Chapter 2

Using immune molecular phenotypes to uncover biological mechanisms of disease-associated genetic loci

Collaboration Note

The custom colocalisation gwas-pw pipeline was designed by Louella Vasquez, which I adapted for my analysis here. Kousik Kundu (Wellcome Sanger Institute) developed custom scripts for visualisation of epigenomic signals such as in Figure 2.10 and performed the analysis with the updated phase 2 of the BLUEPRINT cohort data. Tao Jiang (Department of Public Health and Primary Care) and Klaudia Walter (Wellcome Sanger Institute) advised on the gene-specific regression analyses and Valentina Iotchkova (Weatherall Institute of Molecular Medicine, formerly Wellcome Sanger Institute) on GARFIELD implementation and enrichment analyses. Stephen Watt (Wellcome Sanger Institute) provided the monocyte and neutrophil transcription factor datasets and advised on peak selection and investigation. All other analyses were performed by myself.

We thank all authors and consortia for access to GWAS summary statistics, further details of access can be found in the Methods section and Supplementary information. UKBiLEVE GWAS statistics were obtained through application to the UK Biobank. SLE GWAS was accessed from Immunobase, Coronary artery disease from the CARDIoGRAMplusC4D consortium and Type 2 diabetes summary statistics from the DIAGRAM consortium website. We also thank the International Genomics of Alzheimer's Project (IGAP) for providing summary results data for these analyses (see the Supplementary Information for full information of author contributions and Grants for this study). We also thank the IAMDGC consortium for sharing the summary statistics and effect sizes of advanced age-related macular degeneration GWAS from Fritsche *et al.* (2016). Applications and queries can be made through the following contacts: goncalo@umich.edu, larsf@umich.edu. Further information on AMD statistics: http://amdgenetics.org

2 Using immune molecular phenotypes to uncover biological mechanisms of disease-associated genetic loci

2.1 Introduction

2.1.1 Lessons from genetic and genomic analyses of autoimmune diseases

The study of autoimmune diseases (AID) has generated many important biological insights including demonstrating the central role for the function of multiple immune cell types (Farh et al., 2015, Glinos et al., 2017). Overall, there is a 4.5% prevalence of the 81 identified AID in the general population, which is higher for women (6.4%) than men (2.7%) (Hayter and Cook, 2012). The importance of genetic factors and shared environment is demonstrated by the familial clustering of autoimmune diseases (Gutierrez-Arcelus et al., 2016). Initial linkage studies identified some of these genomic risk regions that had large effect sizes by looking for markers that co-segregated with the disease phenotype. These included the MHC, encoding the major histocompatibility complex, with diseases such as Type 1 diabetes (T1D) (Rich et al., 1984) and systemic lupus erythematosus (SLE) (Gaffney et al., 1998) and the nucleotide-binding oligomerisation domain containing 2 (NOD2) gene with Crohn's disease (Hugot et al., 2001, Gutierrez-Arcelus et al., 2016, de Lange and Barrett, 2015). Strong associations in the MHC region, which contains many immune-related genes, are now well established for a wide range of diseases, such as coeliac disease (CEL), rheumatoid arthritis (RA) and multiple sclerosis (MS) (Sollid et al., 1989, Nepom, 1998, Hollenbach and Oksenberg, 2015). These associations implicate a role for MHC-antigen presentation in triggering the immune response as a general phenomenon in autoimmune disease pathogenesis. Other important genes were also identified through candidate gene studies, which test for association with alleles of genes selected *a priori* (Gutierrez-Arcelus et al., 2016). One example is the *CTLA4* locus, which was associated with T1D and later with autoantibody-positive RA (Nistico et al., 1996, Plenge et al., 2005). *CTLA4* encodes an immunoglobulin superfamily protein expressed on the surface of T helper cells that negatively regulates T cell activation (Nistico et al., 1996, Gutierrez-Arcelus et al., 2016).

The advent of GWAS enabled systematic and unbiased genome-wide searches leading to the identification of hundreds of AID risk loci, many of which are shared between different immune disorders (Gutierrez-Arcelus et al., 2016). The majority of these signals are common (MAF > 5%) with small to moderate effect sizes (OR < 1.6) (Gutierrez-Arcelus et al., 2016). MS risk loci, excluding the MHC region, have odds ratios between 1.1 and 1.6, where an OR of 1 signifies no difference in odds of diseases between cases and controls for that allele (Gutierrez-Arcelus et al., 2016, ImmunoBase, 2017). Such observations supported the common disease-common variant (CDCV) hypothesis, first proposed by Risch and Merikangas in 1996, which suggests that complex disease risk is a result of the accumulation of multiple, low-effect risk factors (Risch and Merikangas, 1996). However, larger sample sizes and higher-powered studies are required to detect rare variants, therefore future efforts may discover that variants with a MAF < 1% also play a role in common diseases (Bomba et al., 2017).

Despite the large number of AID variants now discovered (more than 300 loci (Gutierrez-Arcelus et al., 2016)), a limited degree of the estimated heritability is explained by non-HLA loci (Glinos et al., 2017). Heritability is the proportion of observable phenotypic variation that can be attributed to genetics, which can be estimated from twin or sibling studies (Selmi et al., 2012). A recent GWAS of systemic lupus erythematosus (SLE) including 15,991 controls and 7,219 cases, estimated the heritability explained by 43 identified risk alleles to be 15.3%, with a total estimated heritability of 66% (Bentham et al., 2015). This ‘missing heritability’ may be due to limitations in study power precluding detection of the full effect size and frequency spectrum of variants, particularly rare variants (Vasquez et al., 2016). The combination of multiple studies to increase sample sizes and application of improved imputation methods have been utilised by, for example, the International IBD Genetics Consortium (IIBDGC). These efforts enabled increased loci discovery allowing identification of novel pathways implicated in IBD risk such as cytokine signalling, innate defence and lymphocyte activation (de Lange and Barrett, 2015, Jostins et al., 2012).

Investigation of the biological consequences of known variants has provided important paradigms in the functional interpretation of GWAS SNPs. Haematopoietic cell types have long been known to play key roles in immune responses to infection, homeostatic clearance of cell debris and in regulating the balance between reacting to non-self-antigen; not self-antigen (Vasquez et al., 2016). Genomic data from haematopoietic cells are therefore ideally suited for mechanistic interpretation at AID risk loci. Early studies indicated that AID SNPs affect gene expression in whole blood and PBMCs, for example over half of coeliac GWAS variants were also eQTLs (Dubois et al., 2010, Glinos et al., 2017). These observations have been expanded to a wide-range of AIDs and multiple primary immune cell types and additional regulatory elements such as H3K27ac and TF binding (Farh et al., 2015, Tehranchi et al., 2016, Chen et al., 2016a). RNA splicing can also represent a gene expression-independent regulatory mechanism in genetic disease (Li et al., 2016c, Chen et al., 2016a). Li *et al.* (2016) showed that splicing (s)QTLs independent of eQTLs were enriched in gene bodies, in most cases within the target introns. In addition, the sQTLs were also enriched in AID even when compared to eQTLs (Li et al., 2016c).

Clinical insight can be gleaned from combining GWAS SNPs with immune molecular or functional phenotypes (Barrett et al., 2015). Selection of drug targets based genetic evidence provides promising therapeutic possibilities twice as often as those selected without such prior information (Nelson et al., 2015, Barrett et al., 2015). GWAS of genetic variants pre-determined at birth simulates a randomised clinical trial, where randomisation ensures a balance of all confounders (Evans and Davey Smith, 2015). The advantage of GWAS is that drug administration is not required and individuals have been "exposed" across a lifetime rather than for the length of an RCT (Evans and Davey Smith, 2015, Finan et al., 2017). Integration of LPS-stimulated monocyte eQTL data and GWAS SNPs can aid therapeutic insight demonstrated recently where five IBD risk variants were to found increase gene expression of the integrin genes *ITGA*, *ITGAL*, *ICAM* and *ITGB8* (de Lange et al., 2017). Integrins mediate leucocyte adhesion to inflamed endothelial tissues. Therefore, increased surface levels could contribute to the pro-inflammatory environment observed in IBD patients (de Lange et al., 2017, de Lange and Barrett, 2015). Targeting integrins, for example using monoclonal antibodies, has already shown promising therapeutic results in the context of IBD (de Lange et al., 2017). Clearly, discovering novel common associations and their associated mechanisms can still provide additional clinical insight.

AID GWAS alone have also successfully identified genes and pathways that are current drug targets. For example, the *IL6R* pathway, which contains rheumatoid arthritis risk variants, is targeted by the humanised monoclonal antibody therapy, Tocilizumab (Okada et al., 2014, Law et al., 2014). The same RA GWAS also identified novel risk genes not currently targeted by RA therapies but were used for treating other diseases, offering the potential for the repurposing of licensed drugs (Okada et al., 2014). To further capitalise on the therapeutic potential of GWAS, a new genotyping array was designed to include genes encoding druggable proteins, targets with bioactivity and those with clinical indications of any licensed therapeutics (Finan et al., 2017). GWAS with such an array will enable direct association of variants with druggable genes (Finan et al., 2017).

2.1.2 Expanding the complex disease repertoire for which immune phenotypes can resolve mechanisms

Inflammation has also been shown to be important in disorders not traditionally classified as immune-mediated such as Parkinson's disease (Tufekci et al., 2012) and schizophrenia (Muller et al., 2015), suggesting functional and clinical insight could be gleaned from the application of similar approaches described above. Below, I discuss previous evidence for the role peripheral immune function in the pathogenesis of five diseases, which I focus on in this thesis.

2.1.2.1 Advanced age-related macular degeneration (AMD)

AMD is the leading cause of irreversible blindness later in life in the developed world. AMD affects the central part of the macular and is classified into early, intermediate or advanced based on severity (Pennington and DeAngelis, 2016). The hallmark of AMD is the accumulation of lipid-rich, protein-containing drusen deposits between the retinal pigment epithelium (RPE) and Bruch's membrane (BM) in the retina (Figure 2.1). The RPE forms part of the blood-ocular barrier and performs many important functions including nutrient transport, cytokine release and phagocytosis of fragments released from photoreceptors (Tan et al., 2016). Towards the end stages of the disease, the RPE eventually disintegrates leading to loss of photoreceptors and vision (Figure 2.1). Ordinarily, the RPE and sub-retinal regions are devoid of blood vessels, but in the neovascular ("wet") form of the disease, abnormal growth of blood vessels from the choroid spreads into these regions (Pennington and DeAngelis, 2016) (Figure 2.1). Most current therapies target this growth by inhibiting the angiogenesis-promoting vascular endothelial growth factor A (VEGFA) (Pennington and DeAngelis, 2016). However, disease progression continues for most patients, requiring further treatment.

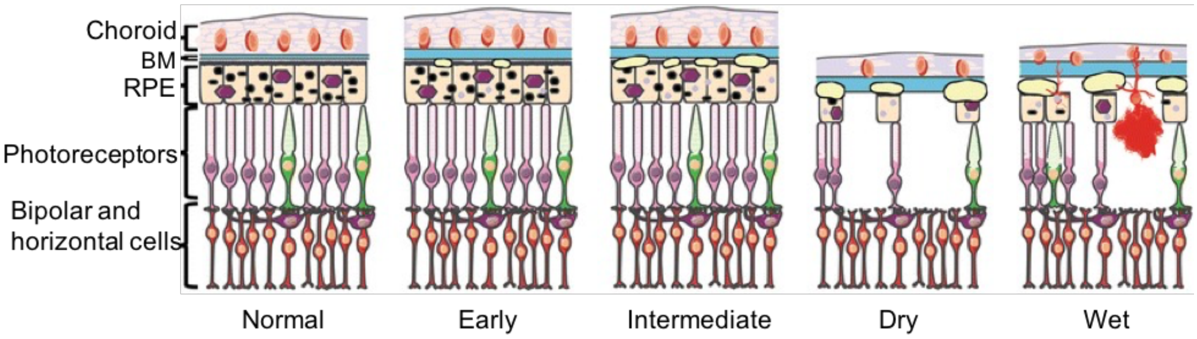


Figure 2.1: Schematic of retinal structure and the effect of AMD pathology
This schematic shows the outer layers of the central retina in normal conditions (left) and the different types of AMD classifications. Small drusen deposits accumulate in the retinal pigment epithelium (RPE) in early AMD, and as the disease progresses, the Bruch's membrane (BM) becomes thicker and additional drusen deposits form. In the later stages of AMD, dry and wet, there is extensive accumulation of drusen deposits, loss of photoreceptors and damage to RPE integrity. The subretinal space refers to the space between the RPE and photoreceptors. In the wet form, choroidal neovascularisation (CNV) occurs. This figure was adapted from (Tan et al., 2016) under the CC license (http://creativecommons.org/licenses/by/4.0/).

AMD is multifactorial with a substantial genetic component, although environmental factors such as smoking and age also contribute (Seddon et al., 2005). Genetic studies have revealed many associated loci and have provided evidence for the involvement of immune components. For example, almost 60% of AMD risk can be explained by variants located near the complement genes *CFH*, *C2/CFB*, *C3*, *CFI* and *C9* (Tan et al., 2016). The risk to AMD also increases with rare alleles. For example, the highly penetrant missense mutation in the complement factor I gene (*CFI*), which functions to inactivate complement pathways, corresponds to an odds ratio of 22.20 (95% CI = 2.98-164.49) (van de Ven et al., 2013, Tan et al., 2016). The p.Gly119Arg mutant protein is expressed and secreted at lower levels (van de Ven et al., 2013).

A recent large-scale GWAS from the International AMD Genomics Consortium (IAMDGC) that included 16,144 advanced AMD patients and 17,832 controls identified 52 independently associated variants in 34 loci (Fritsche et al., 2016). This study confirmed the involvement of inflammation and complement genes (*VTN, CFH, C2/CFH, C3, CFI, C9*) as well as lipid pathway genes (*CETP, LIPC, APOE, ABCA1)* (Fritsche et al., 2016). The complement pathway is an important part of innate immunity that functions to amplify immune responses ultimately resulting in the formation of a membrane attack complex (MAC) at the surface of the pathogen causing cell lysis (Tan et al., 2016). Increased MAC in the retina and drusen has been observed in AMD patients (Tan et al., 2016, Hageman et al., 2001). Although the majority of the complement system is synthesised by the liver, there is also synthesis in the RPE and choroid (Tan et al., 2016, Luo et al., 2013).

Despite initially being considered “immunologically privileged”, the RPE has been shown to contain specialised resident immunocompetent cells including microglia, dendritic cells and perivascular macrophages (Parmeggiani et al., 2012). Para-inflammatory (prolonged inflammation in response to damage)-associated modifications can result in damage to the blood-retinal barrier in AMD as well as microglial activation and recruitment of macrophages (Parmeggiani et al., 2012, Kauppinen et al., 2016). Indeed, immunocompetent cells such as lymphocytes and macrophages have been observed in AMD retinal tissues and isolated mouse bone marrow-derived M1 and M2b macrophages stimulated RPE cells to induce inflammatory cytokine expression and complement factors C3 and CFB (Parmeggiani et al., 2012, Lopez et al., 1991, Luo et al., 2013).

Systemic immune alterations such as increased serum complement components and pro-inflammatory cytokines (IL-1α, IL-1β and IL-17) have been observed in AMD patients (Lechner et al., 2015). Higher numbers of circulating neutrophils were observed in neovascular AMD (nvAMD) patients (Lechner et al., 2015). Also, similarly in nvAMD patients, there was an increased inflammatory transcriptome signature in peripheral blood monocytes, and in an independent study monocytes expressed higher levels of chemokine receptors CCR1, CCR2 and CX3CR1, HLA-DR and phosphorylated STAT3 (Grunin et al., 2016, Grunin et al., 2012, Chen et al., 2016b). Recently, in a study of 161 nvAMD patients and 43 controls, stimulated PBMCs, particularly monocytes, secreted higher levels of the IL8, CCL2 and VEGF compared to controls (Lechner et al., 2017). The pro-inflammatory IL8 and CCL2 promote the recruitment of neutrophils and monocytes and lymphocytes respectively (Lechner et al., 2017). Additional molecular mechanistic insight into monocyte-RPE interactions was shown using a coculture of human CD14$^+$ blood monocytes and primary porcine RPE cells (Mathis et al., 2017). OTX2 is a key TF regulating retinal genes such as retinol dehydrogenase 5 (RDH5), which re-isomerises all-trans-retinal into 11-cis-retinal. It was shown that TNFα, secreted from activated monocytes, mediates the downregulation of *OTX2* and *RDH5* (Mathis et al., 2017).

In summary, immune dysregulation in AMD could be the result of the combined action of resident and infiltrating immune cells as a result of a switch from clearance (of RPE-debris) and immunosuppressive environment to a proinflammatory milieu (Nussenblatt and Ferris, 2007). Genetic variants could disrupt this balance, thereby influencing risk. Definitive assessment of the causality of peripheral immune factors and disease pathogenesis is required, but the observations of systemic immune activation in AMD patients suggests that using the more accessible peripheral immune cells to identify functional mechanisms may provide further insight into the pathogenic process. Some success, for example, in an early pilot phase I/II randomized study of suppression of systemic immune activity, was observed for wet AMD (Nussenblatt et al., 2010). However, other attempts, including the use of eculizumab to inhibit complement have been less successful (Yehoshua et al., 2014). Elucidating the exact mechanisms of peripheral immune involvement may help these efforts.

2.1.2.2 Coronary artery disease (CAD)

Coronary artery disease (CAD) is the most common type of heart disease (Khera and Kathiresan, 2017). Familial history and therefore, a genetic component, has been implicated as an important risk factor (Framingham Heart Study (Watkins and Farrall, 2006), PROCAM study (Assmann et al., 2002), INTERHEART study (Yusuf et al., 2004)). Common CAD is multifactorial and polygenic (Won et al., 2015). Low-density lipoprotein (LDL) cholesterol levels, lipoprotein(a) and BMI have also been demonstrated to be causal risk factors and are themselves under genetic control (Watkins and Farrall, 2006, Do et al., 2013, Clarke et al., 2009, Voight et al., 2012, Khera and Kathiresan, 2017). The INTERHEART study showed that adjustment for the known classical risk factors only marginally reduced the risk (OR from 1.55 to 1.45), suggesting there are other genetic factors involved (Yusuf et al., 2004, Watkins and Farrall, 2006).

There is a well-established causal association of the inflammatory IL6 cytokine pathway with CAD risk (Interleukin-6 Receptor Mendelian Randomisation Analysis Consortium et al., 2012, Il R. Genetics Consortium Emerging Risk Factors Collaboration et al., 2012). The missense SNP, rs2228145, increases the soluble form of the IL6 receptor by improving the efficiency of membrane-bound receptor proteolytic cleavage while also increasing the unique transcript encoding the soluble receptor (Ferreira et al., 2013). Reduced IL6R membrane expression on monocytes and CD4$^+$ T cells results in impaired signalling and IL6 response and reduces CAD risk (also that of RA), clearly demonstrating the pathogenic role of inflammation in disease progression (Ferreira et al., 2013).

Monocytes play a key role in pathological plaque deposition in the coronary arteries, known as atherosclerosis (Ghattas et al., 2013). After recruitment to atherosclerotic lesions, monocytes mature into macrophages (Meeuwsen et al., 2017). Phagocytosis of oxidised LDL stimulates macrophage to form foam cells, which are a constituent of atherosclerotic plaques (Meeuwsen et al., 2017). Recruitment of other immune cells such as neutrophils, mast cells and lymphocytes also contributes to plaque destabilisation eventually leading to plaque rupture (Meeuwsen et al., 2017).

Supporting the importance of leukocyte function, the recent UK Biobank CAD GWAS with 4,831 cases and 115,455 controls identified 15 novel associations including the *ARHGEF26* locus, which is involved in transendothelial migration of leukocytes and encodes the rho guanine nucleotide exchange factor 26 (Klarin et al., 2017). The novel *ARHGEF26* locus was not associated with established risk factors, but previous mouse work did demonstrate a role in atherosclerosis (Klarin et al., 2017). Endogenous siRNA-mediated *ARHGEF26* knock-down decreased leukocyte adhesion to endothelial cells and transendothelial migration (Klarin et al., 2017). Overexpression of the exogenous mutant (Leu29) not only rescued the phenotype but was increased compared to wild-type, which is consistent with a gain of function ARHGEF26 effect associated with rs12493885 (Val29Leu) and increased CAD risk (Klarin et al., 2017). These observations are evidence that dysregulation of leukocyte function is associated with risk of CAD and disease prognosis.

Understanding the contribution of immune processes to CAD risk has important clinical implications. Currently, clinical management of CAD-associated events has improved leading to more than a 50% decrease in age-adjusted mortality rate in the United States (Khera and Kathiresan, 2017). Despite this, CAD is still the biggest cause of death worldwide, and there is a 12% mortality rate within six months of the first coronary event (Meeuwsen et al., 2017). Many available therapies target lipid and thrombosis reduction, but some inflammation-modulating processes are being investigated (Fernandez-Ruiz, 2016). Promising results were reported from a recent clinical trial targeting inflammation called the Canakinumab Antiinflammatory Thrombosis Outcome Study (CANTOS) (Ridker et al., 2017, Harrington, 2017, Couzin-Frankel, 2017). The trial included more than 10,000 heart attack patients with elevated levels of C-reactive protein (CRP). The drug tested, canakinumab, is a human monoclonal antibody that targets the inflammatory cytokine interleukin-1β and is already used in the treatment of systemic juvenile idiopathic arthritis (Harrington, 2017). For patients receiving canakinumab as four infusions each year over 3.5 years, the risk of a second cardiovascular event decreased from 4.5% to 3.86% and the likelihood of angioplasty or cardiac bypass surgery decreased by 30% (Couzin-Frankel, 2017). However, an increased number of deaths from infection among patients who received more doses was observed (Harrington, 2017). While promising and offering a proof of concept for inflammation playing a key role in disease pathogenesis, further research is needed to provide more targeted immune therapeutics which may reduce the risk of mortality related to infection (Harrington, 2017, Couzin-Frankel, 2017). Providing detailed descriptions of the pathways and cell types which may be involved in CAD risk could, therefore, represent an early step in improving therapeutic options for CAD patients.

2.1.2.3 Alzheimer’s disease (AD)

Alzheimer’s disease is the most common form of neurodegenerative dementia characterised by accumulation in the brain of amyloid β (Aβ) plaques and hyper-phosphorylated tau protein aggregates (Pimenova et al., 2017). This chapter focuses on late-onset Alzheimer’s disease (LOAD), which constitutes 99% of AD cases. LOAD is multifactorial with a strong but highly complex genetic component (Pimenova et al., 2017, Gatz et al., 2006).

Involvement of immune components was highlighted in a large-scale meta-analysis of 74,046 individuals (Lambert et al., 2013). For example, a significant intronic SNP was identified in complement factor 1 (CR1). CR1 is expressed on blood cells and specialised brain-resident immune cells known as microglia (Pimenova et al., 2017, Wyss-Coray and Rogers). Other immune loci identified include *ZCWPW1/PILRA/PILRB,* which are monocyte and neutrophil immune infiltration receptors, the *HLA-DRB1/HLA-DRB5* locus of the MHCII region, *CD33* and the *MS4A* gene family, which are both expressed on microglia and myeloid cells (Pimenova et al., 2017, Lambert et al., 2013).

Brain-resident microglia are active immune cells, but there may also be a role for systemic immune responses in disease pathogenesis (Heneka et al., 2015, Czirr and Wyss-Coray, 2012). There are multiple lines of evidence that show that systemic inflammation detrimentally affects brain function and contributes to AD progression (Heneka et al., 2015). Prolonged LPS challenge in amyloid precursor protein (APP) transgenic mice, which contain a mutant APP associated with familiar forms of human AD, has been shown to result in cognitive impairment through increased amyloid deposition (Lee et al., 2008, Czirr and Wyss-Coray, 2012). Increased cognitive decline was observed in AD patients with infection, which correlated with infection TNF levels (Heneka et al., 2015). However, the functionality of peripheral immune cells in AD pathogenesis and their interaction with the brain are not yet clear. There is some evidence that the integrity of the blood-brain barrier (BBB) can be compromised in AD, which could allow infiltration of peripheral monocytes (Zenaro et al., 2017, Heneka et al., 2015). Compelling evidence from both mouse models and from human brain tissue showed that neutrophils could migrate to the central nervous system, which was dependent on the LFA-1 integrin, in turn, triggered by the $A\beta_{42}$ peptide (Zenaro et al., 2015). Depletion of neutrophils resulted in memory improvements in mice. This study suggests that neutrophils could contribute to inflammatory conditions in AD and also damage the BBB (Zenaro et al., 2015).

Integration with genomics data from the Immune Variation project with AD risk loci demonstrated enrichment among monocyte eQTLs, but not T cells, suggesting using monocyte data may help dissect mechanism underlying genetic susceptibility (Raj et al., 2014). Further demonstration of the importance of myeloid cells came from a recent association analysis of age of onset of Alzheimer's disease-defined survival that identified protective rs1057233 (G) associated with reduced *SPI1* expression in human myeloid cells (Huang et al., 2017c). *SPI1* encodes the myeloid master transcription factor, PU.1, which is critical for myeloid lineage differentiation and function (Huang et al., 2017c). PU.1 levels correlated with mouse microglial phagocytic activity and reduced *SPI1* expression corresponded to reduced AD risk (Huang et al., 2017c). Global enrichment of AD heritability in myeloid and B-lymphoid H3K4me1-designated enhancers and in the PU.1 cistrome (i.e. genome-wide binding regions as assayed by ChIP-seq) was also demonstrated (Huang et al., 2017c). Circulating blood monocytes are easily accessed in high numbers compared to brain-derived tissue and share transcriptional patterns with brain-resident microglia (Raj et al., 2014). Therefore, although there is evidence for a role of peripheral immune cells in AD, these cells are also useful as highly related proxies for specialised brain cells (Raj et al., 2014, Proitsi et al., 2014).

Taken together these observations provide evidence for the role of immune cells in AD pathogenesis, suggesting that integrative genomic approaches using immune data may provide insight into pathogenic mechanisms.

2.1.2.4 Lung function and chronic pulmonary obstructive disease (COPD)

Chronic obstructive pulmonary disease (COPD) is an inflammatory airway disease and is the third leading cause of global mortality (Wain et al., 2015). The ratio of two lung measurements, FEV_1/FVC, is used to evaluate airflow obstruction and diagnose COPD. FEV_1 measures how much air is exhaled in one second after maximal inhalation. FVC measures the volume of air exhaled after a maximal inhalation and then a six-second maximal forced exhalation (Weiss, 2010).

Smoking and indoor pollution are major risk factors for disease pathogenesis, but there is also a strong genetic component underlying smoking behaviour and disease risk (Wain et al., 2015, Hukkinen et al., 2011). Evaluation of these lung ratios as quantitative traits has allowed the identification of genetic determinants of lung function in large populations, enabling greater power to detect associations. Ten loci were found to be associated with extremes of the lung ratio FEV_1 in never smokers, and one was located in the MHC region, *HLA-DQB1/HLA-DQA2* (Wain et al., 2015). This important immune locus potentially implicates an immunological role in extremes of lung function and as this locus was also associated with COPD also highlights immune factors in this disorder.

A hallmark of COPD is chronic inflammation but whether this and the role of immune cells are causal independent risk factors for COPD remains to be definitively demonstrated (Brusselle et al., 2011). Cigarette smoke activates innate immune cells, which then stimulate an adaptive immune response (Brusselle et al., 2011). Viral and bacterial infections exacerbate COPD and potentiate the inflammatory environment in the lung (Brusselle et al., 2011). Dysregulation of peripheral immune cells could contribute to the pathogenesis of the disease, either due to the extrapulmonary effects or through infiltration into the lung (Brusselle et al., 2011). Neutrophil infiltration has been linked to tissue destruction and disease progression in COPD patients (Huang et al., 2017a). Understanding potential immune-mediated molecular disruptions associated with the disease can help the design of treatments with the aim to reduce serious symptoms and complications (Brusselle et al., 2011).

2.1.3 Identification of disease-relevant cell types using enrichment approaches

Disease-relevant cell types are not always known and have traditionally been identified through immunology studies, patient observations and mouse models (Glinos et al., 2017). For example, inflammatory cells were identified in the brain lesions of patients and autoreactive T cells responding to myelin antigens were identified in the mouse model for multiple sclerosis (MS), experimental autoimmune encephalomyelitis (EAE) (Fletcher et al., 2010).

With the advent of high-throughput genetic association studies and increased availability of cell-type specific epigenome data from consortia such as the Encode Project Consortium (2012), Roadmap Epigenomics Consortium et al. (2015) and IHEC (Stunnenberg et al., 2016), novel statistical enrichment approaches were developed to assess genomic evidence for the relevance of cell types in complex disease and annotate the putative function of non-coding GWAS variants. These approaches evaluate the statistical significance of a quantified overlap between GWAS SNPs and regulatory annotations. Annotations can take of the form of ChIP-seq binding/signal regions, open chromatin regions or chromatin regulatory states all denoted with the genomic start and end positions of the mapped genomic region (referred to as a peak). An assessment of the significance of enrichment is required given that a large degree of the genome can be bound by these regulatory features, leading to spurious functional assignment occurring by chance if only a simple overlap is applied (Iotchkova et al., 2016).

Earlier functional enrichment approaches, for example as implemented by Maurano *et al.* (2012) demonstrated an enrichment (40%) of significant GWAS variants from 207 diseases and 447 quantitative traits in open chromatin (DHS sites) assayed in 349 tissues, compared to frequency and genomic-location matched SNPs from 1000 Genomes Project (Maurano et al., 2012). This high enrichment highlighted that underlying regulatory function of non-coding SNPs, demarcated by open chromatin, could underpin many complex trait associations. In addition, the authors demonstrated the ability of functional enrichment studies in the *de novo* identification of relevant cell types by comparing the enrichment of progressively more significant trait-associated variants in cell-type specific DHS sites to the proportion of all SNPs from the summary statistics that also overlapped the same DHSs. Higher enrichment for Crohn's disease variants was observed in Th17, and Th1 T cell open chromatin than other immune cell types and multiple sclerosis variants were enriched in B and T cell open chromatin (Maurano et al., 2012). The basis of predicting disease relevant cell types lies in the known cell specificity of regulatory element activity. Therefore, preferential enrichment in regulatory data from certain cell types suggests these variants are more likely functional in those cell types. This approach enables efficient identification of relevant cell types without complex patient, animal or *in vitro* studies by leveraging known disease risk loci and experimentally derived epigenomic data.

However, the above method did not account for linkage disequilibrium (LD) between significant variants (Iotchkova et al., 2016). Chance overlaps are also more likely to occur in regions of highly correlated variants, where high LD between variants can lead to overlaps that are not truly functional (Alasoo et al., 2017). This can inflate enrichment values and generate false positives. Later methods accounting for variant correlation made other interesting observations, that enrichment of disease SNPs in cell specific H3K4me3 regions associated with active transcription was not driven by gene proximity, and could, therefore, highlight potentially causal variants (Trynka et al., 2013). Enrichment of trait-associated SNPs in cell type-specific DHSs was confirmed by an independent method, fgwas, and also observed depletion in repressed chromatin (Pickrell, 2014). Predicted causal SNPs of a variety of autoimmune diseases were shown to be enriched in cell type specific H3K27ac enhancer regions and genetic heritability is also highly enriched in regulatory regions (Farh et al., 2015, Finucane et al., 2015). Collectively, these enrichments suggest that a high proportion of disease risk is mediated by regulatory changes that could ultimately generate variation in gene expression (Chun et al., 2017).

Evidence of enrichment below the genome-wide significance threshold (p values $\leq 5 \times 10^{-08}$) has been observed (Maurano et al., 2012), suggesting that appropriately evaluating enrichment at genome-wide suggestive thresholds (commonly p value $\leq 1 \times 10^{-05}$) could provide biologically relevant observations particularly when limited power precludes discovery of all true associations (Maurano et al., 2012). To address the potential confounding issues and to provide a robust assessment of enrichment at all GWAS significance thresholds, a novel method was developed within the Soranzo team by Valentina Iotchkova, known as GARFIELD (GWAS Analysis of Regulatory or Functional Information Enrichment with LD correction) (Iotchkova et al., 2016) (described in detail in

Materials and Methods). GARFIELD annotates independent variants, accounting for LD as well as for genomic location and minor allele frequency. The odds ratio is calculated for multiple GWAS thresholds, allowing for assessment of informative enrichment at suggestive thresholds, which can highlight novel findings given limited study power. The method confirmed enrichment of Crohn's disease variants in blood DHS (Iotchkova et al., 2016) and demonstrated highly significant enrichment of blood cell indices in corresponding cell-type specific enhancers (Astle et al., 2016).

Recently, the application of the GARFIELD method has been extended to assess the enrichment of GWAS SNPs in molecular QTLs where QTLs were used as regulatory annotations. This approach demonstrated enrichment of neutrophil, monocyte and T cell molecular features in autoimmune diseases including IBD, RA, T1D and MS (Chen et al., 2016a). An independent study of naïve and stimulated iPSC-derived macrophage eQTL and chromatin accessibility QTLs also used GARFIELD to reveal enrichment in autoimmune diseases, Alzheimer's disease and schizophrenia (Alasoo et al., 2017).

2.1.4 Colocalisation methods evaluate shared genetics across different traits

Enrichment methods do not assess whether variation in regulatory function and disease risk or trait variance can be attributed to a single shared variant or whether these are driven by independent effects in the same locus, known as pleiotropy (Chun et al., 2017). Therefore, once enrichment approaches have highlighted relevant cell types, robust approaches are required to identify loci where there is evidence for shared genetic control of multiple traits. This can be achieved by using Bayesian colocalisation methods (described in detail in Materials and Methods) (Pickrell et al., 2016, Guo et al., 2015).

Robust evaluations are particularly important given that eQTLs are widely present across the genome, and the majority of expressed genes are likely to be associated with at least one cis-eQTL (Pai et al., 2015). Guo *et al.* (2015) developed and implemented a Bayesian colocalisation method (*coloc*) using ten immune disease-associated variants (595 in total from 154 regions) and eQTLs from resting and stimulated monocytes and naïve B cells (Guo et al., 2015). They identified 125 eQTLs that overlapped with disease SNPs, but only six genes had evidence for colocalisation (two traits share the same causal variant). A higher proportion, 21% across all AID loci tested colocalised with LCLs, CD4$^+$ T cells and CD14$^+$ monocytes eQTLs (FDR < 5%), identified using an independent method (Chun et al., 2017). For some diseases, this proportion was higher, for example 60% for ulcerative colitis loci colocalised with at least one eQTL (Chun et al., 2017). Overall, across all diseases, 75% of the disease-eQTL pairs were identified as pleiotropic where independent genetic variants within the locus were associated with each trait (Chun et al., 2017). Therefore, the previous usage of colocalisation methods has demonstrated the need for appropriate evaluation of shared trait and QTL loci, demonstrating that simple overlap approaches are prone to false positives.

2.1.5 Aims of this chapter

We previously demonstrated the enrichment of molecular phenotypes in AID variants and described many disease loci where there is evidence of shared molecular mechanisms (Chen et al., 2016a). It is therefore well established that while GWAS provide unbiased systematic identification of disease-associated loci, functional insight must be gleaned through the combination of intermediate phenotypes QTLs. Detailed “omic” data collected from healthy individuals can be used as a powerful tool in understanding disease mechanisms, as these cohorts enable association of variants with intermediate phenotypes which themselves have not been affected by disease status. This limits the confounding factors and possible reverse causation whereby a dysregulated disease state could cause molecular changes rather than those changes being risk factors for the disease. The combination of genomics and functional experiments have also demonstrated how the disruption in the function of peripheral immune cells can contribute to the pathogenesis of a wide-range of diseases.

To expand this analysis, I used the same molecular QTLs from the BLUEPRINT project, but applied them to a collection of four non-autoimmune disorders and one prototypic AID, SLE (Chen et al., 2016a). I used the GARFIELD method to evaluate significant enrichment of GWAS variants in immune molecular QTLs. Following this, I implemented statistical colocalisation methods using gene expression, histone and splicing-associated QTLs in neutrophils, monocytes and T cells with GWAS SNPs. I then further evaluated whether there is evidence that immune phenotypes can, at least in part, aid the development of mechanistic hypotheses underpinning genetic risk at disease loci. I also thoroughly investigated biological mechanisms to provide functional hypotheses that will aid the design of further experimental dissection of disease loci. I demonstrated that this approach required the integration of multiple data sources and analytical approaches in order to provide in-depth molecular insight into a specific disease locus.

2.2 Materials and methods

Quantitative trait loci (QTL) data: Molecular phenotypes from the BLUEPRINT study were used to assign function to disease loci (Chen et al., 2016a). The study design, summarised in Figure 2.2, included generation of a mean read depth of ~7X-whole-genome sequences, transcriptomes (RNA-seq), histone ChIP-seq data (H3K4me1, H3K27ac) and DNA methylation probes (450K array) in population sample of up to 197 healthy individuals. These data were collected in three cell types, CD66b$^+$CD16$^+$ neutrophils, CD14$^+$CD16$^-$ monocytes and CD4$^+$CD45RA$^+$ T cells.

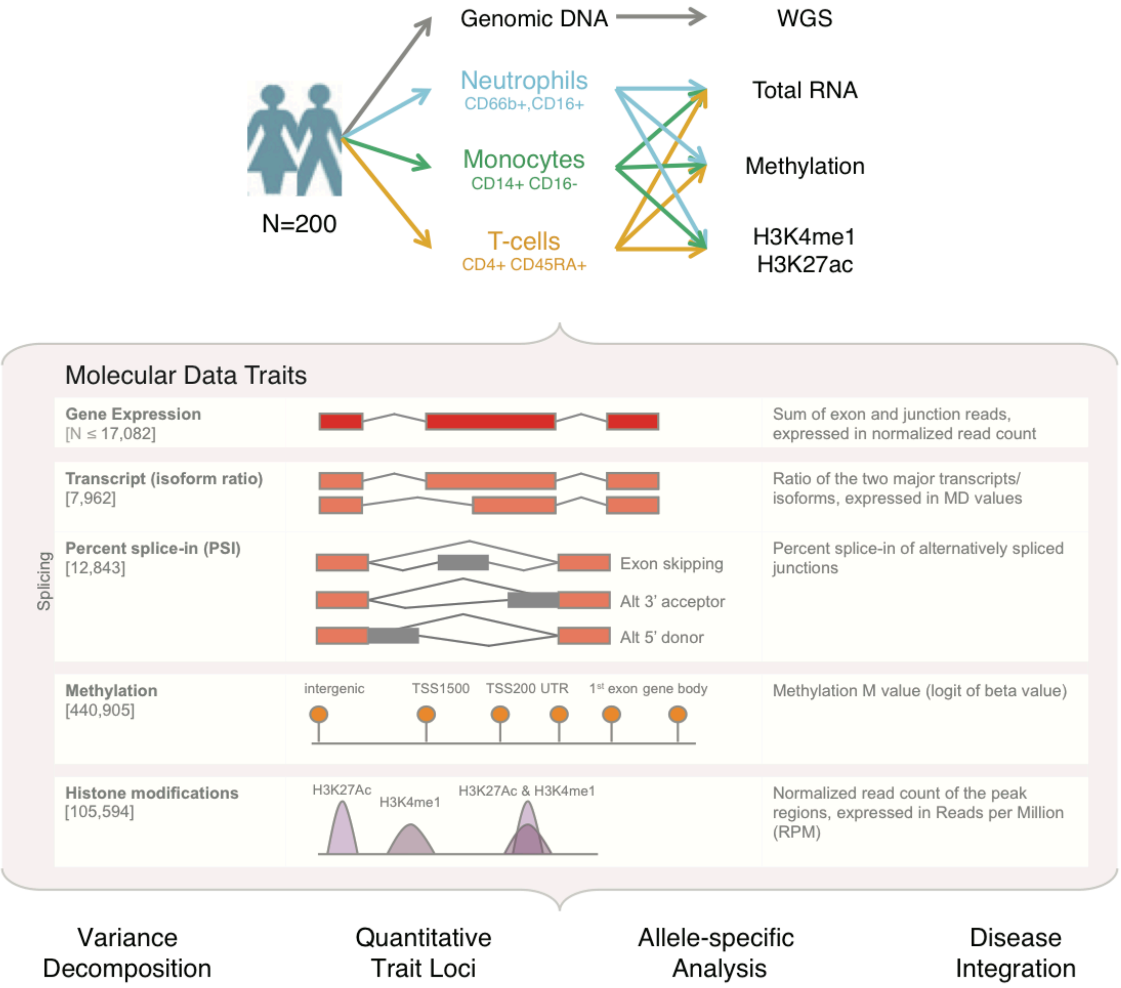


Figure 2.2: Summary of the BLUEPRINT Epigenome variation project
Overview of the study design and molecular traits investigated in BLUEPRINT. Figure reproduced from (Chen et al., 2016a) under the CC license (http://creativecommons.org/licenses/by/4.0/).

For investigating genetic functional mechanisms in this chapter, I focused on gene expression, splicing events (exon skipping or alternative acceptors or donors, Figure 2.2), H3K27ac and H3K4me1 QTLs from all three cell types. I also evaluated disease variant enrichment in methylation QTLs. Full analysis methods are detailed in Chen *et al.* (2016) but briefly, phenotype values of histone signal from ChIP-seq and expression from RNA-seq data were corrected for sequencing centre batch effects. For ChIP-seq data, phenotypes were normalised by the total number of million reads per individual (RPM) and normalised by taking log2. Gene expression was normalised in DESEQ and expressed in units proportional to $\log_2$FPKM, corrected for gene length and sequencing depth. These phenotype values were used for the whole cohort for visualisation and further testing described in this chapter. For QTL association, unknown sources of non-genetic covariation were removed by correcting for the first ten PEER factors (Chen et al., 2016a). Each feature was tested for association with variants within the feature and 1Mb upstream and downstream of the start and end position. The p value of association was corrected for increased false positives due to testing multiple variants by calculating the qvalue. This method controls for the false discovery rate, which is the proportion of false positives generated by testing multiple hypotheses (Bass JDSwcfAJ, 2015). Variants with a qvalue less than 5% were designated significantly associated with the phenotype. For visualising modified histone regions with more resolution, as in Figure 2.10 and 2.17, aligned reads expressed in $\log_2$RPM were used for gene expression and histone signals and calculated across 50 bp non-overlapping sliding windows across the genome.

Disease GWAS datasets: To perform colocalisation methods, I collected summary statistics from GWAS with European cohorts of relatively high sample sizes and power. All GWAS were annotated with the genome build hg19. Z-score was calculated using the supplied effect size estimate divided by the standard error. Advanced age-related macular degeneration summary statistics including beta and standard error estimates were provided by kind permission of the International AMD Genomic Consortium (Fritsche et al., 2016). The study consisted of 12,023,830 variants, 16,144 cases and 17,832 controls. Alzheimer's disease summary statistics were obtained from the International Genomics of Alzheimer's Project (IGAP) stage 1 meta-analysis from 2013 (Lambert et al., 2013). Stage 1 of the study consisted of 7,055,881 SNPs, 17,008 cases and 37,154 controls. Coronary artery disease (CAD) summary statistics were obtained from the CARDIoGRAMplusC4D consortium website from 2015 study (Nikpay et al., 2015). The study consisted of 9,455,778 variants, 60,801 CAD cases and 123,504 controls and the additive model associations were used in this chapter. Summary statistics for systemic lupus erythematosus were obtained from the immunobase website and accessed in October 2016 (ImmunoBase, 2017). The new SLE GWAS consisted of 7,219 cases and 15,991 European controls (Bentham et al., 2015). The summary statistics of extremes of the lung ratio, FEV_1, from never smokers were obtained from the UKBiLEVE project as part of the UK Biobank access (Wain et al., 2015). The initial study included 50,008 individuals, which were further stratified into never or heavy smokers and further into range of FEV_1, for example, 9745 never smokers with low FEV_1, 9827 never smokers with average FEV_1 and 4902 never smokers with high FEV_1 (Wain et al., 2015). Type 2 diabetes summary statistics (Morris et al., 2012) were also used to compare enrichments of other traits, these were accessed from the DIAGRAM consortium website (Stage 1 GWAS) in 2016. Further details of the GWAS summary statistics and website links are detailed in the Supplementary Information.

GWAS enrichment (GARFIELD): GARFIELD (Iotchkova et al., 2016), was implemented to assess significant enrichment of GWAS SNPs with molecular QTLs. In this thesis, the version of the method utilises genome-wide summary statistics to calculate odds ratios for association between an overlap (SNP-QTL annotation) and disease status (significant p-value for the disease/complex trait), as was used in the Astle *et al.* (2016) and Chen *et al.* (2016) studies. The QTL summary statistics are formatted into annotations by selecting all significant QTLs (qvalue < 5%) for each QTL type and each cell type. Where there are duplicated QTLs, due to association with multiple features, the lowest p value is used. The method greedily prunes input GWAS disease/trait-associated variants, retaining the most significant variant and removing variants with LD $r^2 \geq 0.1$. LD tags are pre-calculated using 1Mb windows and the combined UK10K and 1000 genomes Phase 3 panel (Europeans). The process is repeated until no significant variants remain. Independent variants are then overlapped with the annotations of interest by matching genomic positions. SNPs in high LD with the independent variant are also annotated as overlaps ($r^2 \geq 0.8$). Odds ratios are calculated at various GWAS thresholds (1×10^{-08}, 1×10^{-07}, 1×10^{-06}, 1×10^{-05}, 1×10^{-04}, 1×10^{-03}, 1×10^{-02}, 1×10^{-01}, 1). The significance of the odds ratio at each GWAS threshold is calculated using a generalized linear model in a logistic regression approach that controls for LD (number of variant proxies), minor allele frequency and local gene density (variant distance to the TSS) input as categorical variables. The method corrects for multiple annotations tested by applying the Bonferroni correction using the effective number of annotations, which is calculated based on the correlation between annotation-SNP overlap matrices. In this thesis, there was an increased representation of rare variants in the specially designed respiratory array used in the FEV_1 UKBiLeve summary statistics and rare variants were also included in the AMD and CAD studies. In the BLUEPRINT molecular data, the analysis focused on common variants (MAF ≥ 1%). Therefore, variants with a MAF < 1% were filtered from these summary statistics before evaluating enrichment.

Generation of regions for colocalisation input: I performed a locus pre-selection step to test for colocalisation as assessing all regions across the whole genome would constitute a high computational burden given the vast number of QTL associations. I generated a list of regions where significant disease SNPs and index molecular QTLs overlapped, based on the hypothesis that genomic regions associated with molecular feature(s) and a complex trait were more likely to share underlying genetic causes. Specifically, starting from the BLUEPRINT QTL summary statistics, significant QTLs (qvalue $\leq$ 5%) were selected. For each disease-QTL combination, overlaps were annotated if any lead QTL per feature (gene, splicing, methylation probe, H3K27ac or H3K4me1 peak) and proxy variants in high LD ($r^2 \geq$ 0.8), were also significant in the GWAS summary statistics (p value $\leq 5 \times 10^{-08}$). If the overlap occurred with lead QTL proxies, the corresponding lead QTL information was retained. For each lead QTL, all features associated with the unique lead QTLs that overlapped were identified. These were referred to as feature-QTL pairs and represent different regions tested for colocalisation (Figure 2.3). For each pair, the genomic region assessed in colocalisation was defined by the BLUEPRINT QTL testing region (SNPs within the feature and 1 Mb upstream and downstream). Only variants that overlap between the QTL and disease study are evaluated in colocalisation.

Colocalisation of disease and molecular trait-associated loci: To perform colocalisation, I implemented the method, gwas-pw (Pickrell et al., 2016). This is a Bayesian method that assesses whether the overlap observed between a GWAS SNP and molecular trait QTL is due to a sharing of the genetic effect, i.e. the same genetic variant is associated with both the GWAS trait and the molecular trait. The method estimates a probability that the association evidence in a given genomic region falls into one of four models. Model 1 and 2 indicate that the locus contains a single variant associated with the first trait or the second trait only, i.e. there is one association. Model 3 indicates colocalisation where a single variant is associated with both traits. Finally, model 4 indicates that two independent associations exist, where the variant is associated with the first trait and a second, independent variant is associated with the second trait (Figure 2.4).

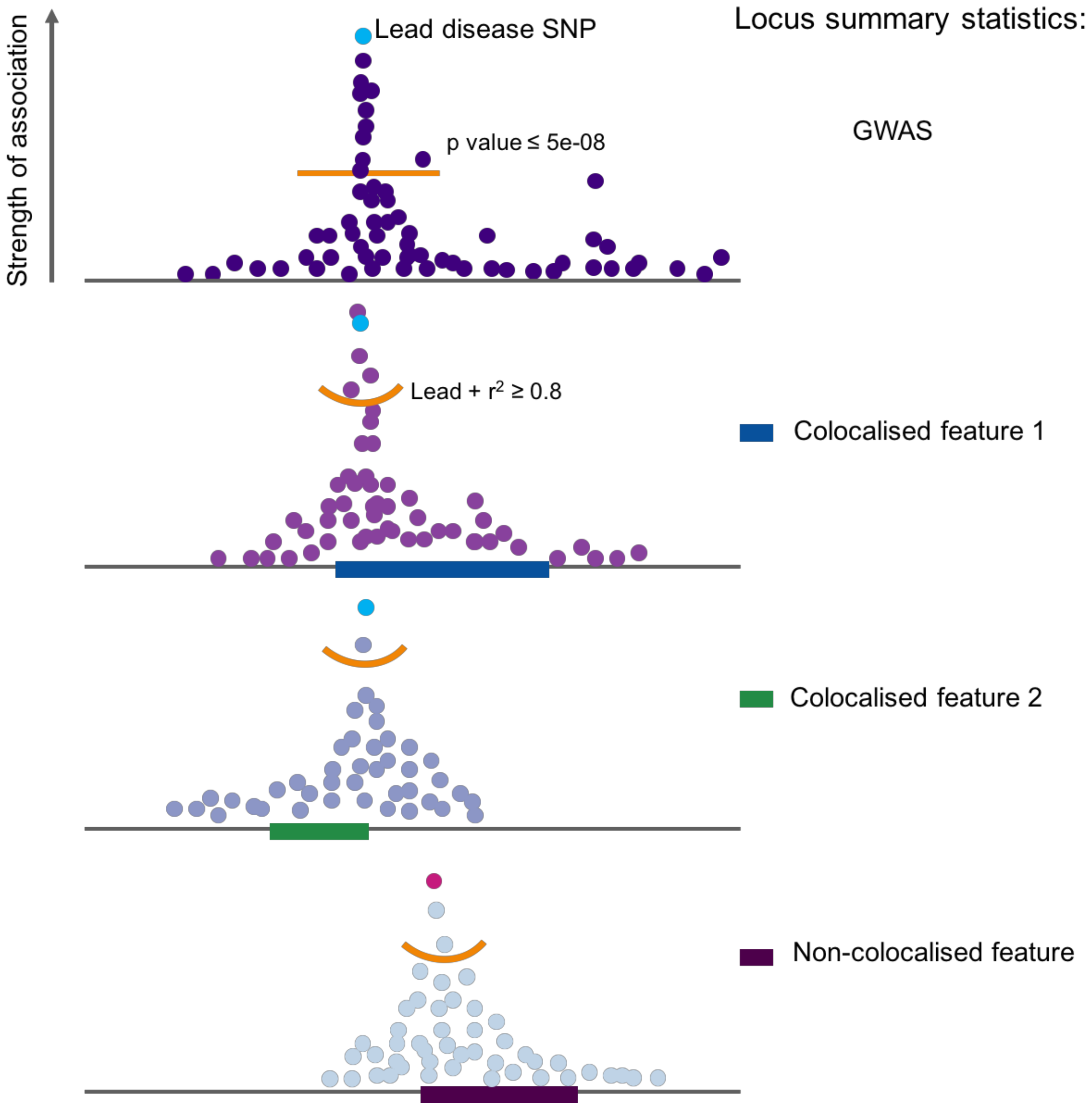


Figure 2.3: Summary of molecular QTL and disease locus colocalisation approach
This schematic summarises the selection of overlapped QTL-SNP regions and assignment of the disease SNP after colocalisation on individually tested regions. Variants tested for association in each set of summary statistics are represented by a coloured dot and the schematics represent manhattan plots with the higher the variant, the more significantly associated with the specified trait. Molecular features, such as genes or histone binding regions are represented by coloured blocks. For each identified GWAS-feature pair, lead QTLs and proxies associated with molecular features (above the curved orange line) are selected and an overlap is called if any of these SNPs are significantly associated with the GWAS trait (p value $\leq 5 \times 10^{-08}$) (above the straight orange line). For each overlap, all features significantly associated (qvalue ≤ 5%) with this lead QTL are assessed for colocalisation (above as the blue and green features). All SNPs within the feature and the cis-genomic region, defined as 1Mb upstream and downstream, are input to colocalisation if shared with the GWAS study. Where many features colocalised with one disease locus, the lead disease SNP in the test window (light blue) is used to assign the previously reported locus. The bottom panel represents a scenario where an overlap may have been detected between a molecular feature SNP and a significant GWAS SNP, but colocalisation was not detected with the disease GWAS signal.

Full details are listed in the Pickrell *et al.* (2016) publication but I briefly summarised the main points and equations from the method below. Here, we assume that the BLUEPRINT and disease cohorts are independent and there is no overlap or correlation. First, the method calculates a Bayes factor, which corresponds to the association evidence for three alternative models (below). Bayes factors for each SNP are approximated from Wakefield [2008]. Equation 1, below, gives the Wakefield approximate Bayes factor for model 1 where the SNP is associated with trait 1.

$$WABF_1 = \sqrt{1-r_1}\exp\left[\frac{Z_1^2}{2}r_1\right] \qquad (1)$$

$$BF^{(1)} = WABF_1 \qquad (2)$$

$$BF^{(2)} = WABF_2 \qquad (3)$$

$$BF^{(3)} = WABF_1 WABF_2\,, \qquad (4)$$

where Z_1 is the Z score estimate (the maximum likelihood estimate of beta divided by standard error, $\sqrt{V_1}$) of each SNP with the trait of interest and $r_1 = \frac{W_1}{V_1 + W_1}$. Bayes factors are averaged over computations with varying W. Therefore, in this method, the Z score for a SNP in the region for each trait is used to evaluate evidence for colocalisation. The three approximate Bayes factors above relate to three models; where the SNP is associated with the first trait (equation 2), second where the SNP is associated with the second trait (equation 3) and the third where a SNP is associated with two traits (equation 4).

Next support for an association is evaluated in a given genomic region, r, for all SNPs in the region. The regional Bayes factor (RBF_r) is evaluated for each model against the null model of no association in the region. The RBF_r evaluates the integral sum of the posterior probability (PP) for all SNPs in the region where the PP is the product of the Bayes factor and the prior probability of the variant being causal in the locus. In the model, all SNPs have equal prior probability of being causal. The method assumes one casual variant and if this is missing, the power to detect a shared genetic effect is reduced (model 3).

$$RBF_r^{(1)} = \sum_{i=1}^{K} \pi_i^{(1)} BF_i^{(1)}$$

$$RBF_r^{(2)} = \sum_{i=1}^{K} \pi_i^{(2)} BF_i^{(2)}$$

$$RBF_r^{(3)} = \sum_{i=1}^{K} \pi_i^{(3)} BF_i^{(3)}$$

$$RBF_r^{(4)} = \sum_{i=1}^{K} \sum_{j=1}^{K} \pi_i^{(1)} \pi_j^{(2)} BF_i^{(1)} BF_j^{(2)} I[i \neq j],$$

where $\pi_i^{(1)}$is the prior probability that SNP i is the causal one under model 1, $\pi_i^{(2)}$, is the prior probability that SNP i is causal under model 2 and $\pi_i^{(3)}$ is the prior probability that SNP i is causal under model 3. $RBF_r^{(4)}$assumes there are two causal SNPs that independently influence the two traits. K refers to the number of SNPs. The SNP priors are set as follows $\pi_i^{(1)} = \pi_i^{(2)} = \pi_i^{(3)} = \frac{1}{K}$.

Next, the prior probability of the regional models is calculated by the method to maximise the log-likelihood function of all SNPs in a region, over all four models. In our case, this is calculated per locus, which is pre-defined as a region with an overlapping molecular QTL and GWAS SNP (Figure 2.3) not genome-wide, given the tendency for QTL testing regions (2 Mb regions) to overlap. Finally, a posterior probability for each model per locus is calculated by multiplying the corresponding model prior probability by the RBF_r. For each locus, four posterior probabilities are generated for model 1, 2, 3 and 4. The PP for model 3 is used to evaluate whether there is evidence for a shared genetic effect.

There are some limitations in this method. It is difficult to distinguish between model 3 and 4 if there is high LD between the lead variants of each of the two traits ($r^2 \geq 0.8$). The model does not provide an estimation and direction of causality between the molecular trait and disease.

The threshold for calling a region colocalised was PP ≥ 0.99 for model 3, as this gives high confidence that there is statistical evidence for an underlying shared genetic signal between two traits. Only regions where there were equal or more than 20 SNPs shared between the disease and molecular datasets were considered as colocalised loci. Locus zoom plots were generated using custom scripts and used to provide visual evidence for colocalisation. The p-value supplied in the GWAS datasets was used in plotting as well as the raw p-value for the molecular QTLs (not 5% qvalue). Only SNPs that were shared between the disease and molecular QTLs were plotted in the locus zoom plot (as these were the input into gwas-pw). Final results excluded the MHC region, which was defined as the region on chr6 between the positions 20000000 and 40000000 based on previous investigations of the region (Trowsdale and Knight, 2013).

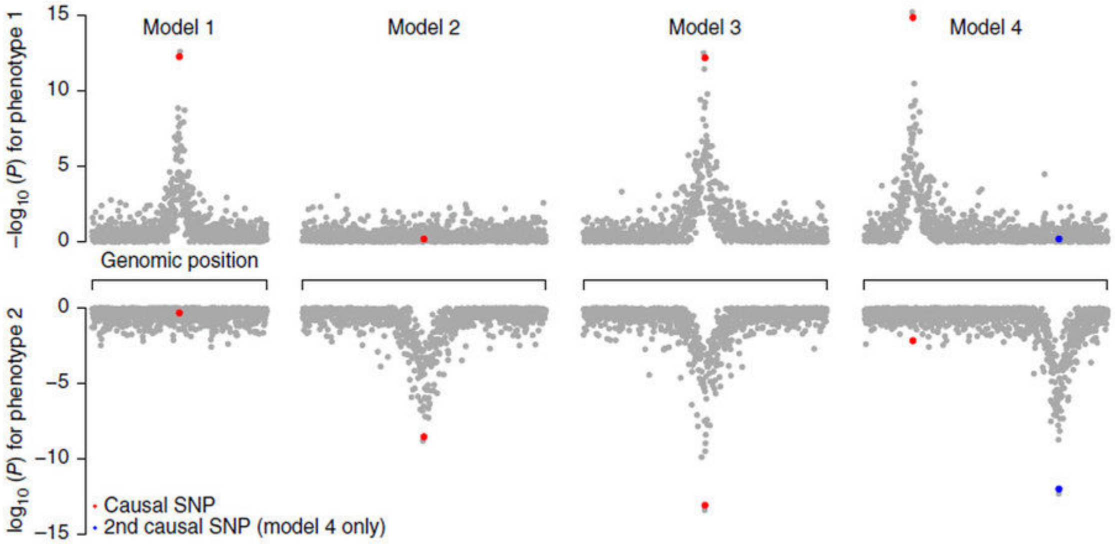


Figure 2.4 Schematic of the four models evaluated by colocalisation methods describing the relationship between two trait associations within a locus
In this analysis, the genomic region above represents a molecular QTL testing region overlapping a GWAS SNP. Model 1 and 2 are single association models, model 3 represents a colocalised region where two traits have shared genetic signals and model 4 is where two independent variants are associated with different traits. Figure reproduced from Pickrell *et al*, 2016 (Pickrell et al., 2016).

Assigning a unique disease locus: Multiple molecular features colocalised with some disease loci. In order to evaluate all possible molecular consequences of each disease loci, a lead disease SNP was assigned to each feature-QTL pair, each assessed independently for colocalisation. Molecular features sharing the same disease SNP were aggregated. The SNP with the lowest p-value in the testing region (+/- 1Mb) was selected from the GWAS summary statistics. The assigned SNP was compared to the reported disease lead SNP. In most cases, there was an exact match between the assigned and reported disease SNP. However, for SLE and AD, summary statistics from the full meta-analysis were not publicly available, instead for both summary statistics from the stage 1 GWAS were available. Reported lead SNPs in the study were based on meta-analysis. In these cases, LD (r^2 1000G) between the assigned and reported disease SNP was used to assign the published locus. For SLE, three regions significant in stage 1, which colocalised with features were removed from the final results as no corresponding region could be identified in the published meta-analysis. The FEV_1 GWAS contained a complex region of extended LD that was classified as the *KANSL1* inversion locus, defined by one conditionally significant genetic signal. The extended LD made it difficult to assign a lead disease SNP in my analyses. I found that features colocalised with FEV_1 disease SNPs rs111907488, rs62060763 and rs2532349 were all located in this region, therefore I combined all of these features into the *KANSL1* locus defined by the study reported lead SNP, rs2532349. Importantly, this method assigns the most significant lead SNP per locus, but the colocalisation could occur between secondary or further independent signals in either GWAS or BLUEPRINT associations. However, the lead SNP is assigned for ease of comparison to previous findings. In depth investigation of each locus is required to assess whether colocalisation occurs with the primary signal.

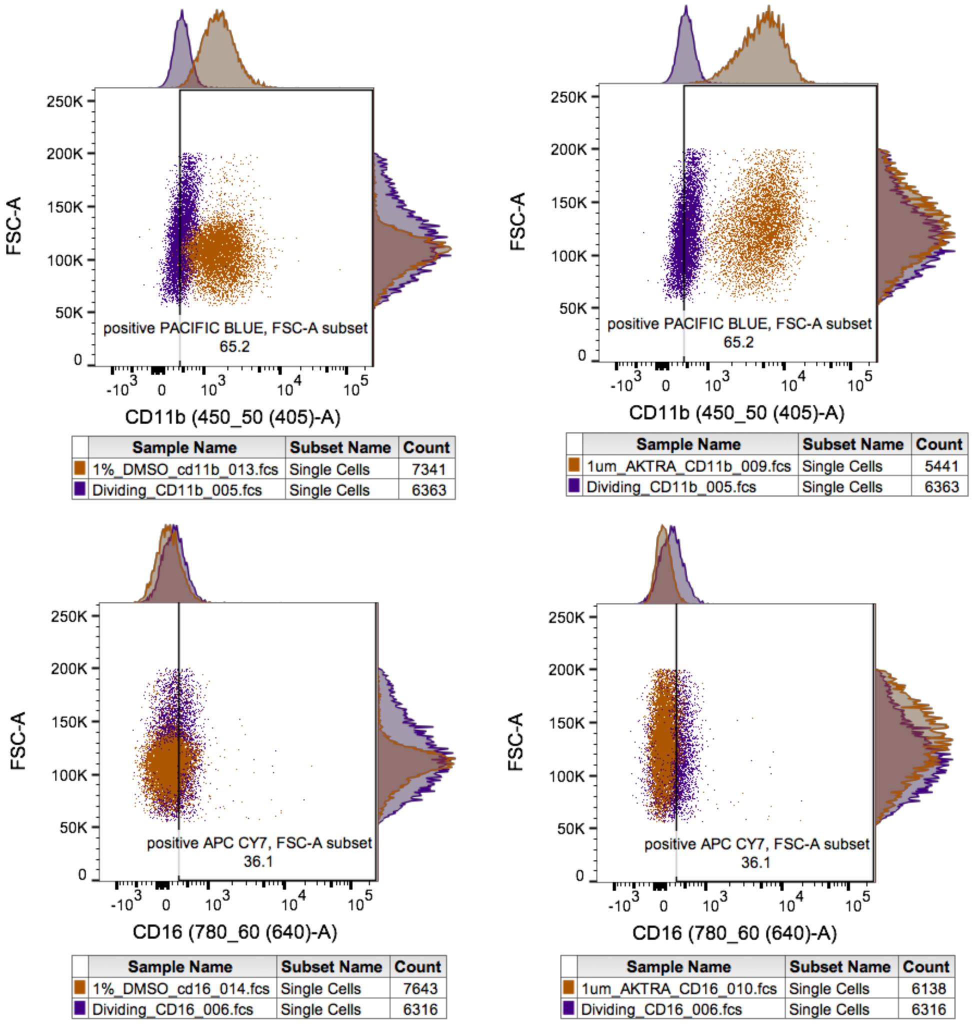
Conditional SNP analysis within phenotype: To gain a better understanding of molecular signals and whether these were shared between cell types, I used GCTA conditional analysis to estimate conditionally independent signals based on association statistics and LD between the variants in the locus. GCTA version 1.25.2 with the --cojo-cond option was implemented using QTL summary statistics from the BLUEPRINT study (Chen et al., 2016a, Yang et al., 2011, Yang et al., 2012). Summary statistics for each feature were input into GCTA. For LD estimation, genotype information from the BLUEPRINT cohort was used in plink hard call format using the --bfile option. Iterative conditional analysis was performed if any SNPs were associated with the phenotype after conditioning on the lead SNP. To evaluate this the q value, which represents the p value corrected for the number of SNPs tested in the region, was calculated using the qvalue R package version 1.43.0 (Bass JDSwcfAJ, 2015). GCTA was then implemented on output summary statistics conditioning on the SNP with the lowest qvalue from the previous conditional analysis. In most cases, the conditional analysis was confirmed by using a simple linear regression model with individual genotype data in R.

Linear regression analysis: For specific loci, as part of an in-depth investigation, various linear regression models were implemented using the lm() function in R. Phenotypes as described above were extracted from the full matrices from the Chen *et al.* (2016) study for features of interest. Units were $\log_2$RPM for the histone signal or normalised expression values, which are proportional to log2FPKM. SNP genotypes were input for every individual where 0 denotes homozygous reference, 1 heterozygous and 2 homozygous alternative, the latter was the defined effect allele for the Chen *et al* (2016) BLUEPRINT study. Genotypes were input as numeric to test for a trend across the 0,1,2 genotypes rather than as a factor where the difference between each genotype level is evaluated. Genotypes and phenotype values were matched using the unique study ID. To evaluate the causal relationships at the AMD *TNFRSF10A* locus, I used phenotype values for 158 individuals for which all phenotype data was available (gene, H3K27ac, H3K4me1). I implemented a two-stage approach to test for causality, first removing the effects of particular phenotypes by including these as covariates in a linear model. I used the Shapiro Wilko test to confirm normality of residuals after correction and used these residuals to test for remaining association with the locus lead SNP genotype.

R^2 and goodness of fit: The model fit was evaluated using the R^2 generated from the lm() function. Briefly R^2 or the coefficient of determination measures the proportion of y that is explained by x, the predictor variable, where a value of 1 demonstrates that x explains all of the variation in y. To evaluate whether fitting of additional covariates improved the model fit, the anova() function in R was used to perform a Chi-squared test to compare nested linear model 1 and linear model 2, related by the inclusion of additional covariates into model 2 (see Section 2.3.5.2). The chi-squared test evaluates whether the reduction in the residual sum of squares is statistically significant or not.

Linkage disequilibrium calculation: LD between variants was assessed using the 1000 genomes panel via the HaploReg resource and denoted as 1000G (Ward and Kellis, 2012). Alternatively, where indicated, LD was calculated from the Astle *et al.* (2016) cohort using PLINK v2 with the flags --ld and --bfile for the input imputed hard call files generated as part of the main GWAS analysis (Astle et al., 2016).

HL60 differentiation model: Additional functional data was required in some cases to further assign mechanism to disease loci. There are known limitations with access to primary human neutrophils and technical difficulties associated with using genomics approaches in these cells. The cell line, HL60, are commonly used as a model for neutrophils but resemble an early population of promyelocytes (Birnie, 1988). To address this and provide an additional high-quality granulocytic dataset, I implemented a well-established method for differentiating the HL60 cell line into a more mature neutrophil-like state and functional phenotype, with the addition of all-trans retinoic acid (ATRA) or dimethyl sulfoxide (DMSO) (Breitman et al., 1980, Chang et al., 2006). HL60 cells were grown at 37^0C in RPMI medium supplemented with 10% FBS and penicillin and streptomycin. Cells were passaged at a cell density of less than 1×10^6 cells per mL media. HL60 cells were seeded at a density of 10×10^6 cells/mL and incubated with either 1 µM ATRA or 1% DMSO for 96 hours. Cells without addition or either ATRA or DMSO were grown for 96hours without media change as a control. Every 24 hours, cells were counted for viability using a C-chip counting chamber and 1:1 dilution of Tryphan blue. After 96 hours, cells were harvested and fixed with 1% formaldehyde. To assess the success of differentiation, cell viability as well as the surface expression of neutrophil marker CD11b were measured. For flow cytometry measurements, 1×10^6 cells were harvested, spun at 1200 rpm for 5 minutes and pellets were resuspended in 100 µL FACS buffer (2% BSA, 0.001% EDTA in D-PBS). Cells were washed and incubated with 5 µL of Fc receptor blocking solution (Human TruStain FcX, BioLegend 422301) for 10 minutes on ice and afterwards washed with FACs buffer (1200 rpm for 5 minutes). 2 µL of the relevant antibody or isotype control was added for 30 minutes at 4^0C (Table 2.1). The stained cells were then washed three times. Unstained and isotype controls were also prepared. Samples were analysed using a BD LSR Fortessa Cell Analyser. In addition, gene expression of candidate genes known to vary in the differentiation process was also evaluated. RNA was extracted using a QIAGEN RNeasy mini kit and treated for DNase using TurboDNase (Life Technologies). Oligo(dT) primers and Superscript II (Invitrogen) were added for 2 hours at 42^0C for reverse transcription. HL60 genomic DNA was used as a control for the RT-PCR standard curve. CT values were calculated and compared to two reference genes, actin and C/EBPβ.



	Sample Name	Subset Name	Count
🟧	1%_DMSO_cd11b_013.fcs	Single Cells	7341
🟪	Dividing_CD11b_005.fcs	Single Cells	6363

	Sample Name	Subset Name	Count
🟧	1um_AKTRA_CD11b_009.fcs	Single Cells	5441
🟪	Dividing_CD11b_005.fcs	Single Cells	6363

	Sample Name	Subset Name	Count
🟧	1%_DMSO_cd16_014.fcs	Single Cells	7643
🟪	Dividing_CD16_006.fcs	Single Cells	6316

	Sample Name	Subset Name	Count
🟧	1um_AKTRA_CD16_010.fcs	Single Cells	6138
🟪	Dividing_CD16_006.fcs	Single Cells	6316

Figure 2.5: CD11b surface expression is increased on HL60 differentiation with DMSO or ATRA

Dot plots and histograms of the fluorescent signal of either CD11b (top panel) or CD16 (bottom panel) measured using flow cytometry. CD11b surface expression is increased on differentiated HL60 cells (DMSO and ATRA) compared to dividing HL60 (undifferentiated). CD16 expression is largely unchanged upon differentiation as has been previously observed (Jacob et al., 2002). Reduced proliferation and known gene expression changes (Lee et al., 2002) were also demonstrated (Supplementary Figures 2.1-2.2).

Chromatin immunoprecipitation followed by next-generation sequencing (ChIP-seq): ChIP experiments were adapted from a previously published protocol (Schmidt et al., 2009), with some modifications. 20 x 10^6 cells were used for TF immunoprecipitation (IP) and 5 x 10^6 cells for histone modifications. Cell pellets were fixed by incubation with 1% formaldehyde for 10 minutes at RT, followed by five minutes with 2.5M glycine. Cells were pelleted and washed with PBS, flash frozen in dry ice and stored at -80°C. Cells were sonicated for eight cycles with 30 seconds on and 45 seconds off at 4°C using a Diagenode PicoRuptor biorupter. Sonication efficiency (150-500bp fragments) was verified using an Agilent DNA bioanalyser. 2.5μg of each antibody was bound to Protein A Dynabeads (Invitrogen) in a 4-hour incubation. Sonicated lysate was added to the antibody-bead mix and incubated overnight at 4°C. The bound-beads were then washed with cold RIPA buffer ten times. Crosslinks were reversed with incubation at 65°C for five hours to elute DNA. Samples were then incubated with 2μl RNase at 37°C for 30 minutes followed by incubation with Proteinase K at 55°C for one hour. Ampure beads (Beckman Coulter, A63881) were added to the DNA in a 1:1.8 ratio and samples were washed twice in cold 70% ethanol. DNA was then eluted in 50 μl elution buffer. Samples were stored at -20°C prior to Illumina library preparation, which was carried out according to the Illumina TruSeq ChIP sample kit protocol. An additional step of enriching fragments through PCR prior to gel purification was added to avoid amplification of contaminants. RNA index sequences were ligated to ChIP-enriched DNA fragments (200-500bp in size) for multiplexed libraries. Libraries were submitted for single-end sequencing with a read length of 50bp using a HiSeq2000. For analysis of sequencing data, reads were first aligned to the human reference genome (hg19 CRCh37) using BWA version 0.6.1 with default parameters (Sanger pipelines). ChIP-seq data analysis was performed using an in-house pipeline developed by Louella Vasquez that implemented standard analysis procedures, described here. Duplicate reads were removed (Picard MarkDuplicates v1.103) and reads with a zero-mapping source were removed. Peaks were called using MACs v2.0.10.20131216 (Zhang et al., 2008) with default parameters using the estimated fragment size evaluated by PhantomPeakQualTools vr18. Narrow flags were used for all factors apart from H3K4me1 for which the broad flag was used. For the background control, sonicated input DNA was used from the respective ATRA or DMSO treatments. Encode quality control metrics (Phantom Quality Tools) were used to evaluate the success of IP as well as visual inspection in the UCSC genome browser. Significant peaks were selected if 1% FDR or less.

Antibody	Supplier	Source/Clone
Anti-PU.1	Santa Cruz, sc22805	Rabbit polyclonal
Anti-CEBPβ	Santa Cruz 150 X	Rabbit polyclonal
Anti-trimethyl histone H3(lys4)	Diagenode C15410003	Rabbit polyclonal
Anti-H3K27acetyl	Abcam ab4729	Rabbit polyclonal
Anti-monomethyl histone H3(lys4)	Diagenode C15410194	Rabbit polyclonal
PE Mouse Anti-Human CD11b	BD Pharmingen 557321	ICRF44
PE Mouse IgG1, K isotype control	BD Pharmingen 556650	MOPC-21
Pacific Blue Anti-Human CD11b	BD 558123	ICRF44
Pacific Blue Mouse IgG1, isotype control	BD 558120	MOPC21
APC CY7 Anti-Human CD16	BD 557758	3G8

Table 2.1: Antibodies used in ChIP-seq and flow cytometry experiments
Antibodies against specific proteins studied, the supplier and reference as well as source or clone for flow cytometry experiments.

2.3 Results

2.3.1 Functional enrichment in five diseases

First, I aimed to assess the enrichment of non-autoimmune disease SNPs in immune molecular QTLs. I used the GARFIELD method (Materials and Methods) to evaluate significant enrichment by calculating the odds ratio (OR) for enrichment of GWAS variants in molecular features such as gene expression, RNA splicing, histone marked regions and lastly DNA methylation from three cell types (Chen et al., 2016a). Higher OR estimates indicated increased odds that an overlap occurs with a significant GWAS variant as opposed to an overlap with a non-significant GWAS variant.

I identified significant enrichment of AD, CAD, FEV$_1$, AMD and SLE variants associated at the GWAS p value threshold of 1 x 10^{-05} in a number of molecular feature types, ranging from four significant features for AD to all features with significant enrichment for SLE. This is in contrast to the complete lack of significant enrichment of immune QTLs with Type 2 diabetes (T2D) SNPs. I included T2D-associated SNPs as a negative control as despite recent links to inflammation, it is currently considered that a disordered metabolic state in Type 2 diabetic patients then leads to immune dysregulation (Hameed et al., 2015). This is supported by low enrichment of T2D variants in immune molecular features, as observed previously in two independent studies, and further confirmed here in my analysis (Figure 2.6) (Chen et al., 2016a, Alasoo et al., 2017).

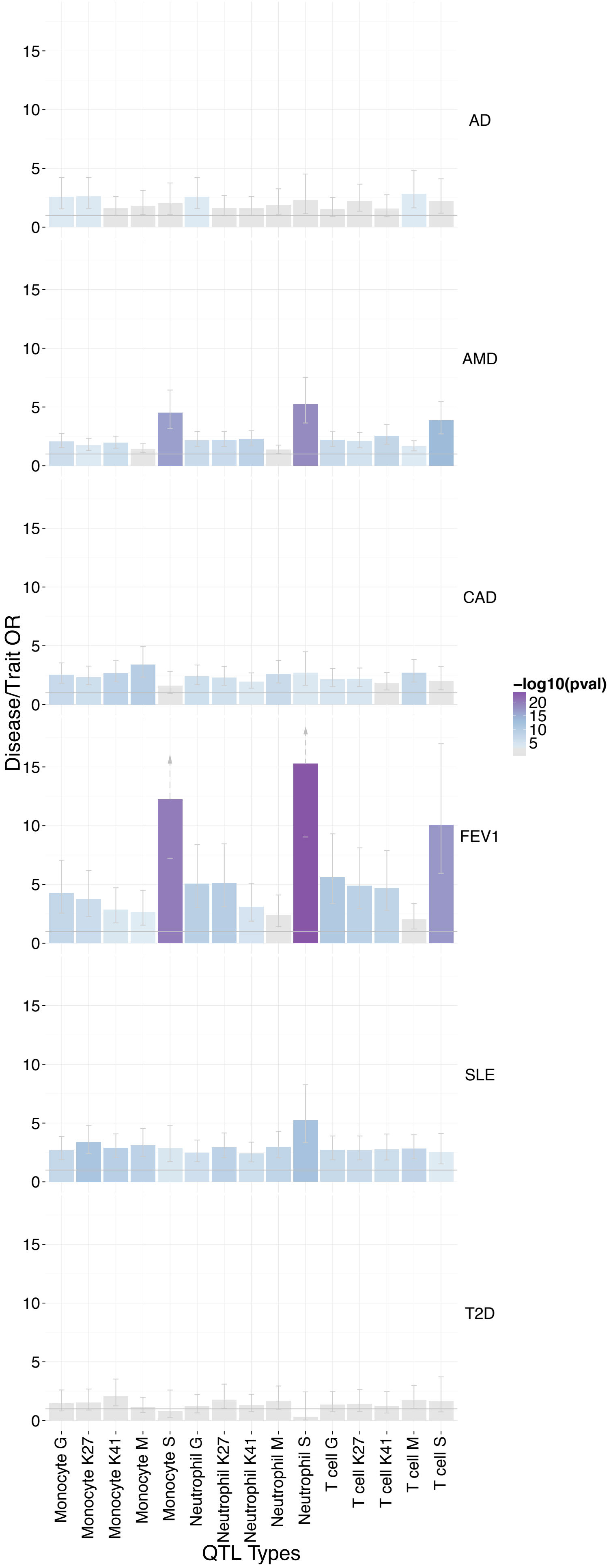
The highest and some of the most significant enrichments were observed with neutrophil, monocyte and T cell splicing QTLs and FEV$_1$-associated variants, with a mean OR of 12.520 across all three cell types and p values ranging from 6.404 x 10^{-18} for T cell splicing QTLs to 2.264 x 10^{-24} for neutrophil splicing QTLs. Significant enrichment in splicing QTLs was also observed for other traits. AMD-associated variants showed the highest enrichment in splicing QTLs from all three cell types (mean OR = 4.541), with the highest OR also in neutrophils. SLE-associated variants also showed the highest enrichment in neutrophil splicing regions (OR = 5.247, p value = 8.505 x 10^{-13}).

However, there were large OR confidence limits for the FEV$_1$ splicing enrichment, which indicated a higher error in this measurement. This could be due to a lower number of variants overlapping these annotations when in comparison to some (but not all) of the disease GWAS. For FEV$_1$, 24 variants were annotated as overlapping monocyte splicing QTLs at the GWAS p value threshold of 1 x 10^{-05}, 24 overlapping T cell splicing QTLs and 24 with neutrophil splicing QTLs, but for AMD-GWAS, 39, 36 and 40 variants overlapped monocyte, neutrophil and T cell splicing QTLs respectively at the same GWAS 1 x 10^{-05} p value threshold.

SLE-associated variants were significantly enriched across all molecular QTL types, whereas other trait variants demonstrated more variable enrichment patterns. The high and significant enrichment across the majority of QTL and cell types was previously observed for other autoimmune diseases including Crohn's disease, rheumatoid arthritis and multiple sclerosis (Chen et al., 2016a). The ubiquitous effect may reflect extensive cross-talk between different immune cell populations or that increased power in both GWAS and QTL studies are required to fully resolve finer immune cell population signatures.

Interestingly, CAD was the only trait for which there appears a possible cell-type specific pattern of enrichment with a consistent significant enrichment for monocyte gene, H3K27ac, H3K4me1 and methylation QTLs, which was in agreement with the established role for monocytes in CAD aetiology (discussed above). The mean OR for the significantly enriched monocyte features, as calculated from the GARFIELD output data was 2.737. Significant neutrophil annotations showed a slightly lower average OR of 2.310, and the equivalent p values of enrichment were less significant than the equivalent p value for monocyte annotations apart from for neutrophil splicing, which was not significant in monocytes. For T cells, only gene, H3K27ac and methylation QTLs showed significant enrichment with CAD variants.

In summary, this approach has demonstrated significant enrichment of non-autoimmune disease variants in molecular QTLs, where greater enrichment was observed in all cases for at least five molecular QTL types than the negative case of T2D. For example, all five disease-associated variants showed significant enrichment in monocyte and neutrophil eQTLs. These significant enrichments suggest that these GWAS variants may result in functional changes in immune cells that could underpin mechanisms at some genetic risk loci. Therefore, using these molecular data to further study disease loci could aid the generation of biological hypotheses of downstream consequences at these loci.



AD
AMD
CAD
FEV1
SLE
T2D
0
5
10
15
Disease/Trait OR
−log10(pval)
20
15
10
5
Monocyte G
Monocyte K27
Monocyte K41
Monocyte M
Monocyte S
Neutrophil G
Neutrophil K27
Neutrophil K41
Neutrophil M
Neutrophil S
T cell G
T cell K27
T cell K41
T cell M
T cell S
QTL Types

Figure 2.6: Enrichment of molecular QTLs in six diseases
Histogram plot summarises the enrichment of molecular QTLs in a range of diseases. Higher enrichment represents a greater overlap of disease SNPs and molecular QTLs and is indicated by an increased odds ratio (OR) on the y-axis calculated by GARFIELD. Represented here is overlap of disease variants at the suggestive GWAS p value threshold, 1 x 10^{-05}. The significance of the overlap is depicted by the colour scale of the plots (purple is more significant), and is corrected for the number of effective annotations (Methods) and for 5 diseases (where significant enrichment was expected, which excludes Type 2 diabetes). Annotations are enriched with OR ≥ 1. The OR confidence limits are reflected by the error bars, with arrows to designate values beyond the maximum OR shown in the figure. Type 2 diabetes was included as a non-immune negative control and as such no significant enrichment was observed for these SNPs. G refers to gene features i.e. eQTLs, K27 to H3K27ac hQTLs, K41 to H3K4me1 hQTLs, M to methylation QTLs and S to splicing QTLs (percent-splice in as defined in the main BLUEPRINT paper and detailed in Methods).

2.3.2 Colocalisation of molecular traits with five diseases

Formal statistical testing is required to evaluate whether the molecular trait and disease associations can be attributed to the same underlying causal genetic variant. I therefore implemented the colocalisation method, gwas-pw, which uses prior information regarding the effect sizes (expressed as a Z-score) of two traits within one locus (detailed in the Materials and Methods). I assessed colocalisation between each of the five complex diseases/traits described above; AD, SLE, CAD, AMD and FEV_1 with four molecular traits, gene expression, H3K72ac & H3K4me1 modifications and RNA splicing from three cell types monocytes, neutrophils and naïve $CD4^+$ T cells (Chen et al., 2016a). The overall aim of this approach was to identify to what extent immune molecular features could explain mechanisms at unique disease loci reported from each GWAS.

Across all five diseases, I identified that 46% (55/120) of previously reported disease loci colocalised with at least one molecular feature using a high posterior probability of colocalisation (PP ≥ 0.99) (Table 2.2). The highest percentage of colocalised loci was observed with systemic lupus erythematosus (SLE) variants (54%), which is expected as SLE is a paradigmatic autoimmune disease displaying strong immune cell involvement (Dai et al., 2014, Farh et al., 2015). SLE was included in this chapter to provide a positive control, as strong enrichment of SLE variants in lymphoid and monocyte enhancers (high H3K27ac) has been previously demonstrated using ROADMAP data (Farh et al., 2015). Neutrophil data was not assessed in this previous study, despite recent observations of the importance of this cell type to SLE pathogenesis (Weidenbusch et al., 2017). Using the BLUEPRINT neutrophil data, therefore, provided the opportunity to gain novel insight into neutrophil-mediated risk at SLE loci.

The lowest percentage of colocalised loci was observed with coronary artery disease (40%). However, of the 43 common GWAS loci excluding the MHC region that were assessed, there were still 17 CAD loci that colocalised with at least one molecular feature. Therefore, this type of analytical approach affords the potential to form molecular hypotheses of up to one-third of the known loci within clearly defined cell type populations facilitating downstream experimentation.

Disease	Disease variants (+MHC)	Disease loci (+MHC)	All coloc (%)	Gene coloc	PSI coloc	H3K4me1 coloc	H3K27ac coloc
AMD	41 (45) *	32 (33)	16 (50)	9	4	10	12
CAD	43 (44)	43 (44)	17 (40)	7	3	12	15
SLE	24 (25) *	24 (25)	13 (54)	6	2	6	9
AD	14 (15)	14 (15)	6 (43)	2	3	4	4
FEV1	7 (9) *	7 (9)	3 (43)	3	1	2	2
All	129 (138)	120 (126)	55 (46)	27	13	34	42

Table 2.2: Number of colocalised disease loci per feature type
Number of colocalised unique disease loci per feature type, per disease. The percentage of all features colocalised per disease is the ratio of colocalised loci over total common loci that were defined in the GWAS excluding the MHC region. For the AMD statistics, 7 variants and 1 locus with a MAF < 1% were also excluded and not counted in the table above. Each disease locus that colocalised with at least one feature across three cell types is counted as one colocalisation. For AMD, the percentage was calculated with respect to the disease loci (32) not the number of independent variants (41), where at some loci there were multiple independent genetic signals. Where some form of conditional analysis was performed and identified further signals this is designated by *.

Nearly half of the disease loci (27/55, 49%) colocalised with at least one eQTL in at least one cell type across all diseases. Of these 27 loci, eight loci colocalised with at least one eQTL in monocytes and no other gene effects in neutrophils or T cells (AD *MS4A6A,* AMD *TBC1D23*, AMD *CETP*, AMD *TNFRSF10A*, CAD *REST*, CAD *PPAP2B*, FEV_1 *TSEN54*, SLE *BANK1*) (Figure 2.8). Two loci, CAD *NT5C2* and FEV_1 *RP11-186N15.3*, colocalised with eQTL effects in neutrophils only i.e. no other colocalised eQTL was detected in either monocytes or T cells and similarly only two loci colocalised with T cell specific eQTLs (AMD *HIGD1AP14* and SLE *BLK*). Where gene effects were not unique to one cell type, in the majority of cases at least one colocalised gene was shared between two or three cell types (80%, 12/15).

In total, there were 13 disease loci that colocalised with at least one percent-splice-in (PSI) effect, which could include exon skipping or alternative donor or acceptor usages. A subset of loci colocalised with eQTLs and splicing effects in the same gene such as AD *MS4A6A*, AMD *PILRB*, CAD *NT5C2* and SLE *IRF7*, which highlighted cases where a disruption of alternative splicing could lead to a change in the overall level of gene expression. Alternatively, five loci colocalised with a splicing QTL but not an eQTL across all cell types. These included AD *CR1*, AMD *NPLOC4*, CAD *MIA3* and CAD *IL6R*, which all colocalised with exon skipping events in the corresponding genes as well as the SLE *FCGR2A* locus*,* which colocalised with exon skipping events in the *FCGR2A* gene in neutrophils and the *FCGR3A* gene in monocytes. Genetic control of splicing has previously been shown to be relevant to disease risk and also often independent of both gene expression and histone activity (Li et al., 2016c). Therefore, investigating splicing events as well as gene expression can highlight additional molecular mechanisms (Odhams et al., 2017, Li et al., 2016c, Chen et al., 2016a).

Next, I evaluated whether there existed any specific cell type patterns across the five diseases. In addition to assessing the number of disease loci that colocalised with *at least* one feature, I counted *all* colocalised gene and splicing QTL effects including where there were multiple features for one disease locus (Figure 2.7). For example, in this analysis, the AMD *TNFRSF10A* locus colocalised with eQTLs in monocytes for three genes: *TNFRSF10A; CHMP7* and *RP11-1149O23.3*, which were counted as three monocyte-specific features. I also limited my analysis to genes and splicing effects to avoid over-inflating the feature counts. In the BLUEPRINT cohort, overlapping histone modification signal peaks from multiple individuals were merged to create a unified peak list facilitating the identification of QTLs across all individuals (example in Figure 2.10) (Chen et al., 2016a). This could generate broad regions that could contain the signal of multiple correlated peaks. Elucidation of the exact putative enhancer region for the corresponding colocalised gene therefore required further molecular dissection (as discussed later in this chapter) (Figure 2.10-12, 2.17). To avoid over-estimating the importance of cell type by counting multiple correlated histone regions I assessed the gene and splicing effects for possible cell-type specific patterns.

18 monocyte features colocalised across all disease loci in total. Nine neutrophil- and nine T cell-derived features also colocalised across all loci (Figure 2.7). There was a higher degree of shared colocalised genes and/or splicing junctions between monocytes and neutrophils (seven) than between monocytes and T cells (three) or between neutrophils and T cells (one). This could be expected given that monocytes and neutrophils are derived from the common myeloid progenitor cells whereas CD4$^+$ T cells differentiate from the common lymphoid progenitor and deviate more in function (Figure 1.5 (Orkin and Zon, 2008)). Interestingly, there were also seven gene or splicing effects that occurred in all three cell types, suggesting these loci may have more ubiquitous effects (AD *EPHA1-AS1*, AMD *PILRB* and *MEPCE* at the same locus, AMD *RP11-644F5.10*, CAD *NT5C2*, CAD *GGCX* and

SLE *IRF7*). Indeed, I observed that the AD *EPHA1-AS1* (*EPHA1* antisense RNA 1) locus lead SNP was also a significant eQTL for this gene across multiple tissues assayed in the (G. TEx Consortium, 2015) including adipose, lung, spleen and whole blood. I observed the same broad tissue effect for the AMD *PILRB* locus, where the lead SNP was also a significant eQTL across more than 15 different tissues including coronary artery, brain, adipose and pancreas. Finally, the CAD *GGCX* locus, was also an eQTL in over ten tissues including aorta artery, stomach, pancreas and adipose. In summary, this approach demonstrated that across all five diseases, the most commonly colocalised features were monocyte-derived.

The majority of disease loci also colocalised with either H3K27ac, H3K4me1 QTLs or both (89%, 49/55). Across all three cell types, 42% (20/55) of loci that colocalised with gene or splicing QTLs also colocalised with a histone QTL from the corresponding cell type in at least one cell type (i.e. where a monocyte eQTL colocalised with either a monocyte H3K27ac or H3K4me1 QTL). In these cases, colocalised histone-bound regions may demarcate putative regulatory regions for the respective colocalised genes (Section 2.3.5 onwards for detailed examples). In addition, 43% (23/55) of disease loci colocalised with a histone QTL but not a gene or splicing QTL. These could represent poised enhancers indicating that the regulated genes are active in other cell types or that these regulatory QTLs affect processes beyond gene expression and splicing (Pai et al., 2015).

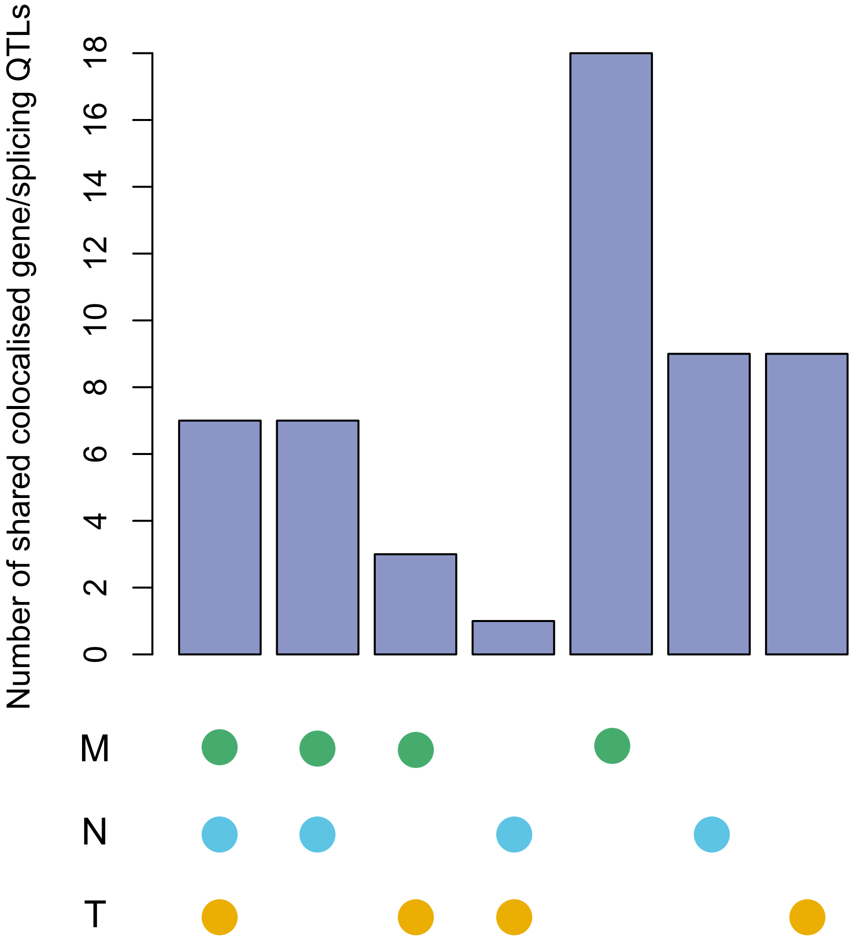


Figure 2.7: Colocalised gene and splicing QTLs per cell type and cell type combination across all diseases
All gene or splicing QTL features are counted for each cell type combination, for example the CAD *GGCX* locus colocalised with *GGCX* eQTLs across all three cell types so was counted as a shared MNT signal, but in the same locus, *VAMP8* eQTL colocalised in T cells only, so this was counted as a T cell (T) specific signal. The highest number of genes and splicing QTL features were in monocytes, demonstrating the potential importance of these cells across all diseases.

2.3.3 Colocalisation of molecular traits reveals potential molecular mechanisms at disease risk loci

I observed a high proportion of disease loci colocalised with molecular features but to form detailed mechanistic hypotheses for specific disease loci, integration of all colocalised features across cell types is required. Below, I discuss specific insights into the potential function of genetic risk loci.

Colocalisation with eQTLs offers the most intuitive interpretation of the molecular consequences at disease risk loci by identifying genes with altered expression levels. Figure 2.8 summarises the multiple molecular features I identified as colocalised with each disease locus. Figure 2.8 also describes the status of the previously predicted gene for the GWAS locus. There were several instances, 18 in total, where the colocalised gene matched previous target predictions for that locus, including AD *MS4A6A*, AMD *SRPK2*, *RDH5*, *CETP*, *CNN2*, *PILRB/A*, *TNFRSF10A*, CAD *NT5C2*, *LIPA*, *REST*, *GGCX*, *PPAP2B*, FEV_1 *TSEN54* and SLE *BANK1*, *IRF7*, *BLK*, *ITGAM*, *UBE2L3*.

All colocalised genes matching previous predictions at SLE loci had direct immune roles, which is expected for this prototypic autoimmune disease, but also in this case likely due to the use of publicly available eQTL data in the initial assignment of GWAS locus by the authors (Bentham et al., 2015). From my analysis, I identified colocalised eQTL genes, *BANK1* and *BLK,* which act in related B cell signalling pathways and are both regulated by type 1 interferons (Delgado-Vega et al., 2010). Here, these gene effects were T cell-mediated; a decrease in T cell *BLK* expression correlated with an increased disease risk, which was also observed in the original GWAS study. I also identified the gene *ITGAM*, where a neutrophil eQTL colocalised at this locus and decreased gene expression corresponded to decreased disease risk (rs9673398, EA = G, eQTL beta = -0.4896, SLE OR = 0.81). *ITGAM* encodes integrin alpha M chain, which forms the leukocyte-specific Mac-1 receptor shown to be important for neutrophil and monocyte-endothelial adherence and phagocytosis (Rebhan et al., 1998). I also identified other well-known immune genes such as *UBE2L3*, which encodes a ubiquitin conjugating enzyme E2 L3 involved in targeting proteins for degradation (Rebhan et al., 1998). Variants within the gene have also been associated with risk of Crohn's disease, coeliac disease and rheumatoid arthritis (Fransen et al., 2010, Zhernakova et al., 2011). Here gene expression was positively correlated with SLE risk, confirming observations from the original GWAS (Bentham et al., 2015) Interestingly, the eQTL was more significant in neutrophils (eQTL lead, rs2298429, p value = 1.188×10^{-30}) than monocytes (eQTL lead rs5749485, p value = 2.901×10^{-04}) and the effect greater in magnitude (neutrophil, EA = G, beta = 1.267, SE = 0.110, monocyte, EA = C, beta = 0.448, SE = 0.124). Neutrophil eQTLs were not assessed as part of locus assignment from the Bentham *et al.* (2015) SLE GWAS, instead stimulated monocytes, B cells, CD14$^+$ monocytes, NK cells and CD4$^+$ T cells were included (Bentham et al., 2015). This suggests that neutrophils may be the effector cell for this disease risk locus, demonstrating the importance of assessing multiple cell types, particularly in this context as neutrophils have been shown to be important in the aetiology of lupus (Weidenbusch et al., 2017).

I also identified colocalised genes with immunological roles for the other traits I investigated. AD risk variants have been identified with within the *MS4A* cluster containing multiple genes encoding accessory proteins that amplify receptor function and regulate immune cell activation and survival (Proitsi et al., 2014). Significantly increased blood *MS4A6A* transcript levels have been associated with common coding SNPs in a cohort of approximately 300 AD patients, suggesting higher protein levels could contribute to the pathogenic pro-inflammatory AD phenotype (Proitsi et al., 2014). The authors found that *MS4A6A* was the only gene significantly differentially expressed, with higher expression in the patient cohort compared to the normal elderly controls. The *MS4A6A* expression effect was associated with SNP genotype but this effect was only significant in the patient group. By contrast, I identified that the AD *MS4A6A* locus colocalised with multiple genes in this locus including *MS4A6A, MS4A4A* and *MS4A4E* in monocytes and a further monocyte splicing QTL for *MS4A6A*. I observed the same positive correlation between gene expression and disease risk, where higher expression of *MS4A6A, MS4A4A* and *MS4A4E* corresponded to increased AD risk. My analysis identified *MS4A* effects in healthy individuals, which suggested in contrast to previous observations that the expression effect could contribute to the risk of AD, rather than reflect an expression effect that is perturbed in the disease state. There is evidence that the effect I identified and the previous effect could represent the same genetic signal, the LD between tested variants was moderate (r^2 0.54-0.61). Higher expression of these genes could, therefore, aid prediction of risk and also provide potential targets to investigate for therapeutic lowering of the expression levels.

For CAD, I confirmed the well-known *IL6R* locus, discussed in Section 2.1, colocalised with an exon skipping event in the *IL6R* gene generating higher levels of the transcript encoding soluble *IL6R*, thus providing clear validation of my analytical approach. Other CAD loci also highlighted the importance of immune activity, for example, I identified that the CAD *PPAP2B* colocalised with monocyte expression of this gene (rs56186267 EA = A, p value = 3.098 x 10^{-10}, beta = 0.885, SE = 0.141) as well as a H3K27ac marked region (p value = 1.884 x 10^{-07}, beta = 0.780, SE = 0.150) towards the 3' end of this gene. The *PPAP2B* intronic SNP, rs72664324, which is in high LD with the lead BLUEPRINT SNP (r^2 = 1, 1000G) has previously been identified as disrupting the binding of a C/EBPβ transcription factor (TF) within a macrophage LDL-induced dynamic open chromatin region (Reschen et al., 2015). Monocytes recruited *in vivo* to atherosclerotic plaques differentiate into macrophages, which are in turn stimulated to form foam cells by uptake of environmental lipids (Reschen et al., 2015). *PPAP2B* encodes the enzyme, LPP3, which deactivates pro-inflammatory mediators, such as those released from foam cells. Increased gene expression confers a protective CAD effect. Similarly, in the colocalised monocyte effects I identified, increased *PPAP2B* gene expression as well as increased signal of the H3K27ac modification (and therefore enhancer activity) corresponds to a decrease in CAD risk. Evidence also suggested that this was a monocyte-specific effect as *PPAP2B* expression was not tested in T cells due to low expression and in neutrophils, the association was not significant after correcting for multiple testing (rs72664324/rs56186267, p value = 1.975×10^{-01}). The classical $CD14^{+}CD16^{-}$ monocytes studied in the BLUEPRINT is known to differentiate into macrophages in inflamed tissues (Ohradanova-Repic et al., 2016). Therefore, the identified monocyte colocalisation at this locus suggests that enhancer activity and *PPAP2B* gene effects may be present in monocytes before stimulation and differentiation to macrophages, which in turn further supported the relevance of these tissues to CAD risk mechanisms.

A subset of AMD loci also colocalised with immune-related genes including the *TNFRSF10A* gene encoding the TRAIL 1 receptor and the *PILRB/A* genes encoding the paired immunoglobulin-like type 2 receptor beta and receptor alpha that regulate immune responses (www.genecards.org), (Rebhan et al., 1998). Similarly, I identified evidence for a possible immunological role for the AMD *CNN2* locus, for which colocalised genes were disparate across all three cell types in this study (Figure 2.8). The locus colocalised with the eQTL for *CNN2* in monocytes, for *CTB-31O20.2* in neutrophils and for *ABCA7* in T cells, a gene located approximately 1 kb downstream of *CNN2* on chr19. The transporter encoded by *ABCA7* is involved in pathogen-mediated phagocytosis, a process which requires actin skeleton reorganisation (Humphries et al., 2015). *CNN2* encodes calponin 2, a protein involved in the structural organisation of actin filaments, suggesting these two genes may have coordinated functions (Humphries et al., 2015). Genes in the *ABCA7* locus, including *CNN2*, have been shown to be involved late-onset Alzheimer's disease aetiology (Humphries et al., 2015), although in this chapter the locus was evaluated in the context of AMD, this suggests a role for this cluster in age-related disorders. Therefore, although divergent colocalised genes across cell types may suggest differential functions in each tissue, a thorough assessment at each locus is required to evaluate whether the combined activity of genes may highlight important disease-relevant pathways.

In addition to immune function, I provided further support for the importance of lipid pathways in CAD and AMD. Important loci included the CAD *LIPA* locus that colocalised with monocyte and T cell expression of the *LIPA* gene and the AMD *CETP* locus colocalised with monocyte expression of the *CETP* gene, which encodes cholesteryl ester transfer protein. *LIPA* encodes lysosomal acid lipase A and the association of the intronic SNP, rs1412444 with increased *LIPA* gene expression levels in blood cell types and increased CAD risk is relatively well established (Zeller et al., 2010, G. TEx Consortium, 2015, Wild et al., 2011). I confirmed this direction of effect in monocytes but observed that in T cells, the effect was less significant and a decreased in *LIPA* expression corresponded to increased CAD risk (rs1412444 monocyte p value=2.044 x 10^{-49}, T cell p value= 4.588 x 10^{-06}). In neutrophils, rs1412444 was not significantly associated with *LIPA* expression (p value = 7.031 x 10^{-02}), suggesting that within the limits of the power of this cohort, the *LIPA* monocyte expression effect was most functionally relevant to CAD risk. Further supporting this is the colocalisation of this locus with a H3K27ac peak (10:90993615:91006217) and a H3K4me1 peak (10:90987967:91024823), which both directly overlap rs1412444. Together the activity of these histones could function as a putative enhancer for the *LIPA* gene as an increase in H3K27ac and H3K4me1 signal corresponds to an increased in *LIPA* expression and increased CAD risk.

I also identified colocalised genes with more general functions. The FEV_1 *TSEN54* locus colocalised with a monocyte eQTL for the *TSEN54* gene encoding a subunit of the tRNA splicing endonuclease complex (Figure 2.8). Despite a more general function, the colocalisation with monocyte expression appeared to reflect a relatively selective cell-type effect within this dataset. *TSEN54* was not tested in neutrophils due to low expression (median $\log_2$FPKM = 2.719, BLUEPRINT cohort). *TSEN54* gene expression was high in T cells and monocytes (median $\log_2$FPKM T cells = 9.433, median $\log_2$FPKM monocytes = 7.067) but the lead T cell eQTL (rs7225469 EA = C, beta = -1.197, SE = 0.176, p value = 9.348 x 10^{-12}) was not highly correlated with the lead FEV_1 variant, rs7218675 (r^2 = 0.26 1000G). These, therefore, may represent independent effects, providing both a hypothesis and a cellular model where decreased *TSEN54* expression in monocytes corresponds to an increase in FEV_1 ratio. The cell-type specific observation is an improvement on the initial observation that the GWAS locus was a significant *TSEN54* eQTL in the heterogeneous mix of whole blood (Wain et al., 2015).

I also identified a monocyte and T cell gene target, *RDH5*, that colocalised with the AMD locus where the encoded protein has a specialised role seemingly localised to the disease tissue (retina). *RDH5* encodes the 11 cis-retinol dehydrogenase enzyme, which catalyses the final step in the synthesis of the mammalian pigment chromophore, 11 cis-retinaldehyde (Liden et al., 2001). Multiple lines of evidence link *RDH5* disruption to impaired eye function; for example, the rare night blindness disorder fundus albipunctatus is caused by *RDH5* mutations and in AMD-like pathology and impaired dark adaptation is observed in the *RDH5* mouse model (MGI:1201412) (Blake et al., 2014, Fritsche et al., 2016).

Here, a decrease in *RDH5* expression in both monocytes (rs3138141 EA = A, p value = 5.512 x 10^{-12}, beta = -0.780, SE = 0.113) and T cells (rs3138141, p value= 1.931 x 10^{-10}, beta = -0.760, SE = 0.119) corresponds to an increase in AMD risk (rs3138141, beta = 0.15). *RDH5* expression is widely expressed beyond ocular tissues and the specialised retinal RPE-tissue (Wang et al., 1999). Interestingly, rs3138141 has shown to be a significant eQTL in over 10 different tissues from the (G. TEx Consortium, 2015) but I did not identify colocalisation of this locus in neutrophils. Instead a seemingly independent QTL was associated with neutrophil *RDH5* expression but not with AMD risk (rs142106092, *RDH5* p=4.536 x 10^{-06}, AMD p=0.147, rs3138141 r^2 < 0.2 1000G). The functional significance of a specialised gene with a fairly ubiquitous effect remains unclear, but I did observe significant enrichment of AMD variants in monocyte eQTLs (Figure 2.6), providing evidence that among the ubiquitous expression, the monocyte-derived expression effect may be more likely to be disease-relevant. Certainly, there is a well-known role for Vitamin A, of which 11-cis-retinal is a derivative, in regulating the immune system and a further immune-link was demonstrated by reduced RPE *RDH5* expression in an *in vitro* system as a result of TNFα secretion from activated pro-inflammatory CD14$^+$ monocytes (Mora et al., 2008, Mathis et al., 2017). Further complicating mechanistic interpretation, this locus colocalised with a more significant gene expression effect in monocyte (p=9.234 x 10^{-34}), T cell (p=6.880 x 10^{-18}) and neutrophil (1.427 x 10^{-24}) of a second gene, *RP11-644F5.10*, for which there is no current characterised function. In addition, *RP11-644F5.10* and *RDH5* are directly overlapping, and rs3138141 is located within both genes.

The disease loci that colocalised with well-studied genes confirmed the validity of my analytical approach. I have also provided functional evidence for loci that may have previously been suggested without eQTL evidence and often based on genomic proximity to the lead SNP. Where eQTL data was used, utilising specific cell populations provides tractable and specific cellular models for functional follow up. Supplementary Figure 2.3 shows the regional association plots for all features colocalised at the loci described here and Supplementary Table 2.2 lists all of the colocalised features for all marks (the most significant colocalised features per disease locus is summarised in Figure 2.8).

	SNP	Locus	Mark	Feature	M	N	T
AD	rs10792832	*PICALM*	gene	*PICALM*	○	◐	○
			H3K27ac	11:85865916:85876777	●	●	○
			H3K27ac	11:85843685:85857459	◐	◐	●
			H3K4me1	11:85823185:85935960	○	●	◐
AD	rs10948363	*CD2AP*	gene	*CD2AP*	◐	◐	○
			H3K4me1	6:47512736:47517132	✕	✕	●
AD	rs11771145	*EPHA1*	gene	*EPHA1*	◐	○	◐
			gene	*EPHA1-AS1*	●	●	◐
			gene	*TAS2R41*	●	●	◐
			gene	*TAS2R62P*	✕	●	◐
			gene	*TAS2R60*	●	●	◐
			H3K27ac	7:143133149:143136359	●	○	◐
			H3K4me1	7:143052447:143144656	●	●	◐
			psi	*EPHA1-AS1*	✕	✕	●
AD	rs6656401	*CR1*	gene	*CR1*	◐	○	○
			psi	*CR1*	●	●	○
AD	rs6733839	*BIN1*	gene	*BIN1*	◐	✕	◐
			H3K27ac	2:127805979:127846262	◐	◐	●
AD	rs983392	*MS4A6A*	gene	*MS4A6A*	●	◐	◐
			gene	*MS4A4E*	●	✕	✕
			gene	*MS4A4A*	●	✕	✕
			H3K27ac	11:60072491:60079697	●	○	○
			H3K27ac	11:59932786:59958180	○	●	○
			H3K4me1	11:59867337:59870007	◐	●	✕
			H3K4me1	11:60097581:60108825	●	○	○
			psi	*MS4A6A*	●	✕	✕
AMD	rs10033900	*CFI*	gene	*CFI*	✕	✕	✕
			gene	*HIGD1AP14*	○	✕	●
AMD	rs11080055	*VTN*	gene	*VTN*	✕	✕	✕
		TMEM97	gene	*TMEM97*	◐	✕	◐
			gene	*SARM1*	○	●	◐
			gene	*TMEM199*	●	●	○
			H3K27ac	17:27592619:27623928	○	○	●
AMD	rs1142	*KMT2E*	gene	*KMT2E*	○	○	○
		SRPK2	gene	*SRPK2*	●	●	◐
			H3K27ac	7:104840590:104849222	●	○	○
			H3K27ac	7:104982791:105001752	○	●	○
			H3K4me1	7:104817678:105045953	●	○	○
AMD	rs140647181	*COL8A1*	gene	*COL8A1*	✕	✕	✕
			gene	*TBC1D23*	●	○	◐
			H3K27ac	3:99927877:99930243	✕	✕	●
			psi	*TBC1D23*	◐	●	◐
AMD	rs142450006	*MMP9*	gene	*MMP9*	◐	○	✕
			H3K4me1	20:43711858:43736607	○	●	◐
AMD	rs1626340	*TGFBR1*	gene	*TGFBR1*	◐	◐	◐
			H3K27ac	9:101863436:101906508	○	●	○
			H3K4me1	9:102394208:102401582	○	○	●
			H3K4me1	9:101924908:101933600	○	●	◐
AMD	rs201459901	*C20orf85*	gene	*C20orf85*	✕	✕	✕
			H3K27ac	20:55937467:55942133	●	○	○
AMD	rs3138141	*RDH5*	gene	*RDH5*	●	◐	●
		CD63	gene	*CD63*	◐	○	○
			gene	*RP11-644F5.10*	●	●	●
			psi	unknown_5611300	✕	●	✕
			psi	unknown_5611300	✕	●	✕
AMD	rs5817082	*CETP*	gene	*CETP*	●	◐	○
			H3K27ac	16:56996497:57122386	●	◐	◐
			H3K4me1	16:56989796:57234898	●	◐	○
AMD	rs61985136	*RAD51B*	gene	*RAD51B*	○	◐	○
			H3K27ac	14:68744958:68766276	○	●	●
			H3K27ac	14:68807602:68809821	●	✕	○
			H3K4me1	14:68786646:68813506	●	○	○
			H3K4me1	14:68705681:68768652	○	●	○
		ADAMTS9-			✕	✕	✕
AMD	rs62247658	*AS2*	gene	*ADAMTS9-AS2*	✕	✕	✕
			H3K27ac	3:64800574:64803236	✕	●	✕
			H3K4me1	3:64807275:64815846	●	◐	◐
AMD	rs6565597	*NPLOC4*	gene	*NPLOC4*	◐	◐	◐
		TSPAN10	gene	*TSPAN10*	✕	✕	◐
			H3K27ac	17:79578768:79583302	◐	○	●
			H3K27ac	17:79585940:79590489	●	●	○
			H3K4me1	17:79594768:79608902	●	○	●
			psi	*NPLOC4*	●	✕	●

M	N	T	
●	●	●	Colocalised
◐	◐	◐	Significant QTL
○	○	○	Non-significant QTL
✕	✕	✕	Not tested

					M	N	T
AMD	rs67538026	*CNN2*	gene	*CNN2*	●	◐	◐
			gene	*ABCA7*	◐	◐	●
			gene	*CTB-31O20.2*	○	●	○
			H3K27ac	19:1024681:1033920	●	○	●
AMD	rs72802342	*CTRB2*	gene	*CTRB2*	–	–	–
		CTRB1	gene	*CTRB1*	–	–	–
			H3K27ac	16:75298651:75302321	–	●	○
			H3K4me1	16:75230954:75238400	○	●	◐
			H3K4me1	16:75294112:75310094	●	◐	○
AMD	rs7803454	*PILRB*	gene	*PILRB*	●	◐	●
		PILRA	gene	*PILRA*	○	○	●
			gene	*AC005071.2*	○	●	–
			gene	*RP11-758P17.2*	○	◐	●
			gene	*ZCWPW1*	○	○	●
			gene	*MEPCE*	●	●	●
			gene	*STAG3*	●	●	◐
			H3K4me1	7:99906073:99912427	●	●	–
			psi	unknown_9993580	●	●	●
			psi	*PILRB*	◐	◐	●
			psi	*PILRB*	●	●	–
AMD	rs79037040	*TNFRSF10A*	gene	*TNFRSF10A*	●	◐	◐
			gene	*RP11-1149O23.3*	●	–	○
			gene	*CHMP7*	●	◐	◐
			H3K27ac	8:23048166:23092260	●	◐	◐
			H3K4me1	8:22998146:23133613	●	◐	○
CAD	chr2:203828796:I	*WDR12*	gene	*WDR12*	○	◐	◐
			gene	*AC073410.1*	○	–	●
			gene	*NBEAL1*	●	◐	◐
			gene	*ALS2CR8*	●	○	◐
			gene	*ICA1L*	●	–	◐
			H3K27ac	2:204364327:204367436	◐	●	○
			H3K4me1	2:204391511:204403180	●	○	○
CAD	rs11065979	*SH2B3*	gene	*SH2B3*	○	◐	○
			H3K27ac	12:112386996:112391985	○	–	●
CAD	rs11191416	*NT5C2*	gene	*NT5C2*	○	●	◐
		CYP17A1	gene	*CYP17A1*	–	–	–
		CNNM2	gene	*CNNM2*	◐	○	○
			gene	*RP11-332O19.2*	○	●	○
			H3K27ac	10:104811999:104815290	○	○	●
			psi	*NT5C2*	●	–	●
CAD	rs1412444	*LIPA*	gene	*LIPA*	●	◐	●
			H3K27ac	10:90248309:90252291	○	◐	●
			H3K27ac	10:90993615:91006217	●	●	○
			H3K4me1	10:90987967:91024823	●	●	○
CAD	rs17087335	*REST*	gene	*REST*	●	◐	◐
		NOA1	gene	*NOA1*	◐	◐	◐
			H3K27ac	4:57823529:57826313	●	●	○
			H3K4me1	4:57820927:57828891	●	●	○
CAD	rs1870634	*CXCL12*	gene	*CXCL12*	–	–	–
			H3K27ac	10:44339141:44344636	●	○	○
			H3K27ac	10:44499917:44501820	–	●	–
			H3K27ac	10:44468665:44477554	◐	◐	●
CAD	rs2487928	*KIAA1462*	gene	*KIAA1462*	–	–	–
			H3K27ac	10:30314435:30318729	○	○	●
			H3K4me1	10:30286485:30293313	○	○	●
CAD	rs28451064	*KCNE2*	gene	*KCNE2*	–	–	–
		(gene desert)	H3K27ac	21:35594186:35597126	●	○	–
			H3K27ac	21:35444064:35452944	◐	●	◐
			H3K27ac	21:35389093:35398453	○	◐	●
			H3K4me1	21:35592772:35599590	●	○	○
CAD	rs4468572	*ADAMTS7*	gene	*ADAMTS7*	–	–	–
			H3K27ac	15:79049034:79056595	○	◐	●
			H3K4me1	15:79121511:79127959	●	●	○
			H3K4me1	15:79029778:79035218	◐	◐	●
CAD	rs56289821	*LDLR*	gene	*LDLR*	○	○	○
			H3K4me1	19:11105519:11214483	◐	●	○
CAD	rs6689306	*IL6R*	gene	*IL6R*	○	○	◐
			H3K27ac	1:154372031:154419908	◐	●	◐
			H3K4me1	1:154342399:154479953	○	●	○
			psi	*IL6R*	◐	◐	●

M	N	T	
●	●	●	Colocalised
◐	◐	◐	Significant QTL
○	○	○	Non-significant QTL
–	–	–	Not tested

CAD	rs67180937	*MIA3*	gene	*MIA3*
			H3K4me1	1:222943024:222949002
			psi	*MIA3*
CAD	rs7212798	*BCAS3*	gene	*BCAS3*
			H3K27ac	17:58166053:58170841
CAD	rs7528419	*SORT1*	gene	*SORT1*
			gene	*PSRC1*
			H3K27ac	1:109109862:109115257
			H3K27ac	1:109812607:109818851
			H3K4me1	1:109779241:109861456
CAD	rs7568458	*VAMP8*	gene	*VAMP8*
		GGCX	gene	*GGCX*
		VAMP5	gene	*VAMP5*
			H3K27ac	2:85760296:85771243
			H3K4me1	2:85523177:85561159
CAD	rs9349379	*PHACTR1*	gene	*PHACTR1*
			H3K27ac	6:12961893:12964068
			H3K4me1	6:12953822:12975623
			H3K4me1	6:13023419:13036119
CAD	rs9970807	*PPAP2B*	gene	*PPAP2B*
			H3K27ac	1:56969801:56978117
			H3K27ac	1:56931078:56933449
FEV1	rs7218675	*TSEN54*	gene	*TSEN54*
FEV1	rs78420228; rs67863175	*CDC123*	gene	*CDC123*
			gene	RP11-186N15.3
			H3K27ac	10:12310277:12315701
			H3K4me1	10:12273289:12320006
SLE	rs10028805	*BANK1*	gene	*BANK1*
			H3K27ac	4:102711289:102714021
			H3K4me1	4:102752891:102758975
			H3K4me1	4:102739046:102740746
SLE	rs10488631	*IRF5*	gene	*IRF5*
			H3K27ac	7:128720370:128724585
			H3K27ac	7:128733525:128737997
SLE	rs10774625	*SH2B3*	gene	*SH2B3*
			H3K27ac	12:112386996:112391985
SLE	rs11644034	*IRF8*	gene	*IRF8*
			H3K4me1	16:85911735:86006346
SLE	rs11889341	*STAT4*	gene	*STAT4*
			H3K27ac	2:190816300:190818184
SLE	rs12802200	*IRF7*	gene	*IRF7*
			gene	*PHRF1*
			gene	*LRRC56*
			gene	*C11orf35*
			H3K27ac	11:600961:621989
			H3K4me1	11:601613:633623
			psi	*IRF7*
SLE	rs1801274	*FCGR2A*	gene	*FCGR2A*
			psi	*FCGR3A*
			psi	*FCGR2A*
SLE	rs2304256	*TYK2*	gene	*TYK2*
SLE	rs2663052	*WDFY4*	gene	*WDFY4*
			H3K27ac	10:49965615:49980674
SLE	rs2732549	*CD44*	gene	*CD44*
			H3K27ac	11:35087097:35089761
SLE	rs2736340	*BLK*	gene	*BLK*
			H3K27ac	8:11348864:11353299
			H3K4me1	8:11345910:11367069
			H3K4me1	8:11336396:11344630
SLE	rs34572943	*ITGAM*	gene	*ITGAM*
			gene	*C16orf58*
			gene	*RP11-388M20.2*
			gene	*RP11-347C12.10*
			H3K4me1	16:31355247:31421179
SLE	rs7444	*UBE2L3*	gene	*UBE2L3*
			H3K27ac	22:21938482:21985305
			H3K4me1	22:21917050:22033004
			H3K4me1	22:22399916:22404237

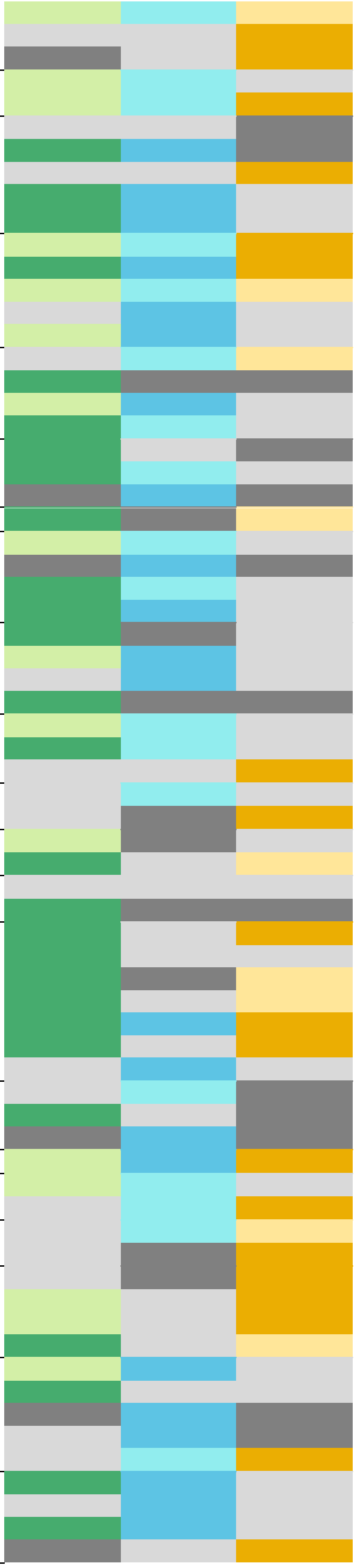


Figure 2.8 Disease loci colocalised with multiple molecular features
Summary of 54 disease loci colocalised with at least one feature. Where multiple features of the same type colocalised, the most significant feature per cell type is given. For each locus, the status of the previously reported locus is given. The most significant SNP per locus as denoted by the study is given. Colocalisation is the dark colour, non-colocalised significant QTLs, lighter colour, non-significant QTL by light grey and not tested, dark grey. The FEV$_1$ *KANSL1* inversion locus is excluded.

2.3.4 Therapeutic utility of colocalised gene targets

I have discussed how genetic studies of human cohorts provide the design of a randomised clinical trial without complications of intervention or reverse causation as genotype is randomly determined at birth (Finan et al., 2017). However, case-control GWAS are still limited in the identification of the exact mechanistic targets. Here, I have leveraged the advantage of GWAS data to identify risk-associated regions together with molecular traits to precisely detect potential putative target genes. I next evaluated whether any of the genes I identified through expression or splicing effects for all colocalised loci (excluding the *KANSL1* inversion locus) were known drug targets for the disease of interest or with other disorders highlighting the potential for drug repurposing (Finan et al., 2017). I collated information from both DrugBank (DrugBank, 2017) and the Open Targets (Open Targets, 2017) platform and identified 11 genes, which are targets for compounds or drugs that are currently under investigation, approved or experimental (Table 2.3) (Law et al., 2014, Koscielny et al., 2017). These included three CAD risk genes; *IL6R, GGCX, NT5C2,* three AMD risk genes; *RDH5*, *TNFRSF10A*, *CETP*, *SRPK2* and four SLE risk genes; *BLK*, *TYK* and the closely related and located *FCGR3A/FCGR2A* (Table 2.3).

I also queried my colocalised genes with a recently updated curation of the “druggable genome” resource from the Finan et al. (2017) study. This resource is composed of three tiers of genes predicted to encode druggable proteins including recent first-in-class drugs, biotherapeutics, drugs in late-phase development at the time of publication, preclinical small molecules with potential to bind proteins as reported in ChEMBL, secreted or plasma membrane-bound proteins that represent ideal targets for monoclonal antibodies and proteins that have greater than 50% identity with approved drug targets (Finan et al., 2017). Using this resource, I identified a further six druggable genes including the AMD *ABCA7* (small molecular or biotherapeutic), CAD *LIPA* (small molecule), CAD *VAMP8* (biotherapeutic), AMD *SRPK2* (small molecule), SLE *ITGAM* (biotherapeutic) and AD *CR1* (biotherapeutic).

In total, there was evidence of potential therapeutic utility for 17 genes identified in my analysis. In conclusion, such colocalisation approaches using molecular QTLs provides an additional layer of genetic and functional evidence for the selection of pre-clinical drug targets while also highlighting potentially affected cell types for further testing (Glinos et al., 2017, Okada et al., 2014).

Gene/Target	Disease	Drug	Target Mechanism	Status	Treated Disease	Accession /ChEMBL
NT5C2- Cytosolic purine 5’-nucleotidase (HGNC Acc:8022)	CAD	ATP, small molecule.	Unknown	-	Nutraceutical	DB00171
		Ribavirin, small molecule, guanosine nucleoside interferes with synthesis of viral mRNA,	Inducer	A	Hepatitis C, viral haemorrhagic fevers	DB00811
IL6R- Interleukin-6 receptor subunit alpha (HGNC Acc:6019)	CAD	Tocilizumab, antagonist, inhibits IL6R alpha subunit	Antibody	A	RA, SJIA, schizophrenia, temporal arteritis, AML, HIV, immune system disease	DB06273
		Sarilumab, antagonist, inhibits IL6R alpha subunit	Antibody	PIV	RA, ankylosing spondylitis, uveitis, immune system disease	DB11767
		SA237, antagonist, IL6Ralpha/GP130	Antibody	PIII	Neuromyelitis optica,	CHEMBL3833307
GGCX- Vitamin K dependent gamma-carboxylase (HGNC Acc:4247)	CAD	Phylloquinone/Vitamin K1 small molecule	Inducer	A	Haemorrhagic conditions	DB01022
		Anisindione, small molecule anticoagulant	Inhibitor	A	Venous thrombosis, embolism	DB01125
		Menadione/Vitamin K3, small molecule	Cofactor	A	Nutraceutical	DB00170
		Coagulation factor VIIa Recombination human promoting hemostasis	Unknown	A	Haemorrhagic complications	DB00036
		Drotrecogin alfa, recombinant activated human protein C	Unknown	A, I, W	Sepsis (withdrawn)	DB00055
		Coagulation Factor IX (Recombinant)	Unknown	A	Factor IX deficiency	DB00100
		Glutamic Acid	Unknown	A	Nutraceutical	DB00142
		Coagulation Factor IX Human, serine protease	Unknown	A	Factor IX deficiency	DB13152
TYK2- Non-receptor tyrosine-protein kinase (HGNC Acc:12440)	SLE	Tofacitinib, small molecule antagonist, inhibits janus kinases	Inhibitor	A, I	RA, immune system disease, UC, psoriasis, CD,	DB08895
		2-(1,1-DIMETHYLETHYL)9-FLUORO-3,6-DIHYDRO-7H-BENZ[H]-IMIDAZ[4,5-F]ISOQUINOLIN-7-ONE, small molecule	Unknown	E	-	DB04716

		Cerdulatinib, small molecule antagonist, tyrosine kinase inhibitor	Inhibitor	PI	Non-hodgkins lymphoma, chronic lymphocytic leukemia	CHEMBL3545284
		Peficitinib, small molecule antagonist, tyrosine kinase inhibitor	Inhibitor	PIII	RA, psoriasis, liver disease	CHEMBL3137308
FCGR3A- Fc fragment of IgG, low affinity IIIa, receptor/CD16a (HGNC Acc:3619) *FCGR2A*- Fc fragment of IgG, low affinity IIa, receptor/CD32 (HGNC Acc:3616)	SLE	Cetuximab, antibody binds EGFr, HER1, c-ErbB-1 and competitively inhibits binding of EGF	Unknown	A	EGFR-expressing metastatic colorectal carcinoma	DB00002
		Etanercept, protein binds to TNF	Unknown	A, I	RA, psoriasis	DB00005
		Immune Globulin Human, antibody mix binds and kills bacteria and viral particles	Antagonist	A, I	Immunodeficiencies	DB00028
		Adalimumab, human monoclonal binds and blocks TNF-alpha reducing inflammation	Unknown	A	RA, CD, psoriatic arthritis, ankylosing spondylitis	DB00051
		Abciximab, antibody binds to glycoprotein IIb/IIIa receptor and inhibits platelet aggregation	Unknown	A	Coronary intervention	DB00054
		Gemtuzumab oxogamicin, antibody binds and kills CD33 leukemic cells	Unknown	A, I, W	AML	DB00056
		Trastuzumba, antibody binds human epidermal GF receptor inhibits proliferation of tumour cells	Unknown	A, I	HER2 Breast cancer	DB00072
		Rituximab, antibody binds CD20 and kills B cells	Unknown	A	CD20^{+}non-hodgkins lymphoma, chronic lymphocytic leukaemia, RA	DB00073
		Basiliximab, immunosuppressive binds IL-2R alpha	Unknown	A, I	Prevent kidney transplant rejection	DB00074
		Muromonab, binds to and kills CD3^{+} cells	Unknown	A, I	Prevent organ rejection	DB00075
BLK- B lymphoid tyrosine kinase (HGNC Acc:1057)	SLE	Dasatinib, small molecule antagonist, SRC inhibitor	Inhibitor	A	Neoplasm, leukaemia, lymphoma	CHEMBL1421
		Ilorasertib, small molecule antagonist, SRC kinase inhibitor	Inhibitor	PII	Neoplasms	CHEMBL1980297
		ENMD-981693, small molecule antagonist, SRC inhibitor	Inhibitor	PII	Pancreatic carcinoma, breast cancer	CHEMBL52885
		XL-228, small molecular, SRC inhibitor	Inhibitor	PI	Lymphoma, leukemia	CHEMBL3545085

TNFRSF10A- Tumor necrosis factor receptor superfamily, member 10a (HGNC Acc:11904)	AMD	Dulanermin, protein, TNFRSF10A/B agonist	Agonist	PII	Non-Hodgkins lymphoma, lung/colorectal carcinoma,	CHEMBL2107846
		Mapatumumab, TNFRSF10A agonist	Agonist	PII	Non-Hodgkins lymphoma,myeloma, various carcinoma	CHEMBL2108621
CETP Cholesteryl ester transfer protein, plasma (HGNC Acc:1869)	AMD	Evacetrapib, small molecule	Inhibitor	PIII	CVD, lipid, hypercholesterolemia	CHEMBL2017179
		Anacetrapib, small molecule	Inhibitor	PIII	CHD, CVD, lipid, hypercholesterolemia	CHEMBL1800807
		Dalcetrapib, small molecule	Inhibitor	PIII	Acute coronary syndrome, CHD, CVD	CHEMBL313006
SRPK2- SRSF protein kinase 2	AMD	Adenine, small molecule	Unknown	A	Nutraceutical	DB00173
		Purvalanol, small molecule	Unknown	E	-	DB02733
		Phosphoaminophosphonic acid-adenylate ester, small molecule inhibits ATPase	Unknown	E	-	DB04395
RDH5- 11-cis retinol dehydrogenase (HGNC Acc:9940)	AMD	NADH, small molecule	-	-	Nutraceutical, possible PD, AD, CVD	DB00157
		Vitamin A, small molecule.	Unknown	A	Nutraceutical	DB00162

Table 2.3: Colocalised genes that are also known drug targets
Table summarises colocalised genes from expression and/or splicing effects that are known drug targets using DrugBank version 5.0.9, accessed August 2017 and Open Targets Platform, accessed October 2017 (Koscielny et al., 2017). The gene name was used to search the platforms for known targets. The drug, status, mechanism of action on target, diseases for which the drug is used, accession number (DrugBank, 2017, Law et al., 2014) or CHEMBL reference (Bento et al., 2014) are listed. Nutraceutical is a food source that provides health benefit. Drug status is listed as A = approved, W = withdrawn, I = investigational, E = experimental or maximum clinical trial either completed, ongoing or terminated (PI/II/III). The diseases for which the drug is currently used or investigated are also given or some example diseases where multiple conditions have been investigated. Disease abbreviations: PD = Parkinson’s disease, AD = Alzheimer’s disease, CVD = cardiovascular disease, RA = rheumatoid arthritis, SJIA = systemic juvenile idiopathic arthritis, CD = Crohn’s disease. SRC inhibitor = src family kinase inhibitor, CD = Corhn’s disease, UC = ulcerative colitis, AML = acute myeloid leukemia.

2.3.5 Complex regulatory mechanisms highlighted through integration of multiple molecular evidence

The combination of gene expression, splicing and histone QTLs enables not only a deeper description of disease risk mechanisms but is also a valuable tool in understanding how genes are regulated under homeostatic conditions. Above I summarised broad themes from my analysis, but a definitive description of the regulatory mechanism at each locus requires in-depth analysis and integration of multiple data sources. Here, I discuss approaches to distinguish likely putative regulatory mechanisms from multiple colocalised phenotypes involving complex histone activity, transcription factor binding and non-coding RNA function, demonstrating the importance of in-depth investigation at every disease locus.

2.3.5.1 Cell-type specific regulatory activity at the CAD *SORT1* locus

Not all loci colocalised with genes that matched previous predictions. One clear example was the CAD *SORT*1 locus, where I identified colocalisation with monocyte and neutrophil *PSRC1* gene expression as well as H3K27ac and H3K4me1 signal in these cell types (Figure 2.9). *SORT1*, a gene which encodes the multi-ligand sortilin receptor protein, has been previously identified as the liver target gene of the causal SNP, rs12740374 (Musunuru et al., 2010). There was also a significant *PSRC1* expression effect in human liver, but the largest observed effect was with the expression of *SORT1* (Musunuru et al., 2010). Increased hepatic *Sort1* expression in mice was also shown to modulate hepatic very-low density lipoprotein (VLDL) secretion resulting in reduced secretion of LDL-C therefore lowering LDL levels, which is known to decrease CAD risk (Musunuru et al., 2010). These effects were recently reproduced in iPSC-differentiated hepatocyte-like cells (HLCs) from 68 lines (Warren et al., 2017). Intracellular metabolites were extracted from these HLCs, and it was demonstrated that in minor allele-carrying individuals (rs12740374, T), there was a significant decrease in lipid metabolites such as triacylglycerol, diacylglycerol and aminoadipic acid, which has been associated with CAD (Warren et al., 2017). There is, therefore, clear evidence that in liver cells, *SORT1* hepatic expression and Sortilin protein levels were significantly associated with rs12740374 genotype and that this protein has a causal role in lipid regulation conferring protection (minor allele, T) to CAD risk. How sortilin exactly modifies lipid phenotypes is not yet clear and will require further experimental investigation (Kjolby et al., 2015).

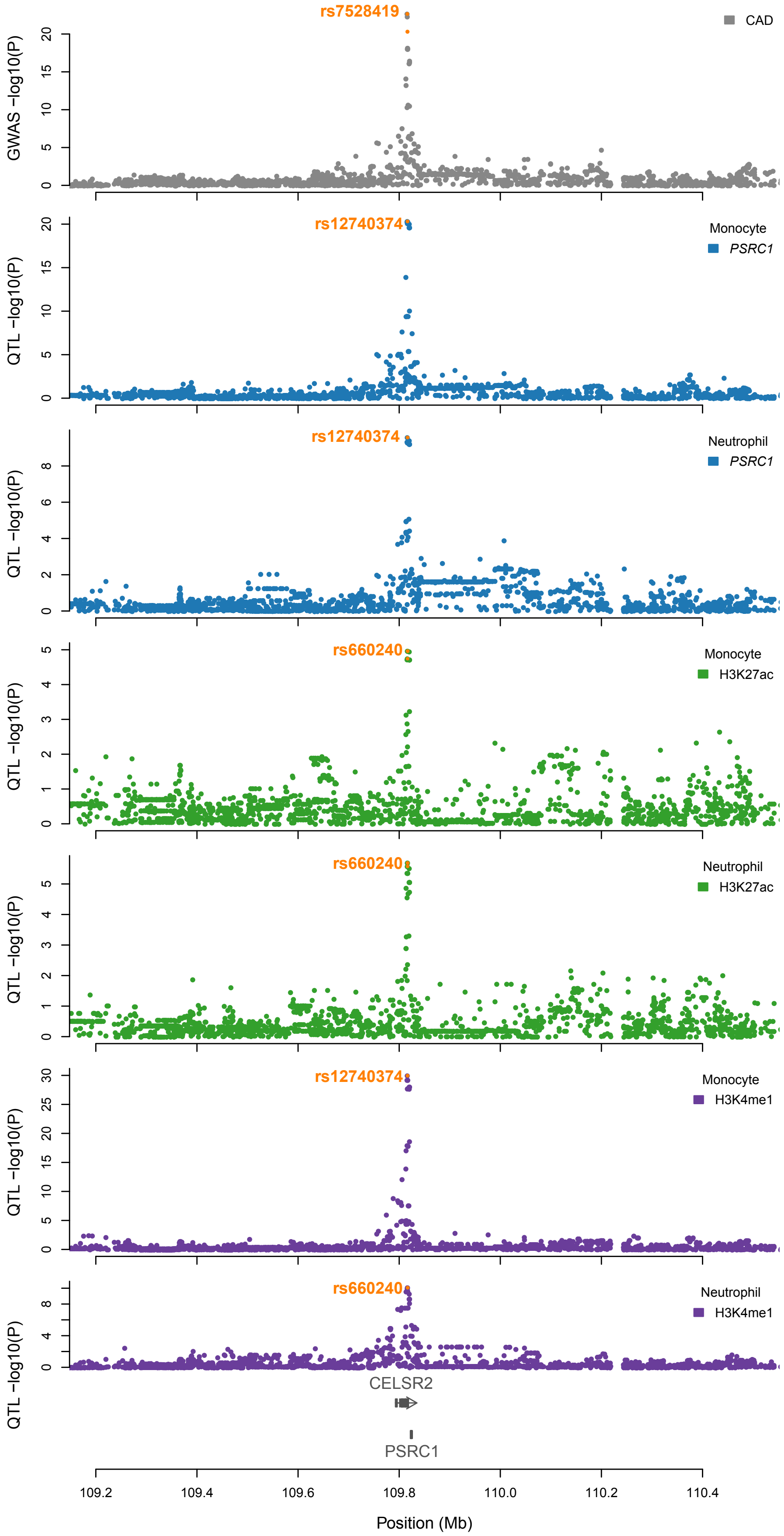


Figure 2.9: Regional association plot of the CAD *SORT1* locus
Locus zoom plots show association strength in -log10(p value) for variants that were shared between studies testing CAD (grey), monocyte gene expression (blue), monocyte H3K27ac (green) and monocyte H3K4me1, (purple). The respective lead SNPs for each feature are highlighted in orange.

In my analysis with the BLUEPRINT cohort, I identified that increased *PSRC1* expression (rs12740374 EA = T, monocyte p value = 4.364 x 10^{-21}, beta = 1.016 SE = 0.108, neutrophil p value = 2.572 x 10^{-10}, beta = 0.723, SE = 0.114) corresponds to a decreased risk of CAD (rs12740374 EA = T, p value = 4.63 x 10^{-23}, beta = -0.114, SE = 0.011). The predicted causal SNP, rs12740374 is located upstream of the *PSRC1* gene in the 3'UTR of the *CELSR2* gene (Figure 2.10). Compared to *SORT1*, very little is known regarding the function of *PSRC1*, which encodes a proline/serine-rich coiled coil protein 1 (Kjolby et al., 2015).

In monocytes and neutrophils, I identified that *SORT1* was expressed, but was not significantly associated with an eQTL. Both *PSRC1* and *SORT1* expression were low in T cells and below the threshold for association testing. The locus also colocalised with monocyte and neutrophil H3K4me1 and H3K27ac QTLs for peaks located just upstream of the *PSRC1* gene (Figure 2.10). The difference in modification level between individuals of discordant genotype was greater for H3K4me1 than H3K27ac (Figure 2.10) and the association more significant in both monocytes and neutrophils (Figure 2.9). Combined with the upstream location relative to the gene of these histone modifications, I postulated that this region acted as an enhancer for both *PSRC1* and *SORT1.* Genetic disruption of H3K4me1 could, in turn, alter downstream *PSRC1* gene expression in monocytes and neutrophils.

I investigated whether differences in regulatory function may explain the difference in the primary gene targets between haematopoietic and hepatic cell types, i.e. why the strongest effect was *PSRC1* in monocytes and neutrophils but as previously established, *SORT1* in liver (Musunuru et al., 2010). First, I used ENCODE ChIP-seq data from the hepatocyte cell line HepG2 to show that there was equivalent histone signal in the region overlapping rs12740374 as well as proximal to the *SORT1* promoter compared to both monocytes and neutrophils (Figure 2.10). Therefore, differences in enhancer activity seemed not explain the observation that *PSRC1* was significantly associated with rs12740374 and not *SORT1*. Instead, additional regulatory factors may generate cell-type genetic regulation of gene expression.

Binding of the liver-enriched transcription factor (TF), C/EBPα, to the motif containing the rs12740374 SNP was previously demonstrated in cultured human hepatocellular carcinoma cells (Hep3B) (Musunuru et al., 2010). Further, the major allele of rs12740374, G, disrupts a nucleotide of the C/EBP motif, G***TT***GCTC*AA*T, where *TT* and *AA* are the consensus nucleotides (Musunuru et al., 2010). The liver-enriched TF, C/EBPα, bound to the minor allele and directly affected *SORT1* expression levels in Hep3B cells (Musunuru et al., 2010). In addition to regulating many metabolic liver genes, C/EBPα is essential for granulopoiesis, regulating important genes such as the granulocyte-stimulating factor receptor (G-CSFR), but its expression decreases as maturation advances (Bardoel et al., 2014, Jakobsen et al., 2013). I hypothesised that binding of C/EBPα may be divergent between hepatocyte and haematopoietic cells as differential TF binding at regulatory regions is important for generating cell-type specific gene expression (Hardison and Taylor, 2012, Heinz et al., 2015). First, I used ChIP-seq data from the Soranzo team of C/EBPα in the monocytic cell line, U937 (unpublished data) to investigate regions of binding. I used two repeats of C/EBPα U937 ChIP-seq data and demonstrated that there was no equivalent C/EBPα binding directly over rs12740374 in U937 cell lines (peaks are represented by blocks as designated by a ChIP-seq peak caller) (Figure 2.10).

I further investigated whether different TFs may bind at the rs12740374 locus, instead of C/EBPα. C/EBPβ is a member of the same basic region leucine zipper-family (bZIP) as C/EBPα and as well as playing an important role in regulating chromatin dynamics and regrowth in liver, it is known to have an important role in haematopoietic cell differentiation and function (Jakobsen et al., 2013, Grontved et al., 2013, Bardoel et al., 2014). C/EBPα and C/EBPβ also bind to the same motif (Jakobsen et al., 2013). In addition, the TF, PU.1, is a crucial factor in promoting lymphomyeloid differentiation and acts as a pioneer factor binding to nucleosomes and preceding deposition of H3K4me1 (Heinz et al., 2010). Based on these observations, I postulated that this highly functional region could be bound by different combinations of lineage-specific master regulators in alternative cell types.

In order to establish whether these factors were bound, I generated C/EBPβ and PU.1 binding data using ChIP-seq in a differentiated HL60 cell line (Materials and Methods). HL60 is an immortalised cancer cell line established from a patient with acute myeloid leukaemia and is thought to resemble the granulocyte precursors, promyelocytes (Birnie, 1988). Using a well-established method, I differentiated HL60 to a more mature neutrophil-like stage by addition of all trans-retinoic acid (ATRA) or dimethyl sulfoxide (DMSO) (Materials and Methods). Using ChIP-seq, I then generated genome-wide binding data for transcription factors PU.1 and C/EBPβ as well as the histone modifications, H3K27ac and H3K4me1.

Figure 2.11 shows that the neutrophil H3K4me1 modification activity is recapitulated in differentiated HL60, confirming that enhancer activity is likely in this region. Further, there are strong binding peaks shown for C/EBPβ and PU.1 in both differentiated HL60 models. This confirms that master haematopoietic regulators are bound in this region in myeloid cells.

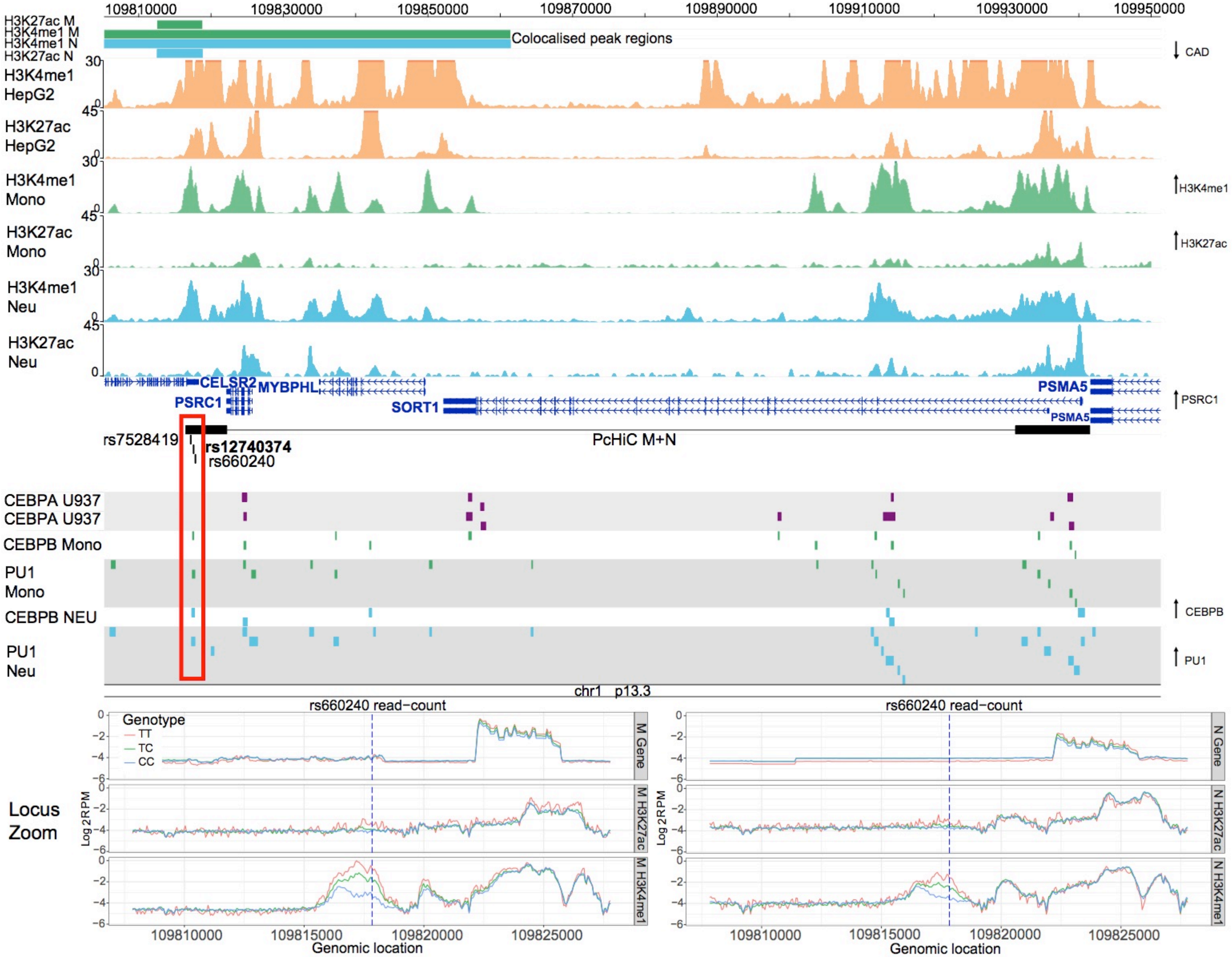


Figure 2.10: Genomic context of the CAD *SORT1* locus

Genome browser plot of the rs12740374 CAD locus with H3K4me1 and H3K27ac colocalised features shown for HepG2 (liver) cells, primary monocytes and neutrophils. Signal peaks are shown for representative individuals from the BLUEPRINT cohort and the x-axis for peaks is given in reads per million. The exact location of the predicted causal SNP, rs12740374, is highlighted in a red box which shows that the SNP intersects with histone marks as well as C/EBPβ and PU.1 in monocytes and neutrophils. C/EBPα binding in the monocyte-like cell line, U937, shows no binding in this region, suggesting differential TF binding at this cis-regulatory element underpins cell-type specific regulation of gene targets. The lower panel shows a zoom in on the region around the SNP upstream of the monocyte and neutrophil colocalised eQTL gene, *PSRC1*. Gene expression, H3K27ac and H3K4me1 signal is shown stratified by genotype. The signal is shown for the lead SNP for the H3K27ac, in this case, rs660240, which is in high LD with rs12740374. The genotype-associated difference in H3K4me1 signal is greater than the signal for H3K27ac, suggesting this is predominantly an enhancer effect due to changes in H3K4me1 activity. The directions of all features at this locus with respect to CAD risk are shown on the right-hand side.

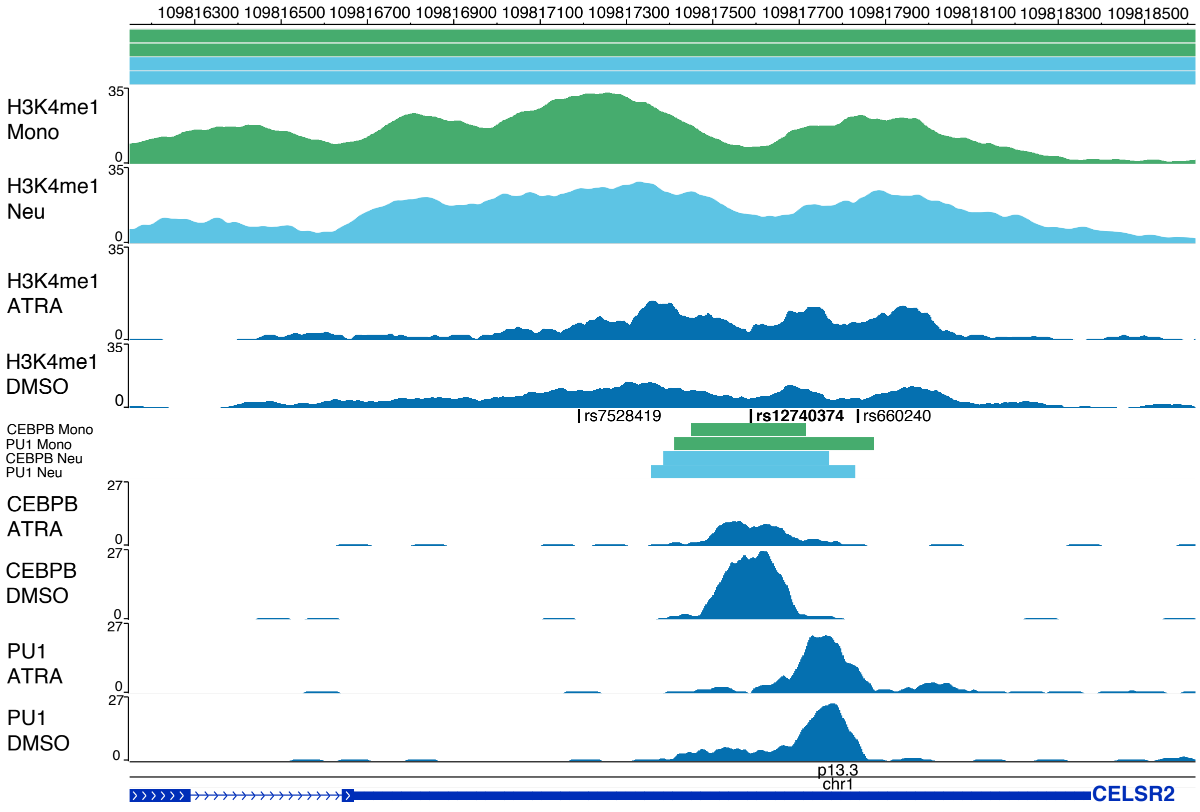


Figure 2.11 C/EBPβ and PU.1 are bound directly over rs12740374 in differentiated HL60 cells

Zoom in of the genomic region of rs12740374, which is located directly within C/EBPβ and PU.1 peaks from differentiated HL60 cells. H3K4me1 peaks for monocyte neutrophil and differentiated HL60 are shown to confirm that, at this locus, this model cell line recapitulates primary human cells. Neighbouring SNPs, rs7528419 (CAD GWAS lead) and rs660240 (certain histone peak lead) are located just outside or on the edge of the peak, showing how using TF data can aid in identifying the putative causal variant at disease loci.

Having demonstrated clear binding signal, I wanted to establish whether rs12740374 disrupts the haematopoietic binding of C/EBPβ and/or PU.1 as was previously shown for C/EBPα in Hep3B cells. In order to assess this, I accessed an unpublished dataset from the Soranzo team of C/EBPβ and PU.1 binding in primary human neutrophils (22 and 93 individuals respectively) and monocytes (nine and ten individuals respectively) (Stephen Watt, manuscript in preparation). Figure 2.12 highlights that in C/EBPβ and PU.1 are also bound directly over rs12740374 in monocytes. The binding of these TFs appeared lower in primary neutrophils than in monocytes and the differentiated HL60 model. This was likely due to the increased technical difficulties associated with applying these approaches in primary neutrophils, as we have observed within our team. This further demonstrates the importance of confirming TF binding using the more tractable, differentiated HL60 model.

Despite this lower level of binding, the higher number of individuals discordant at the rs12740374 genotype in the primary neutrophil cohort enabled me to investigate whether binding was associated with SNP genotype. Figure 2.12 shows primary monocyte and neutrophil binding of C/EBPβ and PU1 around the *PSRC1* locus and also binding of specific peaks stratified by the genotype of this SNP. Both C/EBPβ and PU.1 peaks directly overlapping rs12740374 are significantly associated with genotype as evaluated using linear regression (C/EBPβ, p value = 7.351 x 10^{-04}, PU.1 p value = 1.584 x 10^{-06}). I confirmed that no other immediate surrounding peaks are significantly associated with rs12740374 (Figure 2.12 and data not shown). This evidence suggested that in neutrophils, the major allele of rs12740374 may disrupt binding of PU.1 and C/EBPβ as well as H3K4me1 activity, which could result in disruption of *PSRC1* expression. I also demonstrated that binding of C/EBPβ and PU.1 occurs over rs12740374 in monocytes by showing representative binding of an individual heterozygous for rs12740374 (Figure 2.12). Although there was a limited number of individuals for which monocyte data was available, the concordance of the other molecular effects between monocytes and neutrophils suggests that C/EBPβ and PU.1 binding could also be disrupted in monocytes. However, more individuals would be required to fully validate this effect.

Interestingly, using publicly available promoter-capture HiC (Schofield et al., 2016, Javierre et al., 2016) data, I observed that a significant neutrophil and monocyte chromatin interaction fragment links rs12740374 to the promoter of the *SORT1* gene (Figure 2.10). Despite this physical connection, *SORT1* expression is not significantly associated with this SNP in monocytes or neutrophils. Therefore, this demonstrates that at some loci, the combination of TF bound is an important driving factor over chromatin interactions and enhancer activity in generating cell-type specific gene regulatory mechanisms, although further functional experiments would be required to fully ascertain this potential hierarchical regulation.

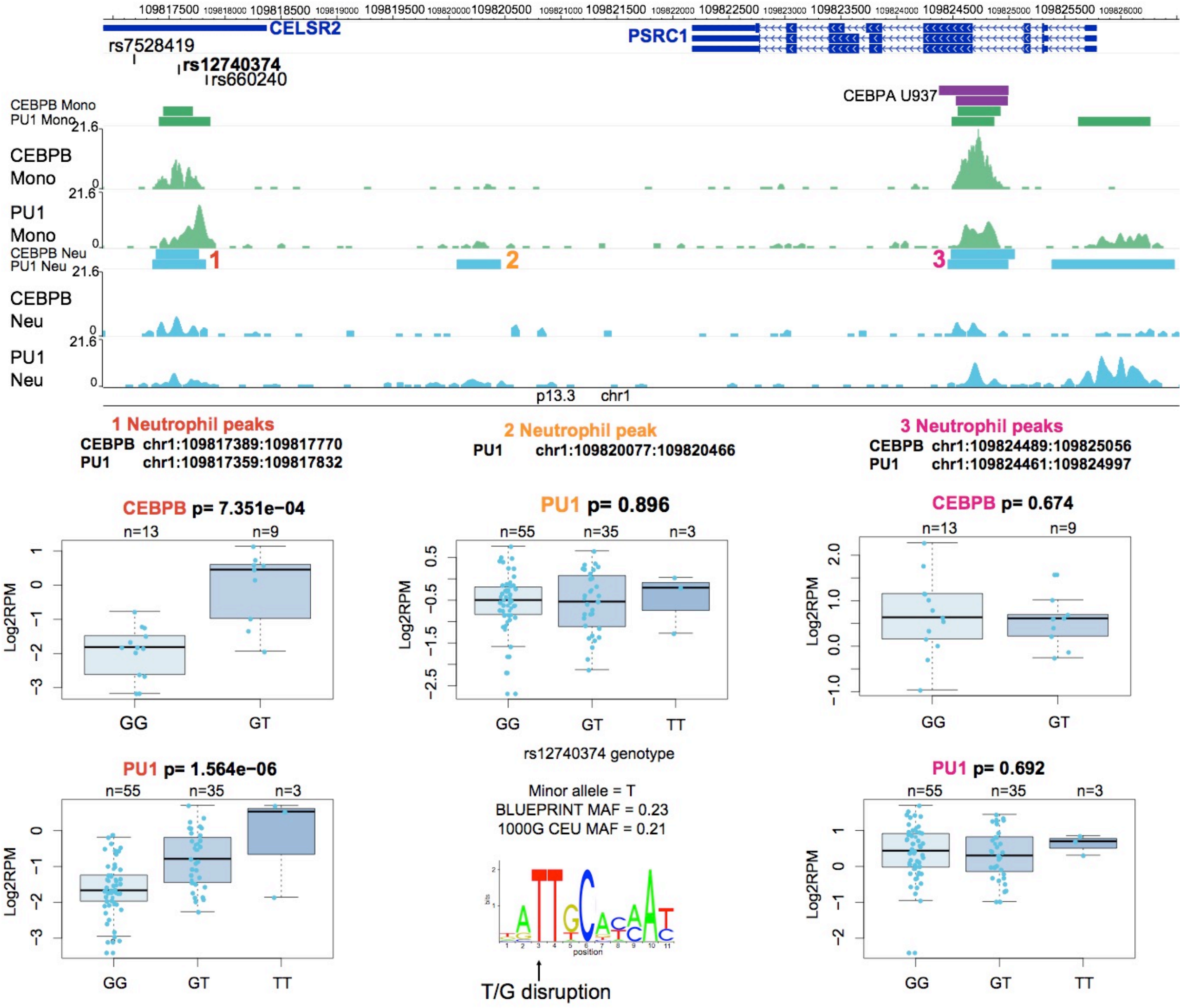


Figure 2.12: Transcription factor binding at the CAD *SORT1/PSRC1* locus in monocytes and neutrophils

Genomic region of *PSRC1* and predicted causal SNP, rs12740374 showing binding of C/EBPβ and PU.1 in monocytes and neutrophils (upper panel). In the lower panel, boxplots show the binding signal in log$_2$RPM of neutrophil TFs at three peaks in the locus stratified by rs12740374 genotype. The p value is shown for the association with genotype as calculated using linear regression on standardised inverse normalised binding values in log$_2$RPM. The only peaks significantly associated with rs12740374 are the C/EBPβ and PU.1 that are bound directly over the SNP. The consensus C/EBPβ motif is also shown with the position of the nucleotide disrupted by rs12740374, where the minor allele T creates the binding site and the major allele, G disrupts the binding site.

2.3.5.2 In-depth dissection of the molecular mechanisms at the AMD *TNFRSF10A* disease locus

I identified colocalisation between the *TNFRSF10A* advanced age-related macular degeneration locus and three monocyte eQTLs; *TNFRSF10A, RP11-1149O23.3* and *CHMP7.* In addition, this locus colocalised with three histone peaks; H3K27ac (8:23048166:23092260), H3K27ac (8.23092704.23132254) and H3K4me1 (8:22998146:23133613) (Figure 2.14). I discuss here approaches to resolve this complex locus and provide paradigms for future efforts to identify mechanisms that influence disease risk in a cell type-specific manner.

TNRFSF10A encodes the TRAILR1 receptor that binds the tumour necrosis factor-related apoptosis-inducing ligand (TRAIL) (Diehl et al., 2004). TRAIL can bind four possible receptors: TNFRSF10A, TNFRSF10B (TRAIL-R2), TNFRSF10C (TRAIL-R3) and TNFRSF10D (TRAIL-R4) (Diehl et al., 2004). TRAIL-R1 and TRAIL-R2 are functional proteins that include an intracellular tail containing the death domain (Figure 2.14) (Diehl et al., 2004). TRAIL-R3, a GPI-linked protein and TRAIL-R4, a truncated protein that misses the death domain in the cytoplasmic tail, are both decoy receptors that do not activate TRAIL-mediated apoptosis and can antagonise TRAILR1-2 signalling (Diehl et al., 2004, Guicciardi and Gores, 2009). Functional TRAILRs activate apoptosis in tumour cells, and were originally not thought to induce cell death in non-transformed cells (Diehl et al., 2004, Liguori et al., 2016). Recently, however, TRAIL susceptibility leading to caspase-8-dependent apoptosis was observed in primary mononuclear phagocytes, where the expression of functional *TNFRSF10A/TRAILR1* was highest compared to the expression on neutrophils and T lymphocytes (Liguori et al., 2016). No caspase-8 activation was observed in neutrophils or lymphocytes (Liguori et al., 2016). Macrophages may be more resistant to death signals as they represent a more activated immune cell than monocytes (Liguori et al., 2016). TRAILR2 seems to have a more important role in stimulating apoptosis than TRAILR1 (Guicciardi and Gores, 2009).

Up-regulation of the *TNFRSF10A/TRAILR1* receptor has been shown to be associated with anti-inflammatory signals such as stimulation by the cytokine IL-10 (Liguori et al., 2016). In *TRAIL-/-* mice, cytokine production from macrophages and dendritic cells was increased and these mice had increased susceptibility to certain immune disorders such as autoimmune arthritis and diabetes (Diehl et al., 2004, Falschlehner et al., 2009). In the MS mouse model, EAE, blocking TRAIL resulted in increased CNS inflammation (Falschlehner et al., 2009). Therefore, in normal immune cells, the TRAIL system seems to exert a regulatory and suppressive role in the functioning of the immune response.

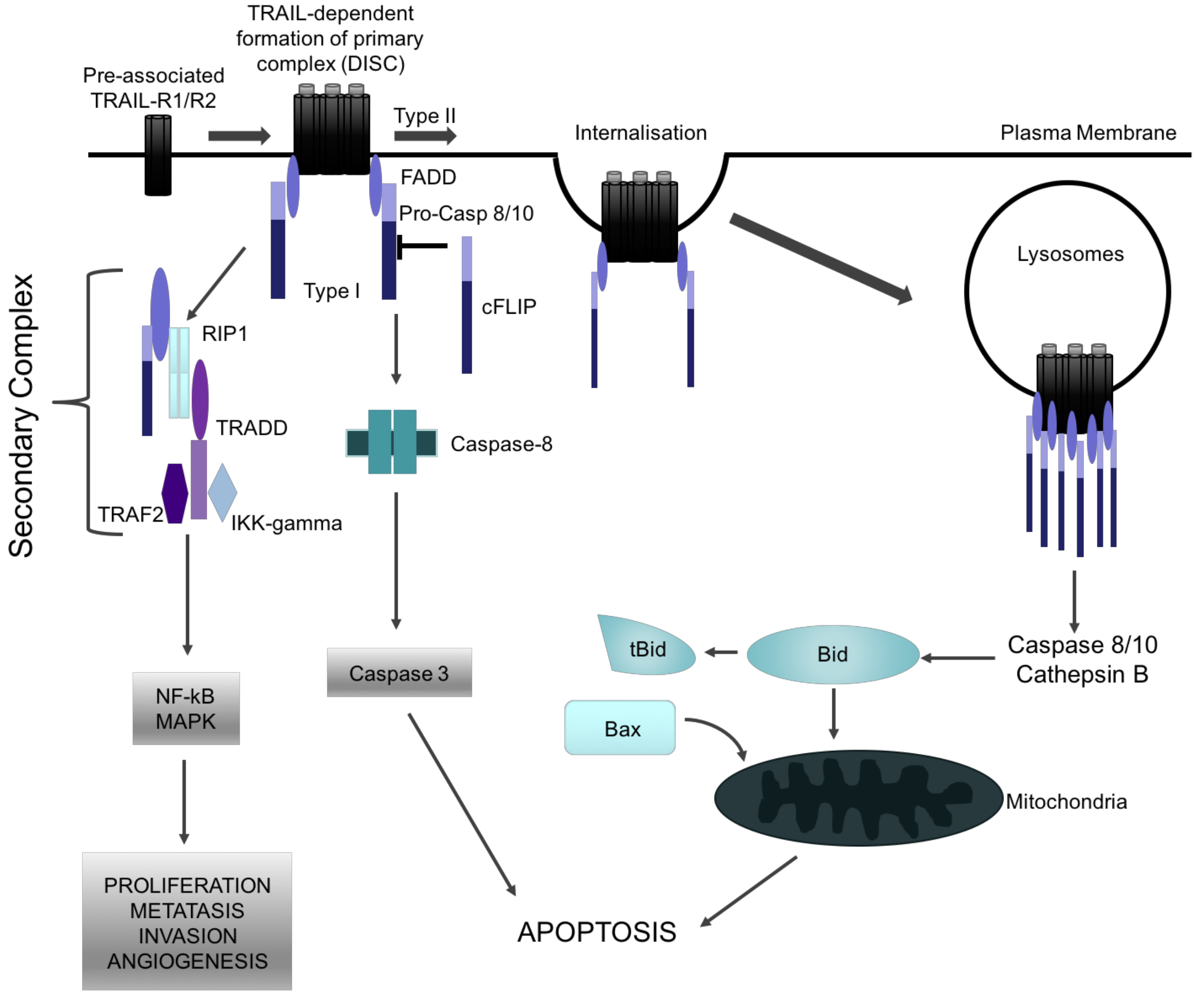


Figure 2.13: TRAILR1/2 signalling pathways
The different pathways stimulated by TRAIL binding to TRAILR1/2, which are often both expressed on the same cell. TRAILR1/2 do not require internalisation for stimulation of apoptosis in type 1 cells, but is essential in hepatocytes, as an example of type 2 cells. In addition to inducing cell death, TRAIL promotes activation of pro-survival mediators such as NF-kB and MAP kinases through a distinct pathway as shown above left. Activation of NF-kB cannot overcome TRAIL-mediated apoptosis in all cell types. Slight differences occur between TRAILR1 and TRAILR2 signalling at the TRAF2 level. TRAILR1 instead activates JNK/SAPK (stress-activated protein kinase) via a TRAF2-MKK4 (mitogen-activated protein/ERK kinase 4)-dependent pathway. Caspases are apoptosis activators. Bid, the truncated Bid (tBid) and Bax are all apoptotic proteins. cFLIP is a caspase 8-like inhibitory protein. RIP1 is the receptor-interacting protein 1, which is a death domain-containing serine/threonine kinase crucial in the balance between death and survival signalling, binds to all death receptors and can stimulate either a death cascade or a survival signal, in this case NF-kB activation by RIP1 activates survival pathways. TRADD is the TNF receptor-associated protein with death domain and acts as an adaptor protein. This figure and associated details described here were adapted from (Guicciardi and Gores, 2009).

I identified that decreased histone modification signal corresponds to decreased expression of both *TNFRSF10A* and *RP11-1149O23.3* genes, which in turn corresponds to an increased AMD risk (Table 2.4 and 2.5). Given the evidence that *TNFRSF10A* functions in monocytes to negatively regulate immune responses, a decrease in expression could result in an increased inflammatory state that over a prolonged period could add to increased risk of AMD.

There were two H3K27ac peaks that colocalised with this disease locus: 8:23048166:23092260 that directly overlapped the SNP and 8:23092704:23132254, located downstream. The peak directly overlapping the SNP was more significantly associated (p value = 3.941 x 10^{-45}, beta = -1.200, SE = 0.085, Table 2.5) than the downstream peak (p value = 1.647 x 10^{-09}, beta = -0.638, SE = 0.106). Therefore, based on location and strength of association, I postulated that the overlapping peak contained a putative regulatory element. Indeed, in previous molecular QTL studies, for example with DNase I hypersensitive (open chromatin) regions, it has been observed that most significant QTLs lie close to the DHS peak and proximal region (target window), specifically 56% of dsQTLs are located within the associated DHS and 67% are within a window of 1 kb around the feature (Degner et al., 2012). In addition, molecular strength of association is known to decay with increasing distance from the SNP (Waszak et al., 2015). I also excluded the colocalisation with the eQTL of the *CHMP7* gene from further analysis as the p values of association was also much less significant than the others (*CHMP7* beta = -0.473, SE = 0.099, value = 1.880 x 10^{-06}, H3K27ac beta = -0.638, SE = 0.1058, p value = 1.647 x 10^{-09}). Figure 2.14 and Table 2.4 and Table 2.5 summarise the association statistics of the four AMD-colocalised features that I investigated further; the two genes *TNFRSF10A* and *RP11-1149O23.3* as well as the H3K4me1 peak (8:22998146:23133613) and the single H3K27ac peak (8:23048166:23092260).

For all monocyte colocalised molecular features, the lead SNP was rs13255394, a common SNP (EAF = 0.575) located just downstream of the *TNFRSF10A* gene start site (Figure 2.17, Table 2.4). I next evaluated whether the lack of colocalisation with either neutrophil or T cell features represented a true cell-type specific disease effect, or whether neutrophil or T cells effects are missed due to a limitation in power.

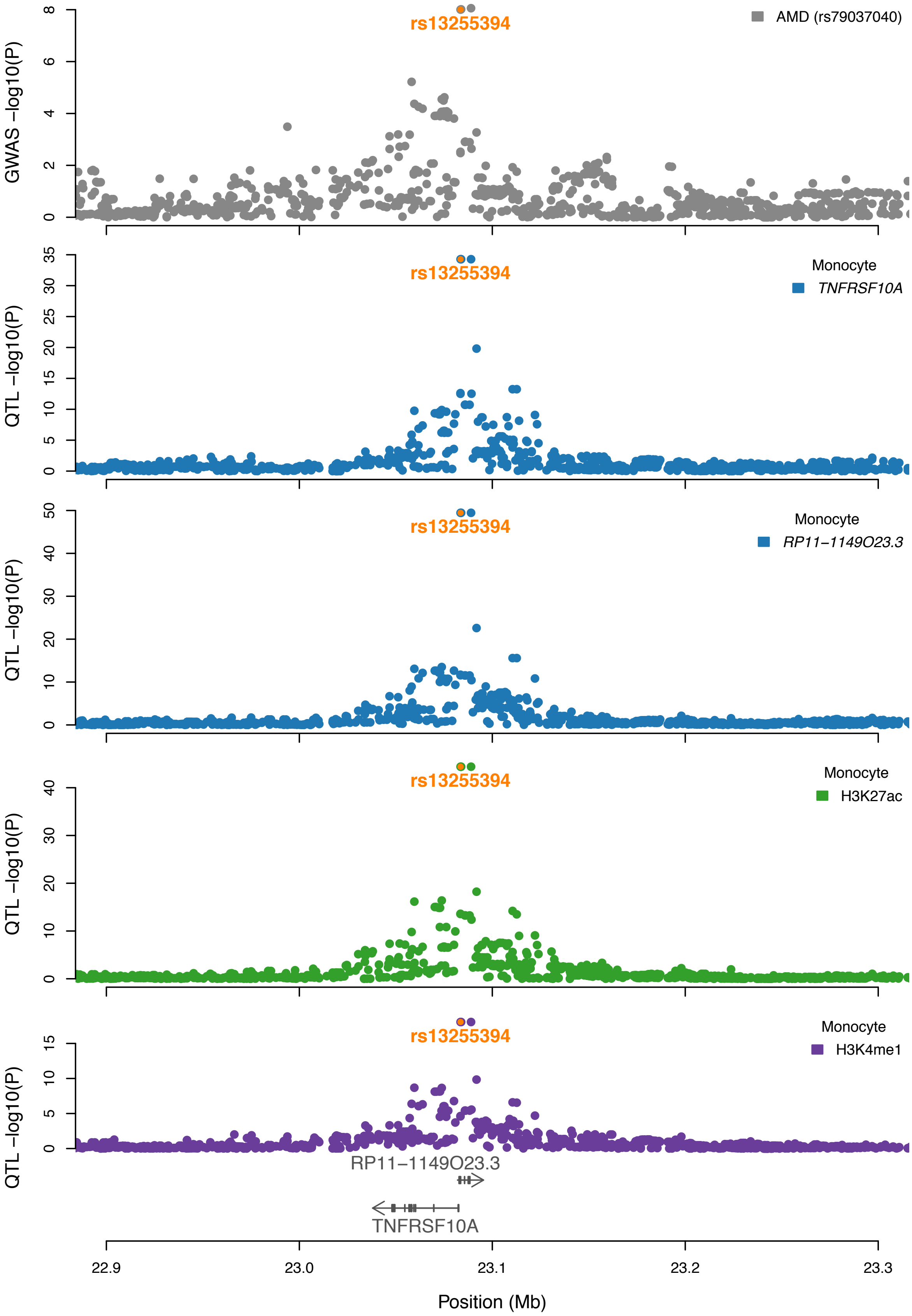


Figure 2.14 Regional association plots for the *TNFRSF10A* locus
Locus zoom plots show association strength in -log10 p value for variants that were shared between studies testing AMD (grey), monocyte gene expression (blue), monocyte H3K27ac, 8:23048166:23092260 (green) and monocyte H3K4me1, 8:22998146:23133613) (purple). The index disease SNP defined for the locus by Fritsche *et al* is rs79037040, but was not tested as part of the Blueprint study. The index molecular SNP, rs13255394 is labelled in orange and shown with respect to the genomic location, within exon 1 of *RP11-1149O23.*

Trait	SNP	R^2	EA/OA	EAF	AMD beta	AMD P	Cell type	Beta (SE)	P
TNFRSF10A	rs13255394 (M lead)	-	C/T	0.575	+	9.92 x 10^{-09}	Monocyte	-1.047 (0.085)	**5.249 x 10^{-35}**
							T cell	0.299 (0.110)	6.493 x 10^{-03}
							Neutrophil	-0.376 (0.096)	1.356 x 10^{-04}
	rs7820465 (T lead)	0.141	A/G	0.23	-	8.64 x 10^{-05}	Monocyte	0.585 (0.117)	5.892 x 10^{-07}
							T cell	-1.164 (0.108)	**3.279 x 10^{-27}**
							Neutrophil	0.536 (0.117)	4.999 x 10^{-06}
	rs4872078 (N lead)	0.005	T/G	0.47	-	1.950 x 10^{-02}	Monocyte	0.132 (0.098)	1.761 x 10^{-01}
							T cell	-0.813 (0.094)	4.833 x 10^{-18}
							Neutrophil	0.841 (0.087)	**6.145 x 10^{-22}**
RP11-1149O23.3	rs13255394 (M lead)	-	C/T	0.575	+	9.92 x 10^{-09}	Monocyte	-1.171 (0.079)	**3.477 x 10^{-50}**
							T cell	-0.295 (0.109)	6.734 x 10^{-03}
							Neutrophil	Not tested	Not tested

Table 2.4: Summary statistics of lead SNPs with gene expression of *TNFRSF10A* and *RP11-1149O23.3* expression, H3K27ac and H3K4me1 modification phenotypes in monocytes, neutrophils and T cells
Association statistics for the cell-specific lead SNPs for BLUEPRINT traits (Chen et al., 2016a). In bold are highlighted the lead associations in that cell type. AMD beta values and standard error estimates can be obtained by application to the IAMDGC consortium.

Trait	SNP	R^2	EA/OA	EAF	AMD beta	AMD P	Cell type	Beta (SE)	P
H3K27ac	rs13255394 (M lead)	-	C/T	0.575	+	9.92 x 10^{-09}	Monocyte	-1.200 (0.085)	**3.941 x 10^{-45}**
							T cell	-0.027 (0.120)	8.201 x 10^{-01}
							Neutrophil	-0.873 (0.096)	6.868 x 10^{-20}
	rs13255997 (T lead)	NT	G/A	0.510	+	8.540 x 10^{-03}	Monocyte	-0.261 (0.105)	1.301 x 10^{-02}
							T cell	0.508 (0.111)	4.669 x 10^{-06}
							Neutrophil	-0.088 (0.103)	3.944 x 10^{-01}
	rs4872090 (N lead)	0.402	A/T	0.763	+	1.240 x 10^{-03}	Monocyte	-0.877 (0.117)	5.407 x 10^{-14}
							T cell	-0.239 (0.130)	6.472 x 10^{-02}
							Neutrophil	-1.107 (0.105)	**4.599 x 10^{-26}**
H3K4me1	rs13255394 (M lead)	-	C/T	0.575	+	9.92 x 10^{-09}	Monocyte	-0.858 (0.097)	8.652 x 10^{-19}
							T cell	-0.054 (0.140)	6.984 x 10^{-01}
							Neutrophil	-0.603 (0.102)	2.944 x 10^{-09}
	rs8192332 (T lead)	NT	T/C	0.288	+	6.930 x 10^{-02}	Monocyte	-0.107 (0.125)	3.936 x 10^{-01}
							T cell	0.651 (0.172)	1.592 x 10^{-04}
							Neutrophil	-0.064 (0.125)	6.091 x 10^{-01}
	rs4872090 (N lead)	0.402	A/T	0.763	+	1.240 x 10^{-03}	Monocyte	-0.546 (0.119)	3.833 x 10^{-06}
							T cell	-0.188 (0.151)	2.142 x 10^{-01}
							Neutrophil	-1.083 (0.108)	**1.583 x 10^{-23}**

Table 2.5: Summary statistics of lead SNPs for H3K27ac with H3K4me1 modification phenotypes in monocytes, neutrophils and T cells

Figures 2.8 and 2.15 show there was a significant eQTL for *TNFRSF10A* in both T cells and neutrophils, but the lead SNPs associated with these signals are different and did not colocalise with AMD at this locus (summarised in Table 2.4). In T cells, the lead SNP was rs7820465 (EA = A, EAF = 0.23, beta = -1.164, SE = 0.108, p value = 3.279 x 10^{-27}, number of individuals = 169). In neutrophils, the lead *TNFRSF10A* eQTL was rs4872078 (EA = T, EAF = 0.47, beta = 0.841, SE = 0.087, p value = 6.145 x 10^{-22}, number of individuals = 196). Neither of these SNPs were significantly associated with AMD (rs7820465 p value = 8.64 x 10^{-05}, rs4872078 p value = 0.020).

I performed conditional analysis using GCTA and the eQTL summary statistics in each cell type (Materials and Methods). I tested for association of remaining SNPs after conditioning on the corresponding lead SNP for each cell type (Table 2.4 and Figure 2.15). In order to evaluate if any residual significant signals remained, I corrected the p value for the number of variants tested in the cis-region using the qvalue R package (Bass JDSwcfAJ, 2015) (Materials and Methods). There were no significant associations after conditioning on the respective lead SNP in monocytes or neutrophils (qvalue < 5%), which was evidence that within the power limitations of the cohort, there was one independent genetic signal in this region driven by the respective lead SNPs. In T cells, after conditioning on the lead SNP, rs7820465, there remained a marginally significant signal driven by the neutrophil lead SNP, rs4872078 (conditional beta = -0.394, conditional SE = 0.098, conditional p value = 5.946 x 10^{-05}, conditional q value = 0.046). I performed an iterative second stage of conditional analysis, using the output summary statistics generated by conditioning on rs7820465. In this second stage, I conditioned on rs4872078 and found no significant associations remained. None of these cell type lead SNPs were highly correlated; I calculated the LD r^2 estimates using the UKBB cohort of nearly 175,000 individuals and found an r^2 of less than 0.2 for each pairwise comparison (Table 2.4). Therefore, the association evidence suggests that expression of *TNFRSF10A* is regulated by varying independent signals across cell types and only the monocyte signal colocalised with AMD risk.

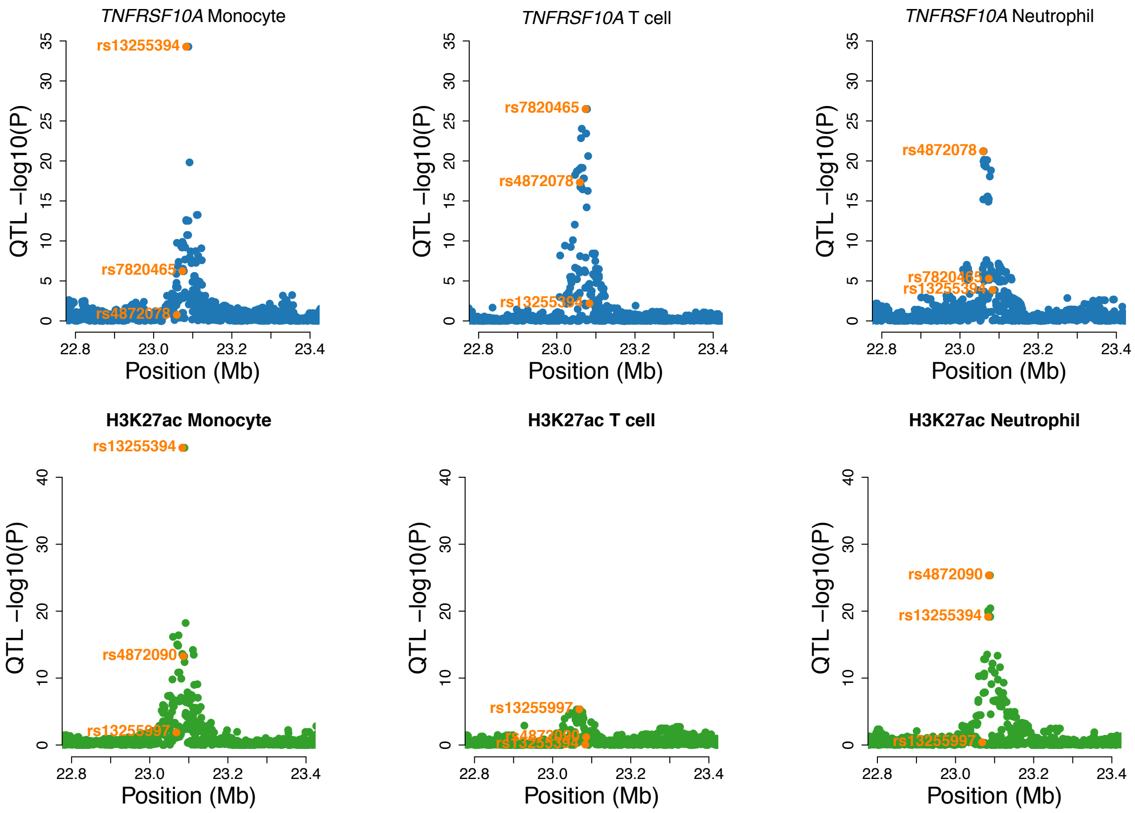


Figure 2.15: Different genetic signals across cell types at the *TNFRSF10A* locus
Regional association plots show the association signals for two colocalised features, *TNFRSF10A* gene expression and H3K27ac signal (8:23048166:23092260) in each cell type. The respective lead SNPs are highlighted in orange in each association plot. For the *TNFRSF10A* gene expression effect, there were three lead SNPs with evidence from LD and conditional analysis suggesting these represented three independent genetic signals explained by rs13255394 (monocytes), rs7820465 (T cells) and rs4872078 (neutrophils and secondary T cells). For the H3K27ac signal effect, evidence of variant LD and conditional analysis suggested that there were two genetic signals (rs13255394/rs4872090 monocytes and neutrophils) and a marginal signal in T cells (rs13255997).

I next investigated the histone modified region, which I postulated, was a regulatory control region for this locus. Monocyte H3K27ac signal across all individuals within the cohort showed greater correlation with monocyte *TNFRSF10A* signal than monocyte H3K4me1 (Figure 2.16, H3K27ac and *TNFRSF10A* Pearson's *r* = 0.567, p value = 1.254 x 10^{-16}, H3K4me1 and *TNFRSF10A* Pearson's *r* = 0.177, p value = 0.03). The higher correlation with H3K27ac than H3K4me1 could reflect the different roles of these histone marks. H3K4me1 is known to demarcate poised enhancers, that may not be active in the current cellular context (Creyghton et al., 2010). H3K27ac marks promoters but also active enhancers when modified in combination with H3K4me1, and therefore is required for active gene expression in specific cellular contexts (Creyghton et al., 2010, Heintzman et al., 2009). Using the Blueprint consortium cohort, we also observed a strong positive correlation between per-allele effect sizes of eQTLs and hQTLs (H3K27ac and H3K4me1) (Chen et al., 2016a). It is highly possible that at this locus, genetic disruption of H3K27ac is functionally more directly linked to gene expression.

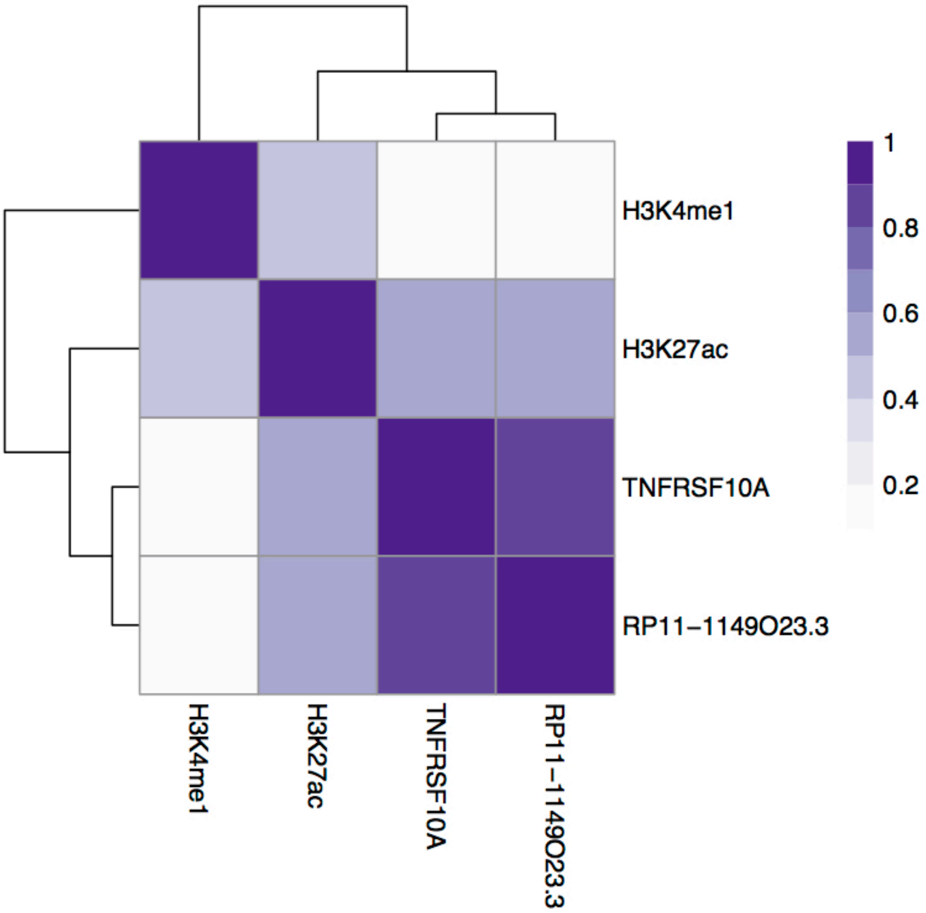


Figure 2.16: Correlation of molecular features at the *TNFRSF10A* locus
Heatmap shows unsupervised hierarchical clustering based on Pearson correlations between the molecular features. The Pearson correlation estimate is plotted between monocyte gene expression values or monocyte histone signal across the 158 individuals for which all of the molecular feature data was available.

In T cells, the H3K27ac signal was both weaker and explained by a different SNP (H3K27ac rs13255997, EA = G, EAF= 0.51, beta = 0.508, SE = 0.111, p value = 4.669 x 10^{-06}, Table 2.5, Figure 2.15) than in monocytes and there were no variants that reached the significant threshold for the H3K4me1 peak. There was evidence of regional histone signal in T cells as shown by the H3K27ac median $\log_2$RPM of 6.333 in T cells and 4.746 in monocytes and H3K4me1 median $\log_2$RPM of 7.246 in T cells and 7.146 in monocytes. This suggests that histone activity is present in both cell types, but the genetic regulatory mechanisms are divergent.

In neutrophils, both peaks were strongly associated with the lead neutrophil eQTL, rs4872090 (Table 2.5, Figure 2.15). The r^2 between the neutrophil histone lead SNP, rs4872090 and the monocyte histone lead SNP, rs13255394 was 0.402, and these SNPs are in close proximity approximately 2.29 kb apart (Figure 2.17). Conditional analysis using GCTA on hQTL summary statistics demonstrated that no significant signal remained after conditioning on the respective lead SNPs in either monocytes or neutrophils (qvalue < 5%). Interestingly, this was evidence that the histone signals can be explained by the same genetic signal in monocytes and neutrophils, despite the observation that the *TNFRSF10A* expression was controlled by independent SNPs across all cell types. Only in monocytes was there evidence for coordinated gene expression, histone activity and disease risk, which were all regulated by same SNP, rs13255394. Given that there was evidence of histone activity in other cell types, similarly, to the CAD *PSRC1* locus explained above, this suggests that there are additional regulatory features that coordinate in generating cell type-specific regulatory mechanisms.

I evaluated whether the additional colocalised gene, *RP11-1149O23.3*, could represent such an additional regulatory mechanism. Figure 2.17 shows that the lead SNP, rs13255394, is located within exon 1 of the non-coding RNA gene, *RP11-1149O23.3*. The p value for association with this RNA was more significant (p value=3.477 x 10^{-50}) and the effect size larger (beta= -1.171, SE = 0.079) than with *TNFRSF10A* expression (Table 2.4), suggesting this is an important functional effect at this locus.

The lead SNP for this locus identified in the original GWAS discovery (Fritsche et al., 2016) was rs79037040 which is also located in exon 1 of *RP11-1149O23.3,* but closer to the TSS of both *RP11-1149O23.3* and *TNFRSF10A* (Figure 2.17). Figure 2.17 shows the raw histone signal (in $\log_2$RPM) stratified by genotype (bottom panel), which enabled identification of the location of disrupted histone activity with a greater resolution (50bp across the genome). It was clear that the position of rs79037040 was directly in the centre of the monocyte and neutrophil H3K27ac peaks. This SNP is in high LD with rs13255394 (r^2 = 0.837, 1000G) suggesting that rs13255394 and rs79037040 likely explain the same genetic signal. rs79037040 was filtered from the first phase of BLUEPRINT cohort analysis due to stringent quality control thresholds. However, efforts within our group have been undertaken to reanalyse this cohort with improved imputation procedures generating a denser SNP panel, which included rs79037040. This analysis, referred to as phase 2, was performed by Kousik Kundu in the Soranzo team. I confirmed that in the phase 2 association testing, rs79037040 was now the lead SNP for all of the colocalised molecular features described thus far (Table 2.6). In addition, the colocalisation method used in this chapter, gwas-pw, calculates the posterior probability (PP) of all tested variants being causal for the two colocalised traits. Colocalisation with the new phase 2 summary statistics, calculated a PP for rs79037040 of 1 for *TNFRSF10A, RP11-1149O23.3,* H3K27ac and H3K4me1 traits. Therefore, all evidence supports that rs79037040 is the single causal variant for AMD risk and all monocyte molecular features.

Feature	SNP (EA/OA)	Beta	SE	P
Mono *TNFRSF10A*	rs79037040 (T/G)	-1.200	0.079	1.540 x 10^{-51}
Mono *RP11-114O23.3*	rs79037040 (T/G)	-1.290	0.073	8.497 x 10^{-70}
Mono H3K27ac (8:23048166:23092260)	rs79037040 (T/G)	-1.291	0.082	7.830 x 10^{-56}
Mono H3K4me1 (8:22998146:23133613)	rs79037040 (T/G)	-0.9856	0.095	2.057 x 10^{-25}
AMD	rs79037040 (T/G)	0.109	0.016	4.5 x 10^{-11}

Table 2.6: Association summary statistics of the lead SNP from Blueprint phase 2 genetic analysis
The AMD disease lead, rs79037040 (later merged into the ID rs13278062) was tested only as part of the latest Blueprint genetic analyses (phase 2). Beta, standard error (SE) and p value for association are listed here. The effect allele, T is associated with a decrease in gene expression and histone signal is also associated with an increase in AMD-risk. The effect allele (T) frequency is 0.556.

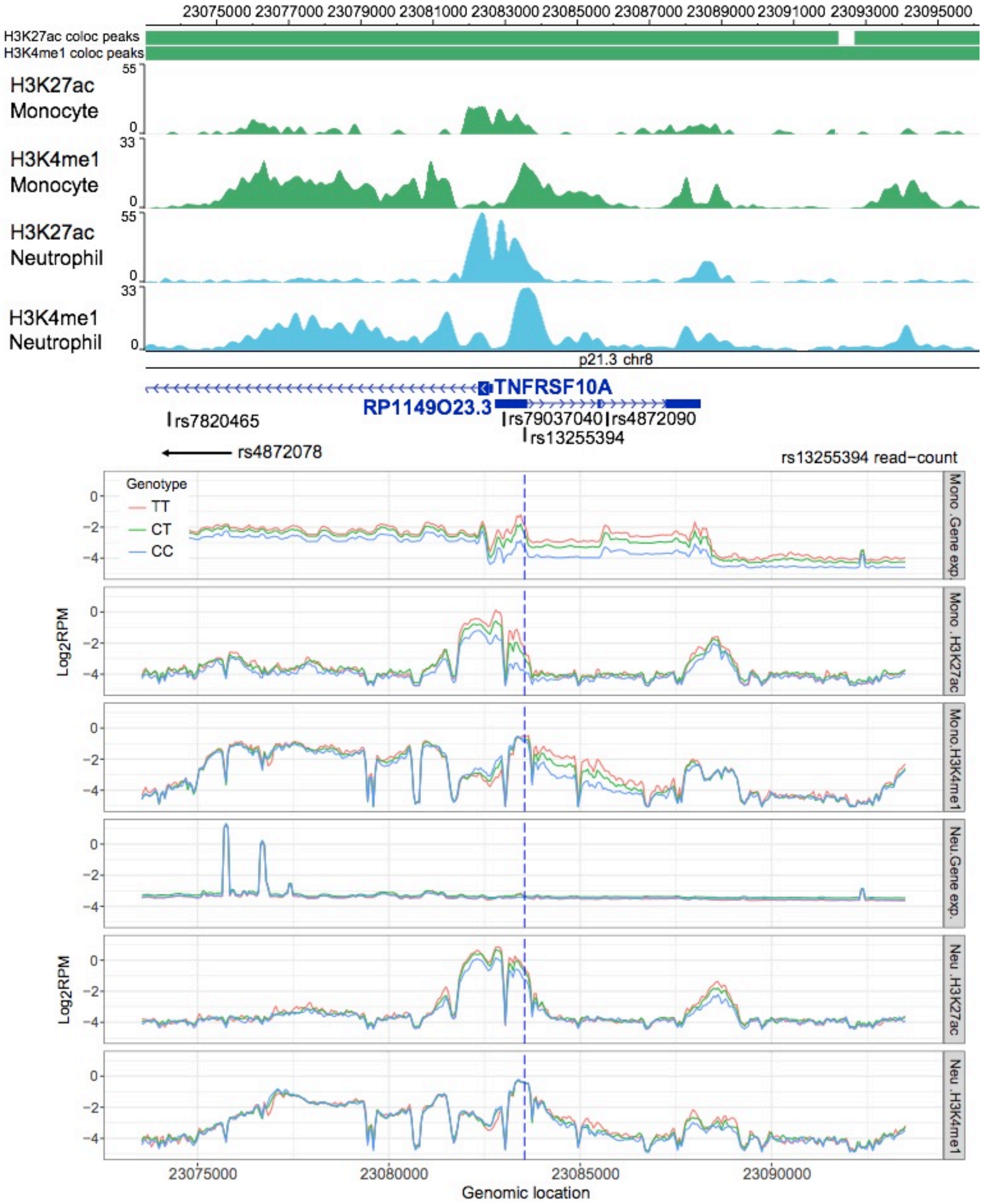


Figure 2.17: Epigenome characteristics of the *TNFRSF10A* locus
Genome browser figure of the genomic region around the *TNFRSF10A* gene and the proximal regulatory RNA, *RP11-1149O23.3* located upstream on the opposite strand (top panel). Peaks shown were generated from a representative individual from the BLUEPRINT cohort predicted to carry the allele associated with the highest signal (Chen et al., 2016a). Locations of the lead QTL variants and their genomic locations are shown. The lead monocyte QTLs, rs13255394 (phase 1) and rs79037040 (phase 2) are located within exon 1 of the RP11-1149O23.3-002/003 transcript. Solid blocks represent the colocalised peak, the second H3K27ac and H3K4me1 peaks extend further downstream than is shown. The bottom panel shows the raw histone signal, in monocytes and neutrophils, in Log$_2$RPM (reads per million) calculated in windows of 50 bp across the genome. The gene expression signal is plotted from the raw signal expressed in Log$_2$RPM. Each genotype of the lead monocyte SNP, rs13255394 is plotted (red for homozygous reference, TT). The genome browser plot (top panel) was generated with custom tracks using the Washu Epigenome browser (Zhou et al., 2011). The raw signal plots were generated using Blueprint data by Kousik Kundu (Chen et al., 2016a).

Less is known about the function of the non-coding RNA, *RP11-114O23.3* compared to the *TNFRSF10A* encoded receptor. However, previously a regulatory relationship between *RP11-114O23.3* and *TNFRSF10A*, two genes located on opposite strands, has been suggested (Zheng et al., 2016). Using microarrays, it was demonstrated that the expression of *RP11-114O23.3* (also known as *LOC389641*) is increased in pancreatic ductal adenocarcinoma (PDAC) tissues and correlated with patient prognosis and high expression levels reduced overall survival (Zheng et al., 2016). *RP11-114O23.3* expression increased proliferation and decreased apoptosis of cancer cell lines (Zheng et al., 2016). The authors identified a 378bp region that contained a highly conserved sequence directly upstream of 5' end of *RP11-114O23.3* that is the reverse complement of the *TNFRSF10A* promoter. In the same study upregulation of *TNFRSF10A* expression was observed in PDAC patient tissue compared to non-tumorous tissues. Crucially, siRNA mediated knock-down of *RP11-114O23.3* in SW1990 cells, significantly decreased *TNFRSF10A* expression but knock-down of *TNFRSF10A* had no effect on the expression of *RP11-114O23.3*. This suggests that *RP11-114O23.3* regulates expression of *TNFRSF10A* through complementary sequence-mediated binding.

These analyses were performed either in cancer cell lines or in PDAC patient tissue. It could be possible that this relationship is not observed in healthy primary immune cells. In addition, to the best of my knowledge, possible genetic regulation of the relationship between these genes has not been previously explored. I therefore sought to investigate the relationship between *RP11-114O23.3* and *TNFRSF10A* in monocytes, neutrophils and T cells.

First, similar to the high significant correlation observed between the *TNFRSF10A* and *RP11-114O23.3* expression in PDAC tissues (r^2=0.606, p<0.001, N = 106 patients), I also identified high correlation between monocyte expression values of the two genes across all BLUEPRINT individuals tested for these monocyte features (r= 0.804, p value = 3.962 x 10^{-45}, Figure 2.18, N = 194). However, in T cells, the correlation between *RP11-1149O23.3* and *TNFRSF10A* expression was lower and less significant than that observed with monocytes (r= 0.172, p value = 0.025, N = 169) (Figure 2.18). Expression of *RP11-1149O23.3* was not significantly associated with the lead T cell *TNFRSF10A* eQTL, when evaluating local FDR (Figure 2.18). In neutrophils, *RP11-1149O23.3* was not tested due to low expression, where the median log2FPKM was 2.672 (N = 196), compared to 5.852 in monocytes and 6.644 in T cells. Combined, this is evidence that the relationship between expression of the two genes and the genetic regulation of each gene was monocyte specific.

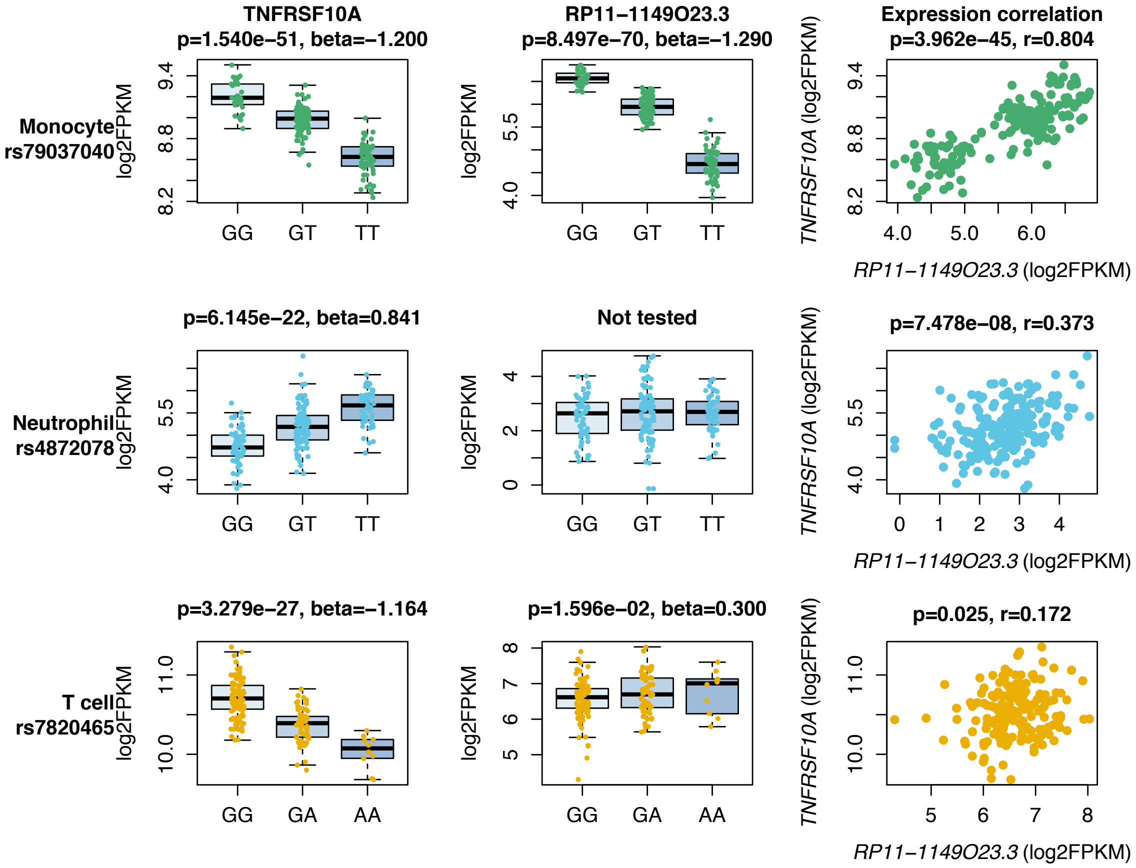


Figure 2.18: Genetic control of *TNFRSF10A* and *RP11-1149O23.3* is not shared across monocytes, T cells and neutrophils
Gene expression of the two genes, in log2FPKM, is stratified by genotype of the respective lead *TNFRSF10A* SNPs in each cell type. The beta and p values of association are shown for each case, as calculated by the BLUEPRINT study (Chen et al., 2016a). The correlation between the gene expression raw signals is shown for each cell type and is strongest in monocytes. Here the phase 2 lead SNP was used, rs79037040, but similar correlation was observed for the phase 1 lead, rs13255394 (data not shown).

I next sought to confirm that the observed correlation in healthy cells was consistent with a hierarchical regulatory mechanism as demonstrated by previous siRNA experiments where knockdown of *RP11-114O23.3* in cancer cells affected *TNFRSF10A* expression and not vice versa (Zheng et al., 2016). To test this, I implemented a linear regression approach to model the gene expression of each gene using expression of the alternative gene as a covariate in the model. In this conditional approach, variation due to one gene was removed by correcting for expression of all individuals in the BLUEPRINT cohort (Materials and Methods). Following this, I tested for association of any variation in expression remaining in the residuals with rs79037040 genotype. I performed the analysis using only the phase 2 lead SNP, given the evidence that rs79037040 was the single causal SNP for all monocyte molecular feature associations.

I first tested the univariate associations with rs79037040 for both genes using the lm() R function and confirmed significant associations observed in the QTL testing method from the BLUEPRINT study (Table 2.7) (Chen et al., 2016a). I applied the two-stage models (above and Materials and Methods) to test if the significant association with the casual SNP remains after removing any variation in *TNFRSF10A* expression that is due to variation in *RP11-1149O23.3* expression levels. I identified the strength of the rs79037040 association with *TNFRSF10A* gene expression decreased and was no longer significant after removing the RNA effect. The p value increased from 2.334 x 10^{-40} to 0.076 (non-significant) and the effect size (beta) decreased by 10-fold (Table 2.7). This demonstrated that *RP11-1149O23.3* contributes a high degree to variation in *TNFRSF10A* expression. I also applied the reverse model: *RP11-1149023.3 ~ TNFRSF10A* expression followed by testing the residuals for association with rs79037040. The results demonstrated that the p value increased from 8.349 x 10^{-66} to 1.124 x 10^{-08}, remaining significant and beta decreased only by 3.5 times. This reduction suggested that there may also be a smaller effect of *TNFRSF10A* expression on that of *RP11-1149023.3,* but the dominant effect is regulation of *RP11-1149023.3* on *TNFRSF10A* expression.

I extended the causality analysis of the different monocyte features to assess whether this approach may indicate that H3K27ac regulation also exerted an effect on gene expression. I applied the following model: *TNFRSF10A* gene expression ~ *RP11-1149O23.3* expression + H3K27ac signal and then tested the residuals for association with rs79037040 (Table 2.6). I used R^2 estimates, adjusted for the number of covariates in the model, to evaluate whether the model fit improves with the addition of the histone modification effect. The adjusted R^2 slightly increased when adding H3K27ac signal as a covariate, from 0.643 to 0.669, and a significant difference was confirmed using the ANOVA significance test for nested models (p value = 2.883 x 10^{-04}, N = 158). Additionally, the inclusion of the colocalised H3K4me1 peak

(8:22998146:23133613) as a covariate in the model, decreased the R^2 estimate to 0.667 and was not significant when compared to the *TNFRSF10A* ~ *RP11-1149O23.3* + H3K27ac model (p value = 0.812, N = 158).

All together, these results suggested that both the RNA and H3K27ac influence gene expression of *TNFRSF10A*, but H3K4me1 has limited effect and that *RP11-1149O23.3* contributes a high degree to the variation in *TNFRSF10A* expression.

Feature (SNP)	Model (covariate)	Beta (SE)	P
TNFRSF10A (rs79037040)	Univariate	-0.303 (0.017)	2.334 x 10^{-40}
	Conditional (*RP11-114O23.3* expression)	-0.031 (0.017)	0.076
	Conditional (*RP11-114O23.3* expression + H3K27ac signal)	-0.025 (0.017)	0.140
RP11-114O23.3 (rs79037040)	Univariate	-1.001 (0.034)	8.349 x 10^{-66}
	Conditional (*TNFRSF10A* expression)	-0.283 (0.047)	1.124 x 10^{-08}
	Conditional (*TNFRSF10A* expression + H3K27ac signal)	-0.264 (0.047)	6.235 x 10^{-08}

Table 2.7: Conditional causality analysis in the *TNFRSF10A* locus
Association results (beta, SE, and p value) from a simple linear regression model using non-transformed phenotype values (as this demonstrated good model fit and normally distributed residuals). Similar trends were also observed for inverse normalised phenotype values. The univariate approach tests for association of the respective gene expression with the genotype of rs79037040 (lead monocyte SNP from Blueprint release 2 and lead AMD SNP). Conditional analysis then tests for association of the gene with genotype whilst conditioning on the expression of the alternative gene. The increase in p value and decrease in significance is greatest when testing for association between the SNP and *TNFRSF10A* expression whilst conditioning on *RP11-114O23.3* expression, which suggests the RNA may be causal for variation in expression of *TNFRSF10A*. The further approach conditions on the gene expression and H3K27ac signal and then tests the resulting residuals for association with the SNP genotypes. For all models, data from 158 individuals was used.

2.4 Discussion

In this chapter, I applied enrichment and colocalisation methods to evaluate the utility of immune molecular phenotypes, specifically of monocytes, neutrophils and naïve CD4 T cells, to dissect mechanisms of disease risk for a variety of disorders. I demonstrated that a high number, 46%, of tested disease loci colocalised with at least one molecular feature in at least one cell type and highlighted many important gene targets, some of which already have therapeutic utility (Table 2.3). Following this, I performed an in-depth analysis of two example disease loci, CAD *SORT1* and AMD *TNFRSF10A* and demonstrated how the integration of multiple data sources is required to generate plausible mechanistic hypotheses.

I identified significant enrichment of GWAS variants in regions of the genome known to be associated with immune molecular traits, particularly of monocyte and neutrophil eQTLs in all of the range of five diseases I studied. The relative absence of strong cell-type specific patterns for most diseases was consistent with previous observations using similar analytical approaches for the same molecular data but with a wider range of classical autoimmune diseases (Chen et al., 2016a). Here, for disease loci associated with coeliac disease, Crohn's disease, inflammatory bowel disease, ulcerative colitis, multiple sclerosis, rheumatoid arthritis or Type 1 diabetes, colocalisation of 54% with eQTLs, 55% with splicing QTLs, 62% with H3K27ac and 54% with H3K4me1 QTLs was observed (Chen et al., 2016a). However, this analysis included the MHC region for all diseases, which may affect the estimates given the complex genetic architecture of this region. Comparison of colocalisation estimates with other studies and/or diseases is challenging, given the different methods available and approaches to evaluate colocalised loci. Using iPSC-differentiated unstimulated and stimulated macrophages, the highest number of colocalised eQTLs or chromatin accessibility (ca)QTLs were observed with inflammatory bowel disease variants (11 and five loci respectively) (Alasoo et al., 2017). The recent G. TEx eQTL analysis identified a similar percentage to my study; 52% of trait-associated variants colocalised with an eQTL in one or more tissues (G. TEx Consortium, 2017).

Many of the colocalised genes identified had well-established or suggested roles in immune function. This is in agreement with increasing insight into the pathogenic involvement of inflammation in wider range of disorders and with the early promise of therapeutically targeting these pathways (Section 2.1). Based on these observations, I concluded that functional insight can be gleaned from using peripheral immune cell types in these diseases, which were traditionally not considered prototypic immune-mediated diseases. I also provided support for the importance of lipid-pathway genes such as the CAD loci colocalised with *LIPA* and the AMD locus colocalised with *CETP*. Through the identification of well-known examples such as these I confirmed the validity of my analytical approach as well as providing further mechanistic evidence for disease loci.

This study is not the first to integrate GWAS loci with molecular features, certainly for autoimmune diseases, this is fairly widespread and as discussed in Section 2.1 has generated important insight (Farh et al., 2015, Chun et al., 2017). Using peripheral whole blood as a tissue source enabled QTL identification in large cohorts of healthy individuals, such as a study from 2013 of 5,311 individuals with replication in 2775 individuals (Westra et al., 2013). These data have been used to dissect gene expression consequences of trait-associated loci with for example inflammatory bowel disease and lung function (FEV$_1$) (Wain et al., 2015, Huang et al., 2017b). In comparison, an advantage of cohorts such as BLUEPRINT is in facilitating identification of the specific cell-type source of the genetic effect. In addition, BLUEPRINT enables the study of neutrophil effects, which are historically understudied despite that the important role in inflammation, immune cross-talk and certain disease aetiology. I highlighted one particular SLE locus, *UBE2L3,* where the strongest colocalised eQTL was from neutrophils compared to previous observations of correlation of this locus with *UBE2L3* expression across monocytes, CD4 T cells, B cells and NK cells at this locus in the original GWAS study (Bentham et al., 2015). Definitive confirmation of disease-relevant mechanisms requires functional validation, but if clear demarcation of cell types is possible for at least a proportion of loci, this is an important preliminary step in designing these experiments and selecting experimental cellular models. Blood is an experimentally tractable and easily accessible tissue source and function is conserved across organisms facilitating the use of animal models (Orkin and Zon, 2008, Vasquez et al., 2016). Providing colocalisations within blood for diseases where human biosamples for other relevant tissues are challenging to obtain, such as brain or ocular tissue, is a clear advantage of these findings.

However, using whole blood and purified cell cohorts to fully resolve functional genetic mechanisms can be thought of as complementary. For example, the smaller sample sizes of cohorts such as BLUEPRINT (N = 200) limits the identification of *trans* QTLs. These are variants that affect molecular features located more than 1Mb away, or even on different chromosomes. Highly powered studies are required to detect these effects due to the increase in the multiple testing burden when expanding the testing window beyond variants in *cis*. *Trans* eQTLs were identified and replicated for 103 independent loci using the large whole-blood cohort described above (Westra et al., 2013). Complex-trait associated variants showed a high number of *trans* eQTL effects. Interesting insights into CAD variants were also gleaned from both *cis*- and *trans*-eQTLs identified in another large microarray gene- and exon-based QTL dataset of whole blood from 5257 individuals (Joehanes et al., 2017).

19,000 independent lead cis-eQTLs were detected compared to just over 6000 *trans*-eQTLs. By overlapping blood eQTLs with CAD SNPs or those in LD $r^2 \geq 0.8$, Joehanes *et al.* (2017) identified genes for 21 of the 58 GWAS loci (Joehanes et al., 2017). Those in agreement with the effects identified in this chapter were eQTLs for *LIPA*, *NT5C2*, *VAMP8 and GGCX, REST* and also the *PSRC1* target at the *SORT1* locus. In my study, only the *GGCX* and *NT5C2* genes showed evidence of either expression or splicing effects across all three of the cell types studied here, enabling some assessment of cell-type specificity of the other loci among the three subsets studied in BLUEPRINT. Joehanes *et al.* (2017) identified *trans* effects at some CAD loci. For example, the evidence suggested that the *VAMP5-VAMP8-GGCX* locus (rs7568458) affected the expression of 5 genes in *trans*; *CASP5, DPEP3*, *CRISPLD2*, *SLC26A8* and *PKN2* (Joehanes et al., 2017). The expression of *CASP5* has been previously shown to be associated with blood pressure, suggesting trait-relevant effects may occur in *trans* (Joehanes et al., 2017). In total, Joehanes *et al.* (2017) identified more CAD SNPs overlapping with gene QTLs than my analysis (21 compared to 8 e/sQTLs here), which could be due to the increased study power but could also represent overlaps occurring by chance without formal assessment such as those applied in colocalisation methods (Joehanes et al., 2017). Future studies with larger cohorts and defined cell populations will combine the advantages of these two study types and provide the opportunity for identifying further regulatory pathways that also influence disease risk. Of course, blood cell types will not be the disease-relevant tissue for all loci, which may be the explanation for why I did not detect colocalisations for all loci. Part of the future work of this thesis will involve fully integrating disease loci with eQTLs from the (G. TEx Consortium, 2015) to evaluate effects across a wider range of tissue types.

A further advantage of the BLUEPRINT cohort is that the it enabled concomitant assessment of multiple regulatory features rather than being limited to gene expression effects. Consideration of the genetic effects on chromatin state enables resolution of regulatory mechanisms at disease loci. It has been demonstrated that transcriptional and local epigenetic states are highly coordinated and that genetically controlled changes in gene expression may occur through disruption of chromatin states (Grubert et al., 2015, Waszak et al., 2015). My analysis focused on the histone modifications, H3K27ac and H3K4me1 and also RNA splicing effects, which were either independent of or in addition eQTL effects depending on the locus. In describing detailed mechanisms, I showed two loci that differed in the strongest chromatin effect being either H3K27ac or H3K4me1. These observations support the use of multiple sources and types of molecular data in fully investigating regulatory function.

The *SORT1* CAD locus colocalised with *PSRC1* eQTL but not a *SORT1* QTL effect in these haematopoietic cell types. That *PSRC1* is regulated (and not *SORT1*) in blood has been observed in independent cohorts (Zeller et al., 2010, Joehanes et al., 2017). However, I demonstrated this effect was also present in neutrophils and absent in CD4$^+$ T cells. I further provided a potential molecular mechanism underlying the differences between hepatocytes and myeloid cells. Principally, this difference seemed to be explained by the binding of important haematopoietic TFs, C/EBPβ and PU.1, to a C/EBP motif disrupted by rs12740374. In liver, C/EBPα was found to be bound at this motif and *SORT1* was the strongest genetically regulated expression effect (Musunuru et al., 2010). Interestingly, using ENCODE data, I found that the liver pioneer factors, FOXA2 and FOXA2, were bound at the rs12740374 locus in HepG2 cells (data not shown) (Odom et al., 2006, Iwafuchi-Doi et al., 2016, Zaret et al., 2008). I postulated that the binding of different pioneer TFs in different cells may promote regulation of the expression of alternative genes. The importance of pioneer factors in directing cell-type specific binding and expression has previously been observed (Heinz et al., 2010, Heinz et al., 2013, Mullen et al., 2011). In this way, SNP-mediated disruption of a TF motif can result in opposing molecular consequences in different cells while the sequence effect remains the same.

The challenge remains of interpreting the role of *PSRC1* in blood cells and whether this gene is causally related to CAD. Gain-of-function studies in mice and genetic findings in human cohorts have both supported an association of the 1p13 minor haplotype (rs12740374 T allele) with increased hepatic *SORT1* expression and decreased LDL-C and VLDL (Musunuru et al., 2010). However, in a mechanism thought to be independent of lipoprotein metabolism, SORT1 mediates LDL uptake in macrophages stimulating their differentiation to foam cells and therefore *promoting* atherosclerosis (Mortensen et al., 2014, Westerterp and Tall, 2015). It is conceivable, therefore, that the differences in response to LDL between liver and myeloid cells manifest in differences in CAD risk. This is supported by the observation of Musunuru *et al.* (2010) that although *PSRC1* expression in human liver was significantly associated with rs12740374, overexpression of *Psrc1* in mouse liver was not associated with any significant changes in total cholesterol (Musunuru et al., 2010). In my analysis, increased *PSRC1* expression was associated with a protective CAD effect. Given little is known regarding the function of *PSRC1*, further work is required to evaluate this effect in haematopoietic cells and whether this effect could also be causal to CAD or whether the hepatocyte *SORT1* effect is the only causal contribution of this locus to disease risk. For example, using CRISPR-Cas9 to knock-out this gene in iPSCs and differentiation to macrophages followed by stimulation with oxidised LDL to promote foam cell formation could highlight whether *PSRC1* is involved through regulating this process. These experimental approaches are well established (Hale et al., 2015, Reschen et al., 2015). Implementation of

Mendelian randomization approaches could also help to ascertain whether the blood *PSRC1* effect of this locus represents a disease causal mechanism or purely pleiotropy.

It would also be interesting to further evaluate the mechanisms through which cell-type specific genetic regulation of gene expression is achieved and confirm that the binding of PU.1 and C/EBPβ is linked to *PSRC1* gene expression. This is important for the design of novel treatments for understanding the effect of a drug on multiple tissues. Specific siRNA knock-down of PU.1, C/EBPβ and C/EBPα in haematopoietic and hepatic cell lines, coupled with an assessment of the effect on *PSRC1* and *SORT1* gene expression, would demonstrate whether the binding of both factors is required for downstream gene expression or whether one TF acts as the putative regulatory factor. These experiments could be performed in the cell line models discussed above; differentiated HL60, the monocyte-like cell line, U937 and the hepatic cell line, HepG2.

I also investigated a monocyte-specific effect at the AMD *TNFRSF10A* locus and identified a putative regulatory element for the *TNFRSF10A* and *RP11-1149O23.3* genes that colocalised with the AMD locus. Although colocalisation was identified between both H3K4me1 and H3K27ac peaks, the effect was more significant with H3K27ac (Figure 2.14, Figure 2.17). In addition, in a linear model of *TNFRSF10A* gene expression, inclusion of H3K27ac improved the model fit, but not inclusion of H3K4me1. I also demonstrated the importance of other regulatory mechanisms at this locus, by identifying the colocalisation of an eQTL for *RP11-1149O23.3* and further that expression of this non-coding RNA explained a significant degree of variation in *TNFRSF10A* expression. This relationship is a “local trans” regulatory mechanism, where a genetic variant affects the expression of one gene, which in turn regulates a proximally located, but distinct gene. Clearly genomic regulation in this locus involved both an RNA and histone effect, but the exact linear relationship between these effects is difficult to ascertain without functional validation. Open chromatin is required for active gene expression of most genes in order to enable access of the RNA polymerase II machinery to the transcription start site (Venters and Pugh, 2009). It is conceivable that *RP11-1149O23.3* could either require established open chromatin to be expressed or to bind this region or could recruit further chromatin remodellers. Using CRISPR to knock-out the RNA and independently the histone region could allow an assessment of the downstream effect on *TNFRSF10A* expression may aid dissection of these relationships. It is thought that disruption of chromatin is proceeded by the alteration of transcription factor binding (McVicker et al., 2013, Kilpinen et al., 2013). Experiments such as ChIP-seq could also be used to identify other bound co-factors. The lead AMD and molecular feature SNP, rs79037040, located within exon 1 of *RP11-1149O23.3* has also been predicted to disrupt the motifs of TFs LXR and NERF1a (Kheradpour and Kellis, 2014). These TFs are both highly expressed in monocytes (>6 and >11 log2FPKM respectively from RNA-seq gene expression from (Chen et al., 2016a). Disruption of TFs bound, and histone activity at the RNA promoter could lead to changes in *RP11-1149O23.3* expression, which further propagate to corresponding changes in the expression of *TNFRSF10A*.

Given that TNFRSF10A is a surface expressed receptor, I postulated that downstream alterations in receptor surface expression or function would ultimately impact AMD risk. Expression of the *TNFRSF10A* gene in the AMD-relevant tissue, peripheral retinal pigment epithelium/choroid/sclera (PRCS) (FPKM PRCS = 1.91), is low compared to AMD-drug target gene, VEGFA (FPKM PRCS = 56.38) (Li et al., 2014), which potentially further provides evidence that the disease-relevant effect of this locus could be exerted in monocytes. TNFRSF10A (TRAIL-R1), TNFRSF10B (TRAIL-R2) and the decoy receptor TNFRSF10D (TRAIL-R4) are all highly expressed on the surface of primary monocytes from healthy individuals, with the highest surface expression observed for TNFRSF10B (Deligezer and Dalay, 2007, Liguori et al., 2016). The disease SNP is not associated with monocyte expression of the other TNF receptors (rs79037040 TRAILR2 p value = 0.976, TRAILR3 p value = 0.144, TRAILR4 = 0.504). Interestingly, despite the significant *TNFRSF10A* eQTLs in neutrophils and T cells, this receptor has been shown to be lowly expressed on freshly isolated primary neutrophils and T cells (Kamohara et al., 2004, Liguori et al., 2016). Instead the decoy protein TNFRSF10C (TRAIL-R3) receptor is highly expressed on the surface of neutrophils and to a lesser extent on the surface of lymphocytes (Kamohara et al., 2004, Liguori et al., 2016). TNFRSF10C expression was associated with a strong eQTL in neutrophils (rs7009522, EA = A, beta = 1.098, p value = 2.519 x 10^{-23}) but was not tested in T cells. Therefore, post-transcriptional processes could play an important role in reducing surface expression of *TNFRSF10A* in certain cell types, highlighting the importance of integrating multiple sources of functional information to interpret the mechanistic consequences at disease loci. Similar experiments to those described in Chapter 4 of this thesis, where surface receptor expression was measured using flow cytometry in a recall-by-genotype design could establish possible differences in surface expression associated with genotype.

Lower expression of *TNFRSF10A* in monocytes corresponded to an increase in AMD risk. It has been postulated that recruitment of blood cells such as macrophages to the damaged retinal tissue in AMD could contribute to a pathogenic pro-inflammatory environment (Nussenblatt and Ferris, 2007). TNFRSF10A is known to be immunosuppressive. The evidence presented here suggests that lower *TNFRSF10A* gene expression could result in a reduced inability to downregulate pro-inflammatory responses in monocytes or in macrophages if differentiated to monocytes. It would be interesting to study these effects in monocyte-derived macrophages to observe if cells with reduced *TNFRSF10A* expression show a more inflammatory profile.

Not all loci that colocalised with histone features also shared a gene effect, either expression or RNA splicing, an observation that was also made in the Chen *et al*. (2016a) study and in other independent cohorts. For example, a study identifying iPSC-differentiated macrophage gene expression and chromatin accessibility QTLs (open chromatin using ATAC-seq) also found that of the 23 caQTLs that colocalised with a GWAS variant, only two of these also colocalised with an eQTL (Alasoo et al., 2017). These regulatory QTLs might impact gene expression in different cells, contexts or affect post-transcriptional processes (Alasoo et al., 2017, Fairfax et al., 2014, Pai et al., 2015).

Colocalisation approaches provide a statistical assessment of regions of the genome that are associated with two traits, but there are still some limitations to this approach. First, the power of the method to detect true colocalisation when the lead variants of each trait are in high LD ($r^2 \geq 0.8$) is limited and the method assumes one causal variant at each locus. Definitive demonstration of causality between traits, specifically whether the shared molecular effect is causal to disease risk, is not possible (discussed in detail in Chapter 5).

Further, the colocalisation method does not distinguish between multiple independent genetic signals, which have been observed at molecular loci and in some cases colocalised with disease variants over or in addition to primary signals (Dobbyn et al., 2017, Ke, 2012). I used the pre-defined association signals from the Chen *et al.* (2016) where multiple independent signals were not investigated. Visual inspection of some colocalised loci in this analysis suggested the colocalisation may not be between the primary molecular association. Supplementary Figure 2.3 shows the colocalisation of the SLE *FCGR2A* locus with a splicing QTL for *FCGR3A* (CD16) in monocytes and *FCGR2A* (CD32) in neutrophils. Both genes are expressed in neutrophils and monocytes (≥ 8 $\log_2$FPKM), and both receptors are expressed on the surface of each cell type (Stenberg et al., 2013, Cooper et al., 2012, Ziegler-Heitbrock, 2007, Devaraj et al., 2013). The disease lead SNP and the *FCGR3A* splicing lead QTL are highly correlated (rs6671847, rs4657041 1000G r^2=0.89) but the lead splicing SNP for *FCGR2A* is not highly correlated with the disease SNP (rs12129787 1000G $r^2 < 0.2$). To assess this, future work will implement conditional analysis to identify independent genetic signals.

In conclusion, I demonstrated that applying colocalisation methods to GWAS and molecular QTL data can provide detailed mechanistic hypotheses at disease risk loci, which are invaluable for facilitating further experimental investigation.