two UC risk loci and one CD risk locus that colocalised with T-cell eQTLs, thus identifying several genes involved in inflammatory or immune pathways with a potential causal role in IBD. Of note, the UC Chromosome 7 rs4728142 risk locus colocalised with an eQTL for *IRF5* in T-memory cells. *IRF5* is a transcription factor which forms one of the major inflammatory pathways, crucial for activation of the

Table 4.5: Colocalisation of non-HLA GWAS risk loci for immune-mediated diseases and T-cell eQTL

Chr	GWAS SNP	Disease	eGene	Cell type	PP4	eQTL Beta	eQTL p-val
1	rs3180018	UC	*GBAP1*	T-reg	0.91	-0.74	3.88E-04
				T-mem	0.98	-1.01	1.30E-10
				CD4+CCR9-	0.98	-0.92	2.01E-07
				CD4+CCR9+	0.98	-0.91	7.66E-07
		UC	*THBS3*	CD4+CCR9-	0.98	0.88	5.59E-07
1	rs2317230	RhA	*FCRL3*	CD8+CCR9-	0.93	0.82	3.29E-04
5	rs7731626	RhA	*IL6ST*	T-reg	0.97	-0.86	2.08E-04
				T-mem	0.90	-0.81	6.16E-04
		RhA	*ANKRD55*	T-mem	1.00	-1.00	5.94E-07
				CD4+CCR9-	1.00	-0.95	9.63E-06
				CD4+CCR9+	0.86	-0.86	1.20E-03
7	rs4728142	UC	*IRF5*	T-mem	0.86	0.77	4.71E-05
11	rs663743	PSC	*AP003774.1*	T-reg	0.99	0.98	7.27E-06
				T-mem	0.95	1.07	1.35E-07
				CD4+CCR9-	0.95	0.96	1.83E-05
11	rs968567	RhA	*FADS1*	T-reg	0.89	1.51	1.38E-07
		RhA	*FADS2*	T-reg	0.98	1.60	2.01E-09
				T-mem	1.00	1.58	2.23E-09
				CD4+CCR9-	0.99	1.58	5.40E-09
				CD4+CCR9+	0.98	1.47	2.82E-07
				CD8+CCR9-	0.96	1.56	9.58E-09
				CD8+CCR9+	0.95	1.46	4.22E-07
12	rs4760341	T1DM	*SUOX*	T-reg	0.80	-0.71	1.06E-03
14	rs941576	T1DM	*WARS*	T-reg	0.85	1.07	3.90E-05
				T-mem	0.93	1.19	1.46E-07
				CD4+CCR9-	0.80	1.35	2.41E-08
				CD8+CCR9-	0.97	-1.03	1.37E-05
19	rs4802307	CD	*PPP5C*	T-mem	0.85	-0.92	1.63E-06
				CD4+CCR9-	0.86	-0.99	2.38E-08
				CD8+CCR9-	0.83	-0.82	1.48E-04
21	rs1893592	PSC	*UBASH3A*	T-mem	0.91	0.93	4.83E-04
22	rs909685	RhA	*SYNGR1*	CD8+CCR9+	0.98	1.14	7.15E-04

UC; Ulcerative colitis, CD; Crohn's Disease, RhA; Rheumatoid arthritis, T1DM; Type 1 diabetes mellitus

pro-inflammatory cytokines IL-6, IL-12 and TNF-α [262, 263]. Its expression is induced in lymphocytes by activation of the Toll-like receptor (TLR) 7 and 9 pathways and polymorphisms within this gene have been associated with SLE, RhA, MS, Sjogren's syndrome, psoriasis and IBD [56, 264]. Although there are no existing drugs targeting this gene, it is widely considered to be a promising future target [265]. Furthermore, the Chromosome 1 rs3180018 UC risk locus colocalised with an eQTL for *GBAP1* in T-reg, T-memory, CD4+CCR9- and CD4+CCR9+ T-cells (PP4≥ 0.91). Interestingly this differs from the previously reported candidate genes for this UC locus, *SCAMP3* and *MUC1*. However a causal role for *GBAP1* in IBD has been further supported by the fact that this same variant has been shown to increase expression of *GBAP1* in a peripheral blood eQTL study of patients with CD, resistant to anti-TNF treatment [266]. *GBAP1* is an expressed pseudogene which is known to regulate *GBA* levels, a gene encoding lysosomal glucocerebrosidase and the major predisposing gene involved in Parkinson's disease (PD) pathogenesis. It functions as a competing-endogenous RNA (ceRNA), acting as a microRNA (miRNA) sponge, resulting in subsequent *GBA* degradation [267]. To date, there have been no studies investigating the potential role of *GBAP1* in relation to IBD pathogenesis, however given the emergence of therapies modulating glucocerebrosidase activity in PD, further investigation of this pathway outside of the central nervous system and in the context of UC pathogenesis may be warranted [268].

4.5 Discussion

In this study, I develop the first eQTL maps of peripheral blood T-cell subsets in patients with PSC. Using recently published methods to estimate patterns of similarity across cell-types and thus improve estimates of effect, I was able to identify >10,000 unique eQTLs in at least one or more of the six T-cell subsets. Furthermore, by performing colocalisation of disease risk loci with eQTLs in PSC-specific T-cell subsets, I was able to identify the genes perturbed by two PSC risk loci, in addition to three IBD, four RhA and two T1DM risk loci.

An important finding from this work is the identification of a lncRNA with a potentially important role in PSC causal pathogenesis. The Chromosome 11 rs663743 PSC risk locus functions as an eQTL of *AP003774.1*, which is highly expressed in PSC-relevant tissues including colon, small intestine and whole blood, as well as T-cells and NK-cells. Indeed, expression of this lncRNA has also been linked to MS, where an MS risk locus in this region has also been shown to decrease the expression level of *AP003774.1* in PBMCs [260]. Further work to fine-map the causal variant for this signal in MS is needed to establish if the same causal variant is responsible for the effects seen in PSC and MS. Nevertheless, these findings suggest that *AP003774.1* may have an important role in the

immune-regulatory pathways of T-cells and further study is warranted to establish how reduced expression of this lncRNA might potentiate increased risk of IMD. One means of identifying other genes within the same biological pathway, affected by this risk locus would be to map *trans*-eQTL in the same cell types. Due to their smaller effect sizes and the large numbers of tests required with all genes across the genome, *trans*-eQTL mapping requires much larger sample sizes than available in this study (although this is less of an issue with a targeted trans-eQTL study). However, the finding of more distant genes affected by this same risk variant may identify the relevant biological pathway for further functional investigation.

The findings of this study confirm *UBASH3A* as an important gene in the causal pathogenesis of PSC, a finding that appears, from the analyses outlined in this thesis, to be specific to T-cells. The PSC risk increasing variant results in a reduction of *UBASH3A* expression at the Chromosome 21 rs1893592 PSC risk locus in T-cells. This same risk locus has been associated with several other IMDs including T1DM, CeD and RhA [169, 216]. Furthermore, an RNA sequencing study has demonstrated reduced expression of *UBASH3A* in the PBMCs of patients with SLE [269]. UBASH3A functions to attenuate the signal transduction of NF-κB upon TCR stimulation, by suppressing the activation of the I-κB kinase complex, lending biological plausibility to its role in IMD pathogenesis [216]. There are, to date, no known drugs targeting the *UBASH3A* gene, however there are several therapeutic options for targeting the NF-κB/I-KKβ pathway. Proteasome inhibitors, such as bortezomib and carfilzomib are known modulators of targets in the NF-κB/I-KKβ pathway. In addition, the widely available drug, acetylsalicylic acid or Aspirin, is an inhibitor of IKKβ [270]. Notably, whilst there have been no randomised controlled trials of Aspirin use in PSC, there are case-control data to support a chemoprotective role for Aspirin in the development of de-novo cholangiocarcinoma, which is one of the serious complications of PSC [271, 272]. Further study of the potential therapeutic effects of Aspirin and other modifiers of the NF-κB/I-KKβ pathway in PSC are therefore warranted.

One of the most important limitations of this, and indeed many eQTL studies, is the sample size. This study included ∼450 samples from ∼75 individuals, which was at the lower limit to powerfully detect a significant number of eQTLs. Using DGE, I demonstrated transcriptional equivalence between T-cell subsets in the PSC-UC and lone UC groups, supporting the amalgamation of disease groups to improve subsequent power to detect eQTLs. One important analysis, which was not possible due to the small sample size in each individual disease group, would be to examine the effects of disease-specific eQTLs. For example, identifying those eQTLs which are active in PSC-UC, but not UC may point to important causal biological pathways for PSC. Despite sample size limitations, the use of stringent quality control measures enabled me to robustly identify a total of ∼3,000 unique eGenes across all T-cell subtypes from the individual cell-type analysis, increasing

this number to ∼10,000 with *mashR*. For those PSC loci which colocalised with an eQTL in one or more T-cell subtype, analysis of the *mashR* eQTL data enabled me to identify PSC risk loci that colocalisaed with the same eQTL across multiple T-cell subtypes. Of the ∼10,000 eQTLs identified in this study, >85% were shared across all six T-cell subtypes. The finding that the majority of eQTLs in this study were shared across all six T-cell subtypes is likely to be explained by the relative similarity between the T-cell phenotypes studied in this analysis; all six cellular subtypes were CD3+ T-cells and four were CD4+. During the design of this study, it was hypothesised that the acquisition and analysis of six different T-cell subsets from each donor would allow the detection of cell-type specific eQTLs. However, the resultant benefit of multiple T-cell subtypes from each donor was, in fact, to enable the estimation of patterns of similarity across conditions or cell-types using *mashR*, to improve accuracy of effect estimates and thus identify greater numbers of eQTLs. In a rare diseases such as PSC, where patient recruitment for sample donation is limited by the number of sample donors, this may be an useful future mechanism to improve eQTL mapping. The vast majority of samples in this study were from patients with active PSC and UC, or UC alone. Whilst it is likely that mapping of eQTLs in cell-types that have been subject to the active disease state may have uncovered some additional eQTLs not active in the healthy state, this may only be evident in studies that are well-powered enough to detect those effects. Although this study focused on deriving samples from PSC and UC patients, I also identified eQTLs that colocalised with RhA and T1DM, suggesting that a more fruitful approach might be to study large cohorts of individuals with RNA-seq data, whether or not they have the disease phenotype.

An important future analysis of this T-cell eQTL data would be to conduct fine-mapping of those colocalising risk loci within the eQTL data. Whereas the previous fine-mapping analysis in Chapter 3 had resolved the Chromosome 21 rs1893592 PSC risk locus to this single causal variant, the Chromosome 11 PSC risk locus was fine-mapped to two potential causal variants. Given that the strengths of association between rs663743 on Chromosome 11 and *AP003774.1* expression are greater than with PSC risk, there is likely to be greater power to fine-map the eQTL data and thus attribute a greater PP of causality to a single causal variant. This would pave the way for future biological studies to analyse the impact of the true causal variant perhaps through CRISPR analysis, or recall by genotype experiments.

The rigorous analysis outlined in this chapter has resulted in the generation of a robust set of eQTL maps for six T-cell subtypes, several of which have not previously been the subject of eQTL mapping efforts, and none of which have been previously mapped in patients with PSC. As demonstrated by the finding of eQTLs that colocalise with other IMD risk loci, the results of these analyses can be relevant and important to variety of IMDs outside of PSC and UC. These eQTL maps, which have revealed important findings

for our understanding of PSC, will also provide a public resource available for further scientific study.

Chapter 5

Conclusions

Our DNA, laid down at conception, gives us an unrivalled opportunity to understand 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 distinguish cause from consequence of disease. The genetic risk loci associated with risk of PSC provide an unrivalled opportunity to further understand the causal biology of this disease, if only we can robustly identify the true causal variants driving these loci and the genes they perturb.

In this thesis, I outline the first study aimed at identifying the true causal variants driving PSC risk loci and the genes they perturb, in an effort to further understand disease biology and identify drug targets. Prior to this study, 23 loci had been associated with PSC risk, the majority of which are in non-coding regions. Using statistical fine-mapping and colocalisation with eQTLs mapped in multiple immune-cells, including self-generated PSC T-cell eQTL maps, I have identified seven downstream genes (*FOXP1, SH2B3, AP003774.1, CCDC88B, PRKD2, ETS2* and *UBASH3A*) affected by six PSC risk loci. Furthermore, I have fine-mapped 15 PSC risk loci to credible sets of causal variants driving each locus. The work outlined in this thesis identifies several genes not previously connected to the causal pathogenesis of this disease, including *ETS2* and *AP003774.1*, as well as identifying several genes (*ETS2, PRKD2* and *UBASH3A*) which warrant further investigation as potential therapeutic targets. Importantly, whilst the work conducted in thesis was not designed to further investigate any of the main hypotheses of PSC pathogenesis, transcriptome analysis and eQTL mapping in CCR9+ effector-memory T-cells did not confirm or refute a specific pathogenic role for these cells in support of the 'gut-homing T-cell' hypothesis.

The overlap in both genetic and immune characteristics of many IMDs and previous success in re-purposing drugs between different IMDs means that for a rare disease such as PSC, the most efficient means of finding a drug that may attenuate disease risk or progression, is through the re-purposing of existing drugs. The results presented in this

thesis identify three genes involved in pathways which are currently the target of existing therapeutic agents or ongoing exploratory studies to develop therapeutic agents. The first of these genes is *UBASH3A*. This study confirms a causal role for *UBASH3A* in PSC risk. My results consistently demonstrate that the fine-mapped Chromosome 21 rs1893592 PSC risk increasing allele acts as an eQTL for reducing *UBASH3A* expression across almost all T-cell sub-types tested in this study, but not in the wide variety of other immune cells analysed. Whilst one criticism of colocalisation analysis across multiple cell types is the finding of multiple eGenes within each locus, consistency of both gene and cell-type, as in this case, increases our confidence that we have identified the true gene or pathway affected by a risk locus. Whilst there are no existing drugs targeting *UBASH3A*, this gene has an important role in the attenuation of the NF-$\varkappa$B/I-KKβ pathway. Proteosome inhibitors (PIs) are an existing group of drugs that target the NF-$\varkappa$B/I-KKβ pathway, and are currently used for the treatment of multiple myeloma and graft-versus-host disease. PIs not only inhibit the activation of NF-$\varkappa$B and release of other pro-inflammatory cytokines, but also induce apoptosis of activated immune cells. Circulating proteosomes have been found in the serum of patients with several IMDs including SLE, RhA, systemic sclerosis and AIH [273, 274] and elevated levels of immunoproteasome are associated with disease progression [275]. It has been hypothesised that these raised levels of circulating proteasomes in IMDs function as auto-antigens [276], with anti-proteasome autoantibodies detected in the serum of patients with RhA, SLE and MS [277, 278]. The immunusuppressive properties of PIs in T-cell-mediated immune responses have been explored to some extent. PI's bortezomib, epoxomycin and lactacystin suppress the activation, proliferation, survival and immune functions of T-helper (Th) cells [279]. In RhA patients, bortezomib, has been shown to inhibit the release of NF-$\varkappa$B-inducible cytokines by activated T-cells [280]. The use of PIs in the treatment of other IMDs such as PSC is a potential avenue for future investigation. Whilst most current experimental evidence has been conducted in RhA, the availability of good first, second and third line immmunsuppressive treatments for RhA means that further investigation of PIs in this disease is of not of great clinical necessity. Furthermore, PIs produce a number of toxic side effects, including (but not limited to) peripheral neuropathy, thrombocytopenia, diarrhoea and an increased risk of infection. Whilst, such a side effects profile may be acceptable for those with a malignant condition such as multiple myeloma, it is perhaps less acceptable for patients living with some chronic IMDs. However, in PSC, a disease with no current therapeutic options and high risk of serious disease complications, the further exploration of PIs as a therapeutic agent is supported by evidence from this study.

The second potential gene for consideration as a therapeutic target is *PRKD2*. The results of this study confirm that the Chromosome 19 PSC non-coding risk locus is an eQTL of *PRKD2*. The fine-mapped PSC risk increasing allele of this locus, reduces expression

of *PRKD2* in monocytes and colonic tissue. It has been recently shown that *PRKD2* has an important role in controlling transition from naïve CD4+ T cells to T-follicular helper (TFH) cells in response to antigen or vaccine stimulus [281]. This is achieved by the direct binding and phosphorylation of Bcl6 by Prkd2, constraining Bcl6 to the cytoplasm, thereby limiting TFH development. Misawa *et al* demonstrated that a *PRKD2* loss of function mutation which results in reduced expression of the Prkd2 protein in mice, allows unrestricted Bcl6 nuclear translocation in Prkd2$^{(-/-)}$ CD4+ T cells. This results in excessive cell-autonomous TFH development and B-cell activation in Prkd2$^{(-/-)}$ spleens and polyclonal hypergammaglobulinemia of IgE, IgG1 and IgA isotypes. This is particularly interesting given that TFH imbalance can contribute to IMD and IgE is often raised in the presence of IMD. Whilst my T-cell study did not find any evidence that this locus was an eQTL of *PRKD2* in T-cell subsets, TFHs were not included within this analysis. Certainly *PRKD2* has an important regulatory role in TFH development and further work examining the therapeutic effects of increasing the kinase activity of Prkd2 in CD4+ T cells as well as monocytes, is warranted, not only for PSC, but also for T1DM for which this is a shared risk locus.

A third potential drug target from this study is the *ETS2* gene. I demonstrate that the fine-mapped Chromosome 21 rs2836883 risk locus is an eQTL for *ETS2*, with the PSC risk increasing allele resulting in increased expression of *ETS2* in monocytes and macrophages. *ETS2* has been found to be up-regulated in a number of cancers, including renal cell carcinoma, prostate cancer and more notably colorectal adenocarcinoma and hepatocellular carcinoma [282–284]. *ETS2* is a transcription factor with an important role in the Ras/Raf/MEK/ERK cascade. It activates the *BCL-2* promoter, which is one of various apoptosis regulating factors that are phosphorylated by the Ras/Raf/MEK/ERK cascade, subsequently inhibiting cellular apoptosis [284]. For this reason, there has been recent interest in *ETS2* inhibitors as a potential means of interrupting the Ras/Raf/MEK/ERK pathway and thus a potential anti-cancer therapy [285]. In PSC, *ETS2* may contribute to several aspects of disease pathogenesis, including the induction of pro-inflammatory cytokine release from macrophages, in addition to IL-2 regulation in the transition of naive Th to Th0 cells upon antigenic stimulation. Furthermore, the role of *ETS2* in the development of inflammation-induced dysplasia is yet to be explored. Therefore, whilst work on ETS2 inhibitors is in its very early stages, further research is warranted to explore mechanisms of ETS2 inhibition and its potential for clinical application in PSC.

PSC is a rare complex disease, which provides many challenges for scientific study. Ultimately the mapping of more eQTL across more cell types and activation states alongside the expansion of GWAS sample sizes and numbers of disease rise loci, holds the key to further understanding the causal pathogenesis of this debilitating disease and the identification of biological pathways for therapeutic target. Common complex diseases

such as IBD, RhA and T1DM have benefited enormously from the genetics revolution. For these diseases, GWAS sample sizes now reaching the tens to hundreds of thousands have lead to the discovery of increasingly large numbers of genetic risk loci. These diseases stand to benefit further from the creation of giant biobanks and consortia, where GWAS can be conducted on an unprecedented scale. For example, the UK Biobank (UKBB) is a health resource that includes clinical phenotype data, multiple biological samples and genotype data from ∼500,000 individuals [286]. For a common complex disease such as IBD, estimated to affect 0.78% of the UK population [287], the UKBB currently includes an additional 6,370 IBD patients whose data can contribute to GWAS meta-analyses. For an even more common disease such as asthma, the UKBB contributes tens of thousands of cases. However, for a rare disease such as PSC with a prevalence of just 1/10,000, the numbers of cases included within the UKBB will be too small to benefit PSC research, especially given the selection bias of the UKBB towards more healthy individuals. Disease-specific initiatives such as the NIHR IBD Bioresource, which has collated biological, genetic and clinical phenotype data for ∼25,000 patients with IBD across the UK, provides a potential resource for further large scale genetic studies in PSC [288]. However disappointingly, whilst the prevalence of PSC in IBD patients predicts that up to ∼1,700 of the 25,000 IBD recruits might have concomitant PSC, current clinical phenotype data identifies only ∼300 PSC cases in the IBD Bioresource. This only serves to highlight the importance of accurate and complete phenotype data in genetic studies of complex disease. The future of PSC research therefore requires ongoing efforts from national and international PSC consortia, such as the UK-PSC consortium and the international PSC Study Group (iPSCSG), to create a large biobank of biological samples, genotype and clinical phenotype data from patients with PSC. Such a biobank could be based upon the NIHR IBD Bioresource model, or indeed the UK-PSC component embedded directly within it. An important question regarding the focus of future genetic studies using such consortia will be whether to concentrate on GWAS, whole-exome or whole-genome sequencing. Whole-exome sequencing (WES) requires the sequencing of just 2% of the genome at greater depth which provides more confidence in calling genotypes at lower frequency SNPs compared to GWAS. Moreover, one captures many more variants in the gene than one could ever capture and impute from a GWAS array. WES also captures rare variants which have fewer LD friends than common variants, and are not so well captured by GWAS. In addition to the above, whole genome sequencing (WGS), allows the interrogation of the many non-coding variants associated with disease risk, in addition to providing more complete coverage of exons than WES [289]. Whilst the most desirable focus for PSC would be on WGS large numbers of PSC samples, WES is hugely advantageous in terms of sequencing time, cost and storage, enabling the analysis of greater numbers of samples where resources are limited.

One group of methods closely related to GWAS–eQTL colocalisation studies, are the transcriptome wide association studies (TWAS), which directly integrate GWAS and gene expression data to identify gene–phenotype associations and prioritise causal genes at GWAS loci. Existing TWAS methods allow the use of individual-level GWAS data [290], or summary-level GWAS data [291, 292]. Firstly, using the gene expression data, a TWAS uses allele counts of genetic variants within 500-1,000Kb of a gene, to learn per-gene predictive models of variation in gene expression. Secondly, this model is then taken forward to predict gene expression for each individual within the GWAS cohort and finally the association between predicted gene expression and the phenotype, is estimated [293]. Thus, TWAS does not test for association with total expression, but rather genotype-predicted expression. However, analogous to the groups of high-LD variants found to be associated with a disease trait in a GWAS, TWAS frequently identify multiple genes per locus, which can be a result of correlated gene expression within a locus [293, 294]. Similar to fine-mapping in GWAS to identify causal variants, methods of fine-mapping causal gene sets have been developed, which model predicted expression correlations in order to assign posterior probabilities of causality to each gene [295]. Due to the variation in gene expression across different cell types, TWAS is however susceptible to the identification of spurious associations with expression data from tissues or cell types that are not mechanistically related to the phenotype. Recent TWAS best practice recommendations therefore suggest the use of only expression data from mechanistically related tissues, even if this results in a smaller sample size. Importantly, as shown in this thesis, it is not always clear which cell types may be the most mechanistically relevant. Finucane *et al* have established one potential method to address this issue, involving stratified LD score regression to test for enrichment of disease heritability in the genomic regions surrounding genes with the highest specific expression in a given tissue [296]. This method could be used in future studies to identify the cell-types most relevant to PSC. As identified in Chapter 4 of this thesis, there is often a lack of publicly available gene expression datasets. One solution to this is to use a similar method to that developed by Barbeira *et al* involving the aggregation of data across all available tissues in a 'tissue-agnostic manner' which can be applied to either individual level or summary-level data [297]. Furthermore, a potentially fruitful future fine-mapping analysis for this study would be to similarly combine data-sets for all traits (both gene expression and genotype) at colocalising risk loci using a model that allows for mixed effect sizes, to perform a fine-mapping meta-analysis for the purposes of better-defining the credible causal variants at each PSC risk locus. Indeed, a similar approach has been used by Westra *et al* in their approach to fine-mapping disease risk loci in RhA and T1DM [114]. Although they did not use colocalisation to prove sharing of disease risk loci in RhA and T1DM, they harnessed the fact that the genetic architecture is somewhat shared between IMDs, combining summary statistics from both diseases to

increase their power to fine-map RhA and T1DM risk loci.

Advances in single-cell RNA sequencing (scRNA-seq) are likely to create several advantages for the study of complex diseases such as PSC. As previously discussed, an important next step in the linkage of disease risk-SNPs to downstream effects on gene expression is to define the cell-types in which disease risk-SNPs affect gene expression levels. The future of RNA sequencing analysis is rapidly moving towards scRNA-seq, which unlike bulk RNA-seq, requires no prior definition of cell types. Furthermore, it allows the analysis of many more cell types with a hypothesis-free approach, potentially limited only by the cellular composition of the input tissue. Using scRNA-seq one can estimate both the cellular composition of the input tissue and the gene expression for discrete cell populations [298]. Furthermore, cell populations are not just limited to discrete populations, but can also be defined along a dynamic continuum and are thus more likely to reflect the dynamics of true human immune cell biology [299]. Several studies have already performed the mapping of eQTLs at the single cell level [298–301]. There are several challenges to eQTL mapping in scRNA-seq data, including the identification and subsequent classification of cells into types or states and the normalisation of gene expression data to account for differences in sequencing depth. The single-cell eQTLGen consortium (sc-eQTLGen), is a large-scale, international collaborative initiative that has been set up to *'identify the upstream interactors and downstream consequences of disease-related genetic variants'* in individual immune cell-types [302]. As part of this effort the sc-eQTLGen aims to address many of the outstanding issues with scRNA-seq data generation and analysis, and to identify new standards for best practice. Whilst sc-eQTL mapping studies are in their infancy, the future application of scRNA-seq and sc-eQTL mapping studies in PSC provides an exciting avenue for the future study of downstream genes affected by PSC risk loci.

PSC is a debilitating disease with serious disease sequelae, for which new therapeutic options are urgently needed. Genetics provides an unrivalled opportunity to improve our mechanistic understanding of the causal pathogenesis and thus identify genes and pathways for potential therapeutic target. In this thesis, I have used genetics to elucidate multiple genes with a causal role in the pathogenesis of this disease, several of which are potential candidates for therapeutic target. Via a combination of experimental and statistical genetic approaches I have addressed and overcome many of the scientific challenges of studying such a rare complex disease. The future of PSC genetic research will continue to benefit enormously from the ongoing advances in computational and experimental research approaches. Alongside rapid developments in disease specific biobanks guaranteeing improved disease sampling as well as technological advances at the single cell level, we will continue to unfurl the complex genetic basis of this disease and move ever closer to a cure for PSC.

Bibliography

[1] U. Broome et al. "Natural history and prognostic factors in 305 Swedish patients with primary sclerosing cholangitis". In: *Gut* 38.4 (1996), pp. 610–5. ISSN: 0017-5749 (Print) 0017-5749 (Linking). URL: http://www.ncbi.nlm.nih.gov/pubmed/8707097.

[2] Elizabeth C Goode et al. "Factors associated with outcomes of patients with primary sclerosing cholangitis and development and validation of a risk scoring system". In: *Hepatology* 69.5 (2019), pp. 2120–2135.

[3] Edward Alabraba et al. "A re-evaluation of the risk factors for the recurrence of primary sclerosing cholangitis in liver allografts". In: *Liver Transplantation* 15.3 (2009), pp. 330–340.

[4] A. Bergquist et al. "Hepatic and extrahepatic malignancies in primary sclerosing cholangitis". In: *J Hepatol* 36.3 (2002), pp. 321–7. ISSN: 0168-8278 (Print) 0168-8278 (Linking). URL: http://www.ncbi.nlm.nih.gov/pubmed/11867174.

[5] M. M. Claessen et al. "High lifetime risk of cancer in primary sclerosing cholangitis". In: *J Hepatol* 50.1 (2009), pp. 158–64. ISSN: 0168-8278 (Print) 0168-8278 (Linking). DOI: 10.1016/j.jhep.2008.08.013. URL: http://www.ncbi.nlm.nih.gov/pubmed/19012991.

[6] Martin Blachier et al. "The burden of liver disease in Europe: a review of available epidemiological data". In: *Journal of hepatology* 58.3 (2013), pp. 593–608.

[7] S Charman, L Copley, C Tovikkai, et al. *UK liver transplant audit (NHS Blood and Transplant). 2012*. 2015.

[8] H. Sano et al. "Clinical characteristics of inflammatory bowel disease associated with primary sclerosing cholangitis". In: *J Hepatobiliary Pancreat Sci* 18.2 (2011), pp. 154–61. ISSN: 1868-6982 (Electronic). DOI: 10.1007/s00534-010-0319-8. URL: http://www.ncbi.nlm.nih.gov/pubmed/20740366.

[9] B. D. Ye et al. “Clinical characteristics of ulcerative colitis associated with primary sclerosing cholangitis in Korea”. In: *Inflamm Bowel Dis* 17.9 (2011), pp. 1901–6. ISSN: 1536-4844 (Electronic) 1078-0998 (Linking). DOI: 10.1002/ibd.21569. URL: http://www.ncbi.nlm.nih.gov/pubmed/21830268.

[10] K. Boonstra et al. “Primary sclerosing cholangitis is associated with a distinct phenotype of inflammatory bowel disease”. In: *Inflamm Bowel Dis* 18.12 (2012), pp. 2270–6. ISSN: 1536-4844 (Electronic) 1078-0998 (Linking). DOI: 10.1002/ibd.22938. URL: http://www.ncbi.nlm.nih.gov/pubmed/22407885.

[11] Jr. Loftus E. V. et al. “PSC-IBD: a unique form of inflammatory bowel disease associated with primary sclerosing cholangitis”. In: *Gut* 54.1 (2005), pp. 91–6. ISSN: 0017-5749 (Print) 0017-5749 (Linking). DOI: 10.1136/gut.2004.046615. URL: http://www.ncbi.nlm.nih.gov/pubmed/15591511.

[12] T. J. Weismuller et al. “Patient Age, Sex, and Inflammatory Bowel Disease Phenotype Associate With Course of Primary Sclerosing Cholangitis”. In: *Gastroenterology* (2017). ISSN: 1528-0012 (Electronic) 0016-5085 (Linking). DOI: 10.1053/j.gastro.2017.02.038. URL: https://www.ncbi.nlm.nih.gov/pubmed/28274849.

[13] E Björnsson et al. “Patients with small duct primary sclerosing cholangitis have a favourable long term prognosis”. In: *Gut* 51.5 (2002), pp. 731–735.

[14] Einar Björnsson et al. “Primary sclerosing cholangitis associated with elevated immunoglobulinG4: clinical characteristics and response to therapy”. In: *American journal of therapeutics* 18.3 (2011), pp. 198–205.

[15] U. Broome et al. “Primary sclerosing cholangitis and ulcerative colitis: evidence for increased neoplastic potential”. In: *Hepatology* 22.5 (1995), pp. 1404–8. ISSN: 0270-9139 (Print) 0270-9139 (Linking). URL: http://www.ncbi.nlm.nih.gov/pubmed/7590655.

[16] U. Beuers. “Drug insight: Mechanisms and sites of action of ursodeoxycholic acid in cholestasis”. In: *Nat Clin Pract Gastroenterol Hepatol* 3.6 (2006), pp. 318–28. ISSN: 1743-4378 (Print) 1743-4378 (Linking). DOI: 10.1038/ncpgasthep0521. URL: http://www.ncbi.nlm.nih.gov/pubmed/16741551.

[17] Elizabeth C Goode and Simon M Rushbrook. “A review of the medical treatment of primary sclerosing cholangitis in the 21st century”. In: *Therapeutic advances in chronic disease* 7.1 (2016), pp. 68–85.

[18] William G Hill, Michael E Goddard, and Peter M Visscher. “Data and theory point to mainly additive genetic variance for complex traits”. In: *PLoS genetics* 4.2 (2008), e1000008.

[19] EA Stahl et al. “Diabetes Genetics Replication and Meta-analysis Consortium; Myocardial Infarction Genetics Consortium (2012). Bayesian inference analyses of the polygenic architecture of rheumatoid arthritis”. In: *Nat. Genet* 44 (), pp. 483–489.

[20] Björn Lindkvist et al. “Incidence and prevalence of primary sclerosing cholangitis in a defined adult population in Sweden”. In: *Hepatology* 52.2 (2010), pp. 571–577.

[21] RAKESH KOCHHAR et al. “Primary sclerosing cholangitis: an experience from India”. In: *Journal of gastroenterology and hepatology* 11.5 (1996), pp. 429–433.

[22] Tiing Leong Ang et al. “Clinical profile of primary sclerosing cholangitis in Singapore”. In: *Journal of gastroenterology and hepatology* 17.8 (2002), pp. 908–913.

[23] Annika Bergquist et al. “Increased prevalence of primary sclerosing cholangitis among first-degree relatives”. In: *Journal of hepatology* 42.2 (2005), pp. 252–256.

[24] Annika Bergquist et al. “Increased risk of primary sclerosing cholangitis and ulcerative colitis in first-degree relatives of patients with primary sclerosing cholangitis”. In: *Clinical gastroenterology and hepatology* 6.8 (2008), pp. 939–943.

[25] Ina Marie Andersen et al. “Effects of coffee consumption, smoking, and hormones on risk for primary sclerosing cholangitis”. In: *Clinical Gastroenterology and Hepatology* 12.6 (2014), pp. 1019–1028.

[26] Kristin G Ardlie, Leonid Kruglyak, and Mark Seielstad. “Patterns of linkage disequilibrium in the human genome”. In: *Nature Reviews Genetics* 3.4 (2002), p. 299.

[27] Mark J Daly et al. “High-resolution haplotype structure in the human genome”. In: *Nature genetics* 29.2 (2001), p. 229.

[28] Jeffrey C Barrett and Lon R Cardon. “Evaluating coverage of genome-wide association studies”. In: *Nature genetics* 38.6 (2006), p. 659.

[29] International HapMap Consortium et al. “A second generation human haplotype map of over 3.1 million SNPs”. In: *Nature* 449.7164 (2007), p. 851.

[30] Clive J Hoggart et al. “Genome-wide significance for dense SNP and resequencing data”. In: *Genetic Epidemiology: The Official Publication of the International Genetic Epidemiology Society* 32.2 (2008), pp. 179–185.

[31] Peter K Gregersen and Lina M Olsson. “Recent advances in the genetics of autoimmune disease”. In: *Annual review of immunology* 27 (2009), pp. 363–391.

[32] Alexandra Zhernakova, Cleo C Van Diemen, and Cisca Wijmenga. “Detecting shared pathogenesis from the shared genetics of immune-related diseases”. In: *Nature Reviews Genetics* 10.1 (2009), p. 43.

[33] Brendan J Keating et al. “Concept, design and implementation of a cardiovascular gene-centric 50 k SNP array for large-scale genomic association studies”. In: *PloS one* 3.10 (2008).

[34] Adrian Cortes and Matthew A Brown. “Promise and pitfalls of the Immunochip”. In: *Arthritis research & therapy* 13.1 (2011), p. 101.

[35] David B Goldstein et al. “Sequencing studies in human genetics: design and interpretation”. In: *Nature Reviews Genetics* 14.7 (2013), p. 460.

[36] Jonathan K Pritchard. “Are rare variants responsible for susceptibility to complex diseases?” In: *The American Journal of Human Genetics* 69.1 (2001), pp. 124–137.

[37] Espen Melum et al. “Genome-wide association analysis in primary sclerosing cholangitis identifies two non-HLA susceptibility loci”. In: *Nature genetics* 43.1 (2011), p. 17.

[38] Trine Folseraas et al. “Extended analysis of a genome-wide association study in primary sclerosing cholangitis detects multiple novel risk loci”. In: *Journal of hepatology* 57.2 (2012), pp. 366–375.

[39] David Ellinghaus et al. “Genome-wide association analysis in Primary sclerosing cholangitis and ulcerative colitis identifies risk loci at GPR35 and TCF4”. In: *Hepatology* 58.3 (2013), pp. 1074–1083.

[40] Jimmy Z Liu et al. “Dense genotyping of immune-related disease regions identifies nine new risk loci for primary sclerosing cholangitis”. In: *Nature genetics* 45.6 (2013), p. 670.

[41] David Ellinghaus et al. “Analysis of five chronic inflammatory diseases identifies 27 new associations and highlights disease-specific patterns at shared loci”. In: *Nature genetics* 48.5 (2016), p. 510.

[42] Sun-Gou Ji et al. “Genome-wide association study of primary sclerosing cholangitis identifies new risk loci and quantifies the genetic relationship with inflammatory bowel disease”. In: *Nature genetics* 49.2 (2017), p. 269.

[43] John Trowsdale and Julian C Knight. “Major histocompatibility complex genomics and human disease”. In: *Annual review of genomics and human genetics* 14 (2013), pp. 301–323.

[44] T. H. Karlsen et al. “Genome-wide association analysis in primary sclerosing cholangitis”. In: *Gastroenterology* 138.3 (2010), pp. 1102–11. ISSN: 1528-0012 (Electronic) 0016-5085 (Linking). DOI: 10.1053/j.gastro.2009.11.046. URL: http://www.ncbi.nlm.nih.gov/pubmed/19944697.

[45] J. Z. Liu et al. “Dense genotyping of immune-related disease regions identifies nine new risk loci for primary sclerosing cholangitis”. In: *Nat Genet* 45.6 (2013), pp. 670–5. ISSN: 1546-1718 (Electronic) 1061-4036 (Linking). DOI: 10.1038/ng.2616. URL: http://www.ncbi.nlm.nih.gov/pubmed/23603763.

[46] Christopher E Lowe et al. “Large-scale genetic fine mapping and genotype-phenotype associations implicate polymorphism in the IL2RA region in type 1 diabetes”. In: *Nature genetics* 39.9 (2007), p. 1074.

[47] Eli A Stahl et al. “Genome-wide association study meta-analysis identifies seven new rheumatoid arthritis risk loci”. In: *Nature genetics* 42.6 (2010), p. 508.

[48] Edward J Carr et al. “Contrasting genetic association of IL2RA with SLE and ANCA–associated vasculitis”. In: *BMC Medical Genetics* 10.1 (2009), p. 22.

[49] Jinguo Wang, Linda S Wicker, and Pere Santamaria. “IL-2 and its high-affinity receptor: genetic control of immunoregulation and autoimmunity”. In: *Seminars in immunology*. Vol. 21. 6. Elsevier. 2009, pp. 363–371.

[50] Li Lu et al. “Hippo signaling is a potent in vivo growth and tumor suppressor pathway in the mammalian liver”. In: *Proceedings of the National Academy of Sciences* 107.4 (2010), pp. 1437–1442.

[51] Koko Katagiri et al. “Mst1 controls lymphocyte trafficking and interstitial motility within lymph nodes”. In: *The EMBO journal* 28.9 (2009), pp. 1319–1331.

[52] Tarana Singh Dang et al. “Defective leukocyte adhesion and chemotaxis contributes to combined immunodeficiency in humans with autosomal recessive MST1 deficiency”. In: *Journal of clinical immunology* 36.2 (2016), pp. 117–122.

[53] Allister J Grant et al. “MAdCAM-1 expressed in chronic inflammatory liver disease supports mucosal lymphocyte adhesion to hepatic endothelium (MAdCAM-1 in chronic inflammatory liver disease)”. In: *Hepatology* 33.5 (2001), pp. 1065–1072.

[54] F Häuser et al. “Macrophage-stimulating protein polymorphism rs3197999 is associated with a gain of function: implications for inflammatory bowel disease”. In: *Genes and immunity* 13.4 (2012), p. 321.

[55] PHILIPPE Goyette et al. “Gene-centric association mapping of chromosome 3p implicates MST1 in IBD pathogenesis”. In: *Mucosal Immunology* 1.2 (2008), p. 131.

[56] Hailiang Huang et al. “Fine-mapping inflammatory bowel disease loci to single-variant resolution”. In: *Nature* 547.7662 (2017), p. 173.

[57] Silvia Fallarini et al. “Expression of functional GPR35 in human iNKT cells”. In: *Biochemical and biophysical research communications* 398.3 (2010), pp. 420–425.

[58] Jinghong Wang et al. “Kynurenic acid as a ligand for orphan G protein-coupled receptor GPR35”. In: *Journal of Biological Chemistry* 281.31 (2006), pp. 22021–22028.

[59] Piotr Paluszkiewicz et al. “High concentration of kynurenic acid in bile and pancreatic juice”. In: *Amino acids* 37.4 (2009), pp. 637–641.

[60] Katrina M de Lange et al. “Genome-wide association study implicates immune activation of multiple integrin genes in inflammatory bowel disease”. In: *Nature genetics* 49.2 (2017), p. 256.

[61] Caroline M Forrest et al. “Purine, kynurenine, neopterin and lipid peroxidation levels in inflammatory bowel disease”. In: *Journal of biomedical science* 9.5 (2002), pp. 436–442.

[62] Georg Schneditz et al. “GPR35 promotes glycolysis, proliferation, and oncogenic signaling by engaging with the sodium potassium pump”. In: *Sci. Signal.* 12.562 (2019), eaau9048.

[63] Marta Wagner et al. “Polymorphisms in CD28, CTLA-4, CD80 and CD86 genes may influence the risk of multiple sclerosis and its age of onset”. In: *Journal of neuroimmunology* 288 (2015), pp. 79–86.

[64] Evaggelia Liaskou et al. “Loss of CD28 expression by liver-infiltrating T cells contributes to pathogenesis of primary sclerosing cholangitis”. In: *Gastroenterology* 147.1 (2014), pp. 221–232.

[65] Kanji Wakabayashi et al. “IL-2 receptor α-/- mice and the development of primary biliary cirrhosis”. In: *Hepatology* 44.5 (2006), pp. 1240–1249.

[66] X Bo et al. “Tumour necrosis factor α impairs function of liver derived T lymphocytes and natural killer cells in patients with primary sclerosing cholangitis”. In: *Gut* 49.1 (2001), pp. 131–141.

[67] Marcial Sebode et al. “Reduced FOXP3+ regulatory T cells in patients with primary sclerosing cholangitis are associated with IL2RA gene polymorphisms”. In: *Journal of hepatology* 60.5 (2014), pp. 1010–1016.

[68] Herbert G Kasler et al. “Histone deacetylase 7 regulates cell survival and TCR signaling in CD4/CD8 double-positive thymocytes”. In: *The Journal of Immunology* (2011), p. 1001179.

[69] Richard N Hanna et al. “The transcription factor NR4A1 (Nur77) controls bone marrow differentiation and the survival of Ly6C- monocytes”. In: *Nature immunology* 12.8 (2011), p. 778.

[70] Franck Dequiedt et al. “Phosphorylation of histone deacetylase 7 by protein kinase D mediates T cell receptor–induced Nur77 expression and apoptosis”. In: *Journal of Experimental Medicine* 201.5 (2005), pp. 793–804.

[71] Franck Dequiedt et al. “HDAC7, a thymus-specific class II histone deacetylase, regulates Nur77 transcription and TCR-mediated apoptosis”. In: *Immunity* 18.5 (2003), pp. 687–698.

[72] Luke Jostins et al. “Host–microbe interactions have shaped the genetic architecture of inflammatory bowel disease”. In: *Nature* 491.7422 (2012), p. 119.

[73] Jimmy Z Liu et al. “Association analyses identify 38 susceptibility loci for inflammatory bowel disease and highlight shared genetic risk across populations”. In: *Nature genetics* 47.9 (2015), p. 979.

[74] Yukinori Okada et al. “HLA-Cw* 1202-B* 5201-DRB1* 1502 haplotype increases risk for ulcerative colitis but reduces risk for Crohn’s disease”. In: *Gastroenterology* 141.3 (2011), pp. 864–871.

[75] Natalie L Berntsen et al. “Association between HLA haplotypes and increased serum levels of IgG4 in patients with primary sclerosing cholangitis”. In: *Gastroenterology* 148.5 (2015), pp. 924–927.

[76] Sigrid Næss et al. “Small duct primary sclerosing cholangitis without inflammatory bowel disease is genetically different from large duct disease”. In: *Liver International* 34.10 (2014), pp. 1488–1495.

[77] Eric J Kunkel et al. “Lymphocyte CC chemokine receptor 9 and epithelial thymus-expressed chemokine (TECK) expression distinguish the small intestinal immune compartment: epithelial expression of tissue-specific chemokines as an organizing principle in regional immunity”. In: *Journal of Experimental Medicine* 192.5 (2000), pp. 761–768.

[78] Palak J Trivedi et al. “Intestinal CCL25 expression is increased in colitis and correlates with inflammatory activity”. In: *Journal of autoimmunity* 68 (2016), pp. 98–104.

[79] Kenneth J Hillan et al. “Expression of the mucosal vascular addressin, MAdCAM-1, in inflammatory liver disease”. In: *Liver* 19.6 (1999), pp. 509–518.

[80] Bertus Eksteen et al. “Hepatic endothelial CCL25 mediates the recruitment of CCR9+ gut-homing lymphocytes to the liver in primary sclerosing cholangitis”. In: *Journal of Experimental Medicine* 200.11 (2004), pp. 1511–1517.

[81] K. M. de Lange et al. “Genome-wide association study implicates immune activation of multiple integrin genes in inflammatory bowel disease”. In: *Nat Genet* 49.2 (2017), pp. 256–261. ISSN: 1546-1718 (Electronic) 1061-4036 (Linking). DOI: 10.1038/ng.3760. URL: https://www.ncbi.nlm.nih.gov/pubmed/28067908.

[82] B. P. Fairfax et al. “Innate immune activity conditions the effect of regulatory variants upon monocyte gene expression”. In: *Science* 343.6175 (2014), p. 1246949. ISSN: 1095-9203 (Electronic) 0036-8075 (Linking). DOI: 10.1126/science.1246949. URL: http://www.ncbi.nlm.nih.gov/pubmed/24604202.

[83] Brian G Feagan et al. “Vedolizumab as induction and maintenance therapy for ulcerative colitis”. In: *New England Journal of Medicine* 369.8 (2013), pp. 699–710.

[84] William J Sandborn et al. “Vedolizumab as induction and maintenance therapy for Crohn’s disease”. In: *New England Journal of Medicine* 369.8 (2013), pp. 711–721.

[85] CS Tse et al. “Effects of vedolizumab, adalimumab and infliximab on biliary inflammation in individuals with primary sclerosing cholangitis and inflammatory bowel disease”. In: *Alimentary pharmacology & therapeutics* 48.2 (2018), pp. 190–195.

[86] B Christensen et al. “Vedolizumab in patients with concurrent primary sclerosing cholangitis and inflammatory bowel disease does not improve liver biochemistry but is safe and effective for the bowel disease”. In: *Alimentary pharmacology & therapeutics* 47.6 (2018), pp. 753–762.

[87] Eva Kristine Klemsdal Henriksen et al. “Gut and liver T-cells of common clonal origin in primary sclerosing cholangitis-inflammatory bowel disease”. In: *Journal of hepatology* 66.1 (2017), pp. 116–122.

[88] Ulrich Beuers et al. “The biliary HCO3- umbrella: a unifying hypothesis on pathogenetic and therapeutic aspects of fibrosing cholangiopathies”. In: *Hepatology* 52.4 (2010), pp. 1489–1496.

[89] Olivier Chazouillères. “Primary sclerosing cholangitis and bile acids”. In: *Clinics and research in hepatology and gastroenterology* 36 (2012), S21–S25.

[90] Peter Fickert et al. “Regurgitation of bile acids from leaky bile ducts causes sclerosing cholangitis in Mdr2 (Abcb4) knockout mice”. In: *Gastroenterology* 127.1 (2004), pp. 261–274.

[91] Tom H Karlsen et al. “Genome-wide association analysis in primary sclerosing cholangitis”. In: *Gastroenterology* 138.3 (2010), pp. 1102–1111.

[92] Keith D Lindor et al. “High-dose ursodeoxycholic acid for the treatment of primary sclerosing cholangitis”. In: *Hepatology* 50.3 (2009), pp. 808–814.

[93] Emmanouil Sinakos et al. “Bile acid changes after high-dose ursodeoxycholic acid treatment in primary sclerosing cholangitis: Relation to disease progression”. In: *Hepatology* 52.1 (2010), pp. 197–203.

[94] A. F. Hofmann et al. “Novel biotransformation and physiological properties of norursodeoxycholic acid in humans”. In: *Hepatology* 42.6 (2005), pp. 1391–8. ISSN: 0270-9139 (Print) 0270-9139 (Linking). DOI: 10.1002/hep.20943. URL: http://www.ncbi.nlm.nih.gov/pubmed/16317695.

[95] Michael Trauner et al. “Potential of nor-ursodeoxycholic acid in cholestatic and metabolic disorders”. In: *Digestive Diseases* 33.3 (2015), pp. 433–439.

[96] Peter Fickert et al. “norUrsodeoxycholic acid improves cholestasis in primary sclerosing cholangitis”. In: *Journal of hepatology* 67.3 (2017), pp. 549–558.

[97] Kris V Kowdley et al. “A randomized trial of obeticholic acid monotherapy in patients with primary biliary cholangitis”. In: *Hepatology* 67.5 (2018), pp. 1890–1902.

[98] Valerio Pontecorvi, Marco Carbone, and Pietro Invernizzi. “The “gut microbiota” hypothesis in primary sclerosing cholangitis”. In: *Annals of translational medicine* 4.24 (2016).

[99] Andrew J Macpherson and Karen Smith. “Mesenteric lymph nodes at the center of immune anatomy”. In: *Journal of Experimental Medicine* 203.3 (2006), pp. 497–500.

[100] Malin EV Johansson et al. “Bacteria penetrate the normally impenetrable inner colon mucus layer in both murine colitis models and patients with ulcerative colitis”. In: *Gut* 63.2 (2014), pp. 281–291.

[101] Andrea Michielan and Renata D’Incà. “Intestinal permeability in inflammatory bowel disease: pathogenesis, clinical evaluation, and therapy of leaky gut”. In: *Mediators of inflammation* 2015 (2015).

[102] Rolf Olsson et al. “Bile duct bacterial isolates in primary sclerosing cholangitis: a study of explanted livers”. In: *Journal of hepatology* 28.3 (1998), pp. 426–432.

[103] Martti Färkkilä et al. “Metronidazole and ursodeoxycholic acid for primary sclerosing cholangitis: A randomized placebo-controlled trial”. In: *Hepatology* 40.6 (2004), pp. 1379–1386.

[104] Iris C Steenstraten et al. “Systematic review with meta-analysis: risk factors for recurrent primary sclerosing cholangitis after liver transplantation”. In: *Alimentary pharmacology & therapeutics* 49.6 (2019), pp. 636–643.

[105] Aleksandar D Kostic, Ramnik J Xavier, and Dirk Gevers. “The microbiome in inflammatory bowel disease: current status and the future ahead”. In: *Gastroenterology* 146.6 (2014), pp. 1489–1499.

[106] Marie Joossens et al. “Dysbiosis of the faecal microbiota in patients with Crohn’s disease and their unaffected relatives”. In: *Gut* 60.5 (2011), pp. 631–637.

[107] Harry Sokol et al. “Faecalibacterium prausnitzii is an anti-inflammatory commensal bacterium identified by gut microbiota analysis of Crohn disease patients”. In: *Proceedings of the National Academy of Sciences* 105.43 (2008), pp. 16731–16736.

[108] Kathleen Machiels et al. “A decrease of the butyrate-producing species Roseburia hominis and Faecalibacterium prausnitzii defines dysbiosis in patients with ulcerative colitis”. In: *Gut* 63.8 (2014), pp. 1275–1283.

[109] João Sabino et al. “Primary sclerosing cholangitis is characterised by intestinal dysbiosis independent from IBD”. In: *Gut* 65.10 (2016), pp. 1681–1689.

[110] Williams Turpin et al. “Association of host genome with intestinal microbial composition in a large healthy cohort”. In: *Nature genetics* 48.11 (2016), p. 1413.

[111] Marc Jan Bonder et al. “The effect of host genetics on the gut microbiome”. In: *Nature genetics* 48.11 (2016), p. 1407.

[112] Dan Knights et al. “Complex host genetics influence the microbiome in inflammatory bowel disease”. In: *Genome medicine* 6.12 (2014), p. 107.

[113] Sarah L Spain and Jeffrey C Barrett. “Strategies for fine-mapping complex traits”. In: *Human molecular genetics* 24.R1 (2015), R111–R119.

[114] Harm-Jan Westra et al. “Fine-mapping and functional studies highlight potential causal variants for rheumatoid arthritis and type 1 diabetes”. In: *Nature genetics* 50.10 (2018), pp. 1366–1374.

[115] Matthew T Maurano et al. “Systematic localization of common disease-associated variation in regulatory DNA”. In: *Science* (2012), p. 1222794.

[116] William Cookson et al. “Mapping complex disease traits with global gene expression”. In: *Nature Reviews Genetics* 10.3 (2009), p. 184.

[117] Dan L Nicolae et al. “Trait-associated SNPs are more likely to be eQTLs: annotation to enhance discovery from GWAS”. In: *PLoS genetics* 6.4 (2010), e1000888.

[118] Harm-Jan Westra et al. “Systematic identification of trans eQTLs as putative drivers of known disease associations”. In: *Nature genetics* 45.10 (2013), p. 1238.

[119] Alexandra C Nica et al. “Candidate causal regulatory effects by integration of expression QTLs with complex trait genetic associations”. In: *PLoS genetics* 6.4 (2010), e1000895.

[120] Barbara E Stranger et al. “Population genomics of human gene expression”. In: *Nature genetics* 39.10 (2007), p. 1217.

[121] Federico Innocenti et al. “Identification, replication, and functional fine-mapping of expression quantitative trait loci in primary human liver tissue”. In: *PLoS genetics* 7.5 (2011), e1002078.

[122] Barbara E Stranger et al. “Relative impact of nucleotide and copy number variation on gene expression phenotypes”. In: *Science* 315.5813 (2007), pp. 848–853.

[123] Jean-Baptiste Veyrieras et al. “High-resolution mapping of expression-QTLs yields insight into human gene regulation”. In: *PLoS genetics* 4.10 (2008), e1000214.

[124] Jacob F Degner et al. “DNase I sensitivity QTLs are a major determinant of human expression variation”. In: *Nature* 482.7385 (2012), p. 390.

[125] Christopher D Brown, Lara M Mangravite, and Barbara E Engelhardt. “Integrative modeling of eQTLs and cis-regulatory elements suggests mechanisms underlying cell type specificity of eQTLs”. In: *PLoS genetics* 9.8 (2013), e1003649.

[126] Scott Smemo et al. “Obesity-associated variants within FTO form long-range functional connections with IRX3”. In: *Nature* 507.7492 (2014), p. 371.

[127] Rudolf SN Fehrmann et al. “Trans-eQTLs reveal that independent genetic variants associated with a complex phenotype converge on intermediate genes, with a major role for the HLA”. In: *PLoS genetics* 7.8 (2011), e1002197.

[128] Matthias Heinig et al. “A trans-acting locus regulates an anti-viral expression network and type 1 diabetes risk”. In: *Nature* 467.7314 (2010), p. 460.

[129] Benjamin P Fairfax et al. “Genetics of gene expression in primary immune cells identifies cell type–specific master regulators and roles of HLA alleles”. In: *Nature genetics* 44.5 (2012), p. 502.

[130] Benjamin P Fairfax et al. “Innate immune activity conditions the effect of regulatory variants upon monocyte gene expression”. In: *Science* 343.6175 (2014), p. 1246949.

[131] Alice Gerrits et al. “Expression quantitative trait loci are highly sensitive to cellular differentiation state”. In: *PLoS genetics* 5.10 (2009), e1000692.

[132] Erin L Heinzen et al. “Tissue-specific genetic control of splicing: implications for the study of complex traits”. In: *PLoS biology* 6.12 (2008), e1000001.

[133] Antigone S Dimas et al. “Common regulatory variation impacts gene expression in a cell type–dependent manner”. In: *Science* 325.5945 (2009), pp. 1246–1250.

[134] Kaur Alasoo et al. “Shared genetic effects on chromatin and gene expression indicate a role for enhancer priming in immune response”. In: *Nature genetics* 50.3 (2018), p. 424.

[135] Bradley E Bernstein, Alexander Meissner, and Eric S Lander. “The mammalian epigenome”. In: *Cell* 128.4 (2007), pp. 669–681.

[136] Zhenhai Zhang and B Franklin Pugh. "High-resolution genome-wide mapping of the primary structure of chromatin". In: *Cell* 144.2 (2011), pp. 175–186.

[137] John Newell-Price, Adrian JL Clark, and Peter King. "DNA methylation and silencing of gene expression". In: *Trends in Endocrinology & Metabolism* 11.4 (2000), pp. 142–148.

[138] Egor Prokhortchouk and Pierre-Antoine Defossez. "The cell biology of DNA methylation in mammals". In: *Biochimica Et Biophysica Acta (BBA)-Molecular Cell Research* 1783.11 (2008), pp. 2167–2173.

[139] Lanlan Shen et al. "Genome-wide profiling of DNA methylation reveals a class of normally methylated CpG island promoters". In: *PLoS genetics* 3.10 (2007), e181.

[140] Robert Illingworth et al. "A novel CpG island set identifies tissue-specific methylation at developmental gene loci". In: *PLoS biology* 6.1 (2008), e22.

[141] Tina Branscombe Miranda and Peter A Jones. "DNA methylation: the nuts and bolts of repression". In: *Journal of cellular physiology* 213.2 (2007), pp. 384–390.

[142] Michael Weber and Dirk Schübeler. "Genomic patterns of DNA methylation: targets and function of an epigenetic mark". In: *Current opinion in cell biology* 19.3 (2007), pp. 273–280.

[143] Eric Hervouet, Francois M Vallette, and Pierre-Francois Cartron. "Dnmt3/transcription factor interactions as crucial players in targeted DNA methylation". In: *Epigenetics* 4.7 (2009), pp. 487–499.

[144] Paul A Wade. "Methyl CpG binding proteins: coupling chromatin architecture to gene regulation". In: *Oncogene* 20.24 (2001), p. 3166.

[145] Allan F McRae et al. "Identification of 55,000 replicated DNA methylation QTL". In: *Scientific Reports* 8.1 (2018), p. 17605.

[146] Marc Jan Bonder et al. "Disease variants alter transcription factor levels and methylation of their binding sites". In: *Nature genetics* 49.1 (2017), p. 131.

[147] Mark C Genovese et al. "Abatacept for rheumatoid arthritis refractory to tumor necrosis factor α inhibition". In: *New England Journal of Medicine* 353.11 (2005), pp. 1114–1123.

[148] Yukinori Okada et al. "Genetics of rheumatoid arthritis contributes to biology and drug discovery". In: *Nature* 506.7488 (2014), p. 376.

[149] William J Sandborn et al. "Ustekinumab induction and maintenance therapy in refractory Crohn's disease". In: *New England Journal of Medicine* 367.16 (2012), pp. 1519–1528.

[150] Andre Franke et al. “Genome-wide meta-analysis increases to 71 the number of confirmed Crohn’s disease susceptibility loci”. In: *Nature genetics* 42.12 (2010), p. 1118.

[151] Robert M Plenge, Edward M Scolnick, and David Altshuler. “Validating therapeutic targets through human genetics”. In: *Nature reviews Drug discovery* 12.8 (2013), pp. 581–594.

[152] Matthew R Nelson et al. “The support of human genetic evidence for approved drug indications”. In: *Nature genetics* 47.8 (2015), p. 856.

[153] UK10K consortium et al. “The UK10K project identifies rare variants in health and disease”. In: *Nature* 526.7571 (2015), p. 82.

[154] 1000 Genomes Project Consortium et al. “An integrated map of genetic variation from 1,092 human genomes”. In: *Nature* 491.7422 (2012), p. 56.

[155] Gleb Kichaev et al. “Integrating functional data to prioritize causal variants in statistical fine-mapping studies”. In: *PLoS genetics* 10.10 (2014).

[156] Farhad Hormozdiari et al. “Identifying causal variants at loci with multiple signals of association”. In: *Genetics* 198.2 (2014), pp. 497–508.

[157] Wenan Chen et al. “Fine mapping causal variants with an approximate Bayesian method using marginal test statistics”. In: *Genetics* 200.3 (2015), pp. 719–736.

[158] Christian Benner et al. “FINEMAP: efficient variable selection using summary data from genome-wide association studies”. In: *Bioinformatics* 32.10 (2016), pp. 1493–1501.

[159] Christian Benner et al. “Prospects of fine-mapping trait-associated genomic regions by using summary statistics from genome-wide association studies”. In: *The American Journal of Human Genetics* 101.4 (2017), pp. 539–551.

[160] Chris Hans, Adrian Dobra, and Mike West. “Shotgun stochastic search for “large p” regression”. In: *Journal of the American Statistical Association* 102.478 (2007), pp. 507–516.

[161] Carrie A Davis et al. “The Encyclopedia of DNA elements (ENCODE): data portal update”. In: *Nucleic acids research* 46.D1 (2018), pp. D794–D801.

[162] Gosia Trynka et al. “Disentangling the effects of colocalizing genomic annotations to functionally prioritize non-coding variants within complex-trait loci”. In: *The American Journal of Human Genetics* 97.1 (2015), pp. 139–152.

[163] Jacob C Ulirsch et al. “Interrogation of human hematopoiesis at single-cell and single-variant resolution”. In: *Nature genetics* 51.4 (2019), p. 683.

[164] Yoav Benjamini and Yosef Hochberg. "Controlling the false discovery rate: a practical and powerful approach to multiple testing". In: *Journal of the Royal statistical society: series B (Methodological)* 57.1 (1995), pp. 289–300.

[165] Haikun Wang et al. "The transcription factor Foxp1 is a critical negative regulator of the differentiation of follicular helper T cells". In: *Nature immunology* 15.7 (2014), p. 667.

[166] Jing He et al. "Circulating precursor CCR7loPD-1hi CXCR5+ CD4+ T cells indicate Tfh cell activity and promote antibody responses upon antigen reexposure". In: *Immunity* 39.4 (2013), pp. 770–781.

[167] Leonie Adam et al. "Follicular T Helper Cell Signatures in Primary Biliary Cholangitis and Primary Sclerosing Cholangitis". In: *Hepatology communications* 2.9 (2018), pp. 1051–1063.

[168] Philipp Rentzsch et al. "CADD: predicting the deleteriousness of variants throughout the human genome". In: *Nucleic acids research* 47.D1 (2018), pp. D886–D894.

[169] Marieke JH Coenen et al. "Common and different genetic background for rheumatoid arthritis and coeliac disease". In: *Human molecular genetics* 18.21 (2009), pp. 4195–4203.

[170] JA Todd et al. "Genetics of Type 1 Diabetes in Finland, Simmonds MJ, Heward JM, Gough SC, Wellcome Trust Case Control Consortium, Dunger DB, Wicker LS, Clayton DG: Robust associations of four new chromosome regions from genome-wide analyses of type 1 diabetes". In: *Nat Genet* 39.7 (2007), pp. 857–864.

[171] Ynto S De Boer et al. "Genome-wide association study identifies variants associated with autoimmune hepatitis type 1". In: *Gastroenterology* 147.2 (2014), pp. 443–452.

[172] H. J. Westra et al. "Systematic identification of trans eQTLs as putative drivers of known disease associations". In: *Nat Genet* 45.10 (2013), pp. 1238–43. ISSN: 1546-1718 (Electronic) 1061-4036 (Linking). DOI: 10.1038/ng.2756. URL: https://www.ncbi.nlm.nih.gov/pubmed/24013639.

[173] Alexandra Zhernakova et al. "Evolutionary and functional analysis of celiac risk loci reveals SH2B3 as a protective factor against bacterial infection". In: *The American Journal of Human Genetics* 86.6 (2010), pp. 970–977.

[174] Marwa Chaouali et al. "Association of STAT4, TGFβ1, SH2B3 and PTPN22 polymorphisms with autoimmune hepatitis". In: *Experimental and molecular pathology* 105.3 (2018), pp. 279–284.

[175] Subra Kugathasan et al. "Loci on 20q13 and 21q22 are associated with pediatric-onset inflammatory bowel disease". In: *Nature genetics* 40.10 (2008), p. 1211.

[176] GTEx Consortium et al. “The Genotype-Tissue Expression (GTEx) pilot analysis: multitissue gene regulation in humans”. In: *Science* 348.6235 (2015), pp. 648–660.

[177] Alexis Battle et al. “Characterizing the genetic basis of transcriptome diversity through RNA-sequencing of 922 individuals”. In: *Genome research* 24.1 (2014), pp. 14–24.

[178] MQ Zhang. “Statistical features of human exons and their flanking regions”. In: *Human molecular genetics* 7.5 (1998), pp. 919–932.

[179] Mingqi Tan et al. “Lysophosphatidylcholine activates a novel PKD2-mediated signaling pathway that controls monocyte migration”. In: *Arteriosclerosis, thrombosis, and vascular biology* 29.9 (2009), pp. 1376–1382.

[180] Xiaona Ge et al. “Angiotensin II directly triggers endothelial exocytosis via protein kinase C-dependent protein kinase D2 activation”. In: *Journal of pharmacological sciences* 105.2 (2007), pp. 168–176.

[181] Xavier Messeguer et al. “PROMO: detection of known transcription regulatory elements using species-tailored searches”. In: *Bioinformatics* 18.2 (2002), pp. 333–334.

[182] Chris Wallace et al. “Statistical colocalization of monocyte gene expression and genetic risk variants for type 1 diabetes”. In: *Human molecular genetics* 21.12 (2012), pp. 2815–2824.

[183] Tsun-Po Yang et al. “Genevar: a database and Java application for the analysis and visualization of SNP-gene associations in eQTL studies”. In: *Bioinformatics* 26.19 (2010), pp. 2474–2476.

[184] Anna L Dixon et al. “A genome-wide association study of global gene expression”. In: *Nature genetics* 39.10 (2007), p. 1202.

[185] Miriam F Moffatt et al. “Genetic variants regulating ORMDL3 expression contribute to the risk of childhood asthma”. In: *Nature* 448.7152 (2007), p. 470.

[186] Alexandra C Nica and Emmanouil T Dermitzakis. “Using gene expression to investigate the genetic basis of complex disorders”. In: *Human molecular genetics* 17.R2 (2008), R129–R134.

[187] Vincent Plagnol et al. “Statistical independence of the colocalized association signals for type 1 diabetes and RPS26 gene expression on chromosome 12q13”. In: *Biostatistics* 10.2 (2008), pp. 327–334.

[188] Chris Wallace. “Statistical testing of shared genetic control for potentially related traits”. In: *Genetic epidemiology* 37.8 (2013), pp. 802–813.

[189] Claudia Giambartolomei et al. “Bayesian test for colocalisation between pairs of genetic association studies using summary statistics”. In: *PLoS genetics* 10.5 (2014), e1004383.

[190] Jun Ding et al. “Gene expression in skin and lymphoblastoid cells: Refined statistical method reveals extensive overlap in cis-eQTL signals”. In: *The American Journal of Human Genetics* 87.6 (2010), pp. 779–789.

[191] Alexander L Richards et al. “Schizophrenia susceptibility alleles are enriched for alleles that affect gene expression in adult human brain”. In: *Molecular psychiatry* 17.2 (2012), p. 193.

[192] Jon Wakefield. “Bayes factors for genome-wide association studies: comparison with P-values”. In: *Genetic Epidemiology: The Official Publication of the International Genetic Epidemiology Society* 33.1 (2009), pp. 79–86.

[193] Mary D Fortune et al. “Statistical colocalization of genetic risk variants for related autoimmune diseases in the context of common controls”. In: *Nature genetics* 47.7 (2015), p. 839.

[194] Urmo Võsa et al. “Unraveling the polygenic architecture of complex traits using blood eQTL meta-analysis”. In: *bioRxiv* (2018), p. 447367.

[195] Hendrik G Stunnenberg et al. “The International Human Epigenome Consortium: a blueprint for scientific collaboration and discovery”. In: *Cell* 167.5 (2016), pp. 1145–1149.

[196] Lu Chen et al. “Genetic drivers of epigenetic and transcriptional variation in human immune cells”. In: *Cell* 167.5 (2016), pp. 1398–1414.

[197] John Lonsdale et al. “The genotype-tissue expression (GTEx) project”. In: *Nature genetics* 45.6 (2013), p. 580.

[198] Sarah Kim-Hellmuth et al. “Genetic regulatory effects modified by immune activation contribute to autoimmune disease associations”. In: *Nature communications* 8.1 (2017), p. 266.

[199] Heather J Cordell et al. “International genome-wide meta-analysis identifies new primary biliary cirrhosis risk loci and targetable pathogenic pathways”. In: *Nature communications* 6 (2015), p. 8019.

[200] Jonathan P Bradfield et al. “A genome-wide meta-analysis of six type 1 diabetes cohorts identifies multiple associated loci”. In: *PLoS genetics* 7.9 (2011), e1002293.

[201] Gosia Trynka et al. “Dense genotyping identifies and localizes multiple common and rare variant association signals in celiac disease”. In: *Nature genetics* 43.12 (2011), p. 1193.

[202] Ashley H Beecham et al. “Analysis of immune-related loci identifies 48 new susceptibility variants for multiple sclerosis”. In: *Nature genetics* 45.11 (2013), p. 1353.

[203] James Bentham et al. “Genetic association analyses implicate aberrant regulation of innate and adaptive immunity genes in the pathogenesis of systemic lupus erythematosus”. In: *Nature genetics* 47.12 (2015), p. 1457.

[204] William J Astle et al. “The allelic landscape of human blood cell trait variation and links to common complex disease”. In: *Cell* 167.5 (2016), pp. 1415–1429.

[205] Sharon A Matthews et al. “Unique functions for protein kinase D1 and protein kinase D2 in mammalian cells”. In: *Biochemical Journal* 432.1 (2010), pp. 153–163.

[206] M Armacki et al. “A novel splice variant of calcium and integrin-binding protein 1 mediates protein kinase D2-stimulated tumour growth by regulating angiogenesis”. In: *Oncogene* 33.9 (2014), p. 1167.

[207] Atsushi Irie et al. “Protein kinase D2 contributes to either IL-2 promoter regulation or induction of cell death upon TCR stimulation depending on its activity in Jurkat cells”. In: *International immunology* 18.12 (2006), pp. 1737–1747.

[208] Kenneth L Rock et al. “Inhibitors of the proteasome block the degradation of most cell proteins and the generation of peptides presented on MHC class I molecules”. In: *Cell* 78.5 (1994), pp. 761–771.

[209] Jian Zhang et al. “Cutting edge: regulation of T cell activation threshold by CD28 costimulation through targeting Cbl-b for ubiquitination”. In: *The Journal of Immunology* 169.5 (2002), pp. 2236–2240.

[210] Melissa Hunter et al. “Survival of monocytes and macrophages and their role in health and disease”. In: *Frontiers in bioscience: a journal and virtual library* 14 (2009), p. 4079.

[211] Guo Wei et al. “Activated Ets2 is required for persistent inflammatory responses in the motheaten viable model”. In: *The Journal of Immunology* 173.2 (2004), pp. 1374–1379.

[212] Ioannis Panagoulias et al. “Transcription factor Ets-2 acts as a preinduction repressor of interleukin-2 (IL-2) transcription in naive T helper lymphocytes”. In: *Journal of Biological Chemistry* 291.52 (2016), pp. 26707–26721.

[213] A. Battle et al. “Characterizing the genetic basis of transcriptome diversity through RNA-sequencing of 922 individuals”. In: *Genome Res* 24.1 (2014), pp. 14–24. ISSN: 1549-5469 (Electronic) 1088-9051 (Linking). DOI: 10.1101/gr.155192.113. URL: https://www.ncbi.nlm.nih.gov/pubmed/24092820.

[214] B. P. Fairfax et al. “Genetics of gene expression in primary immune cells identifies cell type-specific master regulators and roles of HLA alleles”. In: *Nat Genet* 44.5 (2012), pp. 502–10. ISSN: 1546-1718 (Electronic) 1061-4036 (Linking). DOI: 10.1038/ng.2205. URL: http://www.ncbi.nlm.nih.gov/pubmed/22446964.

[215] Tuuli Lappalainen et al. “Transcriptome and genome sequencing uncovers functional variation in humans”. In: *Nature* 501.7468 (2013), pp. 506–511.

[216] Yan Ge et al. “UBASH3A mediates risk for type 1 diabetes through inhibition of T-cell receptor-induced NF-κB signaling”. In: *Diabetes* (2017), db161023.

[217] Yan Ge and Patrick Concannon. “Molecular-genetic characterization of common, noncoding UBASH3A variants associated with type 1 diabetes”. In: *European Journal of Human Genetics* (2018), p. 1.

[218] Julie Devallière and Béatrice Charreau. “The adaptor Lnk (SH2B3): an emerging regulator in vascular cells and a link between immune and inflammatory signaling”. In: *Biochemical pharmacology* 82.10 (2011), pp. 1391–1402.

[219] Robert J Crowder et al. “Dok-6, a Novel p62 Dok family member, promotes Ret-mediated neurite outgrowth”. In: *Journal of Biological Chemistry* 279.40 (2004), pp. 42072–42081.

[220] Jose L Badano et al. “The ciliopathies: an emerging class of human genetic disorders”. In: *Annu. Rev. Genomics Hum. Genet.* 7 (2006), pp. 125–148.

[221] Nicholas F LaRusso and Tetyana V Masyuk. “The role of cilia in the regulation of bile flow”. In: *Digestive Diseases* 29.1 (2011), pp. 6–12.

[222] Jo Knight et al. “Using functional annotation for the empirical determination of Bayes Factors for genome-wide association study analysis”. In: *PLoS One* 6.4 (2011), e14808.

[223] Jian Yang et al. “Conditional and joint multiple-SNP analysis of GWAS summary statistics identifies additional variants influencing complex traits”. In: *Nature genetics* 44.4 (2012), p. 369.

[224] Claudia Giambartolomei et al. “A Bayesian framework for multiple trait colocalization from summary association statistics”. In: *Bioinformatics* 34.15 (2018), pp. 2538–2545.

[225] Chris Wallace et al. “Dissection of a complex disease susceptibility region using a Bayesian stochastic search approach to fine mapping”. In: *PLoS genetics* 11.6 (2015).

[226] Jason Lewis. “Pathological patterns of biliary disease”. In: *Clinical liver disease* 10.5 (2017), p. 107.

[227] Alexander Dobin et al. “STAR: ultrafast universal RNA-seq aligner”. In: *Bioinformatics* 29.1 (2013), pp. 15–21.

[228] Yang Liao, Gordon K Smyth, and Wei Shi. “featureCounts: an efficient general purpose program for assigning sequence reads to genomic features”. In: *Bioinformatics* 30.7 (2013), pp. 923–930.

[229] Jeremy Davis-Turak et al. “Genomics pipelines and data integration: challenges and opportunities in the research setting”. In: *Expert review of molecular diagnostics* 17.3 (2017), pp. 225–237.

[230] Michael I Love, Wolfgang Huber, and Simon Anders. “Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2”. In: *Genome biology* 15.12 (2014), p. 550.

[231] Alexandre Fort et al. “MBV: a method to solve sample mislabeling and detect technical bias in large combined genotype and sequencing assay datasets”. In: *Bioinformatics* 33.12 (2017), pp. 1895–1897.

[232] Franck Rapaport et al. “Comprehensive evaluation of differential gene expression analysis methods for RNA-seq data”. In: *Genome biology* 14.9 (2013), p. 3158.

[233] Alicia Oshlack and Matthew J Wakefield. “Transcript length bias in RNA-seq data confounds systems biology”. In: *Biology direct* 4.1 (2009), p. 14.

[234] Joseph K Pickrell et al. “Understanding mechanisms underlying human gene expression variation with RNA sequencing”. In: *Nature* 464.7289 (2010), p. 768.

[235] Simon Anders and Wolfgang Huber. “Differential expression analysis for sequence count data”. In: *Genome biology* 11.10 (2010), R106.

[236] Jüri Reimand et al. “g: Profiler—a web-based toolset for functional profiling of gene lists from large-scale experiments”. In: *Nucleic acids research* 35.suppl_2 (2007), W193–W200.

[237] Carl A Anderson et al. “Data quality control in genetic case-control association studies”. In: *Nature protocols* 5.9 (2010), p. 1564.

[238] Shane McCarthy et al. “A reference panel of 64,976 haplotypes for genotype imputation”. In: *Nature genetics* 48.10 (2016), p. 1279.

[239] Jonathan RI Coleman et al. “Quality control, imputation and analysis of genome-wide genotyping data from the Illumina HumanCoreExome microarray”. In: *Briefings in functional genomics* 15.4 (2016), pp. 298–304.

[240] Hao Zhao et al. “CrossMap: a versatile tool for coordinate conversion between genome assemblies”. In: *Bioinformatics* 30.7 (2013), pp. 1006–1007.

[241] Olivier Delaneau et al. “A complete tool set for molecular QTL discovery and analysis”. In: *Nature communications* 8.1 (2017), pp. 1–7.

[242] Halit Ongen et al. “Fast and efficient QTL mapper for thousands of molecular phenotypes”. In: *Bioinformatics* 32.10 (2015), pp. 1479–1485.

[243] John D Storey and Robert Tibshirani. “Statistical significance for genomewide studies”. In: *Proceedings of the National Academy of Sciences* 100.16 (2003), pp. 9440–9445.

[244] Timothée Flutre et al. “A statistical framework for joint eQTL analysis in multiple tissues”. In: *PLoS genetics* 9.5 (2013), e1003486.

[245] Yingying Wei, Toyoaki Tenzen, and Hongkai Ji. “Joint analysis of differential gene expression in multiple studies using correlation motifs”. In: *Biostatistics* 16.1 (2014), pp. 31–46.

[246] Sarah M Urbut et al. “Flexible statistical methods for estimating and testing effects in genomic studies with multiple conditions”. In: *Nature genetics* 51.1 (2019), pp. 187–195.

[247] Sarah Kim-Hellmuth et al. “Cell type specific genetic regulation of gene expression across human tissues”. In: *bioRxiv* (2019), p. 806117.

[248] James C Lee et al. “Gene expression profiling of CD8+ T cells predicts prognosis in patients with Crohn disease and ulcerative colitis”. In: *The Journal of clinical investigation* 121.10 (2011), pp. 4170–4179.

[249] Daniele Biasci et al. “A blood-based prognostic biomarker in inflammatory bowel disease”. In: *bioRxiv* (2019), p. 535153.

[250] Y Kim et al. “A meta-analysis of gene expression quantitative trait loci in brain”. In: *Translational psychiatry* 4.10 (2014), e459–e459.

[251] Katharina Schramm et al. “Mapping the genetic architecture of gene regulation in whole blood”. In: *PLoS One* 9.4 (2014).

[252] Yong He et al. “Long noncoding RNAs: Novel insights into hepatocelluar carcinoma”. In: *Cancer letters* 344.1 (2014), pp. 20–27.

[253] Bodu Liu et al. “A cytoplasmic NF-κB interacting long noncoding RNA blocks IκB phosphorylation and suppresses breast cancer metastasis”. In: *Cancer cell* 27.3 (2015), pp. 370–381.

[254] Pin Wang et al. “The STAT3-binding long noncoding RNA lnc-DC controls human dendritic cell differentiation”. In: *Science* 344.6181 (2014), pp. 310–313.

[255] Charles F Spurlock et al. “Expression and functions of long noncoding RNAs during human T helper cell differentiation”. In: *Nature communications* 6.1 (2015), pp. 1–12.

[256] Susan Carpenter et al. “A long noncoding RNA mediates both activation and repression of immune response genes”. In: *science* 341.6147 (2013), pp. 789–792.

[257] Jinsoo Song et al. “PBMC and exosome-derived Hotair is a critical regulator and potent marker for rheumatoid arthritis”. In: *Clinical and experimental medicine* 15.1 (2015), pp. 121–126.

[258] Ming-Chi Lu et al. “Increased expression of long noncoding RNAs LOC100652951 and LOC100506036 in T cells from patients with rheumatoid arthritis facilitates the inflammatory responses”. In: *Immunologic research* 64.2 (2016), pp. 576–583.

[259] Qing-Lin Peng et al. “Transcriptomic profiling of long non-coding RNAs in dermatomyositis by microarray analysis”. In: *Scientific reports* 6 (2016), p. 32818.

[260] Isis Ricaño-Ponce et al. “Refined mapping of autoimmune disease associated genetic variants with gene expression suggests an important role for non-coding RNAs”. In: *Journal of autoimmunity* 68 (2016), pp. 62–74.

[261] Benjamin J Schmiedel et al. “Impact of genetic polymorphisms on human immune cell gene expression”. In: *Cell* 175.6 (2018), pp. 1701–1715.

[262] Akinori Takaoka et al. “Integral role of IRF-5 in the gene induction programme activated by Toll-like receptors”. In: *Nature* 434.7030 (2005), pp. 243–249.

[263] Betsy J Barnes et al. “Multiple regulatory domains of IRF-5 control activation, cellular localization, and induction of chemokines that mediate recruitment of T lymphocytes”. In: *Molecular and Cellular Biology* 22.16 (2002), pp. 5721–5740.

[264] G Gathungu et al. “A two-marker haplotype in the IRF5 gene is associated with inflammatory bowel disease in a North American cohort”. In: *Genes & Immunity* 13.4 (2012), pp. 351–355.

[265] Hannah Almuttaqi and Irina A Udalova. “Advances and challenges in targeting IRF5, a key regulator of inflammation”. In: *The FEBS journal* 286.9 (2019), pp. 1624–1637.

[266] Antonio F Di Narzo et al. “Blood and intestine eQTLs from an anti-TNF-resistant Crohn’s disease cohort inform IBD genetic association loci”. In: *Clinical and translational gastroenterology* 7.6 (2016), e177.

[267] Letizia Straniero et al. “The GBAP1 pseudogene acts as a ceRNA for the glucocerebrosidase gene GBA by sponging miR-22-3p”. In: *Scientific reports* 7.1 (2017), pp. 1–13.

[268] S Pablo Sardi et al. "Augmenting CNS glucocerebrosidase activity as a therapeutic strategy for parkinsonism and other Gaucher-related synucleinopathies". In: *Proceedings of the National Academy of Sciences* 110.9 (2013), pp. 3537–3542.

[269] Jie Liu et al. "Decreased UBASH3A mRNA expression levels in peripheral blood mononuclear cells from patients with systemic lupus erythematosus". In: *Inflammation* 38.5 (2015), pp. 1903–1910.

[270] Rosana HCN Freitas and Carlos AM Fraga. "NF-κB-IKKβ pathway as a target for drug development: Realities, challenges and perspectives". In: *Current drug targets* 19.16 (2018), pp. 1933–1942.

[271] Jonggi Choi et al. "Aspirin use and the risk of cholangiocarcinoma". In: *Hepatology* 64.3 (2016), pp. 785–796.

[272] Jianping Xiong et al. "Aspirin use is associated with a reduced risk of cholangiocarcinoma: a systematic review and meta-analysis". In: *Cancer management and research* 10 (2018), p. 4095.

[273] Karl Egerer et al. "Circulating proteasomes are markers of cell damage and immunologic activity in autoimmune diseases." In: *The Journal of rheumatology* 29.10 (2002), pp. 2045–2052.

[274] Stephan U Sixt and Burkhardt Dahlmann. "Extracellular, circulating proteasomes and ubiquitin—incidence and relevance". In: *Biochimica et Biophysica Acta (BBA)-Molecular Basis of Disease* 1782.12 (2008), pp. 817–823.

[275] Thomas Egerer et al. "Tissue-specific up-regulation of the proteasome subunit β5i (LMP7) in Sjögren's syndrome". In: *Arthritis & Rheumatism: Official Journal of the American College of Rheumatology* 54.5 (2006), pp. 1501–1508.

[276] Sue Ellen Verbrugge et al. "Proteasome inhibitors as experimental therapeutics of autoimmune diseases". In: *Arthritis research & therapy* 17.1 (2015), p. 17.

[277] Michael Brychcy et al. "Anti-20S proteasome autoantibodies inhibit proteasome stimulation by proteasome activator PA28". In: *Arthritis & Rheumatism: Official Journal of the American College of Rheumatology* 54.7 (2006), pp. 2175–2183.

[278] Eugen Feist et al. "Proteasome alpha-type subunit C9 is a primary target of autoantibodies in sera of patients with myositis and systemic lupus erythematosus." In: *The Journal of experimental medicine* 184.4 (1996), pp. 1313–1318.

[279] Carsten Berges et al. "Proteasome inhibition suppresses essential immune functions of human CD4+ T cells". In: *Immunology* 124.2 (2008), pp. 234–246.

[280] JW Van der Heijden et al. "The proteasome inhibitor bortezomib inhibits the release of NFkappaB-inducible cytokines and induces apoptosis of activated T cells from rheumatoid arthritis patients". In: *Clin Exp Rheumatol* 27.1 (2009), pp. 92–8.

[281] Takuma Misawa et al. “Mutual inhibition between Prkd2 and Bcl6 controls T follicular helper cell differentiation”. In: *Science Immunology* 5.43 (2020).

[282] Alba Torres et al. “ETS2 is a prostate basal cell marker and is highly expressed in prostate cancers aberrantly expressing p63”. In: *The Prostate* 78.12 (2018), pp. 896–904.

[283] WD Xi et al. “Bioinformatics analysis of RNA-seq data revealed critical genes in colon adenocarcinoma”. In: *European Review for Medical and Pharmacological Sciences* 21.13 (2017), pp. 3012–3020.

[284] Guang-Wei Zhang et al. “Down-regulation of ETS2 inhibits the invasion and metastasis of renal cell carcinoma cells by inducing EMT via the PI3K/Akt signaling pathway”. In: *Biomedicine & Pharmacotherapy* 104 (2018), pp. 119–126.

[285] Osamu Tetsu and Frank McCormick. “ETS-targeted therapy: can it substitute for MEK inhibitors?” In: *Clinical and translational medicine* 6.1 (2017), pp. 1–9.

[286] Clare Bycroft et al. “The UK Biobank resource with deep phenotyping and genomic data”. In: *Nature* 562.7726 (2018), pp. 203–209.

[287] Gareth-Rhys Jones et al. “Sa1790–Multi-Parameter Data-Sets are Required to Identify the True Prevalence of Ibd: the Lothian IBD Registry (LIBDR)”. In: *Gastroenterology* 156.6 (2019), S–403.

[288] Miles Parkes. “IBD BioResource: an open-access platform of 25 000 patients to accelerate research in Crohn’s and Colitis”. In: *Gut* 68.9 (2019), pp. 1537–1540.

[289] Janine Meienberg et al. “Clinical sequencing: is WGS the better WES?” In: *Human genetics* 135.3 (2016), pp. 359–362.

[290] Eric R Gamazon et al. “A gene-based association method for mapping traits using reference transcriptome data”. In: *Nature genetics* 47.9 (2015), p. 1091.

[291] Alexander Gusev et al. “Integrative approaches for large-scale transcriptome-wide association studies”. In: *Nature genetics* 48.3 (2016), p. 245.

[292] Alvaro N Barbeira et al. “Exploring the phenotypic consequences of tissue specific gene expression variation inferred from GWAS summary statistics”. In: *Nature communications* 9.1 (2018), pp. 1–20.

[293] Michael Wainberg et al. “Opportunities and challenges for transcriptome-wide association studies”. In: *Nature genetics* 51.4 (2019), pp. 592–599.

[294] Chris Wallace. “Eliciting priors and relaxing the single causal variant assumption in colocalisation analyses”. In: *PLoS genetics* 16.4 (2020), e1008720.

[295] Nicholas Mancuso et al. “Probabilistic fine-mapping of transcriptome-wide association studies”. In: *Nature genetics* 51.4 (2019), pp. 675–682.

[296] Hilary K Finucane et al. “Heritability enrichment of specifically expressed genes identifies disease-relevant tissues and cell types”. In: *Nature genetics* 50.4 (2018), pp. 621–629.

[297] Alvaro N Barbeira et al. “Integrating predicted transcriptome from multiple tissues improves association detection”. In: *PLoS genetics* 15.1 (2019), e1007889.

[298] Monique GP van der Wijst et al. “Single-cell RNA sequencing identifies celltype-specific cis-eQTLs and co-expression QTLs”. In: *Nature genetics* 50.4 (2018), pp. 493–497.

[299] Anna SE Cuomo et al. “Single-cell RNA-sequencing of differentiating iPS cells reveals dynamic genetic effects on gene expression”. In: *Nature communications* 11.1 (2020), pp. 1–14.

[300] Abhishek K Sarkar et al. “Discovery and characterization of variance QTLs in human induced pluripotent stem cells”. In: *PLoS genetics* 15.4 (2019).

[301] Hyun Min Kang et al. “Multiplexed droplet single-cell RNA-sequencing using natural genetic variation”. In: *Nature biotechnology* 36.1 (2018), p. 89.

[302] Monique GP van der Wijst et al. “The single-cell eQTLGen consortium”. In: *Elife* 9 (2020), e52155.