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.