Characterising lymphocyte-mediated mechanisms of genetic risk in rheumatoid arthritis

Alexander David Clark

Thesis submitted to Newcastle University for the degree of Doctor of Philosophy

Translational and Clinical Research Institute

November 2019

Abstract

Rheumatoid arthritis (RA) is an immune-mediated inflammatory disease of the synovial joints with a global prevalence of 1%, causing pain, work instability and disability. The condition's genetic aetiology is complex, and genome-wide association studies have highlighted over 100 susceptibility loci. The vast majority of implicated variants are noncoding, posing a major challenge in defining genetic mechanisms of cell-mediated immune dysregulation. The precise contribution to this process of epigenetic factors at a cellular level, such as the addition of methyl groups to DNA, also remains to be deciphered. By conducting comprehensive molecular profiling of circulating immune cells from early arthritis patients, my project aimed to elucidate mechanisms of genetic risk and prioritise causal disease genes in RA. To address this, genome-wide DNA methylation, transcriptome and genotype data were available from peripheral blood CD4$^+$ T cells and B cells of treatment-naïve early arthritis patients. Firstly, a comparison of DNA methylation between patients with early RA and those with other arthropathies was undertaken. Subsequently, the capacity of RA-associated variants to influence DNA methylation by mapping methylation quantitative trait loci (meQTLs) was confirmed. Here, it was observed that disease variants preferentially modified DNA methylation at sites mapping to lymphocyte enhancers and regions flanking transcription start sites, as well as positions bound by the NFκB transcription factor. By integrating transcriptomic data and employing a statistical approach to infer causality, loci were identified at which genetically conferred modifications in DNA methylation regulate transcription of genes including *FCRL3*, *ANKRD55*, *IL6ST*, and *JAZF1* in CD4$^+$ T cells. Finally, in vitro assays were used to validate meQTLs at loci of interest, and to confirm regulatory mechanisms. This work highlights genes and pathways of potential relevance to lymphocyte-mediated pathology in early RA and, potentially, other immune mediated diseases. It has implications for the functional interpretation of genome/epigenome-wide association studies.

Acknowledgements

I would firstly like to thank my supervisors Dr Arthur Pratt, Dr Louise Reynard, and Professor John Isaacs for allowing me the opportunity to undertake my PhD under their guidance. Their support over the past three years has been unwavering, always willing to provide guidance where needed, as well as encouraging independence in my research endeavours and learning. In addition, I am incredibly grateful to the JGW Patterson Foundation for funding this research and making this project possible.

I must also thank colleagues within the Immunotherapy and Skeletal Research Groups, in particular to Julie Diboll, Dr Amy Anderson, Dr Nishanthi Thalayasingam, and Dr Phil Brown, all of whom contributed to sample processing. I am indebted to both Andrew Skelton, who introduced me to bioinformatics analysis during my first few months as a complete novice, and to Najib Naamane, who has always been on hand to answer questions and discuss any issues that I may encounter.

In addition, I am also grateful to collaborators at the University of Manchester for their technical assistance and intellectual input. In particular, I would like to express my thanks to Dr Nisha Nair and Dr Jonathan Massey for performing all of the array-based assays, as well as to Professors Anne Barton and Steve Eyre for help in guiding the project.

I would also like to extend my gratitude towards the research nurses and patient participants at the Freeman Hospital for providing clinical samples. Without their enthusiasm and willingness to support translational research, research in the group would not be possible.

Finally, it is important for me to recognise all of my friends and family, both those in Newcastle and beyond, who have been incredibly supportive, on top of providing a welcome distraction from work when necessary. In particular, I would like to acknowledge the encouragement and advice of my parents, whose support has been integral in my decision to pursue a scientific career.

Table of Contents

Chapter 1 – Introduction

Chapter 2 – Materials & Methods

Chapter 4 – Methylation Quantitative Trait Locus Analysis

Chapter 5 – DNA methylation as a mediator of transcriptional regulation

Chapter 6 – Conclusions and Future Directions

List of Figures

List of Tables

Abbreviations

5’UTR/3’UTR – 5’ untranslated region/3’ untranslated region

5-Aza – 5-Aza-2’-deoxycytidine

5caC – 5-carboxycytosine

5fmC – 5-formylcytosine

5hmC – 5-hydroxymethylcytosine

5mC – 5-methylcytosine

α-KG – α-ketoglutarate

ACPA – Anti-citrullinated protein antibodies

ACR – American College of Rheumatology

AEI – Allelic expression imbalance

AID – autoimmune disease

ANOVA – Analysis of variance

AS – Anyklosing spondylitis

BMIQ – Beta mixture quantile dilation normalisation

CCTF – CCTC-binding factor

cDNA/cRNA – Complementary ribonucleic acid/deoxyribonucleic acid

CGI – CpG Island

ChIP-seq - Chromatin immunoprecipitation with sequencing

CIT – Causal inference test

CpG – Cytosine phosphate Guanosine

CRISPR - Clustered regularly interspaced short palindromic repeats

CTLA-4 – Cytotoxic T-lymphocyte-associated protein 4

DALY – Disease-adjusted life years

DAS28 – Disease activity score at 28 joints

DEPC – Diethyl pyrocarbonate

DHS – DNase I hypersensitivity site

DMARD – Disease-modifying anti-rheumatic drug

DMEM – Dulbecco's modified eagle medium

DMR – Differentially-methylated region

DMP – Differentially-methylated position

DMSO – Dimethyl sulfoxide

DNA – Deoxyribonucleic acid

DNAm – DNA methylation

dNTP – Deoxynucleotide triphosphate

DPBS – Dulbecco's phosphate buffered saline

DTT – Diethylthreitol

DVP – Differentially-variable position

EA – Enteropathic arthritis

EDTA - Ethylenediaminetetraacetic acid

ENCODE – Encyclopaedia of DNA Elements Project

eQTL – Expression quantitative trait locus

eQTM – Expression quantitative trait methylation

eSNP – regulatory eQTL SNP

EULAR – European League against Rheumatism

EWAS – Epigenome-wide association study

FCS – Foetal calf serum

FDR – False discovery rate

FLS – Fibroblast-like synoviocytes

Funnorm – Functional normalisation

GFP – Green fluorescent protein

(G)M-CSF – (Granulocyte)-macrophage colony stimulation factor

GO – Gene Ontology

gRNA – Guide ribonucleic acid

GTEX – Genotype-tissue expression project

GWAS – Genome-wide association study

HDAC – Histone deacetylase

HLA – Human leukocyte antigen

iEVORA – Epigenetic variable outliers for risk prediction algorithm

IFN – Interferon

Ig – Immunoglobulin

IL – Interleukin

IMD – Immune-mediated disease

IQR – Inter-quartile range

JAK – Janus kinase

LCL – Lymphoblastic cell line

LD – Linkage disequilibrium

lncRNA – Long non-coding ribonucleic acid

mAb – Monoclonal antibody

MACS – Magnetic-activated cell sorting

MAF – Minor allele frequency

MBD – Methyl-CpG binding domain

MCP – Metacarpophalangeal

meQTL – Methylation quantitative trait locus

MHC – Major histocompatibility molecule

miRNA – Micro ribonucleic acid

MMP – Matrix metalloproteinase

MR – Mendelian randomization

MS – Multiple sclerosis

NF-κB - Nuclear factor-κB

NK cells – Natural killer cells

Noob – Normal-exponential out-of-band probes normalisation

OA – Osteoarthritis

PAD – Peptidylarginine deiminase

PBMCs – Peripheral blood mononuclear cells

PCA – Principal component analysis

(q)PCR – (Quantitative) Polymerase chain reaction

PIP – Proximal interphalangeal

PsA – Psoriatic arthritis

PVCA – Principal variance component analysis

RA – Rheumatoid arthritis

RANKL - Receptor activator of NF-κB ligand

ReA – Reactive arthritis

RF – Rheumatoid factor

RIN – RNA integrity number

RLM – Relative log methylation

RNA – Ribonucleic acid

SAH – S-adenosylhomocysteine

SAM - S-adenosylmethionine

SBE – Single base extension

scFv – Single chain variable fragment

SE – Shared epitope

SLE – Systemic lupus erythematosus

SNP – Single nucleotide polymorphism

SpA – Spondyloarthropathy

SS - Sjögren syndrome

STAT3/STAT6 - Signal transducer and activator of transcription 3/6

SVA – Surrogate variable analysis

SWAN – Subset-quantile within-array normalisation

T1D – Type 1 diabetes

TAE – Tris-acetate-EDTA

TCR – T cell receptor

TET - Ten-eleven translocation methylcytosine dioxygenase

TF – Transcription factor

TFBS – Transcription factor binding site

T_{FH}/T_{PH} – Follicular helper T cells/Peripheral helper T cells

TGF-β – Transforming growth factor β

$T_{H}1$/$T_{H}2$/$T_{H}17$ – T-helper cell 1/2/17

TNF – Tumour necrosis factor

Treg – Regulatory T cell

TSDR – Regulatory T cell (Treg) specific de-methylated region

TSS – Transcription start site

WGS/WGBS – Whole-genome sequencing/whole-genome bisulphite sequencing

Word Count – 74,312

Chapter 1 - Introduction

1.1 Background

Rheumatoid arthritis (RA) is a chronic autoimmune condition, during which the loss of immune tolerance manifests principally as inflammation at the synovial lining of small joints. It is the most common autoimmune arthropathy, with an estimated prevalence of 0.5-1% in the majority of studied populations, including in the United Kingdom[1-3]. The condition also displays a notable sex bias, with an increased prevalence in females[1, 2].

If inflammation remains unresolved, symptoms progress to structural damage of the cartilage and underlying bone, with accompanying loss of physical function and disability for patients[4]. This can have profound effects at the level of the individual (impaired quality of life, work instability), and in society as a whole (economic burden of medical treatment and lost working days)[5]. The United Kingdom National Audit Office estimates that the financial burden on the National Health Service for direct healthcare provision for RA patients is around £557 million per year (Figure 1.1), with an additional £1.8 billion in economic costs associated with work-related disability[6].

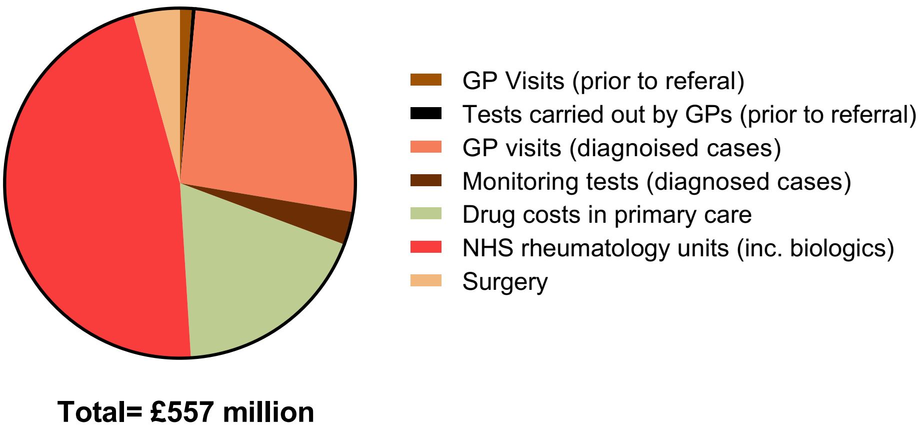


Figure 1.1: The distribution of costs incurred by the United Kingdom National Health Service in the diagnosis and treatment of rheumatoid arthritis. Estimated costs were obtained from the National Audit Office[6].

A recent emphasis on the importance of early diagnosis in RA and the prompt initiation and escalation of treatment according to response, together with the development of biologic drugs that target inflammatory pathways, have greatly enhanced outcomes in RA[7, 8]. Despite this, there remains an unmet clinical need, as approximately half of all patients will respond to current biologic therapies, with an even lower proportion achieving disease-free remission[7, 8]. Furthermore, biomarkers that predict response to different therapies are lacking, meaning that truly tailored treatment strategies during the critical 'window of opportunity' early on in disease remain aspirational[7].

Whilst RA is predominantly considered a condition of the synovial joints, systemic inflammation can result in extra-articular manifestations including, but not limited to, depression, chronic obstructive pulmonary disease, malignancies, and cardiovascular disease[9-11]. Indeed, the latter may be responsible in part for the increased rate of mortality that is observed in RA patients relative to the wider population[12, 13]. This appears to be reduced upon treatment with biologics such as anti-tumour necrosis factor (TNF), albeit only in women, establishing a link between systemic inflammation and RA-associated mortality[14]. The Global Burden of Disease study in 2010 revealed that RA is responsible for 4.8 million disease-adjusted life years (DALYs – years of healthy life lost to disease) worldwide, with 696,000 in Western Europe[15].

1.2 Pathophysiology of RA

The current paradigm in RA is that the autoimmune response occurs following environmental triggers in individuals who are genetically susceptible. In this scenario, preclinical extra-articular immune induction likely occurs prior to inflammation of the joint and chronic synovitis. The most frequently affected sites are the small joints of the hand, such as the proximal interphalangeal (PIP) and metacarpophalangeal (MCP), as well as the wrist, knees, elbow, and shoulders (Figure 1.2). Swelling and tenderness at 28 joints is included in the disease activity score at 28 joints (DAS28), that gives measure of disease activity[16].

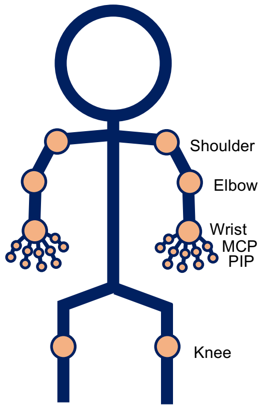


Figure 1.2: Commonly affected rheumatoid arthritis joints. Rheumatoid arthritis typically effects joints symmetrically, and calculation of the DAS28 disease activity score incorporates information on tenderness and swelling in 28 commonly affected joints. MCP = metacarpophalangeal; PIP = proximal interphalangeal.

1.2.1 Autoantibodies

Whilst RA likely encompasses a number of related clinical endotypes, disease is broadly divided into two subcategories based on the presence or absence of autoantibodies. Autoantibodies against the Fc fragment of immunoglobulin G (IgG), named rheumatoid factor (RF), were the first to be associated with RA. These antibodies can be detected in the blood many years before the onset of RA symptoms, and individuals with elevated titres of RF in the serum (>100 IU/ml) have a 26-fold increased risk of developing RA [17, 18].

Subsequent to the discovery of RF, autoantibodies were identified that were reactive against proteins in which post-translational conversion of arginine to citrulline (citrullination) has occurred. This modification is catalysed by a family of enzymes termed peptidylarginine deiminases (PADs), and these anti-citrullinated protein antibodies (ACPAs) are present in approximately 70-80% of RA patients[19, 20]. Importantly, ACPAs are much more specific to RA than is RF, with the latter occurring in a higher proportion of non-RA conditions, and clinical application of a test to detect ACPAs reports a specificity of 96%[19]. The pathogenesis of ACPA seropositive RA is better characterised than that of seronegative disease. ACPA positive patients also have a poorer disease prognosis with increased probability of developing erosion in the joints and radiographic progression[21].

The precise role of autoantibodies in disease induction remains ambiguous. That most autoantibody seropositive RA patients have detectable levels of both of these antibodies, and the formation of ACPAs likely precedes that of RF, suggests a role for the latter in potentiating arthritogenic humoral responses directed at citrullinated self-peptides[17]. One pathogenic mechanism through which autoantibodies may contribute to the triggering of RA is through the formation of immune complexes that promote complement activation and the release of pro-inflammatory mediators by immune cells[22].

1.2.2 Pre-RA: Breakdown in self-tolerance

Observations that both RF and ACPAs are present in the circulation many years prior to the onset of clinical RA indicates an extra-articular break in self-tolerance that precedes synovitis[17, 23]. This stage is often referred to as 'pre-RA', and it is during this period that genetic and environmental interactions manifest as dysregulation of the immune response. Whilst often asymptomatic, patients may also be in the pre-RA phase if they display arthralgia in the absence of clinical arthritis, or an unclassified arthritis that does not yet fulfil RA diagnostic criteria[24].

Seroconversion of patients following the onset of RA is very rare, suggesting a causal role for autoantibodies in driving disease induction[25]. Interestingly, antibody titres gradually increase over years, with epitope spreading and increased avidity occurring in conjunction with the initiation of clinical disease[26-29]. As will be discussed in later sections, the mucosal surfaces, in particular the lung, are the likely sites at which this immune response to self-antigens is initiated. Activation of the immune system, including antigen-presenting cells, at these sites potentially leads to the adaptive immune responses directed against citrullinated self-epitopes[30]. Recent data suggests T cells that are specific to post-translationally modified antigens, such as those that occur in RA, escape negative selection in the thymus that would normally remove autoreactive cells from the circulation[31].

During pre-RA, the cytokine levels in circulation further demonstrate that pre-clinical pathological mechanisms are active in at-risk patients, with elevated serum cytokines and chemokines detected in the circulation relative to healthy individuals. These include interleukin- (IL-)1-α, IL-1β, IL-6, IL-10, TNF-α, and Granulocyte-macrophage colony-stimulating factor (GM-CSF) amongst others, with detectable levels appearing to increase nearer to the time of diagnosis[32]. Prior to the onset of RA, elevated cytokine levels are predominantly characteristic of a CD4$^+$ T cell response[33]. These increases in levels of inflammatory cytokines have also been shown to be more exaggerated in patients with autoantibody seropositive relative to seronegative disease[34]. Nonetheless, cytokines are also prominent in seronegative disease and the upregulation of signal transducer and activator of transcription 3 (STAT3) target genes in CD4$^+$ T cells as a response to IL-6 stimulation has been described in these patients[35].

Consistent with the observation that cytokine levels are elevated in early RA, ACPAs from RA patients can form immune complexes with fibrinogen, which stimulate the secretion of TNF-α by macrophages upon Fc-receptor binding[36], and can also stimulate TNF-α production downstream of nuclear factor-κB (NF-κB) activation in monocytes[37]. Finally, ACPAs are able to activate the complement system, again highlighting potential active roles of these autoantibodies in pathogenesis[38].

Certain cellular compartments are expanded in patients at risk of developing RA. A study of early arthritis patients, healthy controls and autoantibody positive individuals yet to develop arthritis found that B cells were expanded in the inguinal (groin) lymph node of arthritis patients relative to controls[39]. The authors also describe a non-significant trend of a similar B cell expansion in the cohort with circulating autoantibodies, but absent clinical arthritis[39]. A separate investigation has suggested that a population of CD4$^+$ T cells displaying a pro-inflammatory phenotype is more prominent in the lymph nodes of RA patients relative to healthy controls[40].

The data discussed in this section demonstrate that triggering of autoimmunity and an extra-articular inflammatory response occur prior to RA diagnosis. Autoantibodies and cytokines show a clear association with risk of developing RA, though established, causal pathological mechanisms during the early stages of pre-RA remain elusive. As ACPAs are not sufficient to trigger synovitis in isolation, it has been hypothesized that a 'second hit', perhaps as a result of physical trauma or viral infection is necessary in these at-risk patients to develop localized joint inflammation[30]. Although ACPAs are present many years before the onset of clinical RA, recent work has shown that IL-23 activation of IL-17 producing T-helper cell 17 (T_H17) cells in turn results in a shift in the glycosylation profile of IgG antibodies, conferring pro-inflammatory properties and leading to the induction of arthritic symptoms[41]. DNA variants mapping to cytokines or their receptors are implicated in RA susceptibility, such as is the case for the IL-6 receptor (*IL6R*) gene, illustrating the contribution of cytokine signalling pathways to RA susceptibility[42]. The genetic basis of RA provides further clues to the molecular pathways and cell types involved, and the implications for disease mechanisms will be discussed in section 1.3.

1.2.3 Common cellular autoimmune pathways?

RA is one of many common autoimmune conditions, all of which are characterised by an aberrant, self-directed immune response. Whilst the term 'autoimmune disease' (AID) refers to a range of conditions that display markedly distinct clinical manifestations, they are all characterised by this shared immunological aetiology. For example, self-reactive B and T cells in the periphery that produce autoantibodies and pro-inflammatory cytokines are central in the immunopathogenesis of multiple sclerosis (MS) and type 1 diabetes (T1D), conditions for which the target tissues are central nervous system and pancreatic β cells respectively[43, 44].

The potential overlap in autoimmune pathways is particular evident at the genetic level, where disease-associated single nucleotide polymorphisms (SNPs) appear to be involved in modulation of transcriptional activity in immune cells, particularly lymphocytes[45]. The implications of these RA genetic risk factors at the cellular level, and autoimmunity more generally, will be discussed in greater detail in later sections.

It is this autoimmune component that distinguishes RA from the other common form of arthritis in the developed world, osteoarthritis (OA). Whilst immune activation and systemic inflammation have been described to some degree in OA patients, the principal triggers include developmental defects, mechanical stress, and a metabolic shift in chondrocytes as they begin to produce collagen- and aggrecan-degrading enzymes[46]. Because OA presents a somewhat similar phenotype to RA, with the destruction of cartilage and bone, but with distinct aetiologies, samples from OA patients are often used as a control condition in clinical studies of RA.

1.2.4 Early & established RA – synovial inflammation and bone erosion

During the pre-clinical phase of autoantibody-positive RA that precedes the appearance of symptoms in the joint, there does not appear to be any synovitis or marked cellular infiltration[47]. However, as disease progresses, profound changes in the structure and cellular composition of the synovium and joint space shortly precede the appearance of symptoms within the joint (Figure 1.3)[48]. Hyperplasia of cells at the synovial lining occurs as the disease develops. This accompanies extensive angiogenesis and influx of immune cells into the joint, thus creating a pro-inflammatory environment which leads to activation of the fibroblast-like synoviocytes (FLS) which secrete cartilage-degrading matrix metalloproteinases (MMPs)[49]. Membrane-type I matrix metalloproteinase (MMP-14) appears to be the predominant factor secreted by the RA synoviocytes that drives degradation of type I and type II collagen in the cartilage tissue[50]. Cytokines including macrophage colony-stimulating factor (M-CSF) and receptor activator of NF-κB ligand (RANKL), produced by inflammatory cells, promote the formation of bone-degrading osteoclasts[49]. Together with the hyperplastic synovium, these osteoclasts contribute to formation the 'pannus' that invades the cartilage and bone tissue (Figure 1.3). Pro-inflammatory cytokines such as TNF, IL-1 and IL-6 also directly contribute to bone erosion through osteoclast activation, and therapeutic intervention targeting these pathways can slow down structural damage of the joint, as well as suppressing inflammation[51].

Though systemic hallmarks of autoimmunity occur prior to joint-related symptoms, infiltration of T cells into the synovium may represent one of the earliest cell migration events in the transition to synovitis[47]. Recent work has highlighted distinct pathotypes in early RA based on the heterogeneous compositions of infiltrating cells[52]. One such pathotype is characterized by a predominance of lymphoid cells (T cells and B cells), whereas other groups identified were dominated by either myeloid cells, or expansion of resident stromal cells such as fibroblasts (with low levels of immune cells)[52]. In a subgroup of patients, the formation of ectopic lymphoid structures in the joint is accompanied by increased levels of infiltrating T- and B cells, and is also associated with increased inflammatory markers, both in the joint and the periphery[53]. As well as increased lymphocyte infiltrates to the joint tissue, cells undergo metabolic changes that promote a pro-inflammatory phenotype. CD4$^+$ T cells in RA have been shown to display increased levels and activity of the glucose-6-phosphate dehydrogenase[54]. This manifests as hyper proliferation, and skews differentiation of naïve cells towards pro-inflammatory T_H1 and T_H17 phenotypes, characterised by the expression of interferon (IFN)-γ and interleukin-17 (IL-17) cytokines respectively[54].

In addition to their critical roles in antigen presentation and autoantibody secretion, B cells are a major source of cytokines in the synovium, including RANKL that promotes differentiation of bone-eroding osteoclasts from precursor cells[55]. Early work also showed B cells were essential for activation of synovial CD4$^+$ T cells and formation of ectopic synovial germinal centers (sites of B cell proliferation and maturation) in RA[56].

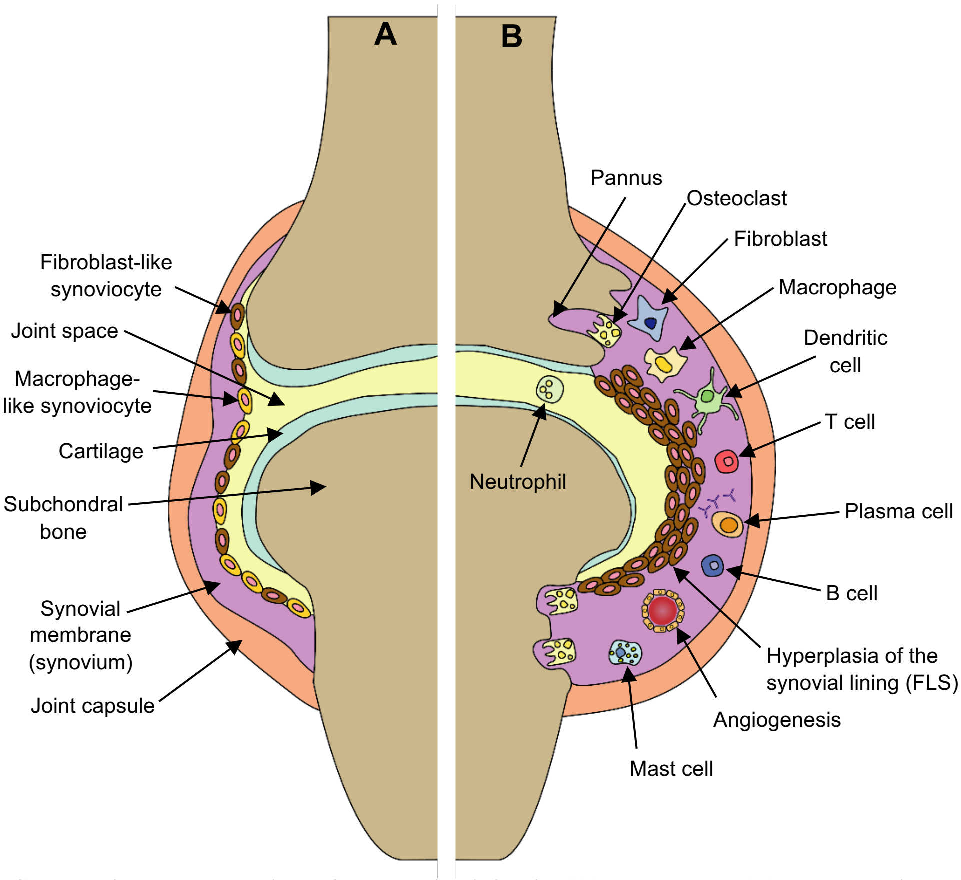


Figure 1.3: Schematic representation of a synovial joint in (A) health and (B) rheumatoid arthritis. In a healthy joint, the bone is covered in a smooth layer of cartilage that acts as a shock absorber and lubricates the joint during movement. The lining of the synovium is composed of a thin layer of synoviocytes, usually no more then 2-3 cells thick. During disease, extensive angiogenesis accompanies influx of a wide range of inflammatory cells into the joint tissue. Hyperplasia occurs at the synovial lining, with fibroblast-like synoviocytes (FLS) becoming hyper-proliferative and promoting inflammation and release of cartilage-degrading enzymes. These FLS, together with the activated bone-degrading osteoclasts contribute to the formation of the pannus which erodes the cartilage and bone. Figure is adapted from Strand, Kimberley & Isaacs (2007)[48].

The stromal compartment of the joint is integral in not only establishing synovial inflammation, but also initiating the processes that result in cartilage and bone damage. Upon stimulation with pro-inflammatory cytokines, the immunosuppressive capacity of synovial fibroblasts from non-inflamed or resolving joints is ablated in patients with RA, the cells losing their ability to suppress endothelial lymphocyte recruitment[57]. It appears that in the early stages of transition to clinical RA, changes that occur in the signaling properties of IL-6 and transforming growth factor (TGF)-β secreted by synovial fibroblasts promotes their pro-inflammatory function[57]. Using an integrated approach, Zhang and colleagues were able to define cells driving inflammation in the synovial tissue at the single cell level, identifying a number of cell states that were overabundant in RA relative to OA controls[58]. These included THY1^{+} HLA-DR^{HI} sublining fibroblasts that were expanded in a subset of RA patients characterized by leukocyte-rich synovia, relative to leukocyte-poor RA and OA patients[58]. Subsequent work has identified two distinct fibroblast subsets that drive distinct pathological processes in RA[59]. These include the THY1^{+} cells of the synovial sublining that were shown to drive inflammation, and the THY1^{-} cells that populate the synovial lining and mediate bone and cartilage erosion[59]. Autoimmune-associated B cells expressing *ITGAX* and *TBX21*, as well as a cluster of CD4^{+} T cells (peripheral helper (T$_{PH}$) and follicular helper (T$_{FH}$)) that expressed *PDCD1* were also amongst the populations highly enriched in the synovium of leukocyte-rich RA patients[58]. Furthermore, a population of CD4^{+} T$_{PH}$ cells that are expanded in the RA synovium have been described, and these cells appear to provide B cell help to promote plasma cell differentiation and antibody production[60]. Taken together, these results indicate a complex cellular milieu within the synovium that drives inflammation at the joint tissue during RA.

The pro-inflammatory environment in the joint causes a functional phenotypic shift in chondrocytes and FLS as they begin to secrete cartilage-degrading enzymes such as MMPs and aggrecanase[61]. Hyperplasia of the synovium ultimately leads to the formation of a structure known as the pannus, at which activated osteoclasts begin to degrade the subchondral bone[61] (Figure 1.3). Pro-inflammatory cytokines such as TNF and IL-1 are also involved in the production of macrophage colony-stimulating factor (M-CSF) and RANKL by FLS and T cells within the synovium, both of which are essential for osteoclastogenesis[61]. There is also evidence to suggest that the development of these bone-degrading osteoclasts is stimulated directly by ACPAs themselves[62].

Understanding the triggers involved in the initial self-directed immune response, as well as transition to the joint and establishment of clinical synovitis, necessitates the dissection of cellular phenotypes in the peripheral blood, as well as the synovial tissue itself. The adaptive immune response is undoubtedly critical in the early disease stages, though the data discussed above also suggest a role, together with the joint resident stromal cells, in mediating pathological processes within the synovial micro-environment. Studying circulating lymphocytes such as CD4$^+$ T cells may also give pathogenic insights relating to the synovium, as a population of these cells appear to be phenotypically analogous to those infiltrating the joint tissue[63].

1.2.5 Clinical presentations and current targets for treatment

Diagnosis of RA in the clinic presents a significant challenge, given the considerable heterogeneity in its clinical presentation, as well as overlap in symptoms with a range of other arthropathies. Indeed, typical symptoms with which an RA patient will present are swollen and tender joints, together with elevated markers of the acute phase response such as C-reactive protein (CRP) and erythrocyte sedimentation rate (ESR), which signify a systemic inflammatory response. Joint involvement, as measured by the number of joints (out of 28 in total) showing swelling or tenderness, can be combined with these measures of inflammation to give a disease activity score (DAS28-CRP, DAS28-ESR). The American College of Rheumatology (ACR) and European League Against Rheumatism (EULAR) diagnostic criteria outline joint involvement and acute phase proteins, together with serological tests of autoantibodies (RF and ACPA) and symptom duration (<6 weeks / ≥6 weeks) as important RA criteria[64]

In addition to presenting a challenge diagnostically, the heterogeneity in RA clinical presentation also results in highly variable treatment responses. The current aim of treatment strategies is to achieve low disease activity (often termed remission) based on the measures described above (DAS28, CRP, ESR) as well as patient-reported general health status. Nonsteroidal anti-inflammatory drugs and steroids such as glucocorticoids are frequently prescribed for short-term suppression of the inflammatory response and amelioration of RA symptoms including pain and stiffness. However, these treatments have no effect on the disease course as they do not target cells or secreted immune modulators involved in disease activity.

The development of disease-modifying anti-rheumatic drugs (DMARDs) has enabled clinicians to target the immune response and halt disease progression. A first line therapy usually prescribed is the conventional synthetic DMARD methotrexate, a folic acid analogue that acts through a number of putative mechanisms including promotion of adenosine signaling, increased production of reactive oxygen species, and downregulation of adhesion molecules[65]. This drug shows a relatively good efficacy in patients, with around 40% achieving a ≥50% improvement in disease activity according to the ACR criteria[65].

An increased understanding of the cell types and cytokines involved in RA pathogenesis has stimulated the development of biological DMARDs (often termed biologics) that target specific immunogenic pathways. Biologics widely used in the clinic include those targeting the cytokines TNF and IL-6. Infliximab, a mouse chimeric monoclonal antibody (mAb) that binds TNF-α, reduced clinical symptoms and improved patient quality of life[66, 67], and humanised analogues and receptor fusion proteins that target the same pathway continue to be used (Figure 1.4). Similar efficacy has been achieved by blocking the IL-6 receptor using the mAb tocilizumab, which can also be effective in patients who fail anti-TNF treatment[68, 69]. These drugs are now routinely used to treat patients in the clinic who show poor response to first line therapies such as methotrexate. That these biologics engender changes in a number of outcomes including cellular infiltration into the synovium, bone destruction, and systemic inflammation confirms the pleiotropic role of key cytokines in pathogenesis.

The central role of lymphocytes in RA also makes them an enticing therapeutic target. Co-stimulation of T cells can be blocked with abatacept, a fusion protein consisting of CTLA-4 and the Fc region of IgG1, which binds CD80/CD86 thus out-competing the CD28 co-stimulatory molecule on T cells. Again, the efficacy of T cell-targeting agents such as abatacept in ameliorating the symptoms associated with RA provides further support a role for these cells in mediating disease processes[70]. An interesting observation is that abatacept may show greater efficacy in ACPA+ than ACPA- RA, suggesting that that T cells have a more prominent role in this particular disease subtype[71]. Depletion of B cells, such as occurs through anti-CD20 targeting by Rituximab, has also proven to be an effective approach[72]. This particular mAb has halted bone erosion and narrowing of the joint space in RA patients and may function in part indirectly through depletion of CD4$^+$ T cells[73, 74].

A range of additional biologics and small molecule inhibitors targeting cytokine and cell signaling pathways have also been developed and evaluated in clinical trials, with varying success in efficacy and translation to clinical application[51]. Given that there are no current biomarkers that predict response to treatment with various available DMARDs, the prescription of expensive biologics to patients is essentially trial and error. Future work must therefore focus on identifying subgroups of patients who are likely to respond to specific treatments, facilitating a stratified therapeutic approach. Moreover, whilst most drugs currently used in the clinic target the immune component of RA, concomitant blockade of pathological mechanisms in the joint tissue itself may prove an effective strategy. For example, inhibition of osteoclast differentiation is possible by targeting RANKL, and hyper proliferative FLS cells can be targeted with small molecule janus kinase (JAK) pathway inhibitors such as tofacitinib. In addition, clinical trials using cellular therapies such as tolerogenic dendritic cells offer promise that re-establishment of immune tolerance may be possible in RA patients[75]. The targets of disease-modifying therapies currently diagnosed for the treatment are summarized in Figure 1.4.

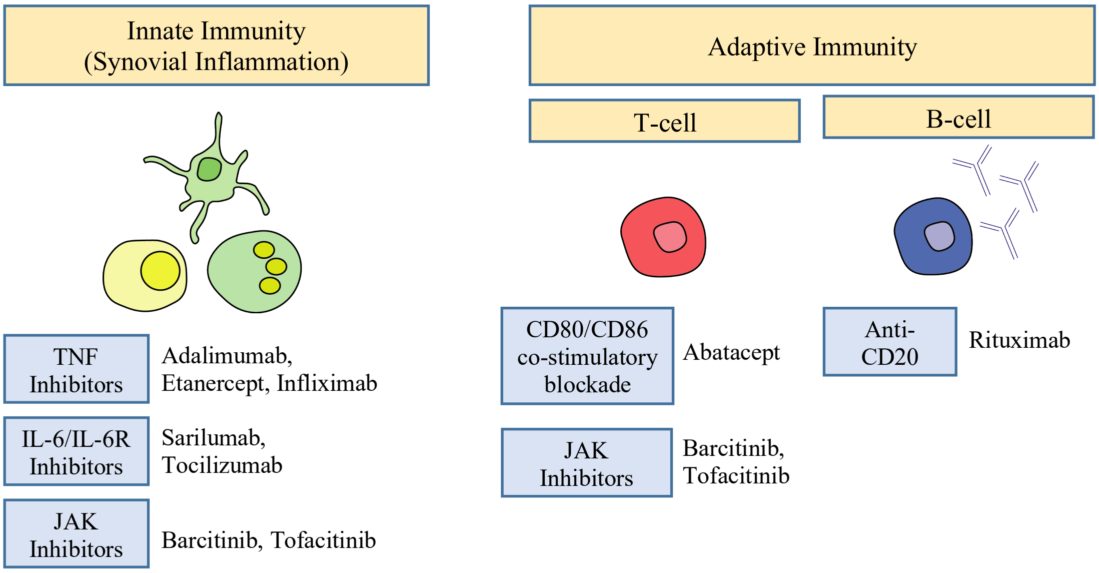


Figure 1.4: Cytokines and cellular targets of disease-modifying drugs currently diagnosed in the clinic for treatment of rheumatoid arthritis. Therapies shown here are those that specifically target immune pathways.

1.3 The genetic basis of RA

The aetiology of RA is complex, with diverse factors contributing to an individual's risk of developing the condition. Epidemiological studies have established clear links between RA and sex (2-3x higher prevalence in females), as well as age (highest prevelance in the fifth decade of life)[3]. Established mechanisms to explain the considerable discrepancy in female and male prevalence remain elusive. The disease also displays clear geographical bias, with the highest rates in Northern Europe and North America (~0.5-1.1%), and the lowest rates in Asia and Africa (~0.1-0.3%)[3]. Whether these geographical disparities represent contrasting environmental exposures or reflect genetic ancestry is unclear. Amongst the environmental risk factors identified to date, smoking provides the strongest evidence from both epidemiological and mechanistic studies, though other exposures have been suggested which may potentially trigger the initial extra-articular immune response (discussed in greater detail throughout section 1.4).

What is clear is that variation in an individual's genome has a considerable impact on susceptibility to RA. A significant heritable component was evident from early observations noting the tendency of RA to aggregate in families. Recently, a large study in Swedish patients established that the odds of developing RA were approximately 3 times higher in individuals who have a first-degree relative with the disease, than in those without such affected relatives[76]. Moreover, the same study found a more pronounced familial aggregation in patients with ACPA+ disease than in those with negative autoantibody status[76]. This is particularly important when studying the aetiology of RA, as it suggests that different genetic risk factors play a role in the development of these distinct disease serotypes.

Twin comparisons have also been useful in estimating the overall heritability of RA. A study of disease discordant twins found that the proportion of the variability in RA susceptibility that is attributable to additive genetic factors was 53% and 65% in United Kingdom and Finnish populations respectively[77]. However, this was a relatively small study and recent estimates place the heritability of ACPA+ RA at 39%[78]. Although a range of figures for RA heritability have been reported, it is believed that roughly 50% of an individual's susceptibility to RA is determined by genetic factors, with this figure dropping to 20% in seronegative RA[76, 79]. However, interpretation of disease heritability from such familial studies is confounded to some degree by the shared environmental factors within families, and this may perhaps explain some of the inflated estimates. Despite the prevalence of RA being higher in females and increasing with age, heritability estimates do not appear to be influenced by sex or patient age at disease onset[76, 77].

1.3.1 The HLA locus

By far the strongest genetic contribution to RA is that of alleles within the Human Leukocyte Antigen (HLA) region of chromosome 6. As early as 1978, it was recognised that the HLA-DRw4 (HLA-DRB1*04) alleles were overrepresented in RA patients compared with healthy controls[80]. This then led to the discovery of a conserved amino acid sequence motif at positions 70-74 of the HLA-DRβ1 chain, termed the 'shared epitope' (SE), that was common to all the disease-associated haplotypes[81]. It was later revealed that genetic risk conferred by alleles at the shared epitope was specific to patients seropositive for ACPAs[82]. Interestingly, the influence of smoking as a risk factor in seropositive disease appears to be linked to the presence of the shared epitope of the HLA-DRB1 molecule, conferring multiplicative disease risk and illustrating interplay between genes and environment [83, 84] (see section 1.5.1).

A more recent study using high-density genotype data attributed the HLA association in RA to a total of five amino acid positions; three in the HLA-DRβ1 chain, as well as single positions in HLA-DPβ1 and HLA-B[85]. Moreover, as these variable positions are located within the antigen-binding groove on the major histocompatibility molecule (MHC), this has potential implications for the presentation of self-peptide to both CD4$^+$ and CD8$^+$ T cells[85].

The mechanism through which sequence variation at the HLA locus contributes to a break in immune tolerance remains unclear. HLA haplotypes also account for considerable genetic risk in autoimmunity more generally, with associations in this locus identified in T1D[86], MS[87], ankylosing spondylitis (AS)[88], and systemic lupus erythematosus (SLE)[89]. Interestingly, the specific alleles conferring risk are often distinct in each disease. For example, in MS the DRB1*15:01 allele displays the strongest association[87], whereas in SLE the HLA associations have been assigned to the class I alleles B*08:01 and B*18:01, and the class II alleles DQB1*02:01, DRB3*02:00 and DQA*01:02[89].

The RA-associated SE alleles at the P4 pocket of MHC class II molecule conferred increased binding affinity to citrullinated peptides as opposed the native arginine-containing peptides[90, 91]. This preferential binding was corroborated in a recent study that characterized hierarchical binding affinities of 34 native/citrullinated self-peptides to HLA-DRB1 molecules[92]. This enhanced binding and subsequent presentation of citrullinated peptides may explain the association between the SE and ACPA+ RA.

Indeed, this is supported by the finding that T cells specific for citrullinated self-peptides, including vimentin and aggrecan, display an activated, immunogenic phenotype[93, 94]. Not only can HLA-DRB1*04 allele carriage enhance affinity of citrullinated peptides for the T cell receptor (TCR), in RA patients CD4$^+$ T cells appear to be more abundant and are enriched for a Th1 memory phenotype relative to healthy individuals with this haplotype[95]. A recent study showed that, in mice carrying a homologous HLA-DRB1*04 allele, immunisation with exogenous PADs resulted in the generation of anti-citrullinated fibrinogen antibodies[96]. Notably, a T cell response to PAD enzymes themselves occurred, and the production of anti-PAD antibodies in the absence of T cell responses to fibrinogen may indicate a mechanism whereby the recognition of citrullinated peptides by T cells requires that they are bound to the PAD enzyme[96]. However, the relevance of this mechanism in the context of RA is yet to be explored.

Despite the considerable contribution of polymorphisms in the HLA locus to RA susceptibility (Odds Ratio ~2.4)[97], and compelling mechanistic links between allelic variants and a break in self-tolerance, this locus still only explains approximately 11% of the heritability in RA (18% for ACPA+, 2% for ACPA-)[98].

1.3.2 The GWAS era

Technological advances in genotyping arrays, together with the inception of genome-wide association studies (GWAS), has galvanized research into the genetic architecture of complex diseases such as RA. The Wellcome Trust Case-Control Consortium association study in 2007 recapitulated the strong genetic associations with *HLA-DRB1* variants, as well as a variant in *PTPN22* – a gene encoding the protein tyrosine phosphatase non-receptor type 22[99]. This study also highlighted putative risk variants mapping to genes including *CTLA4* and *IL2RA*/*IL2RB*[99]. Successive studies have expanded the list of polymorphisms that confer susceptibility in RA, culminating in a recent meta-analysis in European and Asian populations of 29,880 RA cases (88.1% ACPA+, 9.3% ACPA-, 2.6% unknown status) and 73,758 non-RA controls, placing the number of genome-wide significant risk loci at 101[97]. This number continues to grow and more recently has expanded the total risk loci to 106 [100]. In isolation, the risk conferred by these GWAS SNPs is relatively low (odds ratio ~ 1.1-1.3 for most SNPs), and these non-HLA loci identified to date explain ~5% of the RA heritability[97], whereas the HLA variants account for ~13% of the overall RA genetic risk[85].

GWASs to date have predominantly focussed on patients with seropositive disease, and therefore the majority of conclusions from these studies can be extrapolated to this group. Indeed, consistent with the clinical heterogeneity of RA, the genetic aetiology of ACPA+ and ACPA- disease appears to be largely distinct. A comparison of patients categorised into these distinct disease subsets suggests that such differences may largely be accounted for by alleles at the HLA locus[101]. The identification of susceptibility genes outside of the HLA region in ACPA- RA has proved less fruitful in comparison to ACPA+ RA[102]. Whilst the presence of autoantibodies greatly facilitates the reliable diagnosis of seropositive RA, no such biomarker exists for seronegative disease. As a result, misdiagnosis of these patients due to clinical heterogeneity can confound studies of this serotype[103].

Increasing sample sizes and meta-analyses, perhaps with better classification of patients in the clinic and recruitment of larger cohorts, will be required to further decipher the genetic background in seronegative disease. Nonetheless, there are two loci, mapping to *ANKRD55*/*IL6ST* and *BLK*, which have been identified as risk loci in RA independent of serological status, highlighting a potentially common pathways in disease[104, 105].

As with the HLA region, risk variants identified in GWAS would suggest that to some degree AIDs share a common underlying genetic aetiology. By incorporating GWAS data from 7 AIDs (RA, Psoriasis, MS, SLE, Crohn's disease, Coeliac disease, T1D), Cotsapas and colleagues found 47 loci which showed evidence of association with multiple traits, implicating *IL23R*, *PTPN2*, *CTLA4*, *ORMDL3*, and *STAT4* amongst others as genes which may contribute to the autoimmune phenotype of cells in AIDs[106]. There is now also evidence to suggest that many autoimmune-associated pathways may be active in allergic disease, with 29 loci associated with susceptibility to allergy also found to be enriched amongst autoimmune loci[107]. This is congruent with observations that many AIDs appear to co-segregate in families at higher rates than would be expected by chance, as is evident for T1D and primary biliary cirrhosis (pairwise odds ratio (i.e. odds of having the disease in individuals with a family history of the other disease) = 4.6), amongst other diseases[108].

Identifying risk-associated SNPs can begin to give clues about the specific cell types that are involved in triggering the break in self-tolerance, or in driving progression to clinical disease. For example, the risk variant mapping to *CTLA4* is informative given the role of CTLA-4 in suppression of immune responses by regulatory T cells[109]. In general, genes putatively implicated by their proximity to RA risk SNPs suggest the crucial contribution of lymphocyte-mediated immune responses, as would be expected. Amongst the candidate genes identified to date, those with well-established roles in lymphocyte function include *IL2RA*, *STAT4*, and *CCR6* [97, 110-112]. Although over 100 non-HLA genetic risk loci have been identified, little is known about the mechanisms through which these contribute to disease.

1.4 Linking genotype to phenotype – expression Quantitative Trait Loci

Despite the relative success in employing GWASs to uncover genomic loci at which sequence polymorphisms confer increased risk of developing RA, translating this knowledge into an understanding of molecular mechanisms through which genetic risk influences cell-mediated pathogenesis lags some way behind. Given that risk SNPs overwhelmingly map to non-coding regions of the genome, with only 16% being in linkage disequilibrium (r^2 >0.8) with a missense variant[97], assigning which gene within a given locus is responsible for the disease association is troublesome. This is common to all GWAS hits across various autoimmune conditions, with approximately 90% of trait-associated SNPs falling outside of the protein-coding regions - and ~60% mapping to enhancers elements that exhibit cis-regulation of gene expression in a temporal and cell type specific manner[45]. In particular, SNPs associated with RA are over-represented in enhancer elements that are active specifically in T- and B-cells[45]. Together with the observation that 77% of GWAS hits are located within DNase I hypersensitivity sites (DHSs)[113], this would suggest that these variants have a regulatory function, and influence pathogenesis through modulation of transcriptomic profiles in disease-relevant cell types.

Whilst the genome is static and - with a small number of notable exceptions – identical in all cells of the human body, dynamic transcriptional programs dictate cell fate decisions and function. By integrating cell type-specific patterns of transcription factor (TF) binding and chromatin modifications with knowledge of disease-associated risk loci, the relative contribution of immune compartments to aetiology may be inferred. In RA this approach has confirmed the central role of lymphocyte subsets in pathogenesis, with risk-associated SNPs enriched in CD4$^+$ T cell and B cell (super-) enhancer elements that can orchestrate the activation of cell-specific gene expression profiles[45, 114-116]. In addition, GWAS SNPs associated with traits including RA, MS, T1D, asthma, and inflammatory bowel disease (IBD) are enriched in regulatory regions that are active in memory CD4$^+$ T cells shortly following cytokine stimulation[117]. As well as describing a potential function for AID-associated variants early following immune activation, these findings, together with observations that RA variants are enriched at genes and active regulatory regions specific to CD4$^+$ T cells (particularly naïve and regulatory (Treg) subsets), implicate these as critical pathophysiological cells[116, 118]. Consistent with this, a statistical approach to partition disease heritability into regions at which cell-specific gene expression occurs has strongly implicated T cells, and to a slightly lesser extent B cells, in RA pathogenicity[119].

Enhancer elements often reside many kilobases (Kb) from the transcription start site of the genes which they regulate, and can be in intergenic regions or the intron of unrelated genes[120]. For this reason, assigning a given SNP to a candidate gene is not straightforward. The development of technologies that enable the 3D chromatin structure to be assessed, such as chromosome conformation capture based techniques, have facilitated the association of enhancer-related risk variants to gene promoters by assessing physical interactions in 3D space[121]. Mapping long-range enhancer-promoter interactions in T cells has facilitated the nomination of RA candidate genes located up to 1 megabase (Mb) from the regulatory SNP[122].

An important approach to circumvent the issues surrounding non-coding variants, and nominate candidate genes at a given risk locus, is to assess the influence of SNPs on transcriptional activity. Identifying SNPs that regulate gene expression can facilitate the localisation of causal SNPs within extended haplotype blocks, as well as highlighting the genes and pathways involved in pathogenesis. Genomic loci with such regulatory capacity can be mapped genome-wide by seeking associations between allele copy number and transcript levels in large populations of genotyped individuals. These associations are termed expression quantitative trait loci (eQTLs). Such eQTLs can act via distinct mechanisms; be that in *cis*, whereby the gene is located proximal to the regulatory SNP, or in *trans* which occurs over an extended distance, often separate chromosomes[123].

These approaches have proved valuable in bridging the gap between the number of variants known to confer susceptibility, and candidate disease genes. Notably, the predicted capacity of a given non-coding SNP to function as an eQTL is a valuable criterion when selecting variants for more extensive functional studies[124]. Importantly, many eQTL effects can be specific not only to tissue and cell type[125-129], but also to the stage of cellular development[130]. The genotype-tissue expression (GTEX) project has successfully mapped eQTLs in a range of human tissues post-mortem, generating a vital resource for studying the tissue-specific impact of genetic variation on gene expression[131]. However, environmental context, such as immune stimulation during infection or AID, appears to influence these regulatory effects, and may account for inter-individual variability in immune responses[132-134]. For the reasons outlined here, if insights are to be gained into the role of eQTLs in complex disease, such mechanisms should ideally be studied in relevant cell types isolated from appropriate patient cohorts.

1.4.1 eQTL mapping in rheumatoid arthritis

Identification of eQTL effects in a range of cell types has been employed in complex immune-mediated conditions such as RA, with the aim of highlighting disease-relevant pathways for drug targeting or cellular therapy. By isolating pure populations of primary immune cells from healthy individuals, and assessing the capacity of RA risk variants to regulate transcript expression in *cis*, genes can be nominated for which up-/down-regulation impacts cellular phenotype and, consequently, immune responses[118, 128, 135]. Ishigaki and colleagues showed that cell-specific up-regulation of cytokine pathways, namely the TNF pathway in CD4$^+$ T cells, may be dependent on genotype, illustrating how multiple variants can generate a cell-mediated inflammatory response[136].

The first eQTL analysis to be carried out using whole blood transcriptomes from RA patients revealed a greater enrichment of RA-associated SNPs functioning as eQTLs in samples from RA patients relative to those form healthy controls, again emphasizing the importance of disease context[137]. The same study found that eQTL variants identified in whole blood were enriched in enhancer elements that are active in primary blood cells including monocytes, T cells, and B cells, demonstrating a degree of cell-type specificity[137]. One mechanism through which RA risk variants may influence gene expression in T cells is by modifying the accessibility of chromatin, allowing TFs to at bind regulatory elements[138]. Our group has performed the first eQTL analysis in CD4$^+$ T cells and B cells from untreated early arthritis patients, refining the regulatory landscape and revealing genes that warrant follow-up investigation regarding their role in the inception and progression of RA[139].

In conclusion, genetic data in isolation has limited interpretability with regards to identifying the mechanisms of cell-mediated pathogenicity. Integration of multiple layers of molecular data is now integral for understanding how risk-associated polymorphisms alter cellular function during disease, and this will be discussed further in section 1.6. In addition to providing mechanistic insights and highlighting disease-relevant molecular pathways, eQTL effects may also present a potentially useful biomarker when predicting responses of individual patients to biologic drugs[140].

1.5 Environmental risk factors

Though there is clearly a considerable genetic contribution to RA pathogenesis, the concordance rates of ~15% between monozygotic twins indicates that non-genetic exposures are integral in triggering disease induction[98]. There is growing evidence that initial triggering of the immune response occurs at mucosal surfaces where leukocytes are exposed to environmental stimuli, namely the lungs, the gingiva, or the gastrointestinal tract. Whilst the precise factors involved in triggering these responses remain poorly characterised, those for which the most substantial insights exist from epidemiological and mechanistic studies are discussed below.

1.5.1 Smoking

Exposure to cigarette smoke has emerged as a prominent risk factor in RA, with the association confirmed in both twin and case-control studies[141, 142]. In particular, the role of smoking as a determinant of RA susceptibility is confined to seropositive disease, with this effect particularly pronounced in individuals with an extended period of exposure, indicative of a dose-dependent response[142, 143]. Expressing the overall smoking exposure as cigarette pack years by normalising the number of cigarettes to smoking duration (1 pack year = 20 cigarettes per day for one year), revealed a dose-dependent increase in relative risk (relative to those who have never smoked)[142]. Interestingly, the effects resulting from smoking exposure appear to be reversible upon cessation, though this may take from 10-19 years for the elevated risk to subside[142].

Smoking as a risk factor presents a clear example of a gene-environment interaction in RA susceptibility. The influence of smoking as a risk factor is linked to the presence of the SE of HLA-DRB1 molecule[83, 84]. This interaction is only evident in the case of ACPA+ disease, with smokers carrying two copies of the SE alleles having 21-fold higher risk of developing RA relative to non-smokers with no copies of the SE alleles[144]. This is in contrast to a relative increased risk of 5.4 in non-smokers with two SE allele copies, or a 1.5 relative increased risk in smokers without SE alleles[144]. Following on from these discoveries, it has been demonstrated that, regardless of the risk alleles present at the SE locus, the observed interaction with smoking remains[145].

Whilst mechanisms have been proposed, precisely how smoking contributes to autoreactive immune responses remains to be deciphered. There is some evidence to suggest that smoking may lead to the upregulation of PAD enzymes that catalyse the post-translational citrullination of peptides in cells of the bronchiolar lavage[146]. This coincided with increased citrullination in these cells[146]. Indeed, the observation that ACPAs are enriched in the bronchiolar lavage fluid of early ACPA+ RA patients relative to the serum supports the hypothesis that the early dysregulated immune response may occur in the lungs[147]. This is further supported by the observation of immune cell infiltration and local activation of B cell responses in the lung tissue of ACPA+ RA patients[148]. More recently it has been proposed that, whilst environmental triggers including smoking are important in the initial production of ACPAs at mucosal surfaces such as the lung, genetic risk factors at the HLA locus determine the transition to autoimmune disease with associated clinical features[30, 78].

Despite being the most well-established environmental risk factor in RA, further work is necessary to decipher precisely how this exposure may trigger an autoimmune response. Interestingly, smoking also represents a prominent risk factor in autoimmune conditions such as MS[43].

1.5.2 Periodontal disease

An intriguing link between periodontal inflammation, caused by *Porphyromonas gingivalis* bacterial infection, and RA susceptibility has been described[149]. These bacteria express a form of the PAD enzyme that is responsible for the post-translational citrullination of peptides described earlier [150]. Indeed, these bacterial PADs exhibit the potential to citrullinate the human fibrinogen and α-enolase proteins, suggesting they may contribute to the formation of neoantigens that are the target of the autoimmune response during RA[151]. Immunogenicity directed towards *P.gingivalis* is elevated in RA patients relative to controls, as well as in ACPA+ relative to ACPA- patients, and appears to interact with the shared epitope as well as smoking status[152]. More recent work had posited that a distinct pathogen, *Aggregatibacter actinomycetemcomitans*, may trigger citrullination in host neutrophils and thus links periodontal disease to autoimmunity[153]. While perturbations in the species composition of the microbiome have been liked to many AIDs, the mechanistic link between *P. gingivalis,* protein citrullination, and autoimmunity appears to be specific to RA. Nonetheless, periodontal disease resulting from *P. gingivalis* infection can exacerbate insulin resistance in mice fed a high fat diet[154]. This appears to occur via activation of adaptive immune responses, with expanded populations of T cells ($CD4^+$ and $CD8^+$) and B cells in the spleen and local lymph node, highlighting a potential link with T1D[154]. The effect of periodontal infection was not observed in mice fed normal chow, however, suggesting that diet also plays an integral role in this mechanism.

Despite the possible mechanisms above, the link between the oral mucosa and triggering of RA remains tentative, and the association between periodontitis and RA was not replicated in a recent analysis of a Swedish cohort[155]. However, using a metagenomics approach to study the microbiomes at multiple mucosal sites of RA patients and controls, Zhang and colleagues were able to detect dysbiosis at both the oral and gut sites[156]. Notably, microbial balance was partially restored upon treatment with DMARDs[156]. In this study *P.gingivalis* was found to be enriched at the oral sites (saliva and dental plaques) in control samples, potentially discrediting the link between this species and RA[156].

Though a number of exposures have been highlighted in epidemiological studies, and work has now begun to attempt to decipher potential mechanistic links with autoimmunity in RA, the environmental contribution to RA susceptibility remains largely unexplained. Going forward, we must also consider how genetic and environmental risk factors interact, as has been convincingly shown to be the case for smoking and the shared epitope. As will be discussed in the subsequent section, the epigenome may represent an important level at which DNA variants and environmental exposures converge to cause aberrant immune responses.

1.6 Epigenetics

Though risk-associated polymorphisms in RA appear to exert their effect on disease initiation and progression through the modulation of cell-specific transcriptional profiles, twin studies in RA predict that at least 50% of an individual's susceptibility is conferred through non-genetic risk factors.

'Epigenetics' is often used to describe heritable changes in gene expression and/or phenotype that occur in the absence of changes to the primary DNA sequence. Such modifications can influence transcriptional programmes, and as such are tightly involved in the determination of cell fate and function. Genetic variation and environmental exposure may to some extent converge at the level of the epigenome, and as such these epigenetic changes may be impacted by multiple risk factors, in turn regulating gene expression.

Mechanisms through which this epigenetic regulation may occur are chemical modification of either the DNA nucleotides themselves or the histone proteins around which the DNA is packaged in the nucleus, as well as non-coding RNAs which can regulate gene expression at the transcriptional, translational, and post-translational level.

1.6.1 DNA methylation

Perhaps the most frequently studied epigenetic modification with respect to its putative role in disease mechanisms, or clinical utility as a biomarker, is the addition of a methyl group (CH_3) to carbon atom at the 5th position of cytosine nucleotides to create 5-methylcytosine (5mC) residues. In humans, his predominantly occurs in the context of cytosine-guanine dinucleotides (CpGs). Though rare, non-CpG methylation at CHG or CHH sites (here H represents non-guanosine nucleotides) has been described in animals including humans[157]. In human cells, DNA methylation (DNAm) is pervasive throughout the genome, with 5mC found at approximately 70-80% of these CpG sites[158]. A family of highly conserved DNA methyltransferase enzymes (DNMT1, DNMT3A, DNMT3B) catalyse the conversion of cytosine to 5mC in mammalian cells by transfer of the methyl group from a methyl donor – S-adenosylmethionine (SAM) - to the cytosine residue (Figure 1.5A)[159].

Conversely, whilst DNA de-methylation was initially believed to be a, exclusively passive process in humans, the recent discovery of a group of enzymes, termed ten-eleven translocation methylcytosine dioxygenases (TET1, TET2, TET3) that catalyse the conversion of 5mC to 5-hydroxymethylcytosine (5hmC), demonstrates that active de-methylation can occur[160]. Subsequent to their identification, it was revealed that these enzymes are involved in a series of iterative oxidization reactions that convert 5mC to 5hmC, then to 5-formylcytosine (5fC), and finally 5-carboxycytosine (5caC)[161] (Figure 1.5B). As regards active DNA de-methylation, following the oxidization of 5mC by the TET enzyme, the oxidized bases are passively diluted to cytosine by DNA replication (Figure 1.5C). This therefore signifies a mechanism whereby DNA methylation can be reversed in terminally differentiated cells. In certain contexts however, such as the global de-methylation that occurs during re-programming of mouse embryonic stem cells to a pluripotent state, active de-methylation can occur independent of the activity of TET enzymes[162].

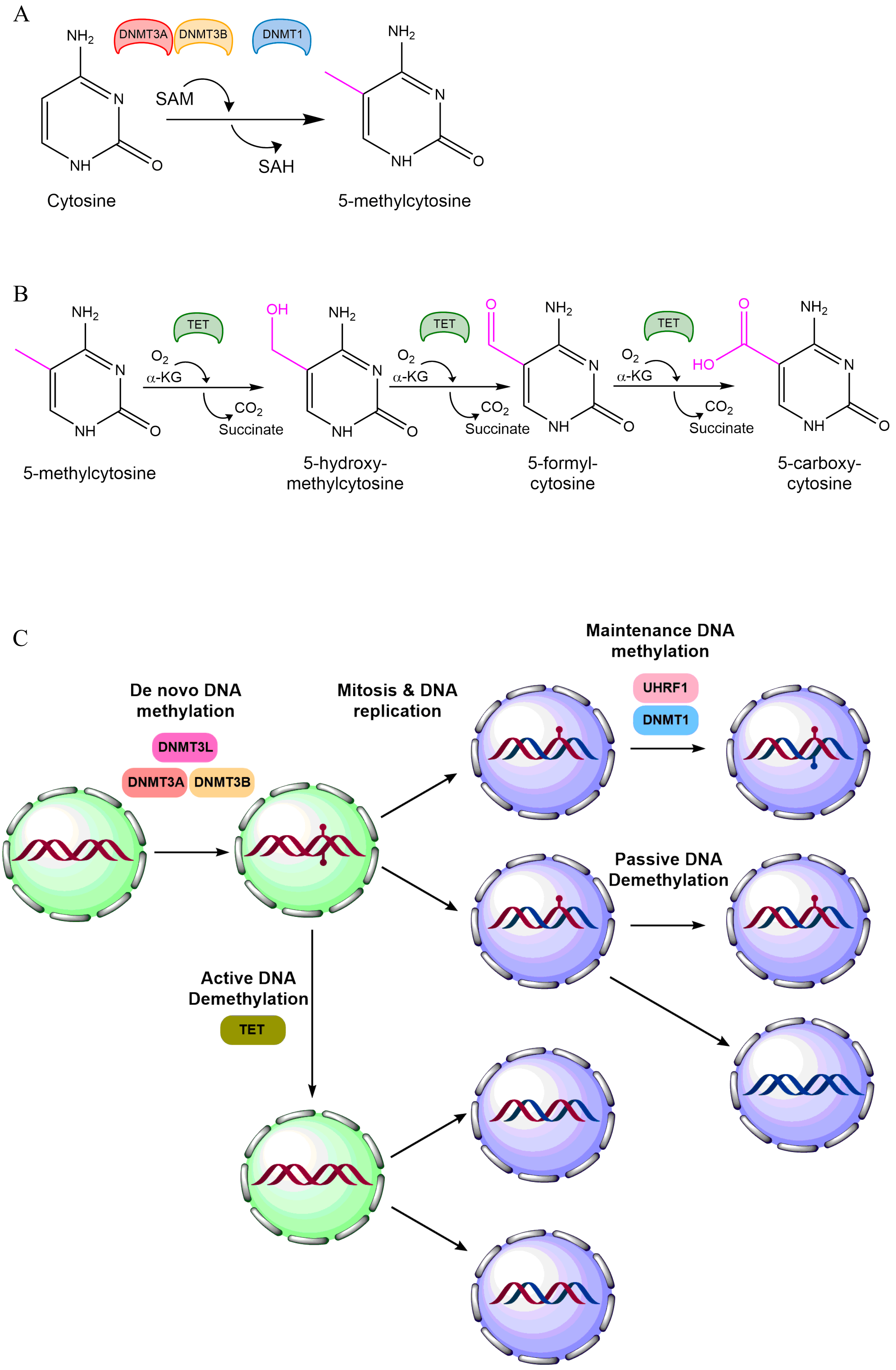

A
DNMT3A
DNMT3B
DNMT1
NH2
N
NH
O
SAM
SAH
Cytosine
5-methylcytosine
B
TET
O2
α-KG
CO2
Succinate
OH
HO
5-methylcytosine
5-hydroxy-methylcytosine
5-formyl-cytosine
5-carboxy-cytosine
C
De novo DNA methylation
DNMT3L
DNMT3A
DNMT3B
Mitosis & DNA replication
Maintenance DNA methylation
UHRF1
DNMT1
Passive DNA Demethylation
Active DNA Demethylation
TET

Figure 1.5 (previous page): Mechanisms of DNA methylation and de-methylation in mammalian cells
A) Active methylation by the DNA methyltransferase (DNMT) enzymes. These enzymes catalyse the conversion of cytosine to 5-methylcytosine. S-adenosylmethionine (SAM) acts as a donor for the methyl group, and in the process is converted to S-adenosylhomocysteine (SAH). (B) Molecular mechanism of ten-eleven translocation (TET) enzyme-mediated active DNA de-methylation by sequential oxidization of 5-methylcytosine to produce 5-hydroxy-methylcytosine, 5-formyl-cytosine and, ultimately, 5-carboxy-cytosine. Each stage of this TET-mediated process requires the substrates oxygen and α-ketoglutarate (α-KG), with the production of carbon dioxide and succinate along with the oxidized nucleotide. (C) Establishment of DNA methylation patterns at a cellular level. De novo DNA methylation occurs through the actions of DNMT3a and DNMT3b. As cells undergo mitosis, the daughter cells each contain one copy of the methylated DNA strand, with the newly synthesized strand lacking this modification. DNMT1 functions to maintain patterns of DNA methylation in daughter cells by recognizing hemi-methylated DNA and adding a methyl group to the newly synthesized strand. In the absence of DNMT1 activity, passive DNA de-methylation occurs as successive rounds of mitosis dilute the methylated DNA, such that after 2 rounds of mitosis, daughter cells completely lacking methylation are generated. Active DNA de-methylation can also occur due to the activity of the TET enzymes. Figure adapted from Wu & Zhang (2014)[159].

The methyltransferase enzymes have distinct roles in the establishment and maintenance of DNAm during the cell cycle. DNMT3a and DNMT3b are responsible for establishing de novo patterns of DNAm in human and other mammalian cells during early development (Figure 1.5C)[163]. The activity of both these de novo methyltransferases can be regulated by a third family member, DNMT3L, which lacks catalytic activity but can stimulate the activity and targeting of the other two family members[164-166].

DNMT1, which was the first DNA methyltransferase enzyme to be identified and cloned, preferentially acts on hemi-methylated sites and is therefore critical in the maintenance of CpG methylation as cells undergo DNA replication and mitosis[167, 168] (Figure 1.5C). The ubiquitin ligase protein UHRF1 is central to this process due to its role in recruiting DNMT1 to sites of hemi-methylated DNA[169]. Given that DNMT1 is necessary for maintenance of methylation across cell division cycles, passive de-methylation can occur in the absence of functional DNMT1 or UHRF1 (Figure 1.5C), such as may be the case during resetting of methylation in the early embryo through nuclear exclusion of these proteins[170].

A novel role for TET enzymes in regulatory T cell (Treg) function has been described, whereby knock out of TET2 and TET3 in mouse Tregs led to the hyper-methylation of the *Foxp3* gene and ablation of their regulatory capacity, with the cells taking on an effector phenotype[171]. This example demonstrates the role of TET-mediated active de-methylation in controlling the differentiation and function of cells, such as those that orchestrate immune responses.

1.6.2 Histone modifications

The DNA in a cell is organised around a core of histone proteins, composed of two copies each of the histones H2A, H2B, H3, and H4, to form a nucleosome. These nucleosomes can then condense even further to form the densely packed heterochromatin through the action of the linker histone H1. Chemical modifications to the histone tails, including acetylation and methylation, can alter the accessibility of TFs to regulatory regions in the DNA, thus dictating patterns of gene expression and cellular phenotype[172].

Histone modifications are useful in delineating the transcriptionally active or repressed regions of the genome. For example, tri-methylation at lysine 4 of histone H3 (H3K4me3) is enriched at active promoters, whereas mono-methylation at this position and acetylation at lysine 27 (H3K27ac3) are associated with enhancer regions and active promoters[173]. Conversely, tri-methylation at lysine 27 of H3 (H3K27me3) is indicative of heterochromatin formation, and thus repression of gene expression[173].

To date, studies of chromatin modifications in RA have largely focussed on the FLS cells. Ai and colleagues extensively profiled the epigenome of these cells from both RA and OA patients, including chromatin immunoprecipitation with sequencing (ChIP-seq) assays to map histone modifications (H3K4me1, H3K4me3, H3K9me3, H3K27ac, H3K27me3, and H3K36me3)[174]. This approach identified genomic regions that share similar epigenetic modifications, and regions harbouring active promoter and enhancer histone marks were identified that differed between RA and OA patients[174]. Upregulation of bone and cartilage-degrading MMPs (including MMP1, MMP3, MMP9, and MMP13) in RA FLS is also associated with histone modifications at their promoters, with increased levels of H3K4me3 (active) and a reduction in H3K27me3 (repressive), which facilitates STAT3 binding downstream of IL-6 signalling[175]. This may also represent a promising therapeutic avenue, and inhibition of histone deacetylase (HDAC) 3 in RA FLS was able to suppress pro-inflammatory type I IFN signalling[176].

1.6.3 Non-coding RNAs

Despite only ~2% of the human genome sequence comprising protein-coding exons, non-coding regions have essential regulatory roles, including non-coding RNAs which are not translated but are central in both development and disease[177]. Micro RNAs (miRNAs) are approximately 22 nucleotides long and function to repress transcription of protein-coding genes by either targeting messenger RNA (mRNA) for cleavage, or repression of translation by ribosomes[177]. Altered expression of a number of miRNAs has been described in both the peripheral blood and synovial tissue of RA patients[178]. For example, miR-146a is upregulated in CD4$^+$ T cells from the peripheral blood and synovial tissue of patients with RA relative to healthy controls[179]. This particular miRNA may contribute to pathogenesis and chronic inflammation in RA by suppressing apoptosis in these cells[179]. These miRNAs may also contribute to the hyper proliferative state that characterises FLSs of the RA joint. A reduction in the expression of miR-34a* in FLSs from patients with RA relative to those with OA can render these cells more resistant to apoptosis[180].

Long non-coding RNAs (lncRNAs), which are >200 nucleotides in length, are involved in a diverse range of regulatory processes, including transcriptional regulation through recruitment of TF and complexes that modify chromatin structure[177]. Through transcriptomic profiling of FLSs from 10 RA patients and 10 healthy controls, Zhang et al. found 135 lncRNAs to be differentially expressed, with 67 up-regulated in RA FLSs[181]. Levels of one of these lncRNAs, ENST00000483588, showed a positive association with both CRP levels and disease activity in these patients, though whether its expression precedes - or is a downstream consequence of - inflammation is unclear[181]. Numerous lncRNAs with potential pathological roles in RA have been described in the blood and synovium, though validation studies and further functional characterisation is required to gain mechanistic insights[182].

1.6.4 DNA methylation in transcriptional regulation

DNAm is typically linked to genome silencing and transcriptional repression, including well-characterised roles in processes such as X chromosome inactivation to ensure transcription from one chromosome copy in females[183]. Whilst the majority of CpG sites in human cells exist in a methylated state, clusters of CpGs generally located at gene promoters, termed CpG islands (CGIs), predominantly remain unmethylated in a state that is associated with active transcription[184]. In more general terms however, the relationship between DNAm and gene expression is complex and relies on intricate spatial and temporal interactions between DNAm, other epigenetic modifications such as histone methylation/acetylation, and TFs. This relationship also appears to be dependent to some extent on the genomic context of a CpG site. Ball et al. (2009) demonstrated that promoter regions of highly expressed genes exhibit low levels of methylation, whereas extensive methylation is present at the gene bodies themselves[185]. In contrast, genes with a lower level of expression showed an intermediate level of methylation, which was relatively constant across both the promoter region and the gene body[185]. Furthermore, the density of CpG sites at a promoter region appears to be intimately related to the methylation levels at those sites. For example, promoter regions with a high CpG content exhibit reduced methylation relative to those with a lower CpG density[185]. The observation that genes can still be transcriptionally active in the presence of promoter methylation is further proof that DNAm should not solely be considered as a repressor of transcription[186]. Interestingly, analysis of tissue-specific DNA methylomes and transcriptomes has revealed that differentially-methylated regions (DMRs) in intragenic regions display the strongest associations with expression, and may therefore be important for instructing tissue-specific patterns of transcription[187].

Whether DNAm plays an active role in regulating transcriptional programmes in the cell, or rather DNA de-methylation actually represents a downstream footprint of TF binding and increased chromatin accessibility is still ambiguous. Most studies *in vivo* report negative correlations between DNAm at CpG sites and transcript levels of proximal genes, though in many instances no method is employed to infer causality or take into account the order in which these events occur. Using infection of dendritic cells as a model system to study the temporal dynamics of DNAm and transcription, Pacis and colleagues were able to show that recruitment of the transcriptional machinery and gene activation precedes enhancer de-methylation[188]. In this particular system it may be that modification of the DNAm status at enhancer elements allows 'priming' of these genes that are upregulated in response to infection, allowing subsequent responses to be more rapid.

However, perhaps the best evidence for an active role of DNAm in coordinating transcriptional programmes comes from the observation that the steric interactions between TFs and DNA is disrupted by the presence or absence of methylation[189]. Whilst DNAm has been linked to repression of transcription, global analysis of the effects of CpG methylation on TF binding has revealed a complex picture whereby methylation at its cognate binding site can either weaken or enhance the binding affinity of different families of TFs[190-192]. Moreover, this modulation of TF binding is also sensitive to the position of the methyl modification across the TF binding site (TFBS) region[192]. TFs also have the capacity to shape local DNAm upon site-specific binding, illustrating that the relationship is bi-directional[193, 194]. Given that the DNMTs and TETs do not display sequence-specific recognition capabilities, it seems likely that TFs are able to bind and recruit the methylation machinery. This would be in agreement with observations that 5hmC, which is produced by the oxidisation of 5mC by TET enzymes (see section 1.6.1), is significantly enriched at TFBSs in the human genome[195].

The precise mechanisms through which DNAm may control transcriptional activity are yet to be comprehensively described. Consistent with the methylation-sensitive binding properties of TFs, DNAm may directly influence expression of genes by interfering with binding of these factors to regulatory elements, as has been described for the E2F family[196]. Alternatively, inhibition of transcription may occur indirectly through the recruitment of proteins that recognise methylated DNA[197]. One such family of proteins are those containing a methyl-CpG binding domain (MBD). Members of this family include MeCP2 and MBD1-6, although not all have been shown to bind methylated DNA *in vitro*[198-200]. In addition to these MBD-containing proteins, the Kaiso family bind methylated CpGs via a C-terminal zinc finger domain[201, 202]. A third family of proteins are able to recognise methylated DNA through binding of a SET and RING finger-associated (SRA) domain[203, 204]. Many of these proteins have been shown to recruit histone remodelling machinery such as HDACs upon binding to methylated DNA, forming repressor complexes that establish stable downregulation of gene expression[203-206]. The interaction of MeCP2 with HDACs also has an important function in regulating alternative splicing of mRNA, with recruitment of MeCP2 to the methylation-enriched alternatively spliced exons promoting their inclusion in the mature transcript[207].

Though clear associations between chemical modification of DNA nucleotides and transcriptional activity have been described *in vivo*, and TF-DNAm interactions described *in vitro*, future work that aims to gain mechanistic insights must attempt to more comprehensively unravel the precise regulatory function of DNAm. Given the role of DNAm in transcriptional regulation, it has emerged as an interesting molecular signature for mechanistic studies into the pathogenesis of complex diseases.

1.6.5 Epigenome-wide association studies

An experimental approach aimed at capturing disease-associated DNAm signatures is to design hypothesis-free studies to compare the methylation levels between patients and controls at multiple sites in parallel. These epigenome-wide association studies (EWASs) seek to identify differentially methylated positions (DMPs) or regions (DMRs) between cases or controls which may highlight disease-relevant molecular pathways or have clinical utility as biomarkers for patient diagnosis or stratification. The most common methods of quantifying DNAm at multiple CpGs in parallel involve first performing bisulphite conversion of the sample DNA to convert unmethylated cytosine (C) residues to uracil (U; Figure 1.6). Following conversion, DNA can be amplified whereby DNA replication results in the conversion of uracil to thymidine (T), and the relative proportions of C to T residues at a given position giving a proxy readout of the methylation levels in the initial pool of sample DNA.

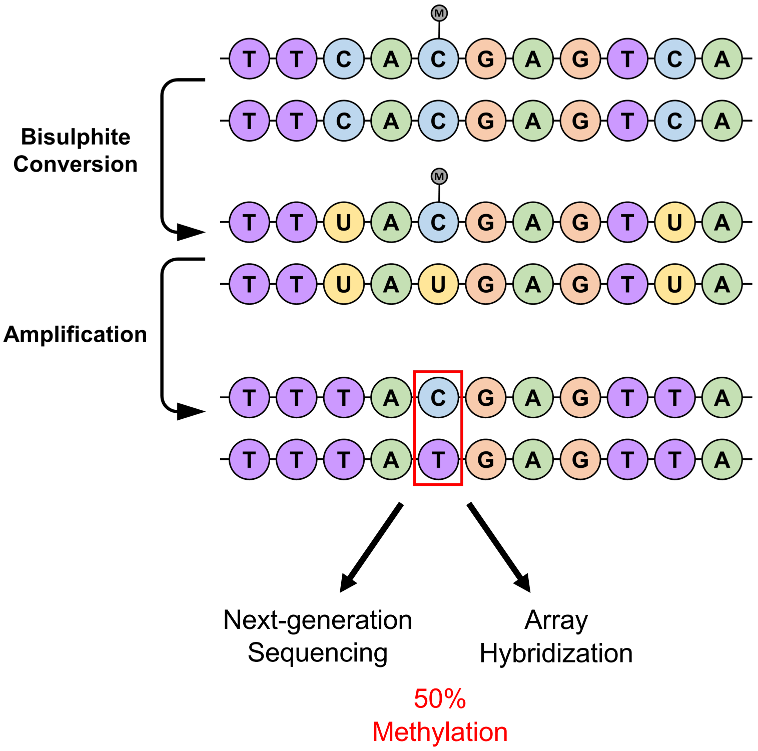


Figure 1.6: Bisulphite conversion of DNA allows for the levels of methylation to be quantified. This method involves treatment of DNA with sodium bisulphite which results in the conversion of unmethylated cytosine (C) residues to uracil (U), whilst methylated cytosine residues are protected against conversion. Following amplification, all unmethylated cytosine residues will be represented by thymidine (T) residues and quantifying the ratio of cytosine to thymidine in the converted DNA by either next-generation sequencing or array hybridisation gives a read out of methylation levels in the original sample.

Methods differ in how nucleotide quantification is performed in the final step, either by whole-genome sequencing (WGS) or array-based methods[208]. Whilst WGS offers the advantage of being able to quantify DNAm at every CpG site genome-wide, the Illumina Infinium™ methylation arrays have proved a popular tool for EWAS given they offer the capacity to assay up to 8 samples in parallel at a relatively low cost. The current platform distributed by Illumina, the Infinium™ MethylationEPIC BeadChip, allows for >850,000 methylation sites to be assayed per human sample. Importantly, the chip is designed to target regions including CGIs (>90% of human CGIs), regions of open chromatin (>66%) and TFBSs (>78%) from the Encyclopaedia of DNA Elements (ENCODE) project[209, 210], as well as enhancers identified by the functional annotation of the mammalian genome 5 (FANTOM5) project[211]. Probes on the MethylationEPIC array predominantly target promoter regions (54%), with the others lying in gene bodies (30%) and intergenic regions (16%)[212].

Whilst the study design of EWAS aims to identify disease-associated epigenetic modifications that occur independently of the genome sequence, all sources of variability in DNAm between individuals diagnosed with a disease and those who are unaffected must be considered. Firstly, given that DNAm is highly cell-type specific, when complex tissues are assayed, the relative proportions of cell types should be taken into account. Another issue is that of reverse causality, as DNAm may be influenced by inflammatory markers such as CRP, which will inevitably be elevated in patients with autoimmune responses relative to healthy controls[213, 214]. Whilst such associations may be interesting with regards to biomarker discovery, they can considerably limit mechanistic insights into the function of DNAm in pathogenesis. For this reason, studying patients in the earliest stages of disease as is possible, as well as those who are yet to receive any form of therapy that alters the course of disease, is desirable. Finally, the impact of functional DNA polymorphisms that shape methylation changes in cis must be taken into account when considering differences in patient and control methylomes, which will be discussed in detail in section 1.6.9. Therefore, if DNAm perturbations are to further our mechanistic understanding in complex diseases, well-designed studies that go beyond simple correlations between methylation levels and phenotypes of interest in heterogeneous tissues must be performed.

1.6.6 Leukocyte DNA methylation and RA

The potential of modifications to the DNA methylome to shape cellular immunity during RA has been appreciated for some time. Perhaps the earliest indication that DNAm changes may be linked to RA pathogenesis was the observation that treating CD4$^+$ T cells with 5-azacytidine (5-Aza), a cytidine analogue that induces global hypo-methylation, produced self-reactive properties, as is characteristic of cells contributing to autoimmune responses in RA[215]. Subsequent to this, it was shown that T cells isolated from RA patients display alterations in their DNA methylation profile, with an overall reduction in the levels observed relative to healthy individuals[216]. Targeted approaches to interrogate methylation signatures at candidate genes led to the identification of a single CpG at the *IL6* gene promoter that was hypo-methylation in peripheral blood mononuclear cells (PBMCs) from RA patients (58% methylation) relative to healthy controls (98% methylation), and linked to increased transcript levels of this gene[217].

The development of technologies that allow genome-wide quantification of DNA methylation levels has enabled EWAS to uncover correlations between DNAm levels in various tissues/cell types and the susceptibility to disease or response to treatment (see section 1.6.5). Liu and colleagues applied such an approach to identify DMPs between RA patients and healthy controls in peripheral blood, most notably falling within the MHC cluster[218]. Importantly, this was one of the first such studies in RA to account for cell type proportions in the sample tissues, estimating cellular compositions using a reference panel and adjusting for these accordingly[218,

[219]. Interestingly, these RA-associated changes appeared to in part mediate genetic risk at this region, and such genetic-epigenetic associations will be discussed in section 1.6.9.

Though methods exists that enable the cellular composition of complex tissues such as whole blood to be estimated based on DNAm profiles[219], distinct DNAm profiles between cell subtypes mean that findings should ideally be validated in purified populations of cells. For example, a number of the associations identified by Liu et al. were successfully replicated in isolated monocytes[218]. A small study in isolated T cells and B cells from 12 RA patients and 12 controls reported 32 and 20 DMPs in these disease-relevant cell types, respectively[220]. A similar approach investigating CD4$^+$ T cell DNAm in individuals of Chinese Han ancestry reported RA hyper-methylation at 383 CpGs, with 785 sites displaying hypo-methylation in the patient cohort[221]. Using larger sample sizes to power DMP discovery, a B cell EWAS identified, and subsequently replicated in a validation cohort, disease-associated DNAm changes at 10 CpG sites[222]. With the exception of one of these sites, they all exhibited increased methylation levels in RA patients, and almost all were also associated with SLE, suggesting common pathways that may be epigenetically dysregulated in both conditions[222].

As well as potentially providing mechanistic insights into disease pathogenesis, DNA methylation may also be useful as a clinical biomarker, particularly of patient response to treatment. Differential methylation at five CpG sites in whole blood has been associated with clinical response to the TNF inhibitor etanercept[223]. Identifying such biomarkers in easily accessible patient tissue such as peripheral blood will be critical in adopting a stratified approach to treatment strategies, particularly in conditions such as RA where predictors of drug response are critically lacking.

1.6.7 Fibroblast DNA methylation and RA

Given the integral role of the FLS cells in orchestrating inflammatory processes and cartilage/bone destruction within the joint tissue itself[224], these cells have been extensively analysed in RA epigenetic studies. RA synovial fibroblasts have reduced DNAm as measured by immunohistochemistry and flow cytometry, in comparison with the same population of cells in OA patients, though this study does not report the specific joints from which the samples originate[225]. Inducing de-methylation by 5-Aza treatment in normal FLS led to these cells adopting an activated phenotype, characterised by overexpression of genes encoding effector proteins such as interleukins and matrix-degrading enzymes[225]. Targeted approaches have suggested that promoter de-methylation drives overexpression of *CXCL12* in RA FLS, with downstream effects on MMP levels[226].

A small study assessing patterns of DNAm genome-wide suggested that positions differentially methylated in RA FLS relative to those from OA patients were pervasive throughout the genome, and were able to successfully discriminate cells isolated from these two groups[227]. It was subsequently discovered that the joint location in which a cell exists can influence its DNA methylome and transcriptome, with these differing not only between RA and OA FLS, but also between RA FLS isolated from hip and knee joints[228]. Principal component analysis (PCA) revealed FLS from RA and OA patients to cluster separately, as well as those from hip and knee sites within each condition, though the number of DMPs between joints in the same condition was lower than those between RA and OA[228]. This site-dependent methylation has been further corroborated by findings that DNAm and histone modification patterns regulating the expression of HOX genes in FLS are joint-specific, and manifests as distinct cellular phenotypes in RA[229].

These studies have culminated in the most extensive epigenomic profiling in RA to date, mapping transcription, DNAm, and chromatin modifications/accessibility in FLSs[174]. This analysis identified clusters of genomic regions that displayed similar epigenetic profiles, and a number of regions were shown to be differentially modified between RA patients and OA controls[174]. The observation that local microenvironments have the capacity to shape the methylome of resident cells further reinforces the importance of studying such effects in the relevant disease stage and tissue, facilitating the distinction between changes involved in the disease pathogenesis and those that occur as a consequence of the inflammatory processes.

1.6.8 Twin studies of DNA methylation in autoimmune disease

One common pitfall of EWAS is a failure to take into account the potential effects of functional DNA variants that can influence DNAm (i.e. methylation quantitative trait loci, see section 1.6.9) - the allelic frequency at such variants may differ between comparator groups, especially for studies with small sample sizes. Such issues can be overcome by studying DNAm differences between disease-discordant monozygotic twins. A recent study which employed such a design, using 79 RA discordant twin pairs, found no CpGs to be differentially-methylated between the RA and control groups in whole blood[230]. There was however a marked increase in methylation variability at a number of positions in the RA twins (differentially variable positions; DVPs), with pathway analysis implicating disruption of stress response and binding of the RUNX3 transcription factor[230]. These findings may indicate that DNA variants drive large site-specific DNAm differences between individuals that are detected in EWAS, whereas external stimuli to which a cell is exposed, for example during early disease, may influence variability in methylation within this genetically determined range.

The hypothesis that disease-associated epigenetic modifications that manifest during complex autoimmunity may reflect underlying genetic architecture is corroborated by another study of T1D-discordant monozygotic twins. DNAm profiling of CD4$^+$ T cells, B cells, and monocytes, identified only a singular DMP in CD4$^+$ T cells and none in the other cell types[231]. Hierarchical clustering revealed that cell type and genetic variation underlie the majority of sample to sample variability, and not disease status (i.e. samples primarily cluster by cell type, and within each cell type twins cluster together)[231]. Interestingly, a large number of DVPs (10,548 B cell, 4,314 CD4$^+$ T cell, 6,508 monocyte) were identified in this study, and the affected twins were strongly enriched for hypervariable positions[231]. A separate study of PBMCs from MS-discordant monozygotic twins again showed DNA methylomes to be largely similar, though in this instance DVPs were significantly over-represented in the healthy twin [232]. These findings, however, clearly demonstrate the need to consider both genetic and non-genetic influences on the epigenome, and this will be discussed in the subsequent section.

To conclude, many lines of evidence suggest a central role for DNAm in the pathogenesis of RA, although the extent to which this is driven by genetic variation or environmental exposures remains ambiguous. Given the advancements in characterizing the genetic architecture of RA, the challenge is now to integrate DNAm with additional omics datasets in relevant cohorts to understand the precise contribution to disease processes beyond simple associations. For example, through combining cell-specific epigenetic and transcriptomic data, and considering the findings in the context of genetic risk factors, more comprehensive mechanistic understanding of the precise role of molecular traits in mediating pathophysiological processes can be achieved.

1.6.9 Methylation Quantitative Trait Loci

Given that changes in DNAm can occur upon cells being subject to extrinsic exposures (most notably cigarette smoke[233-236]), it is enticing to consider such modifications as mediators of environmental risk in complex disease. Additionally, DNAm is important in conferring plasticity of cellular responses. For example, during differentiation of monocytes to dendritic cells, STAT6 confers TET2-dependent demethylation and gene expression downstream of IL-4 signalling[237]. However, as has been referred to in the previous section, DNAm can also be influenced by variants in the DNA sequence, representing an interface between intrinsic genetic risk and external stimuli to which cells are exposed.

Genomic loci that are associated with DNAm levels in cis – termed cis-methylation quantitative trait loci (cis-meQTLs) – have been described genome-wide in a range of tissues and cell types including foetal and adult brain[238-241], adipose tissue[242, 243], pancreatic islets[244], whole blood[245-248], fibroblasts[126, 249], CD4$^+$ T cells[127, 250], monocytes[127], neutrophils[127], and lymphoblastic cell lines (LCLs)[126, 251], amongst others. As with eQTLs, these meQTLs can be mapped by quantifying DNAm genome-wide and seeking associations between allele copy number, for example at disease-associated SNPs, and DNAm levels (Figure 1.7). Early efforts to decipher the causal relationship between genetic variation and gene expression indicated that this likely occurs in several instances via the effects of the former on DNA methylation[252]. Indeed, meQTL variants are often found to co-localise with eQTLs, suggestive of a common underlying mechanism[126, 127, 238-247, 249, 251, 253]. At a given risk locus, a regulatory SNP may be associated with the magnitude of expression at proximal transcripts (cis-eQTL), and of methylation at one or more CpG sites (cis-meQTL) (Figure 1.7).

Seeking associations between DNAm and transcript levels (expression quantitative trait methylation (eQTM)) downstream of QTL effects can further implicate DNAm as a potential mediator of transcriptional regulation at a given locus (Figure 1.7). In whole blood, expression levels at 90% of genes were shown to be associate with an meQTL CpG site in cis, though only 23% of all identified cis-regulated DNAm sites themselves exhibited such a relationship with transcription[245]. Indeed, though co-localisation of eQTLs with meQTLs is frequently observed, the former displays a greater extent of tissue specific activity than the latter, such as has been shown across three immune cells (neutrophils, monocytes, and CD4$^+$ T cells)[127]. Under the assumption that DNAm is a mediator of transcriptional regulation, this signifies that identification of meQTLs is a more proximal molecular trait to genetic variation than is transcript levels, which would be consistent with this greater degree of tissue non-specificity. This would indicate that not all meQTL activity has downstream effects on gene expression in a given tissue, confirming the need to incorporate transcriptomic data in epigenetic studies.

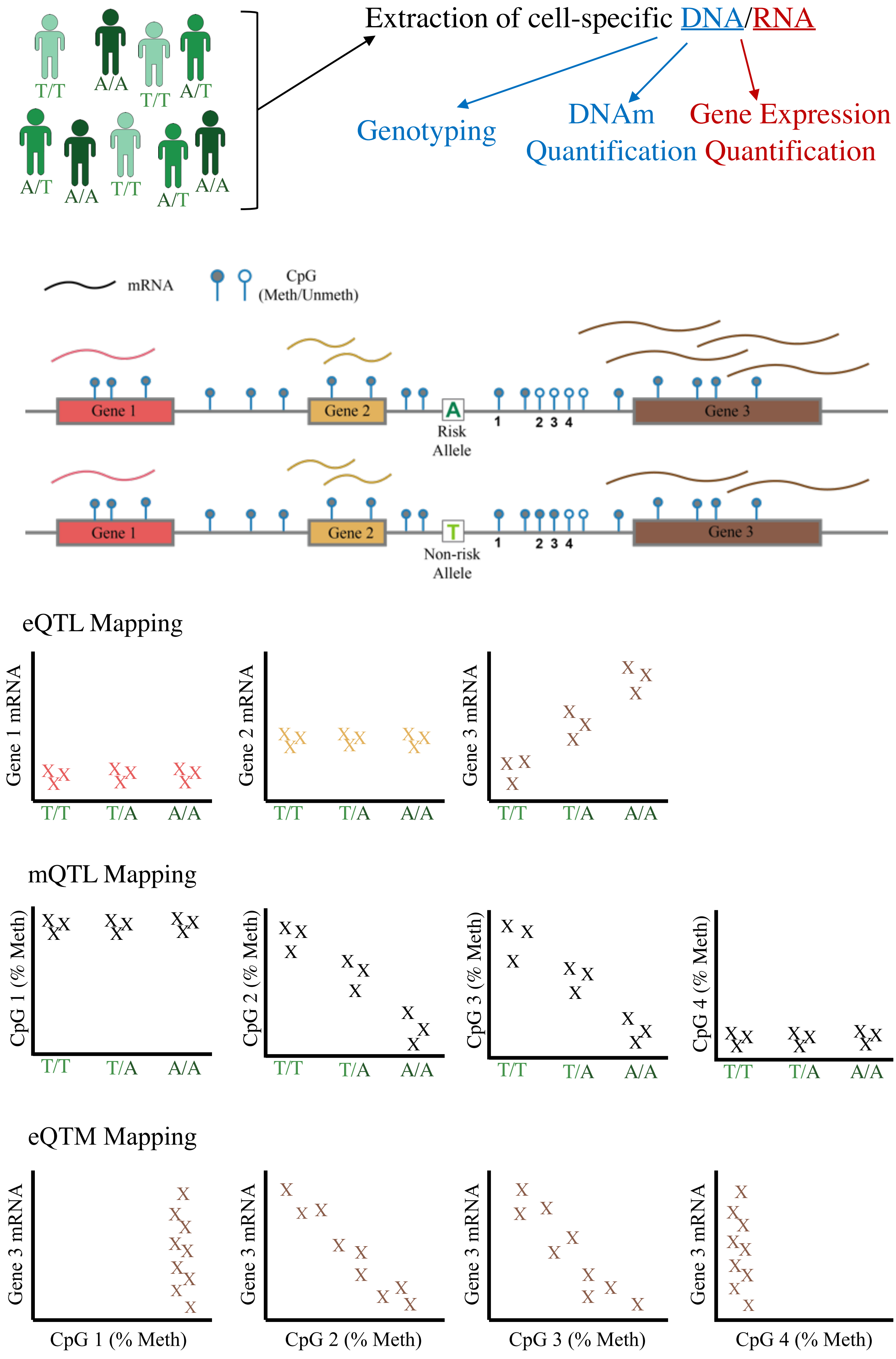

T/T
A/A
T/T
A/T
A/T
A/A
T/T
A/T
A/A
Extraction of cell-specific DNA/RNA
Genotyping
DNAm Quantification
Gene Expression Quantification
mRNA
CpG (Meth/Unmeth)
Gene 1
Gene 2
A
Risk Allele
1
2 3 4
Gene 3
Gene 1
Gene 2
T
Non-risk Allele
1
2 3 4
Gene 3
eQTL Mapping
Gene 1 mRNA
T/T T/A A/A
Gene 2 mRNA
T/T T/A A/A
Gene 3 mRNA
T/T T/A A/A
mQTL Mapping
CpG 1 (% Meth)
T/T T/A A/A
CpG 2 (% Meth)
T/T T/A A/A
CpG 3 (% Meth)
T/T T/A A/A
CpG 4 (% Meth)
T/T T/A A/A
eQTM Mapping
Gene 3 mRNA
CpG 1 (% Meth)
Gene 3 mRNA
CpG 2 (% Meth)
Gene 3 mRNA
CpG 3 (% Meth)
Gene 3 mRNA
CpG 4 (% Meth)

Figure 1.7 (previous page): Mapping quantitative traits to infer mechanisms of cell-mediated pathogenesis. Following the isolation of cell-type specific DNA and RNA, patients are genotyped, and DNA methylation/gene expression data collected, for example using microarray platforms. By seeking linear associations between allele copy number and DNA methylation (methylation quantitative trait locus; meQTL) or gene expression (expression quantitative trait locus; eQTL), regulatory functions of single nucleotide polymorphisms can be unravelled. Following identification of QTL effects, associations between DNA methylation and gene expression (expression quantitative trait methylation; eQTM) indicate either co-regulation of these two molecular signatures, or a causal effect of one on the other downstream of genetic modulation. Though in this example we show associations at a single locus, this can be performed in parallel throughout the genome.

In instances of co-localised meQTL and eQTLs with an observed eQTM effect, a number of possible regulatory mechanisms could explain the co-localised signals (Figure 1.8). It may be that the SNP influences DNAm at a CpG site in cis, which subsequently modulates transcriptional activity (SNP – Methylation – Expression (SME); Figure 1.8A). Alternatively, reverse mediation may occur whereby changes in DNAm arise downstream of transcriptional regulation (SNP – Expression – Methylation (SEM), Fig. 1.8B). Finally, the regulatory SNP may directly influence DNAm and transcription independently, with no effect of DNAm and transcript levels (INDEP, Figure 1.8C).

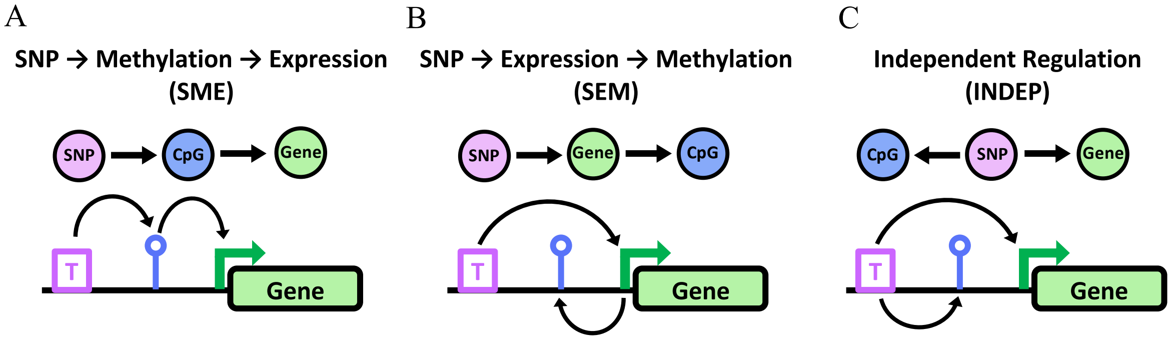


Figure 1.8: Distinct regulatory models that can explain associations between SNP genotype, CpG DNA methylation, and gene expression observed in cross-sectional data. (A) Under the SNP → Methylation → Expression (SME) model, DNA methylation at the CpG site is directly influence by the allele present at the regulatory SNP, which in turn alter expression of the target gene. (B) The SNP → Expression → Methylation (SEM) model, also be referred to as reverse causation, occurs when the SNP directly regulates gene transcription, which in turn influences CpG methylation. (C) The independent (INDEP) model describes a scenario in which DNA methylation has no effect on gene expression, but rather the allelic effect on DNA methylation and expression represent independent associations.

In human hippocampal tissue, regulatory eQTL SNPs (eSNPs) are 6.7-fold enriched for those which function as meQTLs relative those displaying no such regulation of DNAm[241]. In LCLs, 25% of eQTLs also display association with DNAm, which represented significant enrichment for meQTL variants relative to what would be expected by chance[251]. Analysis of eQTLs and meQTLs in whole blood with subsequent testing for co-localization of regulatory variants has revealed that of 6526 independent eSNPs, 5192 (~80%) were also associated with methylation levels at one or more CpG sites in cis, with 54% of these designated as likely sharing a causal regulatory variant[247]. Importantly, 24% of the co-localised molecular trait loci in this study show some evidence of molecular mediation (either SME or SEM), with a Bayesian network analysis revealing that in approximately 90% of cases the SME model was most likely[247]. This SME model has also been favoured as the most likely mechanism of regulation in T cells, whereas in fibroblasts and LCLs the INDEP model was postulated to be more likely[186].

One frequent observation regarding the genomic mapping of meQTL variants is that they are depleted within CGIs (see section 1.6.4) and over represented in intergenic regions [186, 238, 242, 244, 245]. This may indicate that methylation patterns in promoter-associated CGIs are more evolutionary conserved, whereas more distant regulatory elements such as enhancers display more variable, genetically encoded methylation levels. That binding sites of the transcriptional repressor CCCTC-binding factor (CTCF) and other TFs are enriched at meQTLs suggests that binding of trans-acting regulatory proteins may drive DNAm modifications in a tissue-specific manner[254]. Indeed, DNA variants that modify TFBS were statistically enriched amongst those that modulate DNAm at proximal CpG sites[251]. The same study showed that meQTL SNPs are also frequently associated with additional molecular phenotypes, including histone modifications and gene expression, demonstrating that these sequence-dependent regulatory events are highly coordinated[251]. Though many TFs display methylation-sensitive DNA binding activity, it may also be the case that TF binding at many loci prevents DNAm from occurring at that locus, as has been described for SP1/SP3 binding sites[255], indicating that this relationship can be bi-directional.

1.6.10 Methylation quantitative trait locus analysis informs mechanisms of disease risk

Transferring insight from me-/eQTL mapping into studies of disease genetics may help us to delineate the mechanisms through which non-coding variants can contribute to disease pathogenesis. As with eQTL variants, mapping SNPs that influence DNAm provides additional information for prioritizing GWAS variants for downstream functional analysis. Indeed, meQTL SNPs are enriched for GWAS variants conferring susceptibility to common diseases[248, 251, 256]. As such, these variants have been shown to explain a proportion of the heritability of complex diseases, including autoimmune conditions[257]. Using relevant cell types and tissues for seeking co-localization of meQTL and GWAS variants will facilitate the discovery of disease pathways. For example, both meQTL and eQTL SNPs in hippocampal tissue were enriched for schizophrenia-associated GWAS SNPs, providing a mechanistic link with complex disease in a relevant tissue type[241]. In addition, mapping such regulatory effects in lung tissue has revealed putative mechanisms underlying susceptibility to chronic obstructive pulmonary disease[253]. Consistent with these findings, SNPs associated with autoimmune conditions from GWAS studies are overrepresented in differentially methylated regions DMRs that are specifically hypo-methylated in B and T cells relative to other tissue/cell types[258].

A recent study that integrated skeletal muscle molecular data from biopsies of a large cohort found multiple instances of DNAm mediating the genetic risk of traits including height, weight, and type II diabetes through its effect on gene expression[259]. Moreover, early studies of DNA methylomes from RA patients and controls identified patterns of differential methylation that were in part mediated by genetic variants[218]. The RA EWAS in B cells described in section 1.6.5 highlighted that a number of the disease-associated CpG sites may be regulated by SNPs acting in *cis*[222].

In addition to the tissue- and cell-specific nature of DNAm, the highly dynamic nature of this epigenetic modification necessitates the need to consider cellular context, particularly in studies of disease. Further to the joint-specific DNAm patterns identified in synovial fibroblasts[228, 229], there is also some evidence that in autoimmunity, meQTL activity at certain GWAS loci may be specific to patients and absent in healthy control subjects[248]. The appreciation that DNAm patterns can be in part dictated by DNA polymorphisms in a cell-specific and context-specific manner highlights an interesting link between genetic and epigenetic risk factors in RA. Indeed, heritable patterns of methylation in CD4$^+$ T cells, a cell type frequently studied in relation to its role in complex immune-mediated diseases, were largely attributable to regulatory variants in *cis*, with ~75% of heritable CpGs in this cell type being associated with cis-meQTLs[250]. Mapping meQTLs will help not only to decipher functional consequence of GWAS SNPs, but also to interpret EWAS studies that seek to associate epigenetic profiles with disease susceptibility and outcomes.

As correlation between molecular signatures does not inform conclusions regarding causation, numerous statistical approaches can be applied that allow chains of causality to be statistically inferred from observational data. One such method, causal inference testing (CIT), relies on a series of conditional correlations to test the probability that a molecular mediator (i.e. DNAm) influences a phenotype of interest (e.g. gene expression or disease status) downstream of a genetic variant[260]. This test will be discussed in greater detail in Chapter 5 relating to its application for inferring chains of mediation in molecular data collected from the same individuals. Another approach used frequently in epidemiological studies is Mendelian randomisation (MR), which relies on treating genetic variants that have been robustly associated with the trait of interest (i.e. from GWAS) as instrumental variables. This is useful because, unlike measuring other highly-confounded environmental exposures, genetic factors are (excluding somatic mutations) stable from birth across the life course, and each individual is effectively assigned a random genotype at birth (analogous to a randomised control trial)[261]. Trait-associated genetic variants can also be used to anchor the association, given that it is all but certain they are causal for the phenotype of interest and not vice versa (i.e. the measured exposure is not causing site-specific mutations in patients)[261]. The instrumental variable (genetic variant) should thus not be associated with the phenotype of interest (e.g. disease) except through its effect on the measured exposure (e.g. gene expression). Though causality cannot be proved from such cross-sectional patient studies, and ultimately needs to be demonstrated in vitro, these methods are useful for highlighting interesting loci with a high probability of genetic causality through an intermediate.

1.7 Aims & Objectives

1.7.1 Aims

Comprehensively map the lymphocyte DNA methylation landscape in early arthritis, and investigate for the first time the role of DNA methylation as a mediator of RA genetic risk in CD4$^+$ T cells and B cells from an early arthritis cohort.

1.7.2 Specific Objectives

- Assess global differences in CD4$^+$ T cell and B cell DNA methylation between RA patients and those who have other arthritis diagnoses (Chapter 3).
- Integrate paired patient genotype data and perform a methylation quantitative trait locus (meQTL) analysis, generating a comprehensive map of genetic-epigenetic interactions both in *cis* and in *trans* in a highly relevant patient cohort (Chapter 4).
- Leverage GWAS data to assess the capacity of known RA-associated variants to influence lymphocyte DNA methylation. This analysis will also be extended beyond RA risk loci to investigate potential shared pathways of genetic risk in immune-mediated diseases (Chapter 4).
- Assess potential functional implications of genetically encoded DNA methylation variability by integrating publicly available sources of cell-specific chromatin state data as well as transcription factor binding sites. (Chapter 4).

- Incorporate transcriptomic data and employ statistical approaches for inferring molecular causality, identifying loci and genes for which DNA methylation likely mediates RA genetic risk at the transcriptomic level, as well as for immune-mediated disease more generally (Chapter 5).
- Apply *in vitro* techniques to validate associations and regulatory mechanisms at interesting RA loci identified in the above analyses (Chapter 5).

Chapter 2 – Materials & Methods

A note on contributors to laboratory work:

The work described in sections 2.1-2.4 & section 2.8 (excluding 2.8.1) was carried out prior to commencement of the project by colleagues at the Newcastle University musculoskeletal research group (MRG), rheumatologists and research nurses at the Freeman hospital, and collaborators at the University of Manchester Arthritis Research UK Centre for Genetics and Genomics. I am grateful to the following for their contributions to laboratory work that has generated this molecular dataset and made this project possible:

- $CD4^+$ T cell isolation: Dr Amy Anderson, Julie Diboll.
- $CD19^+$ B cell isolation: Julie Diboll, Dr Nishanthi Thalayasingam,
- DNA/RNA extraction from isolated lymphocytes: Dr Amy Anderson, Dr Phil Brown, Dr Arthur Pratt, Dr Nishanthi Thalayasingam.
- RNA processing and transcriptomic arrays: Dr Nisha Nair, Dr Jonathan Massey.
- Genotyping arrays and imputation: Dr Nisha Nair, Dr Jonathan Massey.
- Sample processing and DNAm arrays: Dr Nisha Nair, Dr Jonathan Massey.

I was responsible for all data analysis and experimental work described from section 2.5-2.7 and 2.9-2.15. Cell sorting as described in section 2.15.4 was carried out with assistance from the Newcastle University Flow Cytometry Core Facility.

2.1 Patient Recruitment

Consenting patients were recruited from the Newcastle Early Arthritis Clinic (NEAC) prior to the commencement of any immunomodulatory therapy. Patients received a working diagnosis by a rheumatologist at inception, with all diagnoses confirmed at follow up > 1 year subsequent to the first visit. RA was diagnosed with reference to the 2010 ACR/EULAR criteria[64]. For this study, samples were collected from a cohort of RA patients (both ACPA+ and ACPA-), together with a disease control group comprising patients receiving an alternative diagnosis. This non-RA group included those with spondyloarthropathies (SpA: psoriatic arthritis, enteropathic arthritis, reactive arthritis, and undifferentiated spondyloarthritis), OA and other non-inflammatory causes of arthralgia. Whilst these are all conditions that manifest as joint symptoms, the aetiologies and triggers are distinct from those in RA. For example, psoriatic arthritis (PsA) is a condition that affects up to 30% of patients who suffer psoriasis – an immune-mediated disease affecting the skin[262]. This condition is similar to RA in that it is characterised by synovial inflammation and immune cells infiltrating the joint, with genetic associations such as those in HLA region indicative of an immune-mediated disease[263]. However, a number of distinct features of clinical presentation and pathogenesis of PsA, such as the absence of the specific circulating autoantibodies that characterise a major subgroup of RA patients, indicate differing disease processes. Similarly, OA presents as pain and stiffness accompanied by loss of function, principally in the joints of the hand, knee, and hip. However, to a much greater extent than RA and SpA, the condition is believed to develop as a result of factors such as biomechanical stress on the joints and defects in chondrocyte function that lead to dysregulated cartilage architecture, and a role for the immune system in mediating pathogenesis is considered to be much less pronounced[46, 264]. Selecting patients diagnosed with conditions for which the clinical manifestations broadly reflect those seen in RA patients, but are at least to some degree aetiologically distinct, represents a suitable control group for studying the pathogenesis of RA. The control population, hereon in referred to as the 'non-RA' group, thereby included both patients with primarily immune-mediated, and non-immune-mediated joint pathology.

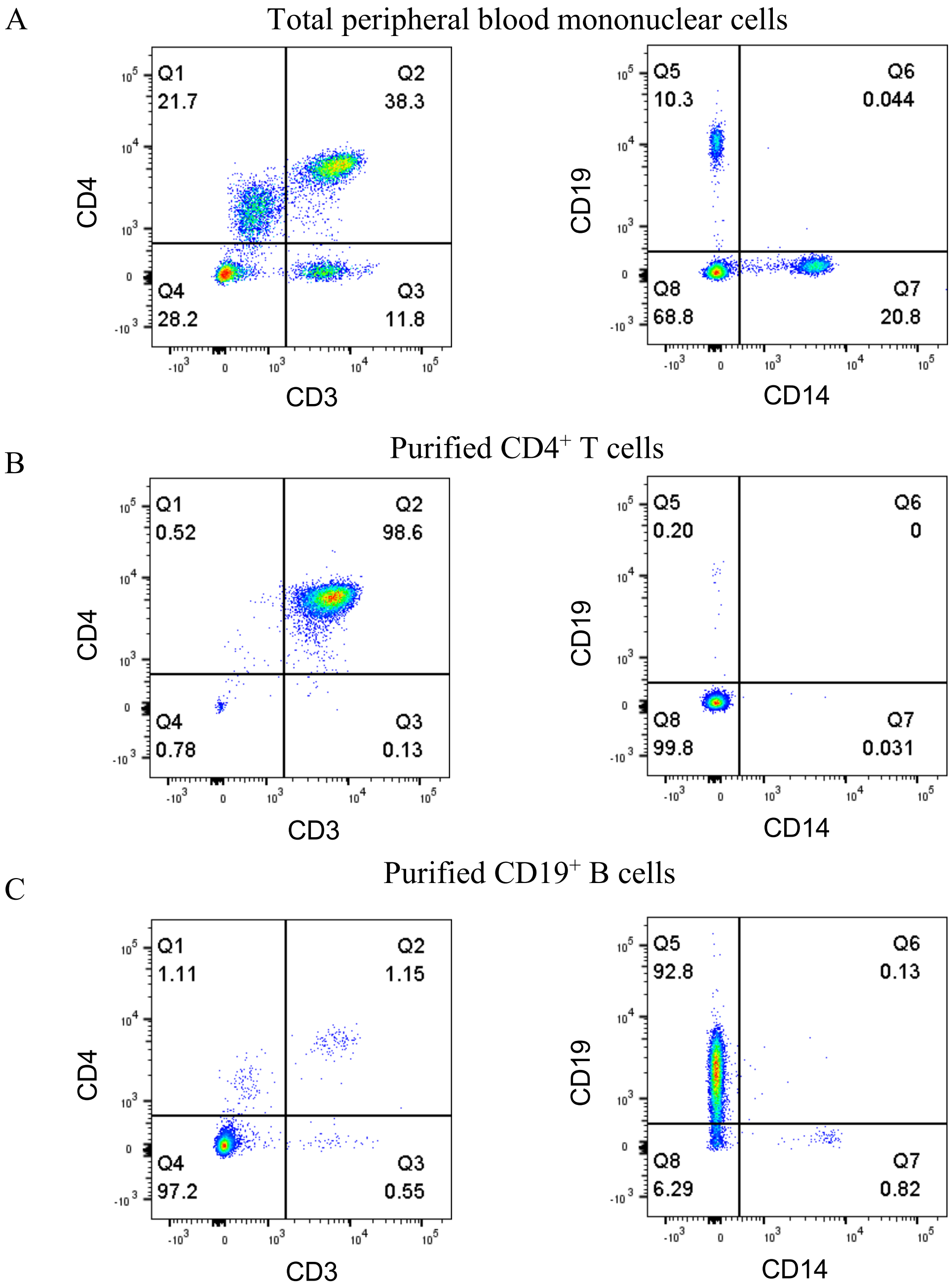
Section 3.2 describes the demographic and clinical characteristics of these cohorts in greater detail (see Table 3.1). Ethical approval for this study was granted by the North Tyneside Regional Ethics Committee, with patients giving written informed consent prior to enrolment (REC: 12/NE/0251).

2.2 Lymphocyte isolation from peripheral blood and DNA/RNA extraction

Peripheral blood was collected from consenting patients attending the NEAC and stored for up to four hours at room temperature prior to processing. CD4$^+$ T cells and CD19$^+$ B cells were isolated from peripheral blood by positive selection using a magnetic bead-based approach. This involves first labelling specific cell subsets using marker-specific antibodies complexed to magnetic beads. Labelled cells are then isolated as they pass through a magnetic column.

For CD4$^+$ T cell isolation, monocytes were initially depleted using RosetteSep™ Human Monocyte Depletion Cocktail (Stemcell Technologies, Cambridge, UK; Cat# 15668), followed by the addition of HetaSep™ solution (Stemcell Technologies, Cat# 07906) and density centrifugation at 50 x g for 5 minutes. CD4$^+$ T cells were then positively isolated from the monocyte-depleted supernatant using the Robosep™ (Stemcell technologies, Cat# 20000) together with the EasySep™ Human Whole Blood CD4 Positive Selection Kit (Stemcell Technologies, Cat# 18082A).

B cells were purified by magnetic-activated cell sorting (MACS) positive selection of cells expressing CD19 – a transmembrane glycoprotein expressed by B cells at all developmental stages, and follicular dendritic cells. Peripheral blood mononuclear cells (PBMCs) were first isolated from patient whole blood by density centrifugation on Lymphoprep medium (Axis-Shield Diagnostics, Dundee, UK; Cat# 07801). Subsequently, B cells were isolated using MACS human CD19 MicroBeads (Miltenyi Biotech, Bergisch Gladbach, Germany; Cat# 130-050-301). After labelling CD19$^+$ cells, magnetic isolation was performed using a LS positive selection column (Miltenyi Biotech; Cat# 130-042-401) with MidiMACS™ separator (Miltenyi Biotech; Cat# 130-042-302). The purity of isolated CD4$^+$ T cell and B cell populations was confirmed by flow cytometry as has been described[139]. Exemplar plots showing proportions of CD4$^+$ T cells and CD19$^+$ B cells in total PBMCs, purified CD4$^+$ T cells, and purified CD19$^+$ B cells from purity checks are shown in Figure 2.1.


A Total peripheral blood mononuclear cells

B Purified CD4$^+$ T cells

C Purified CD19$^+$ B cells

Figure 2.1 (previous page): Flow cytometry purity checks of isolated cell populations. Total proportions of $CD4^+$ T cells (left panels, upper right quadrant; $CD3^+$ $CD4^+$) and B cells (right panels, upper left quadrant; $CD14^-$ $CD19^+$) in (A) peripheral blood mononuclear cells, (B) positively selected $CD4^+$ T cells, and (C) positively selected $CD19^+$ B cells from the same patient. Numbers represent the percentage of all cell events falling within a given quadrant.

Following isolation, cells were lysed by the addition of RLT Plus Buffer (Qiagen, Hilden, Germany) supplemented with 1% v/v β-mercaptoethanol (Sigma-Aldrich, Dorset, UK; Cat# M7154), and passing through the QIAshredder spin column (Qiagen). DNA and RNA were extracted from each sample using the AllPrep DNA/RNA Mini Kit according to the manufacturer's instructions (Qiagen; Cat# 80004).

2.3 Patient genotyping and quality control

Genotyping of patient DNA was performed using the Human CoreExome-24 BeadChip array (Illumina, Cambridge, UK; Cat# 20015246), according to the manufacturer's instructions. In brief, patient genomic DNA was amplified by polymerase chain reaction (PCR), enzymatically fragmented to produce fragments of length 300-600 base pairs, isopropanol precipitated (followed by re-suspension), and hybridised to the BeadChip array. The following day, washing and single-base extension with labelled dideoxynucleotides were performed, with subsequent fluorescent staining of labelled probes. The arrays were then imaged using the iScan System - a laser-based system for detecting fluorescent signals (Illumina).

Quality control (QC) procedures were carried out using the GenomeStudio software (Illumina). Initially samples or SNPs for which the SNP call rate fell below 98% were removed. Subsequently, the cluster separation score was generated to determine how well the three genotypes (major allele homozygote, heterozygote, and minor allele homozygote) for a given SNP separate on a cluster plot, and SNPs for which this metric was <0.4 were excluded. Finally, SNPs with a low minor allele frequency (MAF, < 0.01) in the data were removed. PLINK software (version 1.9) was used to determine potential relatedness between individuals, and in the case of samples with a proportion identity by descent score > 0.2, the sample with the lowest genotype call rate was excluded. This ensures that all individuals included have no greater than a third degree of relatedness. All patients self-reported as being of Northern European descent, and identity-by-state clustering in PLINK revealed no population stratification.

Following genotype QC, haplotype phasing was performed using SHAPEIT2 software[265], and then imputated to the 1000 Genomes reference panel (Phase 3) with IMPUTE2[266]. IMPUTE2 generates an INFO score that describes the confidence with which an imputated genotype is called, and any SNPs with an INFO score less than 0.8 were not included in downstream analyses.

2.4 DNA methylation profiling

Cell-specific DNA was quantified using the Nanodrop ND-1000 Spectrophotometer (ThermoFisher, MA, USA), and 400ng bisulphite-converted using the EZ-96 DNA Methylation Kit (Zymo, Orange, CA; Cat# D5004). DNAm quantification of bisulphite-converted DNA was carried out on the Infinium MethylationEPIC BeadChip Kit following the manufacturer's instructions (Illumina; Cat# WG-317-1003; see section 1.6.5 for further detail on the array). Following hybridisation, washing, and extension steps, the BeadChip arrays were imaged on the iScan system (Illumina), and intensity data extracted in the format of IDAT files using the GenomeStudio Methylation module v1.8 software (Illumina).

2.5 DNA methylation data pre-processing and quality control (QC)

Pre-processing and statistical analysis of DNAm data was performed in the R programming language (version 3.5). Data from CD4$^+$ T cells and B cells was processed and analysed separately. This section will describe a series of sample-level QC checks that were carried out in order that potential mix-ups could be identified, and problematic samples either removed or repeated. A systematic analysis of normalisation methods and an algorithm for detecting residual batch effects will also be introduced, the results of which are presented in Chapter 3.

2.5.1 Sample QC

The minfi package (version 1.28)[267] was used to read in IDAT files exported from the iScan system (section 2.4), and calculate detection p-values to enable the identification of failed probes. Probes with a detection p-value > 0.01 in > 10% of samples were deemed to have failed (intensity signals not significantly different from the negative control background level) and were therefore removed. Initially, though probes which failed in > 10% of samples had been removed, a sample-level probe filter was also applied to remove any samples for which the mean detection p-value across all probes remained at > 0.01 following filtering of failed probes. Using these criteria, four CD4$^+$ T cell samples were excluded (Figure 2.2A), as was one B cell sample (Figure 2.2B).

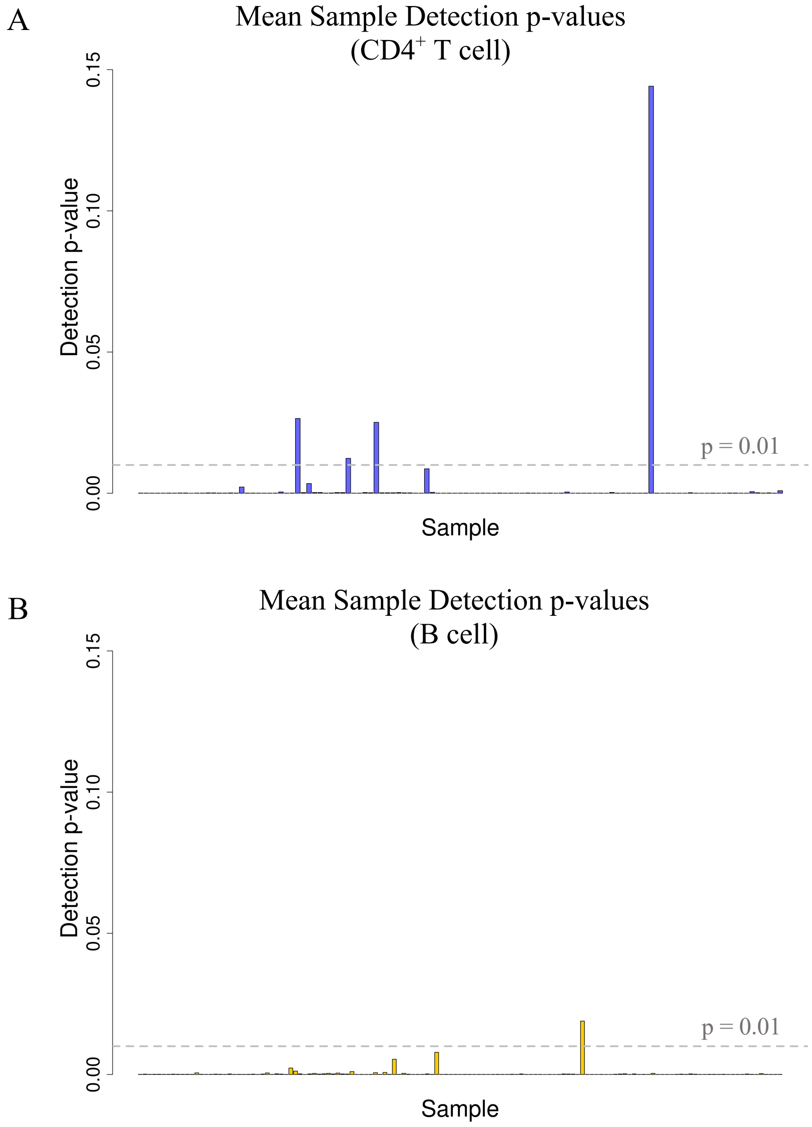


Figure 2.2: Mean DNA methylation probe detection p-values across all samples. Detection p-values based on negative control probe intensity background signal were calculated at (A) 115 CD4$^+$ T cell samples and (B) 136 B cell samples. Following the removal of failed probes, samples for which the mean probe detection p-value was > 0.01 were removed from the analysis. This resulted in the exclusion of four CD4$^+$ T cell samples, and one B cell sample.

2.5.2 Confirming patient identification using SNP probes

The IDAT files extracted following scanning of the MethylationEPIC array contain intensity values for methylated (M) and unmethylated (U) probes at each CpG. These intensities can be converted into a Beta (β)-value that represents the ratio of the methylated signal to the total signal on a scale of 0 (unmethylated) to 1 (methylated):

$$\beta = \frac{\mathrm{M}}{\mathrm{M+U}}$$

To confirm the identity of each sample, the presence of 65 SNP probes on the MethylationEPIC array was exploited. Rather than interrogating methylation levels at genomic CpGs, the probes map to common SNPs and as such β-values at these probes acts as a proxy for the sample genotype. By leveraging SNP probes to identify sample genotype at these SNPs, it is possible to perform sample tracking where genotype data are also available. β-values for these SNP probes were obtained for all samples using the *getSnpBeta* function in minfi. As genotype data were also available for the majority of these patients (95% CD4$^+$ T cell, 92% B cell samples), these β-values could then be mapped back to patient genotype data to confirm that genotype and DNAm samples originate from the same individual. When plotting these SNP β-values against the corresponding SNP genotype, homozygous SNPs should display a bimodal DNAm distribution, with heterozygous SNPs having an intermediate value, as was the case for the CD4$^+$ T cell sample from patient EA1019 (Figure 2.3; left panel). If a sample swap had occurred, the plotted data would deviate from this pattern, as is evident for the B cell data annotated to the same patient (Figure 2.3; right panel).

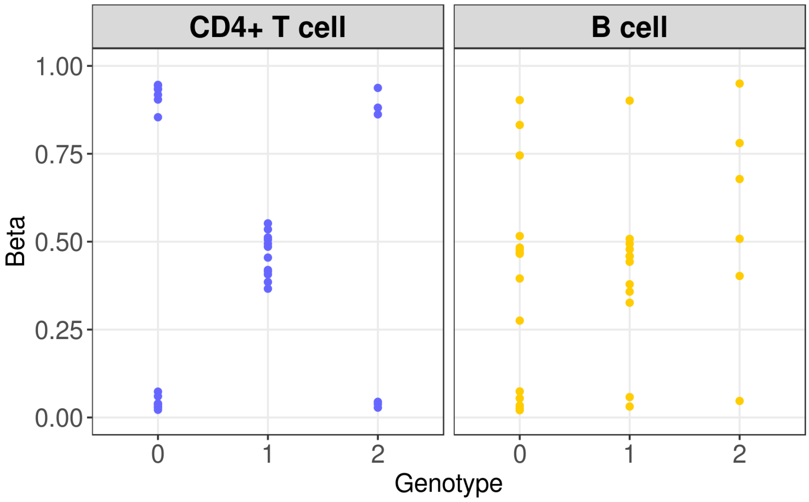


Figure 2.3: Plotting SNP β-values against patient genotype to identify mismatches. Early arthritis patient EA1019 was confirmed as a correct annotation of the CD4$^+$ T cell sample (left panel), but a sample misannotation for the B cell sample (right panel). When plotting SNP β-values against SNP genotypes across all SNPs for a given sample, homozygous SNPs (0 or 2) should display a bimodal β-value distribution, whereas heterozygous SNPs (1) should have intermediate values, as is seen for the CD4$^+$ T cell sample. Deviations from this indicate sample mix-ups during processing or loading the array, as was the case for the B cell sample in this instance.

2.5.3 Estimation of cell type composition

Though cell populations had been isolated by positive bead-based selection, and purity checked by flow cytometry, such checks had not been performed for all samples due to low cell numbers, particularly B cells (no purity check for 11% of CD4$^+$ T cell samples, 36% of B cells samples).

Given that patterns of DNAm are cell-type specific, a complementary approach was therefore employed to assess patterns in the DNAm data itself that may indicate whether a sample mix-up with regards to cell type has occurred or instances of considerable contamination.

For this purpose, data from CD4$^+$ T cell and B cell samples were initially pre-processed and analysed together. In the first instance, PCA was performed for all samples using the *prcomp* function in R (base R version 3.5.3), with the first two principal components visualised using ggplot2 (version 3.2.0) and samples coloured according to their annotated cell type (Figure 2.4). This revealed that samples exhibited marked clustering according to cell type, with principal component 1 (PC1) distinguishing CD4$^+$ T cells and B cells. However, a number of potential incorrect annotations were identified, with some samples appearing to cluster falsely according to the cell type (Figure 2.4).

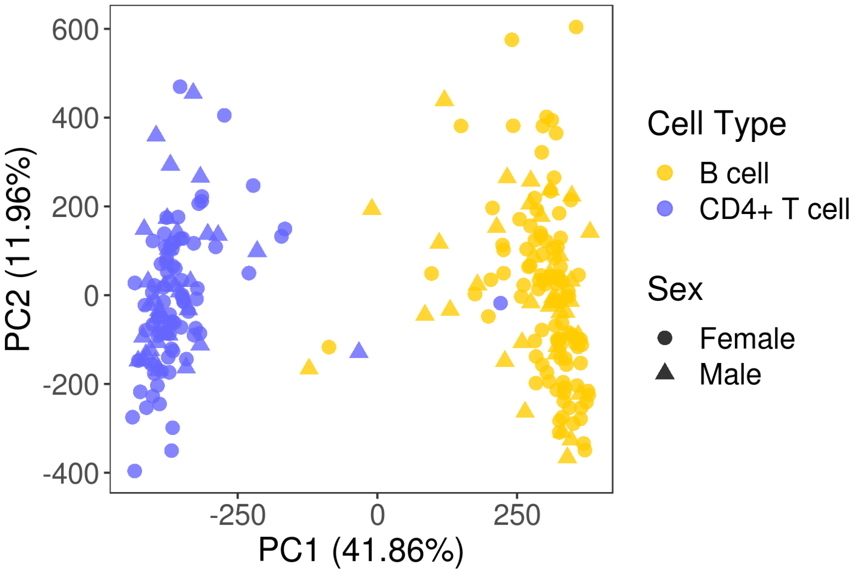


Figure 2.4: Principal component analysis (PCA) of all DNA methylation samples by cell type. A small number of samples appeared to fall within the wrong cluster, indicating either a sample mix-up, or a low purity sample.

To interrogate further, the proportion of each blood cell type in the data was estimated using the Houseman method[219]. This is a reference-based method that utilises methylation signatures from flow-sorted cells to infer proportions of these cell types in whole blood samples[219]. This method was originally developed to quantify leukocyte proportions in whole blood for interpreting EWASs in this complex tissue.

The *estimateCellCounts* function in minfi was used to apply this method to CD4$^+$ T cell and B cell datasets. This function returns an estimate of the total proportion of CD8$^+$ T cells, CD4$^+$ T cells, natural killer (NK) cells, B cells, monocytes, and granulocytes in each sample (values for CD4$^+$ T cell and B cell estimates are given for each relevant sample in Appendix A). These data were plotted, and any samples for which the estimated proportion of the expected cell type fell below 0.65, or that of the alternate cell type was above 0.35, were deemed to be either low-purity or sample mix-ups and as such removed (Figure 2.5; these corresponded to the samples clustering incorrectly in Figure 2.4). Using this approach, two potential CD4$^+$ T cell sample mix-ups were identified (EA0835 & EA2028) which reported higher proportions of B cells than CD4$^+$ T cells (Figure 2.5A). Likewise, in the B cell dataset, five samples (EA1050, EA1067, EA2042, EA2067, EA2075) displayed low (<0.65) B cell proportions and/or high estimates (>0.35) of CD4$^+$ T cells (Figure 2.5B), and were similarly removed.

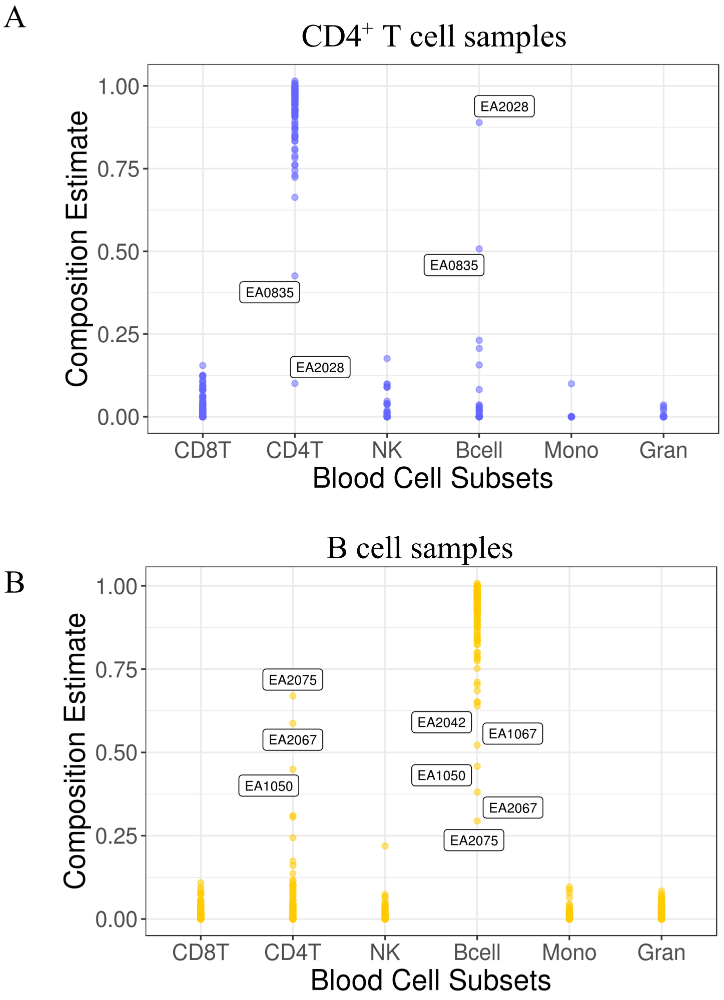


Figure 2.5: Estimation of cell type proportion. The Houseman method was employed to estimate proportions of leukocytes in (A) 111 CD4$^+$ T cell and (B) 135 B cell samples estimated using the reference-based Houseman method for whole blood[216]. Each data point represents a single sample. Early arthritis sample identifiers are given for those samples for which the estimated proportion of the relevant cell type is < 0.65, and where the proportion of the alternate cell type is > 0.35. CD8T = CD8$^+$ T cell; CD4T = CD4$^+$ T cell; NK = Natural killer cell; Bcell = B cell; Mono = Monocytes; Gran = Granulocytes.

The median cell purity based on flow cytometry data was >98% for CD4$^+$ T cell samples (range 79.6 – 99.6%), and >94% for B cell samples (range 60.0 – 98.6%). Though the majority of samples had high purity of >90% as determined by flow cytometry, there were a small number of samples deemed to have low purity (< 90%; one CD4$^+$ T cell sample, six B cell samples), with one particular B cell sample having a purity of 60% (Appendix A). However, given that these samples did not appear as outliers following principal component analysis, and were not deemed to have low purity values (< 80%) based on the cell type proportion estimates, these were not excluded. In addition, the method employed to estimate batch effects in the data (section 2.5.7) will also capture any sample to sample variability associated with differences in cell purity, and as such these can be accounted for in subsequent analyses.

2.5.4 Sex prediction

Signal intensity values from probes mapping to the X and Y chromosome were utilised in order to assign sex to each sample, and map these back to the patient sex in the sample annotation file, again adding an additional level confidence in sample annotations. The median copy numbers (CNs), defined as the median probe intensity (methylated and unmethylated probes) for all probes, were calculated separately for the X and Y chromosomes. The median CN for Y chromosome probes was subtracted from those for X chromosome probes for all samples, with the resulting values plotted for CD4$^+$ T cells (Figure 2.6A) and B cells (Figure 2.6B). For male samples, it would be expected that the median CN (Y) - median CN (X) would be approximately zero, as signals are present from both chromosomes. Conversely, female samples should return a value below zero, given that no Y chromosome intensities are returned for these samples, and the median CN is therefore zero (subtracting the X chromosome median CN from zero will always yield a negative value). A difference in median copy number of -0.2 was used as a threshold for assigning a sample as female (Figure 2.6A-B).

To further corroborate these results from X and Y chromosome intensities, the X chromosome β-values were plotted for CD4$^+$ T cell (Figure 2.6C) and B cell (Figure 2.6D) samples. The difference in the number of X chromosomes copies between female and male samples will result in distinct density plots, with male samples displaying a bimodal distribution (a position can either be methylated (1) or unmethylated (0), whereas the intermediate values from female samples represent the average across both chromosome copies.

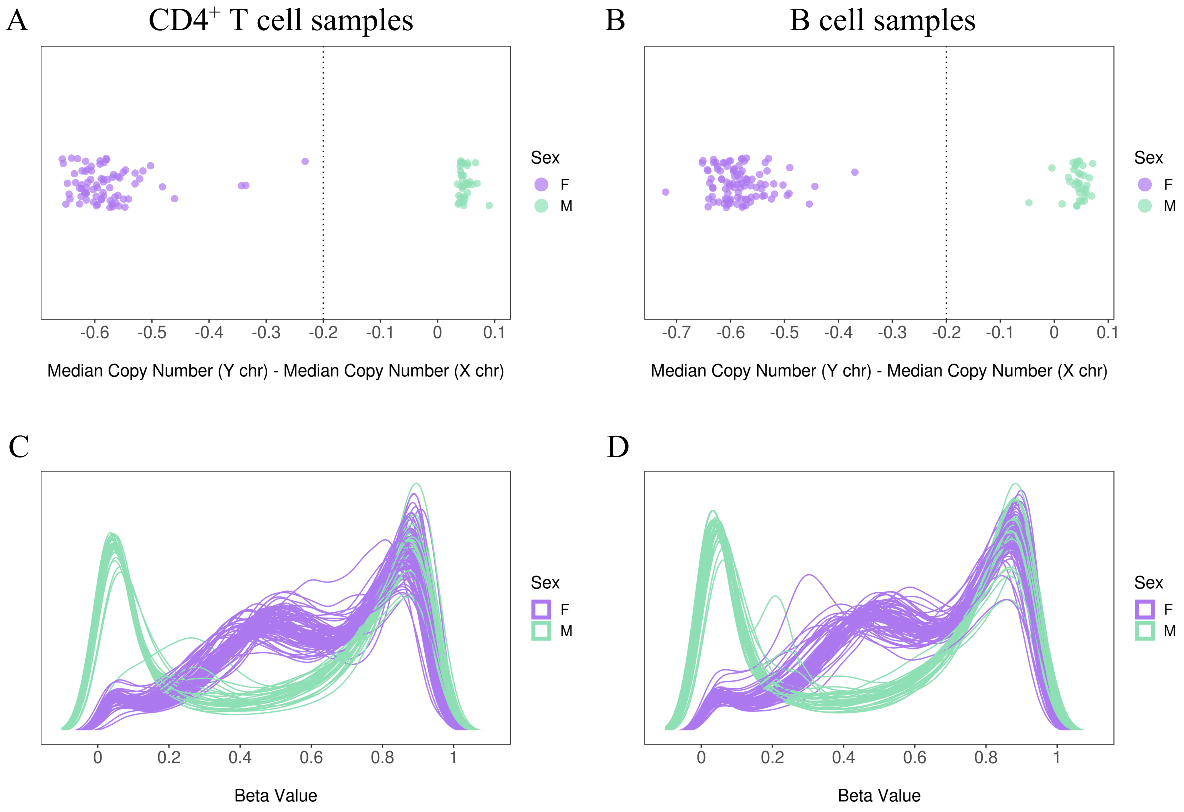


Figure 2.6: Sex prediction using X and Y chromosome probes. Difference in median copy number (sum of methylated and unmethylated channel) between the Y and X chromosomes in (A) CD4$^+$ T cell and (B) B cell samples was used to predict sex. A cut-off of -0.2 was used to assign samples as female. Density plots of β-values from X-chromosome probes in (C) CD4$^+$ T cell and (B) B cell samples confirm patient sex.

2.5.5 DNA methylation data normalisation

Normalisation of data generated by the MethylationEPIC array is an important pre-processing step prior to performing any analyses. The array incorporates two probe type designs (Figure 2.7). The type I design comprises paired beads each hybridized to a 50-nucleotide probe, with the terminal base complimentary to either a C (bisulphite-converted methylated C) or a T (bisulphite-converted unmethylated C). The colour channel is determined by incorporation of fluorescently tagged nucleotides at the single base extension site (Cy3 (green) for C/G and Cy5 (red) for A/T), and both probes therefore fluoresce at the same wavelength. Relative signals from methylated/unmethylated probes at a given locus can be used to determine the overall methylation level. In contrast, type II Infinium design consists of a single bead to interrogate both methylated and unmethylated bisulphite converted DNA, with incorporation of either Cy5-labelled A (unmethylated) or Cy3-labelled G (methylated) nucleotides at the extension site dictating the colour channel. Importantly, the lower dynamic range observed in type II probes can lead to probe-type bias which should be accounted for in data normalisation[268].

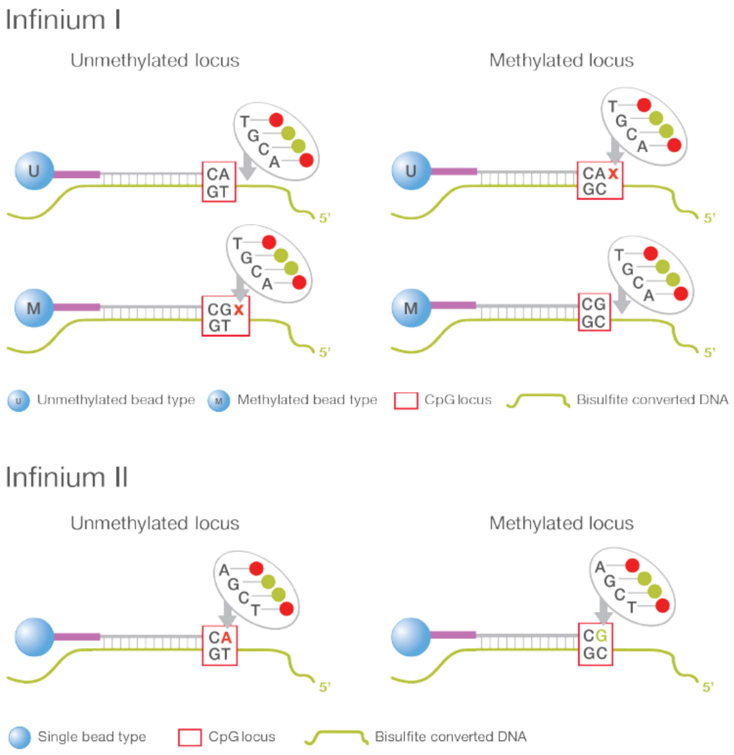


Figure 2.7: Distinct probe chemistry of Infinium type I and type II designs that are present on the MethylationEPIC array. The type I design includes two separate beads, one targeting bisulphite-converted methylated DNA, and the other targeting unmethylated DNA. Incorporation of a base at the single base extension site at either probe causes fluorescence at the same wavelength (green for C/G, red for A/T). Methylation levels at these probes represent the relative signal ratio from each probe. Conversely, the type I design incorporates a single probe for a given locus. Which base is incorporated depends on the sequence of the bisulphite DNA dictates the fluorescence. If bisulphite methylated DNA (C/G) is present, incorporation of the complimentary G/C base generates green fluorescence. If bisulphite unmethylated DNA hybridizes to the probe, incorporation of an A/T base results in generation of a red signal. In this instance, the ratio of green to red fluorescence represents the total methylation. Image taken from the Illumina MethylationEPIC BeadChip Data Sheet[211].

In addition to accounting for this probe type bias inherent in the design of Illumina MethylationEPIC arrays, normalisation methods correct for dye bias by using the positive control probes to correct for the different average intensities that occur between the red and green channels[269]. Furthermore, an additional source of error is the non-specific background fluorescence signals, which can vary from array to array[269]. Negative control probes on the MethylationEPIC array, which do not target CpG sites, allow background fluorescence to be captured as opposed to any biological signal.

Finally, normalisation methods are integral for correcting for effects that can occur when processing multiple samples across separate batches that may result from technical variability such as can occur in sample processing steps. Though good experimental design is essential to reduce the impact of such batch effects (i.e. distributing case and control samples across batches equally), these effects are unavoidable and should be accounted for to prevent bias in downstream analyses. Numerous methods exist for the normalisation of DNAm data collected using the MethylationEPIC array, all of which employ different approaches to adjust for bias in the data.

Three different normalisation methods were applied to the data to test their ability to correct for bias that may arise from factors such as probe type differences and sample-to-sample variability associated with technical batches. These were: (1) normal-exponential using out-of-band probes (noob)[269] with functional normalisation (funnorm)[270], (2) noob with beta mixture quantile dilation (BMIQ)[268], and (3) and subset-quantile within array normalization (SWAN)[271]. The normalisation principles that underpin each of these methods will be discussed further in Chapter 3.4. Funnorm and SWAN were implemented in minfi[267], whilst the wateRmelon package (version 1.26) was used to apply the BMIQ method[272]. The features of each method are discussed in greater detail in Chapter 3 along with the results of the systematic assessment of each method's suitability when applied to the CD4^{+} T cell dataset.

Funnorm corrects for between-array variability by assessing variation in the negative control probes, and utilizing PCA of these probes to adjust for such non-biological variation[270]. The number of control probe principal components to use as a surrogate for technical covariates in funnorm was determined by estimating the dimensionality of the control probe matrix by random matrix theory, as performed using the *EstDimRMT* in the ISVA R package (version 1.9)[273]. As a result, the first three principal components were used to estimate technical covariance for funnorm in both datasets. For BMIQ normalisation, a subset of 10,000 probes of each probe type was used for fitting the 3-state beta mixture model (see Chapter 3.4 for further details).

The suitability of each normalisation method when applied to the DNAm datasets was assessed using a number of different performance metrics (see Chapter 3.4 for results in CD4^{+} T cells). β-value density plots were produced for the raw data, as well as following normalisation, to evaluate the capacity of each method to remove probe bias inherent in the data, with probe type information (type I/type II) extracted from the Illumina annotation file. PCA across all probes, with plotting of the first two PCs for each sample, was subsequently used as a tool to visualise potential batch effects in the data arising from laboratory processing of samples. An additional method used to visualise potential batch effects and their removal following data normalisation was to produce Relative Log Expression (or Relative Log Methylation) plots[274]. Such plots are generated by first obtaining the median logged value for each probe (M-value) across all samples, and then for each sample calculating the deviation in this probe from the median. For each sample, the median and range (interquartile range (IQR) with minimum/maximum values) of these deviations is plotted, and if the probe deviations for a given sample or batch of samples differ considerably from the others, this can represent unwanted variability.

In order to quantify the potential contribution of measured variables to variation in the data identified by PCA, a principal variance component analysis (PVCA) was performed using the pvca package (version 1.22.0). This method combines PCA with variance component analysis (VCA), and first calculates the top PCs, subsequently fitting a mixed linear model to each PC, treating variables of interest as random effects to reveal the proportion of total variability explained. Measured variables of interest that were assessed using this method were sample bisulphite conversion batch, sample position on the BeadChip array, array scanning batch, and the disease diagnosis (RA/non-RA).

To facilitate the identification of technical bias in the data arising from sample processing, technical replicates were included on the BeadChip array, whereby a sample from the same patient was run twice across processing batches. Pearson's correlation across all probes between technical replicates was calculated pre- and post-normalisation to assess the agreement between batches. In addition, sample clustering based on all probes was performed using the flashClust package[275], with the *average* method used to define the distance between clusters. A sample dendrogram was plotted to identify clustering of technical replicates together.

2.5.6 Probe-level filtering

Following data normalisation a series of probe filtering steps were performed to remove any probes that (1) failed the detection p-value cut-off, (2) have been described as cross-hybridizing or multi-mapping in previous studies[212, 276-278], (3) harbour a SNP at a MAF > 0.05 within the probe sequence, single base extension (SBE) site, or CpG site, or (4) map to either of the sex (X or Y) chromosomes. In addition, data normalisation using funnorm often returns a small number of probes with infinite values which cannot be analysed, and as such it was necessary to remove these. Following probe filtering, a total of 709,412 and 710,445 probes for CD4$^+$ T cell and B cell samples, respectively, were included in subsequent analyses. The total number of probes excluded at each stage for both datasets are detailed in Table 2.1.

	Failed (Detection p-value > 0.01)	Cross-reactive	SNP (MAF > 0.05)	X and Y chromosome	Infinite Values	Total Probes Removed	Total Analysis Probes
CD4+ T cells	8,404	51,760	79,388	16,762	133	156,447	**709,412**
B cells	7,001	51,512	79,580	16,826	495	155,414	**710,445**

Table 2.1: Number of MethylationEPIC BeadChip probes removed at each stage of probe filtering in CD4$^+$ T cell and B cell datasets.

After probe filtering, β-values were calculated. In addition, M-values were also calculated as the log2 of the beta value ratio:

$$M\text{-}value = log_2\ (beta\text{-}value / 1 - (beta\text{-}value))$$

M-values are preferable for statistical analyses given that, unlike β-values, they are homoscedastic in nature, such that variability is uniform across the range of values[279].

2.5.7 Batch effect identification in DNA methylation and gene expression data

Residual batch effects that remain in normalised data must be accounted for to avoid the introduction of bias in downstream analyses. Various methods exist that enable potential confounding sources of variation in array data to be accounted for. One popular method, termed ComBat, uses an empirical Bayes approach to adjust data input batches[280]. However, recently it has been shown that when this method is applied together with the specification of a phenotype of interest to be retained, bias can be introduced in downstream differential analyses[281]. An alternative approach is to include known sources of variability (such as the sample processing batch, or biological factors such as age) as covariates in downstream linear models. One drawback to this approach is that, whilst the major sources of variability may be known (for example from PVCA), it is likely that additional technical or biological confounders that have not been directly measured will exist in the data. For this reason, it was decided that the optimal approach would be to use surrogate variable analysis (SVA) to identify such confounders from the data itself, and subsequently include these as covariates in downstream models. This particular approach borrows information across all array probes to estimate sources of variability, termed surrogate variables (SVs), which can then be treated as covariates in subsequent analyses[282]. As well as detecting potential residual sources of variation due to technical batch effects, this approach may also capture biological heterogeneity in the DNAm data, such as different proportions of cell subtypes within each lymphocyte population.

SVA was applied to both normalised DNAm datasets using the sva package (version 3.30.1)[282]. In order to conserve any effects of disease diagnosis (RA/non-RA) for inclusion in downstream analyses, patient diagnosis was included as a variable of interest in the full model, with the null model containing no covariates to allow all variables to be estimated from the data. After applying this method to the data, associations between estimated surrogate variables and measured variables was tested to identify to what extent the SVA was capturing known sources of variability. P-values for associations were generated by linear regression for continuous variables (i.e. CRP levels), analysis of variance (ANOVA) for categorical variables with > 2 categories (i.e. sample processing batch), and Mann-Whitney U test for binary variables (i.e. sex).

When plotting data, the *removeBatchEffects* function in the limma package (version 3.38)[283] was used to adjust for covariates (diagnosis and surrogate variables), although all statistical modelling (differential analyses and quantitative trait locus mapping) was performed using normalised data with these variables provided separately as covariates.

2.6 Differential methylation analysis

2.6.1 Identification of differentially methylated positions

Linear modelling was used to identify any positions that were differentially methylated between RA patients and disease controls in either CD4$^+$ T cells or B cells. Models were fit using the *lmFit* function in the limma package[283], with disease diagnosis (RA/non-RA) treated as the variable of interest and SVs included as covariates in the model. An empirical Bayes method was employed using the *eBayes* limma function to moderate the standard errors[284], as this approach takes into account the probe-wise variability to calculate a moderated t-statistic and rank each CpG by likelihood of being differentially methylated. False discovery rate (FDR) was controlled for using the Benjamini-Hochberg method[285], with an FDR < 0.05 considered to be significantly differentially methylated. To limit the analysis to biologically meaningful differences in DNAm that could feasibly be validated *in vitro*, a delta beta (Δβ, absolute difference between comparator group mean β-values) threshold of 0.05, representing a 5% difference in DNAm, was also applied to filter the results.

2.6.2 Identification of differentially methylated regions

Extended regions of CpGs displaying differential patterns of DNA methylation, termed differentially methylated regions (DMRs), were identified using the DMRcate package in R (version 1.18)[286]. As well as computing the minimum FDR within a DMR, DMRcate also performs a Stouffer transformation which provides a region-specific p-value by taking into account regional correlations in DNAm levels[286]. Initially, an FDR threshold of < 0.05, consistent with the analysis of differentially methylated positions, was used to classify DMPs prior to identification of DMRs. A mean difference in β-value (Δβ) of 0.05 between the RA and non-RA groups across ≥2 CpGs was used for DMR detection.

For discovery purposes, the FDR threshold was relaxed for DMP detection, and non-significant albeit potentially biologically-meaningful DMRs were sought by identifying extended regions of CpGs displaying a consistent Δβ > 0.05 (RA vs. non-RA) across a region of ≥2 CpGs, irrespective of the p-value returned for individual CpGs in the differential analysis.

2.6.3 Identification of differentially variable positions

To identify CpG sites that show differential patterns of variability in DNAm levels between RA patients and disease controls, a test of differential variance, termed the Epigenetic Variable Outliers for Risk Prediction Algorithm (iEVORA) was performed[287]. This algorithm first performs a Bartlett's test to identify instances in which the variance (effectively the spread of values about their mean) differs between the two groups, with an FDR < 0.001 used to define differential variability. Subsequent to this, and to account for the fact that Bartlett's test is susceptible to outlier values skewing group variances, a t-test is performed to rank variable CpGs by the difference between group means, and an unadjusted p-value threshold of < 0.05 used to signify a differentially variable position (DVP). This test was implemented using the *row_ievora* function in the matrixTests package (version 0.1.4). CpGs with a higher variance in RA patients were considered RA hyper-variable, whereas sites for which variance was greater in the non-RA patients were termed RA hypo-variable.

2.6.4 Gene ontology pathway analysis

Gene ontology pathway analysis was performed to identify biological pathways enriched amongst the genes to which either hyper-variable or hypo-variable RA-associated DVPs mapped. Previous studies have shown that bias is introduced when performing such tests for enrichment with array-based DNA methylation data, resulting from the unequal distribution of CpG probes across all genes[288]. The *gometh* function in the missMethyl package (version 1.16.0) was used to apply a modified hypergeometric test that enables the design of the MethylationEPIC array, with differing numbers of probes mapping to each gene, to be accounted for[289, 290].

2.6.5 Assessing the DNA methylation age of early arthritis lymphocytes

To test whether lymphocytes from RA patients exhibited an accelerated biological age relative to those from non-RA patients, the 'epigenetic age' of CD4$^+$ T cell samples and B cell samples was calculated from DNAm data using the *agep* function in the wateRmelon package[272], specifying the use of Horvath's coefficients to calculate the age[291].

To assess whether disease diagnosis had a significant effect on accelerating biological age relative to chronological age, a linear model (*lm* function) was fit to the data to test the effects of chronological age (independent variable) on biological (Horvath) age (dependent variable), with the inclusion of disease diagnosis (RA/non-RA) as an interaction term. Interaction effects (chronological age × disease diagnosis) with a $p < 0.05$ were considered significant.

2.7 Methylation Quantitative Trait Locus Analysis

2.7.1 Mapping of meQTLs in cis and trans

MeQTLs were identified by fitting additive linear models to identify associations between genotype and DNAm levels at CpG sites using the MatrixEQTL package (version 2.2)[292]. MeQTLs were identified in each cell type separately and were mapped both in cis, whereby the CpG site is located < 1MB away from the regulatory SNP, and in trans, which occur over larger distances (> 1MB or separate chromosomes). To reduce the potential for false positive results to arise from outlier samples, the analysis was limited to SNPs for which each genotype was represented by ≥3 patients, or in the absence of any minor allele homozygotes, ≥8 patients. Following this filtering step, a total of 2,901,876 $CD4^+$ T cell SNPs, and 3,035,821 B cell SNPs were included in the meQTL analysis, testing for associations against 709,412 and 710,445 CpGs in an all-against-all analysis (a reduced number of tests in the cis analysis reflected that associations were only mapped for CpGs in a 1Mb window upstream and downstream of each SNP). Covariates passed to the *MatrixEQTL* function were disease diagnosis (RA/non-RA), and the SVs identified using SVA (section 2.5.7).

FDR values were calculated using the Benjamini-Hochberg method[285], and were generated separately for cis and trans associations. To reduce the computational burden associated with calculating FDR values across $> 2 \times 10^{12}$ independent tests, an unadjusted p-value threshold of 1×10^{-2} and 1×10^{-4} for FDR calculation was selected for cis and trans associations, respectively. SNP-CpG associations in cis with an FDR < 0.01 were considered statistically significant, with a threshold of 1×10^{-5} selected for those in trans. A lower p-value cut-off was selected in the trans analysis to reduce the likelihood of false positive associations by setting a higher threshold for evidence of a long-distance association.

2.7.2 Mapping disease-specific meQTLs

An interaction analysis was performed to identify meQTL effects that were either specific to the RA or non-RA comparator groups, or for which the effect size differed significantly between the two groups. The same analysis was performed as for genome-wide detection of meQTLs, albeit with the inclusion of an interaction term (*genotype x diagnosis*) in the linear model. This tests whether or not the effect of the independent variable (SNP genotype) on the dependent variable (DNAm levels) is significantly influenced by the value of a third variable (in this case either a positive or negative diagnosis of RA). As before, the analysis was limited to SNPs for which ≥ 3 patients were represented in *both* the RA and non-RA groups. Whilst FDR values for the initial meQTL analysis were chosen to robustly identify SNP-CpG associations, for the purpose of discovering more subtle effects of diagnosis on these effects, an FDR cut-off of < 0.05 for disease interaction effects at cis loci, and $< 1 \times 10^{-3}$ for those at trans loci, was selected. To mitigate against the potential for false positive effects or those with small differences in effect size, CpGs for which the 10th and 90th percentile DNAm values differed between by a β-value of < 0.05 were excluded.

2.7.3 SNP clumping

Due to patterns of linkage disequilibrium (LD), meQTL analyses often nominate multiple SNPs for a given meQTL effect, all of which are associated with DNAm levels at the same CpG, and it is necessary to remove SNPs tagging the putative causal variant to infer independent associations. To this end, SNP clumping in PLINK was performed to remove tagging SNPs at a given locus. For this purpose, an LD threshold of 0.001 (1000 Genomes Project Phase 3, European (EUR) populations) was used to remove SNPs in a window of 250Kb, each time retaining the SNP displaying the strongest association (lowest p-value) with DNAm levels at that locus. Clumping was performed on both cis- and trans-meQTL results.

2.7.4 Functional annotation of cis-meQTL-associated DNA methylation

To gain additional functional insight from the genomic context of CpG sites regulated in cis, these positions were overlapped with cell specific chromatin state data available from the Roadmap Epigenomics Project[173]. These chromatin states are defined based on patterns of histone modifications, specifically mono-methylation and tri-methylation of lysine 4 on histone 3 (H3K4me1, H3K4me3), as well as tri-methylation of lysine 9 (H3K9me3), lysine 27 (H3K27me3), and lysine 36 (H3K36me3) assayed by ChIP-seq. A total of 15 chromatin states had been defined by relative enrichment of these histone modifications, with chromatin state learning performed using a Hidden Markov Model approach[173]. To aid in the interpretation of

CpG overlaps, these states were collapsed into five functional classes: *transcription start site (TSS)* (active TSS, bivalent/poised TSS*), flanking transcription start site* (flanking active TSS, flanking bivalent TSS), *enhancers* (genic enhancer, enhancer, bivalent enhancer), *transcribed* (transcribed at gene 5' and 3', strong transcription, weak transcription), and, finally, *repressed* (ZNF genes + repeats, heterochromatin, repressed polycomb, weak repressed polycomb, quiescent/low). CD4$^+$ T cell cis-meQTL-associated CpGs (cis-CpGs) were overlapped with chromatin states from peripheral T helper cells (cell type ID = E043) and in some cases primary T regulatory cells (Tregs) from peripheral blood (cell type ID = E044), whereas for B cell CpG overlap was performed using states from primary B cells in peripheral blood (cell type ID = E032).

Cis-CpGs were also assessed in relation to their mapping to CGI features. CGI annotations were obtained from the Illumina MethylationEPIC manifest, which defines feature as mapping to either CGIs, CGI shores (north/south, +/- 0-2Kb of CGI), CGI shelves (north/south, +/- 2-4Kb of CGI), or Open Sea (all other regions, see Figure 2.8).

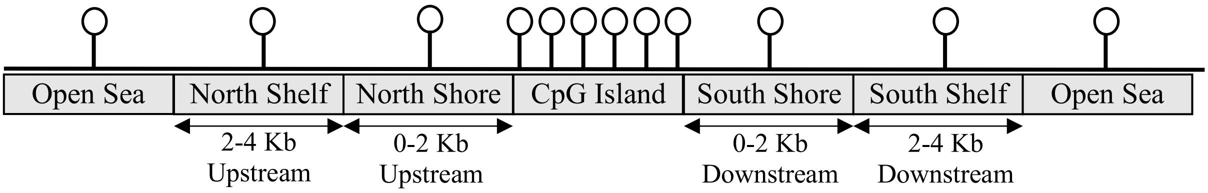


Figure 2.8: CpG islands and related features. The mapping of CpGs to CpG islands and related flanking regions (shores/shelves) is defined in the Illumina Infinium MethylationEPIC annotation file. Open sea regions are classified as all those that are not classified as islands, shores, or shelves.

Enrichment analysis of cis-CpGs at gene features as defined by the University of California Santa Cruz (UCSC) RefGene annotations was also assessed. Cis-CpGs were mapped to either 5' untranslated regions (5'UTR), 200 bases from a transcription start site (TSS200), 1500 bases from a transcription start site (TSS1500), the first Exon, and Exon boundary, the 3'UTR, the gene body, or to an intergenic region (Figure 2.9). Enrichment of cis-CpGs at all features (chromatin states, CGI features, gene features) relative to CpGs not associated with a DNA variant in cis was assessed using a two-way Fisher's exact test in R.

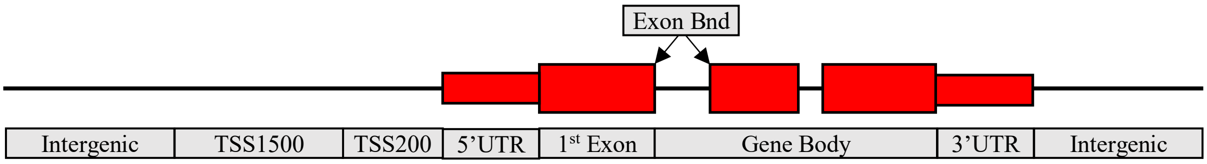


Figure 2.9: UCSC RefGene gene features. TSS1500 = within 1500 base pairs from a transcription start site (TSS); TSS200 = within 200 base pairs from a TSS; 5’UTR = 5’ untranslated region, Exon Bnd = Exon boundary, 3’UTR = 3’ untranslated region. All regions not mapping to these features in relation to a gene are annotated as intergenic.

2.7.5 Co-localisation analysis at disease-associated loci

MeQTLs that putatively underlie disease mechanisms were sought by performing co-localisation analyses of these loci with disease-associated loci from GWASs. GWAS data for four diseases; RA, MS, asthma, and OA were downloaded from the GWAS catalogue[293]. Whilst the focus of this project was on deciphering mechanisms underlying lymphocyte-mediated genetic risk in RA patients, these effects were compared with genetic mechanisms underlying risk to MS and asthma, both of which are considered largely immune-mediated diseases[87, 294]. In addition, the overlap of lymphocyte meQTLs with OA risk loci was assessed; this disease affects the same tissues as RA but the disease mechanisms are distinct, with pathways such as collagen formation and catabolism of the extracellular matrix (rather than lymphocyte dysregulation) believed to be major contributors to pathogenesis[264]. The search terms used to identify reported traits for these four diseases in the GWAS catalogue are listed in Table 2.2. Initially, co-localisation was determined whereby the meQTL variant was in high LD ($r^2 \geq 0.8$ in EUR populations) with the lead variant identified by GWAS.

In instances where the meQTL and RA GWAS signal were found to map to the same locus, additional evidence of co-localisation was sought by performing a Bayesian test for co-localisation in the coloc R package (version 3.2.1)[295]. This approach calculates Bayes factors, which can be thought of as a measure of evidence for a given hypothesis, to compute the posterior probability (PP) of five possible hypotheses:

- H_0 – Neither trait (DNAm levels or RA susceptibility) has a genetic association at this locus.
- H_1 – Trait 1 (DNAm levels), but not trait 2 (RA susceptibility), has a genetic association at this locus.

- H_2 – Trait 2 (RA susceptibility), but not trait 1 (DNAm levels) has a genetic association at this locus.
- H_3 – Both traits have genetic associations at this locus, with independent causal SNPs.
- H_4 – Both traits have genetic associations at this locus, with a shared causal SNP.

Given that genetic associations with both RA genetic susceptibility and DNA methylation have been established prior to applying this test, this approach was used in principle to provide evidence of either H_3 or H_4 regulatory models described above. The test was performed using the *coloc.abf* function, using summary data from MatrixEQTL for the identified meQTL variants, and RA GWAS summary statistics obtained from a meta-analysis[97]. A prior probability of H_4 (PP4) > 0.75 was used as a threshold for good evidence of a shared variant.

Disease	GWAS Catalogue Search Terms
Rheumatoid arthritis (RA)	Rheumatoid arthritis, Rheumatoid arthritis (ACPA-positive), Rheumatoid arthritis (ACPA-negative) Rheumatoid arthritis (rheumatoid factor and/or anti-cyclic citrullinated peptide seropositive.
Multiple Sclerosis (MS)	Multiple sclerosis.
Asthma	Asthma, Adult asthma, Asthma (childhood onset) Asthma & hay fever, Asthma (adult onset), Asthma onset (childhood vs adult), Asthma (age of onset), Asthma (moderate or severe), Allergic disease† (asthma, hay fever, or eczema), Asthma or allergic disease (pleiotropy).
Osteoarthritis (OA)	Osteoarthritis, Osteoarthritis (hand, severe), Osteoarthritis (hip), Osteoarthritis (knee), Osteoarthritis (hip or knee) Osteoarthritis (hospital diagnosed), Osteoarthritis (self-reported), Osteoarthritis (with total joint replacement).

Table 2.2: GWAS Catalogue reported traits. DNA data from GWAS studies of four diseases (rheumatoid arthritis, multiple sclerosis, asthma, and osteoarthritis were downloaded from the GWAS catalogue to perform co-localisation analyses with meQTLs identified in CD4$^+$ T cells and B cells of early arthritis patients.
†This particular GWAS found that genetic associations with asthma exhibited a very strong correlation with other immune-mediate allergies, and a meta-analysis was performed to increase power to detect associations.

2.7.6 Chromatin state and transcription factor binding site enrichment at risk loci

As was performed for cis-CpGs in section 2.7.4, enrichment of cis-CpGs associated with GWAS loci for a given disease at cell-specific chromatin states was assessed. In this instance, enrichment of cis-CpGs associated with risk loci was calculated relative to cis-CpGs that were associated with non-risk meQTLs for that disease, using a two-way Fisher's exact test.

To identify possible enrichment at TFBSs, cis-CpGs were overlapped with ChIP-seq-generated binding site data for 161 transcriptional regulators generated from several human cell lines by the Encode project [209, 210]. TFBS data from all cell types were used to test for enrichment, again using a two-way Fisher's exact test to identify whether the occurrence of any TFBSs at risk-associated cis-meQTL CpGs was over-represented relative to non-risk cis-CpG. To overcome the multiple testing burden associated with testing binding sites at > 100 TFs, a Bonferroni-adjusted p-value was used based on the number of TFs tested.

As with DVPs (see section 2.6.4), CpGs associated with cis-meQTLs were subject to pathway analysis using the *gometh* function in the missMethyl package[289]. Cis-CpGs associated with GWAS risk loci were tested for enrichment of specific biological processes, using as background cis-CpGs outside of the risk loci for the disease being tested.

2.8 Transcriptomic profiling

Genome-wide transcriptomic profiling was performed using the HumanHT-12 v4 BeadChip (Illumina), which includes 47,231 probes, as has been described[139]. The integrity of cell-specific extracted RNA was confirmed using a Bioanalyzer (Agilent, CA, USA), with a median RNA integrity number (RIN) of 10 for CD4$^+$ T cell samples (range 8.2 - 10) and 10 (range 7.9 - 10.0) for B cell samples. Complementary RNA (cRNA) was synthesized from 250ng total RNA using the Illumina® TotalPrep™ RNA Amplification Kit (ThermoFisher; Cat# AMIL1791). cRNA was then hybridised to the BeadChip array and washed following the manufacturer's protocol, with imaging of arrays using the iScan system as for the MethylationEPIC arrays (Illumina). Intensity values for all samples were obtained in Illumina's IDAT file format using the GenomeStudio Gene Expression Module (version 1.8, Illumina), with all subsequent data processing and analysis performed in the R statistical environment.

2.8.1 Pre-processing of transcriptomic data

Transcriptomic data described in section 2.8 had been generated prior to the current study, with quality control performed and integration/normalisation of data from two separate array platforms[139]. For the purpose of the current study, transcriptomic data only from samples for which DNAm and genotype data were available (all of which were profiled using the Illumina HumanHT-12 v4 BeadChip) were extracted and normalised separately from this larger dataset. As with DNAm data from the MethylationEPIC array, transcriptomic data from CD4$^+$ T cell and B cell samples were pre-processed independently.

Initially, as for DNAm data, detection p-values for each probe were extracted across all samples. Unlike for the MethylationEPIC array, where all probes should return a low detection p-value unless failed, low values from the HumanHT-12 array may be indicative of genes with low expression levels in the cell type being studies. For this reason, a more relaxed threshold of > 0.05 in at least 25% of samples was selected to identify failed probes. Samples were then removed for which the mean detection p-value remained > 0.05 after excluding failed probes.

Background correction using negative control probes and quantile normalisation were performed in limma [283] using the *neqc* function. After data normalisation, probes were remove if 1) they returned a failed detection p-value as defined above, 2) the probe quality was deemed as 'Bad' or 'No match' based on probe mappings from the illuminaHumanv4.db package, or 3) they mapped to either the X or the Y chromosome.

As with the DNAm data, SVA was applied to identify residual variability to be accounted for in downstream analyses, with any effects due to diagnosis preserved (see section 2.5.7).

2.9 Expression Quantitative Trait Methylation analysis

The impact of DNA methylation at disease risk cis-meQTL CpGs on transcriptional regulation was assessed through the integration of contemporaneous cell-specific transcriptomic data. Transcript mapping information was first extracted from the Ensembl database using the biomaRt package (version 2.38)[296], and unique Illumina identifiers retrieved for those having a transcription start site (hg19 reference genome) mapping to within ±500Kb of a risk-associated cis-CpG as determined using the GenomicRanges package (version 1.34)[297].

Associations between DNAm at risk cis-CpGs and transcript levels of genes within ±500Kb window, termed cis-eQTMs, were identified by non-parametric Spearman's rho. To account for technical or biological variables, correlations were performed using residuals from linear regression. Firstly, linear models were fit to both DNAm M-values and gene expression datasets using the *lmFit* function in limma, with the inclusion of disease diagnosis and SVs as covariates, and model residuals subsequently extracted. These residuals were then used to compute correlations. Benjamini-Hochberg[285] adjusted p-values, calculated across the total number of transcripts tested for a given CpG (i.e. the number of genes with a 500Kb window upstream and downstream) were calculated, with adjusted p-values < 0.01 considered significant. Though residuals were used for performing eQTM analysis, as is consistent with the input for causal inference testing (see section 2.10), when plotting associations DNAm and gene expression data were adjusted for these covariates using the *removeBatchEffects* function in limma (see section 2.5.4).

2.10 Causal inference testing

Though identification of cis-eQTM effects at CpGs subject to cis-meQTL effects is suggestive of a mechanism whereby DNAm mediates transcriptional regulation by functional SNPs, as discussed in section 1.6.9, a number of possible regulatory models may explain these associations. In order to distinguish instances of methylation-mediated regulation (SNP – Methylation – Expression; SME) from reverse mediation (SNP – Expression – Methylation; SEM) or independent effects (INDEP; see Figure 1.8), a causal inference test (CIT) was implemented[260]. This particular test aims to infer causality by performing four statistical tests based on conditional correlations, all of which must be satisfied to conclude that the mediator of interest (in this case DNAm) effects the phenotypic outcome measure (gene expression)[260]:

1. The SNP is associated with the phenotype (transcript levels).

2. The SNP is associated with the mediator (DNAm) conditional on the phenotype.

3. The mediator is associated with the phenotype conditional on the SNP.

4. The SNP is independent of the phenotype conditional on the mediator.

These four tests are combined into an 'omnibus' p-value, whereby the probability of a causal model is only as strong as the weakest p-value in this chain of four tests. For example, in instances whereby the prevailing model is one of reverse mediation (SEM), condition 2 outlined above would not be satisfied, whereas independent effects on DNAm and transcript levels (INDEP) would not be consistent with condition 3.

CIT was performed on all triplets at cis-meQTL/cis-eQTMs (risk SNP, cis-CpG, transcript) using the *cit.cp* function for testing a continuous outcome measure in the cit package (version 2.2)[298]. As with eQTM analyses, DNAm and gene expression residuals were used as input to the model. One thousand permutations of the test were performed, and the *fdr.cit* function applied to calculate FDR values. DNAm was deemed to likely mediate the SNP effect on gene expression for triplets at which the CIT FDR was < 0.05.

2.11 Validation of cis-meQTLs at loci of interest

Validation of CD4$^+$ T cell cis-meQTL effects was performed at loci of interest using bisulphite pyrosequencing as a targeted approach to quantify DNAm in an independent cohort of 39 patients for whom genotyping had been performed (see section 5.4.2 for further details).

2.11.1 Isolation of genomic DNA and bisulphite conversion

To quantify DNAm levels at a given CpG site, genomic DNA must be bisulphite converted - a process which results in conversion of unmethylated cytosine residues to uracil, whilst methylated residues are protected against this conversion (see section 1.6.5).

CD4$^+$ T cell DNA was available from patients who had not been included in the initial discovery cohort of samples for which DNAm quantification was performed using the MethylationEPIC array. Isolation of these cells and extraction of DNA had been performed exactly as described in section 2.2. DNA concentration was quantified using the Nanodrop ND-1000 Spectrophotometer (ThermoFisher), and 250-300ng of DNA bisulphite-converted and purified using the EZ DNA methylation kit (Zymo Research, Irvine, CA, Cat# D5001), following the manufacturer's protocol. Conversion carried out for 16 hours at 50°C on the Alpha Cycler 1 (PCRmax, Staffordshire, UK), and converted DNA was eluted in 30µl elution buffer.

2.11.2 Quantification of DNA methylation by pyrosequencing

Primers for amplification and pyrosequencing of regions harbouring cg17134153 (hg19 chr1:157,670,328 mapping to *FCRL3*), cg21124310 (hg19 chr5:55,444,106), and cg07522171 (hg19 chr7:28,218,686) were designed using the PyroMark® Assay Design SW 2.0 (Qiagen). The design criteria for primers are listed in Table 2.3. One of the primers in each pair contained a biotin tag to allow for subsequent immobilisation of PCR products on sepharose beads.

	Amplification Primers	**Sequencing Primers**
Minimum Primer Length (nt)	18	15
Maximum Primer Length (nt)	30	25
Maximum Amplicon Length (nt)	400	-
Minimum Melting Temperature (°C)	50	29
Maximum Melting Temperature (°C)	72	59
Maximum GC Difference (%)	50	-
Maximum Distance from Target (nt)	-	10

Table 2.3: Design criteria for bisulphite pyrosequencing primers. A series of parameters were set for designing primers to allow targeted quantification of DNA methylation. Primers were designed using PyroMark® Assay Design SW 2.0 (Qiagen). nt = nucleotides.

Regions of interest were amplified from bisulphite-converted genomic DNA using the Titanium Taq PCR kit (Clontech Laboratories, Mountain View, CA; Cat# 639210). DNA was amplified in a 20µl reaction containing 1µl template DNA (~30ng), 2µl Titanium Taq PCR Buffer (10X), 0.2µM each of forward and reverse primers (listed in Table 2.4; synthesized by Sigma-Aldrich), 0.4µl dNTP mix (final concentration of 0.2mM each of dATP, dCTP, dGTP & dTTP; ThermoFisher; Cat# R0192), 0.4µl Titanium *Taq* DNA polymerase (50X), made up to the final volume with diethyl pyrocarbonate (DEPC)-treated H_2O.

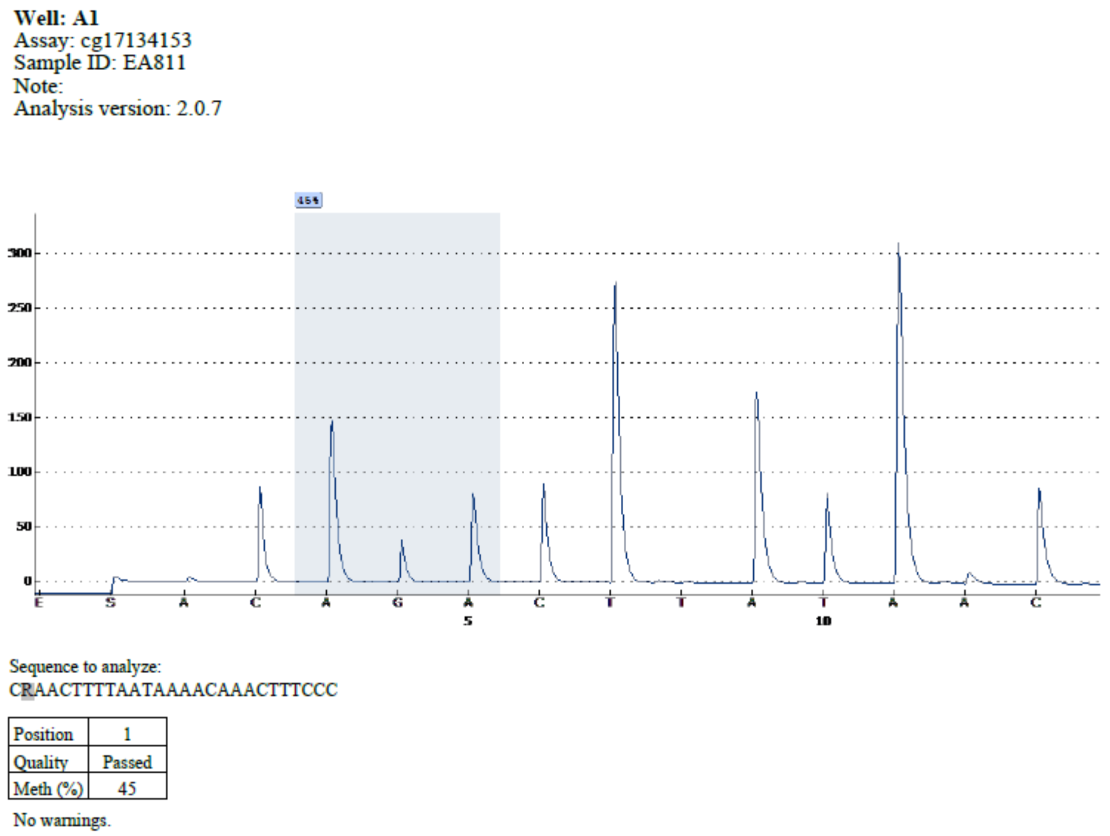
To determine the optimal primer annealing temperature, a temperature gradient PCR reaction was run on an Alpha Cycler 1, with initial denaturation at 95°C for one minute, followed by 40 cycles of 95°C for 15 seconds, 55-70°C for one minute, and 68°C for one minute, with a final extension step at 68°C for 5 minutes. Following amplification, PCR products were run on an agarose gel (3% w/v in Tris-acetate-EDTA (TAE; 40mM Trizma® base (Sigma-Aldrich), 20mM acetic acid (Fisher Scientific, Loughborough, UK), 1mM ethylenediaminetetraacetic acid (EDTA) disodium salt dehydrate; Sigma-Aldrich; Cat# E5134)) buffer with 1% v/v ethidium bromide (EtBr; Sigma-Aldrich; Cat# E1510)) at 100 volts for 90 minutes alongside a Quick-Load® 100bp DNA ladder (New England Biolabs, Ipswich, MA; Cat# N0467S). Bands were visualised using the Odysey® Fc imaging system (LI-COR Biosciences, Lincoln, NE). From the optimisation reaction, the optimal annealing temperature was determined to be 63°C for cg07522171/cg21124310, and 68°C for cg17134153. All subsequent amplifications for bisulphite pyrosequencing were carried out under the above cycling parameters with optimised annealing temperatures. Each sample or condition for which DNAm was to be quantified by pyrosequencing were amplified in duplicate from the same bisulphite DNA template.

	Primer	**Sequence (5' ➞ 3')**
CpG	cg17134153 (Forward)	[Btn]TAGAGGGTTGGGAAAGTTTGT
	cg17134153 (Reverse)	CCACATTCACATTTTCAAAACCCAAAAC
	cg17134153 (Sequencing)	CCCTCCTTCTTAAAAATAAAT
	cg21124310 (Forward)	TTGGAGTTTTATTGAGGGATAAATTGA
	cg21124310 (Reverse)	[Btn]ATATTCCTCCTCACTCTTTAAACC
	cg21124310 (Sequencing)	TGAGGGATAAATTGAGTT
	cg07522171 (Forward)	[Btn]TAAGTAAAGGAGTATAGGGTTTTGTT
	cg07522171 (Reverse)	TACCCCCAAAAAATCCAAATAAATACCATA
	cg07522171 (Sequencing)	CTACAAAATTAAAAAAATAAATCAC
Allelic Expression Quantification	rs7522061 (Forward)	TGGGCTAGGGAATGTGATATG
	rs7522061 (Reverse)	[Btn]TGGCCCCAAAAGCTGTAC
	rs7522061 (Sequencing)	TGGACCATGGAGGAT

Table 2.4: Amplification and sequencing primers used for pyrosequencing in both CpG DNA methylation and allelic quantification assays. [Btn] = Biotin Tag.

DNAm quantification was performed using the PyroMark Q24 MDx system with PyroMark Gold Q96 reagents (Qiagen; Cat# 972804). Following amplification, 10µl PCR product was transferred to a 24-well PCR plate (non-skirted, elevated wells, Starlab, Milton Keynes, UK). 1.5µl Streptavidin-coated Sepharose beads (GE Healthcare, Ursula, Sweden; Cat# 17511301), 40µl PyroMark Binding Buffer (Qiagen), and 28.5µl DEPC-H_2O were added to each well containing PCR product and agitated for 10 minutes at room temperature using an Orbis 700-235 Microplate Shaker (Cole-Parmer Instruments, St. Neots, UK). The beads, complexed to the

PCR product, were captured using the PyroMark Q24 Vacuum Workstation (Qiagen), following which a series of wash steps were performed with 75% ethanol for 5 seconds, PyroMark Denaturing solution for 5 seconds, and 1X PyroMark Wash Buffer for 10 seconds. Beads were then released into the wells of a PyroMark Q24 plate containing 25µl sequencing primer (Table 2.4; synthesized by Sigma-Aldrich) diluted to 0.3µM in PyroMark Annealing buffer, with subsequent incubation at 80°C for 2 minutes. The plate was then loaded into the instrument, together with the PyroMark Q24 cartridge containing appropriate volumes of nucleotides (dATPαs, dCTP, dGTP, dTTP) and the enzyme/substrate mixes. The run was performed using a custom CpG assay created using PyroMark Q24 software (version 2.0.7), and the output T/C ratio used to determine the final DNA methylation value. Samples which passed the quality check (Figure 2.10) and returned duplicate values within a 5% range were included in analyses.



Well: A1
Assay: cg17134153
Sample ID: EA811
Note:
Analysis version: 2.0.7

Sequence to analyze:
CRAACTTTTAATAAAACAAACTTTCCC

Position	1
Quality	Passed
Meth (%)	45

No warnings.

Figure 2.10: Trace plot generated in PyroMark Q24 software. This depicts an exemplar plot from an analysis run on the PyroMark Q24 MDx system (Qiagen). This example shows a sample that passed the quality check and returned a methylation value of 45%.

2.11.3 Validation of Pyrosequencing Assays for DNA methylation Quantification

In order to confirm that pyrosequencing assays designed were able to accurately call allele ratios, and thus reliably quantify DNAm, a validation experiment was set up using artificial mixes of allele ratios to generate a standard curve. gBlocks® gene fragments were designed (synthesized by Integrated DNA Technologies (IDT), Coralville, IA) representing bisulphite-converted double-stranded DNA of methylated and unmethylated CpGs of interest, with 5' and 3' adaptors added to reduce sequence complexity and facilitate fragment synthesis (Table 2.5).

Fragment	Sequence (5' ➞ 3')
cg17134153 (Unmethylated)	TGATGGCTACTGGCTCGGATCCGTTATGGCTAGAGGGTTGGGAAAGTTTGTTTTATTAAAAGTT**TG**ATTTATTTTTAAGAAGGAGGGTAGGAAGTTGTTATTTAGATGAGATTTGTAAGAATTAGAAAAGGGAAGAAGAGTTTAGTGTTATATTTTGTTTTGGGTTTTGAAAATGTGAATGTGGTTATGTCTTTAAATGTTGCATGTATCCGTA
cg17134153 (Methylated)	TGATGGCTACTGGCTCGGATCCGTTATGGCTAGAGGGTTGGGAAAGTTTGTTTTATTAAAAGTT**CG**ATTTATTTTTAAGAAGGAGGGTAGGAAGTTGTTATTTAGATGAGATTTGTAAGAATTAGAAAAGGGAAGAAGAGTTTAGTGTTATATTTTGTTTTGGGTTTTGAAAATGTGAATGTGGTTATGTCTTTAAATGTTGCATGTATCCGTA
cg21124310 (Unmethylated)	TGATGGCTACTGGCTCGGATCCGTTATGGCTTGGAGTTTTATTGAGGGATAAATTGAGTTTT**TG**AAGTATTTAGGAGTTGATAATTTAGTAGTTATTATTAGGTATGTTGTAATAAATATTTAGATGGTTTTGGTGATGGGAGAATTTTATTTTTTTGAAAATTAAAAAGTATTGATTGGTTTAAAGAGTGAGGAGGAATATTTATGTCTTTAAATGTTGCATGTATCCGTA
cg21124310 (Methylated)	TGATGGCTACTGGCTCGGATCCGTTATGGCTTGGAGTTTTATTGAGGGATAAATTGAGTTTT**CG**AAGTATTTAGGAGTTGATAATTTAGTAGTTATTATTAGGTATGTTGTAATAAATATTTAGATGGTTTTGGTGATGGGAGAATTTTATTTTTTTGAAAATTAAAAAGTATTGATTGGTTTAAAGAGTGAGGAGGAATATTTATGTCTTTAAATGTTGCATGTATCCGTA
cg07522171 (Unmethylated)	TGATGGCTACTGGCTCGGATCCGTTATGGCTAAGTAAAGGAGTATAGGGTTTTGTTTTATTTTATTTTTGTATAAATATATAGTAGT**TG**TGATTTATTTTTTTAATTTTGTAGAAATGTGAGTTGTATTTATATGGTTGAATTTATGGTATTTATTTGGATTTTTTGGGGGTATTATGTCTTTAAATGTTGCATGTATCCGTA
cg07522171 (Unmethylated)	TGATGGCTACTGGCTCGGATCCGTTATGGCTAAGTAAAGGAGTATAGGGTTTTGTTTTATTTTATTTTTGTATAAATATATAGTAGT**CG**TGATTTATTTTTTTAATTTTGTAGAAATGTGAGTTGTATTTATATGGTTGAATTTATGGTATTTATTTGGATTTTTTGGGGGTATTATGTCTTTAAATGTTGCATGTATCCGTA

Table 2.5: Fragment sequences used for the validation of pyrosequencing assays targeting CpGs at regions of interest. For each CpG assay, a fragment was designed to represent both the unmethylated (T) and methylated (C) cytosine following bisulphite-conversion of DNA. The CpG site targeted by the respective assay is highlighted in red, and underlined sequences represent 5' and 3' adapters that were added to the ends of probes to facilitate synthesis.

Fragments (250ng) were re-suspended at 2.5ng/µl in 100µl DEPC-H_2O with incubation at 50°C for 20 minutes on a heat block. Solutions were subsequently prepared containing varying proportions of methylated and unmethylated DNA fragments to represent methylation values across the entire range in 10% increments (0 – 100%, Table 2.6). These solutions were then diluted 1:100 in DEPC-H_2O to a concentration of 25pg/µl, and methylation quantified by pyrosequencing as described above in section 2.11.2. Results of the pyrosequencing assay validation are reported together with meQTL validation results in section 5.4.

Methylation (%)	Methylated (C) gBlock (2.5ng/µl)	Unmethylated (T) gBlock (2.5ng/µl)
100	10	0
90	9	1
80	8	2
70	7	3
60	6	4
50	5	5
40	4	6
30	3	7
20	2	8
10	1	9
0	0	10

Table 2.6: Preparation of allele mixes from synthetic bisulphite DNA for the generation of standard curves for validation of CpG assays. Varying proportions of the methylated (C) and unmethylated (T) sequences were mixed to produce solutions representing a pool of cells with DNA methylation values ranging from 0 – 100%. These were sequenced to determine if the CpG assays could correctly call DNA methylation levels at the CpGs of interest.

2.12 Allelic Expression Analysis

To quantify any allelic effects on *FCRL3* gene expression, analysis of allelic expression imbalance (AEI) was performed in patients heterozygous at the regulatory SNP. This technique requires that the regulatory SNP is in high LD with a proxy transcript SNP. By quantifying the levels of the risk allele at this proxy SNP in the mRNA relative to the genomic DNA (gDNA), the extent to which the regulatory SNP confers increased mRNA levels can be quantified. In the genomic DNA of heterozygous individuals, the proportion of each allele copy is 1:1 (i.e. 50% risk allele), and significant deviations from this in the mRNA are indicative of the risk allele conferring either increased or reduced gene expression levels.

This technique was possible at the *FCRL3* locus, given the presence of a proxy SNP (rs7522061) within Exon 4 of the gene that is in high LD ($r^2 > 0.9$ in EUR populations) with the regulatory SNP (rs2210913) identified in the meQTL analysis. The process itself involves reverse-transcription of CD4$^+$ T cell RNA from patients heterozygous at rs2210913, and quantification of allele frequencies in this cDNA and gDNA using pyrosequencing.

2.12.1 Reverse Transcription

To enable allelic expression quantification, CD4$^+$ T cell RNA was first reverse transcribed using the SuperScript™ II reverse transcriptase kit (Invitrogen, Carlsbad, CA; Cat# 18064014). RNA was available from isolated patient CD4$^+$ T cells as described in section 2.2. RNA was quantified using the Nanodrop ND-1000 Spectrophotometer and, for allelic expression analysis, 400ng transferred to a 200µl PCR tube (Starlab) and made up to 6µl with DEPC-H_2O, followed by the addition of 1µl random hexamers (1µg/µl; Invitrogen; Cat# N8080127) and 1µl dNTP mix (0.5mM each in final 20µl volume; ThermoFisher). Samples were incubated at 65°C for 5 minutes and quickly chilled on ice, followed by the addition of 4µl First Strand Buffer (5X), 4µl $MgCl_2$, 2µl dithiothreitol (DTT, 0.1M), and 1µl RNaseOUT™ ribonuclease inhibitor (40U/µl, Invitrogen; Cat# 10777019), with further incubation at 25°C for 1 minute. 1µl SuperScript™ II reverse transcriptase (200U/µl) was added and samples incubated at 25°C for 10 minutes, 42°C for 50 minutes, and 70°C for 10 minutes, with cDNA stored at -20°C until use. As a negative control for reverse transcription, a reaction was set up with 1µl DEPC-H_2O replacing reverse transcriptase enzyme.

2.12.2 PCR amplification and allelic quantification by pyrosequencing

Primers for PCR amplification and allelic quantification by pyrosequencing were designed using the PyroMark® Assay Design SW 2.0 (Qiagen), as for the CpG assays (section 2.11.2). Amplification of the region of interest was carried out in a 20µl reaction using the AmpliTaq Gold Polymerase kit (Applied Biosystems, Foster City, CA; Cat# N8080241). The reaction was set up containing 2µl Buffer II (10X), 1.2mM $MgCl_2$, 0.5µM each of forward and reverse primers (see Table 2.4 for primer sequences), 0.5µl dNTP mix (final concentration of 0.25mM each of dATP, dCTP, dGTP & dTTP), 50U AmpliTaq Gold DNA polymerase, 1µl sample DNA (either gDNA or cDNA), and made up to the final volume with DEPC-treated H_2O. As before, optimal primer annealing temperatures were determine by running a PCR reaction with a temperature gradient (55 - 70°C). Amplification was carried out using the following thermocycling parameters on the G-Storm GS4 Thermal Cycler (G-Storm, Somerton, UK): initial denaturation at 95°C for 10 minutes, after which template was amplified for 45 cycles at 95°C for 15 seconds, 57.5°C for 30 seconds, and 72°C for 5 minutes, with final extension at 72°C for 5 minutes. Amplicons were visualised on the Odysey® Fc imaging system (LI-COR) after being run at 80V for 1 hour on a 3% (w/v in TAE buffer) agarose gel supplemented with 1% (v/v) EtBr.

The percentage of the risk allele (C) at the transcript SNP was then quantified by pyrosequencing exactly as before (see section 2.11.2), with the exception that a custom allelic quantification (AQ) assay was used as opposed to CpG assays. The sequence of the primer used for AQ pyrosequencing at rs7522061 is given in Table 2.4. For each patient sample, gDNA and cDNA were quantified in triplicate and samples for which any one of the triplicates differed from the other two by > 5% were excluded or repeated. In total, allelic quantification at rs7522061 was performed in 33 heterozygous patients.

2.12.3 Validation of pyrosequencing assays for allelic quantification

As with the CpG assays, the pyrosequencing assay for AQ had to be validated to ensure that allelic percentages could be accurately quantified across the entire range (0 – 100%). To validate the assay, a similar approach was used whereby allele mixes were generated with a range of risk allele proportions from 0 – 100%. These mixes were created by combining varying proportions of genomic DNA from patients who were either homozygous at the risk (C) or alternate (T) allele. Allelic proportions within these mixes were quantified by pyrosequencing as described above, and a standard curve produced by plotting the reported percentages from pyrosequencing against the expected percentages based on the prepared mixes to determine the accuracy of the assay. The standard curve generated for validation of this pyrosequencing assay is reported together with the results from the AEI analysis at *FCRL3* in section 5.5.

2.13 Drug-induced DNA hypo-methylation in lymphocyte cell lines

2.13.1 Culture of lymphocyte cell lines

Two lymphocyte cell lines were used to assess the effects of DNAm on gene expression at loci of interest. Jurkat cells (Clone E6-1, TIB-125™; American Type Culture Collection® (ATCC), Manassas, VA) are an acute T cell leukaemia cell line established from peripheral blood of a 14-year-old male patient. Conversely, Ramos (CRL-1596™; ATCC®) cells are a B cell line established from a 3-year-old male patient with Burkitt's lymphoma.

Both cell lines were cultured under identical conditions, maintained in RF10 medium (RPMI-1640 medium (Sigma-Aldrich; Cat# R0883) supplemented with 2mM glutamine, 100U/ml penicillin, 100μg/ml streptomycin, and 10% heat-inactivated foetal calf serum (FCS)) in a CELLSTAR® T75 culture flask (Greiner Bio-One, Stonehouse, UK; Cat# 658170) at 37°C with 5% CO_2. Cells were passaged every 2-3 days by transferring 1ml of the culture into 11ml of fresh RF10 medium. All cell culture procedures described in sections 2.13-2.15 were carried out in a class 2 microbiological safety cabinet (BioMAT; Contained Air Solutions, Middleton, UK).

To re-culture frozen cells, aliquots were removed from liquid nitrogen storage and placed on dry ice for five minutes to allow residual liquid nitrogen to evaporate. Cells were subsequently placed in a water bath at 37°C for 5 minutes until just thawed, and then transferred to a sterile 30ml Polystyrene Universal tube (Starlab; Cat# E1412-3010) containing 20ml of pre-warmed

RF10 medium. Cells were washed twice to remove dimethyl sulfoxide (DMSO) present in the freezing medium by centrifugation at 400 x g for 8 minutes, aspiration of the supernatant, and re-suspension in 20ml of pre-warmed RF10. Following the second wash step, cells were re-suspended in 5ml of warmed RF10 and counted using a Neubaueur counting chamber. Cells were seeded at a density of either 1 x 10^5 cells/ml (Jurkat cells) or 2 x 10^5 cells/ml (Ramos cells).

To freeze cells, freezing medium was prepared with RF10 containing 5% DMSO, and passed through a 0.2µm sterile syringe filter with a polyethersulfone membrane (VWR, Leicestershire, UK; Cat# 514-0073). Cells were centrifuged at 400 x g and 4°C for 5 minutes and re-suspended in freezing media at 2 x 10^6 cells/ml, with 1ml then transferred to a pre-cooled cryogenic vial (2.0ml, self-standing; Corning, Corning, NY; Cat# 431386). The vials containing cells were transferred to a CoolCell® (Corning), frozen at -80°C, and subsequently transferred to storage in liquid nitrogen.

2.13.2 5-Aza-2'-deoxycitidine treatment of cell lines

To assess the impact of global DNAm changes on the transcriptional activity at genes of interest, drug-induced hypo-methylation was achieved by treating lymphocyte cell lines with 5-Aza-2'-deoxycitidine (5-aza, Figure 2.11) – a cytidine analogue that inhibits the activity of the DNA methyltransferase enzymes (DNMTs). Jurkat cells and Ramos cells (see section 2.13.1) were selected as representative T- and B- cell lines and were treated with either 0.25µM or 0.5µM 5-aza for either 48 or 72 hours (Figure 2.12).

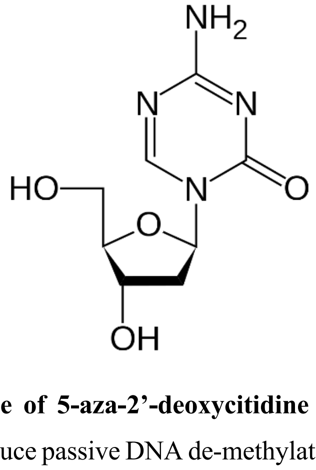


Figure 2.11: Chemical structure of 5-aza-2'-deoxycitidine (Decitabine). This drug is a chemical analogue of cytidine which can induce passive DNA de-methylation through inhibition of DNMT enzyme

Initially, 5mg 5-aza-2'-deoxycitidine (≥97%, Sigma-Aldrich) was dissolved in 1.10ml of 50% DMSO Hybri-Max™ (sterile-filtered, Sigma-Aldrich; Cat# D2650, diluted in DEPC-H_2O) with vortexing to prepare a stock solution of 20mM, which were stored at -20°C for < 1 week prior to use. Working stocks at 200µM and 100µM were prepared immediately prior to treatment of cells by diluting either 10µl of the 20mM stock in 990µl RF10 culture medium, or 5µl of stock in 995µl media. For use as a vehicle control, 10µl 50% DMSO was diluted in 990µl RF10 medium (DMSO control).

Jurkat and Ramos cells cultured in a T75 flask (Greiner Bio-One) with RF10 medium were counted and, for each condition, 2.5 x 10^5 cells in 3ml RF10 plated out in a Falcon® 6-well clear, flat-bottom culture plate (Corning; Cat# 353046) under sterile conditions. Each of the six treatment conditions (DMSO control, 0.25µM 5-aza, and 0.5µM 5-aza for both 48h and 72h treatment) was run in triplicate (Figure 2.12).

In order that all conditions could be harvested simultaneously, 5-aza was added to 72h time point cells immediately after plating on Day 0, with DMSO control added to all 48h cells at this point. Treatment of 48h time point cells with 5-aza began at 24h following plating out of cells, and cells for all conditions subsequently harvested at the 72h time point (Figure 2.12). For each of the treatments (DMSO control, 0.25µM 5-aza, 0.5µM 5-aza), 7.5µl of the prepared working reagent (either DMSO control, 100µM 5-aza, 200µM 5-aza) was added to the respective well. Treatments in all wells were replenished at each 24h time point. At the end of the treatment course, cells were harvested, lysed, and simultaneous extraction of genomic DNA and RNA performed exactly as described in section 2.2, with bisulphite-conversion of DNA (section 2.11.2), and reverse transcription of 600ng RNA (section 2.12.1).

Quantification of DNAm at the CpGs of interest (cg17134153 (*FCRL3*), cg21124310 (*ANKRD55*/*IL6ST*), and cg07522171 (*JAZF1*)) in bisulphite-converted DNA was performed exactly as described in section 2.12.2.

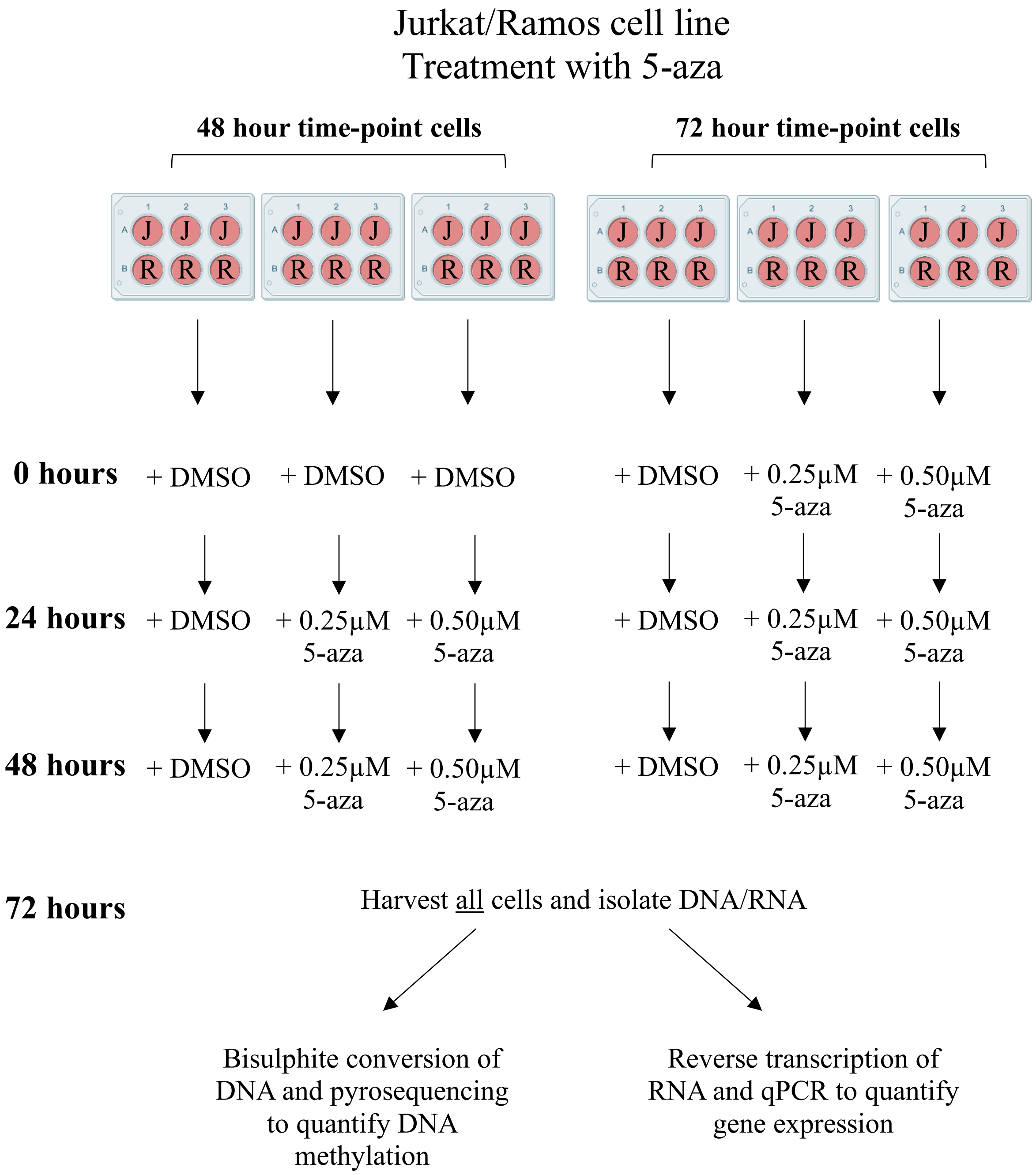


Figure 2.12: 5-Aza-2'-deoxycitidine (5-aza) treatment or Jurkat (J) and Ramos (R) cell lines. Cells were treated with either DMSO (vehicle control), 0.25µM 5-aza, or 0.50µM 5-aza for either 48 hours or 72 hours. All cells were cultured in parallel, with treatment of the 48-hour cells at the 24-hour time point to allow all cells to be harvested at the 72-hour time point. Culture plate logos from BioRender (https://biorender.com).

2.13.3 Quantitative PCR

Expression levels of gene of interest following 5-aza treatment were measured by quantitative PCR (qPCR) with TaqMan Gene Expression Assays (ThermoFisher). This system incorporates exon-spanning, sequence specific primers together with a sequence-specific probe that is 5' labelled with a FAM™ fluorescent dye (Applied Biosystems) and a non-fluorescent quencher.

For qPCR, all samples were run in triplicate, and those for which the cycle threshold (Ct) values of all three technical replicates differed by ≤ 0.5 were included in the analysis. Each amplification was set up in a 20µl reaction containing 10µl TaqMan™ Gene Expression Master Mix (2X, ThermoFisher; Cat# 4369016), 1µl TaqMan Gene Expression Assay (20X, Table 2.7), 4µl DEPC-H_2O, and 5µl cDNA (diluted 1:20 in H_2O). Cycling was performed using the AriaMX Real-time PCR system (Agilent), with initial denaturation at 95°C for 10 minutes, followed by 40 cycles of 95°C for 15 seconds and 60°C for 1 minute. Transcripts of interest were all normalised to transcript levels of the *HPRT1* and *GAPDH* housekeeper genes. Initially, the mean Ct for both housekeeper genes was calculated. This mean housekeeper Ct for a given condition was subsequently subtracted from the Ct value for the gene of interest to give the ΔCt, which was then converted to $2^{-\Delta Ct}$. The mean $2^{-\Delta Ct}$ across triplicates was then calculated for each sample.

Gene	TaqMan Assay ID	RefSeq Transcript (Exon Boundary)
ANKRD55	Hs00902590_m1	NM_024669.2 (11 - 12)
FCRL3	Hs00364720_m1	NM_001320333.1 (10 – 11)
GAPDH	Hs02758991_g1	NM_001256799.2 (6 – 7)
HPRT1	Hs02800695_m1	NM_000194.2 (2 – 3)
IL6ST	Hs00174360_m1	NM_001190981.1 (13 – 14)
JAZF1	Hs00697777_m1	NM_175061.3 (4 – 5)

Table 2.7: TaqMan Gene Expression Assays used for quantifying transcript levels of genes of interest. The TaqMan assay ID represents a unique ID provided for each assay from ThermoFisher. The RefSeq transcript refers to the unique RefSeq ID for the transcript targeted by the assay, together with the exons in the transcript across which the assay spans.

2.14 Luciferase reporter assay

Gene reporter assays were performed with the aim of validating the regulatory potential of SNPs and/or CpG sites identified by the meQTL analysis of patient samples. The *FCRL3* and *JAZF1* regions were selected for luciferase reporter assays, given that the cis-CpGs found to likely mediate transcriptional regulation in these cases map to the gene promoter regions. The pCpGL-basic CG-free plasmid[299] (Figure 2.13) was chosen for this analysis as this plasmid allows for the effect of CpG DNA methylation on reporter activity to be measured. In the instance of the *FCRL3* promoter, the regulatory SNP (rs7528684) and meQTL-associated cis-CpG sites (cg17134153 & cg01045636) were in sufficient proximity to allow both to be cloned into the vector. However, for the *JAZF1* promoter, the CpG was cloned into the vector in isolation, given that the regulatory SNP was at too great a distance to allow the two to be cloned in the same insert.

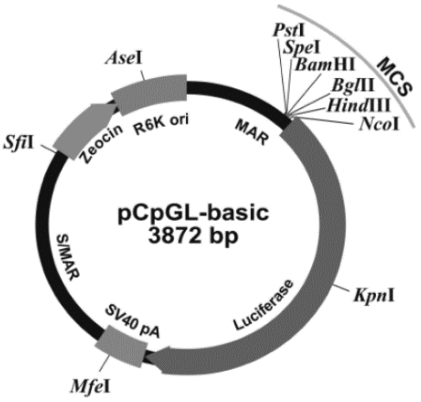


Figure 2.13: pCpGl Basic Vector. The CpG-free vector[299] used for reporter gene assays harbours a luciferase gene to give a readout of transcriptional activity. The multiple cloning site (MCS) upstream of the luciferase gene allows for the promoter region of interest to be cloned into the vector. The plasmid also incorporates a Zeocin™ antibiotic resistance gene to allow for selection of transformed cells on selective media. The presence of the R6K origin of replication allows for plasmid replication in prokaryotes, and the SV40 poly-A sequence signifies the site of transcriptional termination.

2.14.1 Amplification of regions of interest

PCR primers spanning the region of interest were designed using Primer3 software[300] (version 4.1.0; http://primer3.ut.ee), with restriction sites for the SpeI and NcoI (*FCRL3* promoter) or SpeI and HindIII (*JAZF1* promoter) enzymes added to the 5' end of the primers to enable cloning into the CpG-free pCpGl-basic vector[299] (Figure 2.13) in the correct orientation. Primers were optimised as described in section 2.11.2 using patient genomic DNA. In the case of the *FCRL3* promoter, as the regulatory activity of both alleles of the rs7528684 SNP were to be assessed, template DNA from a patient heterozygous at this SNP was used as template for amplification, enabling the generation of inserts harbouring both allele copies.

Regions of interest were amplified using the Phire Hot Start II DNA polymerase (ThermoFisher; Cat# F122S) for high-fidelity amplification. A reaction was set up consisting of 4µl Phire Reaction Buffer (5X), 0.2µM each of forward and reverse primers (Table 2.8), 0.4µl dNTP mix (final concentration of 0.2mM of each), 0.5µl template DNA (~50ng), and 0.4µl Phire Hot Start II DNA Polymerase, and made up to a 20µl volume with DEPC-H_2O. Amplification was performed using the Alpha Cycler 1 (PCRmax) under the following cycling parameters: 98°C for 30 seconds, 35 cycles of 98°C for 5 seconds, 69°C for 5 seconds, and 72°C for 15 seconds, with a final extension step of 72°C for 1 minute. To increase the possibility of obtaining amplicons without de novo mutations introduced during amplification, each PCR was performed in duplicate, with the two reactions combined prior to clean up. Clean-up of PCR products was performed using a QIAquick Gel Extraction Kit (Qiagen; Cat# 28704), following the manufacturer's protocol for clean-up from enzymatic reactions, with elution in 40µl EB buffer.

Primer	Sequence (5' ➞ 3')	Amplified Region (hg19)
FCRL3 (NcoI)	GGGGCCATGGCTCACTTCCCATCCCTTGCT	chr1:157,670,120 – 157,671,093
FCRL3 (SpeI)	GGGGACTAGTAGAACAGTTAGAGGTGCGGG	
JAZF1 (SpeI)	GGGGACTAGTACCCCTGGACCTTTCAACAA	chr7:28,218,982 – 28,218,384
JAZF1 (HindIII)	GGGGAAGCTTCCAAAACTTGCCCAGCTCTT	

Table 2.8: Primers used for amplification of promoter regions of interest prior to cloning into the pCpGl-Basic vector. Underlined are the restriction enzyme recognition sites to enable digestion and ligation into the plasmid.

2.14.2 Digestion of insert and vector

Reactions were set up in parallel to digest the PCR amplicon inserts and the vector with the appropriate restriction enzymes. To digest inserts, a 40µl reaction was set up with 30µl purified PCR product, 4µl CutSmart Buffer® (10X, New England Biolabs; Cat# B7204S), 1.5µl each of either SpeI-HF® and NcoI-HF® (*FCRL3*, New England Biolabs; Cat# R3133S/R3193S) or SpeI-HF® and HindIII-HF® (*JAZF1*, New England Biolabs; Cat# R3133S/R3104S), and 3ul DEPC-H_2O. The pCpGl-basic vector was also digested in a 40µl reaction with 15µl vector (~1µg), again with 4µl CutSmart® Buffer and 1.5µl each of the appropriate restriction enzymes. Digestion was performed in an incubator at 37°C for 16 hours.

Following digestion, the digested inserts and vector were electrophoresed on an agarose gel (1.5% w/v in TAE buffer, 1.5% v/v EtBr) at 100V for 90 minutes, after which the DNA bands of interest were extracted using a scalpel. The DNA was purified from the gel band using the QIAquick Gel Extraction Kit (Qiagen), this time following the instructions for clean-up from low-melt agarose gels, and purified products eluted in 50µl Buffer EB.

2.14.3 Ligation of insert into vector

Following extraction of digested insert and vector from the agarose gel, DNA quantification was performed using the Nanodrop ND-1000 Spectrophotometer to calculate the volume for use in the ligation reaction. To ensure that the insert was in excess of the vector, and thus minimise the number of instances of vector re-ligation, 100ng of insert (0.97Kb (*FCRL3*) / 0.60Kb (*JAZF1*)) was ligated with 200ng vector (3.87Kb), so that the insert was in ~2-3 fold excess. A 20µl reaction was set up with 2µl T4 DNA Ligase Buffer (10X, New England Biolabs), 200ng vector DNA, 100ng insert DNA, 1µl T4 Ligase (New England Biolabs; Cat#

B0202S) and made up to a final volume with DEPC-H_2O. A re-ligation reaction was set up in parallel, replacing the insert with H_2O, in order to assess the efficiency of the ligation of the insert into the digested plasmid. Ligation was performed for 2 hours at room temperature, following which the ligase enzyme was heat inactivated at 65°C for 10 minutes, and the reaction kept on ice until bacterial transformation.

2.14.4 Preparation of LB agar plates and LB broth with antibiotic

To prepare LB media for bacterial selection, 10.3g LB Broth (Lennox, Sigma; Cat# L3022) was made up to 500ml using distilled H_2O (dH_2O) and autoclaved at 121°C. The LB broth was stored at room temperature until use. Prior to use, Zeocin™ (100mg/ml in solution, Invivogen, San Diego, CA; Cat# ant-zn-05) was added to the LB broth at a ratio of 1:3000 v/v, to give a final antibiotic concentration of 33µg/ml.

LB agar antibiotic plates were prepared by mixing 10.3g LB Broth (Lennox, Sigma) and 7.5g Bacto™ Agar (BD Biosciences, San Jose, CA; Cat# 90000-760) and made up to 500ml with distilled H_2O prior to autoclaving at 121°C. The media was allowed to cool to 55°C in a water bath, at which point 166µl Zeocin™ antibiotic was added to a final concentration of 33µg/ml and poured into culture dishes under sterile conditions. Plates were set at room temperature, dried at 37°C in an incubator overnight, and then stored at 4°C until use.

2.14.5 Bacterial transformation, colony picking, and plasmid preparation

ChemiComp GT115 *E. coli* cells (Invivogen; Cat# gt115-11) were removed from storage at -80°C and thawed on ice. 22µl was added to separate 1.5ml Eppendorf tubes for each ligation reaction and placed on ice. 1.5µl of the ligation reaction was added to the appropriate tube containing the cells, gently stirred with a pipette tip, and left on ice for 30 minutes. The cells were heat shocked for 30 seconds at 42°C on a heat block and immediately placed back on ice for 5 minutes, following which 200µl of room temperature S.O.C Medium (ThermoFisher; Cat# 15544034) was added to each tube. The tubes were then placed in an incubator at 37°C with rocking at 225rpm for 1 hour. Two separate LB agar plates supplemented with Zeon™ (33µg/ml) per condition (both vector + insert ligations (*FCRL3*/*JAZF1*), as well as vector re-ligations) were pre-warmed. Plates were streaked with either 30µl or 90µl of transformed cells using a sterilised glass spreader and incubated at 37°C for 16 hours.

Following this incubation, eight individual colonies were picked for each insert using a sterile pipette tip and inoculated in 3ml LB broth with 33µg/ml Zeocin™ in a sterile 15ml polystyrene Falcon tube (ThermoFisher; Cat# 10263041). Each tube inoculated with a single colony was incubated at 37°C for 16 hours with rocking at 225rpm, and 1.5ml transferred to a 1.5ml Eppendorf tube. Glycerol stocks from each colony were prepared by combining 300µl of the bacterial culture with 300µl glycerol (BioUltra, anhydrous, ≥99.5%; Sigma-Aldrich; Cat# 49767), and stocks stored at -80°C.

Bacterial cultures were then miniprepped for plasmid genotyping and sequencing. Bacterial cells were first harvested by centrifugation at 8,000 x g for 4 minutes and the supernatant carefully aspirated. Each bacterial pellet was re-suspended in 100µl P1 + RNase A solution (50mM Tris-HCl, 10mM EDTA, 100µg/ml RNase A) using a vortex. 200µl of P2 solution (200mM NaOH, 1% SDS) was added to each tube and mixed by inverting five times, with cell lysis allowed to progress for 4 minutes, after which 150µl of P3 solution (3M potassium acetate, pH 5.5) was added, and again mixed by inverting the tube. The lysate was centrifuged at 13,000 rpm for 5 minutes to remove cellular debris, and the clear supernatant transferred to a new 1.5ml Eppendorf tube. The centrifugation step was repeated once more, and ~450µl cell lysate added to a sterile 1.5µl tube. To precipitate the plasmid DNA, 1ml 100% ethanol was added to each tube containing the clear supernatant, incubated at 80°C for 20 minutes, followed by centrifugation at 13,000rpm for 10 minutes. The supernatant was carefully aspirated from the DNA pellet and residual ethanol was allowed to evaporate by air drying the DNA pellet at room temperature for 15 minutes, after which the pellet was re-suspended in 80µl DEPC-H_2O and stored at -20°C.

2.14.6 Plasmid genotyping and sequencing

To identify colonies that had been successfully transformed with pCpGl-basic vector containing the correct insert, a genotyping digest reaction was set up with 1µl CutSmart® Buffer, 0.25µl of each restriction enzyme (SpeI-HF® & NcoI-HF® for *FCRL3*, SpeI-HF® & HindIII-HF® for *JAZF1*), 4µl miniprepped plasmid DNA, and 4.5µl DEPC-H_2O. The reaction was incubated at 37°C for 2 hours and run on an agarose gel (2% w/v in TAE Buffer, 1.5% EtBr) at 100V for 90 minutes, with the gel visualised using the Odysey® Fc imaging system. Clones harbouring vector with the ligated insert were identified by the presence of bands at ~3.7Kb (pCpGl-basic) and either ~970bp (*FCRL3*) or ~600bp (*JAZF1*). To check the size of the inserts, digested vectors were run alongside the respective amplicon from PCR.

To confirm insert orientation, identify potential de novo mutations and, in the case of the *FCRL3* promoter, perform SNP haplotyping, plasmids from clones which were found to harbour the vector and insert were diluted 1:1 in DEPC-H_2O and sent for Sanger Sequencing (service provided by Source Bioscience, Nottingham, UK). Sequencing results were analysed using ‘A plasmid Editor’ (ApE) software (version 2.0) and multiple sequence alignment performed using ClustalX[301] (version 2.1) to identify variable positions between clones.

2.14.7 Plasmid maxiprep and in vitro methylation

Following confirmation of plasmid sequences, the appropriate clones were streaked on LB agar plates from glycerol stocks and cultured for 16 hours at 37°C. Colonies were then inoculated in 5ml LB broth with Zeocin (33µg/ml) in a sterile 15ml polystyrene Falcon tube (ThermoFisher), and incubated at 37°C at 225rpm for 6 hours. This culture was then transferred to a 1L conical flask containing 200ml LB broth with Zeocin (33µg/ml) and cultured for a further 16 hours (37°C, 225rpm). Following this incubation period, each culture transferred to four sterile polypropylene 50ml centrifuge tubes (Cole-Parmer; Cat# WZ-06344-29) and centrifuged at 3,500 x g for 20 minutes to pellet the cells. The supernatant was aspirated from the cell pellet and plasmid DNA extracted using the PureYield™ Plasmid Maxiprep System (Promega, Hampshire, UK; Cat# A2939) with the QIAvac 24 Plus vacuum manifold (Qiagen; Cat# 19413) following the manufacturer’s protocol. The harvested cells in separate centrifuge tubes from the same clone were recombined by resuspension in 12ml Cell Resuspension Solution (Promega) prior to plasmid extraction.

40µg plasmid DNA was either methylated or mock-methylated *in vitro* in four separate 120ul reactions per plasmid (10µg DNA per reaction) using the M.SssI CpG methyltransferase enzyme (New England Biolabs; Cat# M0226M). Along with the plasmids containing the inserts of interest, the empty pCpGl-basic plasmid was also methylated/mock-methylated to include as a negative control for normalising luciferase activity. Plasmid DNA (4 reactions per plasmid performed in parallel) was combined with 12µl NEB Buffer 2 (10X, New England Biolabs), 25µl S-adenosylmethionine (SAM, 1.6mM, New England Biolabs), 2µl of either M.SssI enzyme (for methylated plasmids, 20,000U/ml) or DEPC-H_2O (mock-methylated), and then made up to the final volume with DEPC-H_2O. The reactions were incubated for 16 hours at 37°C in a water bath, following which they were replenished with 2µl NEB Buffer 2, 4µl 1.6mM SAM, 2µl of either MSss.I (20,000U/ml) or DEPC-H_2O, and 12µl DEPC-H_2O, with incubation at 37°C for a further 4 hours. Enzymes were heat-inactivated at 65°C for 20 minutes and the plasmid DNA purified using the Wizard® SV Gel and PCR Clean-Up System (Promega; Cat# A9281) following the manufacturer’s instructions, with elution in 40µl nuclease-free H_2O. The four reactions per plasmid run in parallel for each condition were combined following clean-up for downstream transfections.

To confirm complete methylation, both methylated and mock-methylated plasmids were digested with the methylation-sensitive enzyme HpaII, for which restriction sites were present in both the *JAZF1* and *FCRL3* inserts. Plasmid DNA (4µl) was combined with 1µl CutSmart® Buffer, 0.25µl HpaII (New England Biolabs; Cat# R0171S), and 4.75µl DEPC-H_2O, with incubation at 37°C for 2 hours. Digested products were separated on an agarose gel (2% w/v, 1.5% EtBr) at 100V for 90 minutes and the banding pattern visualised using the Odysey® Fc imaging system. After methylation was confirmed, plasmids were stored at -20C.

2.14.8 Culture of HEK-293T cell line

HEK-293T cells are an adherent human cell line derived from an embryonic kidney cell (CRL-3216™, ATCC), and were selected as a suitable transfection host for reporter gene assays. Cells were cultured in a T75 flask with Dulbecco's modified eagle medium (DMEM, ThermoFisher; Cat# 31053028) supplemented with 10% FBS, 2mM L-glutamine, 200IU/ml Penicillin, and 200µg/ml Streptomycin. Cultured cells were maintained at 37°C with 5% CO_2 and passaged when ~80-90% confluent (every 2-3 days) by transferring 2ml of cells to 8ml of fresh medium in a T75 flask. As these cells are adherent, the cells were treated with trypsin before transferring. The medium was removed from the cells and the surface of the flask washed in 10ml Dulbecco's phosphate buffered saline (DPBS, without Mg^{2+} and Ca^{2+}, Sigma-Aldrich; Cat# D8537). Adherent cells were detached from the flask with 1.5ml 0.05% trypsin-EDTA (ThermoFisher; Cat# 25300), placed in the incubator at 37°C for 5 minutes, and the flask tapped to dislodge cells. The trypsin was deactivated by the addition of 8.5ml supplemented DMEM.

2.14.9 Transfection of a HEK-293T cell line

HEK-293T adherent cells were transfected using the nonliposomal transfection reagent FuGENE® HD (Promega; Cat# E2311). Following trypsinisation (see section 2.14.8), cells were transferred to a sterile 50ml Falcon tube (ThermoFisher). The cells were then centrifuged at 15,000 x g for 5 minutes and the medium aspirated, following which they were re-suspended in 6ml medium and counted as before (see section 2.13.1). Cell density was adjusted to 2 x 10^5 cells/ml, and 100µl added to wells of a Corning™ Costar™ 96-well flat bottom culture plate (Fisher Scientific; Cat# 10695951). Prior to transfection, cells were incubated at 37°C for 24 hours. Each well was transfected with 10ul of transfection mix containing 50ng pCpGL-basic plasmid (section 2.14.7), 6ng pRL-TK Renilla Luciferase Control Vector (Promega; Cat# E2231) and 0.15µl FuGENE® HD (Promega) diluted in DMEM (without FCS or antibiotic). Precisely 15 minutes after the addition of FuGENE® HD to the transfection mix, 10µl was added to each well containing 100µl HEK-283T cells (2 x 10^4 cells/well). Transfected cells were then incubated at 37°C with 5% CO_2 for 24 hours before lysing.

2.14.10 Transfection of a Jurkat cell line

Plasmid transfection of Jurkat cells was carried out using the Neon™ Transfection System (ThermoFisher) in conjunction with Neon™ Transfection System 10µl kit reagents (ThermoFisher; Cat# MPK1025). All plasmids were concentrated to 3µg/µl using a DNA120 SpeedVac® System (ThermoFisher) and 4µl combined with 4µl pRL-TK Renilla Luciferase Control Reporter Vector (60ng/µl) to give a final concentration of 1.5µg/µl and 30ng/µl respectively, with 7µl of this plasmid mix then transferred to a 200µl PCR tube (Starlab).

A 24-well flat-bottom culture plate was prepared by adding 500µl of antibiotic-free RF10 media to each well and placed in the incubator to pre-warm for 1 hour. Jurkat cells were cultured in a T75 flask (see section 2.13.1), and the media changed 24 hours prior to transfection. Cells were counted, with12.1 x 10^6 transferred to a universal tube and centrifuged at 500 x g for 8 minutes. The media was aspirated, cells washed by re-suspending in 10ml DPBS (without Mg^{2+} and Ca^{2+}, Sigma-Aldrich) and again centrifuged as before. The DPBS wash step was repeated once more before cells were re-suspended in 550µl Buffer R (ThermoFisher), giving a cell density of 22 x 10^6 cells/ml. 63µl of this prepared cell suspension was transferred to each tube containing the 7µl of the prepared plasmid mixes. Electroporation in a 10µl volume (~2 x 10^5 cells) was carried out using the Neon™ Transfection System according to the manufacturer's protocol, with a program of 1400 voltage pulse, 10ms pulse width, and 3 pulses. Each condition was transfected in triplicate and dispensed into separate wells of the 24-well plate containing 500µl pre-warmed antibiotic-free media. Transfected cells were incubated at 37°C with 5% CO_2 for 24 hours before lysing.

2.14.11 Quantification of reporter gene activity

Twenty-four hours following transfection, luciferase activity was measured using the Dual-Luciferase® Reporter Assay System (Promega; Cat# E1910). Cells were harvested by centrifugation at 500 x g for 8 minutes and aspiration of the supernatant. The cells were washed with PBS and re-suspended in either 100µl (Jurkat) or 30µl (HEK-293T) 1X passive lysis buffer (Promega) and agitated at room temperature for 30 minutes. 20µl of the cell lysate was transferred to separate wells of a 96-well black polystyrene plate with white wells, and both Firefly and Renilla luciferase activity read using the GloMax®-Multi Detection System (Promega) with Dual-Luciferase® Reporter Assay reagents, following the manufacturer's protocol (100µl each of Luciferase Assay Reagent II and Stop & Glo® Reagent, 2 second delay, with a 10-second measurement period). For each transfection reaction, the firefly luciferase activity, indicative of promoter activity for the region of interest, was normalised to the Renilla activity to account for transfection efficacy. Subsequently, the relative luciferase activity for each insert was normalised to the intensity value for the respective methylated/unmethylated empty vector.

2.15 Site-specific DNA-demethylation using a CRISPR-dCas9 system

To assess the potential impact of inducing site-specific de-methylation at *FCRL3*, *ANKRD55* and *JAZF1* CpG sites of interest in a genomic context, a clustered regularly interspaced short palindromic repeats (CRISPR) – nuclease-deficient Cas9 (dCas9) system was employed. The CRISPR-Cas9 genome-editing system was developed by leveraging a bacterial immune mechanism that involves small RNAs (termed CRISPR RNAs (crRNAs)) that recognize DNA from foreign organisms and target the Cas9 endonuclease enzyme to these sequences[302]. This system has enabled targeted cellular genome-editing *in vitro* and *in vivo*[303]. By substituting the active Cas9 enzyme for one that has been edited to become catalytically inactive (dCas9), this system can be leveraged to target a range of effector proteins, including those that modify DNAm, to specific regions in the genome[304].

This system was used here in an attempt to target the catalytic domain of TET1, an enzyme involved in active demethylation, to the loci of interest harbouring cis-CpGs associated with risk loci. The pPlatTET-gRNA2[304] plasmid (Addgene, Watertown, MA, Plasmid #82559) encodes multiple copies of the GCN4 peptide, the nuclease-deficient dCas9, an anti-GCN4 antibody single chain variable fragment (scFv), a copy of the TET1 enzyme which catalyses DNA de-methylation, as well as the green fluorescence protein (Figure 2.14A). Rather than inserting the coding sequence of the guide RNA (gRNA) of choice into the vector itself, a transient transfection approach was adopted with the use of synthetic gRNAs. This involved duplexing a sequence-specific crRNA with a trans-activating crRNA (tracrRNA) to form a guide RNA (gRNA) that can deliver the dCas9 to the intended site (Figure 2.14B). When the gRNA and pPlatTET-gRNA2 plasmid are co-transfected into the host cell, association of the dCas9 with the gRNA localises this fusion protein to the target region of the genome (Figure 2.14C). Subsequent binding of the svFC fragment to the GCN4 peptides delivers the TET1 catalytic domain to the targeted loci, thus inducing de-methylation at proximal CpG sites (Figure 2.14C). The presence of the GFP tag allows transfected cells to be identified by fluorescence.

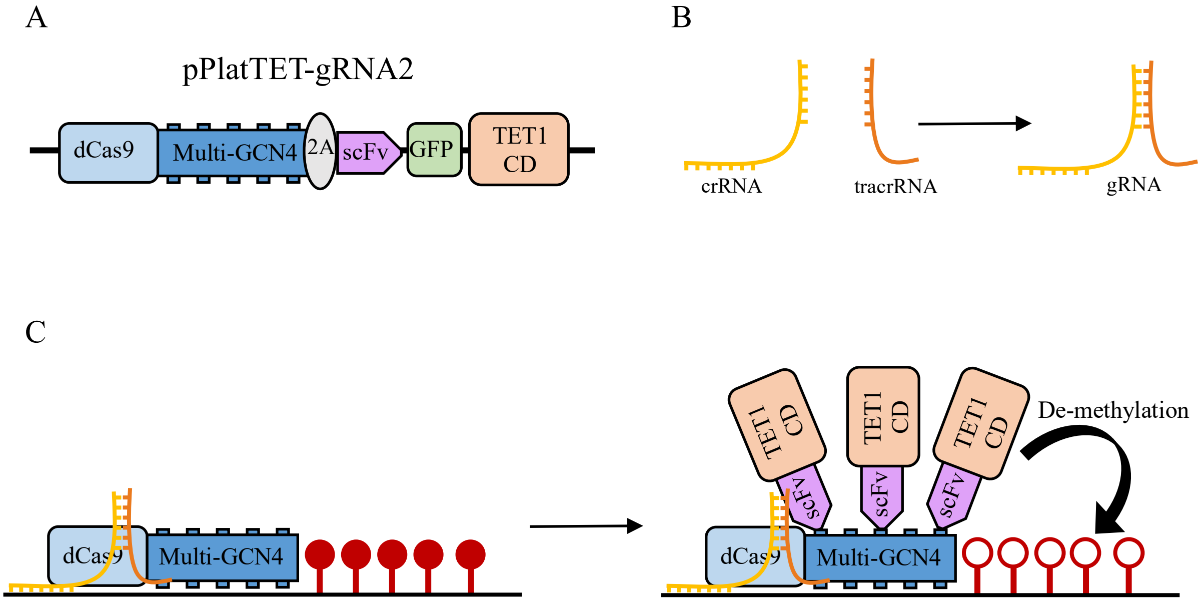


Figure 2.14: mechanism of targeted DNA de-methylation using the pPlatTET-gRNA2 plasmid. A) The pPlatTET-gRNA2 plasmid encodes a copy of the nuclease-deficient Cas9 (dCas9) enzyme, which lacks endonuclease activity, fused to multiple copies of the GCN4 peptide, with a 22 amino acid linker (2A). Also encoded is the catalytic domain (CD) of the TET1 enzyme which induces active de-methylation, fused to the single-chain variable fragment (scFv) of the anti-GCN4 antibody. The presence of a GFP tag allows transfected cells to be identified and sorted. B) Complexing a site-specific CRISPR RNA (crRNA) with a fluorescently labelled trans-activating crRNA (tracrRNA) forms a gRNA duplex which can target the dCas9 to the target sequence. C) Following co-transfection of cells with gRNA targeting the region of interest and the pPlatTET-gRNA2 plasmid, the dCas9 and gRNA form a complex which is targeted to the target site. As the scFv anti-GCN4 regions bind the multi-GCN4 peptide, multiple copies of TET1 are recruited to the region, inducing site-specific DNA de-methylation at neighbouring CpG sites. Figure adapted from Morita et al. (2016)[304].

2.15.1 Plasmid isolation

LB agar plates with antibiotic were prepared as before (section 2.14.4) with the addition of 50µg/ml kanamycin in place of Zeocin™. The agar stab containing cells transformed with pPlatTET-gRNA2 was streaked onto pre-warmed plates using a P20 pipette tip and incubated at 37°C for 16 hours. Colonies were picked and cultured as described before (section 2.14.5), with plasmids again extracted with the PureYield™ Plasmid Maxiprep System (section 2.14.7).

2.15.2 Design of crRNAs and gRNA duplexing

Synthetic gRNAs were designed in a 500bp region encompassing either the *FCRL3*, *ANKRD55*, or *JAZF1* CpG sites of interest using the IDT gRNA design tool (Integrated DNA Technologies (IDT)). gRNAs were selected based on their on-target and off-target scores, as well as the proximity of the annealing site to the CpG. Alt-R® CRISPR-Cas9 crRNAs targeting the sequence of interest were synthesized (Integrated DNA Technologies) with a 3' sequence of 16 nucleotides that is complimentary to the tracrRNA molecule, allowing the crRNA and tracrRNA to be hybridized to produce a gRNA duplex prior to transfection (Table 2.9). Two separate gRNAs were designed to target each region and compare efficacy (gRNA sequences are listed in Table 2.9). Each of the crRNAs (10nmol) was suspended in 100µl Nuclease-Free Duplex Buffer (Integrated DNA Technologies; Cat# 11-01-03-01) to a concentration of 100µM. Alt-R® CRISPR-Cas9 Negative Control crRNA #1 (Integrated DNA Technologies, 2nmol; Cat# 1072544) was suspended in 20µl Duplex Buffer. The fluorescently labelled Alt-R® CRISPR-Cas9 tracrRNA, ATTO™ 550 (Integrated DNA Technologies, 20nmol; Cat# 1072533) was suspended at 100µM in 200µl Duplex Buffer. 20µl of each crRNA (100µM) was combined with 20µl tracrRNA (100µM) in a 200µl PCR tube, heated at 95°C for 5 minutes, and allowed to cool at room temperature for 30 minutes to allow the gRNA duplex to form. The gRNAs were then stored on ice until transfection of cells. Alt-R® CRISPR-Cas9 Electroporation Enhancer (Integrated DNA Technologies, 10nmol; Cat# 1075916) was suspended at 100µM in 100µl Duplex Buffer to produce a stock solution, which was subsequently diluted to a working solution at 10.8µM with Duplex Buffer. 15µl of each duplexed gRNA was combined with 15µl of the pPlatTET-gRNA2 plasmid (3µg/ml) and stored on ice. 15µl of the pPlatTET-gRNA2 plasmid (3µg/ml) was also combined with 15µl Duplex Buffer for single transfection to allow the GFP gate to be set for cell sorting, and similarly 15µl of the negative control gRNA combined with15µl Duplex Buffer to set the ATTO™ 550 gate.

	gRNA	**Sequence**
FCRL3	cg17134153 [1]	/AltR1/**GUGAGUAGAUGGGCUAUAAU**GUUUUAGAGCUAUGCU/AltR2/
	cg17134153 [2]	/AltR1/**CGACUUAUCUCCAAGAAGGA**GUUUUAGAGCUAUGCU/AltR2/
ANKRD55	cg21124310 [1]	/AltR1/**AGUAGAAUUCUCCCGUCACC**GUUUUAGAGCUAUGCU/AltR2/
	cg21124310 [2]	/AltR1/**CUCCAUAACUUGGACACAGA**GUUUUAGAGCUAUGCU/AltR2/
JAZF1	cg07522171 [1]	/AltR1/**UGAUUAUCUCCUAUUCCGGA**GUUUUAGAGCUAUGCU/AltR2/
	cg07522171 [2]	/AltR1/**UCUGGAUUAGACAGCCCCAU**GUUUUAGAGCUAUGCU/AltR2/

Table 2.9: CRISPR RNAs (crRNA) designed to target regions of interest. crRNAs were designed for site-specific DNA de-methylation with the pPlatTET-gRNA2 plasmid. Highlighted in bold are the target sequences that map adjacent to the CpGs of interest, and underlined is the sequence complimentary to the tracrRNA, which allows formation of the gRNA complex (crRNA + tracrRNA). The 5' (AltR1) and 3' (AltR2) chemical modifications protect the RNA from RNases present in the cell (Integrated DNA Technologies)

2.15.3 Transfection of a Jurkat cell line for targeted DNA de-methylation

Co-transfection of Jurkat cells with the pPlatTET-gRNA2 plasmid and gRNA duplexes was performed using the Neon™ Transfection System (ThermoFisher) in conjunction with Neon™ Transfection System 100µl kit reagents. Jurkat cells (clone E6.1) were grown in RF10 media in a T175 culture flask (Cat# 660160) and the media changed 24 hours prior to electroporation. The cells were centrifuged at 500 x g for 8 minutes, the supernatant aspirated, and cells re-suspended in 20ml RF10. A total of 50 x 10^6 cells were transferred to a 30ml universal tube and centrifuged at 500 x g for 8 minutes. The supernatant was aspirated, and the cells washed once with sterile PBS (without Mg^{2+} and Ca^{2+}, Sigma-Aldrich). Cells were then re-suspended at a density of 30 x 10^6 cells/ml in 1667µl Buffer R (ThermoFisher). 20µl of each the gRNA-plasmid mix or gating control prepared in section 2.15.2 was transferred to a 600µl Eppendorf tube and combined with 40µl of the Alt-R® CRISPR-Cas9 Electroporation Enhancer (10.8µM), and 180µl of the Jurkat suspension in Buffer R. Each 100µl electroporation reaction therefore consisted of ~2.25 x 10^6 Jurkat cells, 12.5µg pPlatTET-gRNA2, 2µM gRNA duplex, and 1.8µM Alt-R® CRISPR-Cas9 Electroporation Enhancer. Cells were transfected using the Neon™ Transfection System following the manufacturer's protocol, with a program of 1400V pulse voltage, 10ms pulse width, and 3 pulses. For each condition, two 100µl transfections were performed and transferred to a T25 culture flask containing 5ml pre-warmed antibiotic-free RF10 media. Transfected cells were cultured at 37°C with 5% CO_2 for 24 hours before cell sorting (section 2.15.4).

2.15.4 Cell sorting by flow cytometry and re-culture

Twenty-four hours following transfection, cells were transferred to a universal tube, centrifuged at 500 x g for 8 minutes, and the supernatant aspirated. The cells were then washed once in sterile DPBS (without Mg^{2+} and Ca^{2+}) and suspended in 500µl sterile DPBS with 2% FCS. Prior to sorting, the cells were passed through a 30µM CellTrics® filter (Sysmex, Milton Keynes, UK; Cat# 25004-0042-2316) to remove cell clumps and obtain a suspension of single cells. Cells were sorted using the BD FACSAria™ Fusion Cell Sorter (BD Biosciences) with the assistance of the Newcastle Flow Cytometry Core Facility. Cells transfected with either pPlatTET-gRNA2 plasmid or negative control gRNA alone were used to set the GFP (emission 509nm) and ATTO™ 550 (emission 575nm) gates respectively. Double-positive cells were then sorted into a 12-well plate containing 1ml RF10 without antibiotic. Immediately after sorting, 10,000 cells per condition were split into separate wells of a 48-well plate and the volume in each well made up to 500µl with antibiotic-free RF10. Cells were then incubated at

37°C with 5% CO_2 for 48 hours, after which cells were harvested by centrifugation at 500 x g for 8 minutes.

2.15.5 Nucleic acid extraction and DNAm quantification

Cells were lysed by adding 75µl Buffer RLT Plus (Qiagen, supplemented with 1% v/v β-mercaptoethanol) and vortexing, followed by passing the lysate through a QIAshredder spin column (Qiagen). DNA and RNA were then extracted using the AllPrep® DNA/RNA Micro Kit (Qiagen; Cat# 80284) following the manufacturer's instructions. DNA was eluted in 30µl Buffer EB, with the eluate passed through the column a second time to increase DNA yield. DNA was bisulphite converted as in section 2.11.1, albeit with elution in 23µl of elution buffer given the lower amount of input DNA (100ng). Pyrosequencing was performed as previously described in section 2.11.2, albeit the amount of template DNA was 15ng.

Chapter 3. Systematic Analysis of Lymphocyte DNA Methylation in Early Arthritis

3.1 Introduction

To date, a number of studies have compared the PBMC lymphocyte DNA methylomes of patients with RA to those of healthy individuals in an attempt to describe disease-associated changes[218, 221, 222, 305, 306]. One particular limitation of this approach is that distinguishing disease-specific changes from those that result from systemic inflammation can be a challenge. This chapter will therefore describe a comprehensive genome-wide analysis of DNAm signatures in CD4$^+$ T cells and B cells in the context of early arthritis, with the aim of deciphering whether epigenetic modifications occur specifically in RA. In particular, a control group comprising patients with non-RA arthropathies was selected in an attempt decipher signatures associated with the RA aetiological process, as opposed to those that reflect more general autoimmune responses or tissue damage of the joint. In particular, this study focusses on early arthritis patients who were yet to receive treatment with DMARDS, which target immune cells and secreted cytokines. Ordinarily, such treatments may modify patterns of DNAm, as has been described following treatment with methotrexate[307].

Cell-specific DNAm data at ~850,000 CpGs were quantified in lymphocytes from early RA patients, as well as disease controls, using the Illumina® MethylationEPIC BeadChip microarray. In the first instance, quality control procedures were performed to identify problematic samples or potential mix-ups, and to remove specific probes. A systematic evaluation of the performance of various normalisation methods when applied to the dataset was then carried out. Following this, an EWAS study design was employed to identify any CpG sites that were differentially methylated between early RA patients and the cohort of disease controls. As well as analysing potentially differentially methylated positions (DMPs), extended *regions* displaying differential patterns of methylation between the RA and control groups were sought (differentially methylated regions, DMRs). The analysis was then extended to assess differential variability in DNAm between comparator groups (differentially variable positions; DVPs). Any CpGs found to exhibit differential methylation between RA cases and non-RA controls were then subject to a pathway analysis with the aim of revealing biological pathways that may be perturbed by disease-associated methylation changes. Finally, potential relationships between disease diagnosis and biological age were investigated by assessing the DNA methylation clock.

3.2 Patient recruitment and cohort characteristics

Patients with suspected inflammatory arthritis were recruited from the NEAC at baseline visits following clinical assessment[308]. As described in Chapter 2.1, a cohort of early RA patients, diagnosed according to the ACR/EULAR 2010 classification criteria[64], were recruited alongside a heterogeneous disease control comparator group, comprising patients with non-RA inflammatory and non-inflammatory arthropathies. This group of patients diagnosed with non-RA conditions affecting the joints was selected to give more disease-specific insights relating to the role of DNAm in RA aetiology than would be possible when comparing patients with healthy controls.

Indeed, the disease control (non-RA) comparator group were selected where possible to be matched with the RA group with respect to demographic and clinical characteristics, including age, sex, and markers of acute phase response including CRP and ESR. The characteristics of all recruited patients for whom CD4$^+$ T cell and B cell DNAm data were collected are outlined in Table 3.1.

3.3 Sample quality control

A series of sample-level QC checks were performed to identify problematic samples to be excluded from the analysis, as described in detail in Chapter 2.5. This QC of samples resulted in the removal of four CD4$^+$ T cell samples and one B cell sample based on high mean detection p-value (≥ 0.01), indicating that the intensity values for a high proportion (10%) of probes in the sample were not significantly different from background signal levels. Detection p-values across all probes following removal of these failed samples are shown for CD4$^+$ T cell (Figure 3.1A) and B cell (Figure 3.1B) samples.

Samples were also removed if they displayed aberrant clustering in PCA based on the cell type annotation, and if the cell type proportion estimated using the Houseman method[219] was found to be <65% of the expected cell type . This resulted in a further two CD4$^+$ T cell samples and five B cell samples being removed. A plot of the first and second principal components calculated across samples shows all samples now cluster correctly according to the cell type annotation (Figure 3.2).

	Rheumatoid Arthritis	Disease Controls	P-value†
CD4$^+$ T cells:			
Number of patients	*48*	*67*	
Age	58 (51 - 69)	54 (46 - 63)	NS
Sex (% Female)	67	70	NS
CRP (mg/L)	10 (5 - 13)	8 (5 – 14)	NS
ESR (mm/hr)	21 (10 – 32)	15 (9 - 29)	NS
CCP Positive (%)	52	5	p < 0.0001
RF positive (%)	60	13	p < 0.0001
Tender 28	3 (1 - 11)	2 (0 – 6)	NS
Swollen 28	1 (0 - 3)	0 (0 - 2)	NS
DAS28	4.6 (3.6 – 5.4)	-	-
Diagnosis in disease controls:			
Osteoarthritis		5	
Other Non-Inflammatory Arthritis		6	
Spondyloarthropathy (PsA, ReA, EA)		31	
Crystal Arthropathy		9	
Other Inflammatory Arthritis		8	
Other/Undifferentiated		8	
B cells:			
Number of patients	*52*	*84*	
Age	59 (51 - 68)	54 (45 - 63)	NS
Sex (% Female)	71	71	NS
CRP (mg/L)	10 (5 - 14)	7 (5 – 15)	NS
ESR (mm/hr)	21 (7 - 34)	16 (9 - 30)	NS
CCP Positive (%)	54	6	p < 0.0001
RF positive (%)	60	10	p < 0.0001
Tender 28	3 (1 - 9)	3 (1 – 6)	NS
Swollen 28	1 (0 - 3)	0 (0 - 3)	NS
DAS28	4.3 (3.3 – 5.4)	-	-
Diagnosis in disease controls:			
Osteoarthritis		6	
Other Non-Inflammatory Arthritis		8	
Spondyloarthropathy (PsA, ReA, EA)		42	
Crystal Arthropathy		9	
Lupus/CTD-associated		2	
Other Inflammatory Arthritis		6	
Other/Undifferentiated		11	

Table 3.1 Clinical and phenotypic characteristics of all patients recruited into the study. Values refer to the median (inter-quartile range) for continuous variables. CRP – C-reactive protein; ESR – erythrocyte sedimentation rate; CCP – cyclic citrullinated peptide (clinical test to detect anti-cyclic citrullinated peptide antibodies); RF – rheumatoid factor; DAS28 – disease activity score (28 joints); PsA – psoriatic arthritis; ReA – reactive arthritis; EA – enteropathic arthritis. †P-values were calculated using the student's t-test for continuous data or a Fisher's exact test for categorical data. NS – not significant at $p < 0.05$.

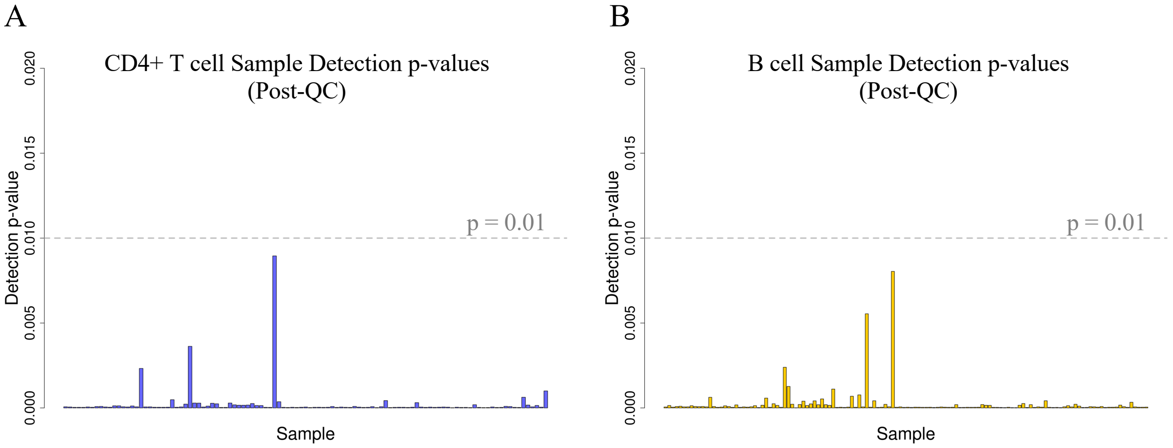


Figure 3.1: Mean sample detection p-values after removing failed samples. Detection p-values across all probes for (A) 109 CD4+ T cell and (B) 130 B cell DNA methylation samples from the MethylationEPIC array. Detection p-values represent the probability that a probe intensity is not significantly higher than background levels detected by the negative control probes on the MethylationEPIC array. Low p-values therefore signify good quality probes, and samples for which the mean detection p-value was < 0.01 passed quality control.

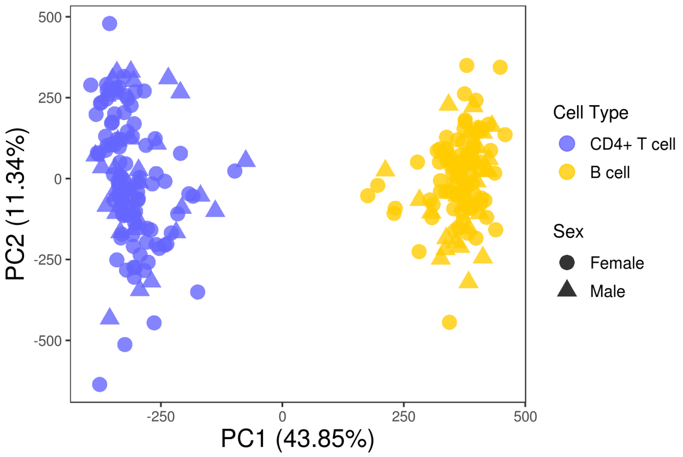


Figure 3.2: Principal component analysis of CD4^{+} T cells and B cells following quality control. The plot depicts all samples included in the analyses following removal of those that failed quality control. Each point represents a single sample, which are coloured according to their cell type annotation, with the shape of each point depicting either a female (circle) or male (triangle) sample.

Following the removal of the samples that failed these QC checks (six CD4^{+} T cell samples, five B cells samples), a total of 109 and 130 samples were available for downstream analyses from CD4^{+} T cells and B cells respectively. The sample information for this cohort post-filtering is provided in Table 3.2. For all subsequent analyses, the CD4^{+} T cell and B cell datasets were pre-processed separately.

	Rheumatoid Arthritis	Disease Controls	P-value†
CD4$^+$ T cells:			
Number of patients	*45*	*64*	
Age	59 (50 - 69)	54 (46 - 62)	NS
Sex (% Female)	69	70	NS
CRP (mg/L)	10 (5 - 13)	8 (5 – 14)	NS
ESR (mm/hr)	21 (10 - 32)	15 (9 - 29)	NS
CCP Positive (%)	51	5	<0.0001
RF positive (%)	58	13	<0.0001
Tender 28	3 (0 - 11)	2 (0 – 6)	NS
Swollen 28	1 (0 - 3)	0 (0 - 2)	NS
DAS28	4.6 (3.6 – 5.4)	-	-
Diagnosis in disease controls:			
Osteoarthritis		5	
Other Non-Inflammatory Arthritis		5	
Spondyloarthropathy (PsA, ReA, EA)		30	
Crystal Arthropathy		9	
Other Inflammatory Arthritis		8	
Other/Undifferentiated		7	
B cells:			
Number of patients	*49*	*81*	
Age	57 (50 - 68)	54 (45 - 63)	NS
Sex (% Female)	76	73	NS
CRP (mg/L)	10 (5 - 13)	7 (5 – 15)	NS
ESR (mm/hr)	21 (6 - 35)	16 (9 - 30)	NS
CCP Positive (%)	58	4	<0.0001
RF positive (%)	61	7	<0.0001
Tender 28	3 (1 - 10)	3 (1 – 6)	NS
Swollen 28	1 (0 - 3)	0 (0 - 2)	NS
DAS28	4.3 (3.2 – 5.4)	-	-
Diagnosis in disease controls:			
Osteoarthritis		6	
Other Non-Inflammatory Arthritis		6	
Spondyloarthropathy (PsA, ReA, EA)		41	
Crystal Arthropathy		8	
Lupus/CTD-associated		2	
Other Inflammatory Arthritis		6	
Other/Undifferentiated		11	

Table 3.2 Clinical and phenotypic characteristics of all patient samples passing quality control. DNA methylation from 109 CD4$^+$ T cell samples and 130 B cell samples were included in the analysis described in this chapter. No significant differences were observed between the rheumatoid arthritis and non-rheumatoid arthritis groups for any of the parameters, with the exclusion of autoantibody status. Values refer to the median (inter-quartile range) for continuous variables. CRP – C-reactive protein; ESR – erythrocyte sedimentation rate; CCP – cyclic citrullinated peptide (clinical test to detect anti-cyclic citrullinated peptide antibodies); RF – rheumatoid factor; DAS28 – disease activity score (28 joints); PsA – psoriatic arthritis; ReA – reactive arthritis; EA – enteropathic arthritis. †P-values were calculated using the student's t-test for continuous data or a Fisher's exact test for categorical data. NS – not significant at $p < 0.05$.

3.4 DNA methylation data normalisation

3.4.1 Normalisation methods for MethylationEPIC array data

Numerous methods exist for the normalisation of DNAm data collected using the MethylationEPIC array, all of which employ different approaches to adjust for probe type bias, background fluorescence, and other batch effects associated with processing of multiple samples (see Chapter 2.5.5).

Normal-exponential using out-of-band probes (noob) normalisation uses type I probes to measure non-specific fluorescence[269]. As both beads (methylated/unmethylated) in the type I design fluoresce at the same wavelength (Cy3 for C/G and Cy5 for A/T), signal from the opposite colour channel can be used to determine background fluorescence. After background correction, noob performs dye bias equalisation using positive control probes to account for different signal intensities from the red and green channels[269].

Beta mixture quantile dilation (BMIQ) applies a three-state (unmethylated, hemi-methylated, fully methylated) mixture model separately to the type I and type II probes, with subsequent adjustment of type II probe β-values to the distribution present in the type I probes[268]. A combination of noob with BMIQ has been shown to perform well in a systematic analysis of normalisation methods across four datasets[309].

Subset-quantile within array normalization (SWAN) assumes that probes with similar CpG content (as a surrogate for regional CpG density) will be biologically similar[271]. Quantiles are calculated for a subset of probes deemed to be biologically similar according to the CpG content, after which the remaining probes are adjusted for each probe type individually using the distribution from the subset[271].

Functional normalization (Funnorm), unlike noob, BMIQ, and SWAN, is a *between-array* normalisation method. Funnorm capitalises on the presence of control probes on the array that capture technical variability independent of any biological effects[270]. Specifically, this method uses principal components from the control probe intensities to remove variability associated with non-biological effects, and is as such an extension to the quantile normalisation methods which force all values from each sample to the same distribution[270]. As with BMIQ, the application of funnorm to data which has been background-corrected using noob performs favourably to the use of this method in isolation[270].

3.4.2 Assessment of methods for data normalisation

Whilst other normalisation methods exist, the three described above (Noob + BMIQ, SWAN, Noob + Funnorm) have proved popular in studies of DNAm and all employ distinct approaches to tackle the issue of non-biological variability in array data. For this reason, the above methods were applied to both DNAm datasets (results for the CD4$^+$ T cell data are shown here), with a range of outcome measures used to assess the suitability of each approach.

For DNAm analysis, intensity values can be converted to beta (β)-values, which represent the ratio (from 0 – 1) of the methylation probe intensity at a particular position relative to the total intensity (methylated and unmethylated; see Chapter 2). However, β-values display heteroscedascity with respect to the more extreme values, with much lower variability at values close to 0 and 1 [279]. For this reason, when performing statistical tests such as differential analysis on DNAm array data, it is recommended to use the log2 ratio of the β-value, referred to as the M-value[279]. SWAN and Funnorm were applied to intensity values, after which β- and M-values were generated, whereas BMIQ requires that normalisation be performed on β-values directly. Though β-values are useful for visualisation purposes and assessing total methylation levels (from methylated (1) to unmethylated (0)), statistical modelling is performed using M-values.

In the first instance, the ability of each method to correct probe type bias was assessed. To visualise the range in methylation values pre- and post-normalisation using the different methods, β-value density plots were produced to visualise the distribution at each probe type (Figure 3.3). In the raw dataset, a reduced dynamic range was observed in type II probes (Figure 3.3A), consistent with previous observations[268]. Though differences in distributions of each probe type are not unexpected, given that they interrogate CpGs in distinct genomic contexts, Noob + BMIQ (Figure 3.3B), SWAN (Figure 3.3C) and Noob + Funnorm (Figure 3.3D) all clearly increase the dynamic range in type II probes relative to raw values. Unsurprisingly, given that BMIQ is a quantile normalisation method applied directly to β-values, this method transforms both probe types to a uniform distribution (Figure 3.3B).

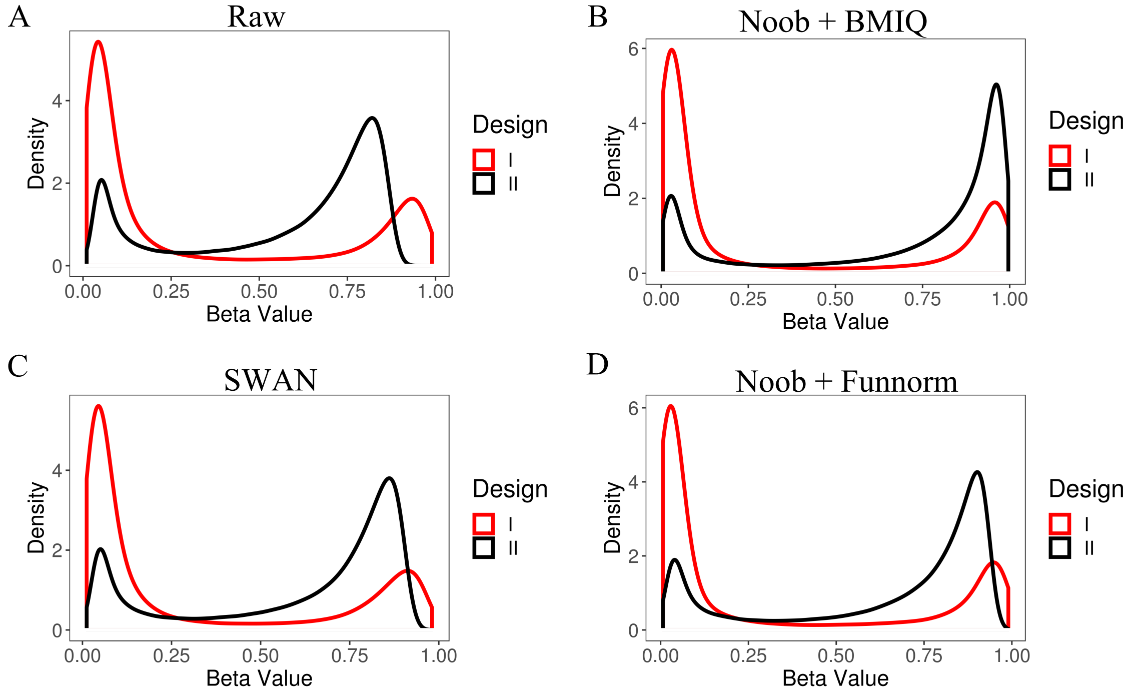


Figure 3.3: β-value density plots for type I and type II probes. DNA methylation (M-value) density plots for CD4^{+} T cell samples (A) pre-normalisation (raw data), as well as following normalisation of data using (B) Normal-exponential using out-of-band probes (noob) with Beta mixture quantile dilation (BMIQ), (C) Subset-quantile within array normalization (SWAN), and (D) noob with functional normalisation (Funnorm).

Subsequently, PCA of M-values was performed to assess global differences between samples that may be driven by technical artefacts such as processing batches. Given that samples were bisulphite converted in separate processing batches, this is likely to represent a major source of sample to sample variation. A degree of separation between bisulphite-conversion batches is observed in the raw data, such as conversion batches 4 and 5 which are largely segregated by the first principal component (PC1; Figure 3.4A). Noob + BMIQ appears to increase the variability associated with bisulphite conversion batch, with the variation in the data associated with PC1 increasing from 24.24% in raw data to 36.84% following application of this normalisation method (Figure 3.4B). Likewise, SWAN offers little improvement in this metric relative to the raw data, with batches exhibiting a degree of segregation, and variance associated with PC1 increase marginally to 25.71% (Figure 3.4C). Conversely, Noob + Funnorm considerably reduces the variability associated with the first PC (18.18%), with a reduction in the separation between conversion batches as observed on the PCA plot (Figure 3.4D). This would indicate that Noob + Funnorm performs favourably to the other methods tested in reducing this source of technical variation.

To further assess unwanted sample to sample variability arising from sample processing, relative log methylation (RLM) plots were generated. Such plots illustrate, for each sample, the deviation of probe values from their median values across all samples. This is effective for visualisation of variability between batches. In the absence of significant sources of non-biological variability, the median value (black line in the centre of each plot; Figure 5.3) for each sample would be centred on zero, whilst variability (height of box and whiskers) would be low. In the raw CD4$^+$ T cell data, divergent samples are detected, particularly in conversion batch 5 (Figure 3.5A). Noob + BMIQ resulted in an increase in the range of probe-wise deviations from the median for all samples (Figure 3.5B). The RLM plot for SWAN-normalised data (Figure 3.5C) appears similar to that for the raw data, whereas Noob + Funnorm considerably reduces the variability observed in relative log methylation across samples, most notably at conversion batch 5 (Figure 3.5D). These findings confirm those from the PCA above that Noob + Funnorm outperforms the other two methods in reducing unwanted variability associated with technical covariates.

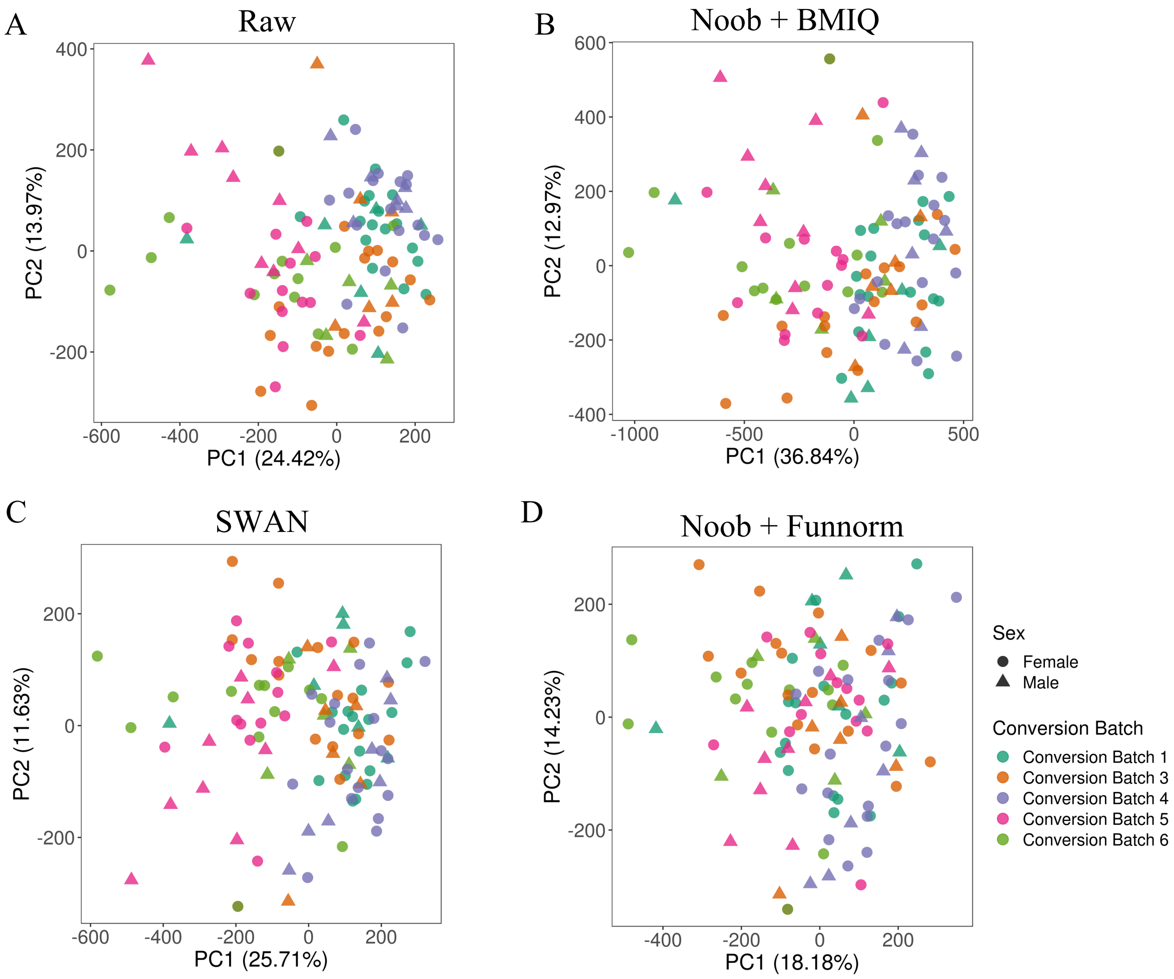


Figure 3.4: Principal component analysis of sample processing batch. CD4$^+$ T cell samples were subject to principal component analysis (A) pre-normalisation (raw data), as well as following normalisation of data using (B) Normal-exponential using out-of-band probes (noob) with Beta mixture quantile dilation (BMIQ), (C) Subset-quantile within array normalization (SWAN), and (D) noob with functional normalisation (Funnorm).Samples are coloured according to the bisulphite conversion batch in which they were processed, likely to represent a major source of non-biological variability, with circle points representing female samples

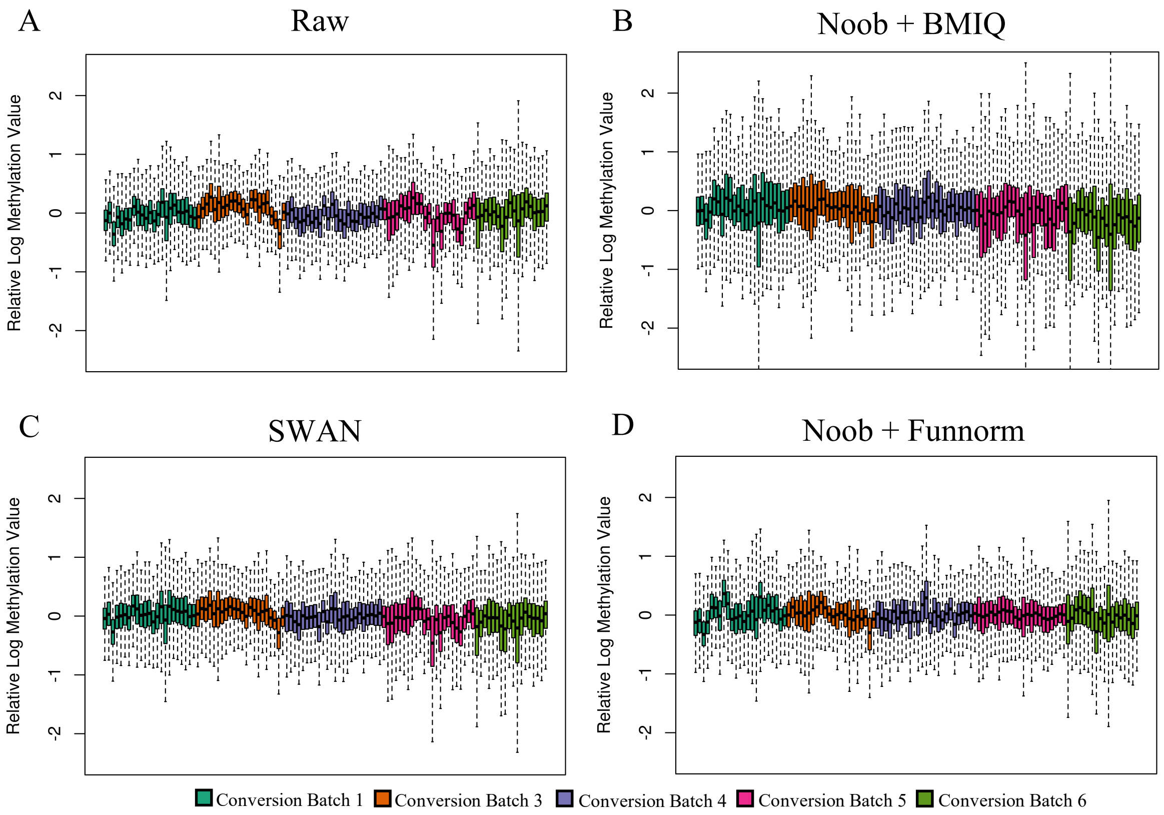


Figure 3.5: Relative log methylation (RLM) plots. CD4$^+$ T cell samples (A) pre-normalisation (raw data), as well as following normalisation of data using (B) Normal-exponential using out-of-band probes (noob) with Beta mixture quantile dilation (BMIQ), (C) Subset-quantile within array normalization (SWAN), and (D) noob with functional normalisation (Funnorm) were visualised as RLM plots. These plots represent, for each sample, the deviation in all probes from that probe's median across all samples. Boxes depict the median probe deviation for a given samples, as well as the lower and upper quartiles. The whiskers extend to the value of probes with the highest deviation (upper whisker for the probe with highest deviation above the sample-wise probe median, and the lower whisker for the probe with the highest deviation below the sample-wise probe median.

In order to quantify the relative sources of variability in the data that were revealed in PCA and RLM plots, a principal variance component analysis (PVCA) was applied to the data. This method first computes principal components in the dataset, and then fits linear models to identify associations between these PCs and measured covariates. Bisulphite conversion batch, sample positions on the array, Illumina scanning batch, and disease diagnosis (RA/non-RA) were selected as covariates of interest for which associations would be tested.

In the raw data, of the factors of interest, bisulphite conversion batch accounts for the highest proportion of the variability (28.0%), with position on the array (Position ID; 3.5%), Illumina scanning batch (Scanning Batch; 4.0%) and disease diagnosis (Diagnosis (RA/non-RA); 0.1%) each explaining a minor proportion of the variance (Figure 3.6A). Residual variance (i.e. that which is not associated with any of the supplied covariates) was 57.0%. Applying Noob +

BMIQ increased the variance that could be attributed to the conversion batch (34.4% vs. 28.0% in raw data; Figure 3.6B), consistent with the PCA results. SWAN was found to result in a reduction in batch-associated variance from 28.0% to 23.7% (Figure 3.6C), though resulted in a marginal increase in scanning batch variance relative to the raw data (4.2% vs 4.0%; Figure 3.6C). Again, Noob + Funnorm performed favourably to the other two methods in this regard, leading to a considerable reduction in variance attributed to both bisulphite conversion batch (18.8% vs. 28.0% in the raw data), as well as a small reduction in the scanning batch variance (2.3% vs. 4.0% in raw data; Figure 3.6D). This method also increased the residual variance from 57.0% (Figure 3.6A) to 62.1% (Figure 3.6D). Consistent across all PVCAs was that the proportion of variance associated with disease diagnosis was negligible.

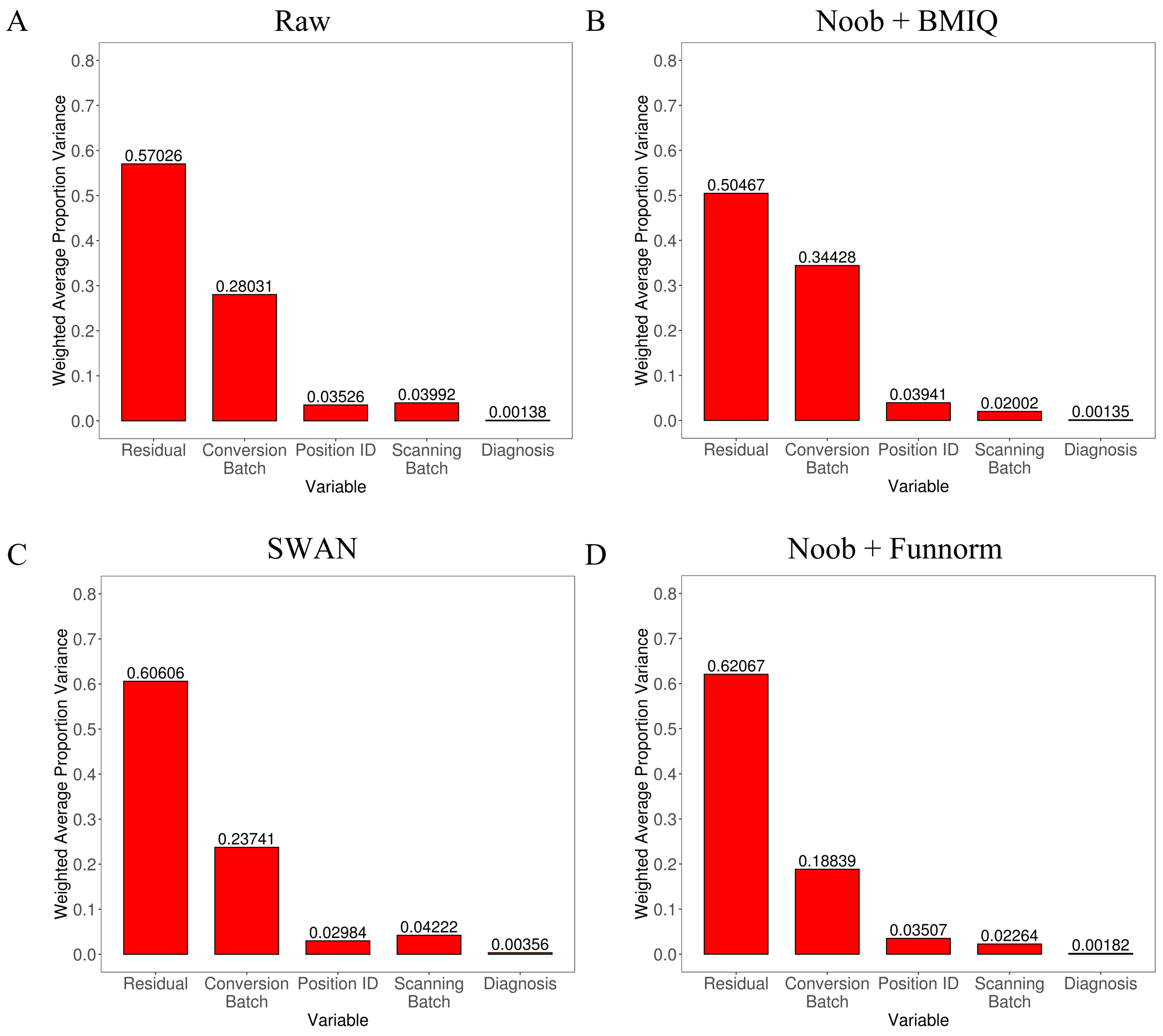


Figure 3.6: Principal variance component analysis (PVCA). PVCA was performed to quantify the relative contribution of factors of interest to data variance in (A) raw data, as well as following normalisation of data using (B) Normal-exponential using out-of-band probes (noob) with Beta mixture quantile dilation (BMIQ), (C) Subset-quantile within array normalization (SWAN), and (D) noob with functional normalisation (Funnorm). The covariates of interest were (bisulphite) conversion batch, the position of a sample of the MethylationEPIC array (Position ID), the scanning batch a sample was run in, and the sample diagnosis (RA/non-RA).

To facilitate the identification of batch to batch variability, a technical replicate was performed, with the same sample processed in separate batches and run on separate arrays. The Pearson's correlation between M-values at all probes in these technical replicates was calculates in the raw data, as well as data which had been normalised using each method. As expected, the correlation between replicates was high in the raw data (r = 0.987855; Figure 3.7A). Both Noob + BMIQ and SWAN had little effect on the correlation between replicates, with the former resulting in a marginal increase (r = 0.987924; Figure 3.7B), and the latter a small decrease (r = 0.987083; Figure 3.7C). Noob + Funnorm increased the correlation between technical replicates (r = 0.989121 vs. 0.987855 in raw data; Figure 3.7D). Following the completion of all quality control checks, and prior to analyses being performed, the sample included as a technical replicate was excluded, retaining the replicate exhibiting the lowest detection p-value.

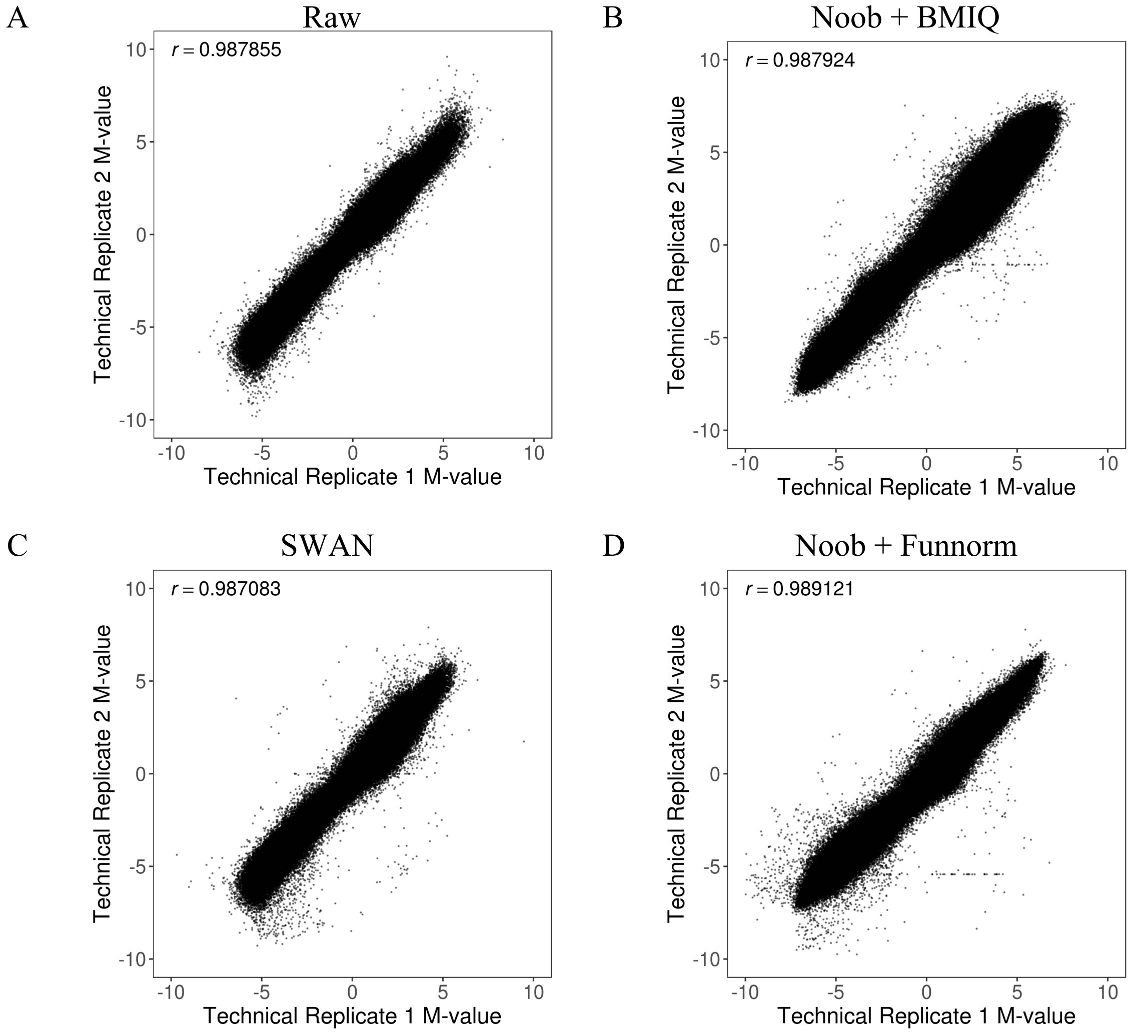


Figure 3.7: Pearson correlation at CD4$^+$ T cell technical replicates. To investigate batch effects, the same sample was run twice on different arrays across sample processing batches. The correlation (Pearson's r) between sample M-values (A) pre-normalisation (raw data), as well as following normalisation of data using (B) Normal-exponential using out-of-band probes (noob) with Beta mixture quantile dilation (BMIQ), (C) Subset-quantile within array normalization (SWAN), and (D) noob with functional normalisation (Funnorm).

Based on these quantitative and qualitative assessments of the normalisation methods described above, it was determined the Noob + Funnorm was the most appropriate procedure for this particular dataset (equivalent plots for the B cell dataset pre- and post-normalisation are presented in Appendix B). As such both CD4$^+$ T cell and B cell datasets were pre-processed accordingly using this method.

Following data normalisation, filtering of probes on the MethylationEPIC array was performed to remove those that failed (low signal; detection p-value), had previously been defined as cross-reactive, harboured a SNP with a minor allele frequency (MAF) > 0.05, or that mapped to the X and Y chromosomes (see section 2.5.6 for a summary of all probes removed). This resulted in a total of 709,412 CD4$^+$ T cell probes and 710,445 B cell probes that passed this filtering step and were included in all subsequent analyses.

3.4.3 Estimating biological and technical confounders

Despite demonstrating that functional normalisation considerably reduces batch-to-batch variability, such effects still persist in normalised data, as is evident from the PVCA plot in which bisulphite conversion batch explains 18.8% of the variability in DNAm data in CD4$^+$ T cells (Figure 3.6D). In addition, hierarchical clustering of samples revealed a degree of clustering by this batch processing variable (Figure 3.8).

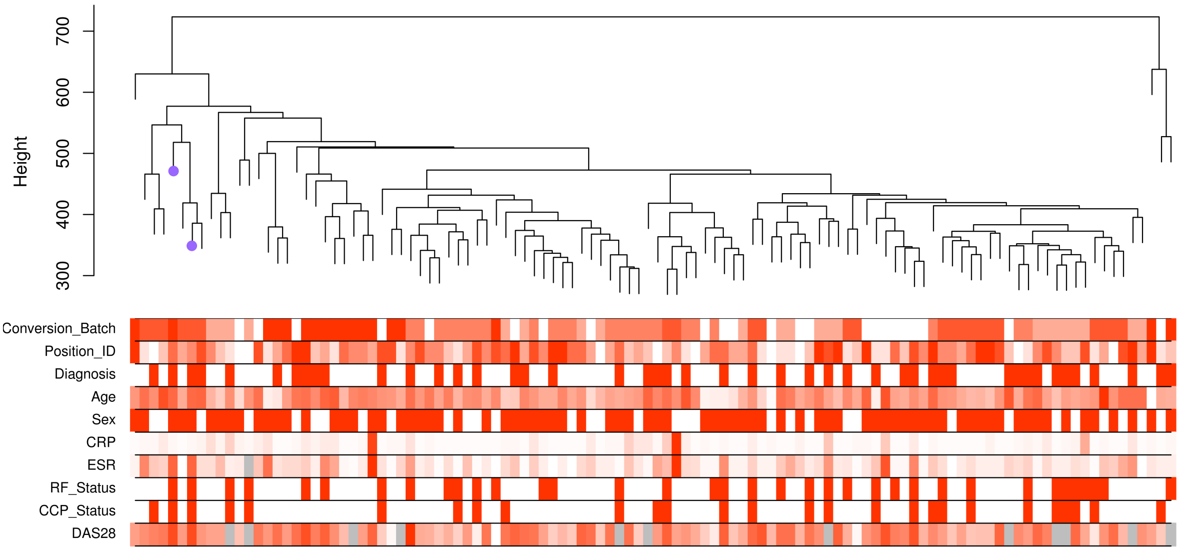


Figure 3.8: Hierarchical clustering of all CD4$^+$ T cell samples. The average distance between cluster elements was used to perform clustering. Purple circles on the dendrogram branches represent the clustering of the two technical replicates run across batches (note: one of the replicates shown here was removed before all subsequent analyses). The cluster dendrogram is plotted adjacent to a heat map of sample demographic, clinical, and processing information. The heat map represents low values as white and high values as red, with intermediate colour gradient indicating increasing values. Diagnosis: red = RA, white = non-RA; Sex: red = female, white = male; RF/CCP status: red = positive, white = negative. Grey tiles on the heat map represent missing data.

To account for remaining technical and biological covariates including batch effects, and the residual 62.1% variance (Figure 3.6D) that may capture unmeasured technical and biological variability, surrogate variable analysis (SVA) was performed. This approach detected 11 surrogate variables (SVs) in the CD4$^+$ T cell DNAm data, and 13 in the B cell data. To check whether SVA was capturing known sources of variability, SVs were tested for associations with measured technical and biological covariates. As expected, the strongest associations were with bisulphite conversion batch in both CD4$^+$ T cells (Figure 3.9A) and B cells (Figure 3.9B). Additionally, the final SV identified for each cell type (SV11 for CD4$^+$ T cells, SV13 for B cells), is associated with patient age. Although this method appears to detect variability in the data that arises from measured covariates, such associations were not evident for all SVs, justifying this approach for the identification of hidden confounders. The surrogate variables identified here were included as covariates in downstream analyses.

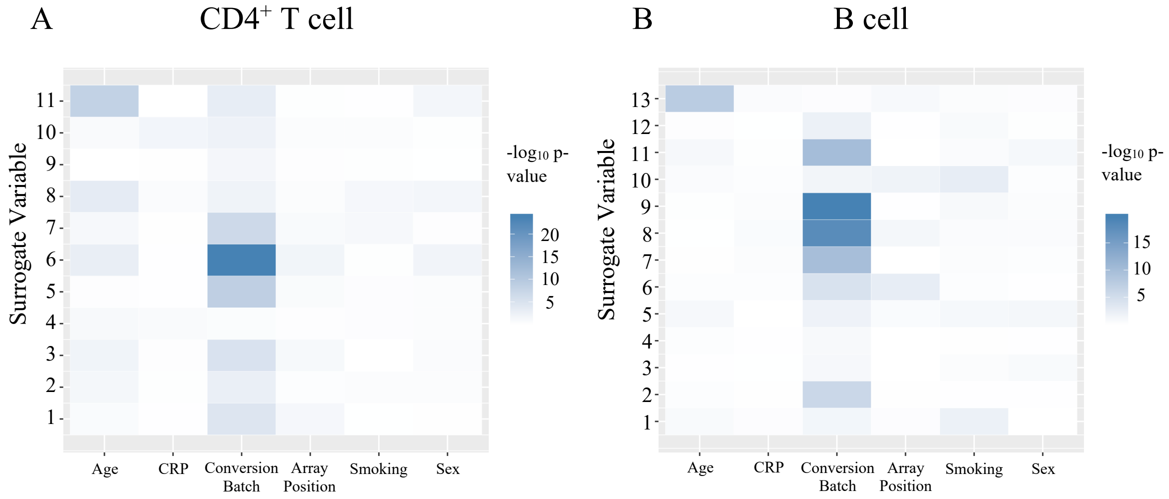


Figure 3.9: Heat map of associations between surrogate variables and measured covariates. Surrogate variables were identified by surrogate variable analysis (SVA) and the association between each surrogate variable and potential sources of variability in the data was calculated using linear regression for continuous variables (Age, CRP), analysis of variance (ANOVA) for categorical variables with > 2 groups (conversion batch, array position), and a Mann-Whitney U test binary variables (smoking, sex). Associations are plotted as the –log$_{10}$ p-value, with white indicating no association and dark blue indicating a strong association.

3.5 DNA methylome comparison of RA and disease control patients

Following the pre-processing and quality control steps as outlined above, a total of 109 CD4$^+$ T cell samples and 130 B cell samples were available for downstream analyses, including 85 patient samples for which paired data were available for both cell types. The demographic and clinical parameters for RA patients and disease controls included in all analyses described in this chapter are reported in Table 3.2. Importantly, the RA patients and non-RA disease control comparator groups were matched for age, sex, and markers of acute phase response (C-reactive protein; CRP, erythrocyte sedimentation rate; ESR; see section 1.2.5). Principal component analysis revealed no clustering of samples based on comparator group (RA or disease controls), suggesting that no global differences exist between either the CD4$^+$ T cell (Figure 3.10A) or B cell (Figure 3.10B) DNA methylomes of these patient groups.

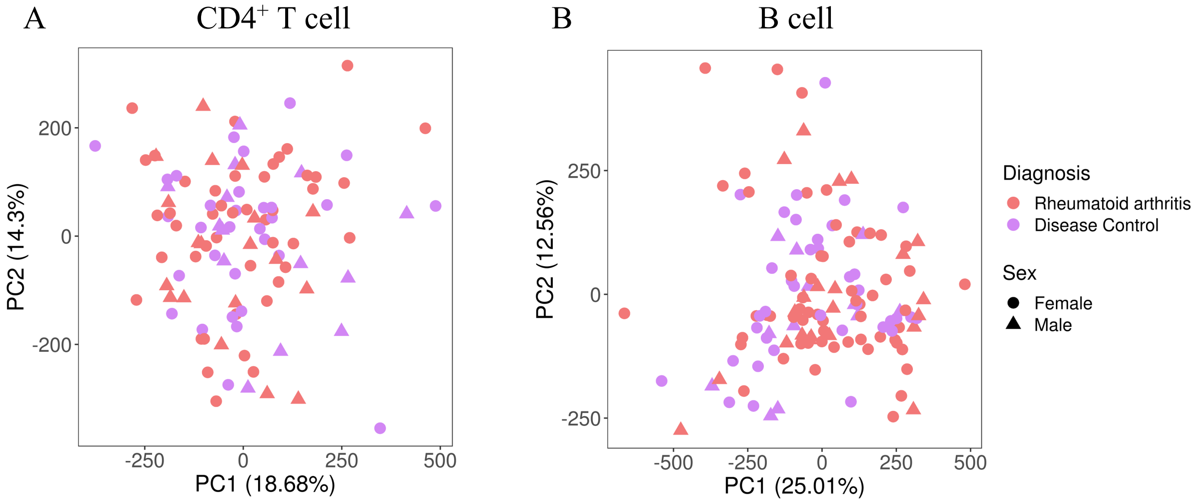


Figure 3.10: Principal component analysis of disease diagnosis. (A) CD4$^+$ T cell samples and (B) B cells samples included in the differential analyses described in this chapter were first assessed by principal component analysis. Each sample is coloured according to the patient diagnosis (red = rheumatoid arthritis, purple = non-rheumatoid arthritis disease control), with the shape depicting female (circle) and male (triangle) patients. Note – the slight differences between panel (A) here and Figure 3.4D reflects the removal of the technical replicate prior to PCA calculation.

3.5.1 RA epigenome-wide association study in CD4$^+$ T cells and B cells

To quantify differential methylation, linear models were fit to M- and β-values in limma[283] for each CpG with an RA vs. non-RA contrast. β-values were used for calculating DNAm differences and data visualisation (converted to % for the latter), whilst statistical modelling was performed using M-values due to their homoscedastic nature[279]. An empirical Bayes method was employed in order to moderate the standard errors[284] as this approach takes into account the probe-wise variability to calculate moderated t-statistic, and rank each CpG by likelihood of being differentially methylated. Surrogate variables, as identified in section 3.4.3 for each cell type, were included as covariates in the design matrix.

The p-value distribution across all differential tests performed (709,412 for CD4$^+$ T cells, Figure 3.11A; 710,445 for B cells, Figure 3.11B) exhibited a uniform distribution across the entire range (0 – 1), as would be expected under the null hypothesis that no differences in DNAm exist between the RA and non-RA groups. Consistent with the lack of enrichment at lower p-values, after controlling for false discovery rate experiment-wide, no CpGs were differentially methylated between RA patients and controls at FDR < 0.05 and $\Delta\beta \geq 0.05$

(equivalent to 5% difference in DNAm group means) in CD4$^+$ T cells (Figure 3.11C). Similarly, the analysis in B cells failed to identify any differentially methylated CpG sites (Figure 3.11D).

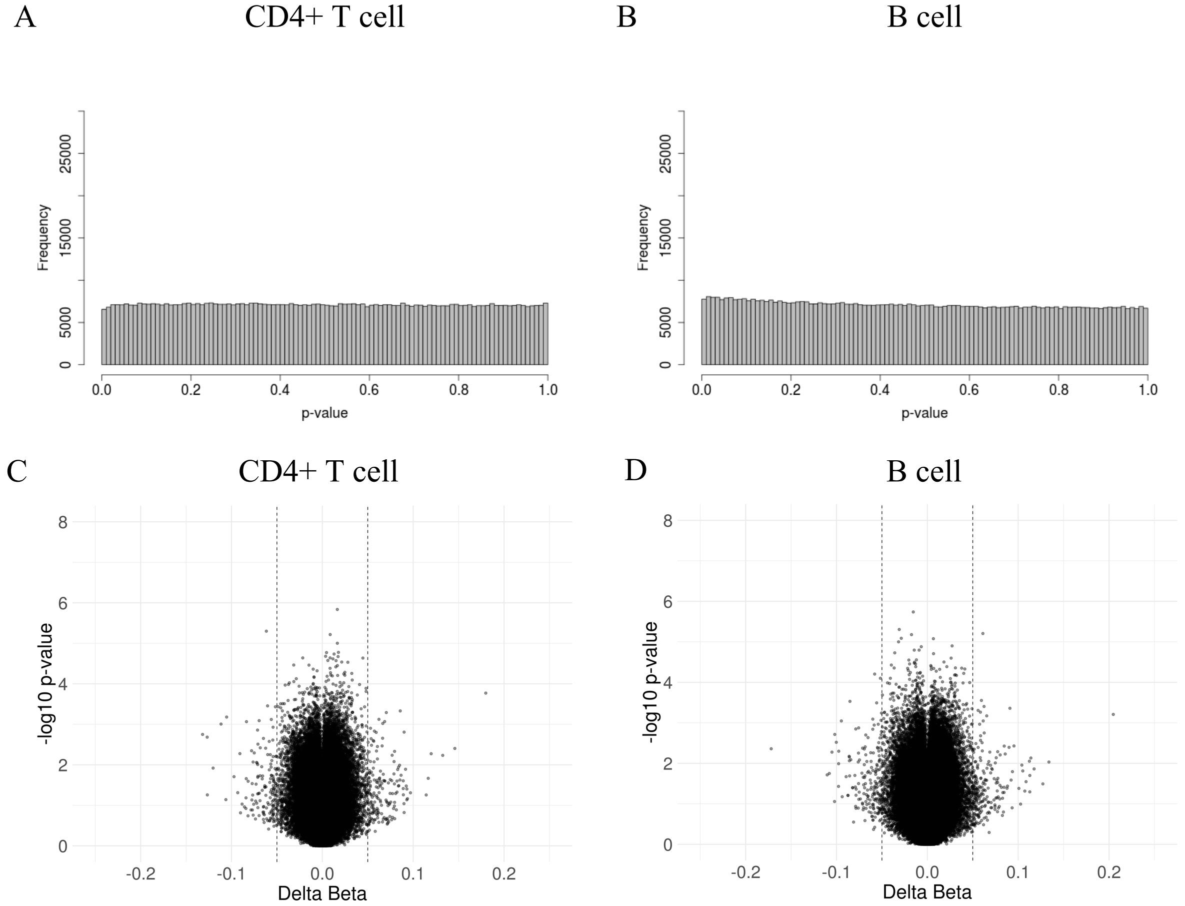


Figure 3.11: Epigenome-wide p-values of associations between lymphocyte DNA methylation and diagnosis in rheumatoid arthritis cases and disease controls. P-value histograms depict the frequency of associations (CpGs) returning a given p-value range for all tests in (A) CD4$^+$ T cell and (B) B cell samples. Each bin on the histogram represents a range of 0.01, with 100 bins across the whole range of p-values (0 – 1). Volcano plots of all tests performed in (C) CD4$^+$ T cell and (D) B cell samples are shown. The delta beta (Δβ; mean difference in DNA methylation between cases and controls at a given position) is plotted against the –log$_{10}$ p-value (un-adjusted). Vertical lines at delta beta -0.05 and 0.05 demarcate the threshold for differential methylation. Data points in black represent those that were not significant at the experiment-wide FDR (< 0.05)

Though no results were significant after controlling false discovery rate, of all CpGs tested in CD4$^+$ T cells, the CpG with the lowest nominal p-value ($p = 5.02 \times 10^{-6}$) in the RA vs. non-RA contrast was cg21289466 (hg19 genome build, chr20: 5,577,716; Figure 3.12A). Furthermore, though no differences were significant to FDR correction, a number of CpGs exhibited a relatively large degree of difference in Δβ values between the comparator groups. Examples included cg11424828 (chr8:2,075,469) which displayed the largest degree of hypo-methylation in RA patients relative to disease controls (Δβ = -0.13, Figure 3.12 B), and cg24245216 (chr19:7,004,657), at which the extent of RA hyper-methylation was greatest (Δβ = 0.18, Figure 3.12C). The 10 CpGs with the lowest nominal p-value (Δβ ≥ 0.05) are reported in Table 3.3.

In the B cell comparison, cg00595050 (chr19:10,398,582) returned the lowest non-significant p-value (3.12D), with cg08736526 (chr4:88,656,433, Δβ = -0.17; Figure 3.12E) and cg20673407 (chr10: 31,040,939, Δβ = 0.20; Figure 3.12F) displaying the greatest degree of RA hypo- and hyper-methylation respectively. The top 10 CpGs with the lowest nominal p-value (Δβ ≥ 0.05) are reported for the CD4$^+$ T cell and B cell analyses in Table 3.3 (the top 100 CpGs ranked by nominal p-value with Δβ ≥ 0.05 are given in Appendix C). Interestingly, a number of these positions (cg24245216 in CD4$^+$ T cells, cg08736526 & cg20673407 in B cells) were associated with proximal genetic variants (see Chapter 4).

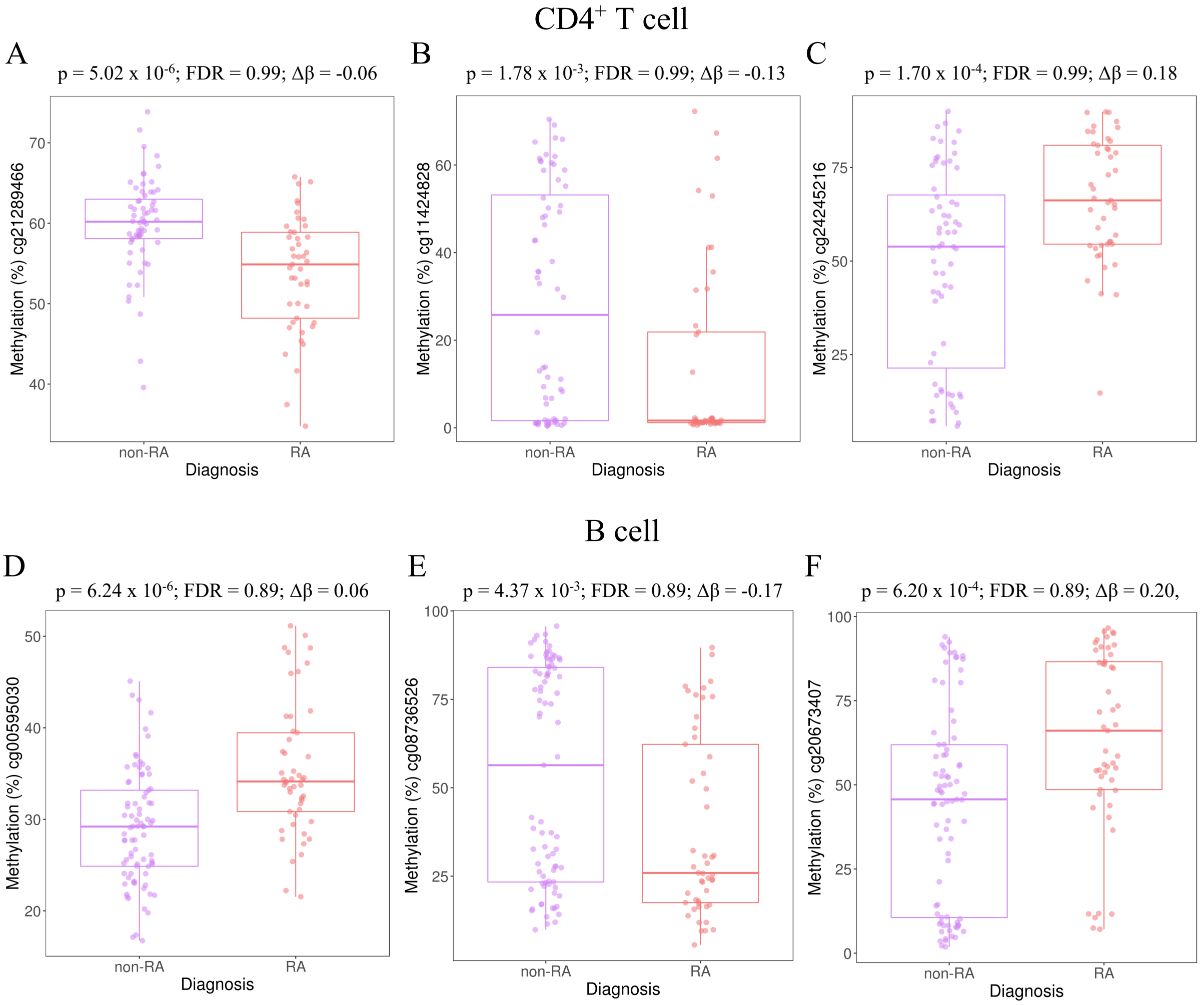


Figure 3.12: Exemplar plots of the top non-significant CpGs in a differential analysis of early rheumatoid arthritis patients and non-rheumatoid arthritis controls. CD4$^+$ T cell plots are shown for CpGs with the (A) lowest un-adjusted p-value, as well as the greatest magnitude of (B) rheumatoid arthritis (RA) hypo-methylation, and (C) RA hyper-methylation (hypo-/hyper-methylation is determined by Δβ). The equivalent plots for the B cell dataset, displaying top (albeit non-significant) hits as determined by (D) un-adjusted p-value, (E) RA hypo-methylation and (F) RA hyper-methylation. Boxplots represent the median value of each group, with the upper and lower box limits extending to the 75th and 25th percentile respectively. Whiskers extend to the maximum and minimum data points that are no greater than 1.5 × the inter-quartile range.

CpG	Coordinates	P-value	FDR	RA vs. non-RA Δβ	UCSC RefGene	Relation to CpG Island
CD4$^+$ T cell						
cg21289466	chr20:5577716	5.02×10^{-6}	0.99	-0.06	*GPCPD1*	Open Sea
cg24245216	chr19:7004657	1.70×10^{-4}	0.99	0.18	-	Open Sea
cg11945167	chr8:4644739	3.50×10^{-4}	0.99	-0.06	*CSMD1*	Open Sea
cg15563420	chr12:86026194	3.70×10^{-4}	0.99	-0.05	-	Open Sea
cg00080972	chr5:178986291	4.67×10^{-4}	0.99	0.09	*RUFY1*	N Shore
cg07612827	chr19:7005180	5.12×10^{-4}	0.99	0.07	*FLJ25758*	Open Sea
cg11787167	chr14:33407370	6.62×10^{-4}	0.99	-0.11	*NPAS3*	S Shelf
cg18471635	chr11:104769411	7.53×10^{-4}	0.99	0.06	*CASP12*	Open Sea
cg03161803	chr6:27649120	7.59×10^{-4}	0.99	-0.06	-	S Shore
cg05287483	chr20:5551376	8.12×10^{-4}	0.99	-0.07	*GPCPD1*	Open Sea
B cell						
cg00595030	chr19:10398582	6.24×10^{-6}	0.89	0.06	*ICAM4*	Island
cg10800620	chr2:196398826	6.26×10^{-5}	0.89	-0.06	-	Open Sea
cg06323052	chr4:56720686	7.94×10^{-5}	0.89	-0.05	*EXOC1*	S Shore
cg25152193	chr1:197874469	2.80×10^{-4}	0.89	0.06	*C1orf53*	S Shelf
cg22901297	chr6:32522795	2.95×10^{-4}	0.89	-0.09	*HLA-DRB6*	Open Sea
cg00538212	chr7:158751591	3.22×10^{-4}	0.89	-0.06	-	N Shore
cg21419137	chr8:87905504	3.95×10^{-4}	0.89	0.06	*CNDB1*	Open Sea
cg16055526	chr6:33083287	4.37×10^{-4}	0.89	0.09	*HLA-DPB2*	N Shore
cg22404498	chr22:32600722	4.94×10^{-4}	0.89	0.06	*RFPL2*	Open Sea
cg08666831	chr19:47507691	5.46×10^{-4}	0.89	0.06	*GRLF1*	Island

Table 3.3: CpGs with the lowest nominal p-values in a rheumatoid arthritis epigenome-wide association study of CD4$^+$ T cells and B cells. The 10 CpGs with the lowest p-values in a differential analysis of early rheumatoid arthritis patients and non-rheumatoid arthritis disease controls. FDR = false discovery rate calculated using the Benjamini-Hochberg method; UCSC RefGene = gene to which the CpG maps based on the Illumina Infinium MethylationEPIC manifest; Relation to CpG Island = CpG island feature (see Chapter 2.7.4) to which the CpG maps based on the MethylationEPIC manifest.

3.5.2 RA differentially methylated regions

Extended regions of CpG sites that display patterns of differential methylation across distinct disease states can be more informative than considering single CpG sites in isolation. These can have important roles in regulating patterns of differential expression that dictate cellular phenotype, as with the Treg-specific de-methylated region (TSDR) that modulates expression of *Foxp3* during development of the Treg lineage [310].

The DMRcate package[286] was used for the identification of DMRs. The first stage in this process involves a differential analysis analogous to that which was described in section 3.5.1, with the aim of identifying DMPs to pass to the *dmrcate* function for DMR identification. However, consistent with the above analysis, this revealed no positions to be differentially methylated between RA cases and non-RA controls.

For discovery purposes, to enable the identification of extended regions that may exhibit a consistent pattern of hyper- or hypo-methylation in RA lymphocytes, this initial stage was forgone. Though such an approach will bias results with respect to generated DMRcate p-values, it allows consistent patterns of Δβ across multiple CpGs to be identified. This function also generates a region-specific p-value, taking into account local correlations between probes, using Stouffer's method[286]. In CD4$^+$ T cells, twelve regions (≥ 2 CpGs) were identified across which multiple CpGs consistently display a non-significant trend of differential methylation (mean Δβ ≥0.05 across the region; Table 3.4). Most notably, a region was identified encompassing 10 CpGs mapping to the promoter of the *RUFY1* gene (chr5:178986131-178987429) that displayed consistently increased levels of hypo-methylation in CD4$^+$ T cells from RA patients relative to controls (Figure 3.13A & Table 3.4). Interestingly, nine of the CpGs in this region displayed nominal significance (unadjusted p-value < 0.01) in the CD4$^+$ T cell EWAS in section 3.5.1 (Appendix C).

In B cells, eleven such non-significant regions were highlighted with consistent trends of either increased or decreased DNAm levels in RA patients compared with controls (Table 3.4). Prominent amongst these was a region spanning an intronic region upstream of the *PIGZ* gene (chr3:196705629-196706839; Table 3.4 & Figure 3.13B). Seven CpGs in this region displayed a trend of RA-hypo-methylation. Interestingly, this was also observed at the same region in CD4$^+$ T cells with a comparable magnitude of RA hypomethylation (regional mean Δβ of -0.063 in CD4$^+$ T cells and -0.059 in B cells; Table 3.4).

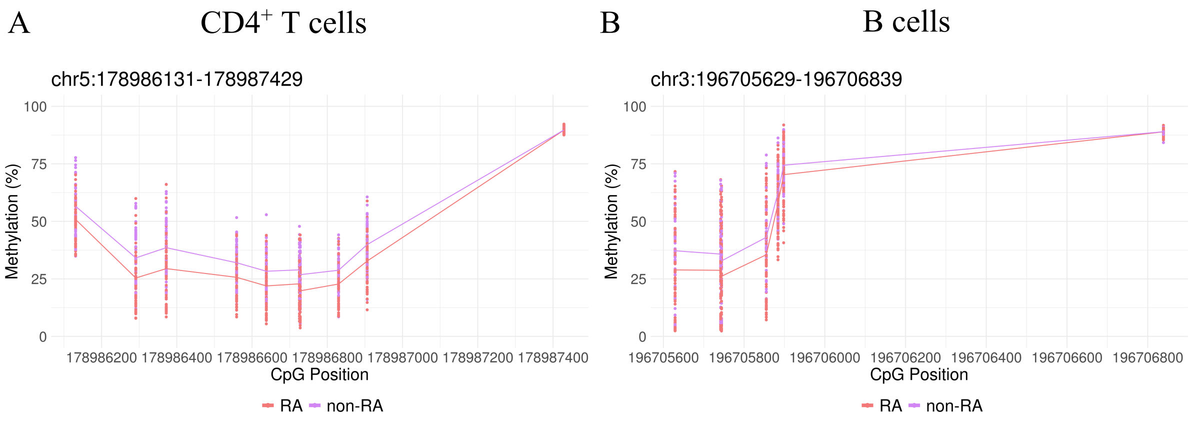


Figure 3.13: Exemplar plots of regions displaying hypo-methylation in rheumatoid arthritis. Extended regions at which ≥ 2 CpGs exhibit a non-significant difference in methylation values in the same direction between cells from rheumatoid arthritis cases and non-rheumatoid arthritis controls. (A) A region on chromosome 5 encompassing 10 CpGs at the promoter of *RUFY1* was found to show a trend of hypo-methylation in CD4$^+$ T cells from rheumatoid arthritis patients. (B) An intronic region harbouring 7 rheumatoid arthritis hypo-methylated CpGs in B cells, as well as CD4$^+$ T cells (DNAm data shown here is from the B cell comparison. The two lines represent the mean DNAm in rheumatoid arthritis patients (red) and non-rheumatoid arthritis controls (purple) at each CpG within the region.

DMR Coordinates	No. of CpGs in DMR	Stouffer p-value*	Max Δβ	Mean Δβ
CD4$^+$ T cell:				
chr5:178986131-178987429	10	1.000	-0.090	-0.060
chr3:196705629-196706839	7	1.000	-0.092	-0.063
chr19:7004657-7005379	4	1.000	-0.180	-0.076
chr8:2074935-2075777	4	1.000	0.132	0.069
chr6:32632643-32633102	3	1.000	-0.146	-0.056
chr19:12876846-12877188	2	1.000	-0.093	-0.082
chr19:35861258-35861642	2	1.000	0.084	0.073
chr4:69435473-69435601	2	1.000	0.069	0.059
chr5:1594715-1594733	2	1.000	-0.069	-0.055
chr6:32552042-32552095	2	1.000	-0.120	-0.100
chr7:24917499-24917750	2	1.000	-0.080	-0.052
chr5:71683884-71683955	2	1.000	0.127	0.064
B cell				
chr3:196705629-196706839	7	1.000	-0.092	-0.059
chr4:25090198-25090665	6	0.999	0.086	0.066
chr12:131519883-131520382	5	0.998	-0.103	-0.077
chr17:5673550-5674234	3	0.984	0.105	0.067
chr13:50194322-50194643	3	0.987	-0.083	-0.073
chr15:81411055-81411066	2	0.959	-0.077	-0.056
chr7:7860864-7861342	2	0.959	0.076	0.050
chr14:106539756-106539897	2	0.961	0.086	0.065
chr5:180402690-180402906	2	0.961	0.071	0.059
chr6:32449961-32450452	2	0.981	0.098	0.059
chr19:15649345-15649508	2	0.993	-0.104	-0.053

Table 3.4: Results from an analysis to detect differentially methylated regions between rheumatoid arthritis cases and non-rheumatoid arthritis disease controls. Results are reported for all regions with ≥ 2 CpGs in CD4$^+$ T cells or B cells that display a non-significant, albeit consistent, trend of either hyper- or hypo-methylation in rheumatoid arthritis cases relative to controls. No. of CpGs in region = the total number of CpGs at which the methylation differences (Δβ ≥ 0.05) are observed; Stouffer p-value = a regional p-value reported by DMRcate that accounts for underlying correlated patterns of DNA methylation across CpGs; Max Δβ = the difference in β-value between cases and controls at the CpG exhibiting the greatest magnitude of hyper-/hypo-methylation; Mean Δβ = the mean difference in β-value between cases and controls across all CpGs within the region.

3.5.3 RA differentially variable positions

The analyses described in sections 3.5.1 & 3.5.2 have revealed that no significant differences exist between RA and non-RA group means at any assayed CpG sites. Subsequently, the possibility that variance in DNAm levels within one particular comparator group may differ significantly from the other, as was highlighted in a recent twin study[230], were considered. This twin study employed a statistical approach that was initially developed to identify heterogeneous patterns of DNAm that may predict risk of cancer progression[287]. This method first applies Bartlett's test to assess whether one group displayed a significantly higher variance in DNAm relative to the other (FDR < 0.001). The variance in data for a given group represents the mean of the squared deviations of each value from a group mean, and as such, higher values indicate a greater degree of variability in the data. Given that Bartlett's test is liable to generate low p-values for variances driven by outlier samples, a second step is performed to rank positions based on the mean group differences between case and control samples. As such, any CpGs passing the initial differential variance threshold were subsequently tested for differences in group means (t-test, $p < 0.05$)[287] to define differentially variable positions (DVPs).

Using these criteria, 291 DVPs were identified between RA and non-RA patients in CD4$^+$ T cells, including both CpG and CpH sites, with 41% (120) displaying patterns of hyper-variability in RA. Exemplar plots of DVPs that were found to be hypo-variable (cg15174564, chr11:120856801; Figure 3.14A) or hyper-variable (cg00647389, chr7:127234990; Figure 3.14B) in CD4$^+$ T cells of RA patients are depicted.

In B cells, 601 such DVPs were highlighted, with 53% (320) showing RA hyper-variability. DNAm values at sites displaying both hypo-variability (Figure 3.14C) and hyper-variability (Figure 3.14D) in B cells from RA patients are shown. The top 10 DVPs ranked by the significant difference between group means are reported for both hyper- and hypo-variable CpGs in each cell type (Table 3.5; top 100 DVPs for each cell type ranked by t-test p-value given in Appendix D).

Of further interest, 15 CpG sites were found to be differentially variable in both CD4$^+$ T cells and B cells, with all but one of these displaying the same pattern of either hyper-/hypo-variability in RA patients relative to controls (Figure 3.15).This suggests that the factors driving variability at some loci may be common to both cell types.

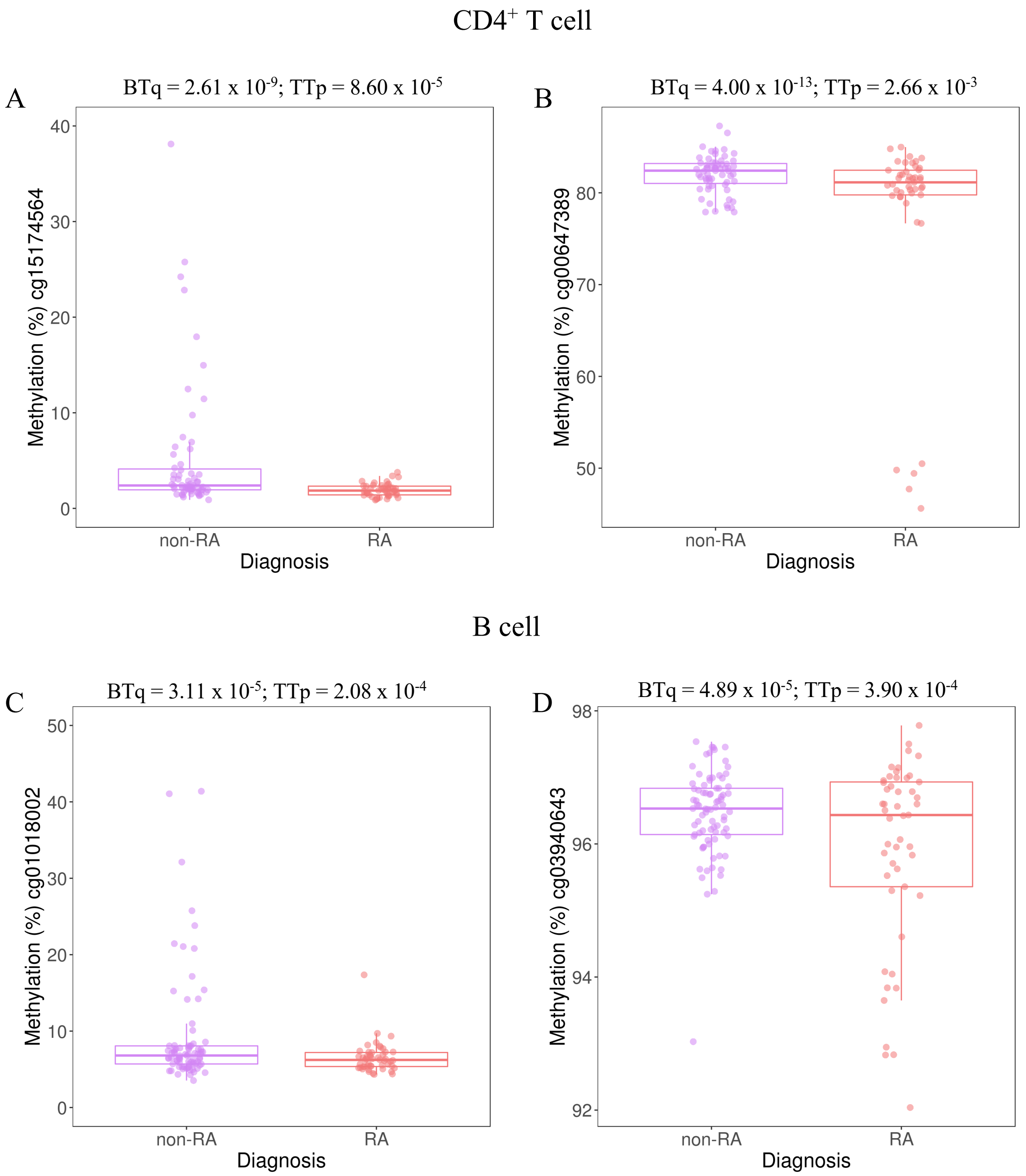


Figure 3.14: Exemplar plots of differentially variable positions between rheumatoid arthritis patients and non-rheumatoid arthritis controls. Plots are shown for CpGs in CD4+ T cells that were found to be either (A) hypo-variable or (B) hyper-variable in rheumatoid arthritis patients, with equivalent B cell plots again depicting (C) rheumatoid arthritis hypo-variable and (D) hyper-variable positions. BTq = Bartlett's q-value, a false discovery rate adjusted measure of group differences in DNA methylation variance; TTp = T-test p-value, applied to test for differences between the group means of any positions found to exhibit differential variance between cases and controls in the Bartlett's test ($q < 0.001$).

CpG	CpG Coordinates	Variance non-RA	Variance RA	p-value (t-test)	q-value (Bartlett's)	UCSC RefGene	Relation to CpG Island
CD4+ T cell – rheumatoid arthritis hyper-variable							
cg15174564	chr11:120856801	1.938	0.214	8.60 x 10^{-5}	2.61 x 10^{-9}	*GRIK4*	Island
cg27284424	chr7:130598669	1.846	0.270	2.45 x 10^{-4}	2.55 x 10^{-7}	*LOC100-506860*	Open Sea
cg04797575	chr4:176709392	0.803	0.160	1.30 x 10^{-3}	2.49 x 10^{-5}	*GPM6A*	Open Sea
ch.8.1995451R	chr8:98920744	1.291	0.233	1.33 x 10^{-3}	5.80 x 10^{-6}	*MATN2*	Open Sea
cg04641168	chr4:57333196	2.376	0.476	1.87 x 10^{-3}	2.61 x 10^{-5}	*SRP72*	N_Shore
cg13396858	chr9:134249466	0.757	0.151	2.24 x 10^{-3}	2.44 x 10^{-5}	-	S_Shore
ch.4.1647744F	chr4:85766242	2.312	0.587	2.47 x 10^{-3}	6.23 x 10^{-4}	*WDYF3*	Open Sea
cg21421501	chr11:61734205	0.620	0.146	2.78 x 10^{-3}	2.46 x 10^{-4}	*FTH1*	N_Shore
cg13565723	chr2:61245319	1.054	0.158	2.85 x 10^{-3}	3.60 x 10^{-7}	*PEX13*	Open Sea
cg07871034	chr19:48103301	0.758	0.092	3.62 x 10^{-3}	1.17 x 10^{-8}	-	N_Shore
CD4+ T cell – rheumatoid arthritis hypo-variable							
cg00647389	chr7:127234990	0.058	0.623	2.66 x 10^{-3}	4.00 x 10^{-13}	*FSCN3*	Open Sea
cg07891761	chr19:35861642	0.165	1.785	3.93 x 10^{-3}	3.12 x 10^{-13}	-	Open Sea
cg18423635	chr6:29869936	0.184	1.062	4.82 x 10^{-3}	1.78 x 10^{-7}	*HCG2P7*	Open Sea
cg13990487	chr1:19420096	0.092	0.344	5.05 x 10^{-3}	2.88 x 10^{-4}	*UBR4*	Open Sea
cg14451627	chr9:115987035	0.299	1.952	5.67 x 10^{-3}	1.63 x 10^{-8}	*SLC31A1*	S_Shelf
cg11460110	chr6:30530458	0.056	0.203	6.71 x 10^{-3}	4.44 x 10^{-4}	*PRR3*	Open Sea
cg20426698	chr3:65960357	0.183	2.000	7.35 x 10^{-3}	2.57 x 10^{-13}	*MAGI1*	Open Sea
cg26312542	chr11:112038104	0.060	0.222	7.40 x 10^{-3}	3.98 x 10^{-4}	*TEX12*	Open Sea
cg02978220	chr1:108337960	0.388	2.513	8.72 x 10^{-3}	1.88 x 10^{-8}	*VAV3*	Open Sea
cg13230994	chr11:66979798	0.039	0.148	8.92 x 10^{-3}	2.61 x 10^{-4}	*KDM2A*	Open Sea
B cell – rheumatoid arthritis hyper-variable							
ch.2.1159565R	chr2:48430755	1.781	0.241	1.39 x 10^{-4}	6.17 x 10^{-9}	-	Open Sea
cg01018002	chr1:28844479	0.825	0.188	2.08 x 10^{-4}	3.11 x 10^{-5}	*SNHG3-RCC1*	N_Shore
cg22797164	chr10:72200612	0.551	0.153	5.12 x 10^{-4}	5.36 x 10^{-4}	*NODAL*	Island
cg15256944	chr17:79651359	0.113	0.028	6.16 x 10^{-4}	8.52 x 10^{-5}	*ARL16*	Island
cg08638512	chr10:120514811	1.207	0.343	8.20 x 10^{-4}	7.22 x 10^{-4}	*C10orf46*	Island
ch.2.1701371R	chr2:75108364	2.069	0.488	1.32 x 10^{-3}	5.36 x 10^{-5}	*HK2*	Open Sea
cg19658926	chr7:26240443	0.323	0.058	2.55 x 10^{-3}	8.79 x 10^{-7}	*CBX3*	Island
cg10570405	chr7:108210304	0.450	0.116	2.76 x 10^{-3}	2.03 x 10^{-4}	*THAP5*	Island
cg06113708	chr10:76996280	0.790	0.045	2.94 x 10^{-3}	3.04 x 10^{-16}	*COMTD1*	S_Shore
cg01915196	chr12:57916109	0.912	0.045	2.95 x 10^{-3}	1.68 x 10^{-17}	*MBD6*	N_Shore
B cell – rheumatoid arthritis hypo-variable							
cg03940643	chr16:11343701	0.168	0.617	3.90 x 10^{-4}	4.89 x 10^{-5}	-	Island
cg10720997	chr1:997858	0.191	1.019	3.93 x 10^{-4}	2.03 x 10^{-8}	-	N_Shore
cg19819559	chr8:107282279	0.293	1.002	8.05 x 10^{-4}	1.65 x 10^{-4}	*OXR1*	Island
cg14268695	chr20:23137044	0.180	0.602	9.92 x 10^{-4}	2.38 x 10^{-4}	-	Open Sea
cg14897833	chr14:52535178	0.106	0.331	9.95 x 10^{-4}	7.04 x 10^{-4}	*NID2*	Island
cg02231880	chr15:33011300	0.127	0.423	1.08 x 10^{-3}	2.74 x 10^{-4}	*LOC100-131315*	Island
cg25308427	chr11:10324739	0.188	0.608	1.12 x 10^{-3}	4.32 x 10^{-4}	-	Island
cg15154191	chr1:3527782	0.071	0.220	1.53 x 10^{-3}	7.96 x 10^{-4}	*MEGF6*	Island
cg02623991	chr19:54926431	0.172	0.623	1.65 x 10^{-3}	6.10 x 10^{-5}	*TTYH1*	N_Shore
cg20803547	chr1:67773440	0.233	0.828	1.71 x 10^{-3}	8.54 x 10^{-5}	*IL12RB2*	Island

Table 3.5: The top 10 differentially variable positions in rheumatoid arthritis lymphocytes. Positions identified as being either hyper- or hypo-variable (Bartlett's test) in CD4+ T cells from rheumatoid arthritis patients, as well as those identified as such in B cells, were ranked by t-test p-value. Variance non-RA = the variance (the average of the squared differences of all samples from the group mean) in DNA methylation M-values in non-rheumatoid arthritis controls; Variance RA = the variance in DNA methylation M-values in rheumatoid arthritis cases; p-value (t-test) = p-value from the t-test of differences in group means between cases and controls; q-value (Bartlett's) = the false discovery rate adjusted p-value from the Bartlett's test of variance; UCSC RefGene = Gene to which the differentially variable position maps based on the Infinium MethylationEPIC manifest; Relation to CpG Island = mapping of differentially variable position to a CpG island or related structure.

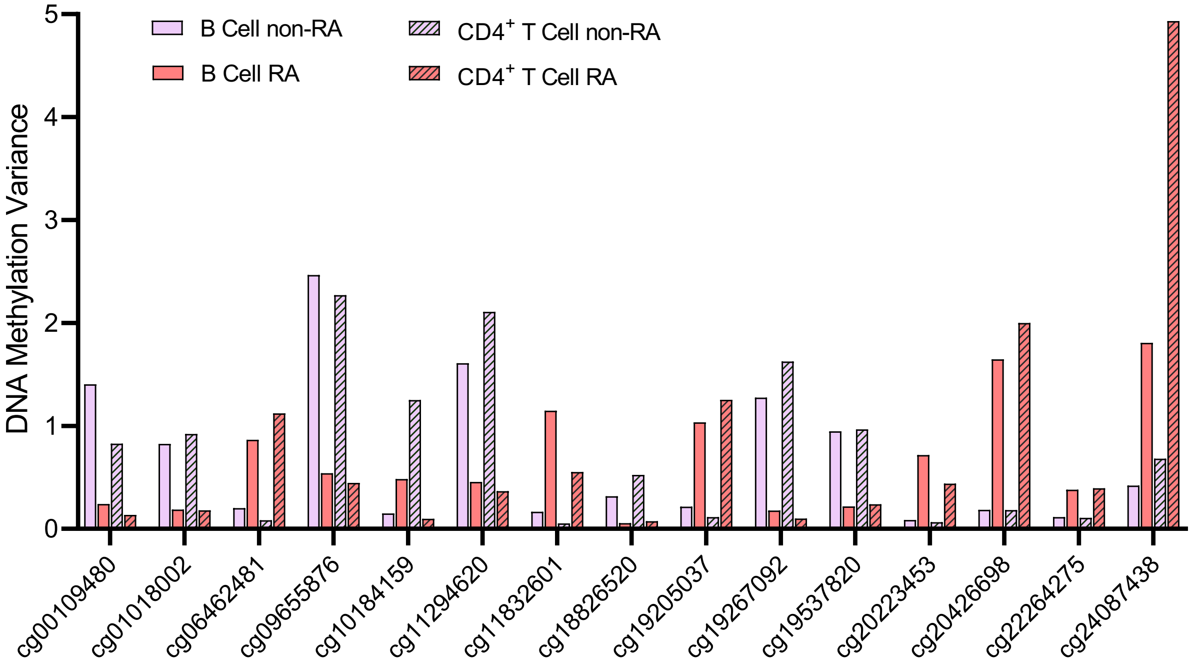


Figure 3.15: Overlapping differentially variable positions in CD4$^+$ T cells and B cells. The variance in DNAm levels is shown for data from rheumatoid arthritis patients and disease controls at CpG sites that were found to be differentially variable in both CD4$^+$ T cells and B cells.

3.5.4 Biological pathway analysis at variable positions

To gain additional functional insight, a pathway enrichment analysis was performed using the *gometh* function in the missMethyl package[289]. This method assigns CpGs to a target gene and performs a hypergeometric test of enrichment for Gene Ontology biological processes (BP), accounting for the number of probes per gene assayed on the MethylationEPIC array (see Chapter 2.6.4 for Methods). DVPs were tested for BP enrichment, with all probes included in the analysis used as background.

In CD4$^+$ T cells, BP pathways that were enriched amongst DVPs found to be hypo-methylated in the RA cohort included those relating to development of the hematopoietic or lymphoid organs, nucleosome assembly, and immune system development (Figure 3.16A). Conversely, RA hyper-methylated DVPs most strongly implicate pathways the cellular response to prostaglandins (Figure 3.16B). In B cells, RA hypo-variable DVPs disproportionately mapped to genes with functions in cellular metabolism and regulation of G0 to G1 transition during the cell cycle (Figure 3.16C), with hyper-variable positions highlighting processes including organisation of the cytoskeleton and cell-cell adhesion, amongst others (Figure 3.16D).

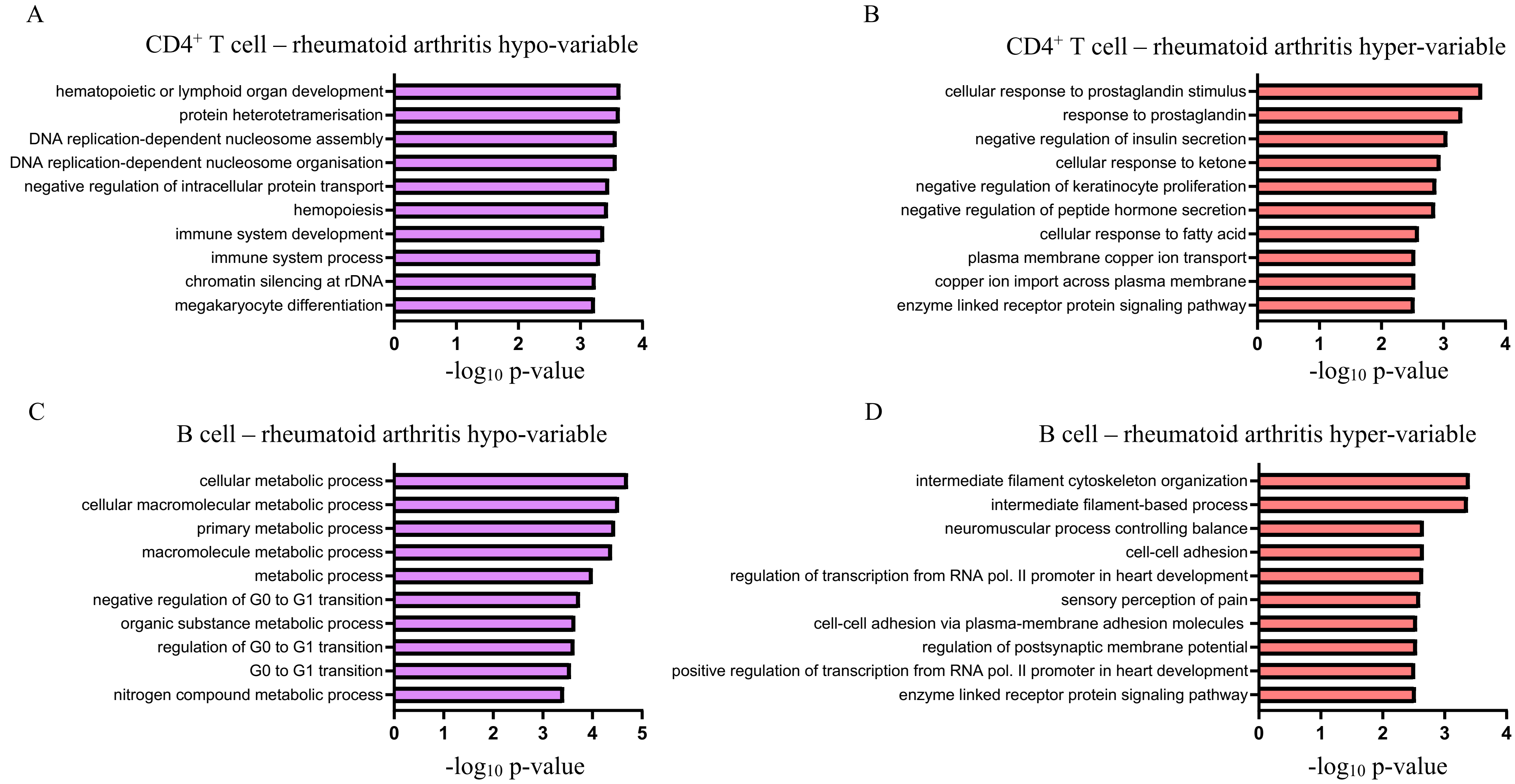


Figure 3.16: Top 10 Gene Ontology biological processes enriched at differentially variable positions. Positions exhibiting differentially variable DNA methylation were mapped to genes and a modified hypergeometric test performed to identify enriched Gene Ontology pathways. Separate analyses were performed for positions that were (A) hypo-methylated and (B) hyper-methylated in rheumatoid arthritis CD4$^+$ T cells, as well as those that were (C) hypo-methylated and (D) hyper-methylated in rheumatoid arthritis B cells. Enrichment analyses were performed using the missMethyl package, using non-variable CpGs included in the analysis as background.

3.6 Associations between RA and epigenetic ageing

'Horvath's clock' or the 'epigenetic clock' refers to a DNAm signature at 353 CpG sites shared across tissues that correlates strongly with chronological age of the individual[291]. This has been of interest clinically given that certain age-related diseases may be associated with an increased 'biological age'. If biological age is reflected in the epigenetic clock then DNAm may represent a better biomarker or predictor of disease risk than chronological age[311].

The epigenetic age of all samples was calculated using the Horvath method[291], and correlated with the chronological age of the patient at the time of samples collection. As expected, the Horvath age of cells correlates strongly with chronological age in both CD4$^+$ T cells (Pearson's r = 0.90; 3.17A) and B cells (Pearson's r = 0.90; 3.17B). In both cell types, a general trend of Horvath age being higher than chronological age at the lower end of the age scale was observed, with the opposite being true at the upper range, as has been described[312].

To assess the extent to which chronological age and disease diagnosis explain variation in epigenetic ageing seen in patient samples, linear regression was performed with the inclusion of these variables as covariates, and Horvath age as the dependent variable. Associations between chronological age and Horvath age in this model were highly significant in both CD4$^+$ T cells ($p = 1.64 \times 10^{-30}$) and B cells (3.65×10^{-40}). An interaction term (chronological age × diagnosis (RA/non-RA)) was included to assess whether or not RA patients exhibited accelerated or decelerated epigenetic ageing. However, no significant interaction effect between these two covariates on the epigenetic age was identified in either CD4$^+$ T cells (p = 0.94; Figure 3.17A) or B cells (p = 0.59; Figure 3.17B).

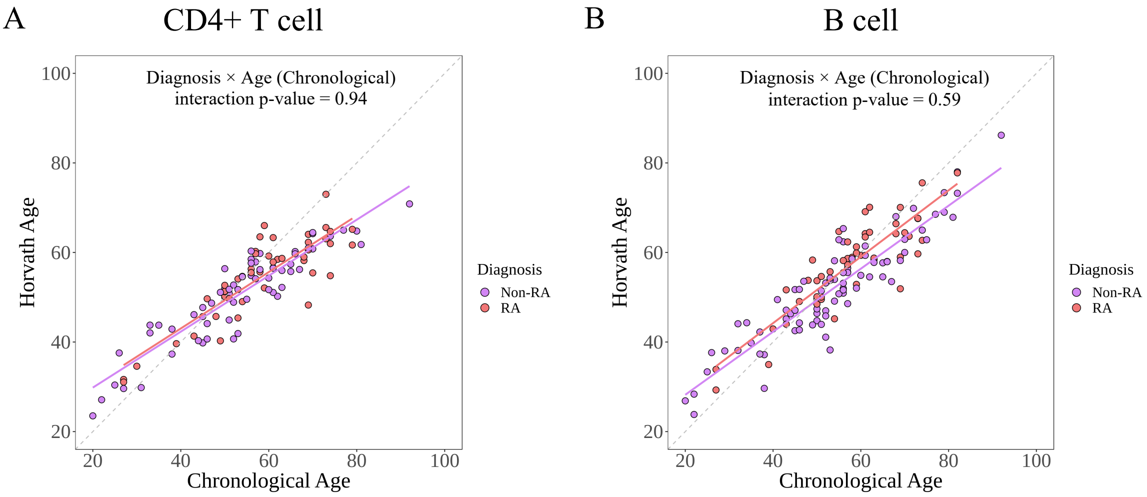


Figure 3.17: Epigenetic ageing in lymphocytes from rheumatoid arthritis and non-rheumatoid arthritis patients. Horvath's method was used to calculate the epigenetic age (Horvath age) of (A) CD4$^+$ T cell and (B) B cell samples, which is plotted against the chronological age of the patient at the time the blood samples were collected. Linear regression (red/purple lines) was performed with an interaction term to identify instances in which disease diagnosis (rheumatoid arthritis or non-rheumatoid arthritis) significantly impacted the association between Horvath age and chronological age. The grey dotted x = y line depicts the point at which the Horvath age is equal to the chronological age.

3.7 Discussion

In this chapter, an extensive characterisation of the CD4$^+$ T cell and B cell DNA methylation landscape in early arthritis has been described. Lymphocyte DNAm in RA patients was compared with a control group diagnosed with diverse rheumatic diseases. This was in contrast to most preceding EWASs which sought epigenetic changes distinguishing RA patients from those with non-inflammatory osteoarthritis, or healthy controls with no clinical symptoms.

The inclusion of this symptomatic comparator group was motivated by a need to distinguish DNAm signatures that predispose to RA or reflect functional disease-specific regulatory modifications, from those that occur more generally under inflammatory conditions. Indeed, by selecting cases and controls who were equivalent with respect to demographic characteristics and markers of active inflammation, as well as being naïve to any disease-modifying therapy, the potential impact of confounding sources of variability was minimised.

CD4$^+$ T cells and B cells represented an appealing candidate cell population for this study given their integral role in the immunopathogenesis of RA, as highlighted in numerous cellular and genetic studies, as well as by the efficacy of biologic therapies targeting these cells (see Chapter 1 for further details).

3.7.1 Epigenome-wide association study of RA fails to replicate findings from previous studies

The first analysis performed in both datasets (CD4$^+$ T-/B-cell) was to employ a typical EWAS design to highlight individual CpG (or CpH) sites at which DNAm levels were associated with RA. However, no such associations were found to be statistically significant after controlling for multiple testing. This finding is at odds with published data from the same or related cell types. In a small study of 23 RA patients and 11 healthy controls, Glossop et al. highlighted differential methylation at 1951 positions in T cells and 2238 in B cells (FDR < 0.05, $\Delta\beta \geq 0.1$), with considerable differences in DNAm ($\Delta\beta \geq 0.2$) at 150 and 113 of these positions, respectively[306]. Furthermore, an additional > 100 CpGs in each cell type were identified as being differentially methylated in a separate comparison of patients with established RA specifically, as opposed the early RA, highlighting the important of disease stage in study design[306]. Using a comparable sample size to Glossop et al., an EWAS in CD4$^+$ T cells from individuals of Han Chinese ethnicity described differential methylation at 1168 CpGs, the majority of which (67%) were hypo-methylated in RA patients relative to age- and sex-matched healthy controls[221]. Despite this, a carefully designed study found no differential methylation in isolated CD4$^+$ T cells from patients with oligoarticular juvenile idiopathic arthritis, a condition pathologically similar to RA that occurs in infants[313].

Focussing on B cells, Julia et al. described 64 DMPs in RA patients (many of whom were receiving DMARD treatment) relative to healthy controls and, crucially, were able to replicate these findings at 10 of these positions in an independent cohort[222]. This study also reported that of these 10 validated RA DMPs, nine of these were also found to display the same association in patients with SLE, a B cell-mediated autoimmunity having overlapping genetic aetiology with RA[222]. Nonetheless, this may highlight similar epigenetic risk mechanisms in two immune-mediated inflammatory disease or reflect similar cellular exposures in the inflammatory milieu. Consistent with this, an EWAS in B cells from multiple sclerosis patients highlighted 5/10 of the CpGs identified by Julia et al. to be differentially methylated in this autoimmune disease affecting the central nervous system[314].

Discrepancies between these preceding studies and the findings described in this chapter may be attributed to a number of factors. Perhaps most likely is the selection of a disease control comparator group that reflects the RA cases with regards to important clinical parameters, as opposed to healthy or non-inflammatory patients. It is possible that DNAm modifications that occur during early RA are in fact shared across autoimmune diseases, as was described for the B cell associations in RA and SLE[222]. In addition, it should also be noted that a small proportion of patients in the disease control group were positive for ACPA antibodies (5% of CD4+ T cell samples, 4% of B cell samples) or RF (13% of CD4+ T cell samples, 7% of B cell samples). Whether these samples represent patients who were misdiagnosed, those in pre-RA and as such are yet to fulfil RA diagnostic criteria, or simply reflect rare cases of these antibodies occurring in individuals without RA is unclear. Nonetheless, if these do reflect true RA cases, then the power to detect RA-associated DNAm differences is reduced by this misclassification.

Though no systematic meta-analysis of shared DMPs across autoimmune traits has been performed, collating findings from various studies has revealed overlap in disease-associated DNAm changes in RA, SLE, and Sjögren syndrome (SS), all of which involve joint-related synptoms[315]. In this scenario, peripheral immune cell methylomes between patients with RA and other inflammatory diseases may be more comparable than those between RA and healthy controls. Consistent with this, epigenetic signatures that are associated with a chronic inflammatory response have been reported, with associations between DNAm at 58 CpGs and CRP levels validated in two separate populations[213]. A study which explored effects beyond CRP to look at other protein markers of inflammation identified associations between leukocyte DNAm and pro-inflammatory cytokines/chemokines, in addition to validating previous CpG associations with CRP[316]. Whether such associations reflect exposure of cells to pro-inflammatory mediators, or the DNAm changes themselves confer such properties on cells is difficult to deduce. Recently it was discovered that 52% of associations between DNAm and protein biomarkers of disease were likely the result of genetic polymorphisms. Using Mendelian randomization (see Chapter 1.6.10) to infer causality, the authors concluded that protein biomarkers influencing DNAm was the most likely scenario, whilst the reverse did not show any associations. Such findings would suggest that the pro-inflammatory environment to which cells are exposed in diseases such as RA can shape the DNA methylome, and as such may influence gene expression and cellular phenotypes.

Nonetheless, DMPs and DMRs have been identified in asymptomatic ACPA+ individuals relative to those without these circulating antibodies, and a small number were similarly found in RA versus healthy controls, suggests that epigenetic modifications that precede clinical symptoms may potentially contribute directly to disease induction[317]. Many of these methylation changes were attributed to genetic variants, emphasising the contribution of both genetic and non-genetic factors in conferring disease associated DNAm (this is discussed further in Chapter 4).

Most EWASs in cells from peripheral blood described previously have been performed using the earlier Illumina platform that quantifies methylation at ~450,000 sites, whereas the current study utilizes the MethylationEPIC platform (~7% of 450K probes are not present on the EPIC array). Whilst this offers the capacity to assess DNAm at much more positions genome-wide (~850,000), the increased multiple testing burden presents additional statistical challenges. Indeed, whilst probe filtering to remove non-variable CpGs has been applied in some studies[220, 317], this approach was not applied here to avoid biasing results.

In conclusion, due to large discrepancies in study designs, including different tissues or cell types being assessed, different technologies or platforms for quantification of DNAm, and distinct statistical approaches to define differential methylation, direct comparison of results across studies is complicated. The majority of findings from RA and other autoimmune EWASs described to date require independent validation, and large-scale meta-analyses will facilitate the identification of small effects that current studies may be under-powered to detect. Further recognition of cell type specificity will also be necessary going forward. In this regard, while the isolation of lymphocyte subsets in this study represents and improvement on the use of whole blood, one limitation is that differences in cellular subtypes was not assessed (i.e. naïve and memory cells, regulatory T cells).

3.7.2 Analysis of differentially methylated regions highlights potential RA-associated perturbations

Though no DMPs were identified, an analysis of putative DMRs was performed by relaxing the FDR threshold for input DMPs. Such observations, even in the absence of statistically significant differential effects, are likely to provide more compelling disease insights than isolated positions showing nominal significance. The identification of such regions harbouring multiple CpGs (ranging from 2-10 CpGs in CD4$^+$ T cells and 2-7 CpGs in B cells) that uniformly displayed a trend of hypo-/hyper-methylation in RA patients relative to controls may highlight relevant pathogenic perturbations.

Interestingly, a DMR encompassing five CpGs that map to this same region of the *RUFY1* promoter shows a similar degree of hypo-methylation ($\Delta\beta = 0.076$) in CD4$^+$ T cells from patients with primary SS, illustrating a potential disease-spanning epigenetic risk mechanism in these rheumatic conditions[318]. *RUFY1* encodes an endosomal protein which may regulate endocytosis following interaction with the tyrosine-protein kinase enzyme Etk[319], though a well-defined role in cellular immunity has not been described.

Also, of note was the observation that a putative intergenic DMR was found to be hypo-methylated to a similar degree in both CD4$^+$ T cells and B cells from RA patients, indicating that DMRs can occur across cell types. Given that datasets for each cell type were derived from the same patient in most cases, genetic influences on DNAm could account for these observations – for example where one allele is over-represented in either comparator cohort by chance. As with DMPs, the extent to which genetic and environmental factors drive disease-associated DMRs is often not considered in EWAS, though identification of such regions that differ in methylation status between RA-discordant monozygotic twins suggests there is a role for the latter[320].

3.7.3 DNA methylation variability as an important epigenetic risk mechanism in RA

Though no DMPs or DMRs were statistically significant at the genome-wide level, a number of loci were identified at which variance in DNAm differed between RA patients and controls. These findings are in agreement with a twin study prioritizing differential variability as an important component of pathological autoimmune responses in RA[230]. A unique feature in twin studies such as this is that the effects of underlying regulatory genetic variants is negated, as genetic variation is matched between the comparator groups. Despite numerous caveats associated with this study, such as analysing whole blood as opposed to individual cell types, this suggests that many DMPs identified in EWAS may be conferred by sequence variation.

This is also supported by results from similar studies of twins discordant for other autoimmune traits. Paul et al. investigated DNAm in three key immune effector cells: CD4$^+$ T cells, B cells, and monocytes, in 52 pairs of monozygotic twins discordant for T1D[231]. Such effects were prevalent in all cell types and were predominantly hyper-variable in the affected twin. Interestingly DVPs were independent of cis-regulatory genetic variants, as would be expected in monozygotic twins, and were found to be depleted in enhancer elements but enriched at active transcription start sites[231].In MS-discordant monozygotic twins, only six DMPs were identified in PBMCs at genome-wide significance, though these were no longer significant after correcting for differing cell type proportions between the cases and controls[232]. Unlike the studies of RA and T1D, however, only 25 DVPs were identified, and the majority of these were hyper-variable in the unaffected twin.

By mapping such variable CpGs to genes and performing pathway analyses, intriguing pathways relating to development of lymphoid organs and the immune system were implicated at RA hypo-variable CpGs. The top pathway associated with CD4$^+$ T cell hyper-variable CpGs in RA were responses to prostaglandin (Figure 3.16B). Indeed, a role for prostaglandins in the differentiation of the T$_H$1 subtype as well as expansion of T$_H$17 cells, both of which contribute to inflammation during autoimmunity, has been described[321]. Amongst the pathways enriched at RA hyper-variable CpGs in B cells were those involved in cytoskeletal organisation and cell-

cell adhesion, suggesting that distinct pathway from CD4$^+$ T cells are perturbed in this cell type, potentially impacting cell migration. The relevance of these DVPs and associated pathways in the pathogenesis of RA warrants further investigation but, if such variability is independent of genetic effects, identifying risk factors that confer such epigenetic variation will be a challenge.

The results in this chapter begin to illustrate the DNA methylation landscape in the context of early arthritis patients. Distinguishing DNA methylation profiles between the two disease-relevant cell types under investigation strongly justifies the approach of isolating individual populations of cells as opposed to considering such modifications solely in the context of whole blood.

An inability to recapitulate previous findings from EWASs of RA lymphocytes when comparing patients to carefully matched controls suggests many changes may not be specific to RA. Despite a considerably larger sample size than some of the early studies that reported DMPs in these cell types, the number of samples here nonetheless limits power to detect subtle disease-specific modifications. Despite this, findings relating to potential DMR and DVP effects may highlight novel genes and pathways in the pathogenesis of RA and generate hypotheses for future work.

As has been alluded to, there is now an appreciation that the widespread association of DMPs and DMRs with DNA sequence variants means that results from EWASs in isolation may be insufficient to gain mechanistic insight into molecular pathogenesis. The following chapter will integrate genetic data from the patients described here to establish the interaction between genetic and epigenetic variation.

Chapter 4 – Methylation Quantitative Trait Locus Analysis

4.1 Introduction

As described in Chapter 1.6.9, modifications in DNAm can be associated with variants in the genome sequence, such as single nucleotide polymorphism (SNPs), with loci exhibiting such effects termed methylation quantitative trait loci (meQTLs). Given that an individual's genotype is established at fertilisation and remains essentially constant throughout life, identifying molecular traits that are associated with disease-associated SNPs is useful for studying aetiological mechanisms.

In addition, DNAm can be affected by the environment to which cells are exposed, acting at the interface of genetic and environmental risk in complex diseases like RA. Indeed, whilst the genome sequence is static, epigenetic modifications such as DNA methylation are dynamic and differ between cell types. Hence, identifying disease-associated meQTLs may reveal cell-specific mechanisms of molecular pathogenesis.

In this chapter, an meQTL analysis in both CD4$^+$ T cells and B cells from early arthritis patients will be described. The results were integrated with known risk loci from genome-wide association studies (GWAS) of RA, in an attempt to establish loci at which altered DNAm might mediate dysregulated cell function, thus contributing to pathogenesis. Beyond RA risk loci, the contribution of DNAm to genetically conferred disease risk is considered more generally by investigating such effects at loci for two immune-mediated diseases, MS and asthma, as well as OA. Findings were considered in the context of other cell-specific regulatory features from publicly available consortia data, such as chromatin states and transcription factor binding. Finally, interaction analyses were performed to explore the possibility that genotype and disease phenotype might interact to shape the DNA methylome in patients with early disease.

4.2 Cis-meQTL mapping in CD4$^+$ T cells and B cells

To quantify the effects of genetic variation on DNAm levels genome-wide in both CD4$^+$ T cells and B cells, an meQTL analysis was performed in each cell type across all patients. DNAm at CpG sites as defined in the previous chapter (709,412 in CD4$^+$ T cells, 710,445 in B cells) was included in the analysis. Genotype data were available for 103 of the CD4$^+$ T cell samples outlined in the previous chapter (94.4%), whereas for B cell data, genotyping had been performed in 119 patients (91.5%). The demographic data for patients included in the meQTL analyses are described in Table 4.1.

	Rheumatoid Arthritis	**Disease Controls**	**P-value†**
CD4+ T cells:			
Number of patients	*43*	*60*	
Age	58 (50 - 69)	54 (46 - 63)	0.13
Sex (% Female)	67%	70%	0.78
CRP (mg/L)	9 (5 - 13)	7.5 (5 – 13.5)	0.24
ESR (mm/hr)	19 (7 - 32)	15 (9 - 29)	0.74
CCP Positive (%)	47%	-	
RF positive (%)	58%	-	
Tender 28	3 (0 - 11)	2 (0 – 5.5)	0.39
Swollen 28	1 (0 - 3)	0 (0 - 2)	0.51
Diagnosis in disease controls:			
Osteoarthritis		8%	
Other Non-Inflammatory Arthritis		7%	
Spondyloarthropathy (PsA, ReA, EA)		45%	
Crystal Arthropathy		15%	
Other Inflammatory Arthritis		13%	
Other/Undifferentiated		12%	
B cells:			
Number of patients:	46	73	
Age	57 (50 - 68)	55 (46 – 64.5)	0.34
Sex (% Female)	74%	71%	0.75
CRP (mg/L)	9 (5 - 13)	7 (5 – 14.5)	0.14
ESR (mm/hr)	19.5 (5.75 – 34.5)	16 (9 – 30.75)	0.99
CCP Positive (%)	54%	-	
RF positive (%)	61%	-	
Tender 28	3 (1 - 9)	3 (1 – 6)	0.50
Swollen 28	1 (0 - 3)	0 (0 - 3)	0.52
Diagnosis in disease controls:			
Osteoarthritis		8%	
Other Non-Inflammatory Arthritis		7%	
Spondyloarthropathy (PsA, ReA, EA)		48%	
Crystal Arthropathy		11%	
Other Inflammatory Arthritis		8%	
Other/Undifferentiated		18%	

Table 4.1: Demographic and clinical characteristics of all patients included in the meQTL analysis of CD4$^+$ T cells and B cells. Values are presented as either percentages of median (inter-quartile range). P-values to identify potential differences between rheumatoid arthritis and control comparator groups were generated using a Mann-Whitney U Test for continuous data, and a Chi-squared test for categorical variables. CRP = C-reactive protein; ESR = erythrocyte sedimentation rate; CCP = cyclic citrullinated peptide; RF = rheumatoid factor; PsA = psoriatic arthritis; ReA = reactive arthritis; EA = enteropathic arthritis.

MeQTL analyses were limited to SNPs for which the minor allele homozygous genotype was represented by three or more patients or, in the absence of any minor allele homozygotes, at least eight heterozygous patients. Following the removal of those which did not satisfy these criteria, genotypes at 2,901,876 SNPs in CD4$^+$ T cells and 3,035,821 SNPs in B cells were included. Associations between SNP genotype and CpG methylation levels occurring either in cis (SNP-CpG distance < 1Mb) or in trans (SNP-CpG distance ≥ 1Mb or mapping to separate chromosomes) were identified by fitting additive linear models in MatrixEQTL[292]. Quantile-quantile (QQ) plots displaying the expected p-values plotted against the observed p-values for all cis (local p-values) and trans (distant p-values) associations are shown for CD4$^+$ T cells (Figure 4.1 A) and B cells (Figure 4.1B). In total, cis p-values were returned for ~1.357 x 10^9 and 1.421 x 10^9 tests in CD4$^+$ T cells and B cells respectively, whereas the number of trans tests performed was 2.057 x 10^{-12} and 2.155 x 10^{-12}.

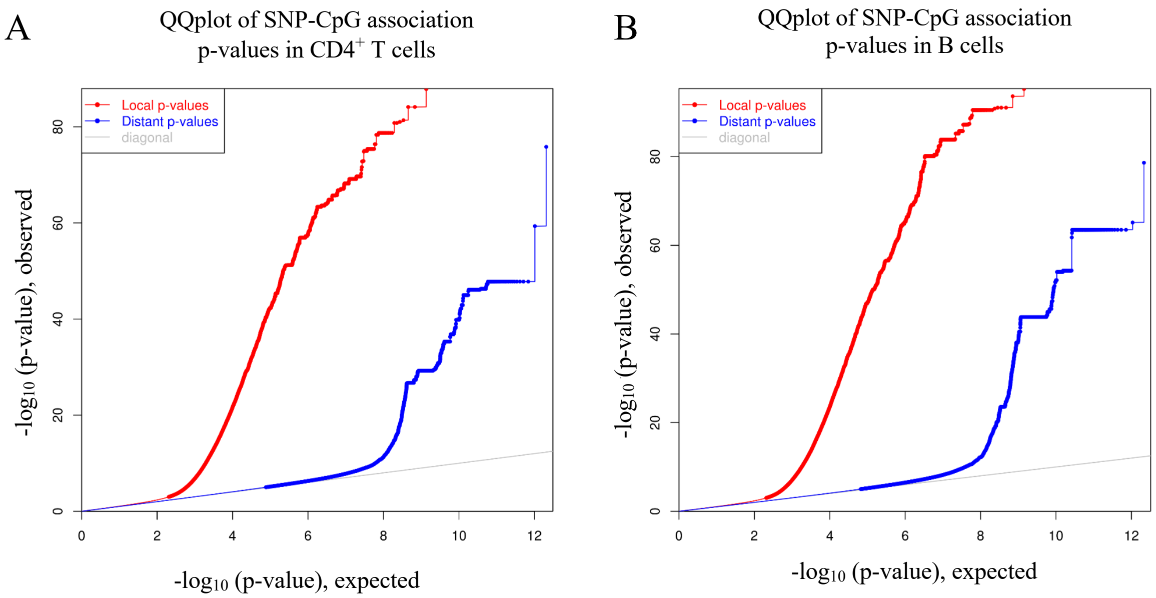


Figure 4.1: QQplots displaying the observed –log$_{10}$ p-values for all SNP-CpG association tests. Linear modelling in MatrixEQTL to test for associations between SNPs and CpGs were performed in cis (local p-values, shown in red) and in trans (distant p-values shown in blue) in **A**) CD4$^+$ T cells and **B**) B cells. The grey x = y line represents that distribution of p-values that would be observed under the null hypothesis that genotype is not associated with DNA methylation levels.

For each type of association (cis and trans), FDR values were calculated separately to adjust for the number of tests, and a threshold of 0.01 selected for cis associations and 1 x 10^{-4} for trans associations. After applying this FDR cut-off, a total of 2,501,652 cis SNP-CpG associations (FDR < 0.01) were identified in CD4$^+$ T cells, with 2,687,897 in B cells. The number of trans associations identified in each cell type were 13,908 and 17,697 in CD4$^+$ cells and B cells respectively (see Figure 4.2 for an overview of the analysis and results described in this chapter).

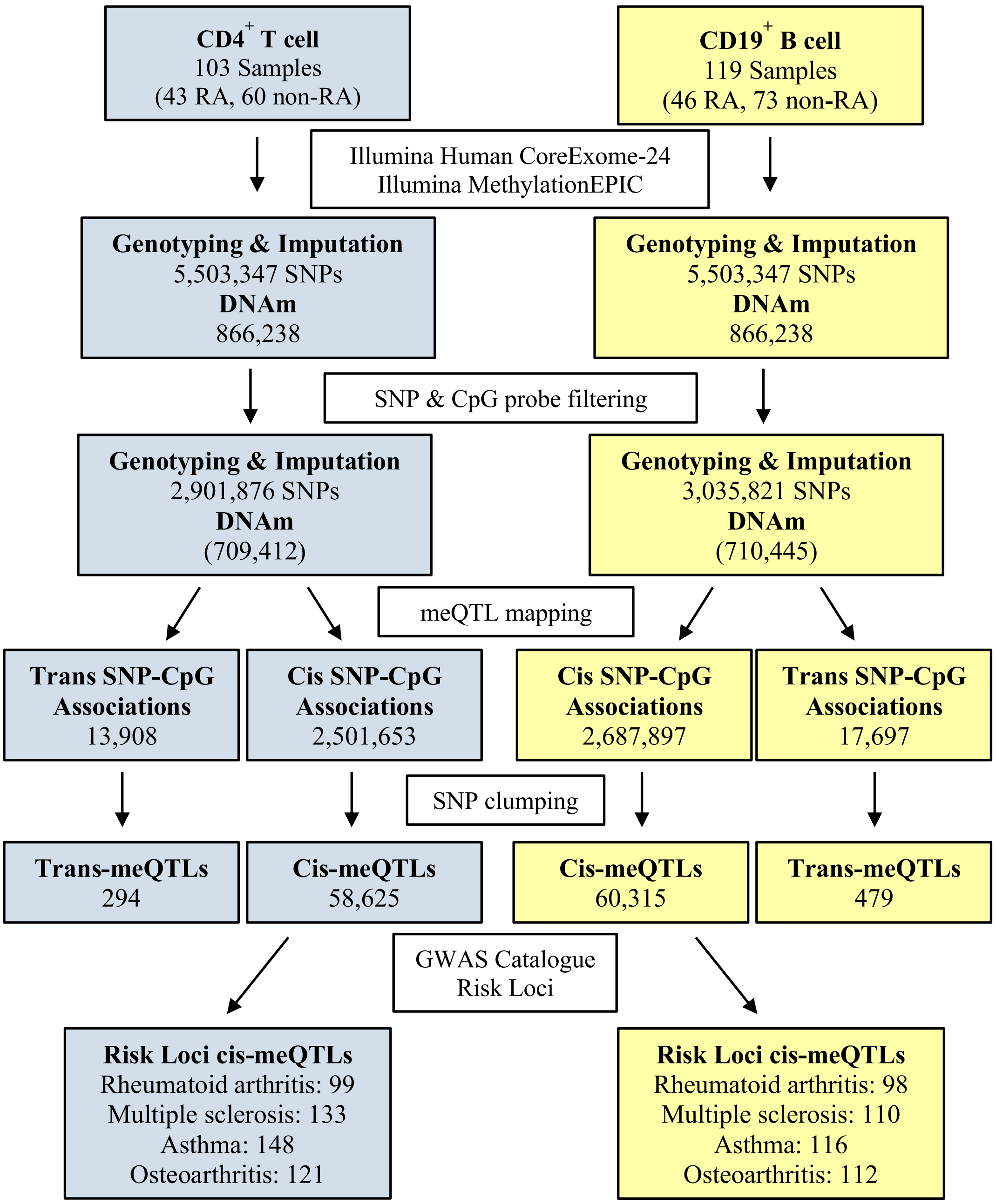


Figure 4.2: Analysis overview and summary of key findings from the meQTL analysis carried out in peripheral CD4$^+$ T cells and B cells from early arthritis patients.

Given that genotypes at SNPs within a LD block are co-inherited, this analysis returns multiple SNPs with FDR < 0.01 for a given CpG, implicating multiple tagging SNPs that are in LD with the causal variant. To collapse these associations into independent signals, SNP clumping was performed – a method which removes all SNPs in LD ($r^2 \geq 0.001$) within a window of 250Kb, retaining the SNP displaying the strongest association (i.e. lowest p-value). SNP clumping reduced the number of independent cis-associations to 58,625 in CD4$^+$ T cells (Figure 4.3 A) and 60,315 in B cells (Figure 4.3 B), which are herein referred to as cis-meQTLs. Conversely, the trans associations were collapsed into 294 (CD4$^+$ T cells; Figure 4.3A) and 479 (B cells; Figure 4.3B) trans-meQTLs.

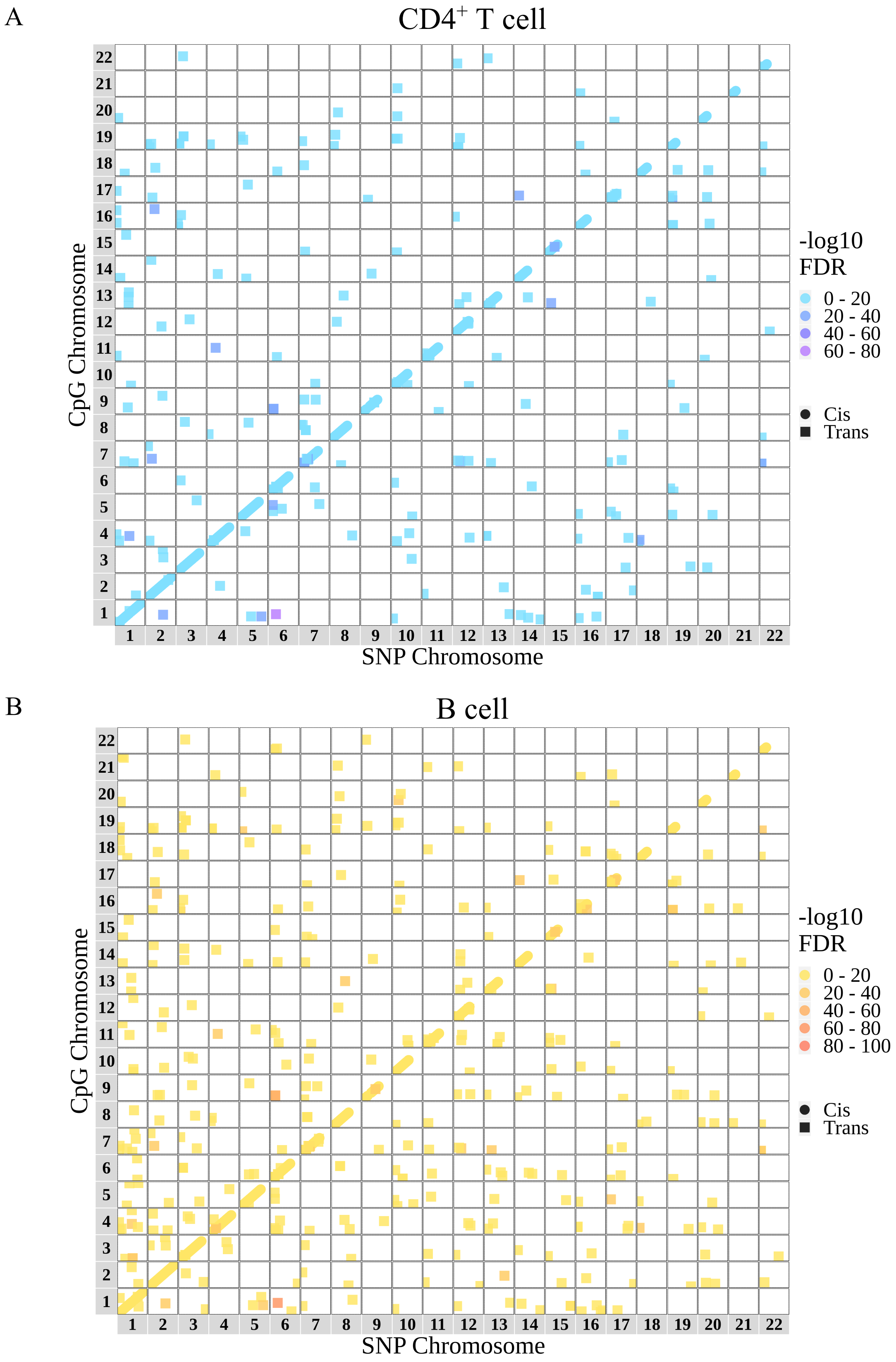


Figure 4.3 Genomic coordinates of meQTLs mapped genome-wide. MeQTLs were mapped in cis and trans in A) CD4$^+$ T cells and B) B cells. Each point represents a significant cis- (shown as circles in the diagonal) or trans- (shown as squares) meQTL association at the selected FDR threshold. The chromosomal coordinates of the regulatory SNP (x axis) are plotted against those for the CpG site (y axis). The colour of each point represents the $-\log_{10}$ p-value of the test, with higher values representing more significant associations.

At CpG sites that were not found to be regulated by SNPs in cis, the β-values (range 0 – 1) displayed a bimodal density distribution, indicating that these sites tend to be fully-methylated (~1) across all samples, or fully de-methylated (~0; Figure 4.4). Contrastingly, CpGs that were associated with cis-meQTLs (cis-CpGs) exhibited more intermediate β-value distributions across probes (Figure 4.4).

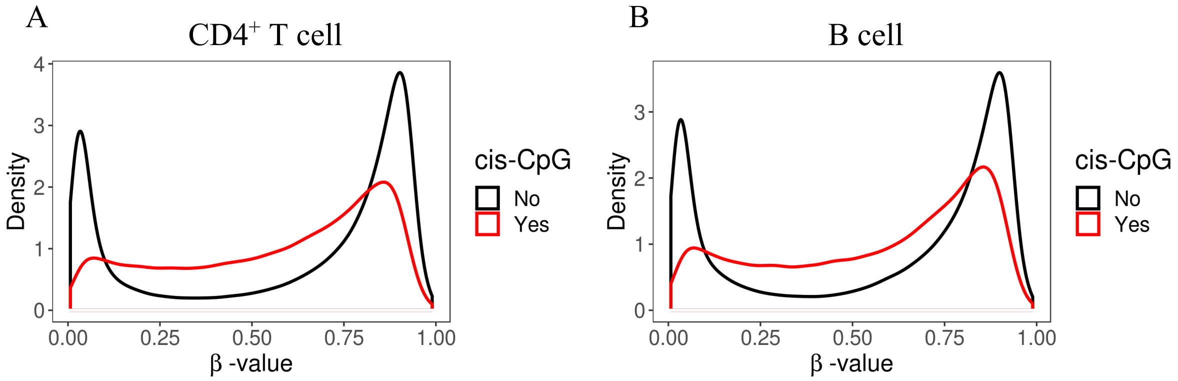


Figure 4.4: Beta value density plots across probes by the presence of absence of cis-regulatory genetic effects. The methylation β-value (range 0 – 1) for probes subject to cis-meQTL effects (cis-CpGs, shown in red), and those that exhibited no such cis-regulation (shown in black).

A preponderance for cis-meQTLs to act over shorter SNP-CpG distances was evident. Indeed, 78% of independent cis associations in CD4$^+$ T cells occurred at a distance of ≤ 100Kb (Figure 4.5A), with 76% of those in B cells mapping to within this distance (Figure 4.5B). Indeed, cis-meQTLs over distances greater than 500Kb were rare, accounting for only ~6% of associations in each cell type (Figure 4.5). Moreover, a general trend whereby the effect size of a cis-meQTL (measured by the absolute regression (β) coefficient) diminished at greater SNP-CpG distances was apparent, again with this observation being consistent across CD4$^+$ T cells (Figure 4.6A) and B cells (Figure 4.6B). These results are consistent with findings from previous studies in a range of tissues[238, 241, 242, 245].

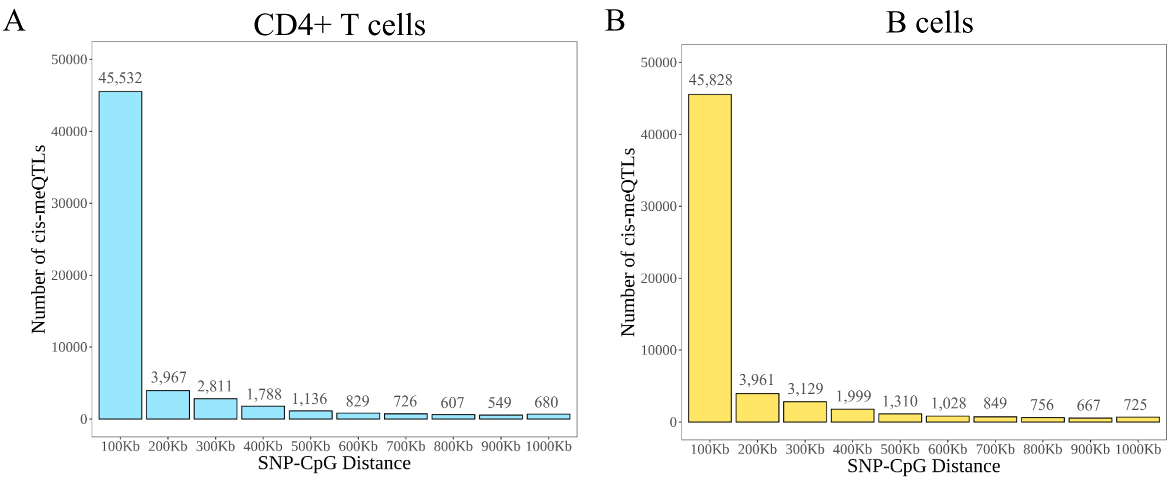


Figure 4.5: Frequency of cis-meQTL effects by SNP-CpG distance. The total number of cis-meQTLs separated into the distance in kilobases (Kb) between the regulatory SNP and CpG site for A) CD4^{+} T cells and B) B cells. Each bin represents the number of cis-meQTLs identified with cis SNP-CpG distance above the previous bin, and up to a distance of the value for that particular bin. For example, the 200Kb bin includes all meQTLs for which the SNP-CpG distances was > 100Kb and ≤ 200Kb.

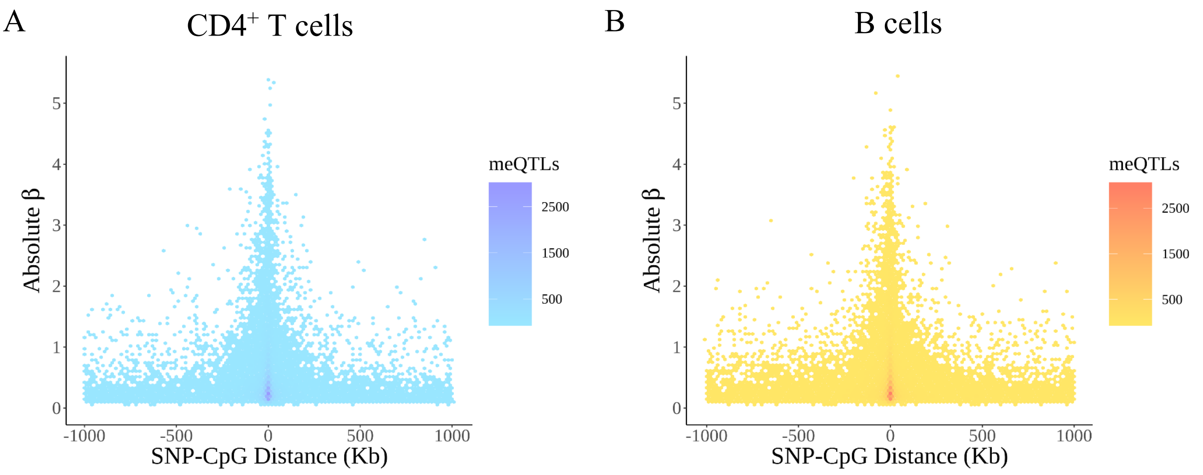


Figure 4.6: cis-meQTL effect size relative to the SNP-CpG distance. The meQTL effects sizes (absolute β) for all cis SNP-CpG associations in A) CD4^{+} T cells and B) B cells are plotted relative to the distance between the lead regulatory SNP and associated CpG site in kilobases (Kb). Effect size is plotted as the absolute regression coefficient (β), indicating the slope of the line from linear models. MeQTLs are plotted in bins, with 100 bins across the length of each axis (i.e. each bin across the x-axis represents a 20Kb SNP-CpG distance). Each point is coloured according to the number of cis-meQTLs falling within that particular bin.

Cis-meQTLs mapped to 53,131 and 53,925 unique CpGs in CD4^{+} T cells and B cells respectively. Previous studies have indicated a considerable degree of cross-tissue effects as regards cis-meQTL activity[126, 127]. Indeed 29,970 CpGs (38.5% of all cis-CpGs) were associated with meQTLs in both cell types (Figure 4.7A). In CD4^{+} T cells, these cis-CpGs were associated with 44,381 SNPs, with 5,495 CpGs associated with > 1 SNP, and 14,244 SNPs associated with > 1 CpG. In B cells, 46,695 cis-regulatory SNPs were identified at these loci, implicating 6,390

CpGs with multiple SNP association and 13,620 SNPs with multiple CpG associations. The number of overlapping cis-CpGs may in fact represent an underestimate of true shared effects between the two cell types, given that two separate analyses were performed with different input data and covariates. Nonetheless, when a cis-meQTL was identified in both cell types, the direction and effect size (regression coefficient (β)) was consistent between cells in a majority of cases (Figure 4.7B). Exemplar plots are shown of cis-meQTLs that were identified uniquely in CD4^{+} T cells (Figure 4.7C), uniquely in B cells (Figure 4.7D), or were found to be active in both cell types (Figure 4.7E).

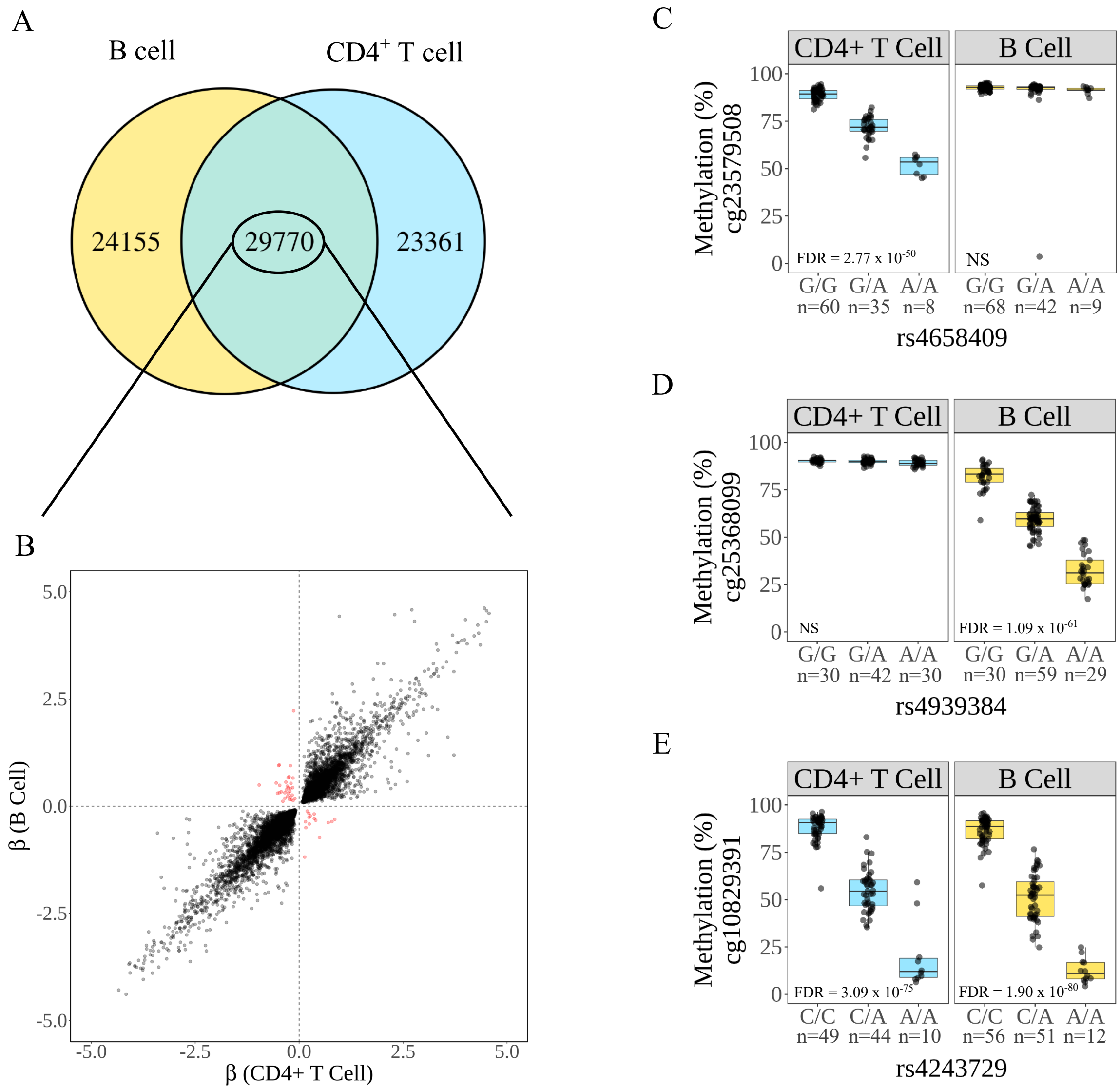


Figure 4.7: Overlapping and unique cis-CpG associations in lymphocyte subtypes. A) The total number of CpGs subject to cis-regulatory effects in CD4^{+} T cells, B cells or both cell types. B) Comparison of effect sizes (regression (β) coefficient) in overlapping cis-meQTL associations between the two cell types. Associations highlighted in red were those whereby the allelic direction of effect differed between the two cell types – i.e. the minor allele conferred increased DNA methylation levels in one cell type, but decreased levels in the other. Exemplar plots are also shown displaying DNA methylation (%) against genotype at cis-meQTLs that were identified uniquely in either (C) CD4^{+} T cells, (D) B cells, or (E) both cell types. NS = not significant (FDR ≥ 0.01).

There were notable exceptions at 47 CpGs whereby the effect of a specific allele on DNAm levels (either conferring increased or decreased levels) appeared to differ between cell types (Figure 4.7B; shown in red). Amongst those exhibiting differential allelic effects were four cis-CpGs (cg04611726, cg03111156, cg14442312, cg21616755) mapping to intron 2 of *CRACR2A* gene on chromosome 12 (Figure 4.8A). The regulatory SNP at this locus (rs241970) exhibited differential allelic effects on DNAm levels at these CpGs between CD4$^+$ T cells and B cells (Figure 4.8B).

Interestingly, the regulatory SNP mapped to a putative CD4$^+$ T cell (Cell ID E043) genic enhancer based on chromatin state information from the Roadmap Epigenomics Consortium[173], whereas the region harbouring rs241970, cg04611726, and cg03111156 is described as 'transcribed' in B cells (Cell ID E032; Figure 4.8A). The cis-CpGs at this locus appear to overlap binding sites of a number of transcription factors, including *IKZF1*, *N2RF1*, *RUNX3*, and *TBX21*. The minor allele (C) at the regulatory SNP reduced DNAm levels at these cis-CpGs in CD4$^+$ T cells, whereas in B cells this allele was associated with an increase in DNAm (Figure 4.8B). Therefore, whilst in general a large overlap was observed in the cis-regulatory landscape of these two lymphocyte subsets, a proportion of cell type-specific effects, including allelic effects that differ between cell types, reinforces the need to study meQTLs in purified populations of cells.

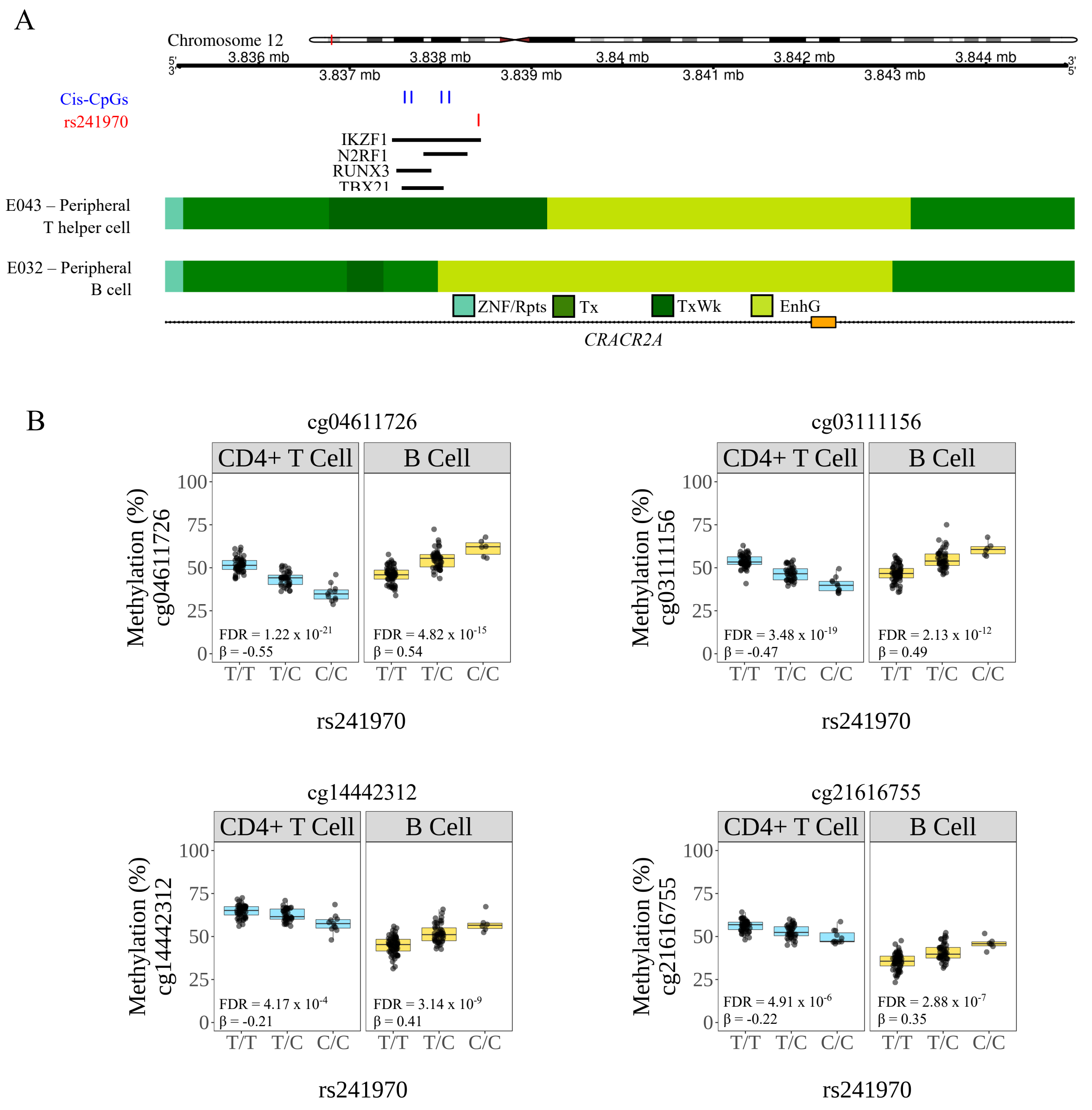


Figure 4.8: Distinct allelic effects on DNA methylation at cis-CpGs mapping to the *CRACR2A* gene in CD4$^+$ T cells and B cells. A) Genomic region plot displaying the *CRACR2A* promoter region and location of the cis-meQTLs on chromosome 12 (region is shown as a red line on the top ideogram plot). The regulatory SNP (rs241970) is shown as a red line with all cis-CpGs (cg04611726, cg03111156, cg14442312, cg21616755) from left to right shown as blue lines. Chromatin state tracks from the Roadmap[173] 15-state learning model are shown for primary T helper cells from peripheral blood (E043) and primary B cells from peripheral blood (E032). Transcription factor binding sites (TFBSs) from the ENCODE ChIP-seq[209,210] experiments in the GM12878 Epstein-bar virus-transformed B cell line (TFBS are shown only for those TFs overlapping the cis-meQTL region with a score > 800*). The orange box on the *CRACR2A* gene track plot represents Exon 2 of this gene. B) Cis-meQTL associations at these cis-CpGs in both cell types. ZNF/Rpts = ZNF genes + repeats; Tx = strong transcription; TxWk = weak transcription; EnhG = genic enhancer.

*Scores designated by the UCSC Browser (range 0-1000) and represent a measure of signal intensity from the analysis pipeline of ENCODE ChIP-seq experiments.

4.3 Disease-specific meQTLs

Numerous studies have described a family of eQTLs that appear to become active upon cell stimulation, such as may occur when a cell is exposed elevated cytokines and extracellular stimuli during an inflammatory response.

To test the hypothesis that certain meQTLs may be impacted by RA-specific factors that lymphocytes are exposed to, an interaction analysis was performed to assess the influence of disease diagnosis (RA or non-RA) on the effect size of cis- and trans-meQTLs. This approach has greater power to detect disease-specific effects than simply running an meQTL analysis in each cohort separately, and detects instances whereby:

- The meQTL is active in on comparator group (RA or non-RA) but not the other (Figure 4.9A).
- The meQTL is active in both groups with an opposing allelic effect on DNAm levels in each (Figure 4.9B).
- The meQTL is active in both groups with a consistent allelic effect, but a significant difference in the magnitude of the allelic effect (Figure 4.10C).

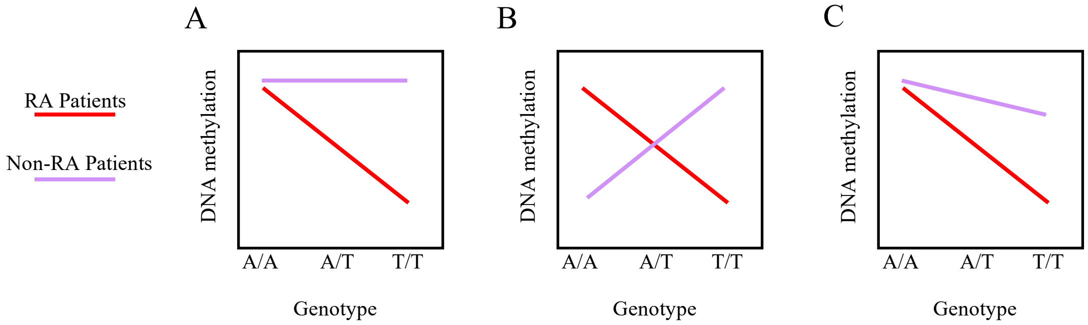


Figure 4.9: Potential disease-specific meQTLs that could be detected by an interaction analysis. A) An meQTL is active in the patient cohort but absent in the controls, B) An meQTL is active in both the patient and control cohorts, though opposing allelic effects on DNAm are present, C) An meQTL is active in both the patient and control cohort, though the magnitude of effect differs (i.e. the minor allele confers a greater reduction in DNAm in one cohort relative to the other).

Only one such interaction effect was identified in cis, which occurred between rs13145446 (chr4:152,729,035) and cg23683081 (chr4:152,682,891) in B cells (Figure 4.10; Table 4.2; genotype × diagnosis interaction FDR = 0.0155). This cis-meQTL displayed opposing allelic effects in each patient cohort, with the minor allele (G) conferring a small increase in DNAm in non-RA patients (p-value = 0.0034), whilst in RA patients this allele was associated with a reduction in DNAm levels (p-value = 2.69 x 10^{-5}; Figure 4.10). This particular CpG maps to region upstream of the *GATB* gene (also names *PET112*), which encodes Glutamyl-TRNA Amidotransferase Subunit B.

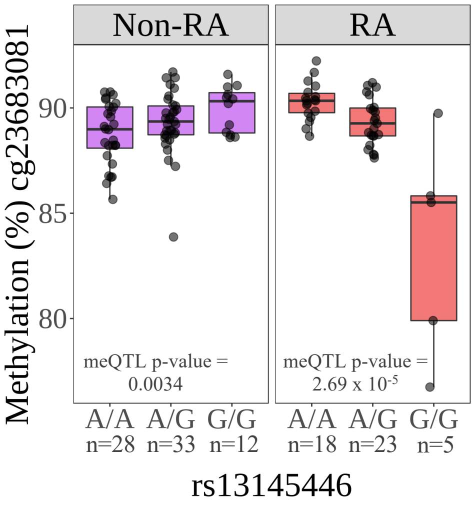


Figure 4.10: Genotype × Diagnosis cis-meQTL interaction in B cells. A cis meQTL was identified at rs13145446 (chr4:152,729,035) at which the minor allele (G) conferred reduced methylation at cg23683081 (chr4:152,682,891).

A number of meQTL interaction effects acting in trans were identified, with three in CD4$^+$ T cells and 21 identified in B cells (Table 4.2). All of the CD4$^+$ T cell interaction trans-meQTLs were inter-chromosomal (Figure 4.11A), as were 20/21 of the B cell trans interactions (Figure 4.11B).

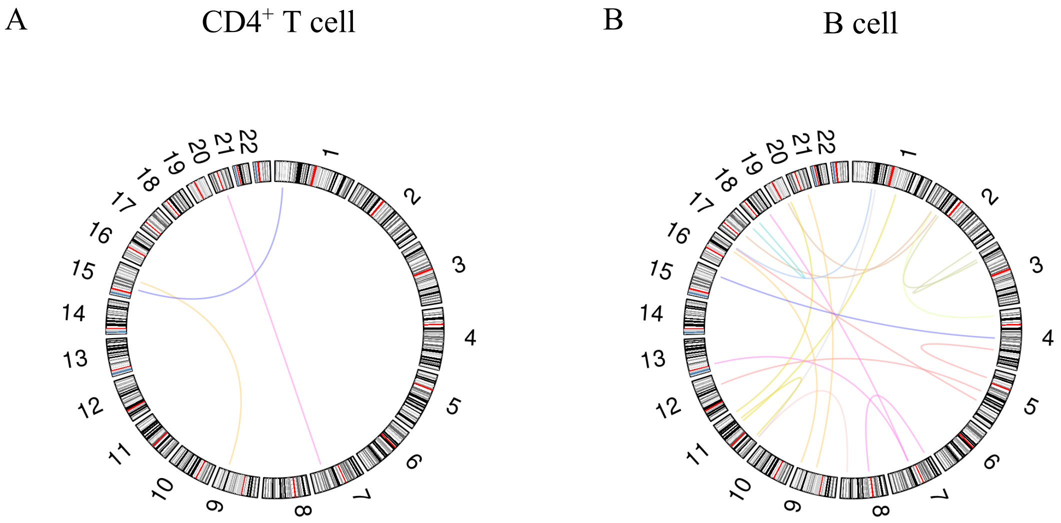


Figure 4.11: Genotype × diagnosis interaction effects occurring in trans. Circos plots displaying the genomic locations of inter-chromosomal trans-meQTLs exhibiting significant interactions in A) CD4$^+$ T cells and B) B cells (See Table 4.2 for details of all interaction effects). The colour of each link depicts the chromosome to which the meQTL variant maps.

SNP	SNP Coord	Minor allele	Major allele	CpG	CpG Coord	P-val Interaction	FDR Interaction	Beta Interaction	P-val RA	Beta RA	P-val nRA	Beta nRA	UCSC Gene
CD4^{+} T cells													
rs7495444	chr15:34897222	T	A	cg20399213	chr1:27020750	1.43 x 10^{-12}	8.20 x 10^{-5}	0.332	8.91 x 10^{-5}	0.175	1.79 x 10^{-5}	-0.122	-
rs2254156	chr9:113189533	A	G	cg26097711	chr15:64749577	2.63 x 10^{-12}	0.0001	-1.051	2.23 x 10^{-8}	-1.029	0.8621	0.013	-
rs969916	chr7:114727332	A	C	cg05886966	chr20:34194255	2.42 x 10^{-12}	0.0001	-0.423	5.38 x 10^{-5}	-0.214	2.20 x 10^{-7}	0.218	*FER1L4*
B cells													
rs13145446	chr4:152729035	G	A	cg23683081	chr4:152682891	3.01 x 10^{-9}	0.0155	-0.456	2.69 x 10^{-5}	-0.338	0.0034	0.114	*GATB*
rs10172895	chr2:34553877	C	T	cg00710196	chr17:36870572	2.96 x 10^{-12}	5.59 x 10^{-5}	-0.543	3.02 x 10^{-7}	-0.395	0.0044	0.121	*MLLT6*
rs1864180	chr5:88421363	A	G	cg23449856	chr12:92474554	9.87 x 10^{-12}	0.0001	0.574	3.40 x 10^{-6}	0.337	8.35 x 10^{-6}	-0.219	*C12orf79*
rs7164870	chr15:86252755	G	A	cg12117684	chr4:96077623	9.07 x 10^{-11}	0.0002	2.315	0.1429	0.262	3.63 x 10^{-11}	-2.024	*BMPR1B*
rs11693561	chr2:56282762	G	A	cg00456774	chr19:40857787	3.19 x 10^{-11}	0.0006	-0.330	1.64 x 10^{-6}	-0.242	7.48 x 10^{-5}	0.106	*PLD3*
rs4843534	chr16:87172625	A	G	cg27626037	chr1:68096297	7.89 x 10^{-11}	0.0008	-0.460	2.22 x 10^{-5}	-0.290	0.0004	0.142	-
rs7802290	chr7:78116293	C	T	cg01625242	chr18:56886915	3.88 x 10^{-11}	0.0005	1.556	6.40 x 10^{-5}	1.236	0.5471	-0.042	*GRP*
rs157624	chr4:16704510	C	T	cg12351039	chr2:38229188	4.89 x 10^{-13}	1.27 x 10^{-5}	-0.334	1.53 x 10^{-7}	-0.220	0.0001	0.105	*RMDN2*
rs11144219	chr9:77698483	A	G	cg23252815	chr20:44420276	3.93 x 10^{-11}	0.0005	-1.770	0.0014	-0.639	1.38 x 10^{-8}	1.138	*WFDC3*
rs1476772	chr7:13152162	T	G	cg15834993	chr8:49826049	6.36 x 10^{-11}	0.0007	0.432	0.0019	0.246	5.23 x 10^{-8}	-0.186	-
rs1286733	chr3:25613372	C	T	cg25371991	chr2:65062088	1.96 x 10^{-11}	0.0002	-0.539	2.71 x 10^{-5}	-0.309	1.72 x 10^{-6}	0.238	-
rs2029543	chr5:125044634	G	A	cg04933361	chr16:83991376	4.76 x 10^{-11}	0.0005	0.400	3.54 x 10^{-5}	0.238	4.01 x 10^{-5}	-0.165	*OSGIN1*
rs7802290	chr7:78116293	C	T	cg11284281	chr13:20967659	5.91 x 10^{-12}	9.80 x 10^{-5}	-1.736	4.58 x 10^{-5}	-1.338	0.3072	0.046	-
rs4798720	chr18:8936107	G	A	cg09382492	chr17:74463681	9.10 x 10^{-11}	0.0009	-0.367	5.46 x 10^{-5}	-0.206	6.51 x 10^{-6}	0.178	*AANAT*
rs8176635	chr9:136152009	A	G	cg27244120	chr16:70610434	7.70 x 10^{-11}	0.0008	0.436	0.0001	0.262	4.21 x 10^{-6}	-0.193	*SF3B3*
rs12283663	chr11:89410440	A	C	cg21464983	chr1:156919932	9.70 x 10^{-11}	0.0010	-0.357	3.21 x 10^{-7}	-0.223	0.0014	0.111	*ARHGEF11*
rs10874175	chr1:81282234	C	T	cg11832601	chr11:8012913	1.60 x 10^{-11}	0.0002	-1.266	6.39 x 10^{-7}	-1.218	0.1016	0.110	*EIF3F*
rs7342215	chr11:129139026	A	G	cg11381655	chr19:51818377	9.55 x 10^{-12}	0.0001	-0.944	2.37 x 10^{-5}	-0.888	0.1496	0.053	*IGLON5*
rs658286	chr11:100942792	G	T	cg16519321	chr11:17741243	2.17 x 10^{-11}	0.0003	-0.340	3.62 x 10^{-7}	-0.271	0.0015	0.092	*MYOD1*
rs12054998	chr5:35839347	C	T	cg02197923	chr4:138622842	2.12 x 10^{-13}	6.42 x 10^{-6}	-0.665	1.76 x 10^{-7}	-0.715	0.6907	0.010	-
rs11775667	chr8:128253752	A	G	cg11832601	chr11:8012913	3.99 x 10^{-11}	0.0005	-1.269	1.52 x 10^{-7}	-1.319	0.6835	0.028	*EIF3F*
rs12487048	chr3:16741847	G	C	cg27391564	chr2:240530497	1.34 x 10^{-15}	8.77 x 10^{-8}	-1.052	9.64 x 10^{-10}	-0.979	0.2509	0.054	-

Table 4.2 (previous page): Me-QTLs (Cis and Trans) exhibiting significant genotype × diagnosis interaction effects in CD4$^+$ T cells and B cells. SNP Coord = Chromosomal co-ordinates of the regulatory SNP; CpG Coord = Chromosomal co-ordinates of the cis-CpG; P-val Interaction = P-value for the interaction effect from the linear model; FDR Interaction = Interaction p-value adjusted for false discovery rate (Benjamini-Hochberg); Beta Interaction = Regression Coefficient (β) of the interaction effect (i.e. impact of disease diagnosis on effect size); P-val RA = P-value of the meQTL association in rheumatoid arthritis patients; Beta RA = Regression Coefficient (β) of the meQTL association in rheumatoid arthritis patients; P-val NRA = P-value of the meQTL association in non-rheumatoid arthritis patients (disease controls); Beta NRA = Regression Coefficient (β) of the meQTL association in non-rheumatoid arthritis patients; UCSC Gene = Gene to which the cis-CpG maps based on the Illumina MethylationEPIC manifest file.

4.4 Trans-meQTL mapping CD4$^+$ T cells and B cells

As mentioned in section 4.2, a relatively small number of trans-meQTLs were identified, mapping to 239 trans-CpGs in CD4$^+$ T cells and 387 in B cells (Figure 4.12A). Of the trans-meQTLs, the majority of these were inter-chromosomal (81.2% in CD4$^+$ T cells, 83.5% in B cells). Regarding trans-CpGs, 139 were subject to trans-meQTL effects in both cell types (Figure 4.12A). An exemplar of one such a shared inter-chromosomal trans-meQTL is shown between rs1655530 (chr1:55,074,463) and cg19586483 (chr6:110,512,070; Figure 4.12B). It should be noted that the possibility cannot be ruled out that trans effects involving SNPs and CpGs mapping to the same chromosome represent long-distance cis effects, beyond the 1Mb pre-defined window for cis associations. Indeed, of such long-distance meQTLs identified in the trans analysis, 53.3% of the regulatory variants in CD4$^+$ T cells were also identified in the cis-meQTL analysis, whilst the proportion in B cells was 65.6%, indicating that in many instances this is likely the case.

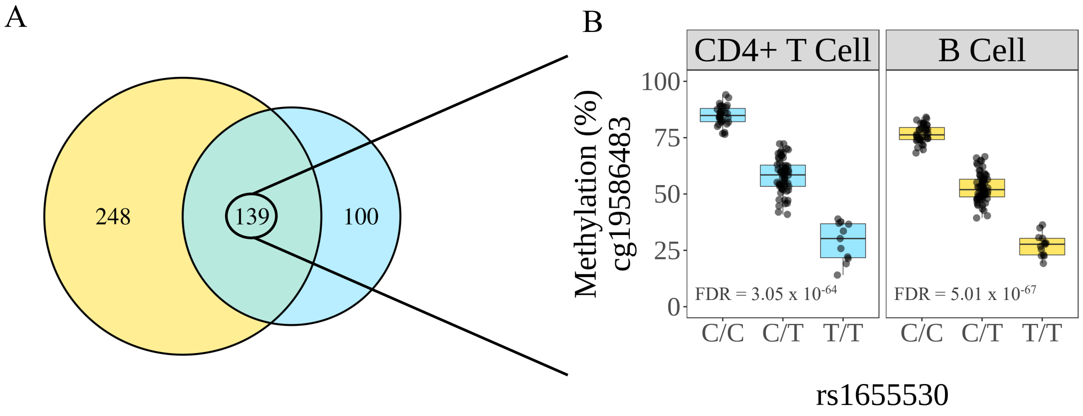


Figure 4.12: Trans-meQTL analysis in CD4+ T cells and B cells. A) Overlap of trans-CpGs identified in either cell type. B) Exemplar plot of inter-chromosomal trans-meQTL identified in both cell types.

Given that the multiple-testing burden associated with identifying associations in trans is considerably greater than for those in cis, the current analysis was mainly powered to map the latter. As such, the remainder of this chapter will largely focus on these associations.

4.5 Functional annotation of cis-meQTLs

The genomic mapping of cis-meQTLs can potentially give insight into their functional consequences at the level of transcriptional regulation. Cis-CpGs that were found to be associated with meQTLs in each lymphocyte cell types, as well as those that were not associated with a genetic variant in cis (non-cis-CpGs) were therefore overlapped with additional marks indicative of regulatory activity from previously annotated large consortia datasets. CpGs were mapped to chromatin states from the Roadmap Epigenomics Project[173] 15-state model, as defined by ChromHMM learning on five histone modifications (see Chapter 2.7.4; Figure 4.13A). Cell-specific chromatin state data from 'primary T helper cells from peripheral blood' (E043) and 'primary B cells from peripheral blood' (E032) were selected for intersecting CD4$^+$ T cell and B cell CpGs, respectively. The 15 chromatin states were aggregated into five states broadly reflecting genomic regions with unique functionality (transcription start sites (Tss); flanking transcription start sites (TssFlnk); transcribed (Tx); enhancer (Enh); and repressed (Rep)). The enrichment of cis-CpGs was compared with all other CpGs which were included in the analysis but found not to be subject to cis-regulatory activity (656,281 in CD4$^+$ T cells, 656,520 in B cells). CpGs were also overlapped with information on CGI features (Figure 4.13B) and UCSC RefGene Genic features (Figure 4.13C) obtained from the Illumina MethylationEPIC manifest.

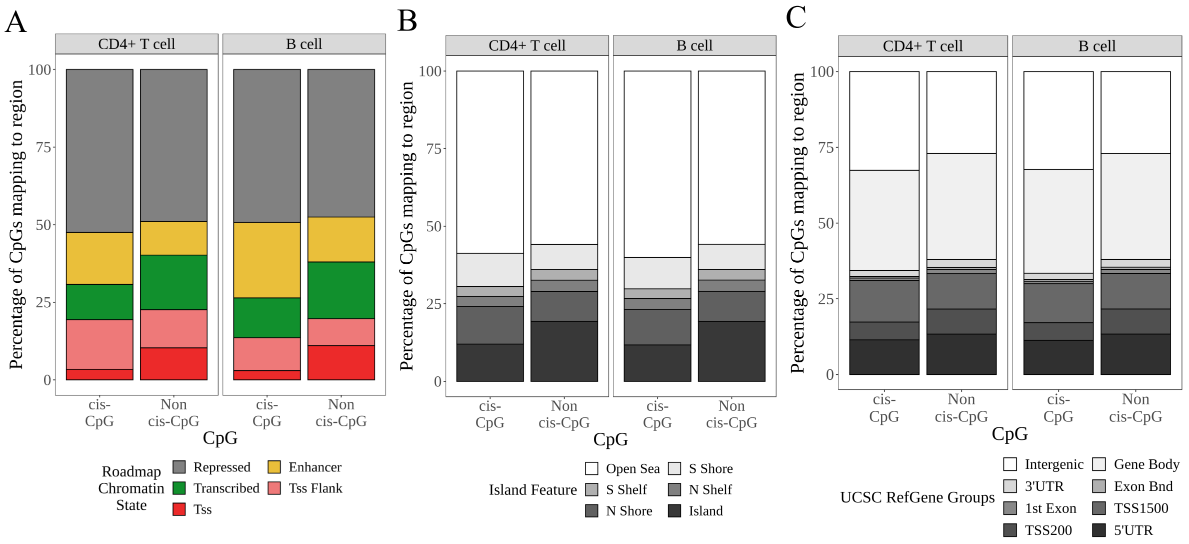
As regards chromatin states, in both CD4$^+$ T cells and B cells, cis-CpGs were found to be significantly enriched in both repressed chromatin and enhancer regions in CD4$^+$ T cells and B cells alike (Figure 4.13A & D-E). Conversely, significant depletion at transcription start sites and transcribed regions was observed, again with this effect being consistent across both cell types (Figure 4.13A & D-E). Though the relative proportions of cis-CpGs were significantly reduced at TSSs, increased levels were found at the regions flanking the TSS, again with this trend observed in CD4$^+$ T cells and B cells (Figure 4.13A & D-E).

CGIs extracted from UCSC annotations[322], as defined in the MethylationEPIC manifest file, are regions with a high density of CpGs found at the promoter of ~60% of human genes[184]. CpGs in the dataset were mapped to CGIs, as well as 'shores' (0-2Kb from a CGI), 'shelves' (2-4Kb from a CGI), and 'open sea' regions (all other mappings; see Chapter 2.7.4). Most strikingly, cis-CpGs were strongly depleted at CGIs in both CD4$^+$ T cells and B cells (Figure

4.13B & D-E). Conversely, significant enrichment of cis-CpGs was evident at the CGI shore regions both 5' (north shore) and 3' (south shore) of CGIs, again with this trend clear in both CD4$^+$ T cells and B cells (Figure 4.13B & D-E). A smaller magnitude of enrichment, albeit still highly significant, was found at 'open sea' regions in both cell types, whereas a small reduction in the proportion of cis-CpGs mapping to shelves was evident (excluding north shelves in B cells; Figure 4.13B & D-E).

Finally, CpG locations relative to specific RefGene sequence features were identified and a test of relative cis-CpG enrichment at each feature performed as before. As would perhaps be expected given the depletion seen at TSS (Figure 4.13A) and CpG islands (Figure 4.13B), CD4$^+$ T cell cis-CpGs were under-represented at regions such as the 5'UTR, the 1st Exon, and the 200bp region immediately upstream of the TSS (Figure 4.13C & D-E). This was also reflected in the B cell data with comparable depletions of cis-CpGs at these regions. Significant depletion was also evident at the 3'UTR genes, as well as the Exon boundaries and gene bodies (Figure 4.13C & D-E). With regards to cis-CpG enrichment, significantly increased mapping to intergenic sites and the regions 200-1500bp upstream of TSSs occurred in both cell types (Figure 4.13C & D-E).

Taken together, these findings illustrate that CpGs subject to cis-regulatory activity by local sequence variation are not randomly distributed throughout the genome, but rather are generally depleted at features associated with gene promoters and enriched at regions adjacent to these promoters, or at greater distances such as enhancers.



D

		Roadmap Chromatin State					Relation to CpG Island						UCSC RefGene Group							
		Rep	Enh	Tx	Tss Flnk	Tss	Open Sea	S Shore	S Shelf	N Shelf	N Shore	Island	Inter-genic	Gene Body	3’ UTR	Exon Bnd	1^{st} Exon	Tss 1500	Tss 200	5’ UTR
CD4+ T cell	Cis-CpG (%)	52.5	16.7	11.4	16.0	3.4	58.7	10.8	3.1	3.2	12.2	12.0	32.6	33.0	2.0	0.5	0.9	13.6	5.9	11.4
	Non cis-CpG (%)	49.0	10.8	17.6	12.3	10.3	55.9	8.2	3.4	3.6	9.6	19.4	27.0	35.0	2.5	0.8	1.3	11.6	8.3	13.4
B cell	Cis-CpG (%)	49.3	24.3	12.8	10.6	3.0	60.0	10.2	3.1	3.5	11.5	11.7	32.3	34.2	2.1	0.5	0.9	13.6	5.9	11.4
	Non cis-CpG (%)	47.5	14.5	18.3	8.7	11.0	55.8	8.2	3.4	3.6	9.7	19.4	27.1	34.9	2.5	0.8	1.3	11.6	8.3	13.4

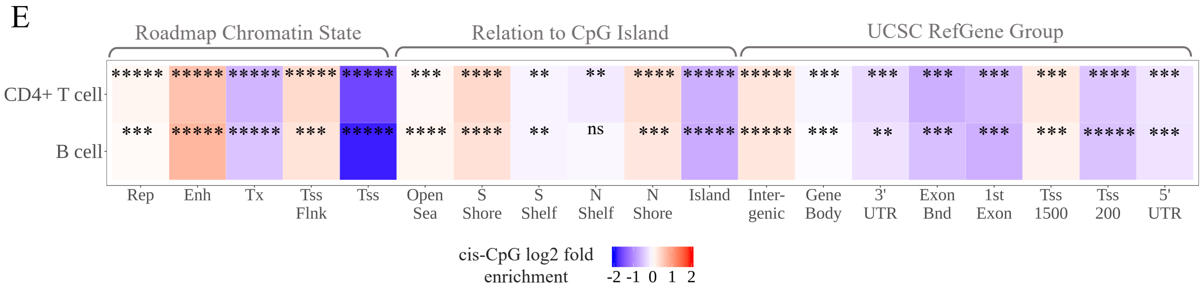


Figure 4.13: Cis-CpG feature enrichment. CpGs at which DNAm levels were associated with cis-meQTLs (cis-CpGs) and those which were not subject to such associations (non cis-CpG) in CD4^{+} T cells and B cells were mapped to (A) Roadmap Epigenomics Project Consortium[172] chromatin states in primary T helper cells from the peripheral blood (cell ID E043; CD4+ T cell CpGs) and primary B cells from the peripheral blood (cell ID E032; B cell CpGs). The 15-state model defined by Roadmap was collapsed into 5 distinct functional states: Transcription start site (TSS), Flanking a TSS (Tss Flnk), Enhancer, Transcribed, and Repressed (see section 2.7.4 for further details on definition of chromatin states). Cis-CpGs and non-cis-CpGs were also mapped to (B) CpG island features (Shore regions are defined as those 0-2Kb from CGI boundaries, with shelves within 2-4Kb., All other regions are defined as open sea.) and (C) UCSC RefGene Gene Features (Exon Bnd = Exon Boundary, 3’UTR/5’UTR = 3’/5’ untranslated region, TSS1500 = 200-1500 bases upstream the gene transcription start site; TSS200 = 0-200 bases upstream of the gene transcription start site). CpG island/RefGene features were obtained from the Illumina Infinium MethylationEPIC manifest. (D) The percentage of cis-CpGs and non-cis-CpGs mapping to each feature in panels A-C for both cell types. (E) Heat map of relative log$_2$-fold enrichment of cis-CpGs at each feature in both cell types, with red depicting enrichment of cis-CpGs at a feature and blue depicting depletion. ***** $p < 1 \times 10^{-100}$; **** $p < 1 \times 10^{-50}$; *** $p < 1 \times 10^{-10}$; ** $p < 0.01$; * $p < 0.05$.

4.6 Co-localisation of cis-meQTLs with RA loci

Whilst GWASs have revealed over 100 genomic loci that appear to confer disease susceptibility in RA[97], linking such genetic modification to molecular pathways that confer dysregulated cellular immunity still remains a difficult challenge. It was therefore hypothesized that many such non-coding variants may in fact modulate DNAm levels at proximal or distal CpG sites, with potential implications for transcriptional regulation. To test this hypothesis, RA GWAS data were downloaded from the GWAS catalogue[293] (see section 2.7.5 for inclusion criteria). In the first instance, co-localisation was defined whereby the meQTL variant mapped to the LD block harbouring the GWAS variant (all SNPs with $r^2 \geq 0.8$ with the lead RA GWAS SNP). After defining RA risk loci using this approach, a total of 104 and 107 such loci were represented in the patient genotype input data for CD4$^+$ T and B cells respectively. In total, 32 risk loci displayed cis-regulatory activity on DNAm levels in CD4$^+$ T cells and 33 in B cells, with 24 of these identified in both cell types (Table 4.3 & Figure 4.14). These effects encompassed a total of 99 CpGs subject to cis-regulation at RA risk loci in CD4$^+$ T cells, and 98 in B cells (Table 4.3)

The cis-CpGs associated with RA risk loci mapped to a number of genes with putative roles in lymphocyte-mediated RA pathogenesis in CD4$^+$ T cells (Figure 4.14A) and B cells (Figure 4.14B; Table 4.3). A number of cis-meQTLs were active in both cell types and were associated with cis-CpGs that mapped to genes including *MMEL1* (cg21621858) and *JAZF1* (cg11187739). In many cases, the cis-CpG exhibiting the strongest association with the cis-meQTL differed between the two cell types. Such examples of this include cis-CpGs mapping to *FCRL3* (cg17134153/cg21721331), *EOMES* (cg21473142/cg20235057), and *SYNGR* (cg24268161/cg15105517) (Figure 4.14). Particularly strong allelic effects on DNAm were observed at the locus on chromosome 11q12.2 harbouring the RA risk variant (rs968567; chr11:61,828,092). Associations with methylation at 25 CpGs in CD4$^+$ T cells and 23 CpGs in B cells were identified at this locus, most of which map to the *FADS2* gene (Table 4.3). Exemplar plots are shown for the SNP-CpG associations for the top three meQTL cis-CpGs in CD4$^+$ T cells; cg21029357 (chr11:61,601,062; Figure 4.15A), cg13299762 (chr11:61,594,708; Figure 4.15B), and cg27386326 (chr11:61,587,980; Figure 4.15C), as well as the top three cis-CpGs in B cells; cg19481605 (chr11:61,596,812; Figure 4.15D), cg13299762 (chr11:61,594,708; Figure 4.15E), and cg01400685 (chr11:61,598,025; Figure 4.15F). Interestingly, the risk allele at this locus conferred both increased DNAm and decreased DNAm levels at different CpGs across this region, such as is seen for cg21029357 (Figure 4.15A) and cg27386326 (Figure 4.15C).

Lead RA SNP	Locus	CpGs associated w/ LD block	CD4+ T cell FDR	CD4+ T cell SNP	r^2 with lead RA SNP	Bayes Coloc. (PP3)	Bayes Coloc. (PP4)	B cell FDR	B cell SNP	r^2 with lead RA SNP	Bayes Coloc. (PP3)	Bayes Coloc. (PP4)	UCSC Gene Symbol†
chr1:2523811	1p36.32	cg21621858	6.99×10^{-28}	rs2260976	0.98	0.99	0.01	1.02×10^{-24}	rs10752747	0.99	0.99	0.01	*MMEL1*
		cg11382394	3.34×10^{-15}	rs2260976	0.98	0.83	0.17	2.61×10^{-18}	rs4648657	0.97	0.33	0.67	*MMEL1*
		cg00205605	4.51×10^{-7}	rs10797431	0.86	0.70	0.30	2.54×10^{-12}	rs10797431	0.86	0.94	0.06	-
		cg18932078	2.21×10^{-11}	rs2843404	0.99	0.44	0.56	4.49×10^{-4}	rs3828154	0.91	0.20	0.79	*MMEL1*
		cg15574442	-	-	-	-	-	1.09×10^{-5}	rs10797431	0.86	0.74	0.26	-
		cg17429686	2.93×10^{-5}	rs60389769	0.96	0.60	0.40	2.18×10^{-5}	rs6667564	0.90	0.04	0.96	-
		cg00070241	8.95×10^{-5}	rs2843404	0.99	0.33	0.66	-	-	-	-	-	*TTC34*
		cg25372239	-	-	-	-	-	4.70×10^{-4}	rs2764845	0.90	0.24	0.76	*PRDM16*
		cg08030037	-	-	-	-	-	0.004	rs10797431	0.90	0.13	0.87	*MMEL1*
		cg05392440	-	-	-	-	-	0.010	rs10797431	0.90	0.15	0.70	-
rs12140275	1p34.3	cg12871964	-	-	-	-	-	9.15×10^{-13}	rs7553012	0.93	0.12	0.88	-
		cg06891056	-	-	-	-	-	1.70×10^{-4}	rs7553012	0.93	0.06	0.94	-
rs624988	1p13.1	cg21456300	0.010	rs771587	0.99	0.01	0.89	-	-	-	-	-	*IGSF3*
rs2317230	1q23.1	cg21721331	6.68×10^{-15}	rs2210913	0.88	0.00	0.99	2.12×10^{-25}	rs2210913	0.88	0.00	0.99	*FCRL3*
		cg19602479	2.53×10^{-12}	rs2210913	0.88	0.01	0.99	5.28×10^{-25}	rs2210913	0.88	0.00	0.99	*FCRL3*
		cg17134153	5.90×10^{-18}	rs2210913	0.88	0.00	0.99	8.42×10^{-5}	rs7522061	0.80	0.05	0.94	*FCRL3*
		cg01045635	1.28×10^{-14}	rs2210913	0.88	0.00	0.99	1.09×10^{-14}	rs7522061	0.80	0.03	0.97	*FCRL3*
		cg08786003	1.71×10^{-10}	rs2210913	0.88	0.00	0.99	-	-	-	-	-	*FCRL3*
		cg15602298	-	-	-	-	-	6.05×10^{-8}	rs7522061	0.80	0.07	0.92	*FCRL3*
		cg25259754	2.79×10^{-7}	rs7522061	0.80	0.02	0.98	1.26×10^{-6}	rs7522061	0.80	0.02	0.98	*FCRL3*
		cg18707136	0.002	rs7522061	0.80	0.04	0.94	-	-	-	-	-	*FCRL1*
		cg04429688	-	-	-	-	-	0.004	rs7522061	0.80	0.01	0.99	*FCRL3*
rs10175798	2p23.1	cg27371770	1.14×10^{-13}	rs10173253	0.94	0.00	0.99	9.75×10^{-4}	rs10173253	0.94	0.00	1.00	-
rs934734	2p14	cg11674355	-	-	-	-	-	8.28×10^{-5}	rs4494728	0.90	0.21	0.79	*SPRED2*
rs3806624	3p24.1	cg20235075	-	-	-	-	-	1.51×10^{-9}	rs9866625	0.96	0.00	1.00	*EOMES*
		cg21473142	3.64×10^{-9}	rs3806624	1.00	0.01	0.99	2.57×10^{-8}	rs3806624	1.00	0.01	0.99	*EOMES*
rs11933540	4p15.2	cg06535121	0.003	rs6448432	0.88	0.01	0.99	-	-	-	-	-	-
rs6859219	5q11.2	cg21124310	9.15×10^{-8}	rs6859219	1.00	0.00	0.99	-	-	-	-	-	*ANKRD55*
		cg10404427	3.16×10^{-5}	rs6859219	1.00	0.00	0.99	-	-	-	-	-	*ANKRD55*
		cg23343972	3.71×10^{-4}	rs6859219	1.00	0.00	0.99	-	-	-	-	-	-

Lead RA SNP	Locus	CpGs associated w/ LD block	CD4+ T cell FDR	CD4+ T cell SNP	r^2 with lead RA SNP	Bayes Coloc. (PP3)	Bayes Coloc. (PP4)	B cell FDR	B cell SNP	r^2 with lead RA SNP	Bayes Coloc. (PP3)	Bayes Coloc. (PP4)	UCSC Gene Symbol†
rs6859219 (continued)	5q11.2	cg15667493	0.009	rs6859219	1.00	0.00	0.99	-	-	-	-	-	*ANKRD55*
		cg00391767	-	-	-	-	-	0.001	rs6859219	1.00	0.00	0.99	*ANKRD55*
		cg15431103	0.0037151	rs6859219	1.00	0.01	0.95	-	-	-	-	-	*ANKRD55*
rs2278600	5q13.2	cg20794793	-	-	-	-	-	0.007	rs10515148	0.80	0.02	0.01	-
rs2561477	5q21.1	cg05225461	7.95 x 10^{-15}	rs28158	0.97	0.03	0.97	2.64 x 10^{-8}	rs2288789	0.95	0.03	0.97	*C5orf30*
		cg02855360	0.003	rs28157	0.95	0.04	0.95	-	-	-	-	-	*C5orf30*
rs2233424	6p21.1	cg05865665	0.003	rs190669824	0.90	0.10	0.29	9.71 x 10^{-4}	rs74950428	0.94	0.18	0.40	*TCTE1*
		cg16224301	-	-	-	-	-	0.009	rs190669824	0.90	0.07	0.42	*DLK2*
rs10499194	6q23.3	cg10612251	0.003	rs617328	0.89	0.05	0.91	-	-	-	-	-	-
rs2451258	6q25.3	cg16640008	0.002	rs2451278	0.98	0.00	0.99	-	-	-	-	-	-
rs1571878	6q27	cg01554751	4.23 x 10^{-8}	rs6907666	0.82	0.13	0.87	-	-	-	-	-	-
		cg19954286	-	-	-	-	-	1.10 x 10^{-4}	rs3093025	0.98	0.02	0.98	*CCR6*
		cg15222091	-	-	-	-	-	2.20 x 10^{-4}	rs3093025	0.98	0.02	0.98	*CCR6*
		cg21794222	-	-	-	-	-	5.35 x 10^{-4}	rs3093025	0.98	0.02	0.98	*CCR6*
		cg16523158	-	-	-	-	-	0.007	rs3093025	0.98	0.05	0.94	*CCR6*
		cg05094429	-	-	-	-	-	0.008	rs3093025	0.98	0.03	0.96	*CCR6*
rs67250450	7p15.1	cg11187739	6.08 x 10^{-19}	rs4722758	0.93	0.01	0.99	3.03 x 10^{-10}	rs4722758	0.93	0.01	0.99	*JAZF1*
		cg07522171	1.89 x 10^{-10}	rs2189966	0.95	0.03	0.97	-	-	-	-	-	*JAZF1-AS1; JAZF1*
		cg16130019	3.07 x 10^{-8}	rs917117	0.93	0.01	0.99	-	-	-	-	-	*JAZF1*
		cg00184826	5.99 x 10^{-8}	rs2893312	1.00	0.02	0.98	-	-	-	-	-	-
		cg26744081	5.35 x 10^{-4}	rs4722758	0.93	0.01	0.97	-	-	-	-	-	*JAZF1*
		cg11724147	7.89 x 10^{-4}	rs4722758	0.93	0.02	0.96	-	-	-	-	-	*HIBADH*
		cg11562379	0.001	rs2893312	1.00	0.02	0.98	-	-	-	-	-	*JAZF1*
		cg08519799	0.003	rs2189966	0.95	0.02	0.97	-	-	-	-	-	*JAZF1-AS1*
chr7:128580042	7q32.1	cg06630958	1.11 x 10^{-8}	rs3807307	1.00	0.99	0.01	-			-	-	*IRF5*
		cg12816198	-	-	-	-	-	3.56 x 10^{-7}	rs4728142	0.83	0.01	0.99	*IRF5*
		cg14349538	2.58 x 10^{-5}	rs3757387	0.92	0.98	0.02	-	-	-	-	-	*IRF5*
		cg24126180	4.01 x 10^{-4}	rs3807306	0.85	0.87	0.05	-	-	-	-	-	*IRF5*
		cg26616347	6.22 x 10^{-4}	rs3807306	0.85	0.92	0.02	-	-	-	-	-	*IRF5*

Lead RA SNP	**Locus**	**CpGs associated w/ LD block**	**CD4+ T cell FDR**	**CD4+ T cell SNP**	**r^2 with lead RA SNP**	**Bayes Coloc. (PP3)**	**Bayes Coloc. (PP4)**	**B cell FDR**	**B cell SNP**	**r^2 with lead RA SNP**	**Bayes Coloc. (PP3)**	**Bayes Coloc. (PP4)**	**UCSC Gene Symbol†**
rs2736337	8p23.1	cg16429190	-	-	-	-	-	3.20 x 10^{-18}	rs2061831	0.99	0.01	0.99	-
		cg21497594	-	-	-	-	-	1.86 x 10^{-9}	rs2061831	0.99	0.01	0.99	*BLK*
		cg01383082	-	-	-	-	-	4.08 x 10^{-9}	rs2618476	0.95	0.02	0.98	*FAM167A*
		cg09528494	5.44 x 10^{-6}	rs922483	0.80	0.01	0.99	1.16 x 10^{-8}	rs2061831	0.99	0.01	0.99	-
		cg11944933	-	-	-	-	-	2.36 x 10^{-6}	rs2618476	0.95	0.02	0.98	*FAM167A*
		cg07959027	4.97 x 10^{-6}	rs2061831	0.99	0.01	0.99	-	-	-	-	-	*BLK*
		cg04986849	-	-	-	-	-	2.77 x 10^{-5}	rs2618476	0.95	0.13	0.86	*BLK*
		cg01527115	-	-	-	-	-	1.60 x 10^{-4}	rs2618476	0.95	0.02	0.98	*FAM167A-AS1*
		cg23507676	-	-	-	-	-	5.92 x 10^{-4}	rs2061831	0.99	0.01	0.99	*BLK*
		cg03002059	-	-	-	-	-	0.003	rs2618473	0.90	0.05	0.93	*BLK*
		cg18591982	0.003	rs2061831	0.99	0.02	0.98	-	-	-	-	-	*BLK*
rs1516971	8q24.21	cg21864152	-	-	-	-	-	0.002	rs28455755	0.97	0.00	0.97	*LINC00824*
rs10985070	9q33.2	cg10737611	1.42 x 10^{-29}	rs2269060	0.97	0.05	0.95	1.55 x 10^{-17}	rs11794516	0.95	0.10	0.89	-
		cg07863409	7.99 x 10^{-8}	rs1930785	0.94	0.06	0.94	-	-	-	-	-	-
rs881375	9q33.2	cg04665046	-	-	-	-	-	2.11 x 10^{-8}	rs7037140	0.99	0.10	0.90	-
		cg21161526	-	-	-	-	-	6.17 x 10^{-8}	rs10760118	0.99	0.10	0.90	-
		cg14580859	1.27 x 10^{-4}	rs2159778	0.96	0.10	0.90	1.92 x 10^{-4}	rs1930778	1.00	0.15	0.84	-
		cg05834805	0.007	rs2241003	0.94	0.10	0.86	-	-	-	-	-	*PHF19*
rs3824660	10p14	cg00296182	-	-	-	-	-	0.003	rs3824660	1.00	0.01	0.98	-
		cg14416352	0.009	rs3802604	0.92	0.03	0.92	-	-	-	-	-	-
rs4918037	10q24.33	cg18735015	9.19 x 10^{-7}	rs7096731	0.96	0.09	0.05	3.04 x 10^{-5}	rs902995	0.92	0.09	0.03	*SH3PXD2A*
		cg00492979	4.74 x 10^{-6}	rs902995	0.92	0.09	0.05	6.77 x 10^{-5}	rs902995	0.92	0.08	0.04	*SH3PXD2A*
		cg20669641	3.10 x 10^{-5}	rs902995	0.92	0.08	0.04	-	-	-	-	-	*SH3PXD2A*
		cg19726408	4.58 x 10^{-5}	rs7096731	0.96	0.09	0.05	1.16 x 10^{-4}	rs902995	0.92	0.08	0.04	*SH3PXD2A*
		cg16669339	2.19 x 10^{-4}	rs902995	0.92	0.08	0.04	-	-	-	-	-	*SH3PXD2A*
rs508970	11q12.2	cg23043946	-	-	-	-	-	3.34 x 10^{-5}	rs508970	1.00	0.05	0.95	*VPS37C*
rs968567	11q12.2	cg21029357	3.31 x 10^{-44}	rs968567	1.00	0.00	1.00	1.51 x 10^{-14}	rs968567	1.00	0.00	1.00	*FADS2*
		cg13299762	5.02 x 10^{-40}	rs968567	1.00	0.00	1.00	3.76 x 10^{-26}	rs968567	1.00	0.00	1.00	*FADS2*
		cg27386326	6.29 x 10^{-32}	rs968567	1.00	0.00	1.00	7.64 x 10^{-16}	rs7943728	0.96	0.02	0.98	-
		cg19481605	2.91 x 10^{-31}	rs968567	1.00	0.00	1.00	1.49 x 10^{-27}	rs968567	1.00	0.00	1.00	*FADS2*

Lead RA SNP	Locus	CpGs associated w/ LD block	CD4+ T cell FDR	CD4+ T cell SNP	r^2 with lead RA SNP	Bayes Coloc. (PP3)	Bayes Coloc. (PP4)	B cell FDR	B cell SNP	r^2 with lead RA SNP	Bayes Coloc. (PP3)	Bayes Coloc. (PP4)	UCSC Gene Symbol†
rs968567 (continued)	11q12.2	cg06781209	2.20 x 10^{-28}	rs61897793	0.95	0.00	1.00	2.90 x 10^{-15}	rs968567	1.00	0.00	1.00	*FADS2*
		cg21409469	4.04 x 10^{-26}	rs968567	1.00	0.00	1.00	3.22 x 10^{-10}	rs7943728	0.96	0.01	0.99	*FADS2*
		cg21709803	1.09 x 10^{-25}	rs968567	1.00	0.00	1.00	2.93 x 10^{-10}	rs7943728	0.96	0.01	0.99	FADS2
		cg00603274	1.41 x 10^{-25}	rs968567	1.00	0.00	1.00	1.04 x 10^{-15}	rs7943728	0.96	0.01	0.99	FADS2
		cg07392590	1.20 x 10^{-24}	rs968567	1.00	0.00	1.00	3.35 x 10^{-14}	rs968567	1.00	0.00	1.00	-
		cg25324164	2.46 x 10^{-24}	rs968567	1.00	0.00	1.00	9.51 x 10^{-9}	rs968567	1.00	0.00	1.00	FADS2
		cg23017530	8.59 x 10^{-23}	rs61897793	0.95	0.00	1.00	5.98 x 10^{-11}	rs968567	1.00	0.00	1.00	FADS2
		cg01400685	1.74 x 10^{-14}	rs61897793	0.95	0.00	1.00	1.05 x 10^{-21}	rs968567	1.00	0.00	1.00	FADS2
		cg20896974	1.91 x 10^{-17}	rs968567	1.00	0.00	1.00	1.95 x 10^{-21}	rs968567	1.00	0.00	1.00	FADS2
		cg14562930	3.89 x 10^{-17}	rs61897793	0.95	0.00	1.00	1.04 x 10^{-7}	rs968567	1.00	0.00	1.00	FADS2
		cg08093537	3.94 x 10^{-17}	rs61897793	0.95	0.00	1.00	3.58 x 10^{-16}	rs968567	1.00	0.00	1.00	FADS2
		cg16213375	2.22 x 10^{-16}	rs61897793	0.95	0.00	1.00	-	-	-	-	-	FADS1
		cg20250926	1.02 x 10^{-12}	rs968567	1.00	0.00	1.00	1.97 x 10^{-5}	rs968567	1.00	0.00	1.00	FADS2
		cg07005513	-	-	-	-	-	1.09 x 10^{-12}	rs968567	1.00	0.00	1.00	FADS2
		cg01556593	1.01 x 10^{-11}	rs61897793	0.95	0.00	1.00	-	-	-	-	-	FADS2
		cg22295169	7.15 x 10^{-11}	rs968567	1.00	0.00	1.00	2.67 x 10^{-4}	rs7943728	0.96	0.01	0.98	FADS2
		cg02213369	-	-	-	-	-	2.13 x 10^{-10}	rs7943728	0.96	0.01	0.99	FADS2
		cg08281583	6.72 x 10^{-7}	rs968567	1.00	0.00	1.00	1.55 x 10^{-9}	rs7943728	0.96	0.01	0.99	FADS2
		cg14911132	3.64 x 10^{-9}	rs968567	1.00	0.00	1.00	1.50 x 10^{-6}	rs968567	1.00	0.00	1.00	FADS2
		cg15454066	-	-	-	-	-	2.49 x 10^{-6}	rs968567	1.00	0.00	1.00	FADS2
		cg19852225	2.98 x 10^{-6}	rs61896141	0.98	0.01	0.99	-	-	-	-	-	FADS2
		cg10069985	1.26 x 10^{-5}	rs968567	1.00	0.00	1.00	-	-	-	-	-	FADS2
		cg19610905	-	-	-	-	-	2.93 x 10^{-4}	rs968567	1.00	0.01	0.99	FADS2
		cg10515671	0.0007603	rs61896141	0.98	0.02	0.98	-	-	-	-	-	FADS1
		cg04010666	0.0098423	rs968567	1.00	0.02	0.92	-	-	-	-	-	TMEM216
rs3781913	11q13.4	cg18574813	8.04 x 10^{-8}	rs12802369	0.90	0.03	0.01	-	-	-	-	-	PDE2A
		cg22635155	-	-	-	-	-	0.002	rs342322	1.00	0.02	0.01	PDE2A
rs4409785	11q21	cg24692812	2.08 x 10^{-8}	rs4409785	1.00	0.00	0.99	-	-	-	-	-	-
		cg00617061	0.002	rs4409785	1.00	0.01	0.70	-	-	-	-	-	-
		cg10402062	-	-	-	-	-	0.006	rs4409785	1.00	0.01	0.77	-

Lead RA SNP	Locus	CpGs associated w/ LD block	CD4+ T cell FDR	CD4+ T cell SNP	r^2 with lead RA SNP	Bayes Coloc. (PP3)	Bayes Coloc. (PP4)	B cell FDR	B cell SNP	r^2 with lead RA SNP	Bayes Coloc. (PP3)	Bayes Coloc. (PP4)	UCSC Gene Symbol†
rs10790268	11q23.3	cg19308663	4.34 x 10^{-35}	rs4938573	0.92	0.99	0.01	2.05 x 10^{-10}	rs4938573	0.92	0.06	0.94	-
		cg00804785	6.78 x 10^{-12}	rs7951740	0.95	0.06	0.94	-	-	-	-	-	-
		cg01440696	0.004	rs73005423	0.79	0.11	0.87	-	-	-	-	-	-
rs773125	12q13.2	cg15120283	2.87 x 10^{-12}	rs11171739	0.81	0.00	1.00	2.30 x 10^{-4}	rs11171739	0.81	0.00	1.00	*RPS26*
		cg18911129	1.69 x 10^{-8}	rs11171739	0.81	0.00	1.00	2.01 x 10^{-4}	rs11171739	0.81	0.00	1.00	-
rs9603616	13q14.11	cg01699148	0.009	rs9532433	0.93	0.04	0.94	-	-	-	-	-	*COG6*
rs1950897	14q24.1	cg04384914	8.94 x 10^{-4}	rs8015139	0.91	0.12	0.87	-	-	-	-	-	*RAD51L1*
rs2841277	14q32.33	cg06920962	-	-	-	-	-	0.001	rs2582531	0.86	0.02	0.01	*PLD4*
rs8032939	15q14	cg17164630	2.18 x 10^{-12}	rs36027443	0.88	0.01	0.99	-	-	-	-	-	*RASGRP1*
		cg22603971	-	-	-	-	-	8.67 x 10^{-7}	rs28536554	0.89	0.01	0.99	*RASGRP1*
		cg25864633	-	-	-	-	-	0.003	rs6495979	0.93	0.01	0.99	*RASGRP1*
rs8026898	15q23	cg04787775	-	-	-	-	-	0.002	rs7170107	0.96	-	-	-
rs4780401	16p13.13	cg00288844	-	-	-	-	-	3.86 x 10^{-4}	rs1579258	0.89	0.03	0.97	*TXNDC11*
rs13330176	16q24.1	cg04454285	1.15 x 10^{-4}	rs12232384	0.86	0.01	0.99	5.24 x 10^{-6}	rs12232384	0.86	0.00	1.00	-
rs1877030	17q12	cg02780210	8.04 x 10^{-6}	rs8073511	0.87	0.91	0.01	-	-	-	-	-	-
chr17:38031857	17q12	cg12749226	-	-	-	-	-	2.88 x 10^{-17}	rs11557466	0.99	0.03	0.97	*ORMDL3*
		cg14348996	1.43 x 10^{-9}	rs11655198	0.86	0.52	0.48	1.04 x 10^{-10}	rs1008723	0.82	0.70	0.30	*GSDMB*
		cg18711369	1.36 x 10^{-7}	rs12946510	0.85	0.14	0.86	8.23 x 10^{-8}	rs9916765	0.81	0.41	0.59	*ORMDL3*
		cg02551532	7.92 x 10^{-7}	rs2952144	0.80	0.62	0.38	-	-	-	-	-	-
		cg13200575	-	-	-	-	-	2.71 x 10^{-6}	rs9903250	0.86	0.53	0.47	-
		cg10909506	4.12 x 10^{-6}	rs12946510	0.85	0.07	0.93	-	-	-	-	-	*ORMDL3*
		cg18691862	-	-	-	-	-	5.03 x 10^{-6}	rs9903250	0.86	0.56	0.44	*LRRC3C*
		cg12655416	-	-	-	-	-	7.20 x 10^{-5}	rs11557466	0.99	0.03	0.97	*ORMDL3*
		cg10057218	1.28 x 10^{-4}	rs11655198	0.86	0.35	0.64	-	-	-	-	-	*GSDMB*
rs11089637	22q11.21	cg19710672	0.001	rs5754387	0.81	0.08	0.91	-	-	-	-	-	*YDJC*
rs1043099	22q12.2	cg04852230	-	-	-	-	-	0.007	rs1043099	1.00	0.22	0.04	-
rs909685	22q13.1	cg15105517	-	-	-	-	-	6.91 x 10^{-43}	rs909685	1.00	0.00	1.00	*SYNGR1*
		cg24268161	1.62 x 10^{-29}	rs909685	1.00	0.00	1.00	1.89 x 10^{-14}	rs909685	1.00	0.00	1.00	*SYNGR1*
		cg22628235	-	-	-	-	-	1.59 x 10^{-6}	rs909685	1.00	0.00	1.00	*SYNGR1*
		cg07919145	-	-	-	-	-	0.004	rs909685	1.00	0.01	0.99	*SYNGR1*

Table 4.3: All rheumatoid arthritis (RA) risk loci harbouring meQTL variants within the linkage disequilibrium block (r^2 > 0.8 in1000 Genomes Project Phase 3 European Populations). The lead RA SNP is that which is reported by the genome-wide association study (GWAS). Linkage disequilibrium r^2 values (based on European populations) between the lead RA risk variant and the lead regulatory meQTL SNP identified in each cell type are reported. Bayes Coloc. PP3 = Bayesian co-localisation posterior probability of the H_3 hypothesis that each association (RA GWAS signal and meQTL) has an independent causal variant mapping to the region; Bayes Coloc. PP4 = Bayesian co-localisation posterior probability of the H_4 hypothesis that each association (RA GWAS signal and meQTL) has a single shared causal variant mapping to the region; UCSC Gene Symbol = UCSC Gene to which the cis-CpG maps as described in the Illumina MethylationEPIC manifest file.

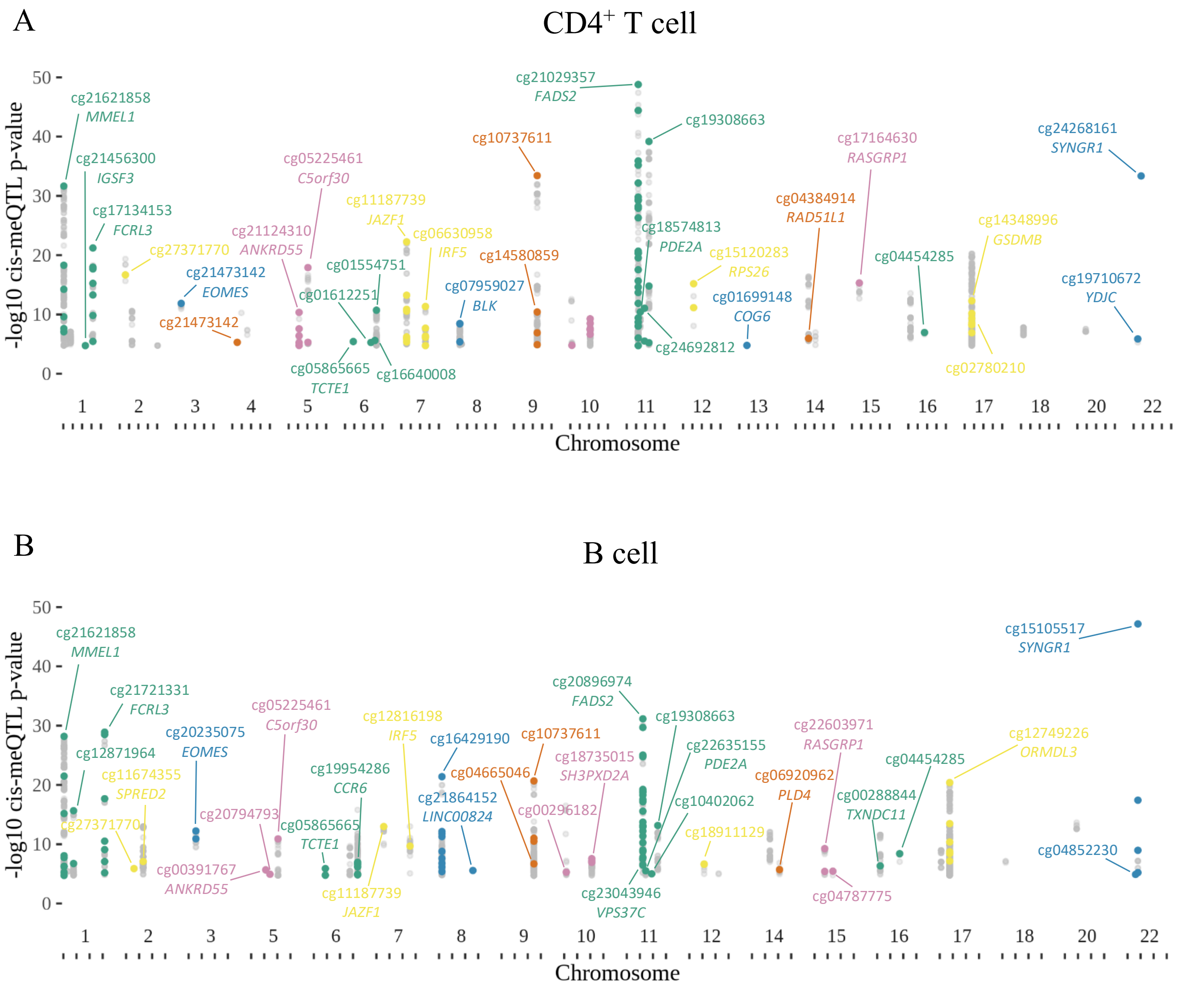


Figure 4.14: Manhattan plot of cis-meQTLs at rheumatoid arthritis risk loci. Associations between regulatory DNA variants at risk loci and CpGs in A) CD4⁺ T cells and B) B cells. Each point represents an association between a variant (chromosomal coordinates on the x-axis) and a cis-CpG probe, with the value on the y axes representing the –log₁₀ nominal association p-value. Points highlighted in colour are those whereby the lead meQTL variant (i.e. that remaining following SNP clumping) mapped to an RA risk locus linkage disequilibrium (LD) block ($r^2 \geq 0.8$ with lead risk variant). Grey points represent those where the SNP association pre-clumping mapped to an RA LD block, but the lead SNP following clumping did not. In instances whereby the CpG is annotated to a UCSC RefGene gene (within 20kb), then the gene name is given next to the CpG.

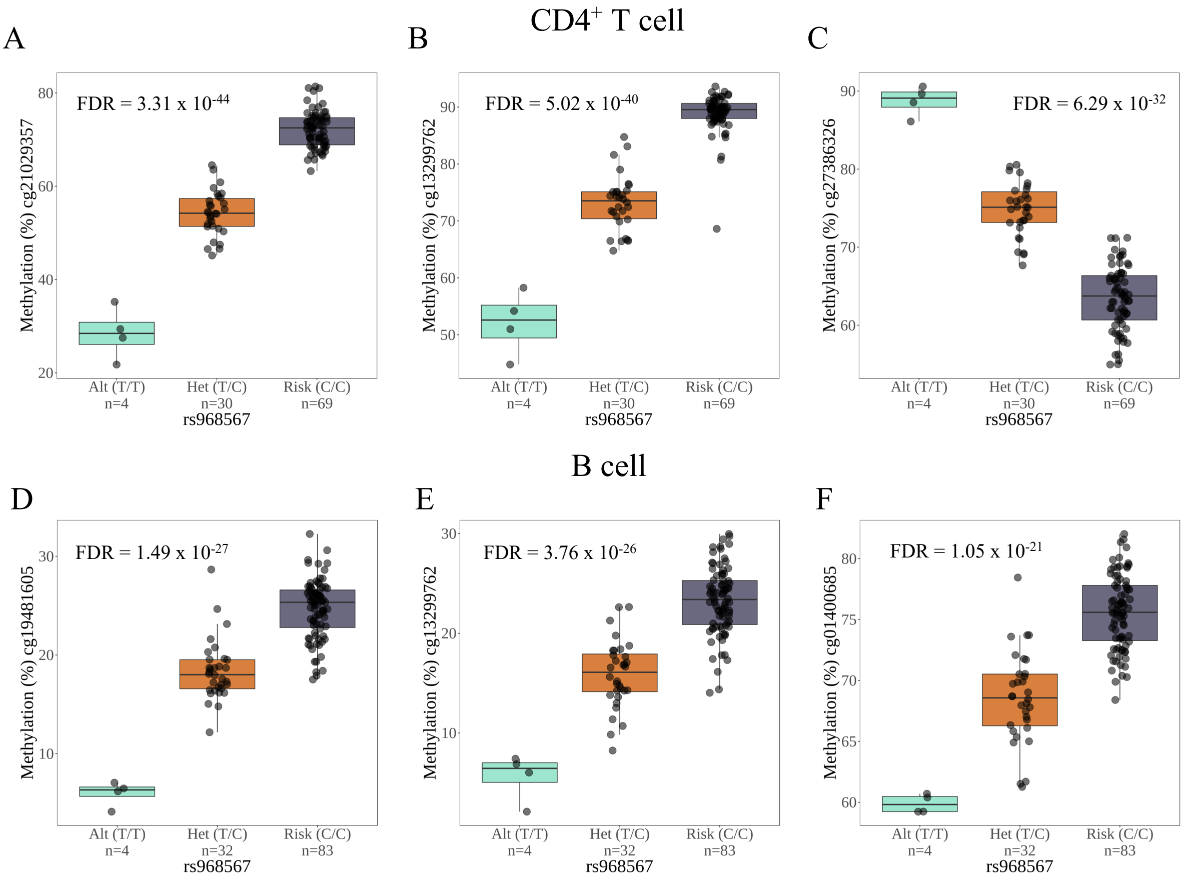


Figure 4.15: Exemplar cis-meQTLs at the rheumatoid arthritis risk locus on chromosome 11q12.2. Associations are shown between the risk allele at rs968567 and the top three cis-CpGs identified in CD4^{+} T cells: (A) cg21029357, (B) cg13299762, and (C) cg27386326, and in B cells: (D) cg19481605, (E) cg13299762, and (F) cg01400685.

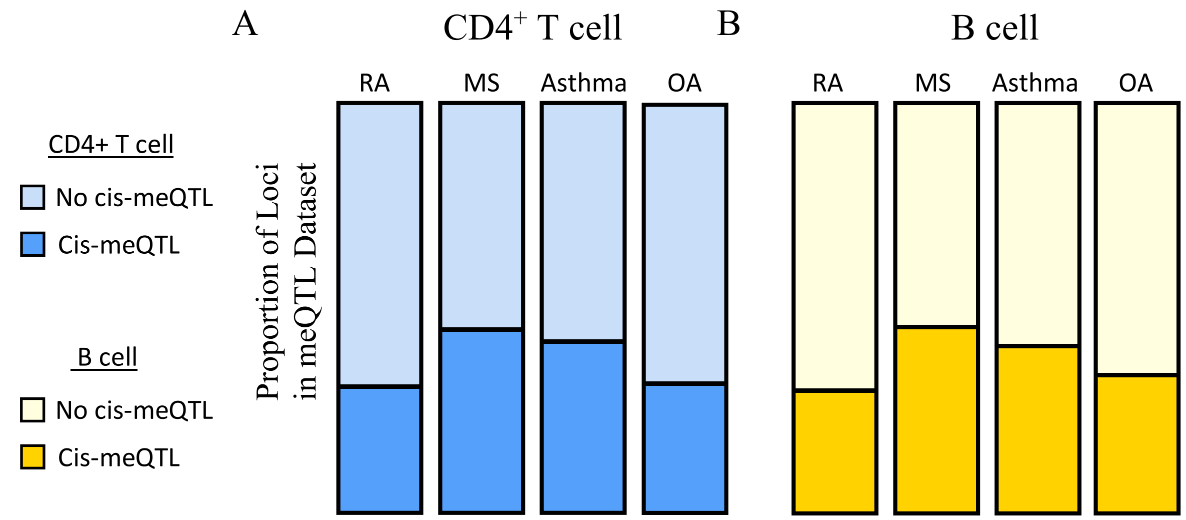
To provide further confidence that a single causal variant is responsible for both the meQTL and RA GWAS signals at a given locus, a Bayesian test for co-localization was performed[295]. This test provides the posterior probability of five individual hypotheses (see section 2.7.5), with H_4 being that of a single causal variant underlying both trait associations. All GWAS summary statistics used for Bayesian co-localisation were obtained from a comprehensive trans-ethnic RA meta-analysis[97]. It should be noted, however, that whilst the majority of RA risk loci intersected with the cis-meQTL results were defined based on this large study, a number were obtained from separate GWASs. Nonetheless, of the 99 independent CD4^{+} T cell SNP-CpG associations identified in cis for which the regulatory SNP mapped to an RA LD block, 75 (75.8%) exhibited strong evidence (posterior probability of H_4 (PP4) > 0.75) of a shared causal variant (Table 4.3). Of the 32 RA risk loci exhibiting cis-meQTL activity in these cells, 26 (81.3%) harboured at least one association with good evidence of co-localisation with the causal RA SNP. Similarly, in B cells 79/98 (80.6%) showed significant evidence of co-localisation, corresponding to 26/33 (78.8%) loci with at least one such association (Table 4.3).

The posterior probabilities (PP) for the hypotheses that the two traits (DNAm and RA susceptibility) have independent causal variants (H_3; PP3) or a single causal variant (H_4; PP4) based on this Bayesian co-localisation analysis are given in Table 4.3.

4.7 Cis-meQTLs associated with additional immune-mediated and joint-related disease

The primary focus of this project was to assess the DNAm landscape in the aetiopathogenesis of RA, having access to biological material from the clinical context of early, untreated arthritis that is most relevant to the understanding of this condition. However, to explore the extent to which putative genetic risk in other immune- or non-immune-mediated diseases functions to modulate DNAm in $CD4^+$ T cells and B cells, the analysis was extended beyond RA. Additional insight was therefore sought by mapping lymphocyte cis-meQTLs to risk loci for other complex genetic diseases. To this end, data for three additional diseases were obtained from the GWAS catalogue. These included multiple sclerosis (MS) – an autoimmune condition of the central nervous system in which $CD4^+$ T cells are believed to function as a pathogenic cell type[87]. Another inflammatory condition included was asthma, which affects the airways and results from a hyper-responsive immune response orchestrated by Type 2 T-helper (T_H2) cells – a $CD4^+$ T cell subset that are important in humoral immunity[294]. Finally, in addition to these two immune-mediated diseases, osteoarthritis (OA) was considered as a condition whereby symptoms manifest at the same tissue as in RA, albeit the aetiology is considered largely distinct, with a lesser or absent role for immune dysregulation.

Of the OA GWAS risk loci represented in the meQTL input data (82 for $CD4^+$ T cells, 83 for B cells), 26 (31.7%) in $CD4^+$ T cells and 28 (33.7%) in B cells were identified as cis-meQTLs, roughly mirroring the proportions seen in RA (Figure 4.16). Interestingly, the proportion of MS and Asthma GWAS loci found to exhibit cis-meQTL activity in $CD4^+$ T cells (MS = 44.7%; Asthma = 41.7%; Figure 4.16A) and B cells (MS = 45.4%; Asthma = 40.7%) was increased relative to RA and OA (Figure 4.16; Figure 4.16B).



Loci in input data:	104	123	139	82	107	130	145	83
Loci with cis-meQTL:	32	55	58	26	33	59	59	26
cis-CpGs at risk loci:	99	133	148	121	98	110	116	112

Figure 4.16: Cis-meQTLs at additional immune-mediated and joint-related genome-wide association loci. Proportion of GWAS loci in meQTL input data exhibiting cis-meQTL activity in A) CD4$^+$ T cells and B) B cells.

4.8 Functional annotation of disease-associated cis-meQTLs

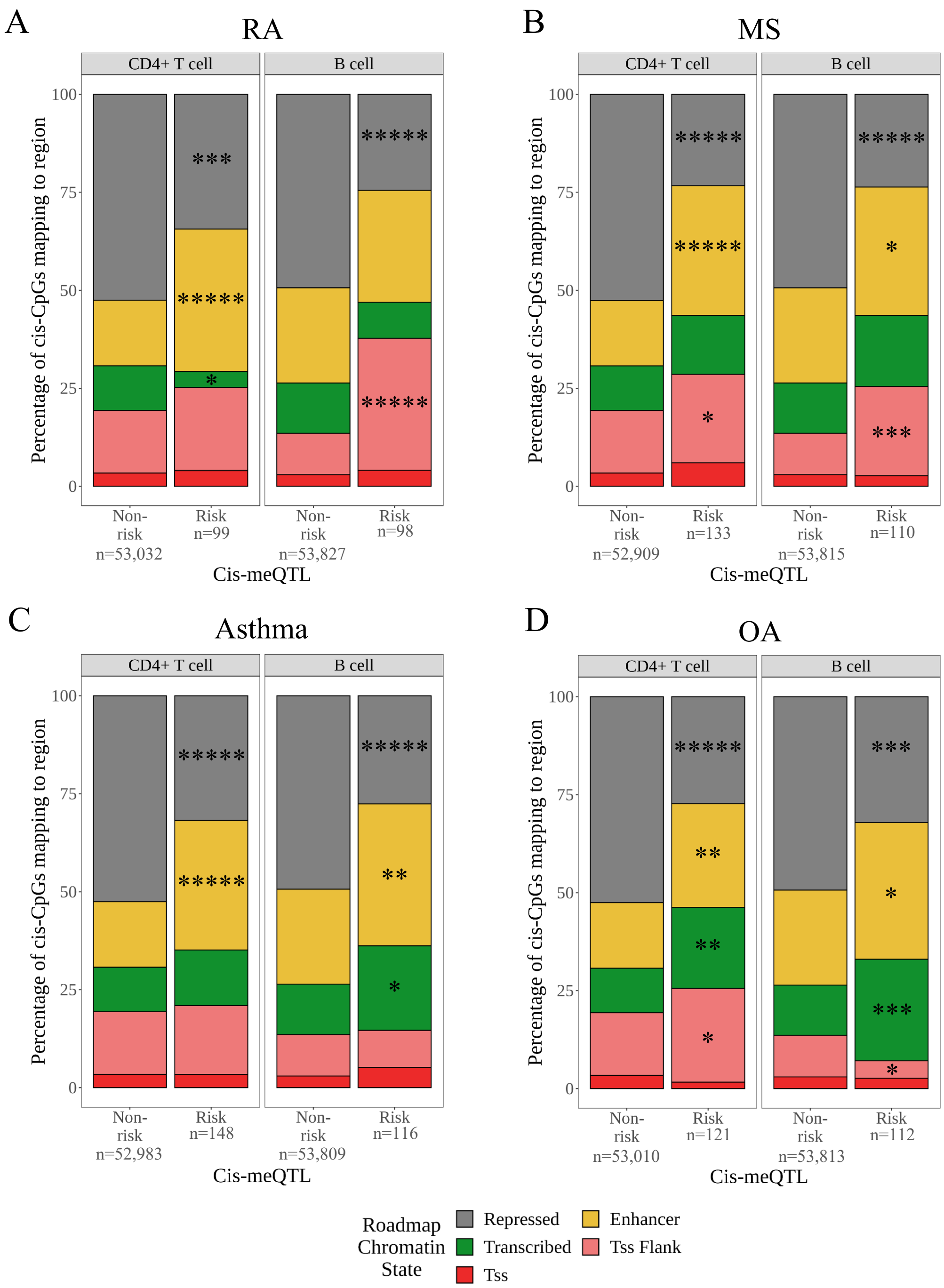
Identifying instances of GWAS loci functioning as cis-meQTLs can provide clues as to the potential mechanisms through which genetic risk modifies cell function. Beyond these associations, mapping disease associated DNAm modifications to specific chromatin states or functional elements such as TFBS can highlight pathways through which these variants impact transcriptional regulation.

4.8.1 Chromatin state enrichment

As was done for cis-meQTLs mapped genome-wide, cis-CpGs associated with the GWAS loci (described in sections 4.6-4.7) were mapped to cell type-specific chromatin state data from the Roadmap Epigenomics Project[173, 293]. Given that the previous analyses demonstrated cis-CpGs to map preferentially to regions such as enhancers, the enrichment of risk-associated cis-CpGs was compared with those regulated by non-risk loci for a given disease (Figure 4.17 & 4.18). For the data discussed in this section, a visual summary displaying the relative enrichments of risk-associated cis-CpGs at chromatin states in both cell types and across all diseases is shown in Figure 4.18.

Across all cell types and all diseases, risk-associated cis-CpGs were significantly depleted in repressed chromatin regions (Figure 4.17A-D). In CD4$^+$ T cells, cis-CpGs were significantly enriched in enhancer elements across all diseases, with this most pronounced at RA loci (2.18-fold enriched, p = 2.34 x 10^{-6}; Figure 4.17A & E), followed by MS (1.98-fold enriched, p = 3.77 x 10^{-6}; Figure 4.17B) and asthma (1.98-fold enriched, p = 1.23 x 10^{-6}; Figure 4.17C & E); although also present, OA cis-CpGs displayed a less marked enhancer enrichment (1.58-fold enriched, p = 0.0069; Figure 4.17D & E). Whilst no differences in the proportions of risk and non-risk cis-CpGs mapping to TSSs was seen for any diseases, an increased proportion of risk cis-CpGs mapping to the TSS flanking regions was observed, and this was particularly notable in B cells at RA (3.18-fold enriched, p = 7.45 x 10^{-10}; Figure 4.17 A & E) and MS (2.14-fold enriched, p = 2.40 x 10^{-4}; Figure 4.17B & E) loci. In both cell types, OA cis-CpGs displayed enrichment at transcribed regions (1.58-fold in CD4$^+$ T cells, p = 0.0069; 2.02-fold in B cells, p = 1.75 x 10^{-4}; Figure 4.17D & E), with the only other instance of this trend being observed at asthma loci in B cells (1.69-fold; p = 7.70 x 10^{-3}; Figure 4.17C & E).

These data collectively suggest that CpG sites at which DNAm levels are regulated in cis by GWAS risk loci are functionally enriched, with consistent depletion of risk-associated cis-CpGs at repressed chromatin across cell types and disease. Enrichment was particularly marked at CD4$^+$ T cell enhancer regions for immune-mediated disease (RA, MS, and asthma) risk loci, consistent with recent chromatin state data highlighting that genetic risk in these diseases highlight a strong T cell component [117].



E

		Rheumatoid arthritis					Multiple sclerosis					Asthma					Osteoarthritis				
		Rep	Enh	Tx	Tss Flnk	Tss	Rep	Enh	Tx	Tss Flnk	Tss	Rep	Enh	Tx	Tss Flnk	Tss	Rep	Enh	Tx	Tss Flnk	Tss
CD4+ T cell	Risk Cis-CpG (%)	34.3	36.4	4.0	21.2	4.0	23.3	33.1	15.0	22.6	6.0	31.8	33.1	14.2	17.6	3.4	27.3	26.4	20.1	24.0	1.7
	Non-risk cis-CpG (%)	52.5	16.7	11.4	16.0	3.4	52.6	16.7	11.4	16.0	3.4	52.5	16.7	11.4	16.0	3.4	52.5	16.7	11.4	16.0	3.4
B cell	Risk Cis-CpG (%)	24.5	28.6	9.2	33.7	4.1	23.6	32.7	18.2	22.7	2.7	27.6	36.2	21.6	10.6	5.2	32.1	34.8	25.9	4.5	2.7
	Non-risk cis-CpG (%)	49.4	24.3	12.8	10.6	3.0	49.4	24.3	12.8	10.6	3.0	49.4	24.3	12.8	10.6	3.0	49.3	24.3	12.8	10.6	3.0

Figure 4.17 (previous page): Proportions of risk-associated and non-risk associated cis-CpGs at cell-specific chromatin states. All cis-CpGs associated with risk loci for (A) rheumatoid arthritis (RA) (B) multiple sclerosis (MS) C) asthma, and D) osteoarthritis (OA), as well as those associated with non-risk loci, were mapped to cell-specific chromatin state data from the Roadmap Epigenomics project[173]. Cis-CpGs identified in CD4+ T cells were mapped to chromatin states of primary T helper cells in peripheral blood (E043) whereas those identified in B cells were mapped to states from primary B cells from peripheral blood (E032). Chromatin states were classified as three separate groups: repressed (grey), enhancer (yellow), transcribed (green), flanking a transcription start site (Tss Flank; pink), and transcription start site (Tss; red). (E) Percentages of risk-associated cis-CpGs and non-risk-associated cis-CpGs mapping to each of the five chromatin states in each cell type. Enrichment of risk-associated cis-CpGs relative to those associated with non-risk loci was performed using a Fisher's exact test. *****p < 0.00001; ****p<0.0001; ***p<0.001; **p<0.01; *p<0.05.

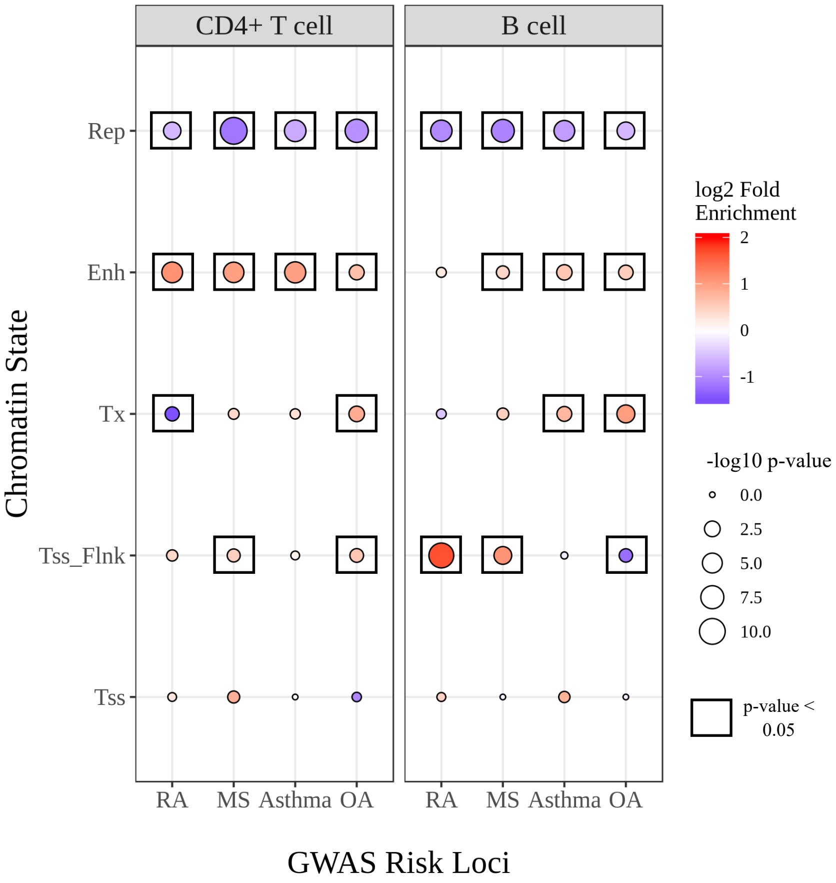


Figure 4.18: Enrichment of risk-associated cis-CpGs at cell specific chromatin states from the Roadmap Epigenomics Consortium[173]. CpGs associated with rheumatoid arthritis (RA), multiple sclerosis (MS), asthma, and osteoarthritis (OA) risk loci were overlapped with chromatin state data from primary T helper cells from peripheral blood (E043) and primary B cells from peripheral blood (E032). Cell states were collapsed into 5 functional groups (see section 2.7.5): Repressed (Rep), Enhancer (Enh), Transcribed (Tx), Flanking a transcription start site (Tss_Flnk), and transcription start site (Tss). Enrichment analysis (Fisher's exact test) was performed using non-risk cis-CpGs for each condition as background. Black boxes indicate enrichments that were significant at $p < 0.05$.

4.8.2 Transcription factor binding site enrichment

One mechanism through which DNAm is thought to exert effects on gene expression is through altering the binding of transcription factors and chromatin remodelling proteins to their cognate motifs (see Chapter 1.6.4). Therefore, following chromatin state mapping, cis-CpGs were overlapped with experimentally-determined transcription factor binding sites (TFBSs) from the Encyclopedia of DNA Elements (ENCODE) Project Consortium[209, 210]. Binding site data had been generated from ChIP-seq experiments and includes binding data for 161 transcriptional regulators. In this analysis, TFBS data across all cell types were used to define binding sites. It is possible that by modifying DNAm at CpG sites falling within TFBSs, risk variants may as such promote or inhibit transcription of proximal genes. Given that TFs often have well-defined roles in cellular responses or functions, identifying enrichments in this way may highlight disease-relevant pathways.

RA risk variants were found to confer altered DNAm at CpGs that were over-represented at RELA (p65 subunit of NFκB) binding sites in CD4$^+$ T cells ($p = 2.15 \times 10^{-4}$) and B cells ($p = 1.22 \times 10^{-4}$) (Figure 4.19 & 4.20). In both cell types, instances of this TFBS occurred ~2.7-fold more frequently at risk-associated cis-CpGs relative to those at non-risk loci (16% of RA risk cis-CpGs vs. 6% of non-risk cis-CpGs Figure 4.19). Indeed, binding sites of this TF were also enriched at cis-CpGs associated with MS loci in CD4$^+$ T cells (2.83-fold enriched, $p = 2.05 \times 10^{-8}$; Figure 4.19), as was also the case for RUNX3 (2.83-fold enriched, $p = 4.10 \times 10^{-5}$; Figure 4.19) and BATF (3.5-fold enriched, $p = 9.86 \times 10^{-5}$; Figure 4.19). The functions of nuclear factor-κB (NF-κB) in regulating immune responses are manifold, and amongst the well-characterised pathways in which this this transcription factor orchestrates transcriptional programs is downstream of T cell receptor stimulation[323]. Similarly, RUNX3 has diverse roles in coordinating immune responses, including the promoting expression of the type 1 CD4$^+$ T helper cell (T_H1) cytokine IFN-γ and repression of the type 2 cytokine IL-4[324]. This TF is an important factor in chromatin modelling that occurs as naïve CD8+ T cells develop into cytotoxic T lymphocytes following stimulation of the T cell receptor[325]. Interestingly, variability in DNAm at RUNX3 binding sites has been associated with susceptibility to RA in a twin study[230].

At MS loci in B cells, cis-CpGs implicated binding sites of TBL1XR1 ($p = 5.84 \times 10^{-5}$), which is required for NF-κB activation (also referred to as TBLR1)[326], and CCNT2 ($p = 1.12 \times 10^{-4}$), encoding the cyclin T2 protein.

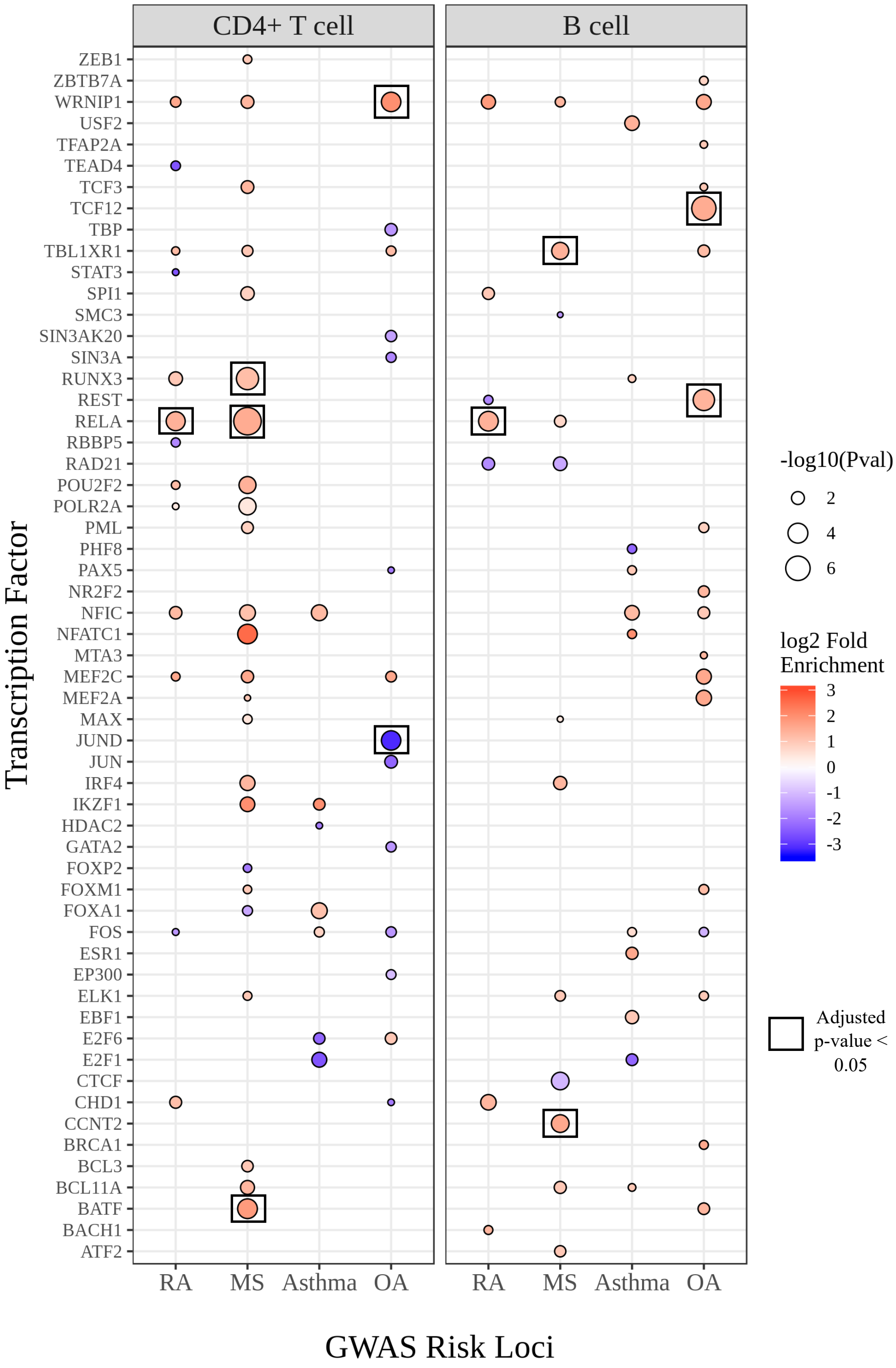


Figure 4.19: Enrichment of risk-associated cis-CpGs at transcription factor binding sites. Cis-meQTL associated CpGs (cis-CpGs) in CD4$^+$ T cells and B cells were overlapped with TFBS data based on data from the ENCODE project[209,210]. Enrichment analyses (Fisher's exact test) were performed to determine whether cis-CpGs associated with risk loci for rheumatoid arthritis (RA), multiple sclerosis (MS), asthma, and osteoarthritis (OA) were over-represented at binding sites for particular transcription factors. Enrichments are shown for all transcription factors for which nominally significant enrichment ($p < 0.05$) was observed in at least one cell type/disease locus. Square boxes surrounding a point indicate that the enrichment was significant following Bonferroni correction (adjusted p-value < 0.05).

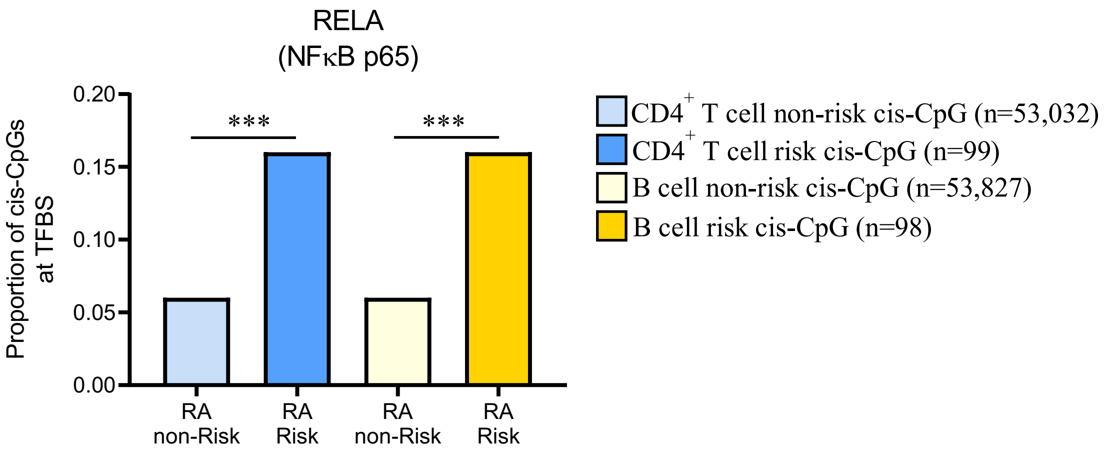


Figure 4.20: Relative proportions of RELA binding sites at rheumatoid arthritis cis-CpGs. The proportions of cis-CpGs overlapping binding sites of RELA – the p65 subunit of NFκB - at sites associated with RA risk loci and non-risk loci in CD4$^+$ T cells and B cells. ***p < 0.001

No TFBS enrichments were robust to stringent Bonferroni correction at Asthma loci in either cell type. Profiles of TFBSs at OA cis-CpGs were largely distinct from those at RA and MS loci, with an over-representation of WRNIP1 in CD4$^+$ T cells (p = 1.47 x 10^{-4}), as well as TCF12 (1.21 x 10^{-6}) and REST (2.62 x 10^{-5}) in B cells (Figure 4.20). WRNIP1 is the Werner helicase interacting protein 1, so named given its interaction with the helicase enzyme which is mutated in Werner syndrome – a rare disease characterised by an accelerated ageing process[327]. This factor is integral in processes such as maintaining stable replication forks during DNA synthesis and cell cycle checkpoint activation in the presence of DNA damage[327]. Conversely, a role for TCF12 has been described in mediating the development of bone-synthesizing osteoblasts from bone marrow-derived stem cells[328]. The only significant depletion of TFBS amongst cis-CpGs was observed for JUND at OA risk loci (9-fold depletion at risk-associated cis-CpGs, p = 1.51 x 10^{-4}) in CD4$^+$ T cells. The proportion of all risk-associated and non-risk-associated cis-CpGs mapping to binding sites of each transcriptional regulator with nominal enrichment/depletion (p < 0.05) are given across the four sets of disease loci in Appendix E.

4.8.3 Gene Ontology Pathway Analysis

To define pathways that were putatively dysregulated downstream of disease-associated DNAm changes, a Gene Ontology (GO) pathway analysis was performed with the *gometh* function in the missMethyl package[289], again using non-risk cis-CpGs as the background to test for enrichments. This method involves mapping CpGs to genes and then applying a modified hypergeometric test, taking into account the number of probes on the MethylationEPIC array mapping to each gene (see Chapter 2.6.4 for details).

Amongst the most significantly enriched processes at RA cis-CpGs in CD4$^+$ T cells were those relating to 'α-linoleic acid metabolic process' (p = 1.23 x 10^{-4}), 'regulation of B cell receptor signalling pathway ($p = 4.29 \times 10^{-4}$), and ‘positive regulation of protein serine/threonine phosphatase activity’ ($p = 0.0014$; Figure 4.21A). As regards MS loci, ‘negative regulation of calcidiol 1-monooxygenase activity’ ($p = 2.34 \times 10^{-5}$), ‘positive regulation of T cell activation’ (5.59×10^{-5}), and ‘positive regulation of leukocyte cell-cell adhesion’ (6.62×10^{-5}) were highlighted as relevant pathways (Figure 4.21B). Cellular metabolic processes were implicated at Asthma loci, with ‘negative regulation of nitrogen compound metabolic process’ ($p = 6.33 \times 10^{-7}$), along with processes relating to T cell development/activation including ‘negative regulation of CD4^{+} αβ T cell differentiation’ (3.63×10^{-6}) and ‘negative regulation of CD4^{+} αβ T cell activation’ ($p = 1.07 \times 10^{-5}$; Figure 4.21C). As would be expected, DNAm changes associated with genetic risk in OA impact distinct molecular pathways to the immune-mediated diseases, with ‘regulation of endopeptidase activity’ ($p = 2.61 \times 10^{-6}$) and ‘activation of cysteine-type endopeptidase activity involved in apoptotic process’ ($p = 2.91 \times 10^{-6}$) prominent amongst enriched processes (Figure 4.21D).

These observations were largely re-capitulated in B cells. For example, analysis of RA-associated cis-CpGs strongly implicates a range of processes related to lymphocyte-mediated immunity, including ‘regulation of B cell receptor signalling pathway’ ($p = 4.61 \times 10^{-4}$), ‘immune effector process’ ($p = 6.73 \times 10^{-4}$), and ‘double-negative stage 3 (DN3) thymocyte differentiation’ ($p = 0.0013$; Figure 4.22 A). As was the case for CD4^{+} T cells, MS-associated cis-CpGs in B cells suggest a role for metabolic processes, with ‘cellular response to organic substance’ ($p = 2.58 \times 10^{-5}$) and ‘regulation of macromolecule metabolic process’ ($p = 2.98 \times 10^{-5}$) featuring as the most enriched pathways (Figure 4.22B). Findings at asthma loci again highlight the pathogenic contribution of adaptive immunity, with enhanced mapping of cis-CpGs to genes involved in ‘cytokine production’ ($p = 6.34 \times 10^{-7}$) and ‘regulation of type 2 immune response’ ($p = 1.63 \times 10^{-5}$; Figure 4.22C). Additionally, analysis of OA loci highlights distinct aetiological processes, with an absence of immune-related pathways, instead implicating ‘positive regulation of apoptotic process’ ($p = 1.19 \times 10^{-5}$), ‘positive regulation of programmed cell death’ ($p = 1.21 \times 10^{-5}$), and ‘negative regulation of TGF-β receptor signalling pathway’ ($p = 1.55 \times 10^{-4}$; Figure 4.22D).

These ontology analyses of DNAm modifications associated with GWAS loci illustrate that for all IMDs, genetic risk variants are associated with DNAm levels at CpGs which could putatively regulate the expression of genes having key roles in immune effector processes. Contrastingly, DNAm modifications associated with OA risk loci map to genes that function in distinct pathways. The top 50 biological processes found to be enriched ($p < 0.01$) at GWAS risk cis-CpGs in each cell type is presented in the Appendix F.

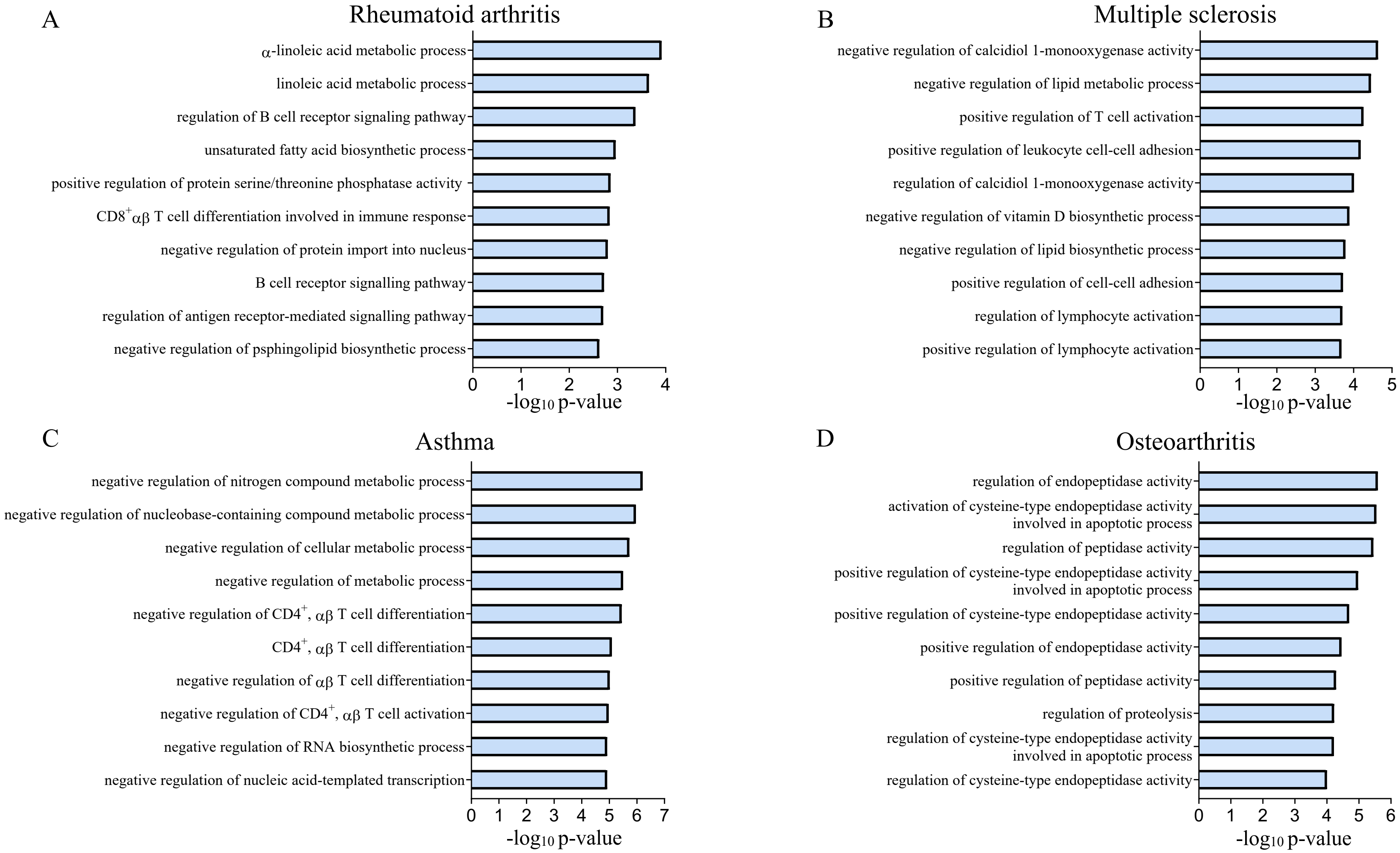


Figure 4.21: Top 10 Enriched 'Biological Process' pathways at risk-associated cis-CpGs in CD4+ T cells. Gene ontology biological process pathway analysis was performed at cis-CpGs associated with A) Rheumatoid arthritis, B) Multiple sclerosis, C) Asthma, and D) Osteoarthritis risk loci using non-risk cis-CpGs as background.

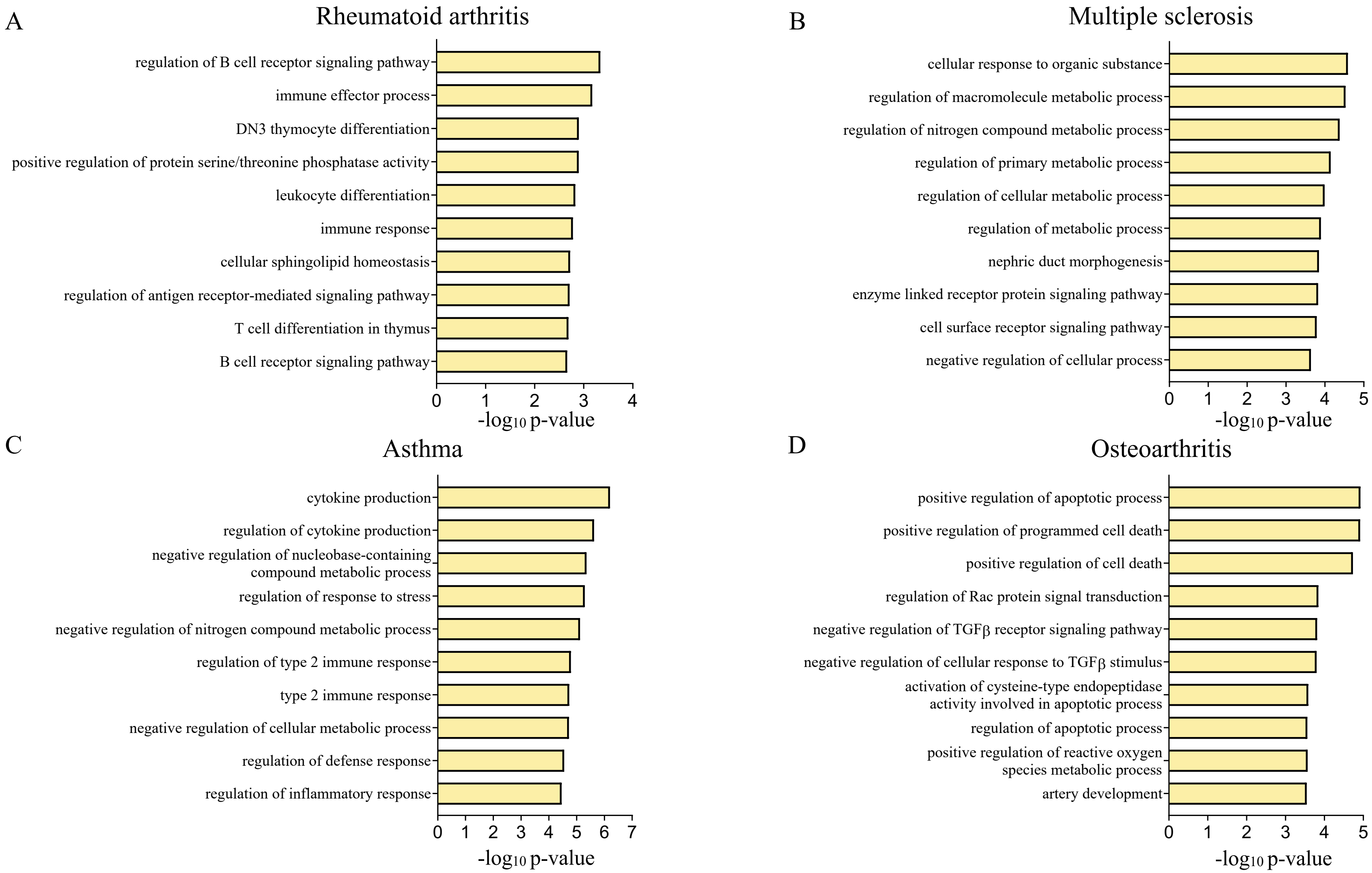


Figure 4.22: Top 10 Enriched 'Biological Process' pathways at risk-associated cis-CpGs in B cells. Gene ontology biological process pathway analysis was performed at cis-CpGs associated with A) Rheumatoid arthritis, B) Multiple sclerosis, C) Asthma, and D) Osteoarthritis risk loci, using non-risk cis-CpGs as background.

4.9 Discussion

This chapter presents a genome-wide meQTL analysis in CD4$^+$ T cells and B cells, the results of which are considered in the context of genetic architecture in complex diseases. The results described here both highlight genetic risk loci across IMDs at which DNAm is likely to have a functional role in conferring pathogenic cellular immune responses, as well as going some way to explaining how sequence polymorphisms may underlie disease-associated epigenetic signatures. As such, the data will provide an important resource for the functional interpretation of not only genome-wide association studies, but also epigenome-wide association studies.

4.9.1 Genetic polymorphisms are a source of variation in lymphocyte DNA methylation

Initial observations were that ~7.5% of CpGs analysed in each cell type were associated with a regulatory variant acting in cis. Even accounting for the relatively small sample size used to map meQTLs in this analysis, these results reveal that variants in the DNA sequence are a major determinant of inter-individual variability in lymphocyte DNAm. Recent studies in whole blood, using larger sample sizes, revealed the proportion of CpGs subject to cis-regulation to be ~12-18%[245, 247]. Estimates of DNAm heritability (H^2) in CD4$^+$ T cells are approximately 0.13 (i.e. 13% of the phenotypic variability in DNAm levels is inherited), with 74% of highly heritable ($H^2 > 0.4$) CpGs being associated with a SNP in cis[250]. Collectively, that a considerable proportion of CpGs present on the Illumina Methylation BeadChip arrays (both 450K and EPIC) are associated with DNA variants in tissues such as whole blood and immune cells reinforces the need to consider such effects when assigning mechanisms to disease associated DNAm changes (see Chapter 3).

It was also discovered that 38.5% of all CpGs associated with a cis-meQTL across this analysis, and over half of those identified in each separate cell type, were common to both lymphocyte subsets. Given the differing number of CD4$^+$ T- and B cell samples available, as well as using differing lists of input SNPs for QTL mapping, this number likely represents an underestimate. Cross-tissue quantification of meQTL effects in three cell types (primary T cells, primary fibroblasts, and lymphoblastoid cell lines (LCLs)) from the same individuals has revealed that 46-80% were shared between at least two of these cell types[126]. The extent of shared effects between any two cells seems to reflect their developmental proximity, with highest degree of sharing (80%) found between T cells and LCLs, both of which develop through a lymphocyte lineage[126]. In agreement with this, the BLUEPRINT epigenome project assessed cell-type specificity of meQTLs in CD4$^+$ T cells, monocytes, and neutrophils, and found sharing between the latter two cells (both myeloid lineage) was greater than of either with T cells (lymphocyte lineage)[127]. Whilst paired (same patient) cell type-specific data were available for the majority of samples in the present analysis (78.6 % of CD4$^+$ T cell samples, 68.1% for B cells), no direct comparison is made between cells of the same individuals, meaning that accurate estimates of cross-tissue effects were not established. Nonetheless, the data reveal that a considerable degree of shared genetic effects on CpG methylation exist between CD4$^+$ T cells and B cells. Despite the extent of shared effects, these results also confirm the critical need for any meQTLs identified in complex tissues to be validated in more uniform populations of cells. It is likely that in studies of whole blood, as has been common for EWASs and meQTL mapping, some cell type-specific effects are obscured by the cell type heterogeneity of this tissue.

As well as QTLs that function in one cell type but not the other, there are those that function in both but exhibit opposing allelic effects. Amongst the small number identified in the current study were those mapping to the *CRACR2A* promoter. *CRACR2A* encodes Calcium Release Activated Channel Regulator 2A, a Rab guanosine triphosphatase (Rab GTPase) expressed in lymphocytes, predominantly CD4$^+$ T cells of the Th1 and Th17 subtypes[329]. The CRACR2A protein is important in intracellular signaling downstream of the T cell receptor, including calcium signaling and activation of the Mitogen-Activated Protein Kinase (MAP kinase) pathway during T cell-mediated immune responses[329]. Whether or not these differing allelic effects between CD4$^+$ T cells and B cells represent functional regulatory differences at this locus will be an interesting question. Such opposing allelic effects have been described at the level of gene expression, with indications that epigenetic mechanisms could be responsible for differing directionality seen between tissues[136, 330].

The finding that cis-CpGs in both cell types were enriched in intergenic regions and depleted in CGIs replicates observations in whole blood[245]. CGIs that occur at many mammalian gene promoters are most often de-methylated in a state associated with active transcription[331]. That promoter-associated CGIs occur at housekeeping genes, and those with important functions in development, likely explains why DNAm patterns are more conserved at these features[331]. As such, cis-meQTL effects on CpGs outside of CGIs, many of which occur in enhancer elements and have a more intermediate level of DNAm across cells, could indicate a role in fine-tuning transcriptional regulation, as opposed to more binary on/off patterns.

4.9.2 Integration of cell-specific meQTL profiles with data from genome-wide association studies can help assign function to non-coding risk variants

The rationale behind assessing co-localisation of cis-meQTL variants with RA GWAS risk loci is that robust disease-associated DNAm changes downstream of known genetic susceptibility can facilitate the identification of regulatory mechanisms at these non-coding variants. Here, intersecting the cis-regulatory landscape in early arthritis lymphocytes with known genetic risk variants from published GWASs revealed the capacity of many such loci to modulate DNAm at neighbouring cis-CpGs. Such an approach has also proved effective in similar studies of relevant tissues for diseases with complex genetic aetiologies. For example, analysis of cartilage samples has revealed cis-meQTLs at 4/16 OA risk loci[332].

An alternative, albeit related approach, is to first identify DMPs/DMRs, and subsequently map these to genetic effects. Though this was not applicable in the case of this project, given the lack of significant differential methylation, it has proved successful in previous studies to define loci at which epigenetic mechanisms interact with genetic risk factors. The first study to directly investigate the potential role of DNAm as a mediator of genetic risk in RA initially performed an EWAS in whole blood to identify DMPs[218]. Subsequently, incorporating case/control genotypes revealed a number of positions, predominantly at the MHC locus, for which DNAm represented a regulatory intermediate of genetic risk in RA[218]. Likewise, of the 10 validated RA B cell differentially-methylated CpGs identified by Julia et al., three were associated with cis-meQTLs at RA risk loci[222]. In a similar study of lymphocytes in primary SS, differentially methylated positions between patients and controls overlapped five GWAS risk loci (though no formal association was tested)[333]. These findings, together with those reported in this chapter, clearly illustrate that identification of meQTLs at a cellular level is necessary for disentangling genetic and epigenetic risk factors, as well as for the functional annotation of non-coding GWAS variants.

A number of the cis-CpGs identified at RA risk loci mapped to genes that we had previously identified as cis-eQTLs in a cohort comprising many of the same patients[139]. These genes included, amongst others, *FCRL3*, *JAZF1*, *FADS2*, and *ORMDL3*, *BLK*, and *SYNGR1*. This highlights a potentially common underlying regulatory mechanism for both molecular traits at these loci. The cis-meQTL mapping to the *FADS2* locus was notable for the large number of associated cis-CpGs, the majority of which exhibit strong allelic effects. MeQTLs at this locus have been reported previously, with DNAm at five CpG sites in the FADS2 promoter downstream of the risk variant (rs968567) influencing co-expression of both *FADS1* and *FADS2*[334]. The authors propose that altered DNAm at these CpG sites disrupts binding of the SREBF2 transcription factor and inhibits expression of the two genes[334]. MeQTL and eQTL mapping thus represent complementary approaches, and highlighting loci that function to modulate both DNAm and gene expression enables regulatory mechanisms to be inferred, as will be discussed in Chapter 5.

4.9.3 Functional enrichment of disease-associated cis-meQTLs suggests a role in transcriptional regulation

The disproportionate occurrence, particularly in CD4$^+$ T cells, of disease associated DNAm modifications at enhancer elements suggests that they may impact gene expression. Indeed, CpGs that influence expression of genes in cis have previously been shown to be enriched within enhancers, as well as CGI shores and gene bodies[126]. Consistent with this, cis-CpGs that are associated with GWAS loci for all diseases (RA, MS, Asthma, OA) were reduced in regions of the genome representing repressed chromatin. CD4$^+$ T cell enhancer enrichment of cis-CpGs was particularly marked at the three IMD loci as compared with OA loci, potentially reflecting the greater contribution of this cell type to pathogenesis of these conditions.

Further evidence for these risk meQTLs disrupting functional DNA elements was the observation that many of the cis-CpGs map to the binding sites of transcriptional regulators. Most notably, at RA loci in both cell types and MS loci in B cells, cis-CpGs were strongly enriched at binding sites of the NFκB p65 subunit (RELA). NFκB activation can be triggered downstream of cytokine signalling, with this transcription factor having key functions in the development and function of immune cells, and is a mediator of multiple RA processes from osteoclast development to autoantibody production[335].

In addition, pathway analyses highlighted genes functioning in intuitive biological processes to be overrepresented amongst those mapping to risk-associated cis-CpGs for each disease. For example, CpGs associated with risk loci for RA, MS, and Asthma were notable in that they were enriched amongst immune-related pathways. These findings are perhaps not surprising, given that GWAS loci for IMDs such as RA themselves map to many genes with crucial roles in T cell and immune function (*CD28*, *CTLA4*, *IL2RA*)[97], and are enriched in T cell specific active regulatory regions[45].

In contrast, those at OA loci instead implicated pathways relating to programmed cell death and TGF-β signalling. This would further suggest that DNAm modifications downstream of risk variants perturb pathways that directly contribute to cell-mediated pathogenesis. In IMDs this involves dysregulation of immune responses, such as the activation of lymphocytes and cytokine production. That lymphocyte cis-meQTLs were, however, prominent at OA risk loci despite these cells having a less pronounced pathophysiological function may reflect that genetic effects on DNAm exhibit less cell type specificity than do effects on transcription[127].

The enrichment of RA- and other IMD-associated cis-CpGs at functional regulatory elements suggests that modified DNAm at these positions may confer upregulation or downregulation of genes responsible for the genetic disease association. Given that additional regulatory mechanisms exist to control cell-specific transcriptional programs, DNAm alterations should be considered in the context of gene expression changes (see Chapter 5).

4.9.4 Scarcity of disease-specific meQTLs

Though meQTL mapping was performed across all patients to increase the power to detect associations, an interaction analysis was also performed to identify putative RA-specific meQTL effects. The rationale behind this study design was to identify context-specific meQTLs, the activity of which is potentially responsive to RA-specific factors during early disease. DNAm can be altered in response to external cellular stimuli as has been shown in a range of experimental models. For example, *in vitro* monocyte stimulation with cytokines that are typically elevated in the circulation of individuals with RA (TNF-α, IFN-α, IFN-γ) can induce TF-dependent hyper- and hypo-methylation at hundreds of CpGs, partially recapitulating patterns seen in patient cells[336]. It is therefore possible that environmentally responsive TFs, displaying preferential binding to specific alleles, may mediate context-dependent meQTLs.

As was suggested for the lack of differential methylation in Chapter 3, it is possible the small number of disease-specific meQTLs revealed in the interaction analysis result from the comparison of RA patients to a disease cohort matched for systemic inflammatory markers. In this scenario it is conceivable that circulating factors such as cytokines, to which the cell may be exposed, are elevated in the control group, such as would be the case in patients diagnosed with psoriatic arthritis for example[263]. Additionally, meQTL associations may be modified by factors that are elevated during systemic inflammatory responses[213]. Though differences in CRP levels between comparator groups had been controlled for in this study, a previous EWAS in IBD found the top differentially methylated position, mapping to the *RPS6KA2* gene, exhibited a strong association with CRP levels[214]

At the level of gene expression, context-dependent eQTLs have been described that are active following stimulation of cells, either by direct stimulation in vitro or using proxy measures[133, 337]. For example, for the *FADS2* locus described earlier (rs968567), levels of the SREBF2 transcription factor that promotes expression of this gene exhibit negative associations with high density lipoprotein cholesterol[133]. This represents an example of an eQTL effect at a RA risk locus that could be modified by an environmental factor. In addition, QTLs may be uniquely active in patients with active disease. One such example is the CD4$^+$ T cell cis-eQTL regulating expression of *CD5* and *CD6* that is specific to patients with inflammatory bowel disease, with the association absent in healthy controls[132]. Cis-eQTLs that respond to CD4$^+$ T cell receptor activation *in vitro* have been suggested to drive inter-individual variability in immune responses[338]. It appears that many eQTLs that become active upon immune stimulation occur at enhancers that are 'primed', whereby the regulatory variant modifies chromatin accessibility and as such facilitates binding of transcription factors upon stimulation[339]. Chromatin accessibility in immune cells is drastically modified upon stimulation, with this effect most marked in T cells and B cells[340]. Interestingly, SNPs conferring RA heritability were most strongly enriched at chromatin regions that were only accessible upon stimulation of T cells with less pronounced, albeit significant, enrichment at accessible regions specific to stimulated B cells[340].

Despite these limited observations in patients, context-specific QTLs described to date have largely been identified in isolated cells using controlled *in vitro* immune stimulation[337]. Identifying disease-specific effects in human cohorts will represent much more of a challenge. A recent study identified cis-meQTLs specific to patients with SLE but not healthy controls, albeit using a less robust approach to the interaction analysis described in this chapter[248]. An analysis of regions of the genome displaying variable DNAm in cord blood from new-borns found those exhibiting genetic (cis) × environmental factor interactions to be enriched at GWAS loci for complex traits (though immune-mediated diseases were not the focus of this analysis)[341].Ultimately, the relative contributions of genetic variation and environmental exposures to DNAm levels in pathogenic cell types will require careful study designs and a combination of *in vitro* and *in vivo* studies (see Chapter 6).

4.9.5 Limitations

One major consideration when interpreting the results presented in this chapter is the size of sample cohorts used for mapping meQTLs, which limits the power to detect such associations. This may in part account for the difference in observed numbers of cis-CpGs and those reported in previous studies of whole blood[245, 247], despite the present study having the advantage of analysing isolated cell types as opposed to homogenous tissues. One other analysis parameter that frequently differs between analyses is the cis-distance defined for mapping cis-meQTLs. In the analysis described in this chapter, a window size of 1Mb either side of a SNP was used to test for associations. However, given that 78% and 76% of cis-associations occurred over a SNP-CpG distance of $\leq$ 100Kb in CD4$^+$ T cells and B cells respectively, decreasing this window size may increase power to detect associations with smaller effect sizes, at the expense of missing long-distance effects.

As meQTL mapping nominates a number of candidate causal SNPs for a given association, the SNP clumping method was used here to remove tagging SNPs and select the top hit to take forward for downstream analyses. However, this approach selects the top SNP based solely on the p-value association, which may not necessarily represent the causal variant. If the true causal SNP is removed at this stage, this could potentially reduce the number of disease co-localised effects. On the other hand, because GWAS risk loci were taken from multiple studies for each trait, the method of using LD blocks to define risk loci does have scope to introduce false positive risk meQTLs. This is because, if the two studies report distinct lead SNPs for the same locus, defining LD blocks ($r^2 \geq 0.8$) based on these variants will return different regions for each locus, with a small chance that the same meQTL co-localisation may be reported twice.

Despite the interesting findings at a number of risk loci, cis-meQTL effects were only identified at ~30% of RA risk loci, with strong statistical evidence of shared effects at ~25%. These findings would, however, be fairly consistent with a study in $CD4^+$ T cells, monocytes, and LCLs, which found that evidence for shared effects between GWAS variants and cis-eQTLs was somewhat limited[342]. It may be that such effects are not picked up because of the limited sample size mentioned above, or because of low minor allele frequencies (the SNP inclusion criteria required ≥ 3 samples per genotype). Indeed, with regard to the latter, regions at which multiple low frequency variants potentially function to modify DNAm levels in cis have been described[343]. As a result, whilst robust associations at a number of loci have been illustrated in this chapter, it cannot be claimed that these results represent a complete map of cis-regulatory DNAm landscape in these cells given the limitations.

Another potential reason for missed effects could be the drawback arising from the MethylationEPIC BeadChip to interrogate DNAm, with identification of meQTL effects limited to ~850,000 pre-defined CpGs. To detect additional regulatory effects would require whole-genome bisulphite sequencing to be performed on samples. Whilst the associated expenses render this method prohibitive for isolated studies across large sample cohorts, coordinated multi-centre efforts will facilitate such approaches moving forward. Ultimately, analyses need to be extended to many cell types, as variants may function as QTLs in cells which are not considered here (e.g. fibroblasts), and may contribute to the pathogenic phenotype of these cells in RA. More comprehensive databases of molecular QTLs from diverse tissues and cell types will assist in unravelling the regulatory potential of GWAS hits. The Genotype-Tissue Expression (GTEX) project is one such example that provides an accessible resource for eQTL data, reporting associations across 42 tissues[344].

Due to the size of each cohort, the experimental design was largely powered to detect associations in cis, given the reduced burden of multiple testing correction involved in these analyses. However, a number of trans-meQTL effects were identified, some with relatively large effect sizes that occurred in both cell types. The necessity for large sample sizes to extensively map trans-meQTLs has hindered these analyses in most studies to date. However, a mechanism has been described whereby disease-associated variants influence DNAm in trans by altering the expression of transcription factors in cis[194]. These transcription factors then act in trans and, upon binding to their cognate sites throughout the genome, putatively modify local patterns of DNAm at these sites[194]. This hypothesis is supported by more recent findings that 28% of trans-meQTL 'hotspots' (i.e. those that are associated with many CpGs in trans) in whole blood are also cis-eQTLs, a large proportion of which regulate the expression of transcription factors[345]. That such trans-meQTLs are enriched at GWAS loci makes the identification of these effects an enticing approach for deciphering additional mechanisms and pathways downstream of genetic risk factors[194]. Nonetheless, the trans-meQTLs described in this chapter likely only represent a fraction of the total effects in these lymphocyte subsets, and moving forward greater sample sizes will be required to comprehensively map these associations.

Despite the limitations described in this section, the DNAm cis-regulatory landscape of two cell types that are central in the development of RA, as well as other autoimmune diseases, is described here for the first time in the context of early arthritis. Though a number of studies have identified disease-associated DNAm changes, some of these can be attributed in part to cis-meQTL regulatory variants. However, whilst DNAm is an important regulatory feature that has emerging roles in complex disease, impacts on cellular phenotype ultimately occur at the level of the transcriptome. For this reason, it is most useful to consider disease-associated DNAm changes in the context of transcriptional regulation, and this will be the focus of the following chapter.

Chapter 5: DNA methylation as a mediator of transcriptional regulation

5.1 Introduction

In the previous chapter, an extensive analysis of genetic effects on DNA methylation (DNAm), termed methylation quantitative trait loci (meQTLs), was described. Inferences regarding the regulatory potential of meQTLs at risk loci were made based on the mapping of such DNA methylation sites to regions harbouring active chromatin marks and transcription factor binding sites. In addition, given that associations had previously been identified between many of these loci and expression levels of candidate genes, a role for DNAm in mediating genetic effects on transcription was postulated. In the current chapter, matched cell-specific transcriptomic data from the same patients was integrated to examine how DNAm impacts the transcription of genes that might contribute to lymphocyte-mediated pathology. In the first instance, associations between cis-CpGs at risk loci and transcript levels of proximal genes, or cis-expression quantitative trait methylations (cis-eQTMs), were sought. Subsequently, a statistical approach was applied to infer mediation between these molecular traits, with the aim of revealing loci at which transcriptomic regulation occurs via a methylation intermediary. After identifying such DNAm-mediated effects at RA risk loci, as well as those of complex disease loci more generally, *in vitro* assays were leveraged to confirm observations at loci of interest in cross-sectional human data, and to attempt validation of regulatory mechanisms.

5.2 Expression Quantitative Trait Methylation Mapping

To identify genes for which transcription is potentially regulated downstream of risk-associated cis-meQTLs, associations between DNAm levels at cis-CpGs and transcripts within a 500Kb window up-/down-stream were identified. Expression data were available for 97.1% (100/103) of CD4$^+$ T cell samples and 91.6% (109/119) of B cell samples. Such associations, referred to as cis-eQTMs, were mapped in both cell types at all CpGs associated with risk loci (RA, MS, Asthma, and OA).

5.2.1 Expression Quantitative Trait Methylations at RA risk loci

In CD4$^+$ T cells, 29 CpG-Gene cis-eQTMs were identified (FDR < 0.01; Benjamini-Hochberg method) at RA risk loci, encompassing 20 CpGs and eight unique genes (Table 5.1). Amongst the genes implicated in this analysis were *ANKRD55*/*IL6ST* at chromosome 5q11.2, *JAZF1* at 7p15.1, and *FCRL3* at 1q23.1. Associations between cis-CpGs (cg10909506 & cg11187739) and *ORMDL3*/*GSDMB* occurred at the risk locus mapping to 17q12 (Figure 5.1). In 79.3% of cases (23/29) a negative association was observed between DNAm levels at the RA cis-CpG and transcript levels of the gene, indicative of a repressive function of DNAm.

CpG	CpG Coord.	Illumina ID	Gene Symbol	P-value	Adjusted p-value	Rho
cg21124310	chr5:55,444,106	ILMN_1798947	*ANKRD55*	4.26 x 10^{-12}	2.13 x 10^{-11}	-0.623
cg23343972	chr5:55,391,305	ILMN_1798947	*ANKRD55*	7.54 x 10^{-11}	1.88 x 10^{-10}	-0.594
cg10404427	chr5:55,443,844	ILMN_1798947	*ANKRD55*	1.16 x 10^{-10}	5.81 x 10^{-10}	-0.589
cg15667493	chr5:55,404,179	ILMN_1798947	*ANKRD55*	5.90 x 10^{-10}	1.47 x 10^{-9}	-0.570
cg15431103	chr5:55,457,410	ILMN_1798947	*ANKRD55*	6.23 x 10^{-8}	3.12 x 10^{-7}	-0.509
cg07522171	chr7:28,218,686	ILMN_1682727	*JAZF1*	4.47 x 10^{-9}	1.79 x 10^{-8}	-0.545
cg11187739	chr7:28,159,128	ILMN_1682727	*JAZF1*	4.96 x 10^{-8}	2.48 x 10^{-7}	-0.513
cg00184826	chr7:28,312,246	ILMN_1682727	*JAZF1*	5.82 x 10^{-5}	1.16 x 10^{-4}	0.391
cg16130019	chr7:28,060,181	ILMN_1682727	*JAZF1*	1.48 x 10^{-4}	7.39 x 10^{-4}	0.371
cg08519779	chr7:28,279,693	ILMN_1682727	*JAZF1*	9.94 x 10^{-4}	0.0040	0.324
cg01045635	chr1:157,670,481	ILMN_1691693	*FCRL3*	1.91 x 10^{-8}	3.82 x 10^{-8}	-0.526
cg17134153	chr1:157,670,328	ILMN_1691693	*FCRL3*	1.98 x 10^{-8}	3.95 x 10^{-8}	-0.526
cg21721331	chr1:157,670,877	ILMN_1691693	*FCRL3*	5.60 x 10^{-7}	1.12 x 10^{-6}	-0.476
cg19602479	chr1:157,670,869	ILMN_1691693	*FCRL3*	4.95 x 10^{-6}	9.89 x 10^{-6}	-0.439
cg08786003	chr1:157,670,710	ILMN_1691693	*FCRL3*	8.88 x 10^{-6}	1.78 x 10^{-5}	-0.428
cg25259754	chr1:157,670,220	ILMN_1691693	*FCRL3*	3.53 x 10^{-4}	7.06 x 10^{-4}	-0.350
cg18707136	chr1:157,790,767	ILMN_1797428	*FCRL3*	7.57 x 10^{-4}	0.0015	0.331
cg10909506	chr17:38,081,995	ILMN_1662174	*ORMDL3*	6.69 x 10^{-7}	1.07 x 10^{-5}	-0.473
cg18711369	chr17:38,081,186	ILMN_1662174	*ORMDL3*	3.16 x 10^{-6}	5.06 x 10^{-5}	-0.447
cg23343972	chr5:55,391,305	ILMN_1849013	*IL6ST*	3.48 x 10^{-5}	5.80 x 10^{-5}	-0.401
cg21124310	chr5:55,444,106	ILMN_1849013	*IL6ST*	3.60 x 10^{-5}	6.01 x 10^{-5}	-0.401
cg10404427	chr5:55,443,844	ILMN_1849013	*IL6ST*	6.25 x 10^{-5}	1.04 x 10^{-4}	-0.389
cg15667493	chr5:55,404,179	ILMN_1849013	*IL6ST*	3.39 x 10^{-4}	5.64 x 10^{-4}	-0.351
cg15431103	chr5:55,457,410	ILMN_1849013	*IL6ST*	3.53 x 10^{-4}	5.88 x 10^{-4}	-0.350
cg18711369	chr17:38,081,186	ILMN_1666206	*GSDMB*	2.94 x 10^{-5}	2.35 x 10^{-4}	-0.405
cg10909506	chr17:38,081,995	ILMN_1666206	*GSDMB*	1.02 x 10^{-4}	8.16 x 10^{-4}	-0.379
cg16213375	chr11:61,584,727	ILMN_1786759	*C11orf10*	1.39 x 10^{-4}	1.39 x 10^{-4}	-0.372
cg11187739	chr7:28,159,128	ILMN_2374770	*TAX1BP1*	6.14 x 10^{-4}	0.0015	0.337
cg07522171	chr7:28,218,686	ILMN_2374770	*TAX1BP1*	0.0044	0.0087	0.283

Table 5.1: CD4^{+} T cell cis-eQTM associations between CpGs associated with RA risk loci and genes ±500Kb. CpG-transcript association p-values were generated by Spearman's rank correlation, with adjustment for the total number of transcripts tested (i.e. number of probes within ±500Kb) for each CpG, using the Benjamini Hochberg method. Associations with an adjusted p-value < 0.01 were considered significant. 'Rho' depicts the Spearman's rho strength of association between DNAm and transcript levels, with negative values (< 0) representing a decrease in transcript levels associated with increased DNAm. Where cis-eQTMs were identified multiple probes mapping to the same gene, the probe exhibiting the strongest association with DNAm levels is reported here. CpG Coord = Genomic coordinated of CpG sites identified as cis-eQTMs; IlluminaID = unique Illumina identified for the transcript probe on the Illumina HumanHT-12 v4 array.

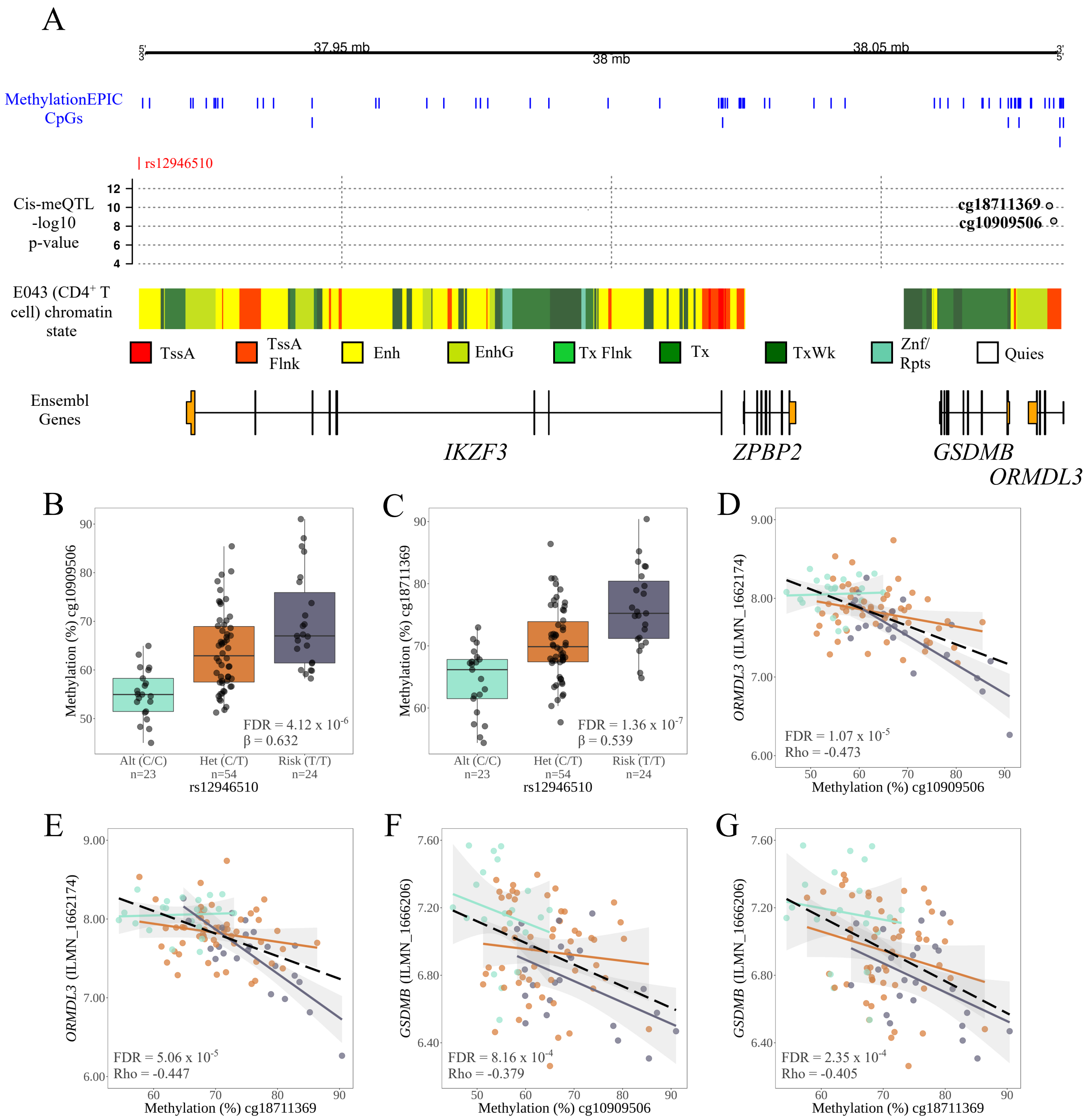


Figure 5.1: Associations between genotype, DNA methylation, and gene expression in CD4$^+$ T cells at the chromosome 17q12 locus. (A) This locus represents a risk factor in rheumatoid arthritis, multiple sclerosis, and asthma. A cis-meQTL was identified with the lead variant, rs12946510, associated with DNA methylation at two CpGs mapping to the *ORMDL3* gene. The risk allele conferred increased methylation at both B) cg10909506 and C) cg18711369. Accordingly, DNAm levels at these CpGs were negatively associated with transcript levels of *ORMDL3* (D-E) and *GSDMB* (F-G). Cis-meQTL p-values were generated by fitting additive linear models in the MatrixEQTL R package with false discovery rate (FDR) controlled using the Benjamini-Hochberg (BH) method. Cis-eQTM p-values were calculated by Spearman's rho correlation, again using the BH method to control FDR. Cis-meQTL boxplots in panels B-C display the median values with the inter-quartile range (IQR). The whiskers extent to maximum and minimum values no greater than 1.5x the upper and lower quantile respectively. Lines in D-G display the linear regression line with confidence intervals for each genotype subset, with the black dotted line representing the linear regression across all samples. Chromatin state data were obtained from the Roadmap Epigenomics Consortium[173]. TssA = active transcription start site (TSS), TssAFlnk = flanking active TSS, Enh = enhancer, EnhG = genic enhancer, TxFlnk = flanking transcription (5'/3'). Tx = strong transcription, TxWk = weak transcription, ZNF/Rpts = ZNF genes and repeats, Quies = quiescent/low.

In B cells, cis-eQTM analysis at RA cis-CpGs revealed 44 such associations mapping to 28 CpGs and eight unique genes (Table 5.2). Amongst these were cis-eQTMs at *FCRL3*, *ORMDL3*, and *GSDMB* that were also identified in CD4$^+$ T cells. Often, the CpG exhibiting the strongest association with transcript levels differed between the two cell types. One such example was the *FCRL3* locus at chromosome 1q23.1, at which the cg01045635 and cg17134153 exhibited the strongest correlation with *FCRL3* transcript levels in CD4$^+$ T cells, whereas in B cells this was most pronounced at cg21721331 and cg19602479 (Figure 5.2). This highlights a potential role for the RA risk variant in mediating expression of *FCRL3* in both cell types through its effects on DNAm at the gene promoter. These cis-CpGs (7 in CD4$^+$ T cells, 6 in B cells) mapped to a transcription start site (TSS) flanking region in B cells based on chromatin state information from the Roadmap Epigenomics Consortium (Primary B cells from peripheral blood, cell ID E032; Figure 5.2 A)[173].

Conversely, the entire *FCRL3*-spanning region was annotated as quiescent in primary T helper cells (Primary T helper cells from peripheral blood, cell ID E043; chromatin state data not shown). However, the same region overlaps a TSS and enhancer in primary regulatory T cells from peripheral blood (Treg, cell ID E044; Figure 5.2A), indicating that the observed effect may be restricted to this subset of CD4$^+$ T cells. Though meQTLs and eQTMs at the *FCRL3* promoter were active in both CD4$^+$ T cells and B cells at cg01045635 (Figure 5.2 B-C), cg17134153 (Figure 5.2 D-E), cg19602479 (Figure 5.2 F-G), and cg21721331 (Figure 5.2 H-I), DNAm levels at all CpGs were higher in the former cell type (Figure 5.2 B, D, F, H), which coincided with lower overall expression of *FCRL3* in this population relative to B cells (Figure 5.2 C, E, G, I).

Genes subject to cis-eQTM regulation exclusively in B cells in this analysis included *BLK*, *CCR6*, *FAM167A*, *IKZF3* and *IRF5* (Table 5.2). At RA-associated cis-eQTMs in B cells, 65.9% represented negative associations between DNAm and transcript levels, largely reflecting what was observed in CD4$^+$ T cells. At some genes, there were instances of both positive and negative putative regulatory effects conferred by cis-CpGs. One such example was at *FAM167A* on chromosome 8 in B cells. At this locus, DNAm at nine cis-CpGs was associated with expression of this gene, six of these displaying a positive correlation with transcript levels and three a negative correlation (Table 5.2).

CpG	CpG Coord.	Illumina ID	Gene Symbol	P-value	Adjusted p-value	Rho
cg16429190	chr8:11,340,719	ILMN_1687213	*FAM167A*	2.92×10^{-14}	2.92×10^{-13}	0.647
cg09528494	chr8:11,338,675	ILMN_3248511	*FAM167A*	3.93×10^{-10}	2.46×10^{-9}	0.555
cg01383082	chr8:11,323,474	ILMN_3248511	*FAM167A*	2.61×10^{-10}	2.61×10^{-9}	-0.559
cg21497594	chr8:11,366,745	ILMN_1687213	*FAM167A*	5.57×10^{-10}	4.27×10^{-9}	0.551
cg04986849	chr8:11,350,092	ILMN_3248511	*FAM167A*	7.81×10^{-8}	7.81×10^{-7}	0.487
cg11944933	chr8:11,320,112	ILMN_1687213	*FAM167A*	1.59×10^{-6}	1.01×10^{-5}	-0.441
cg23507676	chr8:11,395,642	ILMN_3248511	*FAM167A*	9.65×10^{-6}	4.40×10^{-5}	0.410
cg01527115	chr8:11,272,374	ILMN_1687213	*FAM167A*	5.45×10^{-6}	5.21×10^{-5}	-0.420
cg03002059	chr8:11,402,647	ILMN_1687213	*FAM167A*	1.01×10^{-4}	6.97×10^{-4}	0.364
cg12749226	chr17:38,077,703	ILMN_1662174	*ORMDL3*	8.88×10^{-12}	1.60×10^{-10}	-0.595
cg13200575	chr17:38,096,571	ILMN_1662174	*ORMDL3*	6.22×10^{-6}	1.12×10^{-4}	-0.418
cg18691862	chr17:38,096,648	ILMN_1662174	*ORMDL3*	8.07×10^{-6}	1.45×10^{-4}	-0.413
cg14348996	chr17:38,070,021	ILMN_1662174	*ORMDL3*	9.07×10^{-6}	1.54×10^{-4}	0.411
cg18711369	chr17:38,081,186	ILMN_1662174	*ORMDL3*	2.33×10^{-4}	0.0035	-0.346
cg21721331	chr1:157,670,877	ILMN_1691693	*FCRL3*	2.91×10^{-11}	1.75×10^{-10}	-0.583
cg19602479	chr1:157,670,869	ILMN_1691693	*FCRL3*	3.89×10^{-11}	2.33×10^{-10}	-0.580
cg01045635	chr1:157,670,481	ILMN_1699599	*FCRL3*	1.73×10^{-10}	1.04×10^{-9}	-0.564
cg15602298	chr1:157,670,825	ILMN_1797428	*FCRL3*	1.70×10^{8}	1.02×10^{-7}	-0.508
cg25259754	chr1:157,670,220	ILMN_1691693	*FCRL3*	2.00×10^{-5}	1.20×10^{-4}	-0.396
cg17134153	chr1:157,670,328	ILMN_1797428	*FCRL3*	4.32×10^{-5}	2.59×10^{-4}	-0.381
cg16429190	chr8:11,340,719	ILMN_1668277	*BLK*	4.44×10^{-8}	1.48×10^{-7}	-0.495
cg01383082	chr8:11,323,474	ILMN_1668277	*BLK*	5.37×10^{-6}	1.79×10^{-5}	0.420
cg21497594	chr8:11,366,745	ILMN_1668277	*BLK*	8.38×10^{-5}	2.51×10^{-4}	-0.368
cg09528494	chr8:11,338,675	ILMN_1668277	*BLK*	1.49×10^{-4}	4.97×10^{-4}	-0.355
cg23507676	chr8:11,395,642	ILMN_1668277	*BLK*	1.71×10^{-4}	5.14×10^{-4}	-0.352
cg04986849	chr8:11,350,092	ILMN_1668277	*BLK*	2.14×10^{-4}	7.14×10^{-4}	-0.347
cg01527115	chr8:11,272,374	ILMN_1668277	*BLK*	6.27×10^{-4}	0.0021	0.322
cg15222091	chr6:167,536,069	ILMN_1690907	*CCR6*	8.58×10^{-6}	2.57×10^{-5}	-0.412
cg16523158	chr6:167,535,171	ILMN_1690907	*CCR6*	1.04×10^{-5}	3.13×10^{-5}	-0.408
cg19954286	chr6:167,536,056	ILMN_1690907	*CCR6*	4.42×10^{-5}	1.33×10^{-4}	-0.381
cg05094429	chr6:167,536,184	ILMN_1690907	*CCR6*	9.52×10^{-5}	2.86×10^{-4}	-0.365
cg21794222	chr6:167,536,063	ILMN_1690907	*CCR6*	4.67×10^{-4}	0.0014	-0.330
cg14348996	chr17:38,070,021	ILMN_1666206	*GSDMB*	6.67×10^{-5}	5.67×10^{-4}	0.372
cg12655416	chr17:38,077,870	ILMN_1666206	*GSDMB*	1.82×10^{-4}	0.0033	-0.351
cg18711369	chr17:38,081,186	ILMN_1666206	*GSDMB*	3.88×10^{-4}	0.0035	-0.334
cg00288844	chr16:11,835,360	ILMN_1771862	*TXNDC11*	6.84×10^{-5}	8.20×10^{-4}	-0.372
cg12816198	chr7:128,577,593	ILMN_1670576	*IRF5*	1.81×10^{-4}	0.0016	-0.351
cg21497594	chr8:11,366,745	ILMN_1715680	*NEIL2*	8.75×10^{-4}	0.0020	0.314
cg09528494	chr8:11,338,675	ILMN_1724762	*XKR6*	0.0014	0.0035	-0.302
cg21473142	chr3:27,762,095	ILMN_2200917	*SLC4A7*	0.0020	0.0039	0.293
cg14348996	chr17:38,070,021	ILMN_3245973	*MSL1*	7.83×10^{-4}	0.0044	0.317
cg18691862	chr17:38,096,648	ILMN_2300695	*IKZF3*	9.64×10^{-4}	0.0087	0.312
cg18691862	chr17:38,096,648	ILMN_1707448	*CDK12*	0.0015	0.0090	0.301

Table 5.2: B cell cis-eQTM associations between CpGs associated with RA risk loci and genes ±500Kb. CpG-transcript association p-values were generated by Spearman's rank correlation, with adjustment for the total number of transcripts tested (i.e. number of probes within ±500Kb) for each CpG, using the Benjamini Hochberg method. Associations with an adjusted p-value < 0.01 were considered significant. 'Rho' depicts the Spearman's rho strength of association between DNAm and transcript levels, with negative values (< 0) representing a decrease in transcript levels associated with increased DNAm. Where cis-eQTMs were identified multiple probes mapping to the same gene, the probe exhibiting the strongest association with DNAm levels is reported here. CpG Coord = Genomic coordinated of CpG sites identified as cis-eQTMs; IlluminaID = unique Illumina identified for the transcript probe on the Illumina HumanHT-12 v4 array.

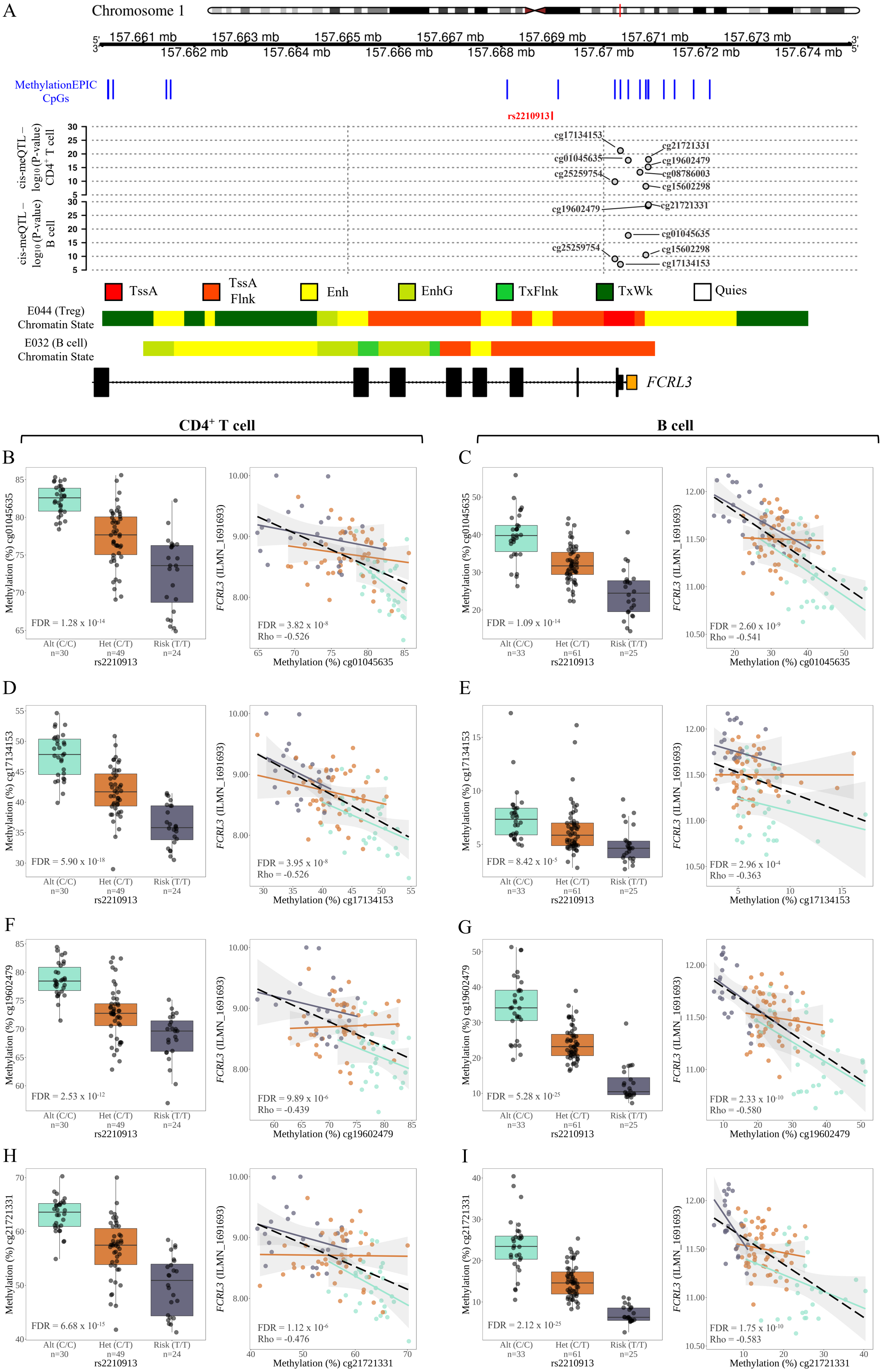

A
Chromosome 1
157.661 mb
157.663 mb
157.665 mb
157.667 mb
157.669 mb
157.671 mb
157.673 mb
157.662 mb
157.664 mb
157.666 mb
157.668 mb
157.67 mb
157.672 mb
157.674 mb
MethylationEPIC CpGs
rs2210913
cis-meQTL – log10 (P-value) CD4+ T cell
cis-meQTL – log10 (P-value) B cell
cg17134153
cg01045635
cg25259754
cg21721331
cg19602479
cg08786003
cg15602298
cg19602479
cg21721331
cg01045635
cg25259754
cg15602298
cg17134153
TssA
TssA Flnk
Enh
EnhG
TxFlnk
TxWk
Quies
E044 (Treg) Chromatin State
E032 (B cell) Chromatin State
FCRL3
CD4+ T cell
B cell
B
C
D
E
F
G
H
I
Methylation (%) cg01045635
FDR = 1.28 x 10-14
Alt (C/C) n=30
Het (C/T) n=49
Risk (T/T) n=24
rs2210913
FCRL3 (ILMN_1691693)
FDR = 3.82 x 10-8
Rho = -0.526
Methylation (%) cg01045635
FDR = 1.09 x 10-14
Alt (C/C) n=33
Het (C/T) n=61
Risk (T/T) n=25
FDR = 2.60 x 10-9
Rho = -0.541
Methylation (%) cg17134153
FDR = 5.90 x 10-18
FDR = 3.95 x 10-8
Rho = -0.526
FDR = 8.42 x 10-5
FDR = 2.96 x 10-4
Rho = -0.363
Methylation (%) cg19602479
FDR = 2.53 x 10-12
FDR = 9.89 x 10-6
Rho = -0.439
FDR = 5.28 x 10-25
FDR = 2.33 x 10-10
Rho = -0.580
Methylation (%) cg21721331
FDR = 6.68 x 10-15
FDR = 1.12 x 10-6
Rho = -0.476
FDR = 2.12 x 10-25
FDR = 1.75 x 10-10
Rho = -0.583

Figure 5.2 (Previous page): cis-meQTL and cis-eQTM associations at the *FCRL3* promoter in CD4$^+$ T cells and B cells. A) A cis-meQTL (lead SNP rs2210913) was identified for which the RA risk allele (T) in CD4$^+$ T cells was associated with reduced methylation at seven cis-CpGs mapping to the *FCRL3* promoter. The same associations were also identified at six out of these seven CpGs in B cells. Based on chromatin state data from the Roadmap Epigenomics Project[173], these cis-CpGs were found to map to active chromatin regions (active transcription start site, TssA; flanking active transcription start site, TssAFlnk; enhancer, Enh) in regulatory T cells (Tregs, E044) and B cells (E032) from peripheral blood. B-I) Plots of cis-meQTL and cis-eQTM associations are shown for each cell type at (B-C) cg01045636 and (D-E) cg17134153 that displayed the strongest DNA methylation-transcript associations in CD4$^+$ T cells, as well as (F-G) cg19602479 and (H-I) cg21721331 for which the associations were strongest in B cells. Left panels in figures B-I represent cis-meQTL associations between the risk SNP and DNAm levels, whereas the right panels are cis-eQTM associations between the risk cis-CpGs and *FCRL3* transcript levels. Cis-meQTL and cis-eQTM p-values, as well as summary statistics for boxplots and lines of best fit are derived as described in Figure 5.1. TssA = active transcription start site (TSS), TssAFlnk = flanking active TSS, Enh = enhancer, EnhG = genic enhancer, TxFlnk = flanking transcription (5'/3'), TxWk = weak transcription, Quies = quiescent/low.

5.2.2 Expression quantitative trait methylations at additional complex disease risk loci.

To extend the meQTL analysis, associations were also mapped at cis-CpGs associated with multiple sclerosis (MS), asthma, and osteoarthritis (OA) risk loci in CD4$^+$ T cells. At MS loci, 39 unique CpG-Gene eQTMs were identified (29 CpGs, 16 genes), whilst at asthma loci 27 such associations were discovered (23 CpGs, 11 genes). Of note, a number of eQTMs were implicated in all three immune-mediated diseases, including those acting on *ORMDL3*/*GSDMB* (cg10909506, cg18711369; Figure 5.1) and *JAZF1* (cg07522171, cg11187739, cg00184826, cg16130019, cg08519779). Additional genes highlighted in these analyses were *ANKRD55*/*IL6ST* (also implicated in RA), *CDK2AP1*, *FAM164A*, and *RSG14* at MS loci, as well as *CD247*, *IL18R1*, and *RERE* at asthma loci. At OA-associated cis-CpGs, 57 cis-eQTMs were associated with 36 CpGs and 7 genes including *LRRC37A4*, *GRINA*, and *PLEC*. Overlap with genes implicated in the immune-mediated diseases (IMDs) was limited, with the only instance being *CDK2AP1* (cg01030110) which was also identified as an MS candidate gene. The full list of cis-eQTMs at all risk loci in CD4$^+$ T cells is given in Appendix G.

As regards cis-eQTMs at risk loci for other IMDs in B cells, 28 MS-associated cis-eQTMs (20 CpGs, 10 genes) and 30 asthma-associated cis-eQTMs (18 CpGs, 10 genes) were identified. As was also seen in CD4$^+$ T cells, transcript levels of *ORMDL3* and *GSDMB* were associated with cis-CpGs at RA, MS, and asthma risk loci. In addition, *FAM164A*, *EAF2*, *RGS1* and *SHMT1* at MS loci, as well as *PGAP3*, *TNFSF4*, *RERE*, and *MRPL45P2* at asthma loci were highlighted as candidate genes in B cells with as potentially being subject to DNAm-mediated regulation in cis (Appendix G). OA-associated cis-CpGs in B cells were associated with 64 eQTMs (46 unique CpGs, 11 unique genes), with only *HIP1R* overlapping with any of the IMDs (MS). All B cell risk cis-eQTMs are reported in Appendix G.

5.3 Causal inference testing

The coincidence of cis-meQTLs at a risk locus with a cis-eQTM association may be indicative of an underlying mechanism whereby DNAm mediates transcriptional regulation by the risk variant, referred to as the SNP → Methylation → Expression (SME) model (see Chapter 1.6.9). However, identification of correlations between molecular traits alone is insufficient to distinguish the SME model from one of reverse causation, whereby modulation of DNAm occurs downstream of direct SNP effects on transcript levels (SNP → Expression → Methylation; SEM), or independent regulation of DNAm and gene expression (INDEP). To reveal associations fitting the SME regulatory model, a causal inference test (CIT)[260] was applied to all triplets (SNP, CpG, transcript) at risk loci (RA, MS, asthma, and OA) for which a cis-meQTL and cis-eQTM effect had been observed. CIT was performed treating the CpG site as a potential mediator and the transcript as the phenotype of interest, with a permutation-based approach to calculate FDR values.

In CD4$^+$ T cells, CIT implicated DNAm as a mediator of genetic risk at five RA risk loci, nominating eight candidate genes (*ANKRD55*, *JAZF1*, *ORMDL3*, *FCRL3*, *IL6ST*, *C11orf10*, *TAX1BP1*, and *GSDMB*) as being subject to DNAm-mediated transcriptional regulation (Table 5.3A; Appendix H for full results). Contrastingly, in B cells CIT only implicated one locus, *FCRL3* at chromosome 1q23.1, as showing strong evidence according to the SME model at FDR < 0.05 (Table 5.3B & Appendix H for full results). Nonetheless, *CCR6*, *IKZF3*, and *ORMDL3* were potentially implicated under this model at nominal significance (CIT p-value < 0.05, FDR < 0.15). These results would therefore suggest that DNAm functions to regulate expression of the *FCRL3* gene in both CD4$^+$ T cells and B cells, and that genetic risk at this particular locus may influence the phenotype of multiple cell types in RA pathogenesis.

A

Gene	CpG	Lead meQTL SNP	Bayes Coloc. PP4†	Locus	CIT P-Value	CIT Permutation FDR
CD4+ T cells:						
ANKRD55	cg21124310	rs6859219	1.00	5q11.2	1.11 x 10^{-4}	7.06 x 10^{-4}
	cg10404427	rs6859219	1.00		0.0057	0.0044
	cg23343972	rs6859219	1.00		0.0062	0.0069
	cg15431103	rs6859219	0.95		0.0496	0.0319
JAZF1	cg07522171	rs2189966	0.97	7p15.1	3.97 x 10^{-4}	0.0035
	cg11187739	rs4722758	0.99		0.0035	0.0044
	cg16130019	rs917117	0.99		0.0529	0.0319
ORMDL3	cg18711369	rs12946510	0.86	17q12	4.46 x 10^{-4}	0.0035
	cg10909506	rs12946510	0.93		0.0016	0.0044
FCRL3	cg17134153	rs2210913	0.99	1q23.1	0.0027	0.0044
	cg01045635	rs2210913	0.99		0.0120	0.0296
IL6ST	cg15431103	rs6859219	0.95	5q11.2	0.0100	0.0296
	cg15667493	rs6859219	0.99		0.0139	0.0296
	cg10404427	rs6859219	1.00		0.0216	0.0305
	cg21124310	rs6859219	1.00		0.0349	0.0305
	cg23343972	rs6859219	1.00		0.0352	0.0305
C11orf10	cg16213375	rs61897793	1.00	11q12.2	0.0163	0.0296
TAX1BP1	cg11187739	rs4722758	0.99	7p15.1	0.0470	0.0305
GSDMB	cg18711369	rs12946510	0.86	17q12	0.0277	0.0305
	cg10909506	rs12946510	0.93		0.0448	0.0305

B

B cells:						
FCRL3	cg19602479	rs2210913	0.99	1q23.1	4.69 x 10^{-4}	0.0420
	cg01045635	rs7522061	0.97		5.49 x 10^{-4}	0.0420
CCR6	cg15222091	rs3093025	0.98	6q27	0.0101	0.0966
	cg19954286	rs3093025	0.98		0.0258	0.1330
	cg05094429	rs3093025	0.96		0.0347	0.1330
IKZF3	cg18691862	rs9903250	0.44	17q12	0.0249	0.1330
ORMDL3	cg12749226	rs11557466	0.97	17q12	0.0249	0.1330

Table 5.3: Causal inference testing (CIT) results for CD4^{+} T cells and B cells. CIT was performed on all triplets (SNP, CpG, and transcript) at RA risk loci exhibiting significant cis-meQTL and cis-eQTM associations. P-values for the CIT were generate using the cit R package[298], and FDR values calculated by performing 1000 permutations of the data. For genes where more than one CpG-Probe association was observed, CIT was performed using both, and the triplet returning the lowest CIT FDR reported in this table. †The prior probability of the meQTL effect and disease association sharing a single causal SNP at a given locus as determined by Bayesian co-localisation analysis (see section 4.6).

Of note, all of the loci at which CIT highlighted a SME model of genetic regulation also exhibited a strong probability (PP4 > 0.85) of co-localisation between the meQTL variant and RA-associated variant. This provided further confidence in concluding that DNAm is an important mediator of genetic risk at these particular loci. An additional analysis was performed to treat transcript levels as mediators with CpG methylation as the outcome measure (i.e. SEM model), though no such instances were identified at FDR < 0.05., suggesting that the remaining associations are likely explained by the INDEP model.

5.3.1 Pleiotropy and the 5q11.2 risk locus

One interesting observation was that DNAm at RA loci exhibited pleiotropic regulation of gene expression at a number of loci in CD4$^+$ T cells. One prominent example was the RA risk locus on chromosome 5q11.2 at which the risk variant (rs6859219) is associated with DNAm at five cis-CpGs (Figure 5.3A), four of which map to the gene body of *ANKRD55* (cg21124310, cg10404427, cg15431103, and cg15667493) and the remaining one being intergenic (cg23343972). CIT inferred that methylation at four of these CpG sites (cg21124310, cg10404427, cg23343972, and cg15431103) likely mediates expression of not only the *ANKRD55* gene, but also *IL6ST* for which the TSS is 238Kb downstream of the *ANKRD55* TSS on the reverse strand (Table 5.3A). In both instances, the risk allele (C) at rs6859219 conferred reduced DNAm at these sites, which was associated with increased expression of the two genes. The two cis-CpGs with the strongest genetic association, as well as the strongest eQTM effect and highest probability of a SME regulatory model for *ANKRD55* (cg21124310, cg10404427; Figure 5.3 A-C) mapped to a CD4$^+$ T cell enhancer in the Intron 5 of *ANKRD55*. The CpG exhibiting the next strongest association with *ANKRD55*, cg23343972 (Figure 5.3D), maps to an intronic enhancer downstream of *ANKRD55* (Figure 5.3A).

Using chromosome conformation data from capture Hi-C experiments in CD4$^+$ T cells[121], it was identified that the intronic enhancer region harbouring cg21124310 and cg10404427, as well as the rs6859219 SNP, interacts with the promoter of *IL6ST* (Figure 5.4). This may therefore explain the co-regulation of both *ANKRD55* and *IL6ST* by these CpGs. Indeed, the cis-CpG for which the SME model was most likely at *IL6ST* based on CIT (cg15431103; Table 5.3 A) was also found to interact with the *IL6ST* promoter (Figure 5.4).

These data indicate that a one-to-one association between genetic risk and transcriptional regulation, whereby genetic risk at a given locus manifests as modified expression of a single gene, may be overly simplistic. Confirming whether or not this regulatory SNP at chromosome 5q11.2 confers RA risk through DNAm-dependent altered expression of both *ANKRD55* and *IL6ST* will require further functional studies in T cells, though both of these genes warrant follow-up. An additional example of a pleiotropic effect was observed for associations at the chromosome 17q12 risk locus illustrated in Figure 5.1. Here, methylation at the cis-CpGs cg18711369 and cg10909506 mediated expression of both *ORMDL3* and *GSDMB* in CD4$^+$ T cells, further highlighting how DNAm can function in gene co-expression.

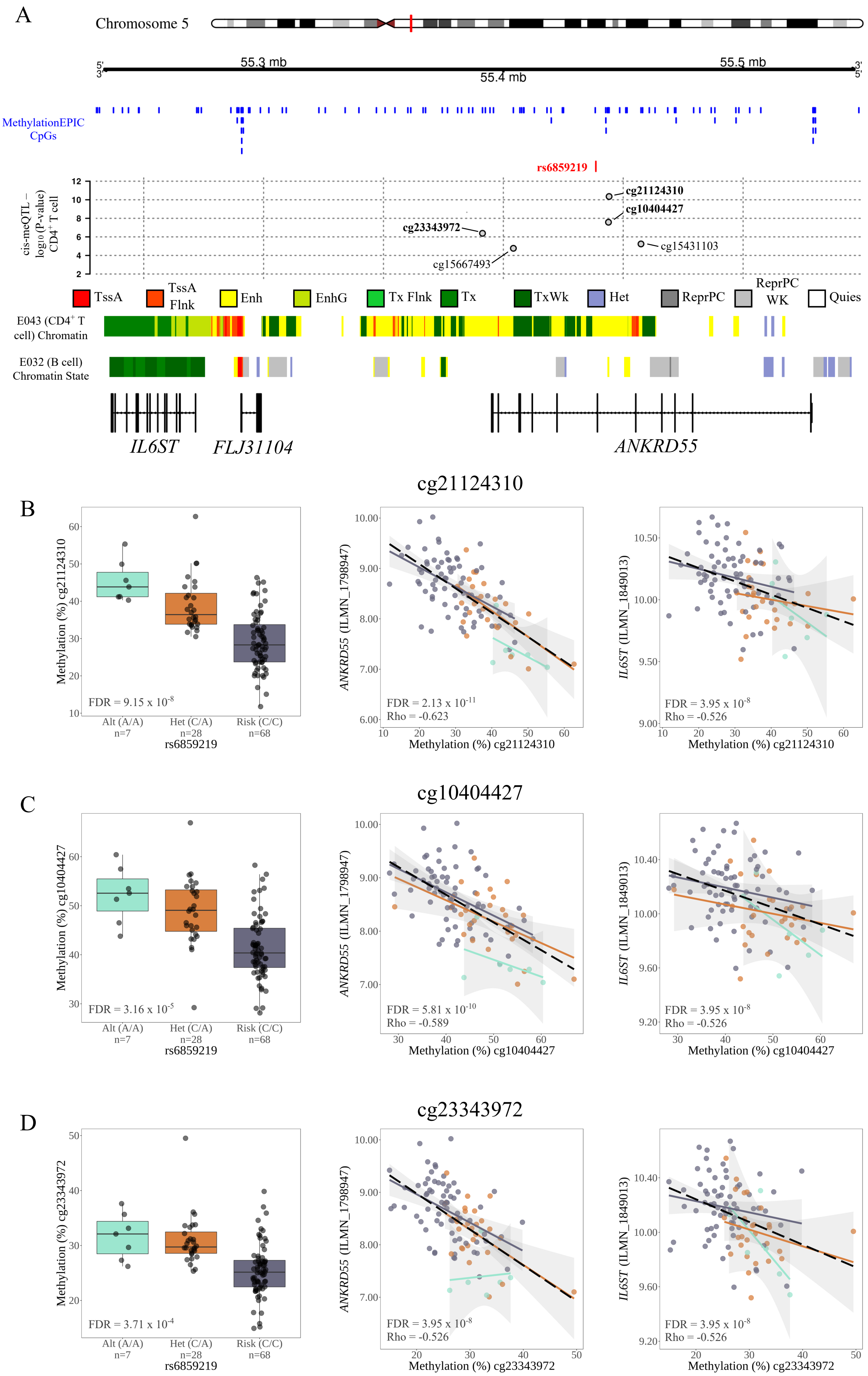

A
Chromosome 5
5'
3'
55.3 mb
55.4 mb
55.5 mb
3'
5'
MethylationEPIC
CpGs
rs6859219
cis-meQTL –
log10 (P-value)
CD4+ T cell
12
10
8
6
4
2
cg21124310
cg10404427
cg23343972
cg15431103
cg15667493
TssA
TssA Flnk
Enh
EnhG
Tx Flnk
Tx
TxWk
Het
ReprPC
ReprPC WK
Quies
E043 (CD4+ T cell) Chromatin
E032 (B cell) Chromatin State
IL6ST
FLJ31104
ANKRD55
cg21124310
B
Methylation (%) cg21124310
60
50
40
30
20
10
FDR = 9.15 x 10-8
Alt (A/A) n=7
Het (C/A) n=28
Risk (C/C) n=68
rs6859219
ANKRD55 (ILMN_1798947)
10.00
9.00
8.00
7.00
6.00
FDR = 2.13 x 10-11
Rho = -0.623
10 20 30 40 50 60
Methylation (%) cg21124310
IL6ST (ILMN_1849013)
10.50
10.00
9.50
9.00
FDR = 3.95 x 10-8
Rho = -0.526
10 20 30 40 50 60
Methylation (%) cg21124310
cg10404427
C
Methylation (%) cg10404427
60
50
40
30
FDR = 3.16 x 10-5
Alt (A/A) n=7
Het (C/A) n=28
Risk (C/C) n=68
rs6859219
ANKRD55 (ILMN_1798947)
10.00
9.00
8.00
7.00
FDR = 5.81 x 10-10
Rho = -0.589
30 40 50 60
Methylation (%) cg10404427
IL6ST (ILMN_1849013)
10.40
10.00
9.60
9.20
FDR = 3.95 x 10-8
Rho = -0.526
30 40 50 60
Methylation (%) cg10404427
cg23343972
D
Methylation (%) cg23343972
50
40
30
20
FDR = 3.71 x 10-4
Alt (A/A) n=7
Het (C/A) n=28
Risk (C/C) n=68
rs6859219
ANKRD55 (ILMN_1798947)
10.00
9.00
8.00
7.00
FDR = 3.95 x 10-8
Rho = -0.526
20 30 40 50
Methylation (%) cg23343972
IL6ST (ILMN_1849013)
10.40
10.00
9.60
9.20
FDR = 3.95 x 10-8
Rho = -0.526
20 30 40 50
Methylation (%) cg23343972

Figure 5.3 (Previous page): cis-meQTL effects mediate the co-expression of *ANKRD55* and *IL6ST* in CD4$^+$ T cells. A) At the RA risk locus on chromosome 5q11.2, the regulatory SNP (rs6859219) is associated with DNAm levels at four intronic cis-CpGs and on intergenic cis-CpG. All of these CpGs map to CD4$^+$ T cell (E044) enhancer elements defined by the Roadmap Epigenomics Project[173]. Interestingly, DNA methylation of four of these cis-CpGs was associated with expression of both *ANKRD55* and *IL6ST* in this cell type as determined by causal inference testing (CIT). Cis-meQTL associations, as well as cis-eQTMs at *ANKRD55* and *IL6ST* are shown for three of the cis-CpGs most strongly implicated in mediating transcript levels by CIT: B) cg21124310, C) cg10404427, and D) cg23343972. Panels in figures B-D represent from left to right: cis-meQTL associations, cis-eQTM associations between the cis-CpG and *ANKRD55*, and cis-eQTM associations between the cis-CpG and *IL6ST*. Cis-meQTL and cis-eQTM p-values, as well as summary statistics for boxplots and lines of best fit are derived as described in Figure 5.1. TssA = Active transcription start site, TssA Flnk = flanking active transcription start site, Enh = enhancer, EnhG = genic enhancer, Tx Flnk = flanking (5'/3') transcribed, Tx = strong transcription, TxWk = weak transcription, Het = Heterochromatin, ReprPC = repressed polycomb, ReprPCWk = weak repressed polycomb, Quies = quiescent/low.

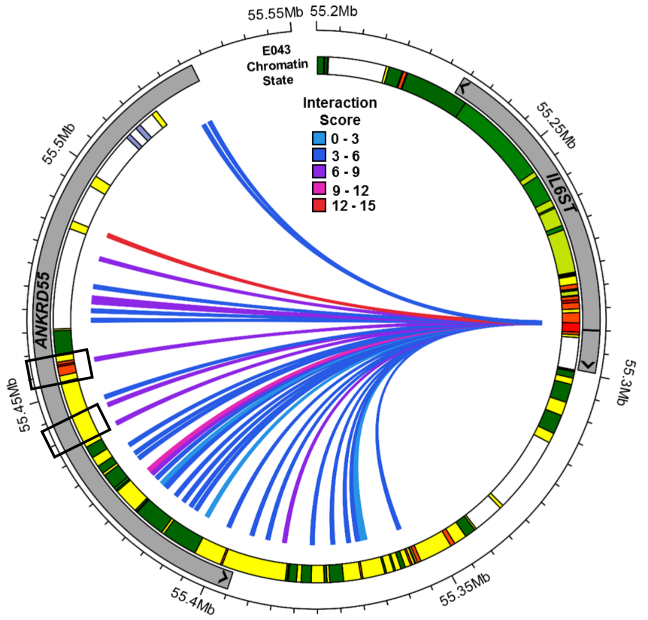


Figure 5.4: Circos plot showing chromosome interactions between intronic cis-CpGs at the *ANKRD55* gene on chromosome 5q11.2 and the *IL6ST* promoter region. Capture HiC interaction data for total CD4$^+$ T cells were obtained from a published study of multiple cell types[121]. The interaction scores had been generated using the CHiCAGO system, and represent the confidence of an interaction being present, with greater confidence placed in higher values[121]. Chromatin state data from the Roadmap Epigenomics project are shown for primary T helper cells from peripheral blood (E043). Black boxes indicate the locations of the two regions harbouring risk-associated cis-CpGs in CD4$^+$ T cells.

5.3.2 DNA methylation-mediated transcriptional regulation explains shared genetic risk of immune-mediated diseases

Given that in chapter 4 it was identified that a significant proportion of risk loci for other IMDs, namely MS and asthma, also function as lymphocyte cis-meQTLs, putative DNAm-mediated transcriptional regulation was also sought at these loci using CIT. At MS loci, DNAm was found to potentially mediate the expression of 11 genes in CD4$^+$ T cells (*ANKRD55*, *JAZF1*, *ORMDL3*, *FAM164A*, *TAX1BP1*, *IL6ST*, *RAB24*, *MRPL45P2*, *SHMT1*, *GSDMB*, and *ZNF688*; Appendix H). No such associations were robust to FDR correction in B cells, although a number of candidate genes were identified at nominal significance (CIT p-value < 0.05), including *SHMT1*, *FAM164A*, *HIP1R*, and *RGS1* (Appendix H). A number of genes were also implicated by the SME model at asthma loci in CD4$^+$ T cells, with *JAZF1*, *ORMDL3*, *CD247*, and *GSDMB* highlighted at these loci. As was the case for MS, no asthma loci were robustly linked to the regulation of B cell gene expression via DNAm, though *GSDMB*, *ORMDL3*, *PGAP3*, and *IKZF3* were amongst those displaying some evidence ($p < 0.05$) of an SME mechanism. CIT was also applied to triplets at OA loci, though no instances of SME regulation of candidate genes were found in either CD4$^+$ T cells or B cells at FDR < 0.05 (Appendices I & J).

The degree of overlap that occurred across IMDs in respect of genes highlighted by CIT in CD4$^+$ T cells was striking. For example, expression of the *JAZF1* gene was linked to DNAm at two CpGs, cg07522121 and cg11187739, both of which were associated with risk loci for RA, MS, and asthma. Though the regulatory SNP associated with each of these cis-CpGs was distinct (rs2189966 for cg07522171, rs4722758 for cg11187739), these are in high linkage disequilibrium ($r^2 = 0.88$ in EUR populations), and as such likely tag the same causal SNP. The risk allele at the regulatory SNPs (C at rs2189966, G at rs4722758) was consistent across all three conditions and was associated with increased DNAm at cg07522171 (Figure 5.5A) and cg11187739 (Figure 5.5 B), respectively. This risk-associated increase in DNAm at both cis-CpGs in turn correlated with reduced transcript levels of the *JAZF1* gene in CD4$^+$ T cells (Figure 5.5C-D).

Additional instances of overlap occurred at *ORMDL3* and *GSDMB*, which were shared across all three IMDs examined, as well as at *ANKRD55*/*IL6ST*, which was common to both RA and MS (Figure 5.6 A; see also section 5.3.1). No overlap was apparent in DNAm-mediated regulatory mechanisms between IMD risk loci and OA (Figure 5.6A).

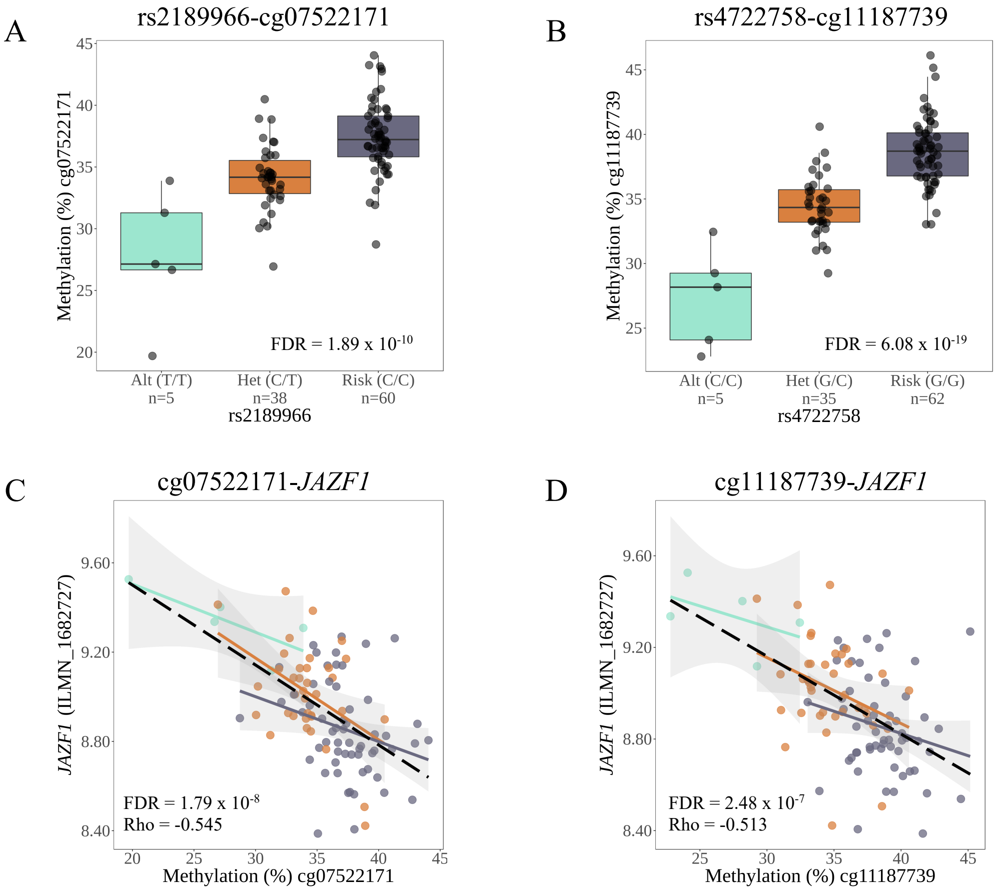


Figure 5.5: Molecular associations between genotype, DNA methylation, and gene expression at chromosome 7p15.1. Cis-meQTLs at (A) cg07522171 and (B) cg11187739 in CD4$^+$ T cells maps to a locus on chromosome 7p15.1 that confers genetic risk in rheumatoid arthritis, multiple sclerosis, and asthma. Cis-eQTM associations between DNA methylation at (C) cg07522171 and (D) cg11187739 highlight an association with *JAZF1* transcript levels, with causal inference testing confirming that methylation at these CpGs regulated expression of this gene in cis. Cis-meQTL and cis-eQTM p-values, as well as summary statistics for boxplots and lines of best fit are derived as described in Figure 5.1.

Unsurprisingly given that only *FCRL3* was identified as a DNAm-mediated candidate gene in B cells, overlap in between IMDs in this cell type was less marked (Figure 5.6B). The SME model at *ORMDL3* which was involved in CD4$^+$ T cell-mediated genetic risk for RA, MS, and asthma was potentially implicated in B cells (CIT p-value < 0.05), though was not robust to permutation-based FDR calculation (Figure 5.6B). *HIP1R* was implicated at nominal significance in OA and MS in B cells, potentially highlighting a mechanistic overlap in genetic risk between these aetiologically distinct conditions (Figure 5.6B)

These results indicate that CD4$^+$ T cell DNAm may have a critical role in mediating the expression of genes that are involved in IMDs generally. This could signify that genetic risk at specific loci acts to perturb pathways that contribute to T cell responses, with additional factors, be they genetic or environmental, determining the clinical manifestations in susceptible individuals.

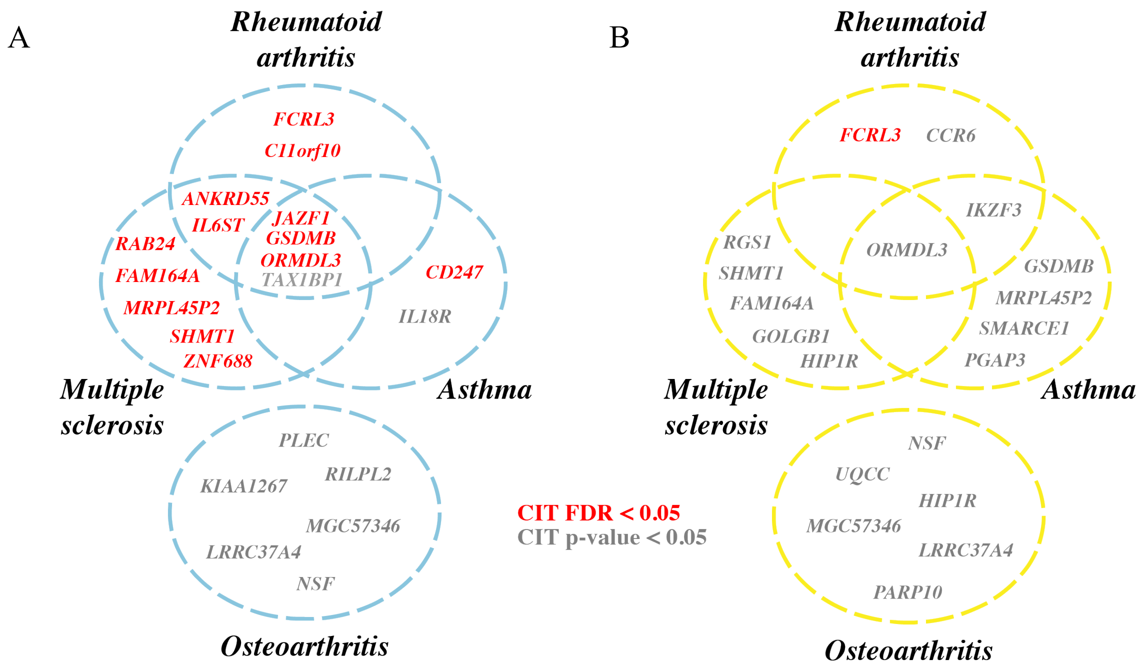


Figure 5.6: Causal inference testing highlights genes for which DNA methylation is a mediator of genetic risk across immune-mediated diseases (IMDs). Genes for which the SNP → Methylation → Expression regulatory model was implicated by causal inference testing at FDR < 0.05 are highlighted in red, whereas those showing suggestive evidence at p < 0.05 are shown in grey. The overlap in methylation-mediated genes between three immune-mediated diseases – rheumatoid arthritis, multiple sclerosis, and asthma is shown, as well as those in osteoarthritis, a condition in which adaptive immunity has a less prominent pathophysiological function.

5.4 Validation of RA-associated cis-meQTL Effects

Several meQTLs at RA-associated loci that were identified in the discovery cohort using the MethylationEPIC BeadChip array were subject to validation using an independent patient cohort, as well as an independent technique for quantifying DNA methylation (DNAm). For the purpose of validation, cg17134153 (*FCRL3*), cg21124310 (*ANKRD55*), and cg07522171 (*JAZF1*) were selected as causally implicated CpGs at which DNAm appears to mediate genetic effects on gene expression. Bisulphite pyrosequencing was employed as a method for accurate, targeted DNAm quantification at these CpGs of interest.

5.4.1 Validation of pyrosequencing assays for DNA methylation quantification

In order to determine whether pyrosequencing assays accurately quantified DNAm at these CpG sites, a standard curve was generated for each assay. This involved generating mixes of synthetic bisulphite DNA fragments harbouring either a C or T allele at the CpG site of interest, representing the methylated and unmethylated cytosine residues, respectively. These allele mixes were generated to mimic a pool of cells with varying DNAm levels from 0 – 100%, at increments of 10%.These were amplified and sequenced in order to assess whether or not the percentage of C allele called by the respective assays reflected that which was expected from the probe mix. Using this approach, it was confirmed that all assays were able to accurately determine the allele proportions across the range of values tested at cg17134153 (Figure 5.7A), cg21124310 (Figure 5.7B), and cg07522171 (Figure 5.7C), thus confirming their suitability for DNAm analysis at the selected cis-CpGs.

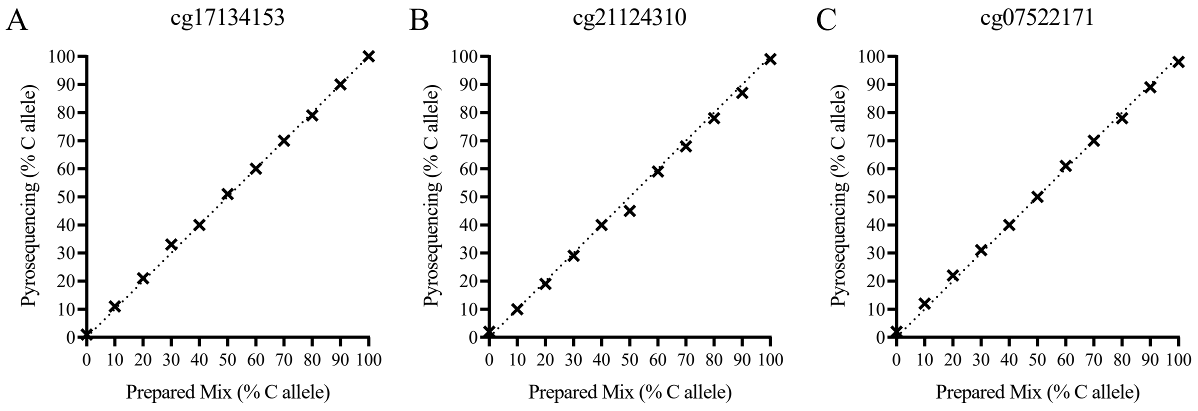


Figure 5.7: Standard curves generated for CpG pyrosequencing assays. Allele mixes were prepared by combining varying proportions of DNA fragments harbouring either a C or T allele at the position of interest. Pyrosequencing was performed on mixes with each custom assay to validate the accuracy of readings. The crosses each represent the mean methylation percentage from two PCR replicated per template, with readings having high accuracy falling on the x=y dotted line.

5.4.2 Independent Validation of cis-meQTL effects

For each CpG site outlined in section 5.4.1, validation cohorts of 39 patients were selected that had not been included in the initial array discovery cohort, but for whom genotype data and CD4$^+$ T cell DNA were available (Table 5.4). Individuals were selected so that each genotype (risk allele homozygote, heterozygote, alternative allele homozygote) at the regulatory SNP (rs2210913, rs6859219, rs2189966) was represented by at least 3 individuals.

	Patient Number	RA (%)	Age	Sex (%F)	CRP
cg17134153 meQTL	39	15	50 (41 – 60)	69	5 (5 – 11)
cg21124310 meQTL	39	15	49 (40 – 65)	74	5 (5 – 9)
cg07522171 meQTL	39	15	48 (37 – 58)	69	5 (5 – 11)

Table 5.4: Demographic and clinical characteristics of independent cohorts for validation of methylation quantitative trait locus (meQTL) effects. Age and CRP levels are presented as group median (inter-quartile range)

DNAm was subsequently quantified in the validation cohort at the three selected CpGs; cg17134153 (*FCRL3*), cg21124310 (*ANKRD55/IL6ST*), and cg07522171 (*JAZF1*). This confirmed the association ($p < 0.05$) between genotype at the respective regulatory SNP, and DNAm at cg17134153 ($p < 1 \times 10^{-4}$; Figure 5.8A), cg21124310 ($p = 4 \times 10^{-4}$; Figure 5.8B), and cg07522171 ($p = 5 \times 10^{-4}$; Figure 5.8C). Importantly, the allelic effect was consistent with that observed in the discovery cohort, with the risk allele conferring either increased DNAm levels (rs2189966/cg07522171) or decreased levels (rs2210913/cg17134153, rs6859219/cg21124310).

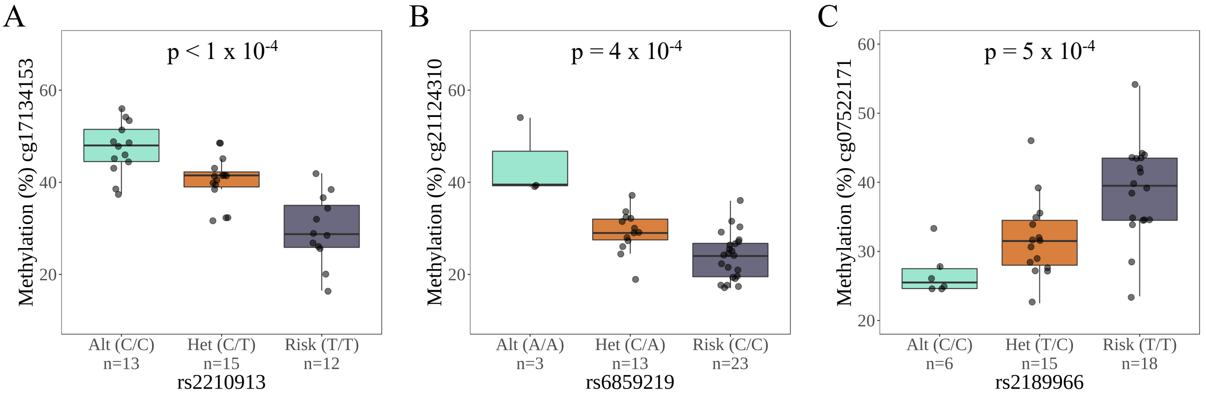


Figure 5.8: Validation of CD4$^+$ T cell cis-meQTL effects at RA risk loci using bisulphite pyrosequencing in an independent cohort of patients. MeQTLs mapping to (A) rs2210913/cg17134153, (B) rs6859219/cg21124310, and (C) rs2189966/cg07522171 were validated in patient isolated CD4$^+$ T cells. P-values were generated using a one-way ANOVA. The box plots indicate the group medians with the 1st and 3rd quantiles, and whiskers extend to the largest and smallest values to a limit of 1.5 times the 25th of 75th quartile respectively.

5.5 Allelic Expression Analysis

The presence of a transcript SNP, rs7522061, in the *FCRL3* coding region in high linkage disequilibrium with the regulatory SNP, rs2210913 ($r^2 = 0.90$), allowed for allelic expression analysis to be performed to confirm eQTL associations at this locus. Thirty-three patients who were heterozygous at the regulatory SNP of interest were selected for allelic expression analysis (Table 5.5). For the other two loci (rs6859219 (*ANKRD55/IL6ST*) and rs2189966 (*JAZF1*)), the absence of an appropriate proxy transcript SNP precludes the above allelic expression analysis. However, expression data from the Illumina HumanHT-12 array was available for independent cohorts of patients to confirm SNP-gene associations (Table 5.5).

	Patient Number	RA (%)	Age	Sex (%F)	CRP
FCRL3 allelic expression	33	52	54 (51 – 69)	85	5 (5 – 10)
ANKRD55/ *IL6ST* eQTL	39	18	49 (40 – 65)	77	5 (5 -10)
JAZF1 eQTL	41	17	49 (37 – 59)	78	5 (5 – 12)

Table 5.5: Demographic and clinical characteristics of independent cohorts for allelic expression analysis and validation of expression quantitative trait locus (meQTL) effects. Age and CRP levels are presented as group median (inter-quartile range)

5.5.1 Validation of a pyrosequencing assay for allelic quantification

Prior to quantifying allelic proportions in the patient genomic DNA (gDNA) and complimentary DNA (cDNA), the pyrosequencing assay was validated to confirm the ability to accurately quantify allelic proportions across the range of values (0 – 100%). Similar to the method used for CpG assay validation, a standard curve was generated by preparing mixes containing allele proportions ranging from 0 to 100%. Sequencing these mixtures confirmed the ability of this assay to accurately call allele ratios at the *FCRL3* transcript SNP (rs7522061), with a linear relationship between expected and observed allele ratios (Figure 5.9A).

5.5.2 Allelic expression analysis at FCRL3

Proportions of the risk allele (C) at rs7522061 were quantified in patient gDNA and cDNA by pyrosequencing in the 33 heterozygous patients. In gDNA, the allele proportions were found to be 52% (range = 50 – 53%), consistent with all patients being heterozygous at this SNP (Figure 5.9B). That the allele ratio was in fact slightly above 50% for these patients suggests a tendency for the assay to over-call the risk allele at intermediate values, as is consistent with the results from the validation (Figure 5.9B). Nonetheless, the risk allele was found to be significantly enriched in the mRNA ($p < 1 \times 10^{-4}$; Figure 5.9B). The proportion of risk allele present in the mRNA ranged from 53% to 76% (median = 64%), indicating that in some cases the transcription of *FCRL3* from the risk-associated allele was over three times that of the non-risk allele (76% vs. 24%). These results confirm that the regulatory SNP highlighted in the meQTL and CIT analyses has the capacity to promote increased transcript levels of the *FCRL3* gene in CD4$^+$ T cells.

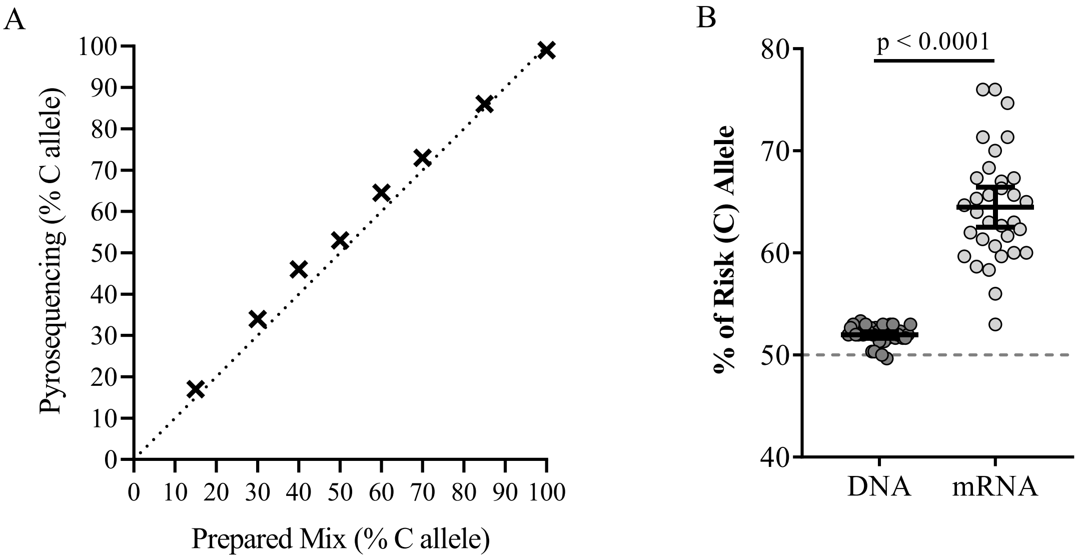


Figure 5.9: Allelic expression analysis at *FCRL3*. A) Standard curve generated by a pyrosequencing assay for allelic expression quantification at rs7522961 (*FCRL3*). Allele mixes were generated by combining various proportions of genomic DNA homozygous for either the C or T allele at rs7522961. Crosses represent the mean value from three PCR repeats, with the x=y dotted line indicating where pyrosequencing values exactly match values in the prepared mix. B) Allelic expression analysis of risk allele proportions at rs7522061 (*FCRL3*) in either the genomic DNA or mRNA (cDNA) in CD4$^+$ T cells isolated from early arthritis patients.

5.5.3 Validation of allelic effects on transcription at additional loci

In the absence of transcript SNPs at the *ANKRD55*/*IL6ST* and *JAZF1* loci, associations between genotype and transcription could be studied. Plotting microarray gene expression values against genotype at the regulatory SNPs for *ANKRD55* (p = 0.0112; Figure 5.10A), *IL6ST* (p = 0.0096; Figure 5.10B), and *JAZF1* (p = 0.0319; Figure 5.10C) across all patients in this validation cohort confirmed such effects at these loci. The association between genotype and expression was inverse to that which occurred between genotype and DNAm, which was consistent with observations from the eQTM analysis that DNAm exhibits an inverse relationship with transcript levels.

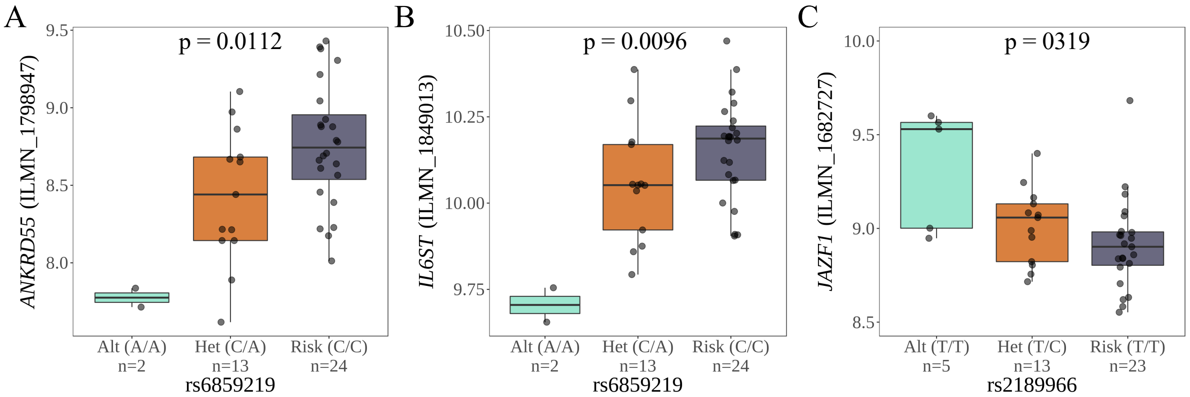


Figure 5.10: Validation of associations between risk genotype and gene expression levels at loci for which allelic expression analyses were not possible. Illumina human HT-12 CD4^{+} T cell gene expression data for an independent cohort of samples that were not included in the initial analysis of DNA methylation and gene expression associations. These associations were validated at (A) rs6859219-*ANKRD55*, (B) rs6859219-*IL6ST*, and (C) rs2189966-*FCRL3*. P-values were generated using a one-way ANOVA. The box plots indicate the group medians with the 1st and 3rd quantiles, and whiskers extend to the largest and smallest values to a limit of 1.5 times the 25th of 75th quartile respectively. P-values were generated using a one-way ANOVA.

5.6 5-Aza-2’-deoxycitidine treatment of lymphocyte cell lines

Given that the *in silico* analyses described a number of genes at which transcription appeared to be regulated downstream of cis-meQTL effects, the transcriptional activity of these genes was assessed following global perturbations in DNAm. To experimentally corroborate that these candidate genes are sensitive to changes in DNAm, two lymphocyte cell lines were treated with 5-Aza-2’-deoxycitidine (5-aza; Decatibine), a cytidine analogue that inhibits activity of DNA methyltransferases (DNMTs), resulting in genome-wide passive CpG hypo-methylation.

Jurkat cells (clone E6-1) - a T cell line from an acute leukaemia patient, and Ramos cells, which are a B cell line from a patient with Burkett’s lymphoma, were both treated with 5-aza. In order to assess the efficacy of varying concentrations of 5-aza, as well as treatment duration, to induce hypo-methylation, the cell lines were treated with either 0.25µM or 0.50µM 5-aza for 48 hours or 72 hours.

5.6.1 Assessment of DNA methylation and gene expression following treatment with 5-Aza-2’-deoxycitidine

Treatment of both Jurkat cells with each concentration of 5-aza (0.25µM, 0.50 µM) for 48 hours resulted in a marked reduction in DNAm at cg17134153 relative to the DMSO control (Figure 5.11A). The DNAm levels at this CpG roughly halved following treatment for 48 hours with

0.25μM 5-aza (43% to 25%), consistent with one round of cellular division (the population doubling time of Jurkat cell clone E6-1 is ~48 hours). Treatment of these cells for 72 hours had little additional effect on DNA de-methylation at cg17134153 beyond what was observed at the 48-hour time point (Figure 5.11A).

In contrast, the levels of DNAm at cg17134153 pre-treatment with 5-aza were considerably lower in Ramos cells (10% DNAm at cg17134153 in Ramos cells; Figure 5.11B). This largely reflected what was observed in primary cells, whereby DNAm levels in primary CD4^{+} T cells ranged from 29-55%, whereas in B cells values were between 2-17% (Figure 5.2D). A smaller magnitude of de-methylation was observed in these cells, with 0.25μM 5-aza treatment for 48 hours reducing methylation levels by 3% at cg17134153 (10% to 7%; Figure 5.11B). Consistent with observations in Jurkat cells, treatment for 72 hours yielded no greater reduction in DNAm levels at this CpG (Figure 5.11B).

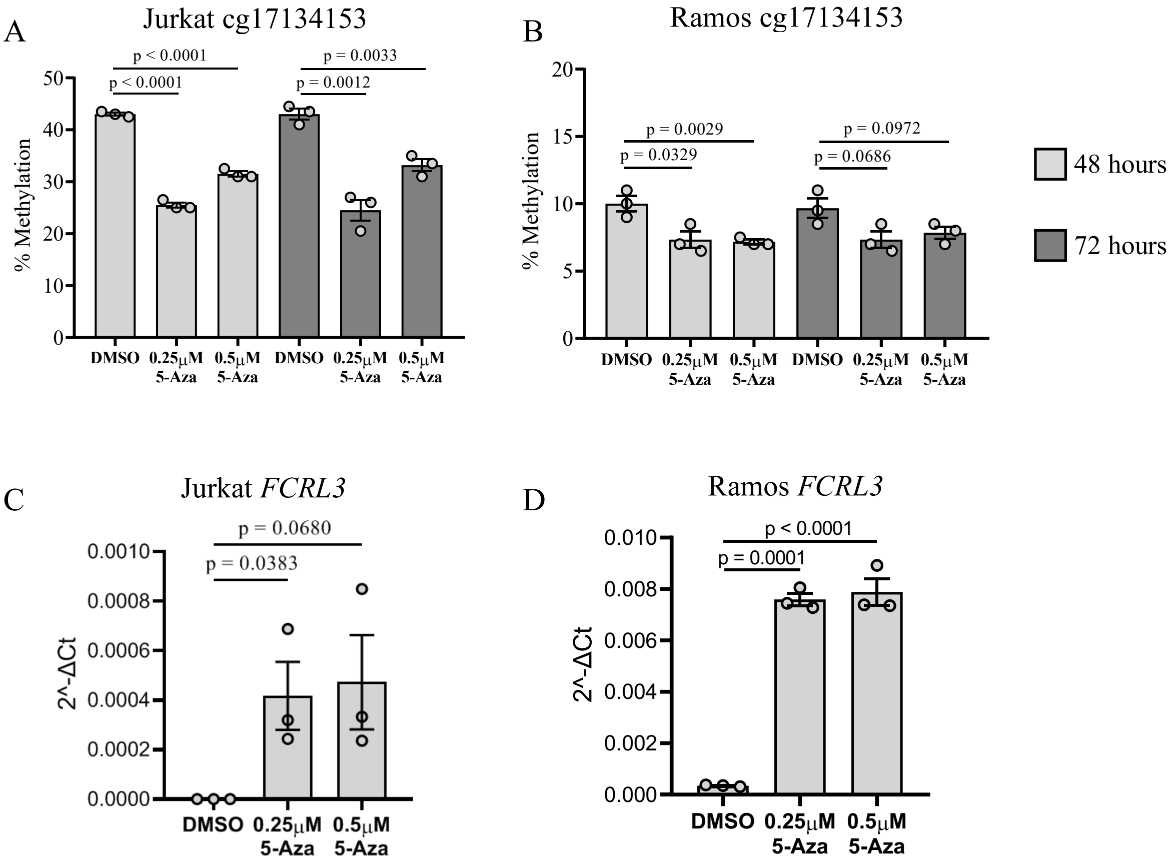


Figure 5.11 DNA methylation at cg17134153 and expression of *FCRL3* in Jurkat and Ramos cells following treatment with 5-Aza-2'-deoxycitidine. After treatment with DMSO (vehicle control) or 5-aza at 0.25μM and 0.50μM for either 48 hours or 72 hours, DNAm was quantified by bisulphite pyrosequencing at cg17134153 in (A) Jurkat cells and (B) Ramos cells. Subsequently, expression of *FCRL3* in (C) Jurkat cells and (D) Ramos cells, treated for 48 hours was measured by qPCR. Bars represent the mean of three biological replicates in each condition, with error bars showing the standard error of the mean. P-values report the difference between each 5-aza treatment condition and the DMSO control using an unpaired t-test.

In Jurkat cells, DNA de-methylation observed at cg17134153 coincided with up-regulation of the *FCRL3* gene in both cell lines. In Jurkat cells, expression of *FCRL3* was not detected in DMSO control cells, with transcripts detected following treatment with both 0.25µM and 0.50µM 5-aza for 48 hours (Figure 5.11C). *FCRL3* expression was detected in untreated Ramos cells, and treatment with 5-aza (0.25µM and 0.50µM) resulted in a 21-fold induction of expression (Figure 5.11D, indicating that *FCRL3* transcription is highly responsive to DNAm in this cell line.

A similar magnitude of de-methylation was observed at cg21124310 in Jurkat cells treated for 48 hours with 0.25µM 5-aza for 48 hours (47% vs. 29% in DMSO controls; Figure 5.12A). Again, treatment for 72 hours resulted in little additional de-methylation upon that which was observed at 48 hours (Figure 5.12A). As was the case for cg17134153, DNAm levels at cg21124310 were considerably lower in untreated Ramos cells (47% in Jurkat vs. 21.5% in Ramos cells; Figure 5.12B). It should be noted that, whilst DNAm at cg21124310 in Jurkat cells (47%) was within the range observed in primary CD4$^+$ T cells (12%-63%), DNAm levels of 21.5% in Ramos cells was considerably lower than was present in primary B cells (81 – 89%). Nonetheless, treatment with 5-aza (0.25µM) for 48 hours did result in a small reduction of 7.5% (21.5% to 14%) at cg21124310 in Ramos cells.

In primary CD4$^+$ T cells, *in silico* analyses suggested that cg21124310 regulated the expression of both *ANKRD55* and *IL6ST*. Here, 5-aza-induced DNA de-methylation coincided with a 12-fold up-regulation of *ANKRD55* expression in Jurkat cells at the 48 hour time point (Figure 5.12C). In contrast, *ANKRD55* expression was absent in Ramos cells, and treatment with 5-aza was unable to induce expression of this gene (Figure 5.12D). As expected, de-methylation at cg21124310 occurred in conjunction with increased levels of the *IL6ST* transcript in Jurkat cells (Figure 5.12E). For this particular gene, treatment with the higher concentration of 5-aza (0.50µM) actually yielded a slightly higher fold induction than the lower concentration (3.5-fold vs. 2.8-fold; Figure 5.12E). Somewhat unexpectedly, treatment with 5-aza inhibited expression of *IL6ST* in Ramos cells (Figure 5.12F), despite DNAm levels being lower in the treated cells.

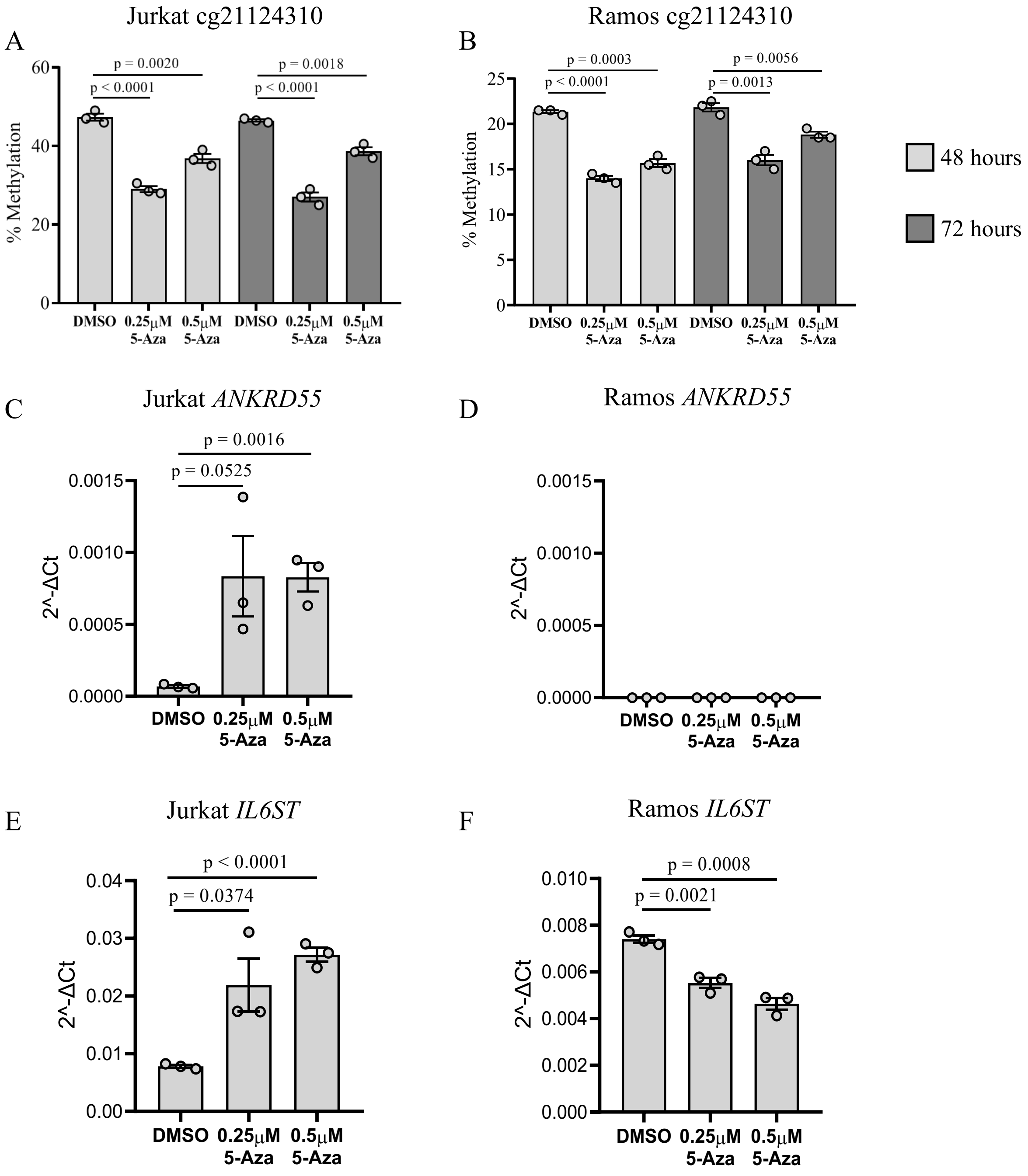


Figure 5.12 DNA methylation at cg21124310 and expression of *ANKRD55* and *IL6ST* in Jurkat and Ramos cells treated with 5-Aza-2'-deoxycitidine. After treatment with DMSO (vehicle control) or 5-aza at 0.25μM and 0.50μM for either 48 hours or 72 hours, DNAm at cg21124310 was quantified by bisulphite pyrosequencing in (A) Jurkat cells and (B) Ramos cells. Subsequently, expression of *ANKRD55* in (C) Jurkat cells and (D) Ramos cells, treated for 48 hours was measured by qPCR. Expression of *IL6ST* was similarly quantified in (E) Jurkat cells and (F) Ramos cells. Bars represent the mean of three biological replicates in each condition, with error bars showing the standard error of the mean. P-values report the difference between each 5-aza treatment condition and the DMSO control using an unpaired t-test.

Finally, 5-aza treatment caused de-methylation at cg07522171 in Jurkat (Figure 5.13A) and Ramos (Figure 5.13B) cells. Contrary to the observations at other CpGs assessed, treatment for 72 hours yielded greater reduction in DNAm than was seen at 48 hours (though this was limited to Jurkat cells). DNAm at cg07522171 in Jurkat cells was highest of the three cis-CpGs, with methylation levels of 81% in untreated cells being considerably higher than those which were seen in primary cells (20%-44%; Figure 5.5A). Patterns of gene expression at this locus largely reflected those which were observed for *IL6ST*, with induction in Jurkats following 5-aza treatment, and a more marked increase at higher 5-aza concentrations (Figure 5.13C). In Ramos cells, *JAZF1* expression followed a pattern analogous to *IL6ST*, with DNA de-methylation unexpectedly resulting in reduced transcript levels (Figure 5.13D).

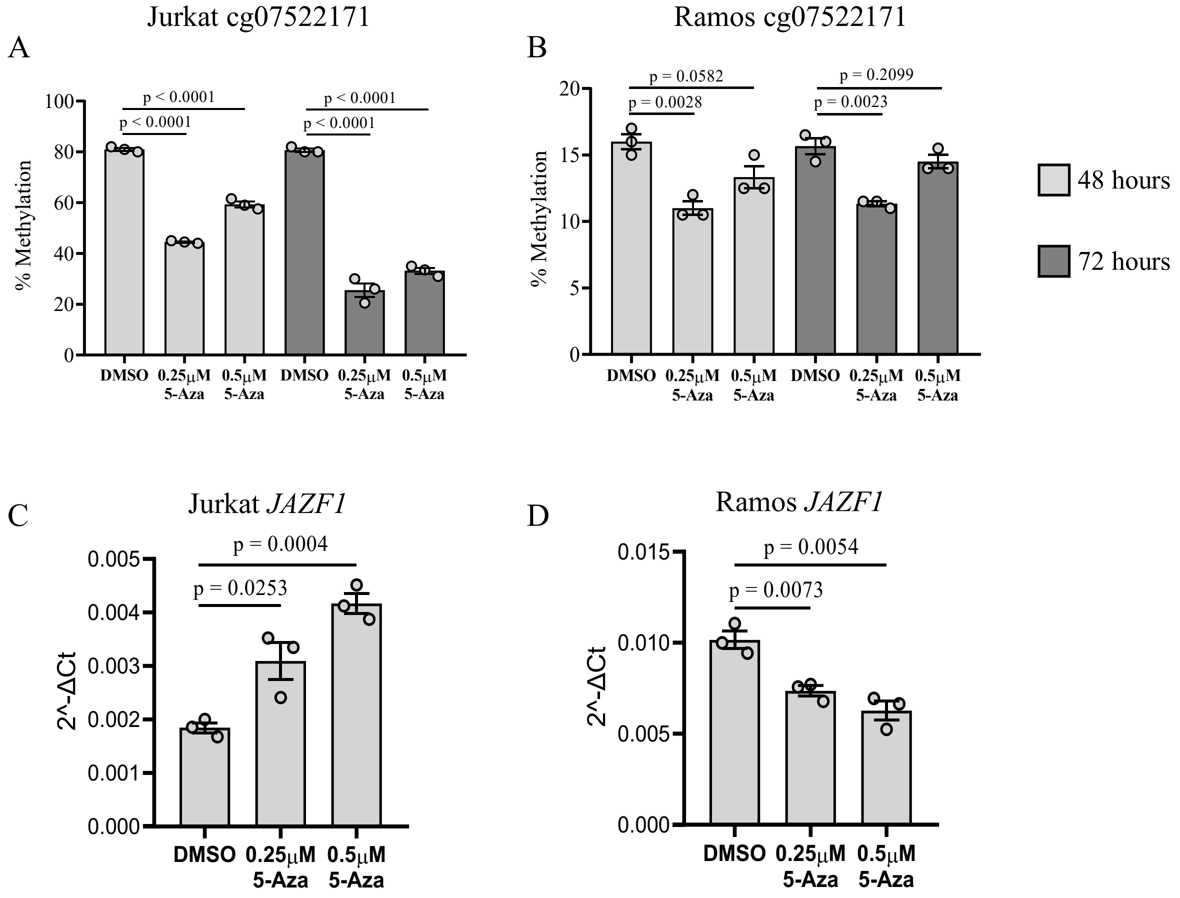


Figure 5.13 DNA methylation at cg07522171 and expression of *JAZF1* in Jurkat and Ramos cells following treatment with 5-Aza-2'-deoxycitidine. After treatment with DMSO (vehicle control) or 5-aza at 0.25µM and 0.50µM for either 48 hours or 72 hours, DNAm was quantified by bisulphite pyrosequencing at cg07522171 in (A) Jurkat cells and (B) Ramos cells. Subsequently, expression of *JAZF1* in (C) Jurkat cells and (D) Ramos cells, treated for 48 hours was measured by qPCR. Bars represent the mean of three biological replicates in each condition, with error bars showing the standard error of the mean. P-values report the difference between each 5-aza treatment condition and the DMSO control using an unpaired t-test.

Indeed, of the CpG-transcript associations investigated following 5-aza treatment, only cg17134153-*FCRL3* exhibited significant associations in primary B cells. To this end, observations in these cell lines reflect that which was observed *in silico*, with *FCRL3* levels up-regulated following DNA de-methylation in T cells and B cells, while associations between DNAm and *ANKRD55/IL6ST/JAZF1* were limited to T cells.

5.7 Validation of DNA methylation-mediated transcriptional regulation by reporter gene assays

The results from the 5-Aza-2'-deoxycitidine treatment of lymphocyte cell lines confirmed that the transcript levels of genes identified at RA risk loci are responsive to changes in DNAm levels. However, given that this treatment induces hypo-methylation genome-wide, these experiments are unable to confirm the region-specific cis-eQTM observations from the patient cohort. In order to validate the influence of such region-specific DNAm on transcription at loci of interest, reporter gene assays were designed. The CpG sites implicated in regulating the expression of *FCRL3* and *JAZF1* in CD4$^+$ T cells map to the promoter regions of these genes, and so luciferase reporter assays were designed to assess the downstream impact of CpG methylation at these sites on transcriptional activity. The lead meQTL SNP (rs2210913) at the *FCRL3* locus is in perfect LD ($r^2 = 1.00$; 1000 Genomes Project Phase 3, European Populations) with a SNP (rs7528684) that has been functionally validated as an regulatory variant regulating transcription of the gene[346]. The risk allele at rs2210913 (T) correlates with the C allele at rs7528684. Importantly, this variant is 488 bases upstream of cg17134153, enabling both the SNP and CpG site to be cloned into a vector, and as such the combinatorial effects of genotype and DNAm can be assessed in conjunction. In the case of the *JAZF1* promoter, the SNP (rs2189966) highlighted in the meQTL analysis is 46,870 bases from the CpG site (cg07522171), and as such the effects of DNAm could only be assessed in isolation.

5.7.1 Amplification of promoter regions

Primers were designed to clone both the promoter region of *FCRL3* (chr1:157,670,120 – 157,671,093) harbouring rs7528684 together with two CpG sites implicated in transcriptional regulation (cg17134153, cg01045635), as well as that of the *JAZF1* promoter (chr7:28,218,982 – 28,218,384) with cg07522171 and no polymorphic positions. The cloned *FCRL3* promoter region also encompassed eight additional CpGs (Figure 5.14A), whereas the *JAZF1* cloned promoter included 14 additional CpGs (Figure 5.14B). These regions were amplified from patient genomic DNA. Given that the allelic effect of rs7528684 was to be quantified, template DNA from individuals who were heterozygous at this SNP (as well as rs945635 which is in high LD ($r^2 > 0.9$) with rs7528684) was selected based on data from genotyping arrays, enabling inserts harbouring each allele to be amplified.

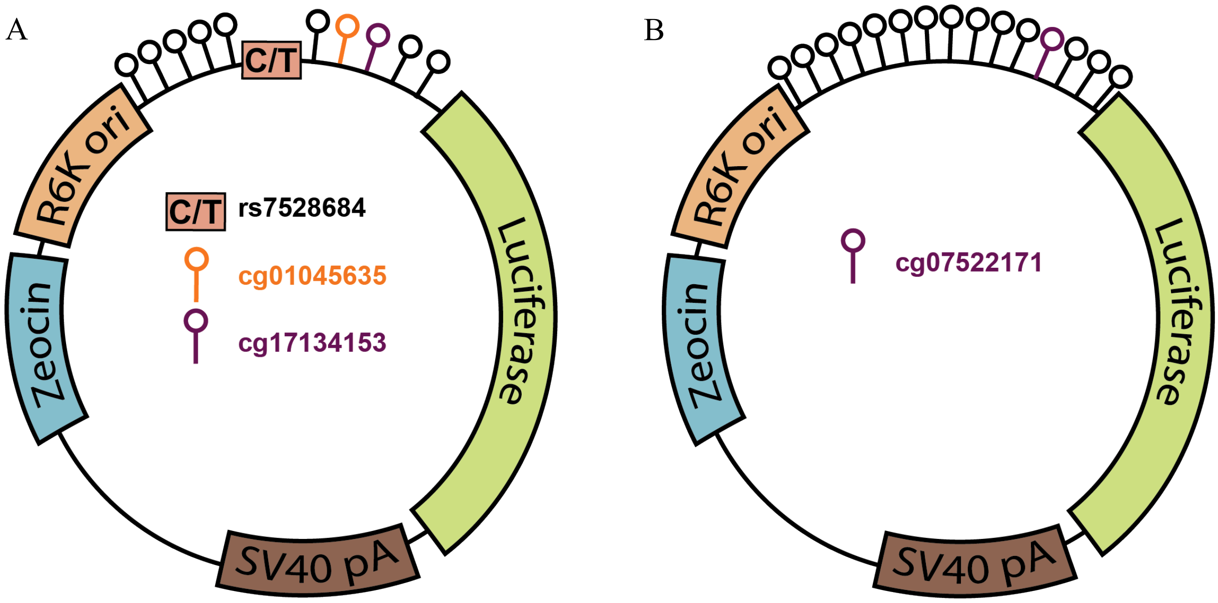


Figure 5.14: pCpGL-basic plasmids harbouring constructs for reporter gene assays. The (A) *FCRL3* and (B) *JAZF1* promoter regions encompassing CpGs implicated in causal inference testing were cloned into the reporter gene vector. Positions of CpGs in the insert relative to one another are accurate, though distances between CpGs are not to scale. Luciferase = luciferase reporter gene giving read-out of transcriptional activity; SV40 pA = Simian virus 40 poly-A transcriptional terminator; Zeocin = Zeocin antibiotic resistance gene; R6K ori = *E. coli* R6K origin of replication.

5.7.2 Genotyping and sequencing of clones

Amplicons were digested using the appropriate restriction enzymes (NocI/SpeI for *FCRL3*, HindIII/SpeI for *JAZF1*) and cloned into the pCpGL-basic vector – a CpG-free vector containing a luciferase reporter gene. After transformation of bacterial cells and selection on antibiotic-supplemented agar plates, colonies were picked and plasmids miniprepped for genotyping. To identify clones transfected with constructs consisting of the insert ligated into the plasmid, miniprepped plasmids were restriction enzyme-digested and electrophoresed through an agarose gel. For the *FCRL3* clones, F9, F10, F11, F12, F13, and F14 were selected for sequencing based on the presence of a band representing the promoter region (Figure 5.15A), whereas for the *JAZF1* insert, clones J1, J2, J3, and J4 were selected on the same basis (Figure 5.15B).

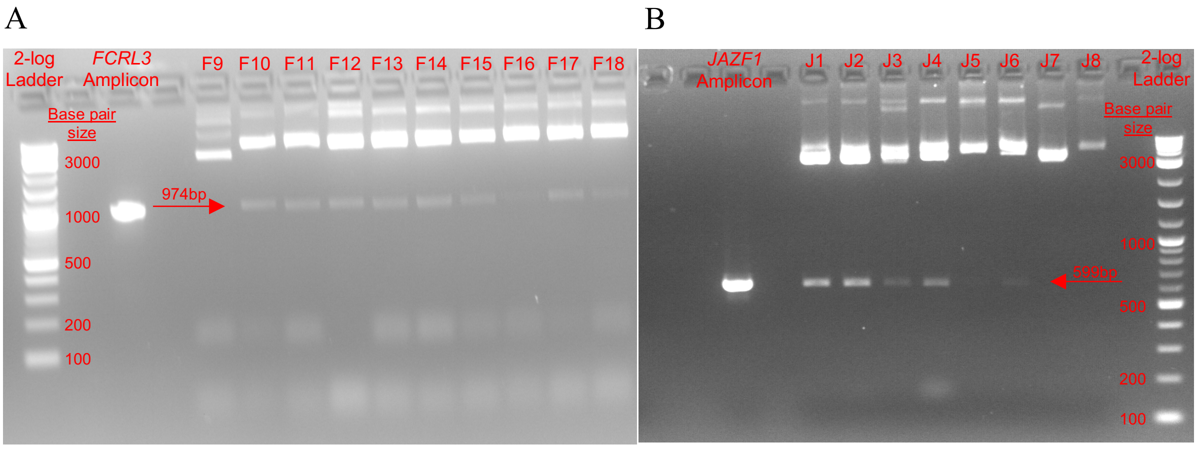


Figure 5.15: Genotype digest of plasmids containing (A) *FCRL3* or (B) *JAZF1* constructs. To identify clones successfully transfected with constructs of interest, isolated plasmids were restriction enzyme-digested and run on an agarose gel. Digested plasmids were run alongside the PCR amplicon for the respective promoter region to confirm the insert size. Plasmids for which inserts were present were selected and sent for Sanger sequencing. Red arrows on each plot indicate the expected size of the digested insert for each construct.

To confirm the correct orientation and sequence of the cloned insert, samples were sent for Sanger sequencing (Source Bioscience). This method is the gold standard for accurate DNA sequencing of short regions (typically up to ~800 bases) and is based on the incorporation of labelled dideoxynucleotides into the DNA chain being synthesized by DNA polymerase. The incorporation of fluorescently-labelled dideoxynucleotides into the DNA strand causes chain termination. As such, performing four reactions in parallel, each with a different labelled dideoxynucleotide (A, C, G, T) enables the nucleotide at each position to be determined. These reactions can generate trace plots as in the example shown in Figure 5.16 for rs7528684 in the *FCRL3* promoter (T allele present in clone F10, C allele in F12), together with the nucleotide sequence. Sequences from each set of clones (*FCRL3* or *JAZF1*) were aligned to identify polymorphic sites, either representing SNPs in the template DNA or mutations arising during amplification or cloning. The *JAZF1* promoter insert contained no SNPs in the template, and all inserts were confirmed to be of the correct orientation and identical to the reference genome sequence. Of the *FCRL3* inserts, the F12 clone had C alleles present at both rs7528684 and rs945635, and all other clones harboured a T and G allele at these positions respectively (Table 5.6). The remaining insert sequence was identical across all clones, including at another SNP (rs11264799) present in this region, as expected based given homozygous template DNA at this position. Clone J1 was therefore selected to be taken forward for reporter gene assays, as were clones F10 and F12 representing both allele copies at the regulatory SNP.

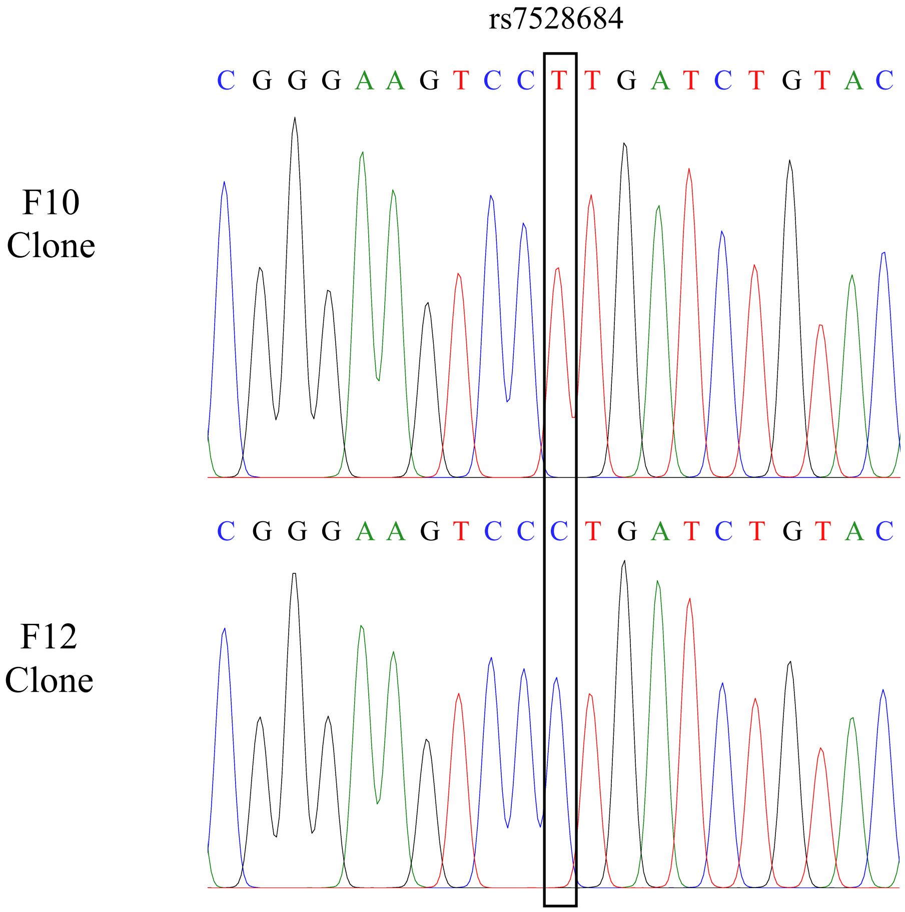


Figure 5.16: Trace plots from Sanger sequencing of constructs harbouring the *FCRL3* promoter insert. The position of rs7528684 (regulatory SNP) is highlighted indicating the presence of differing allele copies at this position in the F10 and F12 clones.

Clone	**rs7528684**	rs11264799	**rs945635**
F10	**T**	G	G
F11	T	G	G
F12	**<u>C</u>**	G	C
F13	T	G	G
F14	T	G	G
F17	T	G	G

Table 5.6: Haplotypes present at variable positions identified in all *FCRL3* clones that were Sanger Sequenced. The promoter region harbours three SNPs (rs7528684, rs11264799, and rs945635), which displayed the same haplotype across all clones with the exception of clone F12, for which the risk allele (C) was present at the regulatory SNP (rs7528684), with a C allele also present at rs945635 which is in perfect linkage disequilibrium with rs7528684. The risk allele (C) at rs7528684 is underlined.

5.7.3 In vitro DNA methylation

The desired clones, together with the empty pCpGL vector, were either methylated *in vitro* using the M.SssI CpG DNA methyltransferase enzyme or mock-methylated (H_20), so that the effect of CpG methylation on transcriptional activity could be assessed. To check whether the plasmids were efficiently methylated, they were digested post-methylation using the HpaII restriction enzyme. This enzyme recognises the sequence CCGG but is methylation sensitive and as such will not digest sequences where a methyl group is present on the cytosine. The empty pCpGL vector has no HpaII recognition sites, and so this plasmid presents in its circular and supercoiled form when run on an agarose gel, regardless of whether methylation or mock-methylation treatment has occurred (Figure 5.17; first pair of lanes).

The introduction of HpaII recognition sites from the insert results in distinct banding patterns for each plasmid treatment. Introduction of HpaII sites in the insert enables digestion of the unmethylated plasmids (Figure 5.17; Cntl lanes). In the case of *JAZF1*, the insert harbours two HpaII sites 345bp apart, resulting in a small band observed in addition to the remainder of the digested plasmid and insert (Figure 5.17, *JAZF1* control lane). For the *FCRL3* inserts, two HpaII sites are also introduced, though they are < 100bp apart and as such the digested plasmids present as the single digested plasmid (Figure 5.17; *FCRL3* (CGC) and *FCRL3* (TGG) control lanes). Conversely, the addition of a methyl group at the HpaII restriction site by M.SssI enzyme

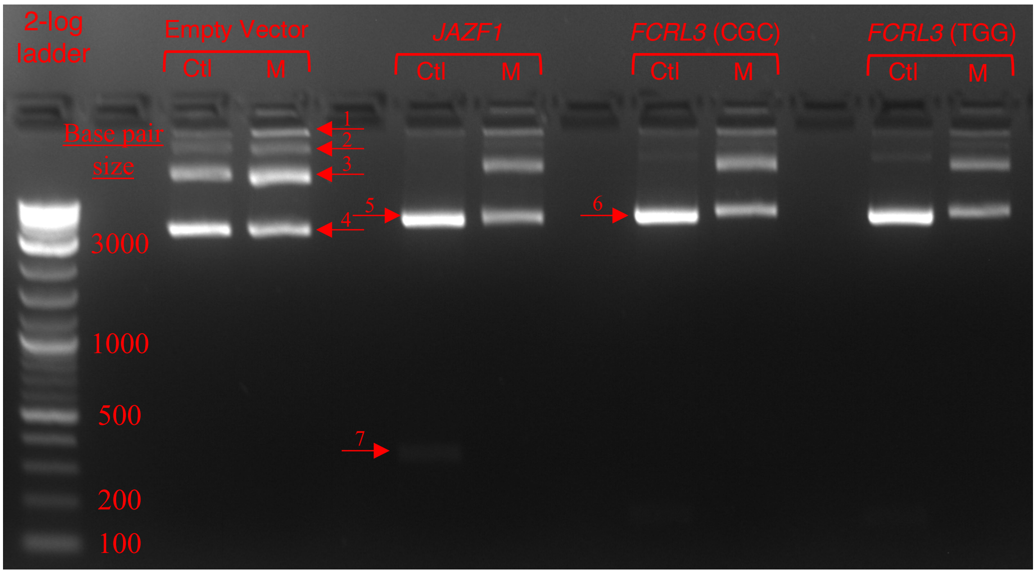


Figure 5.17: Methylation-sensitive restriction enzyme digest with HpaII of *in vitro* methylated (M) and mock-methylated control (Ctl) constructs to confirm DNA methylation. This was done for the empty pCpGL vector, the construct containing the *JAZF1* promoter insert, and the *FCRL3* construct representing both haplotypes (risk – CGC; non-risk - TGG). Red arrows in the lane containing the empty vector (M) represent (1) nicked, (2) linear, (3) supercoiled, and (4) circular single-stranded plasmid DNA. Additional red arrows indicate digested plasmid from *JAZF1* (5) and *FCRL3* (6) clones, as well as the smaller *JAZF1* fragment at 345bp (7).

inhibits the endonuclease activity of this enzyme, preventing plasmid digestion and resulting in the same banding pattern as is seen for the empty plasmid (Figure 5.17; right lane (M) in each pair). This analysis confirmed that M.SssI-treated plasmids containing insert were successfully methylated *in vitro*.

5.7.4 Luciferase reporter assay in HEK-293T cells

Following the promoter regions being cloned into the CpG-free vector and *in vitro* DNA methylation being performed, cell lines were transfected to allow luciferase readings to be taken. Initially, constructs were co-transfected with the control pRL-TK renilla control plasmids into HEK-293T cells, an embryonic kidney cell line, given their relative ease of transfection. Twenty-four hours following transfection, cells were lysed and luciferase activity (Firefly/Renilla) measured.

For the FCRL3 promoter, luciferase reporter activity was not significantly influenced by either genotype nor DNA methylation status (Figure 5.18A). Similarly, DNA methylation had no effect on *JAZF1* promoter activity within this cell type (Figure 5.18B). Indeed, the normalised luciferase activity for both *FCRL3* and *JAZF1* plasmids was comparable to that for the empty vector (RLA normalised to 1), suggesting that these promoters may not be active in this cell type.

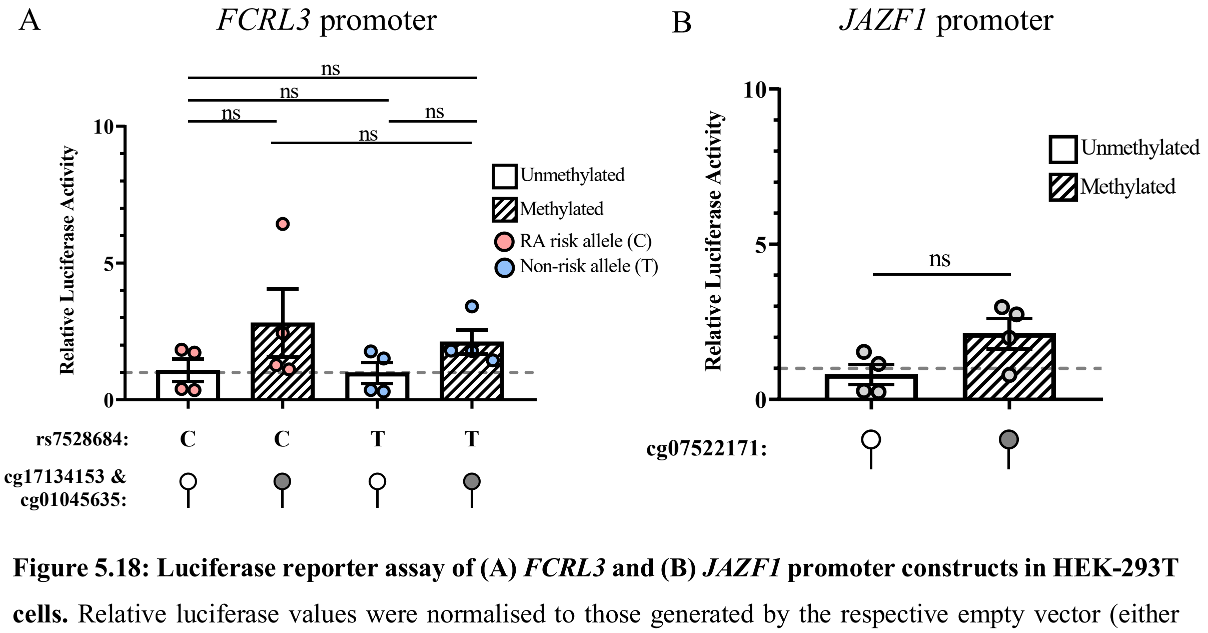


Figure 5.18: Luciferase reporter assay of (A) *FCRL3* and (B) *JAZF1* promoter constructs in HEK-293T cells. Relative luciferase values were normalised to those generated by the respective empty vector (either unmethylated or methylated), which was given a value of 1. Reported p-values were generated by one-way ANOVA with Tukey's multiple comparisons test (A) and a two-tailed t-test (B). Four experiments were performed in total, with a total of 6 wells per condition run in each experiment. Bars represent the mean of the four experiments, and error bars denote the standard error of the mean.

5.7.5 Luciferase reporter assay in Jurkat cells

Following the observations in HEK-293T cells, it was decided to transfect Jurkat cells with plasmids, given that this lymphocyte cell line more closely resembles the CD4$^+$ T cells in which the initial regulatory mechanisms were identified. Jurkat cells were transfected by electroporation using the Neon™ system (ThermoFisher) and, as with HEK-293T cells, lysed 24 hours following transfection with luciferase readings quantified as before.

In this cell line, the relative luciferase activity (RLA) of the plasmid containing the risk allele (C) at rs7528684 was 28.2-fold higher than the empty vector ($p < 0.0001$, one-way with Tukey's multiple comparisons test), confirming that the promoter was active in this cell type (Figure 5.19A). Interestingly, this activity was ~2.3-fold higher the non-risk allele plasmid (Figure 5.19A; RLA 28.2 vs. 12.9; $p < 0.0001$). This is consistent with the observations from allelic analysis of heterozygous patients (Figure 5.9B), whereby risk allele mRNA was on average ~1.8-fold higher than the non-risk allele (64% risk allele at transcript SNP in mRNA), and up to 3.2-fold increased (76% risk allele) in some individuals. Importantly, when plasmids were methylated prior to transfection, the luciferase was only marginally increased relative to the empty vector (RLA ~2.9 for both alleles), confirming that methylation of the *FCRL3* promoter region ablates transcription (Figure 5.19A). Indeed, this ablation was consistent regardless of the allele copy present in the plasmid, validating the mechanism identified *in silico* whereby transcriptional regulation by DNAm modification occurs downstream of the genetic variant.

In contrast to the results observed for *FCRL3* constructs, the luciferase activity in Jurkat cells transfected with the *JAZF1* constructs was comparable to the empty vector (Figure 5.19B; RLA ~0.9 (Unmethylated) vs. 1.2 (Methylated)). This may indicate that the minimal promoter region required for transcription to occur at this locus was not sufficiently captured in the amplified region.

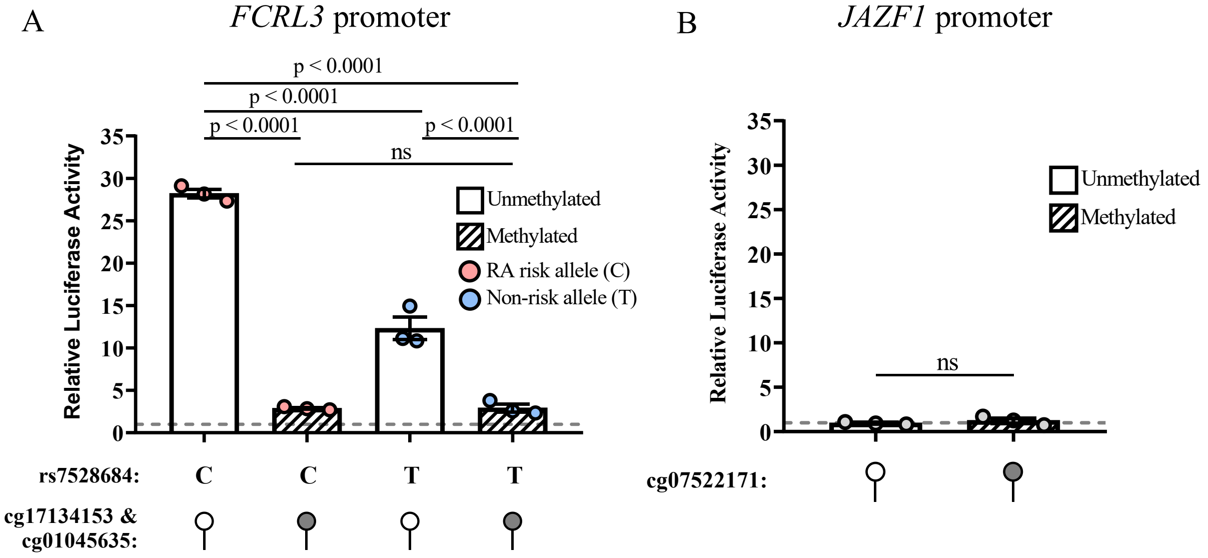


Figure 5.19: Luciferase reporter assay of (A) *FCRL3* and (B) *JAZF1* promoter constructs in Jurkat cells. Relative luciferase values were normalised to those generated by the respective empty vector (either unmethylated or methylated), which was given a value of 1. Reported p-values were generated by one-way ANOVA with Tukey's multiple comparisons (A) and a two-tailed t-test (B). Three separate transfections were run for each condition, and bar represent the mean of these three transfections, with error bars denoting the standard error of the mean.

5.8 Targeted DNA de-methylation using a CRISPR – nuclease-deficient Cas9 system

Targeted epigenetic modification with a CRISPR – nuclease deficient Cas9 (dCas9) system was employed in attempt to induce site-specific DNA de-methylation, which would give insights into the regulatory function of CpG methylation in a genomic context. This particular system consists of dCas9 fused to the catalytic domain of the TET1 demethylating enzyme. This is achieved by the recognition of multiple GCN4 peptide copies on the dCas9 by anti-GCN4 (the pPlatTET-gRNA2 plasmid, see section 2.16)[304]. By targeting this complex to regions harbouring cis-CpG sites using sequence-specific guide RNAs (gRNAs), site-specific DNA de-methylation can potentially be achieved.

5.8.1 Flow cytometry sorting of co-transfected cells

The use of a fluorescently-labelled (ATT0550) gRNAs, together with the presence of a GFP tag expressed by the pPlatTET-gRNA2 plasmid, allowed for co-transfected cells to be sorted by flow cytometry. To set gates for sorting based on ATT0550 and GFP, a fluorescence minus one (FMO) approach was employed; single transfections were set up whereby Jurkat cells were electroporated either with gRNA (ATTO550; Figure 5.20A) alone, or pPlatTET-gRNA2 plasmid (GFP) alone (Figure 5.20B). Twenty-four hours following the transfection of Jurkat cells with pPlatTET-gRNA2 and gRNAs targeting either the cg17134153 (*FCRL3*; Figure

5.20C), cg21124310 (*ANKRD55/IL6ST*; Figure 5.20D), cg07522171 (*JAZF1*; Figure 5.20E) regions, or the non-targeting negative control (Figure 5.20F), double-positive cells were sorted based on the gates set from the single transfections. The recovery of double-positive cell was low (< 4%), largely owing to the fact that the vast majority of cells were negative for GFP.

Double-positive cell were sorted, then split into three separate cultures and incubated for a further 48 hours. Following this incubation, and prior to harvesting and lysis of cells for DNAm analysis, cells were visualised using a fluorescent microscope (Olympus CKX53) to qualitatively assess the expression of GFP by transfected cells. GFP fluorescence of transfected cells was found to be highly variable 48 hours following transfection, as is shown for the cg17134153 gRNA1 (Figure 5.21A) and negative control gRNA (Figure 5.21B) conditions.

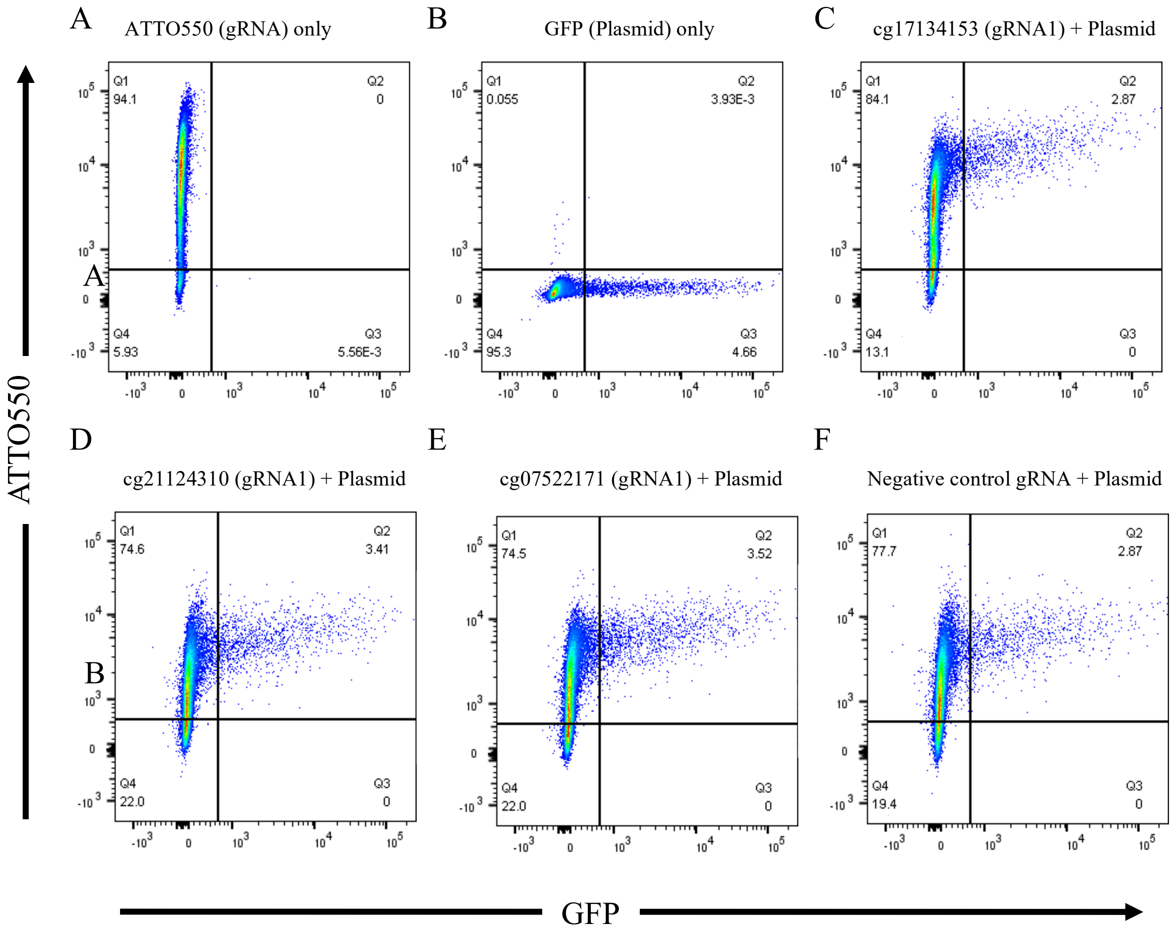


Figure 5.20: Flow sorting of cells co-transfected with the pPlatTET-gRNA2 plasmid and gRNAs. Cells transfected with pPlatTET-gRNA2 plasmid were sorted based on fluorescence of the GFP tag, and those transfected with gRNAs targeting regions harbouring CpGs of interest were sorted based on ATTO550. Gates for sorting GFP and ATTO550 positive cells were determined using a fluorescence minus one (FMO) approach, by performing single transfections with either (A) ATTO550 alone or (B) GFP alone. These gates were then used to sort double-positive cells transfected with plasmid and gRNA targeting either (C) cg17134153, (D) cg21124310, (E) cg07522171, or (F) a non-targeting negative control gRNA.

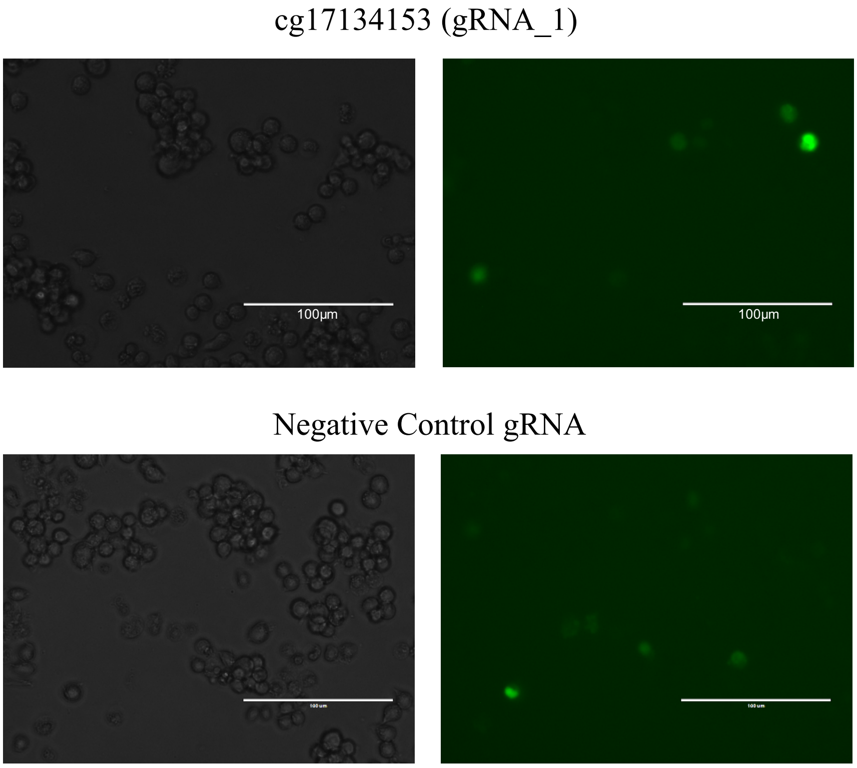


Figure 5.21: Representative microscopy images of co-transfected cells. Bright-field (left panels) and green fluorescence protein (GFP; right panels) images at 48 hours following co-transfection with pPlatTET-gRNA2 and targeting gRNA for (A) cg17134153 (gRNA_1) and (B) negative control.

5.8.2 DNA methylation quantification of co-transfected cells

DNAm was quantified by pyrosequencing 48 hours following sorting of double positive cells. However, methylation at the CpGs of interest was not significantly reduced in the cells transfected with the respective gRNA (Figure 5.22). There was in fact a marginal increase relative to the negative control observed in cells transfected with the gRNA targeting cg17134153 (Figure 5.22A; gRNA_1). At cg21124310, no significant difference was observed between targeting gRNAs and the negative control (Figure 5.22B), as was also the case for cg07522171 (Figure 5.22C). These data would suggest that either the gRNAs are not able to target the regions of interest, or that the system requires further optimisation.

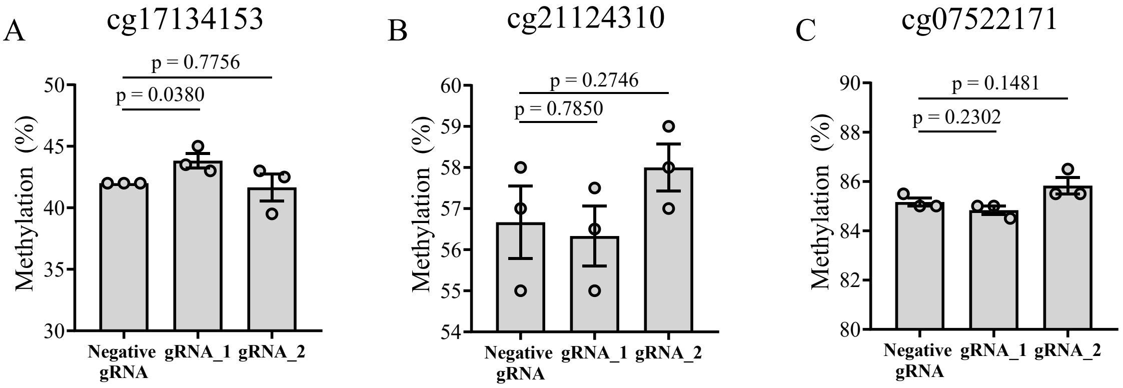


Figure 5.22: DNA methylation quantification in Jurkat cells co-transfected with the pPlatTET-gRNA2 plasmid and synthetic guide RNAs. DNA methylation was quantified by pyrosequencing following transfection with guide RNAs targeting either (A) cg17134153 (*FCRL3*), (B) cg21124310 (*ANKRD55*/*IL6ST*), or (C) cg07522171 (*JAZF1*).

5.9 Discussion

The results described in this Chapter build on the observations in Chapter 4 by demonstrating that DNAm at CpG sites associated with RA risk variants in cis has the capacity to influence transcriptional regulation. In doing so, a number of candidate genes are revealed to be mediated by local CpG methylation at these loci, primarily in CD4$^+$ T cells. This represents the first integrated analysis of genetic risk, DNAm, and gene expression in isolated lymphocytes from a relevant cohort of patients, and as such furthers our understanding of how non-coding risk polymorphisms manifest at the level of cellular phenotype. As well as characterising epigenetic regulatory mechanisms through which RA genetic susceptibility influences pathogenic CD4$^+$ T cell phenotypes, these findings highlight potentially targetable pathways that are relevant to immune dysregulation more generally.

5.9.1 DNA methylation provides mechanistic insight into complex genetic disease mechanisms

The analysis pipeline employed in Chapters 4 & 5 primarily evaluated the role of DNAm in RA pathogenesis, and complex disease more generally, in the context of previously-reported genetic risk loci. Alternative approaches have also been described in which disease-associated methylation changes are first defined based on differential DNAm between patients and controls, irrespective of genetic background. Applying this approach in peripheral blood mononuclear cells, with subsequent eQTM mapping at disease-associated CpGs and causal inference testing (CIT; treating RA diagnosis (case/control) as the outcome phenotype), has led to the discovery of putative novel candidate genes such as *PARP9*[305]. Nonetheless, the advantage of defining disease-associated DNAm modifications based on their association with

DNA variants is a reduction in the levels at which reverse causation can occur. Using patient genotype to infer mediation as opposed to DNAm, which is dynamic with regards to development and tissue context, allows for the associations to be 'anchored' (i.e. all modifications are known to occur downstream of genotypic variation).

Here, CIT was employed to determine the probability of a model under which DNAm acts as a mediator of genetic effects on gene expression. This method requires that all measured variables (genotype, DNAm, gene expression) are collected from the same individual, with a series of conditional correlation tests used to infer causality between molecular measures[260]. One drawback of using the CIT approach to infer causality is that the method is susceptible to measurement error, which could potentially lead to the incorrect model being derived[347]. For this reason, the results described here may represent an underestimate of the true extent of SME regulatory models at RA risk loci. This method is also susceptible to pleitropic effects, whereby genetic variation influences multiple traits which are highly correlated, such as DNAm and chromatin accessibility.

Mendelian randomisation (MR) is an approach that further capitalizes on the static nature of the genome by leveraging genetic variants as instrumental variables to infer causality between a modifiable exposure (i.e. DNAm) and an outcome measure (gene expression or disease)[348]. One advantage of MR is that, unlike CIT, it does not require that all traits are measured in the same individual, and as such can combine large datasets across multiple studies. Indeed, a recent study integrating genotypic variation, DNAm, and gene expression to infer causality in molecular traits associated with skeletal muscle demonstrated that MR and CIT represent complimentary approaches[259].

Ultimately, however, causality in molecular traits that are determined based on cross-sectional patient data will require validation *in vitro*. One such limitation of quantitative trait locus mapping is that this approach is unable to localise the precise causal variant reponsible for the observed effects, and instead nominates risk haplotypes across LD blocks. Following the mapping of cis-meQTLs, the approach taken here was to select the variant exhibiting the strongest association with DNAm levels for subsequent analyses. However, this is unable to conclude causality, particularly in highly polymporphic regions with extensive LD between variants. In addition to this, it has been suggested that susceptibility loci for complex autoimmune trait such as RA may not be attrituable to single variants, but rather multiple variants in LD that map to distinct enhancers and influence gene expression in relevant pathophysiological cell types[349]. The use of statistical approaches to fine-map disease associations, with follow-up functional studies of differential protein binding to risk alleles and enhancer activity in Jurkat cells, has been performed in an attempt to pintpoint causal variants at RA risk loci[350]. This approach was able to provide experimental evidence for functional RA-associated variants at *CD28-CTLA4* and *TNFAIP3*[350]. Further work to map credible causal variants will be necessary to prioritise SNPs for validation studies.

In addition to informing mechanistic understanding of disease genetics, idenfitication of meQTLs provides a complimentary approach to mapping disease heritability beyond considering eQTLs in isolation. As with eQTLs, meQTLs are enriched at disease-associated loci, and mapping molecular QTLs associated with traits such as DNAm and histone modifications provides unique information regarding the disease heritability attributed to particular SNPs[351].

5.9.2 FCRL3 – a shared mechanism of lymphocyte-mediated genetic risk

Only one instance of DNAm mediation at RA risk loci was highlighted in B cells, with this effect mapping to chromosome 1q23.1 associated with the expression of *FCRL3*. This particular observation was compelling given that this locus was also identified as putatively regulating expression of this candidate gene in CD4$^+$ T cells by the same mechanism, with one CpG site (cg01045636) identified as a potential mediator in both cell types.

The mechanism explaining observed associations at the *FCRL3* promoter was also confirmed here experimentally. This was informed by prior work which had localised a functional variant (rs7528684; -169 C → T)[346], which was likely tagging the cis-meQTL regulatory SNP identified in this analysis. As well as confirming associations between this variant and RA susceptibility, this same study also revealed this variant to be a risk factor in autoimmune thyroid disease (Grave's disease & Hashimoto's thyroiditis) as well as systemic lupus erythematosus[346].

Reporter gene assays were able to confirm the allelic effect on transcription, as well as the downstream function of DNAm at cg17134153 and cg01045635 in mediating this effect in CD4$^+$ T cells. It should be noted, however, that given the presence of additional CpGs in the cloned promoter, the regulatory effect cannot be attributed exclusively to these two CpGs. Nonetheless, the functional SNP at this locus maps to a binding site of the NFκB binding site (Figure 5.23; highlighted in the purple box) that has been experimentally determined by ChIP-seq[209, 210]. The first study to describe the rs7528684 functional variant in the *FCRL3* promoter also demonstrated that the risk allele (C) confers increased binding of the NFκB transcription factor[346]. The two CpG sites (cg17134153 & cg01045635) causally implicated in CD4$^+$ T cells at this locus, and confirmed by reporter assays, map to a binding site of RNA polymerase II

(Figure 5.23; highlighted in green boxes). A mechanism may thereby be proposed in which increased NFκB binding at the risk allele inhibts methylation of CpGs at the RNA polymerase II binding site, resulting in the promotion of *FCRL3* transcription. The data presented in this chapter therefore builds on findings from previous work to provide additional mechanistic insight at this autoimmune locus.

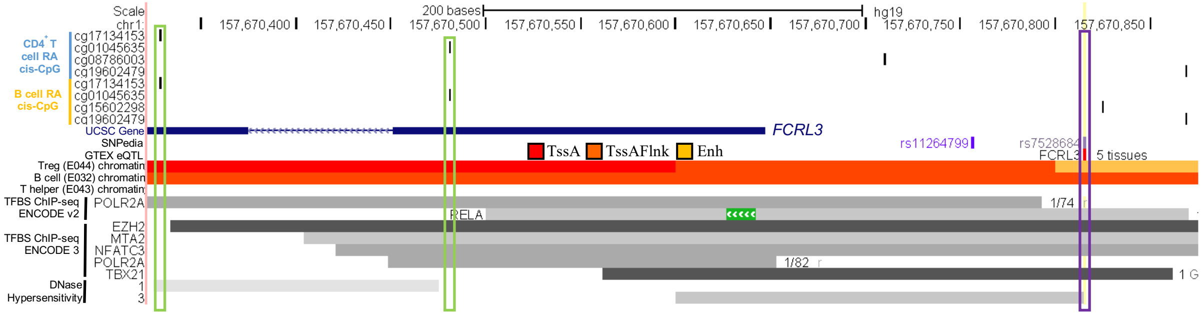


Figure 5.23: DNA methylation and transcription factor binding at the *FCRL3* promoter in CD4⁺ T cells and B cells. Expression of *FCRL3* in both cell types is regulated by an RA risk variant (rs7528684, purple box) at which the risk allele (C) promotes transcription through increased binding affinity of the NFκB transcription factor. Consistent with this, rs7528684 maps to a binding site of RELA (NFκB p65 subunit). This variant appears to influence expression of *FCRL3* in part through modified DNA methylation at CpGs in the promoter region. In CD4⁺ T cells, this occurs via methylation at cg17134153 and cg01045636 (both highlighted in the green boxes), that map to the RNA polymerase II binding site. In B cells, cg01045635 and cg19602479 were implicated as potential regulatory cis-CpGs. Interestingly, the functional variant maps to a regulatory T cell (Treg) enhancer and a B cell active transcription start site flanking (TssAFlnk) region based on chromatin state data, whilst the cis-CpGs map to an active transcription start site (TssA) and TssAFlnk in these cell types respectively. In primary T helper (CD4⁺) cells, this entire region is quiescent, suggesting that the observed regulatory effects in CD4⁺ T cells may be restricted to the Treg subpopulation.

FCRL3 itself encodes a transmembrane glycoprotein that is the third member of the Fc receptor-like (or Fc receptor homologue) family of proteins, each of which are structurally related to the antibody-recognising Fc receptors[352]. To date, eight members of this family have been described (*FCRL1-6*, *FCRLA*, *FCRLB*) and whilst the ligand recognised by the FcRL3 protein remains unknown, the highest levels of expression are seen on B cells[353]. *FCRL3* (as well as *FCRL2* & *FCRL5*) harbours both immunoreceptor tyrosine-based activation motifs (ITAMs) and immunoreceptor tyrosine-based activation motifs (ITIMs) in the intracellular domain, which signal to activate and inhibit immune signaling cascades respectively[353]. This suggests that these receptors may play a role in fine-tuning of adaptive immune signaling.

In the peripheral blood, FcRL3 is expressed by B cells, natural killer (NK) cells, and a population of naturally occurring regulatory T cells (nTregs)[354]. This is in agreement with observations that the regulatory region to which rs7528684 and the cis-CpGs are located is within active chromatin (TSS, flanking TSS, and enhancer) in B cells and Tregs, whereas this region is quiescent in total CD4$^+$ T cells (Figure 5.23). Therefore, whilst our analysis has been performed on total CD4$^+$ T cells, the results at this particular locus actually reflect a mechanism that applies to a subset of these cells. That FcRL3 is not expressed by any other CD4$^+$ T cell subsets likely explains why the observed associations have not been diluted in the larger population of cells. Interestingly, nTregs expressing FcRL3 do not proliferate in response to the T cell mitogen IL-2, suggesting a potential functional defect[354]. Subsequent work has illustrated that this subset of regulatory T cells are reduced in their capacity to attenuate effector T cell responses[354]. In a disease context, increased expression of FcRL3 on Tregs (as well as CD8$^+$ and γδ T cells) in patients harbouring the risk allele at rs7528684 not only correlated positively with DAS28 disease activity, with levels also slightly increased in patients with erosive RA compared with non-erosive disease[355].

In B cells, FcRL3 functions to inhibit signalling at the B cell receptor (BCR), as such preventing activation-induced cell death that would ordinarily occur as a result of BCR stimulation[356]. The authors of this study hypothesised that by increasing the threshold for B cell activation at the BCR, FcRL3 contributes to autoimmunity through a reduction in tolerogenic anergy and deletion of autoreactive B cells. Alternatively, a role for FcRL3 in B cell activation and proliferation in the presence of toll-like receptor 9 (TLR9) co-stimulus has been described[357].

Genetic effects on cell-mediated pathogenesis in RA at this locus may therefore extend beyond a single cell type. The precise function of this gene in lymphocyte biology awaits further clarification, though limited studies to date highlight possible roles in dysregulated B cell and regulatory T cell responses.

5.9.3 Genetic pleitropy highlights additional complexity in disease associations

At the RA risk locus mapping to chromosome 5q11.2, potential pleiotropic regulation of *ANKRD55* and *IL6ST* occurred downstream of cis-meQTLs at four CpG sites. As well as highlighting a long range CpG-promoter interaction in transcriptional regulation, this demonstrates the limitations in attributing genetic risk at GWAS variants to the nearest gene. In addition, the finding at this locus indicates that disease-associated variants may confer risk by altering expression of multiple genes. The pleiotropic effect at chromosome 5q11.2 has been described previously, with the RA risk association and CD4$^+$ T cell eQTL for *ANKRD55* and *IL6ST* determined to have high likelihood of a shared causal variant[342].

This particular locus is notable for its association with both autoantibody-seronegative and autoantibody-seropositive disease[104]. Furthermore, the lead regulatory SNP identified here in the meQTL analysis (rs6859219) has also been associated with MS susceptibility[358]. Fine-mapping studies have narrowed down the list to four credible variants that are likely causal in RA[350]. None of these variants were present in the input genotype dataset used to map meQTLs in this project, however, and so the relative associations with DNAm could not be assessed. Nonetheless, Bayesian co-localisation confirmed that the meQTL signal and disease association likely share a causal variant at this locus, and as such the regulatory SNP identified here (rs6859219) likely tags one of these SNPs. Amongst these credible SNPs defined by Westra et al.[350], two are in high LD ($r^2 = 0.85$) with rs6859219 (rs10213692 & rs71624119), whilst the other two are in relatively low LD (rs7731626, $r^2 = 0.45$ & rs11377254, $r^2 = 0.30$).

IL6ST represents an enticing candidate gene given that it encodes the interleukin 6 signal transducer (gp130). Gp130 is a co-receptor expressed on all cells and binds to IL-6 complexed to either the membrane-bound IL-6 receptor (IL6R, CD126; cis signalling) or the soluble IL-6 receptor (sIL6R; trans signalling)[359]. This transmembrane receptor is therefore essential for the transduction of IL-6 signalling at the cell surface to activate transcriptional programs. The cytokine IL-6 has diverse roles in the pathogenesis of RA, including promotion of chronic inflammation, humoral immunity, and activation of bone-resorbing osteoclasts[360].

The risk variant at this locus was found to regulate expression of *IL6ST* as well as *ANKRD55*, doing so through altered methylation at four CpG sites. Unlike *IL6ST*, the function of *ANKRD55* is unknown, though the gene codes for the Ankyrin Repeat Domain 55. Ankyrin repeats are common protein motifs that function to mediate interactions between proteins[361]. A cis-eQTL at rs6859219 regulating expression of *ANKRD55* has been characterised previously, with this effect limited to CD4$^+$ T cells and absent in CD8$^+$ T cells, CD14$^+$ monocytes, CD19$^+$ B cells, and CD56$^+$ NK cells[362]. This is also congruent with our findings that *ANKRD55* was not expressed in B cells. Deciphering whether or not up-regulation of both genes at this locus increases susceptibility in RA will be an interesting future research question.

5.9.4 DNA methylation implicates common pathways in immune-mediated disease

Extending the analysis beyond RA loci highlighted similar putative mechanisms of DNAm-mediated gene expression at regions conferring susceptibility to MS and asthma, two clinically distinct IMDs. This was particularly pronounced in CD4$^+$ T cells, with CpGs at chromsome 17q12 likely modulating expression levels of *ORMDL3* and *GSDMB*, exemplifying a mechanism of genetic risk that was common to RA, MS, and asthma. Indeed, the findings here confirm the mechanism previously described in whole blood and isolated CD4$^+$ T cells, with these two CpGs found to regulate expression of *ORMDL3* and *GSDMB* in the context of asthma genetic risk[363]. The study by Kothari and colleagues reported an additional three CpGs associated with *ORMDL3* transcript levels, and four for *GSDMB*, likely reflecting the larger sample sizes and use of a targeted, as opposed to genome-wide, approach to identifying molecular associations[363].

The protein product of *ORMDL3*, orosomucoid like 3, is a member of the ORMDL family of transmembrane proteins that are localised to the endoplasmic reticulum[364]. ORMDL3 has been shown to regulate the import of calcium ions into the ER and, in doing so, promoting the unfolded protein response[365]. In CD4$^+$ T cells, ORMDL3 appears to negatively regulate production of IL-2, a cytokine which has diverse roles in the immune system[366]. This therefore highlights a plausible pathway through which decreased *ORMDL3* expression associated with the risk allele may confer immune dysregulation in conditions such as RA, asthma, and MS, as well as a number of additional IMDs such as inflammatory bowel disease[367].

The other gene affected by genetically-conferred DNAm at this locus, *GSDMB*, encodes gasdermin B (GSDMB). The gasdermin proteins, of which GSDMB represents one of five encoded in the human genome (GSDMA, GSDMB, GSDMC, GSDMD, and GSDME/DFNA5), are critical in pyroptosis – a mechanism of programmed cell death that occurs typically upon microbial infection, and drives tissue inflammation[368]. Well-characterised functions of GSDMB in lymphocytes have yet to be defined, though recent work in monocyte cell lines suggests that this protein may contribute to inflammatory disease by promoting non-canonical pyroptosis[369]. In this pathway, following infection GSDMB binds to caspase 4 and increases its enzymatic activity, which subsequently cleaves gasdermin D to induce pyroptosis and the release of pro-inflammatory factors[369].

Another interesting locus that spans all three IMDs included in our analysis was at chromosome 7p15.1, where increased DNAm at three CpG sites (cg07522171, cg11187739, cg16130019) was associated with the risk variant appears to reduce expression of the *JAZF1* gene. This gene was first described upon the discovery that it forms a recurring fusion with another zinc finger gene (*JJAZ1*) during chromosomal translocations in endometrial cancer[370]. Juxtaposed With Another Zinc Finger Protein 1 (JAZF1) has been shown to interact with TAK1, a member of the orphan nuclear receptor family of transcription factors, and as such is also referred to as TAK1-Interacting Protein 27 (TIP27)[371]. The interaction of JAZF1 with TAK1 in the nucleus inhibits the activation of transcription by the latter[371].

Whilst variants mapping to *JAZF1* have been associated with a range of inflammatory conditions, including RA, MS, and asthma as described in this study, little research has examined possible roles in adaptive immunity. In hepatocytes, both in vitro and in vivo mouse data suggest that JAZF1 may play a role in suppressing the production of pro-inflammatory cytokines[372]. This was accompanied by reduced activation of the p38 MAP kinase, JNK, and NFκB signalling proteins[372]. This mechanism was described in the context of systemic inflammation in models of non-alcoholic fatty liver disease, induced *in vitro* by the treatment of cell lines with palmitic acid, or *in vivo* by feeding mice a high fat diet. Whether or not these findings would translate to lymphocytes in the context of autoimmune disease remains to be seen.

These results suggest that DNAm may be crucial in mediating genetic risk spanning multiple immune-mediated diseases. Analysis of shared genetic susceptibility in asthma and autoimmune conditions revealed extensive overlap, with such loci enriched in regulatory regions active in lymphocytes, and pathway analysis implicating T cell function[107]. An analysis of 107 risk SNPs across seven autoimmune diseases found 47 of these (44%) to be associated with more than one condition, likely explaining observations that autoimmunity segregates in families more often than would be expected by chance[106]

5.9.5 Limitations

One limitation of meQTL analyses, and QTL mapping more generally, is the inability of this method to localise the variant responsible for observed associations with molecular traits. Therefore, whilst this approach is useful in elucidating the CpGs and genes involved in disease aetiology, the nomination of multiple putative regulatory variants presents a major obstacle for follow-up functional studies. As a result, a more concerted effort to validate functional variants will be required to confirm these molecular associations. Methods that allow functional activity of thousands of candidate variants to be assessed in parallel have proved successful in unravelling transcriptional regulation at OA loci[373].

In vitro reporter assays can assist in validating allelic effects on transcriptional regulation, as well as indicating whether or not DNAm actively represses this process or simply represents a footprint of active transcription (i.e. allele-specific transcription factor binding). Whilst experimental validation of the regulatory function of a SNP and cis-CpGs on gene expression at the *FCRL3* promoter were successful, effects at other loci were not validated. In the case of *JAZF1*, lack of luciferase activity in Jurkat cells may indicate that the amplified region harbouring the cis-CpGs did not sufficiently capture the gene promoter, or alternatively that this regulatory element is not active in this particular cell line. Therefore, DNAm-mediated regulation of an eQTL effect has been experimentally validated at one RA risk locus, mechanisms at additional loci described here are based exclusively on *in vivo* molecular associations.

Notwithstanding the utility of reporter assays in validating associations, comprehensive evidence of DNAm in a genomic context mediating expression levels of candidate genes would require direct alterations to site-specific methylation in the genome. Attempts to induce DNA de-methylation at CpGs of interest in a Jurkat cell line using a CRISPR-dCas9 system were unsuccessful, precluding the evaluation of such effects. The reasons for the failure of this approach are not clear, but high cell death following plasmid transfection yielded low number of viable co-transfected cells and presented a challenge. Whether the observed lack of de-methylation reflects a problem with the experimental procedure generally, such as the use of co-transfection as opposed to inserting the gRNA sequence into the plasmid, or simply an inability of the gRNAs to target regions of interest, is uncertain at this time. Use of an appropriate positive control that has previously been shown to induce targeted de-methylation would resolve this. Nonetheless, subject to further optimisation (discussed in Chapter 6), this approach has the potential to unequivocally establish the impact of specific cis-CpGs in altering expression levels of genes that are of interest pathophysiologically.

All molecular data integrated in this analysis were generated using array-based technologies. Whilst these technologies allow for affordable quantification of DNAm/transcript levels in many samples, findings are limited to those features included on the array. Particularly in the case of the MethylationEPIC array, whilst this platform offers a considerable increase in coverage relative to the HumanMethylation450 BeadChip that it has superseded, the ~850,000 sites interrogated by this array only represent around 3% of the 28 million CpGs present in the human genome. Whole-genome bisulphite sequencing, as well as methylation-dependent sequencing, a method similar to ChIP-seq that involves isolation and sequencing of methylated DNA, are approaches that can be applied to circumvent this issue. These techniques would also allow the impact on DNAm of disease haplotypes beyond individual SNPs to be extensively assessed. The importance of considering haplotype effects has been demonstrated in a recent study, whereby >36% of methylated regions influenced by trait-associated haplotypes were attributed to non-SNP variation such as insertions/deletions and copy number variants[374].

Finally, the findings described here relate to molecular characterisation of peripheral lymphocytes in RA patients. Whether such conclusions can be extrapolated to cells that migrate into the joint site and exert effector functions within this microenvironment cannot be concluded. Indeed, fibroblasts that exert joint-specific functions have been shown to exhibit distinct epigenetic profiles in relation to DNAm and histone marks[229]. The extent to which the DNA methylome, as well as other epigenetic modifications, confer different cellular properties across distinct tissue sites will be an important consideration for future studies.

Chapter 6 – Conclusions and Future Directions

6.1 Summary and significance of key findings

The aims of this project were to comprehensively map disease-specific lymphocyte DNAm changes in early RA, and to examine the role of DNAm in these cells in mediating previously defined genetic risk in this immune-mediated disease. As such, this represents the first analysis of paired genotype, DNAm, and gene expression data in CD4$^+$ T cells and B cells in a relevant disease context.

The first observation, that no differentially methylated positions were identified in either cell type, was contrary to findings described in a number of previous studies in cells of both the peripheral blood and stromal tissues. The lack of significant differences in DNAm levels between patients and controls at any CpG sites likely reflects to some degree the use of a matched disease control group of non-RA arthritis patients in place of healthy controls. Matching for acute phase response (as well as age and sex) and the use of robust statistical approaches were features of the work presented here, and it is conceivable that confounding sources of differential methylation between patients and controls have been under-reported in this field to date. The presence of differential variability in DNAm in RA patients does however suggest a disease-specific deviation from the stable patterns of DNAm in health (or in the case of the present study, non-RA arthropathies). However, whilst the identification of differentially variable positions has emerged as an important tool in the selection of disease-specific DNAm features, most large differences in methylation observed between patients and controls are believed to be driven by genetic variants (meQTLs) and cell type heterogeneity[375].

Genetically conferred patterns of DNAm were approximated by mapping meQTLs in cis and trans across all patients in both cell types. Such effects in cis were present at ~7.5% of all CpGs tested, indicating that genetic variants are a prominent source of variability in DNAm levels genome-wide in CD4$^+$ T cells and B cells, even in the relatively small sample cohort included here. Importantly, genetic variants that had previously been identified at susceptibility loci for RA, as well as other immune-mediated (MS, asthma) and non-immune-mediated (osteoarthritis) diseases were found to function as meQTLs in these cell types. Many of the highlighted RA loci were previously revealed to exhibit cis-regulatory activity on expression of candidate genes in these cells[139], and DNAm sites regulated in cis are enriched at active chromatin regions and binding sites of transcription factors, suggesting DNAm modifications impact downstream gene transcription. These findings confirmed the utility of cell-specific DNAm profiling and meQTL mapping in assigning regulatory function to non-coding risk variants. Having performed the analysis in cell types known to mediate RA pathogenesis, and focussed on relevant patient cohorts, cis genetic-epigenetic interactions in the context of early RA have been described here for the first time.

Finally, given the findings described above, integration of matched transcriptomic data allowed for DNAm changes associated with risk loci to be related to transcript levels of proximal candidate genes at these loci. By identifying such associations, and using a causal inference test to infer molecular mediation by DNAm, a number of candidate genes (most prominently in CD4$^+$ T cells) were implicated downstream of risk-associated cis-meQTL effects. This analysis also highlighted overlap in methylation-mediated genes between RA and other immune-mediated diseases, suggesting common regulatory effects that perturb lymphocyte function across multiple conditions.

Collectively, the results presented here add to our understanding of mechanisms through which non-coding genetic risk in RA engenders pathological adaptive immune responses in individuals harbouring risk alleles. Given lymphocytes are a popular cell substrate for EWASs in RA and other lymphocyte-mediated autoimmune diseases, the cis-genetic effects described here represent an important reference point when assigning differential patterns of methylation to sequence variants. The presence of DNAm modifications that influence transcription at RA risk loci adds credence to the candidacy of putatively causal genes for functional validation.

6.2 Outstanding questions for future study

In addition to answering some important mechanistic questions regarding genetic risk in complex immune mediated diseases, the work presented here leaves some questions unanswered, and generates potentially interesting hypotheses for future study.

6.2.1 What are the relative genetic and non-genetic influences on DNA methylation in early arthritis?

This study is unique in that lymphocyte DNAm changes associated with RA have been assessed in an EWAS, with genetic effects on DNAm in the same cohort quantified by mapping meQTLs genome-wide, as well as at established risk loci. Whilst meQTL mapping revealed that risk alleles at RA-associated loci can influence DNAm in cis, with strong allelic effects at some loci, such effects were identified across patients with varying arthropathies, and as such were not RA-specific. Whether differential DNAm in RA patients is conferred by genetic or environmental risk factors, and to what extent these two interact, remains largely unknown, and is usually not addressed in typical EWASs.

Studies comparing whole blood of monozygotic twins with dizygotic twins showed that the overall contribution of genetics to inter-individual variability in DNAm was low compared with non-shared environmental effects[376]. However, at positions with variable, intermediate (mean DNAm 20%-80%) DNAm levels, the contribution of genetic factors to variation in methylation was considerably higher[376]. This is consistent with the findings from the meQTL analysis in CD4$^+$ T cells and B cells described here, with meQTL-associated CpGs having more intermediate levels of DNAm than those not associated with genetic variation.

Previous observations that CpG sites at which DNAm is highly heritable were enriched at CpG island (CGI) shores, intergenic regions, and distal promoter regions, with a depletion in CGIs and shelves was consistent with the mapping of meQTLs described in Chapter 5[377]. These findings demonstrate that the contribution of genetic factors to DNAm differs depending on the genomic context. Interactions between environmental RA risk factors and genetic variants that influence DNAm may represent a mechanism through which dynamic epigenetic changes act to integrate genetic and non-genetic sources of risk in complex autoimmune disease.

There is now good evidence that age, sex, and smoking – all of which represent risk factors in RA – can interact with genetic effects to shape the DNA methylome[376, 377]. Isolated instances of such effects in RA have now been described, with DNA methylation at a CpG in the MHC region potentially mediating a gene-smoking interaction that confers risk of developing ACPA+ RA[378]. Interaction effects between disease diagnosis and genetic variation on DNAm were not prevalent in the analysis described in the present study, and this may reflect the fact that diagnosis (RA or non-RA) represents a confluence of many such risk factors. Collecting large cohorts of comprehensively phenotyped patients with detailed data for non-genetic risk factors such as smoking may allow for gene-environment effects on DNAm to be better characterized. This will help to reveal the extent to which DNAm modifications associated with RA are mediated solely by genetics, as opposed to a combination of genetic and environmental exposures. For example, the CD4$^+$ T cell meQTL described in chapter 5 at which cis-CpG DNAm regulates transcription of *ANKRD55* and *IL6ST* maps to a chromatin region that becomes accessible upon stimulation of CD4$^+$ T cell subsets[340].

Allele-specific methylation/expression (ASM/ASE) may also facilitate the identification of such interactions between genetics and environment. This involves comparing relative methylation/expression associated with each allele copy in individuals who are heterozygous at the regulatory SNP, and as such controls for inter-individual confounding variability (relative effects are measured within the same individual). This approach has been applied to human gene expression data, and offered increased sensitivity upon eQTL interaction analyses, identifying 35 genetic × environment effects[379]. Employing studies of ASM in conjunction with meQTL analysis can yield greater discovery power than using either in isolation, and tissue-specific ASM has been used to facilitate the identification of causal variants in complex neurological disease[254]. Systematic identification of meQTLs/ASM that are impacted by environmental factors would provide further clues as to the extent to which the DNA methylome functions at the interface of genetic and environmental risk during complex diseases such as RA.

6.2.2 Are DNA methylation patterns in peripheral lymphocytes reflective of joint-homing cells?

In RA, though the initial break in immune tolerance occurs in the primary and secondary lymphoid organs, and immune activation with autoantibody production is initiated in the periphery, adaptive immune cells exert important effector functions in the joint tissue[380]. Though studies of epigenetic signatures in FLS have given insights into DNAm changes that occur in cells resident within the synovial tissue[174, 229], no such analysis has been performed of migratory lymphocytes in the joint tissue itself.

Nonetheless, pro-inflammatory immune cells within the RA joint have been described at the single cell level[58], and epigenetic characterisation of these cells would undoubtedly aid in understanding whether DNAm changes precede this transition to a pro-inflammatory phenotype. Indeed, the single-cell analysis of joint cells has described three distinct CD4$^+$ T cell subsets, with expression profiles distinguishing these as cells with different effector functions[58]. A population of CD4$^+$ T cells that are expanded in RA synovium express markers consistent with an increased capacity to migrate to inflamed tissue sites, and promote local B cell autoantibody production[60]. It may therefore be the case that cell subtype-specific changes that occur in the synovium are missed by profiling bulk lymphocyte populations in the blood. A study aimed at comprehensive epigenetic profiling of peripheral and joint-resident lymphocytes in RA patients, perhaps at the single cell level, would be an interesting approach to answer such questions. Indeed, extending analyses to whole-genome quantification of DNAm using WGBS will be a worthwhile endeavour to extend analyses beyond pre-determined probes on an array, and capture changes at regions that are not targeted by such technologies as the MethylationEPIC array.

In the current study, patients with early RA were specifically recruited to facilitate the identification of DNAm changes that precede long-term chronic inflammation. However, given that DNAm modifications in a given cell are temporally dynamic (i.e. may differ throughout stages of disease), a longitudinal assessment of patients from pre-RA to established RA would allow for the distinction of alterations that precede RA from those that are a consequence of disease. As such, moving forwards more comprehensive insights into epigenetic mechanisms of disease will require that such changes are considered not only in space (tissue), but also in time (disease stage).

6.2.3 How do changes in DNA methylation coordinate with other epigenetic mechanisms of transcriptional regulation?

Regulation of gene expression is a complex process that involves coordinated changes in biochemical processes that can be directly or indirectly influenced by variants in the genome sequence. DNAm therefore likely represents a proxy read out of multiple molecular processes that occur during transcriptional regulation, including chemical modification of histone proteins, accessibility of chromatin, and binding of transcription factors. This is consistent with the results described in Chapter 4 indicating that CpGs associated with RA genetic risk loci are over-represented in enhancer regions and at the binding sites of some transcription factors, including NFκB. Such enrichments were, however, assigned based on data from publicly available consortia datasets. To gain a more comprehensive picture of the regulatory mechanisms that control the levels of gene expression at risk loci in immune-mediated diseases such as RA, multiple sources of evidence would ideally be profiled in patient cohorts.

Indeed, RA-associated variants identified in GWASs are particularly enriched in regions of chromatin that become active upon stimulation via the T cell co-receptors (CD3/CD28) as well as with polarizing cytokines[117, 340]. This would suggest that certain regulatory variants may exert their effects only after the cell has encountered a stimulus, meaning that meQTLs conditional on cellular activation state may well exist and contribute to expression levels of candidate genes. Given that most large-scale consortia efforts to map chromatin modifications or binding events of transcriptional regulators are performed in unstimulated cell lines or healthy human subjects, localisation of cis-regulatory effects to such regions may be overlooked.

Advancements in technology now allow for multiple markers of transcriptional regulation to be concurrently profiled in cell types of interest. Methods have been developed that enable not only profiling of DNA methylation and transcription from an individual cell, but also assays to define areas of accessible chromatin, providing a powerful approach for integrating multiple layers of regulatory information[381]. Whether DNAm modifications at risk loci occur at an earlier time point to the increase in chromatin accessibility and transcription factor binding, or represent a secondary event that has subsequent effects on transcription, remains to be determined.

6.2.4 How do risk-associated modifications to DNA methylation impact cellular phenotype in immune-mediated diseases?

The focus of the work described in this project was to define the mechanisms through which DNAm can contribute to transcriptional regulation of candidate genes at known genetic risk loci. Ultimately, however, genetic risk in complex disease mediates pathogenesis when such regulatory effects manifest at the level of cellular function. As was discussed in Chapter 5, many of the candidate genes highlighted as being subjected to DNAm-mediated regulation in this analysis have poorly defined immunobiological functions. Functional cellular studies must now follow to investigate precisely how the up-/down-regulation of these genes leads to dysregulated adaptive immunity.

The final section of Chapter 5 discussed an attempt to induce site-specific DNA de-methylation using a CRISPR-dCas9 delivery system to deliver the TET1 catalytic domain to regions harbouring CpGs of interest. Whilst unsuccessful, further optimisation of this method in lymphocyte cell lines will be valuable in not only validating the regulatory capacity of DNAm at these sites, but also in defining the impact of such modifications on cellular function. To this end, cell lysates from cells subject to this treatment have been stored, which will allow secreted proteins such as pro-inflammatory cytokines to be measured.

Using the same CRISPR-dCas9 system to induce de-methylation at the *FOXP3* promoter and Treg-specific de-methylated region (TSDR) in Jurkat cells was found to upregulate expression of *FOXP3* and confer regulatory T cell properties in this cell line[382]. The gRNA sequences used in this study have now been obtained for use as a positive control in future optimisations. These findings also raise the prospect that targeted DNAm modification may be an effective therapeutic strategy. Unlike techniques such as RNA interference which target transcript levels directly, DNAm is maintained in daughter cells following mitosis, and as a result represents a more stable method of controlling gene expression. In addition, as DNAm is dynamic and can be reversed, targeting such modifications at loci associated with risk variants should represent a safer approach than direct genome editing. Novel systems that allow co-delivery of the TET1 catalytic domain together with proteins that promote activity of this de-methylating enzyme will increase efficacy of these targeted approaches[383]. As well as the delivery of TET1 catalytic domains, the use of DNMT methyltransferase domains will enable targeted methylation.

Following validation of such effects in vitro, trials in animal models will be necessary to confirm the feasibility and efficacy of this approach for translation to human studies. Currently, conclusive evidence that the presence or absence of DNAm at specific locus can confer autoimmune properties on a cell are still lacking, though regions highlighted in this study warrant further characterisation. Whether or not such an approach proves attainable in a clinical setting remains to be seen. Nonetheless, combining DNAm with other sources of molecular data clearly represents a valuable means of discovering aetiological genes and pathways in RA and related immune-mediated disease.

The analyses presented in this thesis represent one step in an ongoing effort to link complex genetic risk in RA to cellular pathways that may serve as targets for therapeutic intervention, or be useful biomarkers of disease prognosis or treatment response. Such efforts are now an important focus of functional genomics in the post-GWAS era, where assigning cellular phenotype to static genotypes remains a considerable challenge in translating GWAS discoveries into patient benefit.

References

1. Silman AJ, Pearson JE. Epidemiology and genetics of rheumatoid arthritis. Arthritis Research & Therapy 2002; 4:S265-S72.
2. Symmons D, Turner G, Webb R, Asten P, Barrett E, Lunt M, et al. The prevalence of rheumatoid arthritis in the United Kingdom: new estimates for a new century. Rheumatology 2002; 41:793-800.
3. Alamanos Y, Drosos AA. Epidemiology of adult rheumatoid arthritis. Autoimmunity Reviews 2005; 4:130-6.
4. Welsing PMJ, van Gestel AM, Swinkels HL, Kiemeney L, van Riel P. The relationship between disease activity, joint destruction, and functional capacity over the course of rheumatoid arthritis. Arthritis and Rheumatism 2001; 44:2009-17.
5. Lundkvist J, Kastang F, Kobelt G. The burden of rheumatoid arthritis and access to treatment: health burden and costs. European Journal of Health Economics 2008; 8:S49-S60.
6. National Audit Office. Services for People with Rheumatoid Arthritis. Report by the Comptroller and Auditor General. 2009.
7. Isaacs JD. The changing face of rheumatoid arthritis: sustained remission for all? Nature Reviews Immunology 2010; 10:605-11.
8. van Vollenhoven RF. Treatment of rheumatoid arthritis: state of the art 2009. Nature Reviews Rheumatology 2009; 5:531-41.
9. Wolfe F, Freundlich B, Straus WL. Increase in cardiovascular and cerebrovascular disease prevalence in rheumatoid arthritis. Journal of Rheumatology 2003; 30:36-40.
10. Hochberg MC, Johnston SS, John AK. The incidence and prevalence of extra-articular and systemic manifestations in a cohort of newly-diagnosed patients with rheumatoid arthritis between 1999 and 2006. Current Medical Research and Opinion 2008; 24:469-80.
11. Dougados M, Soubrier M, Antunez A, Balint P, Balsa A, Buch MH, et al. Prevalence of comorbidities in rheumatoid arthritis and evaluation of their monitoring: results of an international, cross-sectional study (COMORA). Annals of the Rheumatic Diseases 2014; 73:62-8.
12. Gonzalez A, Kremers HM, Crowson CS, Nicola PJ, Davis JM, Therneau TM, et al. The widening mortality gap between rheumatoid arthritis patients and the general population. Arthritis and Rheumatism 2007; 56:3583-7.

13. Levy L, Fautrel B, Barnetche T, Schaeverbeke T. Incidence and risk of fatal myocardial infarction and stroke events in rheumatoid arthritis patients. A systematic review of the literature. Clinical and Experimental Rheumatology 2008; 26:673-9.

14. Jacobsson LTH, Turesson C, Nilsson JA, Petersson IF, Lindqvist E, Saxne T, et al. Treatment with TNF blockers and mortality risk in patients with rheumatoid arthritis. Annals of the Rheumatic Diseases 2007; 66:670-5.

15. Cross M, Smith E, Hoy D, Carmona L, Wolfe F, Vos T, et al. The global burden of rheumatoid arthritis: estimates from the Global Burden of Disease 2010 study. Annals of the Rheumatic Diseases 2014; 73:1316-22.

16. Prevoo MLL, Vanthof MA, Kuper HH, Vanleeuwen MA, Vandeputte LBA, Vanriel P. Modified disease-activity scores that include 28-joint counts - development and validation in a prospective longitudinal-study of patients with rheumatoid-arthritis. Arthritis and Rheumatism 1995; 38:44-8.

17. Nielen MMJ, van Schaardenburg D, Reesink HW, van de Stadt RJ, van der Horst-Bruinsma IE, de Koning M, et al. Specific autoantibodies precede the symptoms of rheumatoid arthritis - A study of serial measurements in blood donors. Arthritis and Rheumatism 2004; 50:380-6.

18. Nielsen SF, Bojesen SE, Schnohr P, Nordestgaard BG. Elevated rheumatoid factor and long term risk of rheumatoid arthritis: a prospective cohort study. British Medical Journal 2012; 345:9.

19. Schellekens GA, Visser H, de Jong BAW, Van den Hoogen FHJ, Hazes JMW, Breedveld FC, et al. The diagnostic properties of rheumatoid arthritis antibodies recognizing a cyclic citrullinated peptide. Arthritis and Rheumatism 2000; 43:155-63.

20. Firestein GS, McInnes IB. Immunopathogenesis of Rheumatoid Arthritis. Immunity 2017; 46:183-96.

21. Machold KP, Stamm TA, Nell VPK, Pflugbeil S, Aletaha D, Steiner G, et al. Very recent onset rheumatoid arthritis: clinical and serological patient characteristics associated with radiographic progression over the first years of disease. Rheumatology 2007; 46:342-9.

22. Toes R, Pisetsky DS. Pathogenic effector functions of ACPA: Where do we stand? Annals of the Rheumatic Diseases 2019; 78:716-21.

23. Rantapaa-Dahlqvist S, de Jong BAW, Berglin E, Hallmans G, Wadell G, Stenlund H, et al. Antibodies against cyclic citrullinated peptide and IgA rheumatoid factor predict the development of rheumatoid arthritis. Arthritis and Rheumatism 2003; 48:2741-9.

24. Gerlag DM, Raza K, van Baarsen LGM, Brouwer E, Buckley CD, Burmester GR, et al. EULAR recommendations for terminology and research in individuals at risk of rheumatoid arthritis: report from the Study Group for Risk Factors for Rheumatoid Arthritis. Annals of the Rheumatic Diseases 2012; 71:638-41.

25. Ronnelid J, Wick MC, Lampa J, Lindblad S, Nordmark B, Klareskog L, et al. Longitudinal analysis of citrullinated protein/peptide antibodies (anti-CP) during 5 year follow up in early rheumatoid arthritis: anti-CP status predicts worse disease activity and greater radiological progression. Annals of the Rheumatic Diseases 2005; 64:1744-9.

26. Brink M, Hansson M, Mathsson L, Jakobsson PJ, Holmdahl R, Hallmans G, et al. Multiplex Analyses of Antibodies Against Citrullinated Peptides in Individuals Prior to Development of Rheumatoid Arthritis. Arthritis and Rheumatism 2013; 65:899-910.

27. van der Woude D, Dahlqvist SR, Ioan-Facsinay A, Onnekink C, Schwarte CM, Verpoort KN, et al. Epitope spreading of the anti-citrullinated protein antibody response occurs before disease onset and is associated with the disease course of early arthritis. Annals of the Rheumatic Diseases 2010; 69:1554-61.

28. Sokolove J, Bromberg R, Deane KD, Lahey LJ, Derber LA, Chandra PE, et al. Autoantibody Epitope Spreading in the Pre-Clinical Phase Predicts Progression to Rheumatoid Arthritis. Plos One 2012; 7:9.

29. Suwannalai P, van de Stadt LA, Radner H, Steiner G, El-Gabalawy HS, Jol-van der Zijde CM, et al. Avidity maturation of anti-citrullinated protein antibodies in rheumatoid arthritis. Arthritis and Rheumatism 2012; 64:1323-8.

30. Malmstrom V, Catrina AI, Klareskog L. The immunopathogenesis of seropositive rheumatoid arthritis: from triggering to targeting. Nature Reviews Immunology 2017; 17:60-75.

31. Raposo B, Merky P, Lundqvist C, Yamada H, Urbonaviciute V, Niaudet C, et al. T cells specific for post-translational modifications escape intrathymic tolerance induction. Nature Communications 2018; 9:11.

32. Deane KD, O'Donnell CI, Hueber W, Majka DS, Lazar AA, Derber LA, et al. The Number of Elevated Cytokines and Chemokines in Preclinical Seropositive Rheumatoid Arthritis Predicts Time to Diagnosis in an Age-Dependent Manner. Arthritis and Rheumatism 2010; 62:3161-72.

33. Kokkonen H, Soderstrom I, Rocklov J, Hallmans G, Lejon K, Dahlqvist SR. Up-Regulation of Cytokines and Chemokines Predates the Onset of Rheumatoid Arthritis. Arthritis and Rheumatism 2010; 62:383-91.

34. Mateen S, Moin S, Shahzad S, Khan AQ. Level of inflammatory cytokines in rheumatoid arthritis patients: Correlation with 25-hydroxy vitamin D and reactive oxygen species. Plos One 2017; 12:11.

35. Anderson AE, Pratt AG, Sedhom MAK, Doran JP, Routledge C, Hargreaves B, et al. IL-6-driven STAT signalling in circulating CD4+lymphocytes is a marker for early anticitrullinated peptide antibody-negative rheumatoid arthritis. Annals of the Rheumatic Diseases 2016; 75:466-73.

36. Clavel C, Nogueira L, Laurent L, Lobagiu C, Vincent C, Sebbag M, et al. Induction of macrophage secretion of tumor necrosis factor a through Fc gamma receptor IIa engagement by rheumatoid arthritis-specific autoantibodies to citrullinated proteins complexed with fibrinogen. Arthritis and Rheumatism 2008; 58:678-88.

37. Lu MC, Lai NS, Yu HC, Huang HB, Hsieh SC, Yu CL. Anti-Citrullinated Protein Antibodies Bind Surface-Expressed Citrullinated Grp78 on Monocyte/Macrophages and Stimulate Tumor Necrosis Factor alpha Production. Arthritis and Rheumatism 2010; 62:1213-23.

38. Trouw LA, Haisma EM, Levarht EWN, van der Woude D, Ioan-Facsinay A, Daha VR, et al. Anti-Cyclic Citrullinated Peptide Antibodies From Rheumatoid Arthritis Patients Activate Complement via Both the Classical and Alternative Pathways. Arthritis and Rheumatism 2009; 60:1923-31.

39. van Baarsen LGM, de Hair MJH, Ramwadhdoebe TH, Zijlstra I, Maas M, Gerlag DM, et al. The cellular composition of lymph nodes in the earliest phase of inflammatory arthritis. Annals of the Rheumatic Diseases 2013; 72:1420-4.

40. Ramwadhdoebe TH, Hahnlein J, Maijer KI, van Boven LJ, Gerlag DM, Tak PP, et al. Lymph node biopsy analysis reveals an altered immunoregulatory balance already during the at-risk phase of autoantibody positive rheumatoid arthritis. European Journal of Immunology 2016; 46:2812-21.

41. Pfeifle R, Rothe T, Ipseiz N, Scherer HU, Culemann S, Harre U, et al. Regulation of autoantibody activity by the IL-23-T(H)17 axis determines the onset of autoimmune disease. Nature Immunology 2017; 18:104-13.

42. Ferreira RC, Freitag DF, Cutler AJ, Howson JMM, Rainbow DB, Smyth DJ, et al. Functional IL6R 358Ala Allele Impairs Classical IL-6 Receptor Signaling and Influences Risk of Diverse Inflammatory Diseases. Plos Genetics 2013; 9:12.

43. Dendrou CA, Fugger L, Friese MA. Immunopathology of multiple sclerosis. Nature Reviews Immunology 2015; 15:545-58.

44. Paschou SA, Papadopoulou-Marketou N, Chrousos GP, Kanaka-Gantenbein C. On type 1 diabetes mellitus pathogenesis. Endocrine Connections 2018; 7:R38-R46.

45. Farh KKH, Marson A, Zhu J, Kleinewietfeld M, Housley WJ, Beik S, et al. Genetic and epigenetic fine mapping of causal autoimmune disease variants. Nature 2015; 518:337-43.

46. Glyn-Jones S, Palmer AJR, Agricola R, Price AJ, Vincent TL, Weinans H, et al. Osteoarthritis. Lancet 2015; 386:376-87.

47. de Hair MJH, van de Sande MGH, Ramwadhdoebe TH, Hansson M, Landewe R, van der Leij C, et al. Features of the Synovium of Individuals at Risk of Developing Rheumatoid Arthritis. Arthritis & Rheumatology 2014; 66:513-22.

48. Strand V, Kimberly R, Isaacs JD. Biologic therapies in rheumatology: lessons learned, future directions. Nature Reviews Drug Discovery 2007; 6:75-92.

49. McInnes IB, Schett G. Mechanisms of Disease The Pathogenesis of Rheumatoid Arthritis. New England Journal of Medicine 2011; 365:2205-19.

50. Sabeh F, Fox D, Weiss SJ. Membrane-Type I Matrix Metalloproteinase-Dependent Regulation of Rheumatoid Arthritis Synoviocyte Function. Journal of Immunology 2010; 184:6396-406.

51. McInnes IB, Schett G. Pathogenetic insights from the treatment of rheumatoid arthritis. Lancet 2017; 389:2328-37.

52. Humby F, Lewis M, Ramamoorthi N, Hackney JA, Barnes MR, Bombardieri M, et al. Synovial cellular and molecular signatures stratify clinical response to csDMARD therapy and predict radiographic progression in early rheumatoid arthritis patients. Annals of the Rheumatic Diseases 2019; 78:761-72.

53. Thurlings RM, Wijbrandts CA, Mebius RE, Cantaert T, Dinant HJ, van der Pouw-Kraan T, et al. Synovial lymphoid neogenesis does not define a specific clinical rheumatoid arthritis phenotype. Arthritis and Rheumatism 2008; 58:1582-9.

54. Yang Z, Shen Y, Oishi H, Matteson EL, Tian L, Goronzy JJ, et al. Restoring oxidant signaling suppresses proarthritogenic T cell effector functions in rheumatoid arthritis. Science Translational Medicine 2016; 8:13.

55. Yeo L, Toellner KM, Salmon M, Filer A, Buckley CD, Raza K, et al. Cytokine mRNA profiling identifies B cells as a major source of RANKL in rheumatoid arthritis. Annals of the Rheumatic Diseases 2011; 70:2022-8.

56. Takemura S, Klimiuk PA, Braun A, Goronzy JJ, Weyand CM. T cell activation in rheumatoid synovium is B cell dependent. Journal of Immunology 2001; 167:4710-8.

57. Filer A, Ward LSC, Kemble S, Davies CS, Munir H, Rogers R, et al. Identification of a transitional fibroblast function in very early rheumatoid arthritis. Annals of the Rheumatic Diseases 2017; 76:2105-12.

58. Zhang F, Wei K, Slowikowski K, Fonseka CY, Rao DA, Kelly S, et al. Defining inflammatory cell states in rheumatoid arthritis joint synovial tissues by integrating single-cell transcriptomics and mass cytometry. Nature immunology 2019; 20:928-42.

59. Croft AP, Campos J, Jansen K, Turner JD, Marshall J, Attar M, et al. Distinct fibroblast subsets drive inflammation and damage in arthritis. Nature 2019; 570:246-51.

60. Rao DA, Gurish MF, Marshall JL, Slowikowski K, Fonseka CY, Liu YY, et al. Pathologically expanded peripheral T helper cell subset drives B cells in rheumatoid arthritis. Nature 2017; 542:110-+.

61. McInnes IB, Schett G. Cytokines in the pathogenesis of rheumatoid arthritis. Nature Reviews Immunology 2007; 7:429-42.

62. Harre U, Georgess D, Bang H, Bozec A, Axmann R, Ossipova E, et al. Induction of osteoclastogenesis and bone loss by human autoantibodies against citrullinated vimentin. Journal of Clinical Investigation 2012; 122:1791-802.

63. Spreafico R, Rossetti M, van Loosdregt J, Wallace CA, Massa M, Magni-Manzoni S, et al. A circulating reservoir of pathogenic-like CD4(+) T cells shares a genetic and phenotypic signature with the inflamed synovial micro-environment. Annals of the Rheumatic Diseases 2016; 75:459-65.

64. Aletaha D, Neogi T, Silman AJ, Funovits J, Felson DT, Bingham CO, et al. 2010 Rheumatoid Arthritis Classification Criteria An American College of Rheumatology/European League Against Rheumatism Collaborative Initiative. Arthritis and Rheumatism 2010; 62:2569-81.

65. Brown PM, Pratt AG, Isaacs JD. Mechanism of action of methotrexate in rheumatoid arthritis, and the search for biomarkers. Nature Reviews Rheumatology 2016; 12:731-42.

66. Maini R, St Clair EW, Breedveld F, Furst D, Kalden J, Weisman M, et al. Infliximab (chimeric anti-tumour necrosis factor alpha monoclonal antibody) versus placebo in rheumatoid arthritis patients receiving concomitant methotrexate: a randomised phase III trial. Lancet 1999; 354:1932-9.

67. Maini RN, Breedveld FC, Kalden JR, Smolen JS, Furst D, Weisman MH, et al. Sustained improvement over two years in physical function, structural damage, and signs and symptoms among patients with rheumatoid arthritis treated with infliximab and methotrexate. Arthritis and Rheumatism 2004; 50:1051-65.

68. Genovese MC, McKay JD, Nasonov EL, Mysler EF, da Silva NA, Alecock E, et al. Interleukin-6 Receptor Inhibition With Tocilizumab Reduces Disease Activity in Rheumatoid Arthritis With Inadequate Response to Disease-Modifying Antirheumatic Drugs The Tocilizumab in Combination With Traditional Disease-Modifying Antirheumatic Drug Therapy Study. Arthritis and Rheumatism 2008; 58:2968-80.

69. Emery P, Keystone E, Tony HP, Cantagrel A, van Vollenhoven R, Sanchez A, et al. IL-6 receptor inhibition with tocilizumab improves treatment outcomes in patients with rheumatoid arthritis refractory to anti-tumour necrosis factor biologicals: results from a 24-week multicentre randomised placebo-controlled trial. Annals of the Rheumatic Diseases 2008; 67:1516-23.

70. Maxwell LJ, Singh JA. Abatacept for Rheumatoid Arthritis: A Cochrane Systematic Review. Journal of Rheumatology 2010; 37:234-45.

71. Gottenberg JE, Ravaud P, Cantagrel A, Combe B, Flipo RM, Schaeverbeke T, et al. Positivity for anti-cyclic citrullinated peptide is associated with a better response to abatacept: data from the 'Orencia and Rheumatoid Arthritis' registry. Annals of the Rheumatic Diseases 2012; 71:1815-9.

72. Edwards JCW, Szczepanski L, Szechinski J, Filipowicz-Sosnowska A, Emery P, Close DR, et al. Efficacy of B-cell-targeted therapy with rituximab in patients with rheumatoid arthritis. New England Journal of Medicine 2004; 350:2572-81.

73. Cohen SB, Keystone E, Genovese MC, Emery P, Peterfy C, Tak PP, et al. Continued inhibition of structural damage over 2 years in patients with rheumatoid arthritis treated with rituximab in combination with methotrexate. Annals of the Rheumatic Diseases 2010; 69:1158-61.

74. Melet J, Mulleman D, Goupille P, Ribourtout B, Watier H, Thibault G. Rituximab-Induced T Cell Depletion in Patients With Rheumatoid Arthritis Association With Clinical Response. Arthritis and Rheumatism 2013; 65:2783-90.

75. Bell GM, Anderson AE, Diboll J, Reece R, Eltherington O, Harry RA, et al. Autologous tolerogenic dendritic cells for rheumatoid and inflammatory arthritis. Annals of the Rheumatic Diseases 2017; 76:227-34.

76. Frisell T, Holmqvist M, Kallberg H, Klareskog L, Alfredsson L, Askling J. Familial Risks and Heritability of Rheumatoid Arthritis Role of Rheumatoid Factor/Anti-

Citrullinated Protein Antibody Status, Number and Type of Affected Relatives, Sex, and Age. Arthritis and Rheumatism 2013; 65:2773-82.

77. MacGregor AJ, Snieder H, Rigby AS, Koskenvuo M, Kaprio J, Aho K, et al. Characterizing the quantitative genetic contribution to rheumatoid arthritis using data from twins. Arthritis and Rheumatism 2000; 43:30-7.

78. Hensvold AH, Magnusson PKE, Joshua V, Hansson M, Israelsson L, Ferreira R, et al. Environmental and genetic factors in the development of anticitrullinated protein antibodies (ACPAs) and ACPA-positive rheumatoid arthritis: an epidemiological investigation in twins. Annals of the Rheumatic Diseases 2015; 74:375-80.

79. Speed D, Hemani G, Johnson MR, Balding DJ. Improved Heritability Estimation from Genome-wide SNPs. American Journal of Human Genetics 2012; 91:1011-21.

80. Stastny P. Association of B-cell alloantigen DRW4 with rheumatoid-arthritis. New England Journal of Medicine 1978; 298:869-71.

81. Gregersen PK, Silver J, Winchester RJ. The shared epitope hypothesis - an approach to understanding the molecular-genetics of susceptibility to rheumatoid-arthritis. Arthritis and Rheumatism 1987; 30:1205-13.

82. Huizinga TWJ, Amos CI, van der Helm-van Mil AHM, Chen W, van Gaalen FA, Jawaheer D, et al. Refining the complex rheumatoid arthritis phenotype based on specificity of the HLA-DRB1 shared epitope for antibodies to citrullinated proteins. Arthritis and Rheumatism 2005; 52:3433-8.

83. Padyukov L, Silva C, Stolt P, Alfredsson L, Klareskog L, Epidemiological Invest R. A gene-environment interaction between smoking and shared epitope genes in HLA-DR provides a high risk of seropositive rheumatoid arthritis. Arthritis and Rheumatism 2004; 50:3085-92.

84. Linn-Rasker SP, van der Helm-van Mil AHM, van Gaalen FA, Kloppenburg M, de Vries RRP, le Cessie S, et al. Smoking is a risk factor for anti-CCP antibodies only in rheumatoid arthritis patients who carry HLA-DRB1 shared epitope alleles. Annals of the Rheumatic Diseases 2006; 65:366-71.

85. Raychaudhuri S, Sandor C, Stahl EA, Freudenberg J, Lee HS, Jia XM, et al. Five amino acids in three HLA proteins explain most of the association between MHC and seropositive rheumatoid arthritis. Nature Genetics 2012; 44:291-U91.

86. Pociot F, Lernmark A. Genetic risk factors for type 1 diabetes. Lancet 2016; 387:2331-9.

87. Sawcer S, Hellenthal G, Pirinen M, Spencer CCA, Patsopoulos NA, Moutsianas L, et al. Genetic risk and a primary role for cell-mediated immune mechanisms in multiple sclerosis. Nature 2011; 476:214-9.

88. Cortes A, Hadler J, Pointon JP, Robinson PC, Karaderi T, Leo P, et al. Identification of multiple risk variants for ankylosing spondylitis through high-density genotyping of immune-related loci. Nature Genetics 2013; 45:730-+.

89. Bentham J, Morris DL, Graham DSC, Pinder CL, Tombleson P, Behrens TW, et al. Genetic association analyses implicate aberrant regulation of innate and adaptive immunity genes in the pathogenesis of systemic lupus erythematosus. Nature Genetics 2015; 47:1457-+.

90. Hill JA, Southwood S, Sette A, Jevnikar AM, Bell DA, Cairns E. Cutting edge: The conversion of arginine to citrulline allows for a high-affinity peptide interaction with the rheumatoid arthritis-associated HLA-DRB1*0401 MHC class II molecule. Journal of Immunology 2003; 171:538-41.

91. Scally SW, Petersen J, Law SC, Dudek NL, Nel HJ, Loh KL, et al. A molecular basis for the association of the HLA-DRB1 locus, citrullination, and rheumatoid arthritis. Journal of Experimental Medicine 2013; 210:2569-82.

92. Ting YT, Petersen J, Ramarathinam SH, Scally SW, Loh KL, Thomas R, et al. The interplay between citrullination and HLA-DRB1 polymorphism in shaping peptide binding hierarchies in rheumatoid arthritis. Journal of Biological Chemistry 2018; 293:3236-51.

93. Snir O, Rieck M, Gebe JA, Yue BB, Rawlings CA, Nepom G, et al. Identification and Functional Characterization of T Cells Reactive to Citrullinated Vimentin in HLA-DRB1*0401-Positive Humanized Mice and Rheumatoid Arthritis Patients. Arthritis and Rheumatism 2011; 63:2873-83.

94. Law SC, Street S, Yu CHA, Capini C, Ramnoruth S, Nel HJ, et al. T-cell autoreactivity to citrullinated autoantigenic peptides in rheumatoid arthritis patients carrying HLA-DRB1 shared epitope alleles. Arthritis Research & Therapy 2012; 14:12.

95. James EA, Rieck M, Pieper J, Gebe JA, Yue BB, Tatum M, et al. Citrulline-Specific Th1 Cells Are Increased in Rheumatoid Arthritis and Their Frequency Is Influenced by Disease Duration and Therapy. Arthritis & Rheumatology 2014; 66:1712-22.

96. Arnoux F, Mariot C, Peen E, Lambert NC, Balandraud N, Roudier J, et al. Peptidyl arginine deiminase immunization induces anticitrullinated protein antibodies in mice

with particular MHC types. Proceedings of the National Academy of Sciences of the United States of America 2017; 114:E10169-E77.

97. Okada Y, Wu D, Trynka G, Raj T, Terao C, Ikari K, et al. Genetics of rheumatoid arthritis contributes to biology and drug discovery. Nature 2014; 506:376-+.

98. van der Woude D, Houwing-Duistermaat JJ, Toes REM, Huizinga TWJ, Thomson W, Worthington J, et al. Quantitative Heritability of Anti-Citrullinated Protein Antibody-Positive and Anti-Citrullinated Protein Antibody-Negative Rheumatoid Arthritis. Arthritis and Rheumatism 2009; 60:916-23.

99. Burton PR, Clayton DG, Cardon LR, Craddock N, Deloukas P, Duncanson A, et al. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. Nature 2007; 447:661-78.

100. Okada Y, Eyre S, Suzuki A, Kochi Y, Yamamoto K. Genetics of rheumatoid arthritis: 2018 status. Annals of the Rheumatic Diseases 2019; 78:446-53.

101. Padyukov L, Seielstad M, Ong RTH, Ding B, Ronnelid J, Seddighzadeh M, et al. A genome-wide association study suggests contrasting associations in ACPA-positive versus ACPA-negative rheumatoid arthritis. Annals of the Rheumatic Diseases 2011; 70:259-65.

102. Bossini-Castillo L, de Kovel C, Kallberg H, van 't Slot R, Italiaander A, Coenen M, et al. A genome-wide association study of rheumatoid arthritis without antibodies against citrullinated peptides. Annals of the Rheumatic Diseases 2015; 74:10.

103. Terao C, Raychaudhuri S, Gregersen PK. Recent Advances in Defining the Genetic Basis of Rheumatoid Arthritis. In: Chakravarti A, Green E, editors. Annual Review of Genomics and Human Genetics, Vol 17. Palo Alto: Annual Reviews; 2016. p. 273-301.

104. Viatte S, Plant D, Bowes J, Lunt M, Eyre S, Barton A, et al. Genetic markers of rheumatoid arthritis susceptibility in anti-citrullinated peptide antibody negative patients. Annals of the Rheumatic Diseases 2012; 71:1984-90.

105. Viatte S, Massey J, Bowes J, Duffus K, Eyre S, Barton A, et al. Replication of Associations of Genetic Loci Outside the HLA Region With Susceptibility to Anti-Cyclic Citrullinated Peptide-Negative Rheumatoid Arthritis. Arthritis & Rheumatology 2016; 68:1603-13.

106. Cotsapas C, Voight BF, Rossin E, Lage K, Neale BM, Wallace C, et al. Pervasive Sharing of Genetic Effects in Autoimmune Disease. Plos Genetics 2011; 7:8.

107. Kreiner E, Waage J, Standl M, Brix S, Pers TH, Alves AC, et al. Shared genetic variants suggest common pathways in allergy and autoimmune diseases. Journal of Allergy and Clinical Immunology 2017; 140:771-81.

108. Eaton WW, Rose NR, Kalaydjian A, Pedersen MG, Mortensen PB. Epidemiology of autoimmune diseases in Denmark. Journal of Autoimmunity 2007; 29:1-9.

109. Wing K, Onishi Y, Prieto-Martin P, Yamaguchi T, Miyara M, Fehervari Z, et al. CTLA-4 control over Foxp3(+) regulatory T cell function. Science 2008; 322:271-5.

110. Fontenot JD, Rasmussen JP, Gavin MA, Rudensky AY. A function for interleukin 2 in Foxp3-expressing regulatory T cells. Nature Immunology 2005; 6:1142-51.

111. Yamazaki T, Yang XO, Chung Y, Fukunaga A, Nurieva R, Pappu B, et al. CCR6 Regulates the Migration of Inflammatory and Regulatory T Cells. Journal of Immunology 2008; 181:8391-401.

112. Mathur AN, Chang HC, Zisoulis DG, Stritesky GL, Yu Q, O'Malley JT, et al. Stat3 and Stat4 direct development of IL-17-secreting Th cells. Journal of Immunology 2007; 178:4901-7.

113. Maurano MT, Humbert R, Rynes E, Thurman RE, Haugen E, Wang H, et al. Systematic Localization of Common Disease-Associated Variation in Regulatory DNA. Science 2012; 337:1190-5.

114. Vahedi G, Kanno Y, Furumoto Y, Jiang K, Parker SCJ, Erdos MR, et al. Super-enhancers delineate disease-associated regulatory nodes in T cells. Nature 2015; 520:558-+.

115. Hnisz D, Abraham BJ, Lee TI, Lau A, Saint-Andre V, Sigova AA, et al. Super-Enhancers in the Control of Cell Identity and Disease. Cell 2013; 155:934-47.

116. Trynka G, Sandor C, Han B, Xu H, Stranger BE, Liu XS, et al. Chromatin marks identify critical cell types for fine mapping complex trait variants. Nature Genetics 2013; 45:124-30.

117. Soskic B, Cano-Gamez E, Smyth DJ, Rowan WC, Nakic N, Esparza-Gordillo J, et al. Chromatin activity at GWAS loci identifies T cell states driving complex immune diseases. Nature Genetics 2019; 51:1486-+.

118. Hu XL, Kim H, Raj T, Brennan PJ, Trynka G, Teslovich N, et al. Regulation of Gene Expression in Autoimmune Disease Loci and the Genetic Basis of Proliferation in CD4(+) Effector Memory T Cells. Plos Genetics 2014; 10:13.

119. Finucane HK, Reshef YA, Anttila V, Slowikowski K, Gusev A, Byrnes A, et al. Heritability enrichment of specifically expressed genes identifies disease-relevant tissues and cell types. Nature Genetics 2018; 50:621-+.

120. Schoenfelder S, Fraser P. Long-range enhancer–promoter contacts in gene expression control. Nature Reviews Genetics 2019.

121. Javierre BM, Burren OS, Wilder SP, Kreuzhuber R, Hill SM, Sewitz S, et al. Lineage-Specific Genome Architecture Links Enhancers and Non-coding Disease Variants to Target Gene Promoters. Cell 2016; 167:1369-+.

122. Mumbach MR, Satpathy AT, Boyle EA, Dai C, Gowen BG, Cho SW, et al. Enhancer connectome in primary human cells identifies target genes of disease-associated DNA elements. Nature Genetics 2017; 49:1602-+.

123. Albert FW, Kruglyak L. The role of regulatory variation in complex traits and disease. Nature Reviews Genetics 2015; 16:197-212.

124. Croteau-Chonka DC, Rogers AJ, Raj T, McGeachie MJ, Qiu WL, Ziniti JP, et al. Expression Quantitative Trait Loci Information Improves Predictive Modeling of Disease Relevance of Non-Coding Genetic Variation. Plos One 2015; 10:20.

125. Dimas AS, Deutsch S, Stranger BE, Montgomery SB, Borel C, Attar-Cohen H, et al. Common Regulatory Variation Impacts Gene Expression in a Cell Type-Dependent Manner. Science 2009; 325:1246-50.

126. Gutierrez-Arcelus M, Ongen H, Lappalainen T, Montgomery SB, Buil A, Yurovsky A, et al. Tissue-Specific Effects of Genetic and Epigenetic Variation on Gene Regulation and Splicing. Plos Genetics 2015; 11:25.

127. Chen L, Ge B, Casale FP, Vasquez L, Kwan T, Garrido-Martin D, et al. Genetic Drivers of Epigenetic and Transcriptional Variation in Human Immune Cells. Cell 2016; 167:1398-+.

128. Fairfax BP, Makino S, Radhakrishnan J, Plant K, Leslie S, Dilthey A, et al. Genetics of gene expression in primary immune cells identifies cell type-specific master regulators and roles of HLA alleles. Nature Genetics 2012; 44:502-+.

129. Fu JY, Wolfs MGM, Deelen P, Westra HJ, Fehrmann RSN, Meerman GJT, et al. Unraveling the Regulatory Mechanisms Underlying Tissue-Dependent Genetic Variation of Gene Expression. Plos Genetics 2012; 8:14.

130. Gerrits A, Li Y, Tesson BM, Bystrykh LV, Weersing E, Ausema A, et al. Expression Quantitative Trait Loci Are Highly Sensitive to Cellular Differentiation State. Plos Genetics 2009; 5:8.

131. Lonsdale J, Thomas J, Salvatore M, Phillips R, Lo E, Shad S, et al. The Genotype-Tissue Expression (GTEx) project. Nature Genetics 2013; 45:580-5.

132. Peters JE, Lyons PA, Lee JC, Richard AC, Fortune MD, Newcombe PJ, et al. Insight into Genotype-Phenotype Associations through eQTL Mapping in Multiple Cell Types in Health and Immune-Mediated Disease. Plos Genetics 2016; 12:29.

133. Zhernakova DV, Deelen P, Vermaat M, van Iterson M, van Galen M, Arindrarto W, et al. Identification of context-dependent expression quantitative trait loci in whole blood. Nature Genetics 2017; 49:139-45.

134. Kim-Hellmuth S, Bechheim M, Putz B, Mohammadi P, Nedelec Y, Giangreco N, et al. Genetic regulatory effects modified by immune activation contribute to autoimmune disease associations. Nature Communications 2017; 8:10.

135. Kasela S, Kisand K, Tserel L, Kaleviste E, Remm A, Fischer K, et al. Pathogenic implications for autoimmune mechanisms derived by comparative eQTL analysis of CD4(+) versus CD8(+) T cells. Plos Genetics 2017; 13:21.

136. Ishigaki K, Kochi Y, Suzuki A, Tsuchida Y, Tsuchiya H, Sumitomo S, et al. Polygenic burdens on cell-specific pathways underlie the risk of rheumatoid arthritis. Nature Genetics 2017; 49:1120-+.

137. Walsh AM, Whitaker JW, Huang CC, Cherkas Y, Lamberth SL, Brodmerkel C, et al. Integrative genomic deconvolution of rheumatoid arthritis GWAS loci into gene and cell type associations. Genome Biology 2016; 17:16.

138. Gate RE, Cheng CS, Aiden AP, Siba A, Tabaka M, Lituiev D, et al. Genetic determinants of co-accessible chromatin regions in activated T cells across humans. Nature Genetics 2018; 50:1140-+.

139. Thalayasingam N, Nair N, Skelton AJ, Massey J, Anderson AE, Clark AD, et al. CD4+and B Lymphocyte Expression Quantitative Traits at Rheumatoid Arthritis Risk Loci in Patients With Untreated Early Arthritis Implications for Causal Gene Identification. Arthritis & Rheumatology 2018; 70:361-70.

140. Spiliopoulou A, Colombo M, Plant D, Nair N, Cui J, Coenen MJH, et al. Association of response to TNF inhibitors in rheumatoid arthritis with quantitative trait loci for <em>CD40</em> and CD39. Annals of the Rheumatic Diseases 2019:annrheumdis-2018-214877.

141. Silman AJ, Newman J, MacGregor AJ. Cigarette smoking increases the risk of rheumatoid arthritis - Results from a nationwide study of disease-discordant twins. Arthritis and Rheumatism 1996; 39:732-5.

142. Stolt P, Bengtsson C, Nordmark B, Lindblad S, Lundberg I, Klareskog L, et al. Quantification of the influence of cigarette smoking on rheumatoid arthritis: results

from a population based case-control study, using incident cases. Annals of the Rheumatic Diseases 2003; 62:835-41.

143. Kallberg H, Ding B, Padyukov L, Bengtsson C, Ronnelid J, Klareskog L, et al. Smoking is a major preventable risk factor for rheumatoid arthritis: estimations of risks after various exposures to cigarette smoke. Annals of the Rheumatic Diseases 2011; 70:508-11.

144. Klareskog L, Stolt P, Lundberg K, Kallberg H, Bengtsson C, Grunewald J, et al. A new model for an etiology of rheumatoid arthritis. Arthritis and Rheumatism 2006; 54:38-46.

145. Lundstrom E, Kallberg H, Alfredsson L, Klareskog L, Padyukov L. Gene-Environment Interaction Between the DRB1 Shared Epitope and Smoking in the Risk of Anti-Citrullinated Protein Antibody-Positive Rheumatoid Arthritis. Arthritis and Rheumatism 2009; 60:1597-603.

146. Makrygiannakis D, Hermansson M, Ulfgren AK, Nicholas AP, Zendman AJW, Eklund A, et al. Smoking increases peptidylarginine deiminase 2 enzyme expression in human lungs and increases citrullination in BAL cells. Annals of the Rheumatic Diseases 2008; 67:1488-92.

147. Reynisdottir G, Karimi R, Joshua V, Olsen H, Hensvold AH, Harju A, et al. Structural Changes and Antibody Enrichment in the Lungs Are Early Features of Anti-Citrullinated Protein Antibody-Positive Rheumatoid Arthritis. Arthritis & Rheumatology 2014; 66:31-9.

148. Reynisdottir G, Olsen H, Joshua V, Engstrom M, Forsslund H, Karimi R, et al. Signs of immune activation and local inflammation are present in the bronchial tissue of patients with untreated early rheumatoid arthritis. Annals of the Rheumatic Diseases 2016; 75:1722-7.

149. de Pablo P, Dietrich T, McAlindon TE. Association of periodontal disease and tooth loss with rheumatoid arthritis in the US population. Journal of Rheumatology 2008; 35:70-6.

150. McGraw WT, Potempa J, Farley D, Travis J. Purification, characterization, and sequence analysis of a potential virulence factor from Porphyromonas gingivalis, peptidylarginine deiminase. Infection and Immunity 1999; 67:3248-56.

151. Wegner N, Wait R, Sroka A, Eick S, Nguyen KA, Lundberg K, et al. Peptidylarginine Deiminase From Porphyromonas gingivalis Citrullinates Human Fibrinogen and alpha-Enolase Implications for Autoimmunity in Rheumatoid Arthritis. Arthritis and Rheumatism 2010; 62:2662-72.

152. Kharlamova N, Jiang X, Sherina N, Potempa B, Israelsson L, Quirke AM, et al. Antibodies to Porphyromonas gingivalis Indicate Interaction Between Oral Infection, Smoking, and Risk Genes in Rheumatoid Arthritis Etiology. Arthritis & Rheumatology 2016; 68:604-13.

153. Konig MF, Abusleme L, Reinholdt J, Palmer RJ, Teles RP, Sampson K, et al. Aggregatibacter actinomycetemcomitans-induced hypercitrullination links periodontal infection to autoimmunity in rheumatoid arthritis. Science Translational Medicine 2016; 8:12.

154. Blasco-Baque V, Garidou L, Pomie C, Escoula Q, Loubieres P, Le Gall-David S, et al. Periodontitis induced by Porphyromonas gingivalis drives periodontal microbiota dysbiosis and insulin resistance via an impaired adaptive immune response. Gut 2017; 66:872-85.

155. Eriksson K, Nise L, Kats A, Luttropp E, Catrina AI, Askling J, et al. Prevalence of Periodontitis in Patients with Established Rheumatoid Arthritis: A Swedish Population Based Case-Control Study. Plos One 2016; 11:16.

156. Zhang X, Zhang DY, Jia HJ, Feng Q, Wang DH, Liang D, et al. The oral and gut microbiomes are perturbed in rheumatoid arthritis and partly normalized after treatment. Nature Medicine 2015; 21:895-905.

157. Zhu H, Wang GH, Qian J. Transcription factors as readers and effectors of DNA methylation. Nature Reviews Genetics 2016; 17:551-65.

158. Bird A. DNA methylation patterns and epigenetic memory. Genes & Development 2002; 16:6-21.

159. Wu H, Zhang Y. Reversing DNA Methylation: Mechanisms, Genomics, and Biological Functions. Cell 2014; 156:45-68.

160. Tahiliani M, Koh KP, Shen YH, Pastor WA, Bandukwala H, Brudno Y, et al. Conversion of 5-Methylcytosine to 5-Hydroxymethylcytosine in Mammalian DNA by MLL Partner TET1. Science 2009; 324:930-5.

161. Ito S, Shen L, Dai Q, Wu SC, Collins LB, Swenberg JA, et al. Tet Proteins Can Convert 5-Methylcytosine to 5-Formylcytosine and 5-Carboxylcytosine. Science 2011; 333:1300-3.

162. von Meyenn F, Iurlaro M, Habibi E, Liu NQ, Salehzadeh-Yazdi A, Santos F, et al. Impairment of DNA Methylation Maintenance Is the Main Cause of Global Demethylation in Naive Embryonic Stem Cells. Molecular Cell 2016; 62:848-61.

163. Okano M, Bell DW, Haber DA, Li E. DNA methyltransferases Dnmt3a and Dnmt3b are essential for de novo methylation and mammalian development. Cell 1999; 99:247-57.

164. Aapola U, Shibuya K, Scott HS, Ollila J, Vihinen M, Heino M, et al. Isolation and initial characterization of a novel zinc finger gene, DNMT3L, on 21q22.3, related to the cytosine-5-methyltransferase 3 gene family. Genomics 2000; 65:293-8.

165. Chedin F, Lieber MR, Hsieh CL. The DNA methyltransferase-like protein DNMT3L stimulates de novo methylation by Dnmt3a. Proceedings of the National Academy of Sciences of the United States of America 2002; 99:16916-21.

166. Wienholz BL, Kareta MS, Moarefi AH, Gordon CA, Ginno PA, Chedin F. DNMT3L Modulates Significant and Distinct Flanking Sequence Preference for DNA Methylation by DNMT3A and DNMT3B In Vivo. Plos Genetics 2010; 6:15.

167. Bestor T, Laudano A, Mattaliano R, Ingram V. Cloning and sequencing of a cDNA-encoding DNA methyltransferase of mouse cells - the carboxyl-terminal domain of the mammalian enzymes is related to bacterial restriction methyltransferases. Journal of Molecular Biology 1988; 203:971-83.

168. Hermann A, Goyal R, Jeltsch A. The Dnmt1 DNA-(cytosine-C5)-methyltransferase methylates DNA processively with high preference for hemimethylated target sites. Journal of Biological Chemistry 2004; 279:48350-9.

169. Bostick M, Kim JK, Esteve PO, Clark A, Pradhan S, Jacobsen SE. UHRF1 plays a role in maintaining DNA methylation in mammalian cells. Science 2007; 317:1760-4.

170. Seisenberger S, Peat JR, Reik W. Conceptual links between DNA methylation reprogramming in the early embryo and primordial germ cells. Current Opinion in Cell Biology 2013; 25:281-8.

171. Yue XJ, Lio CWJ, Samaniego-Castruita D, Li X, Rao A. Loss of TET2 and TET3 in regulatory T cells unleashes effector function. Nature Communications 2019; 10:14.

172. Bannister AJ, Kouzarides T. Regulation of chromatin by histone modifications. Cell Research 2011; 21:381-95.

173. Kundaje A, Meuleman W, Ernst J, Bilenky M, Yen A, Heravi-Moussavi A, et al. Integrative analysis of 111 reference human epigenomes. Nature 2015; 518:317-30.

174. Ai RZ, Laragione T, Hammaker D, Boyle DL, Wildberg A, Maeshima K, et al. Comprehensive epigenetic landscape of rheumatoid arthritis fibroblast-like synoviocytes. Nature Communications 2018; 9:11.

175. Araki Y, Wada TT, Aizaki Y, Sato K, Yokota K, Fujimoto K, et al. Histone Methylation and STAT-3 Differentially Regulate Interleukin-6-Induced Matrix

Metalloproteinase Gene Activation in Rheumatoid Arthritis Synovial Fibroblasts. Arthritis & Rheumatology 2016; 68:1111-23.

176. Angiolilli C, Kabala PA, Grabiec AM, Van Baarsen IM, Ferguson BS, Garcia S, et al. Histone deacetylase 3 regulates the inflammatory gene expression programme of rheumatoid arthritis fibroblast-like synoviocytes. Annals of the Rheumatic Diseases 2017; 76:277-85.

177. Esteller M. Non-coding RNAs in human disease. Nature Reviews Genetics 2011; 12:861-74.

178. Moran-Moguel MC, Petarra-del Rio S, Mayorquin-Galvan EE, Zavala-Cerna MG. Rheumatoid Arthritis and miRNAs: A Critical Review through a Functional View. Journal of Immunology Research 2018:16.

179. Li JY, Wan Y, Guo QY, Zou LY, Zhang JY, Fang YF, et al. Altered microRNA expression profile with miR-146a upregulation in CD4(+) T cells from patients with rheumatoid arthritis. Arthritis Research & Therapy 2010; 12:12.

180. Niederer F, Trenkmann M, Ospelt C, Karouzakis E, Neidhart M, Stanczyk J, et al. Down-regulation of microRNA-34a* in rheumatoid arthritis synovial fibroblasts promotes apoptosis resistance. Arthritis and Rheumatism 2012; 64:1771-9.

181. Zhang Y, Xu YZ, Sun N, Liu JH, Chen FF, Guan XL, et al. Long noncoding RNA expression profile in fibroblast-like synoviocytes from patients with rheumatoid arthritis. Arthritis Research & Therapy 2016; 18:10.

182. Gao YZ, Li SS, Zhang ZJ, Yu XH, Zheng JF. The Role of Long Non-coding RNAs in the Pathogenesis of RA, SLE, and SS. Frontiers in Medicine 2018; 5:14.

183. Li E, Zhang Y. DNA Methylation in Mammals. Cold Spring Harbor Perspectives in Biology 2014; 6:21.

184. Jones PA. Functions of DNA methylation: islands, start sites, gene bodies and beyond. Nature Reviews Genetics 2012; 13:484-92.

185. Ball MP, Li JB, Gao Y, Lee JH, LeProust EM, Park IH, et al. Targeted and genome-scale strategies reveal gene-body methylation signatures in human cells. Nature Biotechnology 2009; 27:361-8.

186. Gutierrez-Arcelus M, Lappalainen T, Montgomery SB, Buil A, Ongen H, Yurovsky A, et al. Passive and active DNA methylation and the interplay with genetic variation in gene regulation. Elife 2013; 2:18.

187. Schultz MD, He YP, Whitaker JW, Hariharan M, Mukamel EA, Leung D, et al. Human body epigenome maps reveal noncanonical DNA methylation variation. Nature 2015; 523:212-U189.

188. Pacis A, Mailhot-Leonard F, Tailleux L, Randolph HE, Yotova V, Dumaine A, et al. Gene activation precedes DNA demethylation in response to infection in human dendritic cells. Proceedings of the National Academy of Sciences of the United States of America 2019; 116:6938-43.

189. Domcke S, Bardet AF, Ginno PA, Hartl D, Burger L, Schubeler D. Competition between DNA methylation and transcription factors determines binding of NRF1. Nature 2015; 528:575-+.

190. Yin YM, Morgunova E, Jolma A, Kaasinen E, Sahu B, Khund-Sayeed S, et al. Impact of cytosine methylation on DNA binding specificities of human transcription factors. Science 2017; 356:15.

191. Zuo Z, Roy B, Chang YK, Granas D, Stormo GD. Measuring quantitative effects of methylation on transcription factor-DNA binding affinity. Science Advances 2017; 3:11.

192. Kribelbauer JF, Laptenko O, Chen SY, Martini GD, Freed-Pastor WA, Prives C, et al. Quantitative Analysis of the DNA Methylation Sensitivity of Transcription Factor Complexes. Cell Reports 2017; 19:2383-95.

193. Stadler MB, Murr R, Burger L, Ivanek R, Lienert F, Scholer A, et al. DNA-binding factors shape the mouse methylome at distal regulatory regions. Nature 2011; 480:490-5.

194. Bonder MJ, Luijk R, Zhernakova DV, Moed M, Deelen P, Vermaat M, et al. Disease variants alter transcription factor levels and methylation of their binding sites. Nature Genetics 2017; 49:131-8.

195. Stroud H, Feng SH, Kinney SM, Pradhan S, Jacobsen SE. 5-Hydroxymethylcytosine is associated with enhancers and gene bodies in human embryonic stem cells. Genome Biology 2011; 12:8.

196. Campanero MR, Armstrong MI, Flemington EK. CPG methylation as a mechanism for the regulation of E2F activity. Proceedings of the National Academy of Sciences of the United States of America 2000; 97:6481-6.

197. Tate PH, Bird AP. Effects of DNA methylation on DNA-binding proteins and gene-expression. Current Opinion in Genetics & Development 1993; 3:226-31.

198. Hendrich B, Bird A. Identification and characterization of a family of mammalian methyl-CpG binding proteins. Molecular and Cellular Biology 1998; 18:6538-47.

199. Roloff TC, Ropers HH, Nuber UA. Comparative study of methyl-CpG-binding domain proteins. Bmc Genomics 2003; 4:9.

200. Laget S, Joulie M, Le Masson F, Sasai N, Christians E, Pradhan S, et al. The Human Proteins MBD5 and MBD6 Associate with Heterochromatin but They Do Not Bind Methylated DNA. Plos One 2010; 5:11.

201. Prokhortchouk A, Hendrich B, Jorgensen H, Ruzov A, Wilm M, Georgiev G, et al. The p120 catenin partner Kaiso is a DNA methylation-dependent transcriptional repressor. Genes & Development 2001; 15:1613-8.

202. Filion GJP, Zhenilo S, Salozhin S, Yamada D, Prokhortchouk E, Defossez PA. A family of human zinc finger proteins that bind methylated DNA and repress transcription. Molecular and Cellular Biology 2006; 26:169-81.

203. Unoki M, Nishidate T, Nakamura Y. ICBP90, an E2F-1 target, recruits HDAC1 and binds to methyl-CpG through its SRA domain. Oncogene 2004; 23:7601-10.

204. Johnson LM, Bostick M, Zhang XY, Kraft E, Henderson I, Callis J, et al. The SRA methyl-cytosine-binding domain links DNA and histone methylation. Current Biology 2007; 17:379-84.

205. Nan XS, Ng HH, Johnson CA, Laherty CD, Turner BM, Eisenman RN, et al. Transcriptional repression by the methyl-CpG-binding protein MeCP2 involves a histone deacetylase complex. Nature 1998; 393:386-9.

206. Fuks F, Hurd PJ, Wolf D, Nan XS, Bird AP, Kouzarides T. The Methyl-CpG-binding protein MeCP2 links DNA methylation to histone methylation. Journal of Biological Chemistry 2003; 278:4035-40.

207. Maunakea AK, Chepelev I, Cui KR, Zhao KJ. Intragenic DNA methylation modulates alternative splicing by recruiting MeCP2 to promote exon recognition. Cell Research 2013; 23:1256-69.

208. Laird PW. Principles and challenges of genome-wide DNA methylation analysis. Nature Reviews Genetics 2010; 11:191-203.

209. Wang J, Zhuang JL, Iyer S, Lin XY, Whitfield TW, Greven MC, et al. Sequence features and chromatin structure around the genomic regions bound by 119 human transcription factors. Genome Research 2012; 22:1798-812.

210. Gerstein MB, Kundaje A, Hariharan M, Landt SG, Yan KK, Cheng C, et al. Architecture of the human regulatory network derived from ENCODE data. Nature 2012; 489:91-100.

211. Illumina MethylationEPIC BeadChip Data Sheet. 2015.

212. Pidsley R, Zotenko E, Peters TJ, Lawrence MG, Risbridger GP, Molloy P, et al. Critical evaluation of the Illumina MethylationEPIC BeadChip microarray for whole-genome DNA methylation profiling. Genome Biology 2016; 17:17.

213. Ligthart S, Marzi C, Aslibekyan S, Mendelson MM, Conneely KN, Tanaka T, et al. DNA methylation signatures of chronic low-grade inflammation are associated with complex diseases. Genome Biology 2016; 17:15.

214. Ventham NT, Kennedy NA, Adams AT, Kalla R, Heath S, O'Leary KR, et al. Integrative epigenome-wide analysis demonstrates that DNA methylation may mediate genetic risk in inflammatory bowel disease. Nature Communications 2016; 7:14.

215. Richardson B. Effect of an inhibitor of dna methylation on t-cells .2. 5-azacytidine induces self-reactivity in antigen-specific T4+-cells. Human Immunology 1986; 17:456-70.

216. Richardson B, Scheinbart L, Strahler J, Gross L, Hanash S, Johnson M. Evidence for impaired T-cell DNA methylation in systemic lupus-erythematosus and rheumatoid-arthritis. Arthritis and Rheumatism 1990; 33:1665-73.

217. Nile CJ, Read RC, Akil M, Duff GW, Wilson AG. Methylation status of a single CpG site in the IL6 promoter is related to IL6 messenger RNA levels and rheumatoid arthritis. Arthritis and Rheumatism 2008; 58:2686-93.

218. Liu Y, Aryee MJ, Padyukov L, Fallin MD, Hesselberg E, Runarsson A, et al. Epigenome-wide association data implicate DNA methylation as an intermediary of genetic risk in rheumatoid arthritis. Nature Biotechnology 2013; 31:142-7.

219. Houseman EA, Accomando WP, Koestler DC, Christensen BC, Marsit CJ, Nelson HH, et al. DNA methylation arrays as surrogate measures of cell mixture distribution. Bmc Bioinformatics 2012; 13:16.

220. Glossop JR, Emes RD, Nixon NB, Haworth KE, Packham JC, Dawes PT, et al. Genome-wide DNA methylation profiling in rheumatoid arthritis identifies disease-associated methylation changes that are distinct to individual T- and B-lymphocyte populations. Epigenetics 2014; 9:1228-37.

221. Guo SC, Zhu Q, Jiang T, Wang RS, Shen Y, Zhu X, et al. Genome-wide DNA methylation patterns in CD4+T cells from Chinese Han patients with rheumatoid arthritis. Modern Rheumatology 2017; 27:441-7.

222. Julia A, Absher D, Lopez-Lasanta M, Palau N, Pluma A, Jones LW, et al. Epigenome-wide association study of rheumatoid arthritis identifies differentially methylated loci in B cells. Human Molecular Genetics 2017; 26:2803-11.

223. Plant D, Webster A, Nair N, Oliver J, Smith SL, Eyre S, et al. Differential Methylation as a Biomarker of Response to Etanercept in Patients With Rheumatoid Arthritis. Arthritis & Rheumatology 2016; 68:1353-60.

224. Bartok B, Firestein GS. Fibroblast-like synoviocytes: key effector cells in rheumatoid arthritis. Immunological Reviews 2010; 233:233-55.

225. Karouzakis E, Gay RE, Michel BA, Gay S, Neidhart M. DNA Hypomethylation in Rheumatoid Arthritis Synovial Fibroblasts. Arthritis and Rheumatism 2009; 60:3613-22.

226. Karouzakis E, Rengel Y, Jungel A, Kolling C, Gay RE, Michel BA, et al. DNA methylation regulates the expression of CXCL12 in rheumatoid arthritis synovial fibroblasts. Genes and Immunity 2011; 12:643-52.

227. Nakano K, Whitaker JW, Boyle DL, Wang W, Firestein GS. DNA methylome signature in rheumatoid arthritis. Annals of the Rheumatic Diseases 2013; 72:110-7.

228. Ai R, Hammaker D, Boyle DL, Morgan R, Walsh AM, Fan SC, et al. Joint-specific DNA methylation and transcriptome signatures in rheumatoid arthritis identify distinct pathogenic processes. Nature Communications 2016; 7:9.

229. Frank-Bertoncelj M, Trenkmann M, Klein K, Karouzakis E, Rehrauer H, Bratus A, et al. Epigenetically-driven anatomical diversity of synovial fibroblasts guides joint-specific fibroblast functions. Nature Communications 2017; 8:14.

230. Webster AP, Plant D, Ecker S, Zufferey F, Bell JT, Feber A, et al. Increased DNA methylation variability in rheumatoid arthritis-discordant monozygotic twins. Genome Medicine 2018; 10:12.

231. Paul DS, Teschendorff AE, Dang MAN, Lowe R, Hawa MI, Ecker S, et al. Increased DNA methylation variability in type 1 diabetes across three immune effector cell types. Nature Communications 2016; 7:11.

232. Souren NY, Gerdes LA, Lutsik P, Gasparoni G, Beltran E, Salhab A, et al. DNA methylation signatures of monozygotic twins clinically discordant for multiple sclerosis. Nature Communications 2019; 10:12.

233. Breton CV, Byun HM, Wenten M, Pan F, Yang A, Gilliland FD. Prenatal Tobacco Smoke Exposure Affects Global and Gene-specific DNA Methylation. American Journal of Respiratory and Critical Care Medicine 2009; 180:462-7.

234. Breitling LP, Yang RX, Korn B, Burwinkel B, Brenner H. Tobacco-Smoking-Related Differential DNA Methylation: 27K Discovery and Replication. American Journal of Human Genetics 2011; 88:450-7.

235. Zeilinger S, Kuhnel B, Klopp N, Baurecht H, Kleinschmidt A, Gieger C, et al. Tobacco Smoking Leads to Extensive Genome-Wide Changes in DNA Methylation. Plos One 2013; 8:14.

236. Ambatipudi S, Cuenin C, Hernandez-Vargas H, Ghantous A, Le Calvez-Kelm F, Kaaks R, et al. Tobacco smoking-associated genome-wide DNA methylation changes in the EPIC study. Epigenomics 2016; 8:599-618.

237. Vento-Tormo R, Company C, Rodriguez-Ubreva J, de la Rica L, Urquiza JM, Javierre BM, et al. IL-4 orchestrates STAT6-mediated DNA demethylation leading to dendritic cell differentiation. Genome Biology 2016; 17:18.

238. Gibbs JR, van der Brug MP, Hernandez DG, Traynor BJ, Nalls MA, Lai SL, et al. Abundant Quantitative Trait Loci Exist for DNA Methylation and Gene Expression in Human Brain. Plos Genetics 2010; 6:13.

239. Zhang DD, Cheng LJ, Badner JA, Chen C, Chen Q, Luo W, et al. Genetic Control of Individual Differences in Gene-Specific Methylation in Human Brain. American Journal of Human Genetics 2010; 86:411-9.

240. Hannon E, Spiers H, Viana J, Pidsley R, Burrage J, Murphy TM, et al. Methylation QTLs in the developing brain and their enrichment in schizophrenia risk loci. Nature Neuroscience 2016; 19:48-+.

241. Schulz H, Ruppert AK, Herms S, Wolf C, Mirza-Schreiber N, Stegle O, et al. Genome-wide mapping of genetic determinants influencing DNA methylation and gene expression in human hippocampus. Nature Communications 2017; 8:11.

242. Volkov P, Olsson AH, Gillberg L, Jorgensen SW, Brons C, Eriksson KF, et al. A Genome-Wide mQTL Analysis in Human Adipose Tissue Identifies Genetic Variants Associated with DNA Methylation, Gene Expression and Metabolic Traits. Plos One 2016; 11:31.

243. Grundberg E, Meduri E, Sandling JK, Hedman AK, Keildson S, Buil A, et al. Global Analysis of DNA Methylation Variation in Adipose Tissue from Twins Reveals Links to Disease-Associated Variants in Distal Regulatory Elements. American Journal of Human Genetics 2013; 93:876-90.

244. Olsson AH, Volkov P, Bacos K, Dayeh T, Hall E. Genome-Wide Associations between Genetic and Epigenetic Variation Influence mRNA Expression and Insulin Secretion in Human Pancreatic Islets (vol 10, e1004735, 2014). Plos Genetics 2014; 10:4.

245. Hannon E, Gorrie-Stone TJ, Smart MC, Burrage J, Hughes A, Bao YC, et al. Leveraging DNA-Methylation Quantitative-Trait Loci to Characterize the Relationship between Methylomic Variation, Gene Expression, and Complex Traits. American Journal of Human Genetics 2018; 103:654-65.

246. Gaunt TR, Shihab HA, Hemani G, Min JL, Woodward G, Lyttleton O, et al. Systematic identification of genetic influences on methylation across the human life course. Genome Biology 2016; 17:14.

247. Pierce BL, Tong L, Argos M, Demanelis K, Jasmine F, Rakibuz-Zaman M, et al. Co-occurring expression and methylation QTLs allow detection of common causal variants and shared biological mechanisms. Nature Communications 2018; 9:12.

248. Imgenberg-Kreuz J, Almlof JC, Leonard D, Alexsson A, Nordmark G, Eloranta ML, et al. DNA methylation mapping identifies gene regulatory effects in patients with systemic lupus erythematosus. Annals of the Rheumatic Diseases 2018; 77:736-43.

249. Wagner JR, Busche S, Ge B, Kwan T, Pastinen T, Blanchette M. The relationship between DNA methylation, genetic and expression inter-individual variation in untransformed human fibroblasts. Genome Biology 2014; 15:17.

250. Day K, Waite LL, Alonso A, Irvin MR, Zhi D, Thibeault KS, et al. Heritable DNA Methylation in CD4(+) Cells among Complex Families Displays Genetic and Non-Genetic Effects. Plos One 2016; 11:20.

251. Banovich NE, Lan X, McVicker G, van de Geijn B, Degner JF, Blischak JD, et al. Methylation QTLs Are Associated with Coordinated Changes in Transcription Factor Binding, Histone Modifications, and Gene Expression Levels. Plos Genetics 2014; 10:12.

252. van Eijk KR, de Jong S, Boks MPM, Langeveld T, Colas F, Veldink JH, et al. Genetic analysis of DNA methylation and gene expression levels in whole blood of healthy human subjects. Bmc Genomics 2012; 13:13.

253. Morrow JD, Glass K, Cho MH, Hersh CP, Pinto-Plata V, Celli B, et al. Human Lung DNA Methylation Quantitative Trait Loci Colocalize with Chronic Obstructive Pulmonary Disease Genome-Wide Association Loci. American Journal of Respiratory and Critical Care Medicine 2018; 197:1275-84.

254. Do C, Lang CF, Lin J, Darbary H, Krupska I, Gaba A, et al. Mechanisms and Disease Associations of Haplotype-Dependent Allele-Specific DNA Methylation. American Journal of Human Genetics 2016; 98:934-55.

255. Boumber YA, Kondo Y, Chen X, Shen L, Guo Y, Tellez C, et al. An Sp1/Sp3 Binding Polymorphism Confers Methylation Protection. Plos Genetics 2008; 4:10.

256. Gamazon ER, Badner JA, Cheng L, Zhang C, Zhang D, Cox NJ, et al. Enrichment of cis-regulatory gene expression SNPs and methylation quantitative trait loci among bipolar disorder susceptibility variants. Molecular Psychiatry 2013; 18:340-6.

257. McRae AF, Marioni RE, Shah S, Yang J, Powell JE, Harris SE, et al. Identification of 55,000 Replicated DNA Methylation QTL. Scientific Reports 2018; 8:9.

258. Ziller MJ, Gu HC, Muller F, Donaghey J, Tsai LTY, Kohlbacher O, et al. Charting a dynamic DNA methylation landscape of the human genome. Nature 2013; 500:477-81.

259. Taylor DL, Jackson AU, Narisu N, Hemani G, Erdos MR, Chines PS, et al. Integrative analysis of gene expression, DNA methylation, physiological traits, and genetic variation in human skeletal muscle. Proceedings of the National Academy of Sciences of the United States of America 2019; 116:10883-8.

260. Millstein J, Zhang B, Zhu J, Schadt EE. Disentangling molecular relationships with a causal inference test. Bmc Genetics 2009; 10:15.

261. Davey Smith G, Hemani G. Mendelian randomization: genetic anchors for causal inference in epidemiological studies. Human Molecular Genetics 2014; 23:R89-R98.

262. Ritchlin CT, Colbert RA, Gladman DD. Psoriatic Arthritis. New England Journal of Medicine 2017; 376:957-70.

263. Veale DJ, Fearon U. Psoriatic arthritis 1 The pathogenesis of psoriatic arthritis. Lancet 2018; 391:2273-84.

264. Tachmazidou I, Hatzikotoulas K, Southam L, Esparza-Gordillo J, Haberland V, Zheng J, et al. Identification of new therapeutic targets for osteoarthritis through genome-wide analyses of UK Biobank data. Nature Genetics 2019; 51:230-+.

265. Delaneau O, Marchini J, Zagury JF. A linear complexity phasing method for thousands of genomes. Nature Methods 2012; 9:179-81.

266. Howie BN, Donnelly P, Marchini J. A Flexible and Accurate Genotype Imputation Method for the Next Generation of Genome-Wide Association Studies. Plos Genetics 2009; 5:15.

267. Aryee MJ, Jaffe AE, Corrada-Bravo H, Ladd-Acosta C, Feinberg AP, Hansen KD, et al. Minfi: a flexible and comprehensive Bioconductor package for the analysis of Infinium DNA methylation microarrays. Bioinformatics 2014; 30:1363-9.

268. Teschendorff AE, Marabita F, Lechner M, Bartlett T, Tegner J, Gomez-Cabrero D, et al. A beta-mixture quantile normalization method for correcting probe design bias in Illumina Infinium 450 k DNA methylation data. Bioinformatics 2013; 29:189-96.

269. Triche TJ, Weisenberger DJ, Van Den Berg D, Laird PW, Siegmund KD. Low-level processing of Illumina Infinium DNA Methylation BeadArrays. Nucleic Acids Research 2013; 41:11.

270. Fortin JP, Labbe A, Lemire M, Zanke BW, Hudson TJ, Fertig EJ, et al. Functional normalization of 450k methylation array data improves replication in large cancer studies. Genome Biology 2014; 15:17.

271. Maksimovic J, Gordon L, Oshlack A. SWAN: Subset-quantile Within Array Normalization for Illumina Infinium HumanMethylation450 BeadChips. Genome Biology 2012; 13:12.

272. Pidsley R, Wong CCY, Volta M, Lunnon K, Mill J, Schalkwyk LC. A data-driven approach to preprocessing Illumina 450K methylation array data. Bmc Genomics 2013; 14:10.

273. Teschendorff AE, Zhuang J, Widschwendter M. Independent surrogate variable analysis to deconvolve confounding factors in large-scale microarray profiling studies. Bioinformatics 2011; 27:1496-505.

274. Gandolfo LC, Speed TP. RLE plots: Visualizing unwanted variation in high dimensional data. Plos One 2018; 13:9.

275. Langfelder P, Horvath S. Fast R Functions for Robust Correlations and Hierarchical Clustering. Journal of Statistical Software 2012; 46:1-17.

276. Benton MC, Johnstone A, Eccles D, Harmon B, Hayes MT, Lea RA, et al. An analysis of DNA methylation in human adipose tissue reveals differential modification of obesity genes before and after gastric bypass and weight loss. Genome Biology 2015; 16:21.

277. Chen YA, Lemire M, Choufani S, Butcher DT, Grafodatskaya D, Zanke BW, et al. Discovery of cross-reactive probes and polymorphic CpGs in the Illumina Infinium HumanMethylation450 microarray. Epigenetics 2013; 8:203-9.

278. McCartney DL, Walker RM, Morris SW, McIntosh AM, Porteous DJ, Evans KL. Identification of polymorphic and off-target probe binding sites on the Illumina Infinium MethylationEPIC BeadChip. Genomics Data 2016; 9:22-4.

279. Du P, Zhang XA, Huang CC, Jafari N, Kibbe WA, Hou LF, et al. Comparison of Beta-value and M-value methods for quantifying methylation levels by microarray analysis. Bmc Bioinformatics 2010; 11:9.

280. Johnson WE, Li C, Rabinovic A. Adjusting batch effects in microarray expression data using empirical Bayes methods. Biostatistics 2007; 8:118-27.

281. Nygaard V, Rodland EA, Hovig E. Methods that remove batch effects while retaining group differences may lead to exaggerated confidence in downstream analyses. Biostatistics 2016; 17:29-39.

282. Leek JT, Storey JD. Capturing heterogeneity in gene expression studies by surrogate variable analysis. Plos Genetics 2007; 3:1724-35.

283. Ritchie ME, Phipson B, Wu D, Hu YF, Law CW, Shi W, et al. limma powers differential expression analyses for RNA-sequencing and microarray studies. Nucleic Acids Research 2015; 43:13.

284. Smyth Gordon K. Linear Models and Empirical Bayes Methods for Assessing Differential Expression in Microarray Experiments. Statistical Applications in Genetics and Molecular Biology, 2004:1.

285. Benjamini Y, Hochberg Y. Controlling the false discovery rate - a practical and powerful approach to multiple testing. Journal of the Royal Statistical Society Series B-Statistical Methodology 1995; 57:289-300.

286. Peters TJ, Buckley MJ, Statham AL, Pidsley R, Samaras K, Lord RV, et al. De novo identification of differentially methylated regions in the human genome. Epigenetics & Chromatin 2015; 8:16.

287. Teschendorff AE, Gao Y, Jones A, Ruebner M, Beckmann MW, Wachter DL, et al. DNA methylation outliers in normal breast tissue identify field defects that are enriched in cancer. Nature Communications 2016; 7:12.

288. Geeleher P, Hartnett L, Egan LJ, Golden A, Ali RAR, Seoighe C. Gene-set analysis is severely biased when applied to genome-wide methylation data. Bioinformatics 2013; 29:1851-7.

289. Phipson B, Maksimovic J, Oshlack A. missMethyl: an R package for analyzing data from Illumina's HumanMethylation450 platform. Bioinformatics 2016; 32:286-8.

290. Young MD, Wakefield MJ, Smyth GK, Oshlack A. Gene ontology analysis for RNA-seq: accounting for selection bias. Genome Biology 2010; 11:12.

291. Horvath S. DNA methylation age of human tissues and cell types. Genome Biology 2013; 14:19.

292. Shabalin AA. Matrix eQTL: ultra fast eQTL analysis via large matrix operations. Bioinformatics 2012; 28:1353-8.

293. MacArthur J, Bowler E, Cerezo M, Gil L, Hall P, Hastings E, et al. The new NHGRI-EBI Catalog of published genome-wide association studies (GWAS Catalog). Nucleic Acids Research 2017; 45:D896-D901.

294. Holgate ST. Innate and adaptive immune responses in asthma. Nature Medicine 2012; 18:673-83.

295. Giambartolomei C, Vukcevic D, Schadt EE, Franke L, Hingorani AD, Wallace C, et al. Bayesian Test for Colocalisation between Pairs of Genetic Association Studies Using Summary Statistics. Plos Genetics 2014; 10:15.

296. Durinck S, Spellman PT, Birney E, Huber W. Mapping identifiers for the integration of genomic datasets with the R/Bioconductor package biomaRt. Nature Protocols 2009; 4:1184-91.

297. Lawrence M, Huber W, Pages H, Aboyoun P, Carlson M, Gentleman R, et al. Software for Computing and Annotating Genomic Ranges. Plos Computational Biology 2013; 9:10.

298. Millstein J, Chen GK, Breton CV. cit: hypothesis testing software for mediation analysis in genomic applications. Bioinformatics 2016; 32:2364-5.

299. Klug M, Rehli M. Functional Analysis of Promoter CpG Methylation Using a CpG-Free Luciferase Reporter Vector. Epigenetics 2006; 1:127-30.

300. Untergasser A, Cutcutache I, Koressaar T, Ye J, Faircloth BC, Remm M, et al. Primer3-new capabilities and interfaces. Nucleic Acids Research 2012; 40:12.

301. Larkin MA, Blackshields G, Brown NP, Chenna R, McGettigan PA, McWilliam H, et al. Clustal W and clustal X version 2.0. Bioinformatics 2007; 23:2947-8.

302. Jinek M, Chylinski K, Fonfara I, Hauer M, Doudna JA, Charpentier E. A Programmable Dual-RNA-Guided DNA Endonuclease in Adaptive Bacterial Immunity. Science 2012; 337:816-21.

303. Akcakaya P, Bobbin ML, Guo JA, Malagon-Lopez J, Clement K, Garcia SP, et al. In vivo CRISPR editing with no detectable genome-wide off-target mutations. Nature 2018; 561:416-+.

304. Morita S, Noguchi H, Horii T, Nakabayashi K, Kimura M, Okamura K, et al. Targeted DNA demethylation in vivo using dCas9-peptide repeat and scFv-TET1 catalytic domain fusions. Nature Biotechnology 2016; 34:1060-5.

305. Zhu H, Wu LF, Mo XB, Lu X, Tang H, Zhu XW, et al. Rheumatoid arthritis-associated DNA methylation sites in peripheral blood mononuclear cells. Annals of the Rheumatic Diseases 2019; 78:36-42.

306. Glossop JR, Emes RD, Nixon NB, Packham JC, Fryer AA, Mattey DL, et al. Genome-wide profiling in treatment-naive early rheumatoid arthritis reveals DNA methylome changes in T and B lymphocytes. Epigenomics 2016; 8:209-24.

307. de Andres MC, Perez-Pampin E, Calaza M, Santaclara FJ, Ortea I, Gomez-Reino JJ, et al. Assessment of global DNA methylation in peripheral blood cell subpopulations of

early rheumatoid arthritis before and after methotrexate. Arthritis Research & Therapy 2015; 17:9.

308. Iqbal K, Lendrem DW, Hargreaves B, Isaacs JD, Thompson B, Pratt AG. Routine musculoskeletal ultrasound findings impact diagnostic decisions maximally in autoantibody-seronegative early arthritis patients. Rheumatology 2019; 58:1268-73.

309. Liu J, Siegmund KD. An evaluation of processing methods for HumanMethylation450 BeadChip data. Bmc Genomics 2016; 17:11.

310. Polansky JK, Kretschmer K, Freyer J, Floess S, Garbe A, Baron U, et al. DNA methylation controls Foxp3 gene expression. European Journal of Immunology 2008; 38:1654-63.

311. Field AE, Robertson NA, Wang T, Havas A, Ideker T, Adams PD. DNA Methylation Clocks in Aging: Categories, Causes, and Consequences. Molecular Cell 2018; 71:882-95.

312. Marioni RE, Suderman M, Chen BH, Horvath S, Bandinelli S, Morris T, et al. Tracking the Epigenetic Clock Across the Human Life Course: A Meta-analysis of Longitudinal Cohort Data. Journals of Gerontology Series a-Biological Sciences and Medical Sciences 2019; 74:57-61.

313. Chavez-Valencia RA, Chiaroni-Clarke RC, Martino DJ, Munro JE, Allen RC, Akikusa JD, et al. The DNA methylation landscape of CD4(+) T cells in oligoarticular juvenile idiopathic arthritis. Journal of Autoimmunity 2018; 86:29-38.

314. Maltby VE, Lea RA, Graves MC, Sanders KA, Benton MC, Tajouri L, et al. Genome-wide DNA methylation changes in CD19(+) B cells from relapsing-remitting multiple sclerosis patients. Scientific Reports 2018; 8:10.

315. Carnero-Montoro E, Alarcon-Riquelme ME. Epigenome-wide association studies for systemic autoimmune diseases: The road behind and the road ahead. Clinical Immunology 2018; 196:21-33.

316. Myte R, Sundkvist A, Van Guelpen B, Harlid S. Circulating levels of inflammatory markers and DNA methylation, an analysis of repeated samples from a population based cohort. Epigenetics 2019; 14:649-59.

317. Shao XJ, Hudson M, Colmegna I, Greenwood CMT, Fritzler MJ, Awadalla P, et al. Rheumatoid arthritis-relevant DNA methylation changes identified in ACPA-positive asymptomatic individuals using methylome capture sequencing. Clinical Epigenetics 2019; 11:11.

318. Altorok N, Coit P, Hughes T, Koelsch KA, Stone DU, Rasmussen A, et al. Genome-Wide DNA Methylation Patterns in Naive CD4+T Cells From Patients With Primary Sjogren's Syndrome. Arthritis & Rheumatology 2014; 66:731-9.

319. Yang HB, Kim O, Wu J, Qiu Y. Interaction between tyrosine kinase Etk and a RUN domain- and FYVE domain-containing protein RUFY1 - A possible role of Etk in regulation of vesicle trafficking. Journal of Biological Chemistry 2002; 277:30219-26.

320. Svendsen AJ, Gervin K, Lyle R, Christiansen L, Kyvik K, Junker P, et al. Differentially Methylated DNA Regions in Monozygotic Twin Pairs Discordant for Rheumatoid Arthritis: An Epigenome-Wide study. Frontiers in Immunology 2016; 7:10.

321. Yao CC, Sakata D, Esaki Y, Li YX, Matsuoka T, Kuroiwa K, et al. Prostaglandin E-2-EP4 signaling promotes immune inflammation through T(H)1 cell differentiation and T(H)17 cell expansion. Nature Medicine 2009; 15:633-U141.

322. Bock C, Walter J, Paulsen M, Lengauer T. CpG island mapping by epigenome prediction. Plos Computational Biology 2007; 3:1055-70.

323. Oh H, Ghosh S. NF-kappa B: roles and regulation in different CD4(+) T-cell subsets. Immunological Reviews 2013; 252:41-51.

324. Djuretic IM, Levanon D, Negreanu V, Groner Y, Rao A, Ansel KM. Transcription factors T-bet and Runx3 cooperate to activate lfng and silence ll4 in T helper type 1 cells. Nature Immunology 2007; 8:145-53.

325. Wang DP, Diao HT, Getzler AJ, Rogal W, Frederick MA, Milner J, et al. The Transcription Factor Runx3 Establishes-Chromatin Accessibility of cis-Regulatory Landscapes that Drive Memory Cytotoxic T Lymphocyte Formation. Immunity 2018; 48:659-+.

326. Perissi V, Aggarwal A, Glass CK, Rose DW, Rosenfeld MG. A corepressor/coactivator exchange complex required for transcriptional activation by nuclear receptors and other regulated transcription factors. Cell 2004; 116:511-26.

327. Yoshimura A, Seki M, Enomoto T. The role of WRNIP1 in genome maintenance. Cell Cycle 2017; 16:515-21.

328. Yi SQ, Yu M, Yang S, Miron RJ, Zhang YF. Tcf12, A Member of Basic Helix-Loop-Helix Transcription Factors, Mediates Bone Marrow Mesenchymal Stem Cell Osteogenic Differentiation In Vitro and In Vivo. Stem Cells 2017; 35:386-97.

329. Woo JS, Srikanth S, Kim KD, Elsaesser H, Lu J, Pellegrini M, et al. CRACR2A-Mediated TCR Signaling Promotes Local Effector Th1 and Th17 Responses. Journal of Immunology 2018; 201:1174-85.

330. Mizuno A, Okada Y. Biological characterization of expression quantitative trait loci (eQTLs) showing tissue-specific opposite directional effects. European Journal of Human Genetics 2019; 27:1745-56.

331. Greenberg MVC, Bourc'his D. The diverse roles of DNA methylation in mammalian development and disease. Nature Reviews Molecular Cell Biology 2019; 20:590-607.

332. Rushton MD, Reynard LN, Young DA, Shepherd C, Aubourg G, Gee F, et al. Methylation quantitative trait locus analysis of osteoarthritis links epigenetics with genetic risk. Human Molecular Genetics 2015; 24:7432-44.

333. Miceli-Richard C, Wang-Renault SF, Boudaoud S, Busato F, Lallemand C, Bethune K, et al. Overlap between differentially methylated DNA regions in blood B lymphocytes and genetic at-risk loci in primary Sjogren's syndrome. Annals of the Rheumatic Diseases 2016; 75:933-40.

334. Wu Y, Zeng J, Zhang FT, Zhu ZH, Qi T, Zheng ZL, et al. Integrative analysis of omics summary data reveals putative mechanisms underlying complex traits. Nature Communications 2018; 9:14.

335. Liu T, Zhang L, Joo D, Sun S-C. NF-κB signaling in inflammation. Signal Transduction and Targeted Therapy 2017; 2:17023.

336. Rodríguez-Ubreva J, de la Calle-Fabregat C, Li T, Ciudad L, Ballestar ML, Català-Moll F, et al. Inflammatory cytokines shape a changing DNA methylome in monocytes mirroring disease activity in rheumatoid arthritis. Annals of the Rheumatic Diseases 2019; 78:1505-16.

337. Fairfax BP, Humburg P, Makino S, Naranbhai V, Wong D, Lau E, et al. Innate Immune Activity Conditions the Effect of Regulatory Variants upon Monocyte Gene Expression. Science 2014; 343:1118-+.

338. Ye CJ, Feng T, Kwon HK, Raj T, Wilson M, Asinovski N, et al. Intersection of population variation and autoimmunity genetics in human T cell activation. Science 2014; 345:1311-+.

339. Alasoo K, Rodrigues J, Mukhopadhyay S, Knights AJ, Mann AL, Kundu K, et al. Shared genetic effects on chromatin and gene expression indicate a role for enhancer priming in immune response. Nature Genetics 2018; 50:424-+.

340. Calderon D, Nguyen MLT, Mezger A, Kathiria A, Muller F, Nguyen V, et al. Landscape of stimulation-responsive chromatin across diverse human immune cells. Nature Genetics 2019; 51:1494-+.

341. Czamara D, Eraslan G, Page CM, Lahti J, Lahti-Pulkkinen M, Hamalainen E, et al. Integrated analysis of environmental and genetic influences on cord blood DNA methylation in new-borns. Nature Communications 2019; 10:18.

342. Chun S, Casparino A, Patsopoulos NA, Croteau-Chonka DC, Raby BA, De Jager PL, et al. Limited statistical evidence for shared genetic effects of eQTLs and autoimmune-disease-associated loci in three major immune-cell types. Nature Genetics 2017; 49:600-+.

343. Richardson TG, Shihab HA, Hemani G, Zheng J, Hannon E, Mill J, et al. Collapsed methylation quantitative trait loci analysis for low frequency and rare variants. Human Molecular Genetics 2016; 25:4339-49.

344. Aguet F, Brown AA, Castel SE, Davis JR, He Y, Jo B, et al. Genetic effects on gene expression across human tissues. Nature 2017; 550:204-+.

345. Huan TX, Joehanes R, Song C, Peng F, Guo YC, Mendelson M, et al. Genome-wide identification of DNA methylation QTLs in whole blood highlights pathways for cardiovascular disease. Nature Communications 2019; 10:14.

346. Kochi Y, Yamada R, Suzuki A, Harley JB, Shirasawa S, Sawada T, et al. A functional variant in FCRL3, encoding Fc receptor-like 3, is associated with rheumatoid arthritis and several autoimmunities. Nature Genetics 2005; 37:478-85.

347. Hemani G, Tilling K, Smith GD. Orienting the causal relationship between imprecisely measured traits using GWAS summary data. Plos Genetics 2017; 13:22.

348. Relton CL, Smith GD. Two-step epigenetic Mendelian randomization: a strategy for establishing the causal role of epigenetic processes in pathways to disease. International Journal of Epidemiology 2012; 41:161-76.

349. Corradin O, Saiakhova A, Akhtar-Zaidi B, Myeroff L, Willis J, Iari RCS, et al. Combinatorial effects of multiple enhancer variants in linkage disequilibrium dictate levels of gene expression to confer susceptibility to common traits. Genome Research 2014; 24:1-13.

350. Westra HJ, Martinez-Bonet M, Onengut-Gumuscu S, Lee A, Luol Y, Teslovich N, et al. Fine-mapping and functional studies highlight potential causal variants for rheumatoid arthritis and type 1 diabetes. Nature Genetics 2018; 50:1366-+.

351. Hormozdiari F, Gazal S, van de Geijn B, Finucane HK, Ju CJT, Loh PR, et al. Leveraging molecular quantitative trait loci to understand the genetic architecture of diseases and complex traits. Nature Genetics 2018; 50:1041-+.

352. Davis RS, Wang YH, Kubagawa H, Cooper MD. Identification of a family of Fc receptor homologs with preferential B cell expression. Proceedings of the National Academy of Sciences of the United States of America 2001; 98:9772-7.

353. Davis RS. Fc receptor-like molecules. Annual Review of Immunology 2007; 25:525-60.

354. Nagata S, Ise T, Pastan I. Fc Receptor-Like 3 Protein Expressed on IL-2 Nonresponsive Subset of Human Regulatory T Cells. Journal of Immunology 2009; 182:7518-26.

355. Bajpai UD, Swainson LA, Mold JE, Graf JD, Imboden JB, McCune JM. A functional variant in FCRl3 is associated with higher fc receptor-like 3 expression on T cell subsets and rheumatoid arthritis disease activity. Arthritis and Rheumatism 2012; 64:2451-9.

356. Kochi Y, Myouzen K, Yamada R, Suzuki A, Kurosaki T, Nakamura Y, et al. FCRL3, an Autoimmune Susceptibility Gene, Has Inhibitory Potential on B-Cell Receptor-Mediated Signaling. Journal of Immunology 2009; 183:5502-10.

357. Li FJ, Schreeder DM, Li R, Wu JR, Davis RS. FCRL3 promotes TLR9-induced B-cell activation and suppresses plasma cell differentiation. European Journal of Immunology 2013; 43:2980-+.

358. Alloza I, Otaegui D, de Lapuente AL, Antiguedad A, Varade J, Nunez C, et al. ANKRD55 and DHCR7 are novel multiple sclerosis risk loci. Genes and Immunity 2012; 13:253-7.

359. Rose-John S. IL-6 Trans-Signaling via the Soluble IL-6 Receptor: Importance for the Pro-Inflammatory Activities of IL-6. International Journal of Biological Sciences 2012; 8:1237-47.

360. Srirangan S, Choy EH. The role of Interleukin 6 in the pathophysiology of rheumatoid arthritis. Therapeutic Advances in Musculoskeletal Disease 2010; 2:247-56.

361. Li JN, Mahajan A, Tsai MD. Ankyrin repeat: A unique motif mediating protein-protein interactions. Biochemistry 2006; 45:15168-78.

362. de Lapuente AL, Feliu A, Ugidos N, Mecha M, Mena J, Astobiza I, et al. Novel Insights into the Multiple Sclerosis Risk Gene ANKRD55 (vol 196, pg 4553, 2016). Journal of Immunology 2016; 197:4177-.

363. Kothari PH, Qiu WL, Croteau-Chonka DC, Martinez FD, Liu AH, Lemanske RF, et al. Role of local CpG DNA methylation in mediating the 17q21 asthma susceptibility gasdermin B (GSDMB)/ORMDL sphingolipid biosynthesis regulator 3 (ORMDL3)

expression quantitative trait locus. Journal of Allergy and Clinical Immunology 2018; 141:2282-6.

364. Hjelmqvist L, Tuson M, Marfany G, Herrero E, Balcells S, Gonzalez-Duarte R. ORMDL proteins are a conserved new family of endoplasmic reticulum membrane proteins. Genome Biology 2002; 3:16.

365. Cantero-Recasens G, Fandos C, Rubio-Moscardo F, Valverde MA, Vicente R. The asthma-associated ORMDL3 gene product regulates endoplasmic reticulum-mediated calcium signaling and cellular stress. Human Molecular Genetics 2010; 19:111-21.

366. Schmiedel BJ, Seumois G, Samaniego-Castruita D, Cayford J, Schulten V, Chavez L, et al. 17q21 asthma-risk variants switch CTCF binding and regulate IL-2 production by T cells. Nature Communications 2016; 7:14.

367. Liu JZ, van Sommeren S, Huang HL, Ng SC, Alberts R, Takahashi A, et al. Association analyses identify 38 susceptibility loci for inflammatory bowel disease and highlight shared genetic risk across populations. Nature Genetics 2015; 47:979-+.

368. Ding JJ, Wang K, Liu W, She Y, Sun Q, Shi JJ, et al. Pore-forming activity and structural autoinhibition of the gasdermin family. Nature 2016; 535:111-+.

369. Chen Q, Shi PL, Wang YF, Zou DY, Wu XW, Wang DY, et al. GSDMB promotes non-canonical pyroptosis by enhancing caspase-4 activity. Journal of Molecular Cell Biology 2019; 11:496-508.

370. Koontz JI, Soreng AL, Nucci M, Kuo FC, Pauwels P, van den Berghe H, et al. Frequent fusion of the JAZF1 and JJAZ1 genes in endometrial stromal tumors. Proceedings of the National Academy of Sciences of the United States of America 2001; 98:6348-53.

371. Nakajima T, Fujino S, Nakanishi G, Kim YS, Jetten AM. TIP27: a novel repressor of the nuclear orphan receptor TAK1/TR4. Nucleic Acids Research 2004; 32:4194-204.

372. Yang ML, Dai JH, Jia YJ, Suo LQZ, Li SB, Guo YS, et al. Overexpression of juxtaposed with another zinc finger gene 1 reduces proinflammatory cytokine release via inhibition of stress-activated protein kinases and nuclear factor-kappa B. Febs Journal 2014; 281:3193-205.

373. Klein JC, Keith A, Rice SJ, Shepherd C, Agarwal V, Loughlin J, et al. Functional testing of thousands of osteoarthritis-associated variants for regulatory activity. Nature Communications 2019; 10:9.

374. Bell CG, Gao F, Yuan W, Roos L, Acton RJ, Xia YD, et al. Obligatory and facilitative allelic variation in the DNA methylome within common disease-associated loci. Nature Communications 2018; 9:13.

375. Teschendorff AE, Relton CL. Statistical and integrative system-level analysis of DNA methylation data. Nature Reviews Genetics 2018; 19:129-47.

376. Hannon E, Knox O, Sugden K, Burrage J, Wong CCY, Belsky DW, et al. Characterizing genetic and environmental influences on variable DNA methylation using monozygotic and dizygotic twins. Plos Genetics 2018; 14:27.

377. van Dongen J, Nivard MG, Willemsen G, Hottenga JJ, Helmer Q, Dolan CV, et al. Genetic and environmental influences interact with age and sex in shaping the human methylome. Nature Communications 2016; 7:13.

378. Meng WD, Zhu ZH, Jiang X, Too CL, Uebe S, Jagodic M, et al. DNA methylation mediates genotype and smoking interaction in the development of anti-citrullinated peptide antibody-positive rheumatoid arthritis. Arthritis Research & Therapy 2017; 19:10.

379. Knowles DA, Davis JR, Edgington H, Raj A, Fave MJ, Zhu XW, et al. Allele-specific expression reveals interactions between genetic variation and environment. Nature Methods 2017; 14:699-+.

380. Smolen JS, Aletaha D, McInnes IB. Rheumatoid arthritis. Lancet 2016; 388:2023-38.

381. Clark SJ, Argelaguet R, Kapourani CA, Stubbs TM, Lee HJ, Alda-Catalinas C, et al. scNMT-seq enables joint profiling of chromatin accessibility DNA methylation and transcription in single cells. Nature Communications 2018; 9:9.

382. Jeffries MA. Epigenetic editing: How cutting-edge targeted epigenetic modification might provide novel avenues for autoimmune disease therapy. Clinical Immunology 2018; 196:49-58.

383. Taghbalout A, Du MH, Jillette N, Rosikiewicz W, Rath A, Heinen CD, et al. Enhanced CRISPR-based DNA demethylation by Casilio-ME-mediated RNA-guided coupling of methylcytosine oxidation and DNA repair pathways. Nature Communications 2019; 10:12.

Appendices

Appendix A – Patient Phenotype Data and Sample processing Information

All patient phenotype data

Clinical characteristics of each patient included in the study. EA ID = Unique anonymised identifier given to all patients attending the early arthritis clinic; Sex (F = Female, M = Male); Diagnosis = Most recent patient diagnosis (CA = Crystal arthritis, EA = Enteropathic arthritis, L/CTD = Lupus/other connective tissue disease-associated, NIA = Non-inflammatory arthritis, OA = Osteoarthritis, OIA = Other inflammatory arthritis, PsA = Psoriatic arthritis, RA = Rheumatoid arthritis, ReA = Reactive arthritis, UIA = Undifferentiated inflammatory arthritis, USpA = Undifferentiated spondyloarthropathy); CRP = C-reactive protein (mg/L of blood); ESR = erythrocyte sedimentation rate (mm/hour); CCP = cyclic citrullinated peptide test (clinical test for anti-cyclic citrullinated protein antibodies, positive (+) or negative (-)); RF = Rheumatoid factor (positive (+) or negative (-)); Tender 28 = Tender joint count (0 – 28), Swollen 28 = Swollen joint count (0-28; see Chapter 1.2); DAS28 = Disease activity score at 28 joints; Symptom duration weeks = patient-reported symptom duration in number of weeks.

EA ID	Age	Sex	Diagnosis	RA/ nRA	Smoking	CRP	ESR	CCP	RF	Tender 28	Swollen 28	DAS28	Symptom weeks
809	51	F	OIA	nRA	No	12	29	-	-	0	0	3.11	12
810	56	M	RA	RA	No	7	13	+	+	13	3	5.03	52
813	50	F	UIA	nRA	No	11	10	-	+	2	0	3.64	28
814	31	F	CA	nRA	No	7	9	-	-	1	0	2.53	NA
815	70	F	OIA	nRA	Yes	18	33	-	-	7	0	4.84	24
826	52	F	OA	nRA	No	5	7	-	NA	15	0	4.74	NA
827	50	F	RA	RA	Yes	5	7	-	-	13	3	4.51	12
829	80	F	OIA	nRA	No	52	94	-	-	4	2	6.05	16
834	59	F	RA	RA	No	22	67	+	+	18	0	6.05	24
835	82	M	UIA	nRA	No	43	37	-	-	0	0	2.86	16
837	33	F	UIA	nRA	No	14	13	-	-	0	0	2.69	24
838	27	F	RA	RA	Yes	6	34	+	+	3	1	4.63	52
839	33	M	ReA	nRA	No	20	28	-	-	8	3	5.14	20
840	68	M	RA	RA	No	10	23	-	-	6	3	5.16	52
841	22	M	ReA	nRA	No	11	8	-	-	2	0	2.99	8
842	52	F	EA	nRA	No	0	7	-	-	3	0	3.05	52
844	58	F	RA	RA	No	6	18	+	+	0	0	2.30	52
845	67	F	OIA	nRA	No	8	18	-	-	2	1	3.22	52
854	49	M	RA	RA	Yes	5	10	+	-	1	0	2.45	28
867	63	F	ReA	nRA	No	64	28	-	-	2	0	3.80	9
868	56	M	OA	nRA	No	0	1	-	-	0	1	1.15	52
874	25	F	UIA	nRA	No	10	6	-	-	0	0	2.29	1
878	74	F	PsA	nRA	No	9	16	-	-	3	1	3.95	36
879	79	M	ReA	nRA	No	0	1	-	-	0	3	1.46	12
882	27	F	RA	RA	No	9	63	+	+	10	0	5.71	4

884	41	F	ReA	nRA	No	9	30	-	-	6	0	3.98	8
891	26	F	OIA	nRA	No	30	78	-	-	0	0	4.03	4
892	71	M	RA	RA	Yes	6	6	+	+	1	1	2.47	10
893	43	F	PsA	nRA	No	10	35	-	-	4	3	5.19	12
896	30	F	RA	RA	No	7	18	-	-	11	6	5.17	12
898	54	F	ReA	nRA	No	5	23	-	-	3	0	3.86	12
905	56	F	USpA	nRA	Yes	5	8	-	-	4	0	3.68	32
912	50	F	OIA	nRA	Yes	5	13	-	-	4	2	3.59	12
915	39	F	RA	RA	Yes	11	33	-	-	11	7	5.49	8
923	79	F	ReA	nRA	No	6	30	-	-	6	3	5.05	24
926	61	M	RA	RA	No	12	12	-	-	7	1	4.73	52
929	48	F	RA	RA	No	11	12	-	+	0	0	NA	12
930	69	F	RA	RA	Yes	19	29	-	+	20	0	6.15	12
932	73	F	RA	RA	Yes	10	86	+	+	13	0	6.34	36
934	79	M	RA	RA	No	5	24	-	-	0	0	NA	7
935	63	F	PsA	nRA	No	5	4	-	-	0	0	NA	10
937	77	F	CA	nRA	No	20	49	-	-	0	2	4.27	28
938	74	M	RA	RA	No	20	65	-	+	5	3	5.65	4
944	51	F	RA	RA	No	11	45	-	+	1	2	4.32	6
945	61	M	UIA	nRA	Yes	5	1	-	+	0	0	0.70	4
946	43	M	RA	RA	No	7	12	+	+	0	0	NA	NA
948	53	F	RA	RA	No	53	59	+	+	0	0	NA	3
954	45	F	PsA	nRA	No	5	7	-	-	3	1	2.77	52
957	57	M	RA	RA	No	10	23	+	+	17	2	5.66	24
962	65	M	PsA	nRA	No	8	14	-	-	2	2	3.74	4
965	57	M	OA	nRA	No	8	4	-	-	0	0	0.97	NA
967	62	F	PsA	nRA	Yes	171	111	-	-	1	5	5.49	7
969	38	F	ReA	nRA	No	14	38	-	-	0	0	3.55	NA
973	50	F	UIA	nRA	No	9	24	-	-	0	0	NA	4
974	69	F	RA	RA	No	24	59	-	-	0	0	NA	6
975	54	F	RA	RA	No	9	11	+	+	0	0	NA	24
978	62	F	RA	RA	No	5	5	-	-	17	6	5.19	18
980	20	F	ReA	nRA	No	8	15	-	-	4	1	3.65	8
992	60	F	RA	RA	No	45	10	+	+	0	0	NA	16
995	73	F	RA	RA	No	9	6	-	-	18	8	5.44	14
996	53	F	RA	RA	No	5	4	-	-	3	3	2.54	24
1000	74	F	RA	RA	No	26	7	-	-	0	0	NA	3
1003	51	M	RA	RA	Yes	13	21	+	+	8	5	5.36	52
1005	69	F	RA	RA	Yes	5	32	-	+	7	1	5.11	NA
1008	70	M	PsA	nRA	No	5	1	-	-	18	4	4.08	NA
1010	50	F	RA	RA	Yes	9	21	-	-	4	3	4.37	12
1019	38	F	OA	nRA	Yes	8	23	+	-	7	4	5.05	12
1020	45	M	UIA	nRA	No	5	10	-	-	0	0	2.40	24

1022	53	F	USpA	nRA	No	21	22	-	-	8	0	4.66	3
1032	68	F	RA	RA	No	13	26	+	+	2	0	3.87	NA
1037	65	M	CA	nRA	No	7	48	-	-	4	4	5.10	16
1042	50	F	RA	RA	No	10	5	+	+	0	0	NA	52
1050	58	M	CA	nRA	No	8	NA	+	+	2	3	NA	2
1051	69	F	RA	RA	Yes	29	54	-	+	16	12	7.15	12
1054	57	F	OA	nRA	No	7	22	-	-	13	2	5.49	8
1058	44	F	NIA	nRA	No	5	14	-	-	22	2	5.64	4
1067	79	M	RA	RA	No	17	22	-	-	2	1	4.05	8
1070	46	F	PsA	nRA	No	5	8	-	-	1	1	3.35	52
1072	50	F	RA	RA	Yes	5	1	-	+	11	0	2.31	12
1076	66	F	ReA	nRA	Yes	6	19	-	-	1	9	4.19	12
1080	52	F	PsA	nRA	No	45	33	-	-	2	4	5.09	9
1083	63	F	RA	RA	No	11	50	+	-	1	1	3.91	8
1085	43	F	PsA	nRA	No	15	12	-	-	0	0	2.68	16
1087	55	M	CA	nRA	No	24	15	-	-	0	0	2.32	24
1088	54	F	ReA	nRA	No	76	15	-	+	1	0	3.48	4
1094	61	M	RA	RA	Yes	7	4	+	+	0	0	1.26	6
2010	46	F	RA	RA	Yes	5	4	+	+	1	1	2.24	52
2012	63	F	ReA	nRA	No	5	10	-	-	3	0	2.96	6
2013	73	F	RA	RA	No	10	31	+	-	3	1	4.40	12
2028	37	F	NIA	nRA	No	5	2	-	-	0	0	1.24	32
2029	59	M	RA	RA	Yes	10	32	+	-	2	0	4.60	12
2030	57	M	RA	RA	Yes	5	2	-	-	3	0	2.63	52
2034	56	F	UIA	nRA	No	5	10	-	-	1	1	3.36	24
2036	61	M	PsA	nRA	No	5	19	-	-	4	0	3.37	20
2040	53	F	OA	nRA	No	9	10	-	-	21	1	4.99	52
2042	61	M	RA	RA	No	5	15	-	+	2	0	3.28	4
2044	56	F	PsA	nRA	Yes	23	29	-	-	6	0	4.97	2
2045	55	M	RA	RA	No	5	20	+	+	3	0	3.40	26
2047	51	M	PsA	nRA	No	5	1	-	-	8	3	2.50	12
2052	50	F	NIA	nRA	No	5	14	-	-	15	0	5.08	12
2054	66	M	CA	nRA	No	6	20	-	+	1	2	3.23	3
2062	60	M	CA	nRA	No	7	35	-	-	1	4	3.85	4
2067	27	F	NIA	nRA	No	5	6	-	+	0	0	1.37	2
2072	58	F	RA	RA	No	5	19	+	+	1	1	3.74	52
2075	81	M	ReA	nRA	No	20	26	-	-	6	6	4.95	9
2078	92	M	CA	nRA	No	11	34	-	+	1	6	4.16	7
2086	45	F	NIA	nRA	Yes	6	20	+	+	0	0	2.60	6
2087	47	F	PsA	nRA	No	5	12	-	-	1	0	2.75	5
2090	57	F	RA	RA	No	5	5	+	+	0	0	1.55	52
2133	70	F	RA	RA	Yes	48	37	+	+	6	6	5.64	52
2140	74	M	RA	RA	No	12	21	-	-	3	7	4.04	24

2144	51	M	OIA	nRA	No	10	4	-	-	0	0	1.08	34
2145	49	F	EA	nRA	No	5	1	-	-	19	2	3.82	52
2146	56	F	NIA	nRA	Yes	5	12	-	-	19	0	5.58	52
2148	60	M	ReA	nRA	No	189	113	-	-	0	0	3.60	3
2166	57	F	RA	RA	Yes	62	91	-	-	1	0	4.17	12
2197	57	F	NIA	nRA	No	9	27	-	-	0	0	3.01	32
2200	46	M	PsA	nRA	No	15	20	-	-	7	6	4.66	2
2203	50	F	RA	RA	No	8	4	+	-	1	1	2.82	52
2222	46	F	ReA	nRA	No	33	21	-	-	2	0	3.62	2
2231	56	F	RA	RA	No	16	27	+	+	12	6	5.35	8
2233	58	F	OIA	nRA	Yes	4	30	-	-	7	1	5.26	52
2257	35	F	ReA	nRA	No	4	9	-	-	20	12	5.91	5
2261	69	M	CA	nRA	No	4	47	-	-	0	0	2.98	7
2264	70	F	EA	nRA	No	8	67	-	-	1	2	4.21	5
2265	51	F	PsA	nRA	No	4	9	-	-	3	3	3.55	6
2281	56	F	OIA	nRA	Yes	93	40	-	+	0	1	3.30	16
2305	59	F	L/CTD	nRA	No	10	27	-	-	5	4	5.30	52
2311	40	F	RA	RA	Yes	4	23	+	+	0	0	2.61	24
2322	68	F	CA	nRA	No	7	5	-	+	4	3	3.99	12
2330	32	F	NIA	nRA	Yes	4	9	-	-	7	0	3.20	8
2345	82	F	RA	RA	No	56	16	-	-	18	15	6.54	3
2348	62	M	RA	RA	No	13	5	-	-	7	7	4.19	8
2367	82	M	USpA	nRA	No	75	99	-	-	0	0	3.50	26
2368	61	F	RA	RA	No	12	35	+	+	2	1	4.26	8
2369	59	F	RA	RA	No	16	43	+	+	0	0	3.40	14
2378	75	F	ReA	nRA	No	23	39	-	-	5	8	5.03	4
2379	72	M	ReA	nRA	No	4	38	-	-	1	4	3.67	4
2390	29	M	OIA	nRA	No	5	9	-	-	0	0	2.38	3
2416	37	F	L/CTD	nRA	No	23	41	-	-	1	0	4.14	8
2437	56	M	ReA	nRA	No	54	16	-	-	2	0	3.26	8
2439	55	F	ReA	nRA	No	4	12	+	-	20	4	5.41	10
2476	81	F	UIA	nRA	No	15	38	-	-	3	0	4.31	12
2493	54	F	ReA	nRA	No	4	7	-	-	3	0	3.09	6
2506	61	M	UIA	nRA	No	15	31	-	-	3	0	3.91	8
2507	22	F	ReA	nRA	No	4	8	-	-	1	0	2.34	5
2510	51	F	RA	RA	No	4	2	+	+	17	0	3.47	24
2511	53	F	UIA	nRA	No	16	40	-	-	14	0	5.57	12
2519	32	F	UIA	nRA	No	4	2	-	-	3	0	2.58	20
2520	45	F	UIA	nRA	No	4	16	-	-	17	0	5.13	52
2548	34	F	PsA	nRA	No	9	2	-	-	2	2	2.07	24
2549	52	F	UIA	nRA	No	5	2	+	-	6	0	2.84	52
2759	68	F	RA	RA	Yes	7	75	-	+	2	10	4.84	8
2767	43	F	RA	RA	Yes	4	2	-	-	5	2	3.11	52

Sample processing and quality control data for CD4+ T cell and B cell samples. EA ID = Unique anonymised identifier given to all patients attending the early arthritis clinic. QC = Quality control indicating whether a sample passed all quality control checks or failed at any check; Purity Meth % = The purity of target cells in each sample preparation as determined from the DNA methylation data using the Houseman reference-based method[219]; Purity Flow % = The purity of target cells in each sample preparation as determined by flow cytometry(see Chapter 2.2); Conv. Batch = Bisulphite conversion batch for each sample; iScan Batch = Batch in which each sample was read on the iScan system; Array ID = unique identifier given the each MethylationEPIC array; Array Position = Physical position of each sample on the respective array (from row 1 to row 8); Diff. Analysis/meQTL/eQTM = inclusion of a given sample in differential analyses (Chapter 3)/meQTL mapping (Chapter 4)/eQTM mapping (Chapter 5) (✓ = included, × = not included).

CD4+ T cell sample processing and quality control information

EA ID	QC	CD4T Purity Meth (%)	CD4T Purity Flow (%)	Conv. Batch	iScan Batch	Array ID	Array Position	Diff. Analysis	meQTL	eQTM
809	Pass	0.99	98.9	1	1	1	1	✓	✓	✓
810	Pass	1	99.2	1	1	1	3	✓	✓	✓
813	Pass	0.87	99.3	1	1	1	5	✓	✓	✓
814	Pass	0.99	98.8	1	1	2	4	✓	✓	✓
815	Pass	1	99.4	1	1	1	7	✓	✓	✓
826	Pass	0.97	98.5	1	1	2	6	✓	✓	✓
827	Pass	0.93	97.9	1	1	3	3	✓	✓	✓
829	Pass	0.99	96.7	1	1	2	2	✓	✓	✓
834	Pass	0.94	97.4	1	1	3	7	✓	×	×
835	Fail	0.42	97.4	1	1	4	6	×	×	×
837	Pass	0.92	98	1	1	3	1	✓	×	×
838	Pass	0.97	97	1	1	4	2	✓	✓	✓
839	Pass	0.98	98.1	1	1	4	4	✓	×	×
840	Pass	0.99	98.5	1	1	4	8	✓	✓	✓
841	Pass	0.99	98.8	1	1	5	1	✓	✓	✓
842	Pass	0.9	98	1	1	3	5	✓	✓	✓
844	Pass	0.84	97.3	1	1	5	7	✓	✓	✓
845	Pass	0.93	98.6	1	1	5	5	✓	✓	✓
854	Pass	0.93	98.6	1	1	6	2	✓	✓	✓
874	Pass	0.98	96.8	1	1	6	4	✓	×	×
878	Pass	0.93	96.8	1	1	6	8	✓	✓	✓
882	Pass	0.98	98.6	1	1	6	6	✓	✓	✓
891	Pass	0.96	98.3	6	4	7	2	✓	✓	✓
896	Pass	0.97	NA	6	4	7	4	✓	✓	✓
898	Pass	0.93	NA	6	4	7	6	✓	✓	✓
905	Pass	0.76	95.5	6	4	7	8	✓	✓	✓
912	Pass	0.99	97.7	5	3	8	4	✓	✓	✓
915	Pass	0.85	99	1	1	2	8	✓	✓	✓

923	Pass	0.73	79.6	6	4	9	1	✓	×	×
926	Pass	0.8	86.6	6	4	10	3	✓	✓	✓
929	Pass	0.85	97.5	6	4	9	3	✓	✓	✓
930	Pass	0.73	91.9	6	4	11	8	✓	✓	✓
932	Pass	0.66	94.3	6	4	12	6	✓	✓	✓
934	Pass	0.99	98.9	1	1	5	3	✓	✓	✓
935	Pass	0.96	97.9	6	4	9	5	✓	✓	✓
937	Pass	0.96	98.3	6	4	11	2	✓	✓	✓
938	Pass	0.95	97.9	6	4	11	4	✓	✓	✓
944	Pass	0.68	96.1	6	4	10	1	✓	×	×
945	Pass	0.93	98.2	6	4	9	7	✓	✓	✓
946	Pass	0.99	97.5	3	2	13	5	✓	✓	✓
948	Pass	0.98	NA	3	2	14	7	✓	✓	×
954	Pass	0.96	99.5	6	4	11	6	✓	✓	✓
957	Pass	0.78	98.4	3	2	14	1	✓	×	×
962	Pass	0.98	99.3	6	4	10	5	✓	✓	✓
965	Pass	0.91	96.4	6	4	10	7	✓	✓	✓
967	Pass	0.92	99	6	4	12	4	✓	✓	✓
969	Pass	0.91	96.4	3	2	14	3	✓	✓	✓
973	Pass	0.84	98.4	6	4	12	2	✓	✓	✓
974	Pass	0.97	99.1	3	2	15	8	✓	✓	✓
975	Pass	0.98	99.1	3	2	13	3	✓	✓	✓
978	Pass	0.89	99.2	3	2	16	4	✓	✓	✓
980	Pass	0.87	NA	6	4	12	8	✓	✓	✓
992	Pass	0.94	98	3	2	15	2	✓	✓	✓
995	Pass	0.97	97.4	3	2	17	5	✓	✓	✓
996	Pass	0.92	98.1	4	2	18	5	✓	✓	✓
1000	Pass	1	98.7	4	2	19	3	✓	✓	✓
1003	Pass	0.98	98.1	3	2	17	3	✓	✓	✓
1005	Pass	0.98	96	3	2	20	2	✓	✓	✓
1008	Pass	0.98	98.3	3	2	15	4	✓	✓	✓
1010	Pass	0.99	98.1	4	2	18	1	✓	✓	✓
1019	Pass	0.95	NA	3	2	15	6	✓	✓	✓
1020	Pass	0.96	NA	3	2	13	1	✓	✓	✓
1022	Pass	0.97	NA	3	2	13	7	✓	✓	✓
1032	Pass	0.96	NA	4	2	21	2	✓	✓	✓
1037	Pass	1	99.2	3	2	16	2	✓	✓	✓
1042	Pass	0.95	99.3	4	2	21	6	✓	✓	✓
1050	Pass	0.92	97.9	3	2	17	1	✓	✓	✓
1051	Pass	0.83	99.5	4	2	19	5	✓	✓	✓
1054	Pass	0.99	98.8	3	2	16	6	✓	✓	✓
1058	Pass	0.97	99.1	3	2	16	8	✓	✓	✓
1067	Pass	1	99.6	4	2	22	2	✓	✓	✓
1070	Pass	1	99.1	3	2	20	4	✓	✓	✓

1072	Pass	0.89	97	4	2	22	8	✓	✓	✓
1076	Pass	0.98	98.3	3	2	20	8	✓	✓	✓
1080	Pass	0.98	97.7	4	2	18	3	✓	✓	✓
1083	Pass	0.99	98.6	4	2	23	1	✓	✓	✓
1085	Pass	0.96	98.3	5	3	24	8	✓	✓	✓
1087	Pass	0.99	98.6	4	2	21	4	✓	✓	✓
1088	Pass	0.94	97.6	3	2	17	7	✓	✓	✓
1094	Pass	0.81	94.8	4	2	23	7	✓	✓	✓
2010	Pass	0.76	98.2	4	2	25	2	✓	✓	✓
2012	Pass	0.93	97.4	4	2	21	8	✓	✓	✓
2013	Pass	0.98	96.1	4	2	25	8	✓	×	×
2028	Fail	0.1	95.8	3	2	20	6	×	×	×
2029	Pass	0.79	93.4	5	3	26	1	✓	✓	✓
2030	Pass	1	97.2	5	3	26	5	✓	✓	✓
2034	Pass	0.83	97.6	4	2	18	7	✓	✓	✓
2036	Pass	1	NA	4	2	19	1	✓	✓	×
2040	Pass	0.95	97.6	4	2	22	4	✓	✓	✓
2042	Pass	0.87	93.4	5	3	27	2	✓	✓	✓
2044	Pass	0.87	99.1	4	2	19	7	✓	✓	✓
2045	Fail	0.52	91.7	5	3	28	1	×	×	×
2047	Pass	0.98	98.5	4	2	22	6	✓	✓	✓
2052	Pass	0.73	96	4	2	23	5	✓	✓	✓
2054	Pass	0.93	97.1	4	2	23	3	✓	✓	✓
2062	Pass	0.95	96.8	4	2	25	4	✓	✓	✓
2067	Pass	0.98	96	4	2	25	6	✓	✓	✓
2072	Pass	0.88	95.2	5	3	28	5	✓	✓	✓
2075	Pass	0.99	97.4	5	3	26	3	✓	✓	✓
2078	Pass	1	96.7	5	3	27	4	✓	✓	✓
2086	Pass	0.99	98.5	5	3	27	6	✓	✓	✓
2087	Pass	0.97	98.7	5	3	27	8	✓	✓	✓
2090	Pass	0.95	96.5	5	3	24	2	✓	✓	✓
2133	Pass	1	97.8	5	3	29	3	✓	✓	×
2140	Pass	0.78	95	5	3	28	7	✓	✓	✓
2144	Pass	0.91	91.2	5	3	28	3	✓	✓	✓
2145	Pass	0.75	98.6	5	3	24	4	✓	✓	✓
2146	Pass	0.95	98.1	5	3	26	7	✓	×	×
2148	Pass	0.99	98.5	5	3	29	5	✓	✓	✓
2222	Pass	0.97	NA	3	2	14	5	✓	✓	✓
2231	Pass	1	NA	5	3	24	6	✓	✓	✓
2233	Pass	0.98	NA	5	3	29	1	✓	✓	✓
2257	Pass	0.96	98.8	5	3	30	2	✓	✓	✓
2261	Pass	0.98	97.8	5	3	29	7	✓	✓	✓
2264	Pass	0.92	97.9	5	3	8	2	✓	✓	✓

B cell sample processing and quality control information

EA ID	QC	B Purity Meth (%)	B Purity Flow (%)	Conv. Batch	iScan Batch	Array ID	Array Position	Diff. Analysis	meQTL	eQTM
826	Pass	0.97	84.5	1	1	1	2	✓	✓	×
827	Pass	0.94	NA	1	1	1	6	✓	✓	✓
834	Pass	0.98	NA	5	3	31	8	✓	✓	×
835	Pass	0.88	92.1	1	1	1	4	✓	✓	×
838	Pass	0.92	90.5	1	1	2	1	✓	✓	×
841	Pass	0.82	NA	1	1	1	8	✓	×	×
842	Pass	0.88	71.5	1	1	2	3	✓	✓	✓
845	Pass	0.84	NA	1	1	2	7	✓	✓	✓
854	Pass	0.7	NA	1	1	2	5	✓	✓	✓
867	Pass	1	91.7	1	1	3	8	✓	✓	✓
868	Pass	0.98	93	1	1	3	2	✓	✓	✓
874	Pass	0.98	96.7	1	1	4	3	✓	✓	✓
878	Pass	0.84	92.7	1	1	3	4	✓	✓	✓
879	Pass	0.88	NA	1	1	4	1	✓	✓	✓
882	Pass	0.98	95.3	1	1	4	5	✓	✓	✓
884	Pass	0.96	NA	1	1	4	7	✓	✓	✓
891	Pass	0.98	95	1	1	5	6	✓	✓	✓
892	Pass	0.95	93.2	1	1	5	4	✓	✓	✓
893	Pass	0.97	95.1	1	1	5	2	✓	✓	✓
898	Pass	0.96	91.4	1	1	5	8	✓	✓	×
905	Pass	0.99	91.8	1	1	6	3	✓	✓	✓
912	Pass	0.97	92.8	1	1	6	5	✓	✓	✓
915	Pass	0.95	96	1	1	3	6	✓	✓	✓
923	Pass	0.86	NA	1	1	6	7	✓	✓	✓
926	Pass	0.78	NA	6	4	7	1	✓	✓	✓
929	Pass	0.92	88.2	6	4	7	7	✓	✓	✓
930	Pass	0.92	94.8	6	4	9	6	✓	✓	✓
932	Pass	0.9	92.8	6	4	11	1	✓	✓	✓
937	Pass	0.96	98.6	6	4	7	5	✓	✓	✓
938	Pass	0.8	NA	1	1	6	1	✓	✓	✓
944	Pass	0.79	NA	6	4	9	2	✓	✓	✓
946	Pass	0.96	91.4	6	4	7	3	✓	✓	✓
948	Pass	0.92	94.1	3	2	15	1	✓	✓	✓
954	Pass	0.89	91.6	6	4	9	4	✓	✓	✓
957	Pass	0.98	95.6	6	4	10	6	✓	✓	✓
962	Pass	0.67	93.2	6	4	9	8	✓	✓	✓
965	Pass	0.85	NA	6	4	11	3	✓	✓	✓
967	Pass	0.71	87.5	6	4	11	7	✓	✓	✓
969	Pass	0.92	93.2	6	4	11	5	✓	✓	✓
973	Pass	0.93	95.3	6	4	10	4	✓	✓	✓

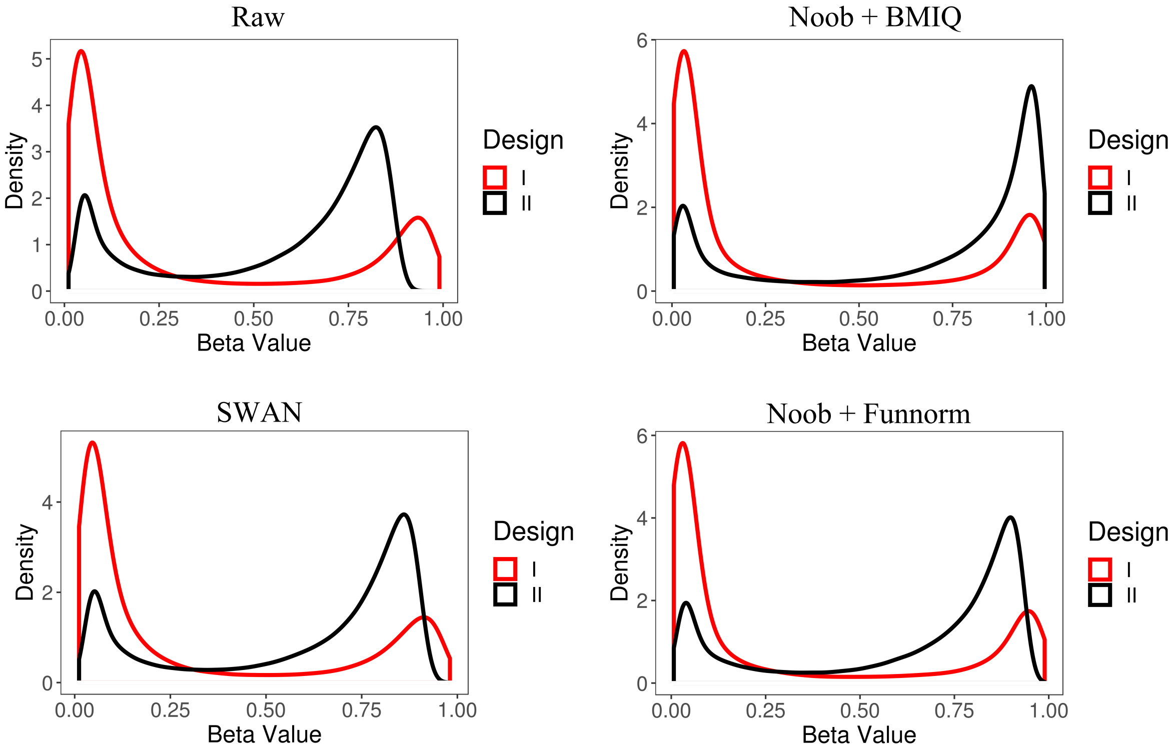
974	Pass	0.65	94.8	6	4	10	8	✓	✓	✓
975	Pass	0.97	96.3	6	4	12	3	✓	✓	✓
978	Pass	0.96	94	3	2	14	8	✓	✓	✓
980	Pass	0.96	97	6	4	12	1	✓	✓	✓
992	Pass	0.95	96.5	3	2	14	4	✓	✓	✓
995	Pass	0.9	NA	3	2	16	1	✓	✓	✓
996	Pass	1	97.7	4	2	21	1	✓	✓	✓
1000	Pass	0.8	94	4	2	25	3	✓	✓	✓
1003	Pass	0.97	96.7	3	2	13	6	✓	✓	✓
1005	Pass	0.94	93.2	5	3	32	7	✓	✓	✓
1008	Pass	0.9	NA	6	4	12	5	✓	✓	✓
1010	Pass	0.84	91.4	3	2	16	7	✓	×	×
1019	Pass	0.84	NA	6	4	12	7	✓	✓	✓
1020	Pass	0.75	NA	3	2	14	6	✓	✓	✓
1022	Pass	0.83	NA	3	2	14	2	✓	×	×
1032	Pass	0.92	NA	4	2	22	5	✓	✓	✓
1037	Pass	0.86	NA	6	4	10	2	✓	✓	✓
1050	Fail	0.46	95.7	3	2	15	3	×	✓	✓
1054	Pass	0.99	NA	3	2	15	5	✓	✓	✓
1058	Pass	0.98	NA	3	2	15	7	✓	✓	✓
1067	Fail	0.52	NA	3	2	20	7	×	✓	✓
1070	Pass	0.92	92.2	3	2	13	2	✓	✓	✓
1072	Pass	0.92	96.7	3	2	17	8	✓	✓	✓
1076	Pass	0.83	NA	3	2	16	3	✓	✓	✓
1080	Pass	1	97.8	3	2	16	5	✓	✓	✓
1083	Pass	0.97	94.6	3	2	20	5	✓	✓	✓
1085	Pass	0.99	NA	3	2	13	4	✓	✓	✓
1087	Pass	1	98	3	2	17	6	✓	✓	✓
1088	Pass	0.93	90.6	4	2	21	5	✓	×	×
1094	Pass	0.93	91.7	4	2	18	8	✓	×	×
2010	Pass	0.92	NA	4	2	19	6	✓	✓	✓
2012	Pass	0.91	92.9	3	2	17	4	✓	✓	✓
2013	Pass	0.94	NA	4	2	23	4	✓	✓	✓
EA2028	Pass	0.99	94.6	3	2	13	8	✓	✓	✓
2029	Pass	0.93	91.1	4	2	18	4	✓	✓	✓
2030	Pass	0.98	93.4	4	2	21	7	✓	×	×
2034	Pass	0.97	95.3	3	2	20	1	✓	✓	×
2036	Pass	0.96	NA	5	3	30	3	✓	✓	✓
2040	Pass	0.96	95.6	3	2	17	2	✓	✓	✓
2042	Fail	0.64	NA	4	2	25	5	×	×	×
2044	Pass	0.79	NA	4	2	22	3	✓	✓	✓
2045	Pass	0.88	95	5	3	26	8	✓	✓	✓
2047	Pass	1	97.9	3	2	20	3	✓	✓	✓
2052	Pass	0.94	93.3	4	2	18	6	✓	×	×

2054	Pass	0.92	94.9	4	2	19	4	✓	✓	✓
2062	Pass	0.93	93.6	4	2	22	7	✓	×	×
2067	Fail	0.38	95.5	4	2	21	3	×	✓	✓
2072	Pass	0.96	93	5	3	32	3	✓	✓	✓
2075	Fail	0.29	NA	4	2	23	6	×	✓	×
2078	Pass	0.9	94.5	4	2	25	7	✓	✓	✓
2086	Pass	0.97	95.6	5	3	26	4	✓	✓	✓
2087	Pass	0.91	93.9	5	3	26	6	✓	✓	✓
2090	Pass	0.69	NA	5	3	8	6	✓	✓	✓
2133	Pass	0.99	94.9	5	3	31	4	✓	×	×
2140	Fail	0.74	NA	5	3	28	2	×	✓	✓
2145	Pass	0.98	95.7	5	3	27	7	✓	✓	✓
2146	Pass	0.97	93.4	4	2	18	2	✓	×	×
2166	Pass	0.94	94.1	5	3	27	5	✓	✓	×
2197	Pass	0.88	97	4	2	19	2	✓	✓	✓
2200	Pass	1	96.9	5	3	28	6	✓	×	×
2203	Pass	0.93	97.5	5	3	28	8	✓	✓	✓
2222	Pass	0.93	NA	5	3	24	5	✓	✓	✓
2231	Pass	0.95	NA	5	3	24	7	✓	✓	✓
2233	Pass	0.99	96	5	3	32	5	✓	✓	✓
2257	Pass	0.97	95.1	4	2	22	1	✓	✓	✓
2261	Pass	0.99	95.9	5	3	29	4	✓	✓	✓
2264	Pass	0.98	94.4	5	3	8	7	✓	✓	✓
2265	Pass	0.81	60	5	3	32	1	✓	✓	×
2281	Pass	0.85	87.8	5	3	31	2	✓	✓	✓
2305	Pass	0.93	92.6	4	2	19	8	✓	✓	✓
2311	Pass	1	98.1	5	3	29	6	✓	✓	✓
2322	Pass	0.94	90.9	5	3	31	6	✓	✓	✓
2330	Pass	0.98	97.3	4	2	23	2	✓	✓	✓
2345	Pass	0.87	91.3	4	2	23	8	✓	✓	✓
2348	Pass	0.94	94.7	4	2	25	1	✓	✓	✓
2367	Pass	0.96	95.5	5	3	27	1	✓	✓	✓
2368	Pass	0.89	91.5	5	3	26	2	✓	✓	✓
2369	Pass	0.99	93.9	5	3	8	3	✓	×	×
2378	Pass	0.97	94.4	5	3	27	3	✓	×	×
2379	Pass	0.97	NA	5	3	28	4	✓	✓	✓
2390	Pass	0.99	NA	5	3	24	3	✓	✓	✓
2416	Pass	0.98	NA	5	3	29	2	✓	×	×
2437	Pass	0.92	NA	5	3	29	8	✓	✓	✓
2439	Pass	0.97	NA	5	3	8	1	✓	✓	✓
2476	Pass	0.97	NA	5	3	32	4	✓	✓	✓
2493	Pass	0.99	NA	5	3	32	6	✓	✓	✓
2506	Pass	0.99	NA	5	3	31	7	✓	✓	✓
2507	Pass	0.99	NA	5	3	8	5	✓	✓	✓

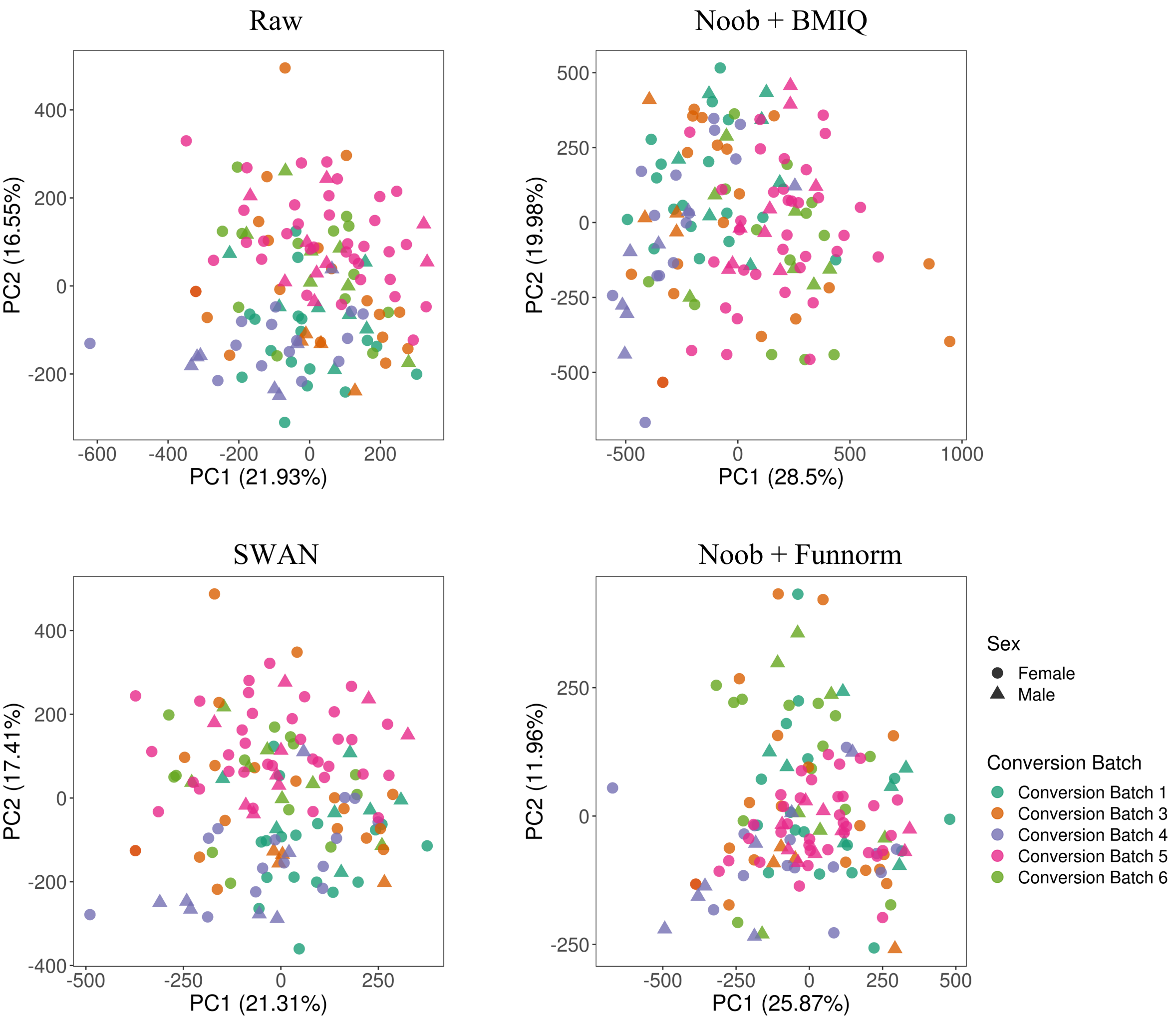
2510	Pass	0.99	NA	5	3	8	8	✓	✓	✓
2511	Pass	0.98	NA	5	3	32	2	✓	✓	✓
2519	Pass	0.93	NA	5	3	32	8	✓	✓	✓
2520	Pass	0.98	NA	5	3	31	1	✓	×	×
2548	Pass	1	NA	5	3	31	3	✓	×	×
2549	Pass	0.95	NA	5	3	31	5	✓	✓	✓
2759	Pass	1	NA	5	3	30	1	✓	✓	✓
2767	Pass	0.97	NA	5	3	30	4	✓	✓	×

Appendix B – Comparison of Normalisation Methods for B cell samples

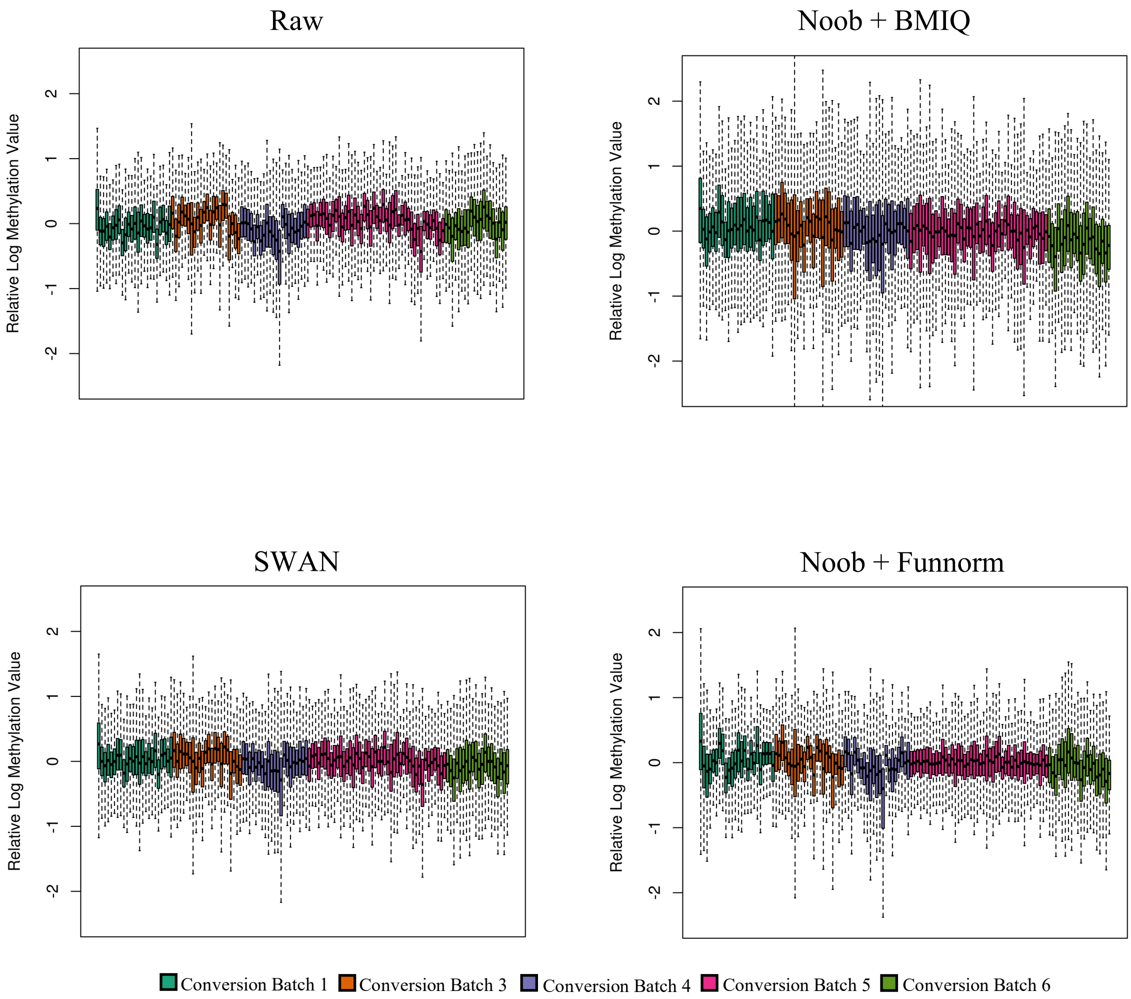
Plots displayed here are analogous to those presented for the CD4^{+} T cell samples in Chapter 3.4.2



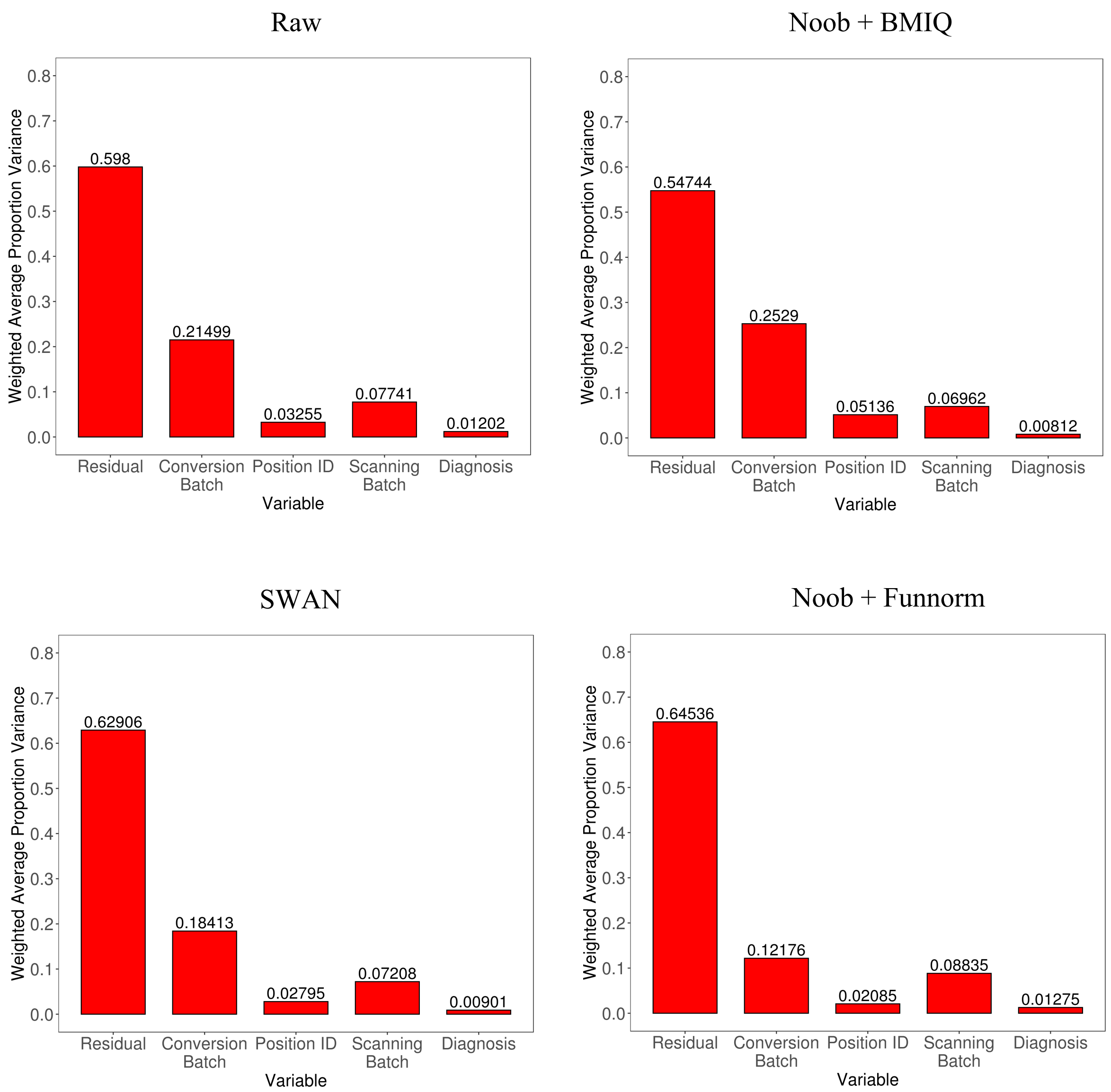
DNA methylation (β-value) for each Illumina probe type (Type I & II) in raw B cell data (pre-normalisation), as well as following normalisation of data using Normal-exponential using out-of-band probes (noob) with Beta mixture quantile dilation (BMIQ), Subset-quantile within array normalization (SWAN), and noob with functional normalisation (Funnorm).



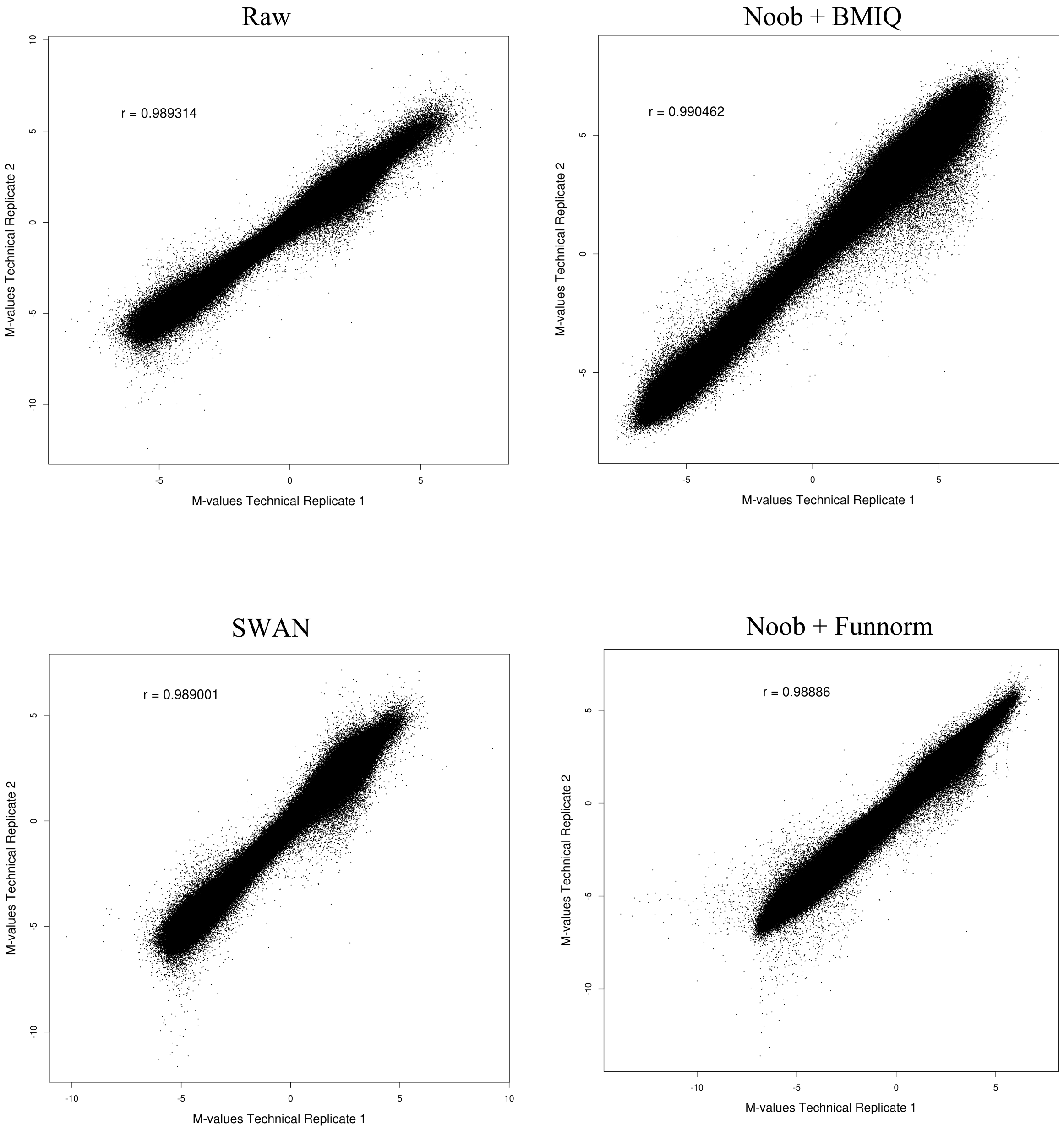
Principal component analysis of sample processing batch in raw B cell data (pre-normalisation), as well as following normalisation of data using Normal-exponential using out-of-band probes (noob) with Beta mixture quantile dilation (BMIQ), Subset-quantile within array normalization (SWAN), and noob with functional normalisation (Funnorm).



Relative log methylation (RLM; see Chapter 3.4.2 for further details) of sample processing batch in raw B cell data (pre-normalisation), as well as following normalisation of data using Normal-exponential using out-of-band probes (noob) with Beta mixture quantile dilation (BMIQ), Subset-quantile within array normalization (SWAN), and noob with functional normalisation (Funnorm).



Principal variance component analysis (PVCA) of raw B cell data (pre-normalisation), as well as following normalisation of data using Normal-exponential using out-of-band probes (noob) with Beta mixture quantile dilation (BMIQ), Subset-quantile within array normalization (SWAN), and noob with functional normalisation (Funnorm).



Pearson's correlation of B cell technical replicates in raw B cell data (pre-normalisation), as well as following normalisation of data using Normal-exponential using out-of-band probes (noob) with Beta mixture quantile dilation (BMIQ), Subset-quantile within array normalization (SWAN), and noob with functional normalisation (Funnorm).

Appendix C – Differentially-methylated position analysis

Top 100 CpGs by nominal p-value from the RA vs. non-RA differential analysis in CD4^{+} T cell and B cell samples. Chrom = CpG chromosome; Pos = CpG base pair position on chromosome; Δβ = Difference in mean methylation (β-value, 0-1) in RA patients relative to the non-RA controls; UCSC RefGene Name = RefGene to which the CpG maps; Relation to CpG Island = mapping of CpG to CpG island feature (N_Shore = North shore, N_Shelf = North shelf, S_Shelf = South shelf, S_Shore = South shore, see Chapter 2.7.4); RefGene feature = mapping of CpG genic feature (TSS200 – 0-200 base pairs from transcription start site, TSS1500 = 200-1500 base pairs from transcription start site, UTR = untranslated region, see Chapter 2.7.4).

Top 100 CpGs ranked by nominal p-value from the CD4^{+} T cell RA vs. non-RA differential analysis (Δβ ≥ 0.05)

CpG	Chrom	Pos	P-value	Adjusted p-value	Δβ (RA vs. non-RA	UCSC RefGene	Relation to CpG Island	RefGene Feature
cg21289466	20	5,577,716	5.02E-06	0.993	0.062	GPCPD1	OpenSea	Body
cg24245216	19	7,004,657	1.70E-04	0.993	-0.180	-	OpenSea	Intergenic
cg11945167	8	4,644,739	3.50E-04	0.993	0.060	CSMD1	OpenSea	Body
cg15563420	12	86,026,194	3.70E-04	0.993	0.052	-	OpenSea	Intergenic
cg00080972	5	178,986,291	4.67E-04	0.993	-0.086	RUFY1	N_Shore	TSS1500
cg07612827	19	7,005,180	5.12E-04	0.993	-0.070	FLJ25758	OpenSea	Body
cg11787167	14	33,407,370	6.62E-04	0.993	0.105	NPAS3	S_Shelf	TSS1500
cg18471635	11	104,769,411	7.53E-04	0.993	-0.062	CASP12	OpenSea	TSS200
cg03161803	6	27,649,120	7.59E-04	0.993	0.056	-	S_Shore	Intergenic
cg05287483	20	5,551,376	8.12E-04	0.993	0.070	GPCPD1	OpenSea	Body
cg26516362	5	178,986,906	8.55E-04	0.993	-0.069	RUFY1	Island	5'UTR
cg07891761	19	35,861,642	8.60E-04	0.993	0.084	-	OpenSea	Intergenic
cg05457628	5	178,986,728	9.52E-04	0.993	-0.067	RUFY1	Island	TSS200
cg17285144	22	22,532,640	9.85E-04	0.993	0.112	-	OpenSea	Intergenic
cg02671281	9	95,783,395	1.18E-03	0.993	0.054	FGD3	OpenSea	Body
cg16766914	17	55,962,703	1.25E-03	0.993	-0.053	CUEDC1	Island	Body
cg14451627	9	115,987,035	1.39E-03	0.993	0.071	SLC31A1	S_Shelf	5'UTR
cg22764044	5	178,986,830	1.47E-03	0.993	-0.057	RUFY1	Island	1stExon
cg25658438	5	178,986,372	1.56E-03	0.993	-0.090	RUFY1	N_Shore	TSS1500
cg06118287	5	178,986,559	1.58E-03	0.993	-0.059	RUFY1	Island	TSS200
cg11424828	8	2,075,469	1.78E-03	0.993	0.132	MYOM2	Island	Body
cg08366828	5	71,683,884	2.07E-03	0.993	0.127	-	OpenSea	Intergenic
cg05624577	15	81,411,055	2.21E-03	0.993	-0.066	-	Island	Intergenic
cg15975806	13	110,319,607	2.23E-03	0.993	0.065	-	OpenSea	Intergenic
cg05056638	8	24,800,824	2.26E-03	0.993	-0.054	-	S_Shore	Intergenic
cg17951445	4	30,842,020	2.45E-03	0.993	0.055	PCDH7	OpenSea	Body
cg25456477	12	86,230,367	2.52E-03	0.993	-0.056	RASSF9	OpenSea	TSS200
cg09060608	5	178,986,726	2.62E-03	0.993	-0.057	RUFY1	Island	TSS200
cg16060867	19	31,576,236	2.64E-03	0.993	0.058	-	OpenSea	Intergenic

cg07053672	17	32,957,113	2.78E-03	0.993	-0.050	TMEM132E	S_Shelf	Body
cg10653456	12	57,561,993	3.09E-03	0.993	0.071	LRP1	OpenSea	Body
cg19827854	7	17,868,819	3.29E-03	0.993	-0.054	SNX13	OpenSea	Body
cg11399539	18	19,882,372	3.41E-03	0.993	-0.066	-	OpenSea	Intergenic
cg08058472	5	178,986,638	3.53E-03	0.993	-0.061	RUFY1	Island	TSS200
cg16896868	13	110,319,562	3.80E-03	0.993	0.068	-	OpenSea	Intergenic
cg26203582	7	38,236,541	3.88E-03	0.993	0.062	STARD3NL	OpenSea	5'UTR
cg13423887	6	32,632,694	3.94E-03	0.993	-0.146	HLA-DQB1	Island	Body
cg02978220	1	108,337,960	4.50E-03	0.993	0.078	VAV3	OpenSea	Body
cg19626725	5	178,986,131	4.56E-03	0.993	-0.060	RUFY1	N_Shore	TSS1500
cg14762436	7	24,917,750	4.65E-03	0.993	-0.080	OSBPL3	OpenSea	Body
cg15037581	13	94,535,214	4.79E-03	0.993	-0.051	GPC6	OpenSea	Body
cg12448452	10	59,715,224	5.08E-03	0.993	0.075	-	OpenSea	Intergenic
cg24087438	17	3,704,482	5.29E-03	0.993	0.091	ITGAE	OpenSea	1stExon
cg09139047	6	32,552,042	5.32E-03	0.993	-0.120	HLA-DRB1	Island	Body
cg25817165	18	72,167,213	5.59E-03	0.993	0.060	CNDP2	S_Shelf	1stExon
cg20426698	3	65,960,357	5.61E-03	0.993	0.057	MAGI1	OpenSea	Body
cg19683494	5	74,908,142	5.80E-03	0.993	-0.077	-	S_Shore	Intergenic
cg02839725	17	30,823,006	5.84E-03	0.993	-0.133	MYO1D	Island	Body
cg03458265	14	81,408,650	6.11E-03	0.993	0.051	-	S_Shore	Intergenic
cg19205037	2	8,736,346	6.20E-03	0.993	0.051	-	OpenSea	Intergenic
cg23043514	5	179,870,089	6.26E-03	0.993	0.054	-	OpenSea	Intergenic
cg24413597	1	162,155,427	7.16E-03	0.993	-0.061	NOS1AP	OpenSea	Body
cg08214808	11	45,922,166	7.58E-03	0.993	-0.050	MAPK8IP1	Island	Body
cg13422161	12	52,773,842	7.83E-03	0.993	0.056	KRT84	OpenSea	Body
cg05387464	2	9,956,256	8.29E-03	0.993	-0.053	-	OpenSea	Intergenic
cg08221350	10	134,983,838	8.32E-03	0.993	-0.079	KNDC1	S_Shelf	Body
cg08358620	12	86,230,403	8.45E-03	0.993	-0.057	RASSF9	OpenSea	TSS200
cg10120522	3	69,891,885	8.54E-03	0.993	0.076	MITF	OpenSea	Body
cg03607220	6	32,526,263	8.71E-03	0.993	0.072	HLA-DRB6	OpenSea	Body
cg08627981	20	1,757,237	9.04E-03	0.993	0.072	LOC100289473	N_Shore	Body
cg15134106	1	95,974,671	9.56E-03	0.993	0.057	LOC100996635	OpenSea	TSS1500
cg18304483	19	35,853,396	9.85E-03	0.993	0.071	-	OpenSea	Intergenic
cg17212350	7	87,075,377	1.02E-02	0.993	-0.085	ABCB4	OpenSea	Body
cg16677969	10	85,677,559	1.12E-02	0.993	0.067	-	OpenSea	Intergenic
cg16820615	13	114,884,918	1.14E-02	0.993	0.056	RASA3	OpenSea	Body
cg12100178	1	14,437,715	1.14E-02	0.993	-0.084	-	OpenSea	Intergenic
cg03299990	20	1,757,570	1.20E-02	0.993	0.120	-	N_Shore	Intergenic
cg27467591	5	7,046,940	1.21E-02	0.993	-0.056	-	OpenSea	Intergenic
cg24915592	13	110,319,578	1.24E-02	0.993	0.052	-	OpenSea	Intergenic
cg21434132	3	196,705,742	1.29E-02	0.993	-0.092	-	OpenSea	Intergenic
cg16762802	19	15,649,508	1.31E-02	0.993	-0.083	CYP4F22	OpenSea	Body
cg19105674	13	114,305,226	1.33E-02	0.993	0.058	ATP4B	S_Shelf	Body
cg16425314	4	97,273,416	1.34E-02	0.993	0.051	-	OpenSea	Intergenic
cg12162424	21	15,646,312	1.39E-02	0.993	-0.052	ABCC13	OpenSea	Body
cg03351301	17	46,969,163	1.43E-02	0.993	-0.064	ATP5G1	N_Shore	TSS1500

cg21144063	7	64,035,529	1.52E-02	0.993	-0.054	-	OpenSea	Intergenic
cg16399632	4	1,244,006	1.58E-02	0.993	-0.052	CTBP1	Island	TSS1500
cg19922198	6	51,963,711	1.61E-02	0.993	-0.052	-	OpenSea	Intergenic
cg22334681	3	14,257,893	1.68E-02	0.993	-0.079	-	OpenSea	Intergenic
cg05740244	11	18,434,015	1.71E-02	0.993	-0.067	LDHC	OpenSea	5'UTR
cg12584458	6	28,447,107	1.72E-02	0.993	0.050	-	OpenSea	Intergenic
cg11488033	3	196,705,898	1.80E-02	0.993	-0.056	-	OpenSea	Intergenic
cg07870920	4	121,569,769	1.81E-02	0.993	-0.087	-	OpenSea	Intergenic
cg16345520	14	60,432,168	1.81E-02	0.993	0.056	LRRC9	OpenSea	Body
cg16792464	1	22,250,284	1.84E-02	0.993	0.056	HSPG2	OpenSea	Body
cg03570708	5	133,135,701	1.86E-02	0.993	0.069	-	OpenSea	Intergenic
cg05107650	3	141,160,397	1.88E-02	0.993	0.066	ZBTB38	OpenSea	5'UTR
cg16597280	14	60,432,675	1.90E-02	0.993	0.054	LRRC9	OpenSea	Body
cg03483727	15	33,487,681	1.92E-02	0.993	0.071	FMN1	S_Shore	TSS1500
cg21847720	8	2,075,777	1.96E-02	0.993	0.097	MYOM2	Island	Body
cg13892472	8	139,847,905	2.04E-02	0.993	0.071	COL22A1	OpenSea	Body
cg00169354	20	54,922,827	2.06E-02	0.993	-0.051	-	S_Shelf	Intergenic
cg24445388	1	2,084,391	2.12E-02	0.993	-0.063	PRKCZ	S_Shore	Body
cg16814680	8	91,681,699	2.15E-02	0.993	-0.117	-	OpenSea	Intergenic
cg25051134	2	181,938,260	2.16E-02	0.993	-0.051	-	OpenSea	Intergenic
cg26197874	9	137,003,621	2.17E-02	0.993	0.054	WDR5	S_Shore	TSS1500
cg22444562	2	207,090,465	2.19E-02	0.993	0.084	GPR1-AS	OpenSea	Body
cg04278296	7	36,696,838	2.20E-02	0.993	0.057	AOAH	OpenSea	Body
cg07721756	7	48,019,419	2.21E-02	0.993	-0.063	HUS1	S_Shore	TSS200
cg06339162	4	37,938,710	2.22E-02	0.993	0.066	TBC1D1	OpenSea	Body

Top 100 CpGs ranked by nominal p-value from the B cell RA vs. non-RA differential analysis (Δβ ≥ 0.05)

CpG	Chrom	Pos	P-value	Adjusted p-value	Δβ (RA vs. non-RA	UCSC RefGene	Relation to CpG Island	RefGene Feature
cg00595030	19	10,398,582	6.24E-06	0.890	-0.061	ICAM4	Island	3'UTR
cg10800620	2	196,398,826	6.26E-05	0.890	0.058	-	OpenSea	Intergenic
cg06323052	4	56,720,686	7.94E-05	0.890	0.052	EXOC1	S_Shore	5'UTR
cg25152193	1	197,874,469	2.80E-04	0.890	-0.060	C1orf53	S_Shelf	Body
cg22901297	6	32,522,795	2.95E-04	0.890	0.085	HLA-DRB6	OpenSea	Body
cg00538212	7	158,751,591	3.22E-04	0.890	0.056	-	N_Shore	Intergenic
cg21419137	8	87,905,504	3.95E-04	0.890	-0.062	CNBD1	OpenSea	Body
cg16055526	6	33,083,287	4.37E-04	0.890	-0.091	HLA-DPB2	N_Shore	Body
cg22404498	22	32,600,722	4.94E-04	0.890	-0.056	RFPL2	OpenSea	TSS1500
cg08666831	19	47,507,691	5.46E-04	0.890	-0.058	GRLF1	Island	3'UTR
cg17688837	4	25,090,665	5.74E-04	0.890	0.051	-	S_Shore	Intergenic
cg20673407	10	31,040,939	6.20E-04	0.890	-0.205	-	OpenSea	Intergenic
cg11655243	5	140,778,396	6.73E-04	0.890	-0.065	PCDHGA4	N_Shore	Body
cg17633222	10	63,808,249	8.11E-04	0.890	0.060	ARID5B	OpenSea	TSS1500

cg27619385	1	32,729,615	8.66E-04	0.890	0.051	LCK	OpenSea	5'UTR
cg16263980	12	114,079,546	8.98E-04	0.890	-0.061	-	OpenSea	Intergenic
cg25299227	14	76,940,165	8.98E-04	0.890	0.095	ESRRB	OpenSea	Body
cg18848287	7	5,111,641	1.08E-03	0.890	-0.058	LOC389458	Island	TSS200
cg03055671	3	172,231,528	1.09E-03	0.890	0.054	TNFSF10	OpenSea	Body
cg26262055	4	25,090,618	1.14E-03	0.890	0.059	-	S_Shore	Intergenic
cg19312667	4	25,090,491	1.33E-03	0.890	0.076	-	Island	Intergenic
cg03935359	1	68,363,156	1.42E-03	0.890	0.079	-	OpenSea	Intergenic
cg03803940	7	158,532,542	1.44E-03	0.890	0.060	ESYT2	OpenSea	Body
cg20988098	6	157,931,791	1.65E-03	0.890	0.051	ZDHHC14	OpenSea	Body
cg26660177	18	24,791,755	1.78E-03	0.890	-0.063	-	OpenSea	Intergenic
cg17686260	10	131,412,764	1.93E-03	0.890	0.102	MGMT	OpenSea	Body
cg19336497	11	14,380,999	1.96E-03	0.890	0.051	RRAS2	S_Shore	TSS1500
cg24118521	13	47,472,330	2.00E-03	0.890	0.068	HTR2A	OpenSea	TSS1500
cg05056638	8	24,800,824	2.09E-03	0.890	-0.059	-	S_Shore	Intergenic
cg09083742	8	59,043,488	2.32E-03	0.890	0.065	FAM110B	OpenSea	5'UTR
cg00415333	6	137,493,017	2.36E-03	0.890	0.052	IL22RA2	OpenSea	5'UTR
cg08247564	12	51,835,517	2.70E-03	0.890	-0.059	SLC4A8	OpenSea	5'UTR
cg20690458	13	111,174,402	3.03E-03	0.890	0.100	-	OpenSea	Intergenic
cg10315076	4	56,251,808	3.09E-03	0.890	0.053	SRD5A3-AS1	OpenSea	TSS200
cg11195926	4	145,621,327	3.10E-03	0.890	-0.060	HHIP	OpenSea	Body
cg08249075	1	58,824,309	3.25E-03	0.890	-0.053	-	OpenSea	Intergenic
cg02573091	5	74,908,125	3.72E-03	0.890	-0.095	-	S_Shore	Intergenic
cg24635402	19	44,352,649	3.75E-03	0.890	-0.085	ZNF283	OpenSea	Body
cg10644073	15	86,018,136	3.92E-03	0.890	0.068	AKAP13	OpenSea	5'UTR
cg19325793	6	33,082,165	4.03E-03	0.890	-0.090	HLA-DPB2	N_Shelf	Body
cg25371762	21	47,674,425	4.05E-03	0.890	0.055	MCM3AP	OpenSea	ExonBnd
cg12681972	6	26,225,299	4.05E-03	0.890	-0.056	HIST1H3E	N_Shore	TSS200
cg09219839	7	158,574,563	4.14E-03	0.890	0.052	ESYT2	OpenSea	Body
cg21169940	1	202,251,908	4.22E-03	0.890	0.071	LGR6	OpenSea	Body
cg02079122	1	240,262,933	4.24E-03	0.890	0.055	FMN2	OpenSea	Body
cg08736526	4	88,656,433	4.37E-03	0.890	0.172	-	OpenSea	Intergenic
cg02428792	4	25,090,597	4.43E-03	0.890	0.069	-	S_Shore	Intergenic
cg17758363	20	13,199,787	4.48E-03	0.890	-0.052	-	N_Shore	Intergenic
cg16131272	8	58,201,111	4.54E-03	0.890	-0.050	-	OpenSea	Intergenic
cg21104965	6	117,869,453	4.93E-03	0.890	-0.060	DCBLD1	Island	Body
cg06759629	4	25,090,198	4.99E-03	0.890	0.053	-	Island	Intergenic
cg07891761	19	35,861,642	5.03E-03	0.890	0.098	-	OpenSea	Intergenic
cg05624577	15	81,411,055	5.16E-03	0.890	-0.077	-	Island	Intergenic
cg15365744	12	106,849,290	5.28E-03	0.890	-0.069	POLR3B	OpenSea	Body
cg24587835	17	5,674,234	5.30E-03	0.890	0.105	LOC339166	OpenSea	TSS1500
cg11969609	10	118,047,768	5.35E-03	0.890	-0.051	-	OpenSea	Intergenic
cg27401488	5	133,899,839	5.47E-03	0.890	0.062	JADE2	OpenSea	Body
cg23881368	13	47,472,343	5.54E-03	0.890	0.059	HTR2A	OpenSea	TSS1500
cg13219362	7	7,860,864	5.59E-03	0.890	0.076	UMAD1	OpenSea	Body
cg12242038	5	135,418,206	5.88E-03	0.890	-0.086	-	S_Shore	Intergenic

cg16896868	13	110,319,562	6.17E-03	0.890	0.068	-	OpenSea	Intergenic
cg11198398	10	31,054,241	6.20E-03	0.890	-0.080	-	OpenSea	Intergenic
cg12624096	1	211,780,386	6.50E-03	0.890	0.053	-	OpenSea	Intergenic
cg14771766	9	82,345,430	6.70E-03	0.890	0.086	-	OpenSea	Intergenic
cg00901687	17	48,585,270	7.00E-03	0.890	0.086	MYCBPAP	N_Shore	TSS1500
cg19626725	5	178,986,131	7.24E-03	0.890	-0.058	RUFY1	N_Shore	TSS1500
cg11724562	11	71,196,642	7.28E-03	0.890	-0.076	NADSYN1	OpenSea	Body
cg26893861	17	41,843,967	7.39E-03	0.890	-0.114	DUSP3	OpenSea	3'UTR
cg10142271	7	44,766,894	8.05E-03	0.890	0.052	-	OpenSea	Intergenic
cg08366828	5	71,683,884	8.15E-03	0.890	0.084	-	OpenSea	Intergenic
cg11440486	17	48,585,216	8.25E-03	0.890	0.075	MYCBPAP	N_Shore	TSS1500
cg17133734	15	86,042,851	8.52E-03	0.890	0.053	AKAP13	OpenSea	Body
cg06485460	18	2,084,004	8.53E-03	0.890	-0.059	-	OpenSea	Intergenic
cg01218619	4	25,090,298	8.64E-03	0.890	0.086	-	Island	Intergenic
cg21598751	1	159,791,473	8.88E-03	0.890	-0.112	-	OpenSea	Intergenic
cg18297445	15	81,578,048	8.95E-03	0.890	0.062	IL16	OpenSea	ExonBnd
cg06339162	4	37,938,710	9.00E-03	0.890	0.067	TBC1D1	OpenSea	Body
cg15322070	8	96,814,318	9.01E-03	0.890	0.053	C8orf37-AS1	OpenSea	Body
cg03489016	2	230,124,603	9.03E-03	0.890	-0.068	PID1	OpenSea	Body
cg06771126	4	57,547,699	9.09E-03	0.890	-0.053	HOPX	OpenSea	5'UTR
cg16591159	16	31,487,813	9.20E-03	0.890	-0.052	TGFB1I1	Island	Body
cg18200810	13	47,472,200	9.27E-03	0.890	0.076	HTR2A	OpenSea	TSS1500
cg02839725	17	30,823,006	9.28E-03	0.890	-0.134	MYO1D	Island	Body
cg21004358	4	187,422,119	9.77E-03	0.890	-0.075	-	Island	Intergenic
cg13081526	6	32,449,961	1.02E-02	0.890	0.098	-	OpenSea	Intergenic
cg26911830	1	157,219,222	1.06E-02	0.890	-0.067	-	OpenSea	Intergenic
cg09589057	14	106,351,578	1.07E-02	0.890	0.056	-	OpenSea	Intergenic
cg20111217	17	48,585,264	1.08E-02	0.890	0.069	MYCBPAP	N_Shore	TSS1500
cg03834793	4	25,310,627	1.10E-02	0.890	-0.056	-	N_Shelf	Intergenic
cg16762802	19	15,649,508	1.11E-02	0.890	-0.104	CYP4F22	OpenSea	Body
cg14481689	19	42,873,953	1.12E-02	0.890	-0.070	MEGF8	Island	Body
cg21964148	12	92,735,364	1.15E-02	0.890	0.055	-	OpenSea	Intergenic
cg17403731	19	613,505	1.23E-02	0.890	-0.055	HCN2	Island	Body
cg06683719	17	5,673,550	1.23E-02	0.890	0.051	-	OpenSea	Intergenic
cg14659662	1	54,151,053	1.27E-02	0.890	0.061	GLIS1	OpenSea	5'UTR
cg19683494	5	74,908,142	1.32E-02	0.890	-0.066	-	S_Shore	Intergenic
cg03545643	8	43,554,101	1.33E-02	0.890	-0.055	-	OpenSea	Intergenic
cg00101154	16	420,108	1.35E-02	0.890	-0.071	MRPL28	Island	Body
cg10596483	8	143,751,796	1.38E-02	0.890	-0.118	JRK	S_Shore	TSS1500
cg25813936	9	36,276,879	1.38E-02	0.890	0.056	GNE	OpenSea	Body

Appendix D – Differentially-variable position analysis

Top 100 CpGs by t-test p-value found to be differentially-variable (either RA hyper-variable or RA hypo-variable) in CD4+ T cells and B cells as identified by iEVORA (see Chapter 2.6.3). Var RA = DNA methylation variance in RA patients; Var nRA = DNA methylation variance in non-RA controls; P-value TT = T-test p-value; P-value Bt = Bartlett's test p-value; Adjusted p-value Bt = Adjusted Bartlett's test p-value; UCSC RefGene = gene to which the variable CpG maps.

Top 100 significant differentially variable positions in CD4+ T cells between RA patients and non-RA controls ranked by t-test p-value

CpG	Var RA	Var nRA	P-value TT	P-Value Bt	Adiusted P-value Bt	UCSC RefGene
cg15174564	1.94	0.21	8.60E-05	4.64E-12	2.61E-09	GRIK4
cg27284424	1.85	0.27	2.45E-04	7.37E-10	2.55E-07	LOC100506860
cg04797575	0.80	0.16	1.30E-03	1.27E-07	2.49E-05	GPM6A
ch.8.1995451R	1.29	0.23	1.33E-03	2.45E-08	5.80E-06	MATN2
cg04641168	2.38	0.48	1.87E-03	1.34E-07	2.61E-05	SRP72
cg13396858	0.76	0.15	2.24E-03	1.24E-07	2.44E-05	-
ch.4.1647744F	2.31	0.59	2.47E-03	4.80E-06	6.23E-04	WDFY3
cg00647389	0.06	0.62	2.66E-03	2.67E-16	4.00E-13	FSCN3
cg21421501	0.62	0.15	2.78E-03	1.67E-06	2.46E-04	FTH1
cg13565723	1.05	0.16	2.85E-03	1.08E-09	3.60E-07	PEX13
cg07871034	0.76	0.09	3.62E-03	2.48E-11	1.17E-08	-
cg26679958	1.96	0.19	3.74E-03	5.19E-13	3.70E-10	LEF1
cg13213009	0.89	0.20	3.92E-03	7.04E-07	1.15E-04	LY96
cg07891761	0.16	1.79	3.93E-03	2.00E-16	3.11E-13	-
cg21391815	0.69	0.13	4.22E-03	4.13E-08	9.22E-06	CGGBP1
cg07870920	3.91	0.79	4.45E-03	1.60E-07	3.06E-05	-
cg23032110	0.07	0.02	4.78E-03	7.91E-06	9.69E-04	LOC643339
cg18423635	0.18	1.06	4.82E-03	4.96E-10	1.78E-07	HCG2P7
cg07911953	1.15	0.24	4.83E-03	3.35E-07	5.92E-05	ADK
cg13990487	0.09	0.34	5.05E-03	1.99E-06	2.88E-04	UBR4
cg22052758	0.79	0.19	5.16E-03	2.08E-06	2.99E-04	ACCN2
cg04752352	0.29	0.07	5.25E-03	7.59E-07	1.22E-04	SERGEF
cg14451627	0.30	1.95	5.67E-03	3.56E-11	1.63E-08	SLC31A1
cg13956671	0.76	0.09	6.03E-03	6.64E-12	3.59E-09	CUL2
cg19537820	0.97	0.24	6.07E-03	3.53E-06	4.76E-04	REG4
cg13887021	1.06	0.21	6.44E-03	1.41E-07	2.74E-05	-
cg26994227	0.14	0.04	6.54E-03	6.85E-06	8.56E-04	RAPGEF6
cg02632185	0.34	0.04	6.55E-03	5.72E-12	3.14E-09	MAST4
cg11460110	0.06	0.20	6.71E-03	3.26E-06	4.44E-04	PRR3
cg09437423	1.04	0.26	6.73E-03	4.30E-06	5.65E-04	TRAT1
cg20426698	0.18	2.00	7.35E-03	1.64E-16	2.57E-13	MAGI1

cg26312542	0.06	0.22	7.40E-03	2.87E-06	3.98E-04	TEX12
cg16401214	0.31	0.07	7.76E-03	5.17E-07	8.73E-05	C16orf63
cg22875643	0.40	0.07	7.82E-03	1.28E-08	3.27E-06	MIER1
ch.4.255198F	1.48	0.18	7.85E-03	2.80E-11	1.30E-08	TBC1D14
ch.2.237314361R	0.49	0.08	8.27E-03	4.94E-09	1.39E-06	-
cg02978220	0.39	2.51	8.72E-03	4.17E-11	1.88E-08	VAV3
cg13230994	0.04	0.15	8.92E-03	1.79E-06	2.61E-04	KDM2A
cg10653456	0.34	1.97	9.39E-03	4.17E-10	1.52E-07	LRP1
cg18316974	0.96	0.13	1.01E-02	2.42E-10	9.28E-08	GFI1
cg14568768	0.93	0.14	1.04E-02	2.02E-09	6.23E-07	LCORL
cg17364114	0.10	0.39	1.05E-02	7.22E-07	1.17E-04	SLC22A3
cg23269169	1.03	0.10	1.07E-02	3.08E-13	2.33E-10	C1orf144
cg00591556	0.09	0.31	1.07E-02	7.20E-06	8.94E-04	HS6ST3
cg20096979	1.32	0.27	1.09E-02	2.08E-07	3.87E-05	HNRNPH3
cg16842060	0.43	0.10	1.09E-02	2.22E-06	3.17E-04	NFE2L2
cg11759930	0.03	0.13	1.11E-02	7.73E-07	1.24E-04	-
cg05770946	0.78	0.11	1.12E-02	3.80E-10	1.40E-07	SIM1
cg02221805	1.20	0.22	1.17E-02	2.87E-08	6.68E-06	TRIM59
ch.9.103369821R	0.60	0.13	1.18E-02	3.74E-07	6.53E-05	-
cg04951677	0.09	0.37	1.19E-02	6.09E-07	1.01E-04	-
ch.9.25704165R	2.00	0.38	1.22E-02	5.18E-08	1.13E-05	-
cg14360268	1.44	0.30	1.25E-02	2.02E-07	3.78E-05	C17orf62
cg13714403	1.49	0.16	1.27E-02	5.05E-12	2.82E-09	SIK3
cg22435810	0.39	0.09	1.28E-02	2.09E-06	3.01E-04	SMCR7
cg13483603	1.32	0.19	1.32E-02	9.12E-10	3.09E-07	OAS1
cg14365785	0.04	0.26	1.37E-02	1.27E-10	5.13E-08	MTUS2
cg14986104	0.93	0.16	1.38E-02	8.08E-09	2.17E-06	LINC01229
cg09817993	0.77	0.14	1.44E-02	1.91E-08	4.65E-06	GNL3
cg24079043	0.14	0.76	1.46E-02	1.68E-09	5.27E-07	MYO1E
cg04963948	0.86	0.22	1.49E-02	4.87E-06	6.32E-04	TPD52
cg04674291	0.08	1.20	1.49E-02	6.31E-20	2.01E-16	KSR2
cg21848624	1.88	0.46	1.51E-02	2.67E-06	3.72E-04	NFATC1
cg07872987	1.99	0.52	1.52E-02	8.03E-06	9.82E-04	CD96
cg24087438	0.68	4.93	1.52E-02	3.53E-12	2.05E-09	ITGAE
cg23460084	0.64	0.11	1.52E-02	1.28E-08	3.27E-06	CLPTM1
cg00686761	0.46	0.12	1.52E-02	6.74E-06	8.43E-04	CEP192
cg01273790	1.20	0.07	1.53E-02	3.56E-17	6.31E-14	PCBP2
cg02081454	0.80	0.18	1.53E-02	1.15E-06	1.77E-04	POLR2M
ch.3.3584022R	1.24	0.15	1.53E-02	2.11E-11	1.01E-08	ABCC5
cg13729466	1.73	0.43	1.54E-02	3.28E-06	4.47E-04	SDCCAG8
cg19239506	0.07	0.41	1.54E-02	9.85E-11	4.11E-08	SIK2
cg06462481	0.08	1.12	1.54E-02	7.93E-19	2.00E-15	TNC
ch.10.91757602F	1.28	0.33	1.56E-02	5.78E-06	7.35E-04	-
cg22336141	0.19	0.65	1.56E-02	6.72E-06	8.40E-04	-
cg19205037	0.11	1.25	1.57E-02	1.30E-16	2.11E-13	-
cg12815829	0.89	0.17	1.57E-02	6.13E-08	1.31E-05	FAM107B

cg22577685	0.93	0.17	1.58E-02	2.66E-08	6.24E-06	-
cg11779273	0.11	1.05	1.60E-02	1.77E-15	2.28E-12	INTU
cg17582160	0.06	0.39	1.60E-02	1.55E-11	7.69E-09	-
cg02648471	1.19	0.16	1.61E-02	1.94E-10	7.61E-08	-
cg08215379	0.35	1.37	1.61E-02	7.79E-07	1.25E-04	KLHDC4
cg00817232	0.20	0.84	1.62E-02	4.12E-07	7.13E-05	-
cg10515751	0.11	0.59	1.65E-02	1.63E-09	5.16E-07	EPB41L3
cg24021101	0.96	0.18	1.66E-02	4.99E-08	1.09E-05	SLC12A6
cg18342183	0.08	0.93	1.69E-02	2.19E-17	4.12E-14	-
cg08870433	0.05	0.41	1.76E-02	4.05E-13	2.97E-10	ZNF57
cg00799171	0.42	0.07	1.80E-02	3.05E-09	9.02E-07	ZNF608
ch.7.135065R	0.12	0.69	1.83E-02	2.48E-10	9.50E-08	TTYH3
cg22324028	0.27	0.96	1.86E-02	6.28E-06	7.91E-04	NBAS
cg16520768	0.70	0.14	1.87E-02	1.18E-07	2.33E-05	LIG1
cg03941040	0.19	1.31	1.88E-02	1.48E-11	7.43E-09	TFAM
cg10184159	1.25	0.10	1.88E-02	5.86E-15	6.62E-12	NBEAL1
cg16177573	0.68	0.12	1.89E-02	2.71E-08	6.35E-06	I-DL
cg13318572	0.62	0.10	1.92E-02	2.20E-09	6.73E-07	MYOM2
cg13113115	0.05	0.18	1.92E-02	1.56E-06	2.31E-04	TFCP2
cg22047295	0.72	0.19	1.94E-02	7.63E-06	9.41E-04	WIPF1
cg25766763	0.04	0.13	1.99E-02	3.66E-06	4.91E-04	DCK
cg09006572	0.14	0.55	1.99E-02	1.07E-06	1.67E-04	-
cg10659652	0.29	1.06	1.99E-02	2.86E-06	3.96E-04	LOC286016

Top 100 significant differentially variable positions in B cells between RA patients and non-RA controls ranked by t-test p-value

CpG	Var RA	Var nRA	P-value TT	P-Value Bt	Adiusted P-value Bt	UCSC RefGene
ch.2.1159565R	1.78	0.24	1.39E-04	1.53E-11	6.17E-09	-
cg01018002	0.82	0.19	2.08E-04	1.96E-07	3.11E-05	SNHG3-RCC1
cg03940643	0.17	0.62	3.90E-04	3.25E-07	4.89E-05	-
cg10720997	0.19	1.02	3.93E-04	5.62E-11	2.03E-08	-
cg22797164	0.55	0.15	5.12E-04	4.95E-06	5.36E-04	NODAL
cg15256944	0.11	0.03	6.16E-04	6.08E-07	8.52E-05	ARL16
cg19819559	0.29	1.00	8.05E-04	1.29E-06	1.65E-04	OXR1
cg08638512	1.21	0.34	8.20E-04	6.98E-06	7.22E-04	C10orf46
cg14268695	0.18	0.60	9.92E-04	1.95E-06	2.38E-04	-
cg14897833	0.11	0.33	9.95E-04	6.77E-06	7.04E-04	NID2
cg02231880	0.13	0.42	1.08E-03	2.29E-06	2.74E-04	LOC100131315
cg25308427	0.19	0.61	1.12E-03	3.86E-06	4.32E-04	-
ch.2.1701371R	2.07	0.49	1.32E-03	3.60E-07	5.36E-05	HK2
cg15154191	0.07	0.22	1.53E-03	7.81E-06	7.96E-04	MEGF6
cg02623991	0.17	0.62	1.65E-03	4.18E-07	6.10E-05	TTYH1

cg20803547	0.23	0.83	1.71E-03	6.10E-07	8.54E-05	IL12RB2
cg02033258	0.23	0.71	1.82E-03	5.83E-06	6.18E-04	PDLIM4
cg04970117	0.08	0.36	1.99E-03	1.45E-08	3.05E-06	SLC6A20
cg19658926	0.32	0.06	2.55E-03	3.62E-09	8.79E-07	CBX3
cg10570405	0.45	0.12	2.76E-03	1.63E-06	2.03E-04	THAP5
cg01579289	0.13	0.45	2.87E-03	1.15E-06	1.50E-04	-
cg06113708	0.79	0.04	2.94E-03	1.87E-19	3.04E-16	COMTD1
cg01915196	0.91	0.04	2.95E-03	8.39E-21	1.68E-17	MBD6
cg14838086	0.61	0.15	3.10E-03	8.79E-07	1.18E-04	BAT5
cg10919329	0.27	0.07	3.18E-03	1.60E-06	2.00E-04	SLC35A1
cg08053598	0.19	0.03	3.35E-03	5.98E-10	1.74E-07	L3MBTL2
cg27369452	0.25	0.81	3.38E-03	2.92E-06	3.38E-04	DNMT3A
cg12518834	1.04	0.05	3.48E-03	2.54E-21	5.31E-18	RALGPS2
cg02664993	0.14	0.44	3.68E-03	8.58E-06	8.62E-04	CTSA
cg11039604	0.96	0.19	3.70E-03	1.15E-08	2.48E-06	-
cg03246739	0.12	0.38	3.84E-03	4.27E-06	4.72E-04	LOC100131315
cg22021178	0.30	0.07	3.87E-03	5.08E-07	7.25E-05	TGFB2
cg15769279	0.26	0.05	3.91E-03	1.16E-08	2.49E-06	RNF214
cg27471464	0.30	0.07	3.91E-03	3.65E-07	5.43E-05	CCDC137
cg22264275	0.12	0.38	3.93E-03	2.65E-06	3.11E-04	MYO7A
cg23170029	0.13	0.51	4.08E-03	7.49E-08	1.33E-05	JPH1
cg12267948	0.15	0.69	4.26E-03	2.86E-09	7.12E-07	PRDM16
cg22413388	0.13	0.44	4.28E-03	1.50E-06	1.89E-04	WNT7B
cg27041026	0.09	0.93	4.28E-03	1.15E-18	1.68E-15	-
cg07157430	0.20	0.87	4.29E-03	5.07E-09	1.19E-06	-
cg01796104	0.42	1.53	4.38E-03	3.83E-07	5.65E-05	CNKSR3
cg19489797	0.39	1.32	4.45E-03	1.45E-06	1.83E-04	DNMT3A
cg16300637	0.15	0.51	4.47E-03	2.84E-06	3.30E-04	-
cg03328984	0.20	0.75	4.52E-03	1.89E-07	3.03E-05	VIPR2
cg06244627	0.72	0.14	4.78E-03	2.18E-08	4.37E-06	MMP2
cg03149173	0.10	0.31	4.94E-03	4.99E-06	5.40E-04	LINC00899
cg20021244	0.25	0.07	5.07E-03	5.74E-06	6.10E-04	APBB3
cg11342941	3.35	0.88	5.19E-03	2.02E-06	2.46E-04	AGPAT4
cg23945725	0.27	0.05	5.27E-03	1.47E-09	3.91E-07	AHI1
cg10323257	0.16	0.67	5.28E-03	1.31E-08	2.79E-06	-
cg18135502	0.54	0.06	5.36E-03	1.02E-13	6.30E-11	LHPP
cg20308895	0.14	0.50	5.39E-03	8.54E-07	1.15E-04	IRX6
cg10667205	0.11	0.34	5.48E-03	4.21E-06	4.67E-04	-
cg08115387	0.21	0.83	5.49E-03	3.89E-08	7.40E-06	-
cg14486812	0.85	0.19	5.54E-03	1.67E-07	2.71E-05	H3F3A
cg01414845	1.14	0.23	5.54E-03	2.51E-08	4.96E-06	NFYC
cg08686311	0.10	0.32	5.60E-03	2.75E-06	3.21E-04	FAM110A
cg07987842	0.89	0.24	5.62E-03	2.31E-06	2.77E-04	ZNF687
cg08593712	0.48	0.10	5.64E-03	2.26E-08	4.53E-06	ZNF839
cg07923233	0.15	0.51	5.76E-03	1.21E-06	1.56E-04	-
cg27086028	0.05	0.17	5.89E-03	8.03E-06	8.15E-04	SHANK1

cg13355248	0.17	0.85	5.93E-03	5.21E-10	1.54E-07	NPTX1
cg20401551	0.21	0.74	5.94E-03	5.31E-07	7.54E-05	SCARF2
cg10711035	0.95	0.27	5.95E-03	6.28E-06	6.60E-04	TAPBP
cg15261499	0.46	0.06	6.04E-03	7.21E-12	3.08E-09	-
cg01391506	0.90	0.20	6.13E-03	1.60E-07	2.62E-05	-
cg03707742	1.13	0.23	6.16E-03	2.31E-08	4.62E-06	TM2D3
cg01567084	1.21	0.21	6.22E-03	1.47E-09	3.93E-07	DYNC2H1
cg25340966	0.17	0.73	6.24E-03	6.87E-09	1.56E-06	TBX15
cg24000099	0.68	0.03	6.33E-03	4.46E-22	1.06E-18	DAGLB
cg19537820	0.95	0.22	6.35E-03	2.30E-07	3.59E-05	REG4
cg14270124	0.36	0.04	6.36E-03	7.93E-13	4.08E-10	PRDX5
cg13686615	0.25	0.90	6.47E-03	3.15E-07	4.77E-05	-
cg09550697	0.05	0.18	6.56E-03	4.50E-07	6.52E-05	PLEC1
cg25082959	0.11	0.37	6.62E-03	9.43E-07	1.26E-04	FGF14
cg19651132	0.14	0.52	6.74E-03	1.76E-07	2.84E-05	KC-1
cg07413609	0.57	0.09	6.77E-03	1.88E-10	6.08E-08	GLI3
cg27391564	0.19	0.63	6.78E-03	1.86E-06	2.29E-04	-
cg16797003	0.18	0.69	7.00E-03	8.71E-08	1.52E-05	-
cg10315076	0.23	1.36	7.01E-03	3.15E-12	1.44E-09	SRD5A3-AS1
cg14500336	0.09	0.59	7.03E-03	1.60E-13	9.50E-11	-
cg03861428	0.35	0.06	7.04E-03	3.83E-09	9.25E-07	SDHB
cg03489016	1.55	0.37	7.16E-03	4.29E-07	6.24E-05	PID1
cg11640253	0.14	0.71	7.32E-03	3.28E-10	1.02E-07	UPF1
cg09567642	0.48	0.07	7.33E-03	7.56E-11	2.66E-08	BDP1
cg07833382	1.44	0.22	7.35E-03	1.73E-10	5.67E-08	SCPEP1
cg04399418	0.39	0.05	7.37E-03	9.83E-13	4.94E-10	-
cg26783464	0.28	0.08	7.37E-03	5.29E-06	5.68E-04	ANKRD35
cg02047489	0.17	0.64	7.63E-03	1.45E-07	2.40E-05	F2R
cg22387890	0.25	1.08	7.63E-03	1.14E-08	2.46E-06	USP42
cg23356977	0.65	0.15	7.65E-03	3.15E-07	4.76E-05	BRD7
cg06427816	0.21	0.97	7.77E-03	3.36E-09	8.24E-07	ABR
cg08826127	0.08	0.37	7.91E-03	1.46E-09	3.89E-07	SLC6A20
cg06652112	0.80	0.14	7.94E-03	2.83E-09	7.05E-07	HES1
cg10489986	0.08	0.37	7.99E-03	1.82E-09	4.76E-07	TRNP1
cg13696940	0.09	0.38	8.05E-03	6.75E-09	1.53E-06	RASL10A
cg10384554	0.21	0.83	8.16E-03	6.36E-08	1.15E-05	IRAK3
ch.6.156307177R	1.41	0.33	8.45E-03	2.83E-07	4.34E-05	-
cg10500173	1.30	0.13	8.47E-03	5.87E-14	3.79E-11	TACC1
cg14085523	0.89	0.17	8.53E-03	9.04E-09	2.00E-06	-

Appendix E – Transcriptional regulator binding site enrichment

Binding sites of transcriptional regulators displaying nominally significant (p < 0.05) at cis-CpGs associated with rheumatoid arthritis, multiple sclerosis, asthma, and osteoarthritis risk loci. TR = Transcriptional regulator name; Prop RA cis-CpG = proportion of all risk-associated cis-CpGs mapping to the binding site; Prop nRA cis-CpG = proportion of all non risk-associated cis-CpGs mapping to the binding site; P-value = p-value for the two-way Fisher's exact test of enrichment. Regulators that exhibited significant enrichment/depletion following Bonferroni correction for the number of proteins tested are highlighted in bold.

All transcriptional regulators exhibiting nominally significant (p < 0.05) enrichment or depletion at risk-associated cis-CpGs in CD4$^+$ T cells.

Rheumatoid arthritis			
TR	**Prop RA cis-CpG**	**Prop nRA cis-CpG**	**P-value**
RELA	**0.16**	**0.06**	**2.15E-04**
RUNX3	0.14	0.07	7.12E-03
NFIC	0.12	0.05	1.21E-02
CHD1	0.09	0.04	1.50E-02
WRNIP1	0.06	0.02	2.40E-02
TEAD4	0.01	0.06	3.12E-02
RBBP5	0.02	0.07	3.40E-02
MEF2C	0.03	0.01	3.66E-02
POU2F2	0.07	0.03	3.74E-02
TBL1XR1	0.07	0.03	4.06E-02
FOS	0.03	0.09	4.70E-02
STAT3	0.01	0.06	4.74E-02
POLR2A	0.39	0.3	4.76E-02

Multiple sclerosis			
TR	**Prop RA cis-CpG**	**Prop nRA cis-CpG**	**P-value**
RUNX3	**0.17**	**0.07**	**4.10E-05**
BATF	**0.08**	**0.02**	**9.86E-05**
RELA	**0.14**	**0.06**	**3.91E-04**
NFATC1	0.06	0.01	1.03E-03
IKZF1	0.05	0.01	2.56E-03
BCL11A	0.06	0.02	2.73E-03
BCL3	0.07	0.03	3.21E-03
MEF2A	0.05	0.02	1.01E-02
SPI1	0.09	0.05	1.14E-02
WRNIP1	0.06	0.02	1.15E-02
TBL1XR1	0.07	0.03	1.24E-02
IRF4	0.05	0.02	1.51E-02
NFYB	0.06	0.02	1.57E-02
MEF2C	0.03	0.01	1.80E-02
POLR2A	0.39	0.30	1.97E-02
ATF2	0.09	0.04	2.03E-02
NFIC	0.10	0.05	2.22E-02
CTCF	0.10	0.17	2.39E-02
ELK1	0.04	0.02	2.45E-02
SMC3	0.11	0.06	2.77E-02
POU2F2	0.07	0.03	4.13E-02
TCF3	0.04	0.02	4.77E-02
MAZ	0.14	0.09	4.96E-02

Asthma			
TR	**Prop RA cis-CpG**	**Prop nRA cis-CpG**	**P-value**
NFIC	0.12	0.05	2.02E-03
FOXA1	0.11	0.05	2.10E-03
E2F1	0.01	0.06	3.70E-03
IKZF1	0.04	0.01	1.75E-02
E2F6	0.01	0.05	1.84E-02
FOS	0.14	0.08	2.83E-02
HDAC2	0.01	0.04	4.78E-02

Osteoarthritis			
TR	**Prop RA cis-CpG**	**Prop nRA cis-CpG**	**P-value**
WRNIP1	**0.08**	**0.02**	**1.47E-04**
JUND	**0.01**	**0.09**	**1.51E-04**
JUN	0.01	0.05	1.10E-02
TBP	0.03	0.09	1.34E-02
E2F6	0.1	0.05	1.63E-02
SIN3AK20	0.03	0.08	1.89E-02
MEF2C	0.03	0.01	2.26E-02
FOS	0.03	0.09	2.46E-02
GATA2	0.02	0.06	2.65E-02
SIN3A	0.02	0.07	2.70E-02
TBL1XR1	0.07	0.03	2.92E-02
EP300	0.06	0.12	2.95E-02
CHD1	0.01	0.04	4.76E-02
PAX5	0.01	0.04	4.81E-02

All transcriptional regulators exhibiting nominally significant (p < 0.05) enrichment or depletion at risk-association cis-CpGs in B cells.

Rheumatoid arthritis			
TR	**Prop RA cis-CpG**	**Prop nRA cis-CpG**	**P-value**
RELA	**0.16**	**0.06**	**1.22E-04**
CHD1	0.1	0.04	2.66E-03
WRNIP1	0.07	0.02	5.48E-03
RAD21	0.03	0.1	1.22E-02
SPI1	0.1	0.05	1.46E-02
REST	0.02	0.07	3.39E-02
BACH1	0.05	0.02	3.60E-02

Multiple sclerosis			
TR	**Prop RA cis-CpG**	**Prop nRA cis-CpG**	**P-value**
TBL1XR1	**0.12**	**0.03**	**5.84E-05**
CCNT2	**0.12**	**0.03**	**1.12E-04**
ATF2	0.12	0.04	1.10E-03
MAZ	0.17	0.09	3.91E-03
RAD21	0.03	0.10	4.50E-03
BHLHE40	0.13	0.06	5.25E-03
BCL11A	0.05	0.02	7.30E-03
SP4	0.06	0.02	7.86E-03
CTCF	0.08	0.17	8.08E-03
WRNIP1	0.06	0.02	1.04E-02
ELK1	0.05	0.02	1.21E-02
IRF4	0.05	0.02	1.28E-02
E2F4	0.07	0.03	2.32E-02
RCOR1	0.12	0.06	3.06E-02
RUNX3	0.12	0.06	3.36E-02
IKZF1	0.04	0.01	4.40E-02
MXI1	0.13	0.07	4.66E-02
FOXA1	0.01	0.05	4.76E-02
NFE2	0.02	0.00	4.79E-02

Asthma			
TR	**Prop RA cis-CpG**	**Prop nRA cis-CpG**	**P-value**
NFIC	0.12	0.05	4.48E-03
USF2	0.08	0.03	4.52E-03
EBF1	0.1	0.05	8.77E-03
ESR1	0.06	0.02	1.39E-02
E2F1	0.01	0.05	1.61E-02
PHF8	0.01	0.05	3.28E-02
NFATC1	0.04	0.01	3.42E-02
FOS	0.14	0.09	3.55E-02
PAX5	0.08	0.04	3.57E-02
BCL11A	0.04	0.02	4.35E-02
RUNX3	0.11	0.06	4.36E-02

Osteoarthritis			
TR	**Prop RA cis-CpG**	**Prop nRA cis-CpG**	**P-value**
TCF12	**0.17**	**0.06**	**1.21E-06**
REST	**0.18**	**0.07**	**2.62E-05**
MEF2A	0.06	0.02	2.94E-03
MEF2C	0.03	0.01	3.52E-03
WRNIP1	0.06	0.02	3.97E-03
TBL1XR1	0.07	0.03	1.55E-02
NFIC	0.1	0.05	1.56E-02
BATF	0.05	0.02	1.69E-02
NR2F2	0.05	0.02	1.93E-02
PML	0.09	0.05	2.78E-02
FOXM1	0.07	0.03	2.85E-02
FOS	0.04	0.09	3.28E-02
ELK1	0.04	0.02	3.31E-02
BRCA1	0.03	0.01	3.52E-02
ZBTB7A	0.07	0.04	3.69E-02
TCF3	0.04	0.02	4.37E-02
TFAP2A	0.06	0.03	4.40E-02
MTA3	0.05	0.02	4.65E-02

Appendix F – Gene Ontology Biological Process pathway analysis

Gene Ontology (GO) Biological Pathway enrichment at risk-associated cis-CpGs in CD4$^+$ T cells and B cells. Number of Genes in process = The number of genes annotated to a particular biological pathway in GO; Number of genes at cis-CpG = number of genes mapping to the risk-associated cis-CpG; P-value = Enrichment p-value calculated using a modified hypergeometric test to account for bias in the number of CpG probes mapping to each gene (see Chapter 2.6.4).

The top 50 Gene Ontology Biological Processes enriched at risk-associated cis-CpGs in CD4$^+$ T cells relative to non-risk-associated cis-CpGs

Rheumatoid arthritis (CD4$^+$ T cell)			
Biological Process Term	**Number of Genes in Process**	**Number of Genes at cis-CpG**	**P-value**
alpha-linolenic acid metabolic process	9	2	1.23E-04
linoleic acid metabolic process	11	2	2.25E-04
regulation of B cell receptor signaling pathway	14	2	4.29E-04
unsaturated fatty acid biosynthetic process	28	2	1.10E-03
positive regulation of protein serine/threonine phosphatase activity	1	1	1.41E-03
CD8-positive, alpha-beta T cell differentiation involved in immune response	1	1	1.48E-03
negative regulation of protein import into nucleus, translocation	1	1	1.61E-03
B cell receptor signaling pathway	33	2	1.93E-03
regulation of antigen receptor-mediated signaling pathway	35	2	2.00E-03
negative regulation of sphingolipid biosynthetic process	1	1	2.40E-03
cellular sphingolipid homeostasis	1	1	2.40E-03
negative regulation of ceramide biosynthetic process	1	1	2.40E-03
cellular response to 2,3,7,8-tetrachlorodibenzodioxine	2	1	2.83E-03
synaptic vesicle membrane organization	2	1	3.30E-03
valine catabolic process	2	1	3.33E-03
mesodermal to mesenchymal transition involved in gastrulation	2	1	3.56E-03
immune system process	1589	8	3.70E-03
regulation of toll-like receptor 9 signaling pathway	2	1	3.95E-03
unsaturated fatty acid metabolic process	58	2	4.38E-03
negative regulation of immunoglobulin production	3	1	4.39E-03
immune effector process	652	5	4.40E-03
positive regulation of histone H3-K27 methylation	3	1	4.50E-03
cerebral cortex regionalization	3	1	4.56E-03
endodermal cell fate specification	3	1	5.01E-03
leukocyte activation	689	5	5.50E-03
negative regulation of B cell receptor signaling pathway	3	1	5.54E-03
long-chain fatty acid metabolic process	64	2	5.57E-03
response to cGMP	4	1	5.82E-03
cellular response to cGMP	4	1	5.82E-03
regulation of histone H3-K27 methylation	4	1	6.07E-03
response to 2,3,7,8-tetrachlorodibenzodioxine	4	1	6.89E-03
negative regulation of cGMP-mediated signaling	4	1	7.39E-03
response to peptidoglycan	5	1	7.89E-03
immune response	1077	6	8.24E-03
valine metabolic process	5	1	8.69E-03
osteoclast fusion	4	1	9.21E-03

fatty acid biosynthetic process	87	2	9.44E-03
positive regulation of vascular permeability	6	1	9.72E-03

Multiple Sclerosis (CD4^{+} T cell)

Biological Process Term	Number of Genes in Process	Number of Genes at cis-CpG	P-value
negative regulation of calcidiol 1-monooxygenase activity	2	2	2.34E-05
negative regulation of lipid metabolic process	49	4	3.58E-05
positive regulation of T cell activation	105	5	5.59E-05
positive regulation of leukocyte cell-cell adhesion	108	5	6.62E-05
regulation of calcidiol 1-monooxygenase activity	4	2	9.90E-05
negative regulation of vitamin D biosynthetic process	4	2	1.31E-04
negative regulation of lipid biosynthetic process	30	3	1.66E-04
positive regulation of cell-cell adhesion	135	5	1.91E-04
regulation of lymphocyte activation	221	6	1.99E-04
positive regulation of lymphocyte activation	140	5	2.10E-04
negative regulation of vitamin metabolic process	5	2	2.24E-04
negative regulation of hormone biosynthetic process	6	2	2.27E-04
positive regulation of immune system process	538	9	2.55E-04
positive regulation of leukocyte differentiation	78	4	2.55E-04
regulation of vitamin D biosynthetic process	6	2	2.65E-04
innate immune response-activating signal transduction	142	5	2.90E-04
leukocyte activation	684	10	2.91E-04
negative regulation of cellular process	2712	22	2.95E-04
fat-soluble vitamin biosynthetic process	7	2	3.08E-04
vitamin D biosynthetic process	7	2	3.08E-04
regulation of leukocyte cell-cell adhesion	152	5	3.27E-04
regulation of DNA-binding transcription factor activity	243	6	3.29E-04
activation of innate immune response	148	5	3.42E-04
negative regulation of hormone metabolic process	7	2	3.44E-04
negative regulation of alcohol biosynthetic process	8	2	3.47E-04
positive regulation of cell adhesion	240	6	4.13E-04
positive regulation of leukocyte activation	162	5	4.15E-04
regulation of T cell activation	166	5	4.77E-04
positive regulation of cell activation	171	5	5.22E-04
positive regulation of hemopoiesis	96	4	5.28E-04
regulation of multicellular organismal process	1857	17	5.56E-04
regulation of vitamin metabolic process	8	2	5.57E-04
negative regulation of monooxygenase activity	9	2	5.91E-04
regulation of leukocyte activation	269	6	5.91E-04
stimulatory C-type lectin receptor signaling pathway	42	3	6.08E-04
regulation of cytokine production	375	7	6.34E-04
negative regulation of biological process	3077	23	6.57E-04
leukocyte cell-cell adhesion	176	5	6.64E-04
positive regulation of innate immune response	176	5	6.91E-04
Fc-epsilon receptor signaling pathway	42	3	6.96E-04
innate immune response activating cell surface receptor signaling pathway	45	3	7.34E-04
positive regulation of T cell differentiation	45	3	7.61E-04
platelet-derived growth factor receptor signaling pathway	42	3	7.82E-04
transcription, DNA-templated	1942	17	8.15E-04
immune system process	1584	15	8.43E-04
regulation of cell differentiation	1074	12	8.44E-04
cell surface receptor signaling pathway	1736	16	8.52E-04
cell activation	783	10	8.77E-04
regulation of immune system process	783	10	8.85E-04
regulation of macromolecule biosynthetic process	2153	18	8.86E-04

Asthma (CD4$^+$ T cell)			
Biological Process Term	**Number of Genes in Process**	**Number of Genes at cis-CpG**	**P-value**
negative regulation of nitrogen compound metabolic process	1307	17	6.33E-07
negative regulation of nucleobase-containing compound metabolic process	769	13	1.11E-06
negative regulation of cellular metabolic process	1410	17	1.93E-06
negative regulation of metabolic process	1670	18	3.24E-06
negative regulation of CD4-positive, alpha-beta T cell differentiation	10	3	3.63E-06
CD4-positive, alpha-beta T cell differentiation	40	4	8.26E-06
negative regulation of alpha-beta T cell differentiation	13	3	9.86E-06
negative regulation of CD4-positive, alpha-beta T cell activation	14	3	1.07E-05
negative regulation of RNA biosynthetic process	670	11	1.22E-05
negative regulation of nucleic acid-templated transcription	670	11	1.22E-05
CD4-positive, alpha-beta T cell activation	47	4	1.40E-05
negative regulation of cellular biosynthetic process	841	12	1.60E-05
negative regulation of biosynthetic process	856	12	1.89E-05
negative regulation of macromolecule metabolic process	1521	16	1.92E-05
negative regulation of RNA metabolic process	706	11	1.95E-05
positive regulation of CD8-positive, alpha-beta T cell differentiation	2	2	3.07E-05
negative regulation of T cell differentiation	19	3	3.17E-05
alpha-beta T cell differentiation	57	4	3.19E-05
negative regulation of alpha-beta T cell activation	21	3	4.44E-05
positive regulation of CD8-positive, alpha-beta T cell activation	3	2	5.55E-05
negative regulation of transcription, DNA-templated	648	10	5.57E-05
regulation of CD8-positive, alpha-beta T cell differentiation	3	2	5.79E-05
negative regulation of biological process	3077	23	5.90E-05
negative regulation of macromolecule biosynthetic process	808	11	6.00E-05
T cell differentiation	132	5	6.56E-05
regulation of CD4-positive, alpha-beta T cell differentiation	25	3	6.57E-05
negative regulation of lymphocyte differentiation	26	3	6.64E-05
positive regulation of alpha-beta T cell differentiation	25	3	7.44E-05
alpha-beta T cell activation	76	4	1.04E-04
T-helper 1 cell differentiation	8	2	1.16E-04
regulation of CD4-positive, alpha-beta T cell activation	32	3	1.24E-04
positive regulation of alpha-beta T cell activation	32	3	1.46E-04
regulation of alpha-beta T cell differentiation	33	3	1.49E-04
regulation of macromolecule metabolic process	3401	23	1.54E-04
T cell activation	251	6	1.65E-04
T cell differentiation involved in immune response	37	3	1.84E-04
negative regulation of cellular macromolecule biosynthetic process	770	10	2.03E-04
peripheral nervous system neuron differentiation	6	2	2.09E-04
peripheral nervous system neuron development	6	2	2.09E-04
regulation of metabolic process	3705	24	2.13E-04
regulation of nitrogen compound metabolic process	3209	22	2.27E-04
interleukin-2 production	39	3	2.61E-04
regulation of macromolecule biosynthetic process	2152	17	2.87E-04
regulation of immune response	500	8	2.91E-04
transcription, DNA-templated	1943	16	3.06E-04
positive regulation of protein localization to cell surface	8	2	3.13E-04
regulation of nucleobase-containing compound metabolic process	2160	17	3.27E-04
nucleic acid-templated transcription	1966	16	3.59E-04
lymphocyte differentiation	189	5	3.64E-04
CD8-positive, alpha-beta T cell differentiation	8	2	3.67E-04

Osteoarthritis (CD4$^+$ T cell)			
Biological Process Term	**Number of Genes in Process**	**Number of Genes at cis-CpG**	**P-value**
regulation of endopeptidase activity	247	6	2.61E-06
activation of cysteine-type endopeptidase activity involved in apoptotic process	64	4	2.91E-06
regulation of peptidase activity	260	6	3.60E-06
positive regulation of cysteine-type endopeptidase activity involved in apoptotic process	88	4	1.08E-05
positive regulation of cysteine-type endopeptidase activity	100	4	2.06E-05
positive regulation of endopeptidase activity	115	4	3.57E-05
positive regulation of peptidase activity	126	4	5.28E-05
regulation of proteolysis	437	6	6.09E-05
regulation of cysteine-type endopeptidase activity involved in apoptotic process	131	4	6.15E-05
regulation of cysteine-type endopeptidase activity	146	4	1.01E-04
regulation of hydrolase activity	784	7	2.55E-04
positive regulation of apoptotic process	390	5	3.83E-04
positive regulation of programmed cell death	393	5	3.94E-04
positive regulation of proteolysis	224	4	4.56E-04
negative regulation of protein metabolic process	646	6	5.18E-04
positive regulation of cell death	422	5	5.48E-04
digestive system development	93	3	6.94E-04
positive regulation of apoptotic signaling pathway	104	3	7.47E-04
hyaluronan metabolic process	23	2	7.95E-04
positive regulation of hydrolase activity	478	5	1.21E-03
regulation of catalytic activity	1371	8	1.41E-03
positive regulation of supramolecular fiber organization	140	3	1.50E-03
regulation of clathrin coat assembly	1	1	1.61E-03
positive regulation of clathrin coat assembly	1	1	1.61E-03
establishment of protein localization to plasma membrane	33	2	1.62E-03
positive regulation of cytoskeleton organization	150	3	1.77E-03
negative regulation of lung blood pressure	1	1	2.00E-03
negative regulation of Arp2/3 complex-mediated actin nucleation	2	1	2.15E-03
regulation of cellular protein metabolic process	1509	8	2.19E-03
regulation of molecular function	1868	9	2.30E-03
regulation of diacylglycerol kinase activity	1	1	2.31E-03
positive regulation of diacylglycerol kinase activity	1	1	2.31E-03
transforming growth factor beta3 production	2	1	2.40E-03
regulation of transforming growth factor beta3 production	2	1	2.40E-03
positive regulation of transforming growth factor beta3 production	2	1	2.40E-03
retinoic acid receptor signaling pathway involved in somitogenesis	1	1	2.51E-03
somitogenesis	42	2	2.54E-03
positive regulation of catalytic activity	856	6	2.67E-03
negative regulation of cellular protein metabolic process	609	5	2.88E-03
protein transport from ciliary membrane to plasma membrane	2	1	2.92E-03
somite development	49	2	3.30E-03
regulation of protein metabolic process	1614	8	3.40E-03
positive regulation of platelet-derived growth factor receptor-beta signaling pathway	2	1	3.41E-03
regulation of chromatin assembly	2	1	3.47E-03
regulation of lung blood pressure	3	1	3.81E-03
activation of cysteine-type endopeptidase activity involved in apoptotic signaling pathway	3	1	4.21E-03
positive regulation of reactive oxygen species metabolic process	58	2	4.22E-03
negative regulation of molecular function	653	5	4.42E-03

ureter maturation	2	1	4.53E-03
negative regulation of actin nucleation	4	1	4.55E-03

The top 50 Gene Ontology Biological Processes enriched at risk-associated cis-CpGs in B cells relative to non-risk-associated cis-CpGs

Rheumatoid arthritis (B cell)			
Biological Process Term	**Number of Genes in Process**	**Number of Genes at cis-CpG**	**P-value**
regulation of B cell receptor signaling pathway	18	2	4.61E-04
immune effector process	676	6	6.73E-04
DN3 thymocyte differentiation	1	1	1.27E-03
positive regulation of protein serine/threonine phosphatase activity	1	1	1.27E-03
leukocyte differentiation	297	4	1.50E-03
immune response	1120	7	1.67E-03
cellular sphingolipid homeostasis	1	1	1.91E-03
regulation of antigen receptor-mediated signaling pathway	39	2	1.95E-03
T cell differentiation in thymus	39	2	2.06E-03
B cell receptor signaling pathway	37	2	2.19E-03
negative regulation of protein import into nucleus, translocation	1	1	2.45E-03
cellular response to 2,3,7,8-tetrachlorodibenzodioxine	1	1	2.45E-03
regulation of dendritic cell chemotaxis	2	1	2.47E-03
positive regulation of dendritic cell chemotaxis	2	1	2.47E-03
positive regulation of flagellated sperm motility involved in capacitation	2	1	2.57E-03
DN2 thymocyte differentiation	2	1	2.68E-03
immunoglobulin production	54	2	2.80E-03
negative regulation of sphingolipid biosynthetic process	2	1	3.25E-03
negative regulation of ceramide biosynthetic process	2	1	3.25E-03
synaptic vesicle membrane organization	2	1	3.82E-03
thymocyte migration	3	1	4.06E-03
lymphocyte differentiation	194	3	4.07E-03
immune system process	1677	8	4.35E-03
leukocyte activation involved in immune response	410	4	4.74E-03
cell activation involved in immune response	413	4	4.86E-03
leukocyte activation	708	5	5.82E-03
isotype switching to IgA isotypes	5	1	6.06E-03
leukocyte mediated immunity	444	4	6.09E-03
positive regulation of interferon-gamma secretion	4	1	6.30E-03
response to 2,3,7,8-tetrachlorodibenzodioxine	3	1	6.52E-03
negative regulation of B cell receptor signaling pathway	4	1	6.62E-03
response to cGMP	4	1	6.95E-03
cellular response to cGMP	4	1	6.95E-03
negative regulation of immunoglobulin production	5	1	7.19E-03
regulation of interferon-gamma secretion	5	1	7.34E-03
regulation of toll-like receptor 9 signaling pathway	5	1	7.55E-03
positive regulation of defense response	253	3	7.71E-03
regulation of fat cell differentiation	78	2	8.45E-03
positive regulation of brown fat cell differentiation	5	1	9.08E-03
positive regulation of flagellated sperm motility	6	1	9.36E-03
hemopoiesis	496	4	9.47E-03
negative regulation of signal transduction	773	5	9.69E-03
response to peptidoglycan	7	1	9.76E-03
negative regulation of cGMP-mediated signaling	5	1	9.84E-03
osteoclast fusion	4	1	9.91E-03

Multiple Sclerosis (B cell)			
Biological Process Term	Number of Genes in Process	Number of Genes at cis-CpG	P-value
cellular response to organic substance	1533	15	2.58E-05
regulation of macromolecule metabolic process	3501	23	2.98E-05
regulation of nitrogen compound metabolic process	3289	22	4.21E-05
regulation of primary metabolic process	3395	22	7.12E-05
regulation of cellular metabolic process	3465	22	1.02E-04
regulation of metabolic process	3804	23	1.28E-04
nephric duct morphogenesis	6	2	1.43E-04
enzyme linked receptor protein signaling pathway	648	9	1.50E-04
cell surface receptor signaling pathway	1766	15	1.63E-04
negative regulation of cellular process	2801	19	2.30E-04
response to organic substance	1864	15	2.34E-04
response to stimulus	5114	27	2.36E-04
cellular response to chemical stimulus	1856	15	2.40E-04
nephric duct development	8	2	2.80E-04
cellular response to stimulus	4226	24	2.86E-04
regulation of mitochondrion organization	106	4	3.17E-04
regulation of sphingolipid biosynthetic process	9	2	3.24E-04
regulation of membrane lipid metabolic process	9	2	3.24E-04
regulation of ceramide biosynthetic process	9	2	3.24E-04
platelet-derived growth factor receptor signaling pathway	44	3	3.77E-04
negative regulation of biological process	3177	20	3.84E-04
signal transduction	3436	21	3.98E-04
regulation of cell differentiation	1106	11	4.15E-04
regulation of signal transduction	1934	15	4.22E-04
mitophagy	10	2	4.24E-04
regulation of mitophagy	10	2	4.24E-04
signaling	3724	22	4.27E-04
asymmetric cell division	10	2	4.49E-04
macromolecule metabolic process	5378	27	4.61E-04
cell communication	3755	22	4.80E-04
regulation of gene expression	2543	17	5.73E-04
positive regulation of metabolic process	2083	15	8.20E-04
positive regulation of developmental process	851	9	1.03E-03
negative regulation of cellular metabolic process	1467	12	1.03E-03
positive regulation of cellular process	3133	19	1.06E-03
negative regulation of lymphocyte mediated immunity	17	2	1.14E-03
positive regulation of mitochondrion organization	67	3	1.16E-03
cell death	1270	11	1.20E-03
response to retinoic acid	68	3	1.21E-03
regulation of multicellular organismal process	1905	14	1.23E-03
STAT cascade	77	3	1.48E-03
negative regulation of adaptive immune response based on somatic recombination of immune receptors built from immunoglobulin superfamily domains	20	2	1.49E-03
nephron epithelium development	71	3	1.49E-03
regulation of cell communication	2177	15	1.51E-03
positive regulation of biological process	3517	20	1.65E-03
regulation of signaling	2198	15	1.69E-03
positive regulation of cell communication	1109	10	1.73E-03
regulation of response to stimulus	2461	16	1.75E-03
negative regulation of adaptive immune response	22	2	1.75E-03
gene expression	3077	18	1.76E-03

Asthma (B cell)

Biological Process Term	Number of Genes in Process	Number of Genes at cis-CpG	P-value
cytokine production	430	10	6.34E-07
regulation of cytokine production	388	9	2.37E-06
negative regulation of nucleobase-containing compound metabolic process	816	12	4.44E-06
regulation of response to stress	820	12	5.11E-06
negative regulation of nitrogen compound metabolic process	1364	15	7.60E-06
regulation of type 2 immune response	17	3	1.63E-05
type 2 immune response	18	3	1.86E-05
negative regulation of cellular metabolic process	1466	15	1.91E-05
regulation of defense response	412	8	2.82E-05
regulation of inflammatory response	210	6	3.50E-05
response to stress	2173	18	3.66E-05
regulation of CD4-positive, alpha-beta T cell differentiation	23	3	3.76E-05
regulation of immune system process	849	11	4.22E-05
negative regulation of RNA biosynthetic process	705	10	4.81E-05
negative regulation of nucleic acid-templated transcription	705	10	4.81E-05
negative regulation of type 2 immune response	4	2	5.23E-05
negative regulation of cellular biosynthetic process	879	11	5.28E-05
positive regulation of response to stimulus	1359	14	5.45E-05
negative regulation of biological process	3180	22	6.05E-05
negative regulation of biosynthetic process	894	11	6.08E-05
interleukin-5 secretion	4	2	6.66E-05
regulation of interleukin-5 secretion	4	2	6.66E-05
positive regulation of interleukin-5 secretion	4	2	6.66E-05
positive regulation of immune system process	587	9	6.84E-05
negative regulation of RNA metabolic process	746	10	7.44E-05
isotype switching to IgG isotypes	6	2	7.60E-05
regulation of isotype switching to IgG isotypes	6	2	7.60E-05
regulation of response to stimulus	2462	19	8.65E-05
regulation of CD4-positive, alpha-beta T cell activation	30	3	8.83E-05
T-helper 2 cell cytokine production	6	2	9.79E-05
regulation of T-helper 2 cell cytokine production	6	2	9.79E-05
regulation of response to external stimulus	480	8	1.18E-04
negative regulation of cellular process	2804	20	1.26E-04
negative regulation of metabolic process	1741	15	1.26E-04
multi-organism process	1340	13	1.40E-04
T cell activation	264	6	1.68E-04
CD4-positive, alpha-beta T cell differentiation	37	3	1.72E-04
interleukin-2 production	35	3	1.75E-04
regulation of alpha-beta T cell differentiation	34	3	1.79E-04
negative regulation of macromolecule metabolic process	1598	14	1.85E-04
regulation of leukocyte cell-cell adhesion	174	5	1.89E-04
regulation of macromolecule metabolic process	3502	22	1.90E-04
negative regulation of macromolecule biosynthetic process	842	10	1.94E-04
regulation of immune response	544	8	2.13E-04
negative regulation of transcription, DNA-templated	684	9	2.20E-04
regulation of lymphocyte differentiation	95	4	2.46E-04
regulation of T cell activation	184	5	2.58E-04
regulation of DNA recombination	45	3	2.64E-04
regulation of leukocyte activation	289	6	2.66E-04
regulation of nitrogen compound metabolic process	3290	21	2.68E-04

Osteoarthritis (B cell)			
Biological Process Term	**Number of Genes in Process**	**Number of Genes at cis-CpG**	**P-value**
positive regulation of apoptotic process	390	7	1.19E-05
positive regulation of programmed cell death	392	7	1.21E-05
positive regulation of cell death	424	7	1.86E-05
regulation of Rac protein signal transduction	7	2	1.44E-04
negative regulation of transforming growth factor beta receptor signaling pathway	46	3	1.55E-04
negative regulation of cellular response to transforming growth factor beta stimulus	47	3	1.60E-04
activation of cysteine-type endopeptidase activity involved in apoptotic process	62	3	2.61E-04
positive regulation of reactive oxygen species metabolic process	58	3	2.71E-04
regulation of apoptotic process	889	8	2.75E-04
artery development	65	3	2.87E-04
regulation of programmed cell death	899	8	2.99E-04
regulation of clathrin-dependent endocytosis	14	2	3.69E-04
regulation of transforming growth factor beta receptor signaling pathway	68	3	4.31E-04
regulation of cellular response to transforming growth factor beta stimulus	69	3	4.41E-04
positive regulation of hydrolase activity	482	6	4.67E-04
negative regulation of cellular protein localization	73	3	5.06E-04
regulation of cell death	973	8	5.15E-04
negative regulation of transmembrane receptor protein serine/threonine kinase signaling pathway	73	3	5.55E-04
liver development	81	3	5.69E-04
receptor-mediated endocytosis	179	4	5.74E-04
hepaticobiliary system development	82	3	5.80E-04
astrocyte development	19	2	5.95E-04
endocrine system development	83	3	6.72E-04
positive regulation of cysteine-type endopeptidase activity involved in apoptotic process	91	3	7.73E-04
heart development	363	5	8.50E-04
adrenal gland development	19	2	9.36E-04
regulation of transforming growth factor beta production	24	2	9.48E-04
regulation of hydrolase activity	794	7	1.06E-03
negative regulation of cellular response to growth factor stimulus	92	3	1.11E-03
positive regulation of cysteine-type endopeptidase activity	103	3	1.15E-03
forebrain astrocyte differentiation	1	1	1.16E-03
forebrain astrocyte development	1	1	1.16E-03
gamma-aminobutyric acid secretion, neurotransmission	1	1	1.16E-03
negative regulation of establishment of protein localization	105	3	1.18E-03
transforming growth factor beta production	26	2	1.23E-03
apoptotic process	1116	8	1.26E-03
digestive system development	98	3	1.26E-03
anatomical structure maturation	105	3	1.29E-03
positive regulation of nitric oxide biosynthetic process	25	2	1.41E-03
positive regulation of nitric oxide metabolic process	25	2	1.41E-03
negative regulation of protein metabolic process	658	6	1.41E-03
regulation of endopeptidase activity	254	4	1.42E-03
Rac protein signal transduction	22	2	1.44E-03
positive regulation of apoptotic signaling pathway	109	3	1.48E-03
regulation of reactive oxygen species metabolic process	111	3	1.52E-03
positive regulation of endopeptidase activity	118	3	1.64E-03
clathrin-dependent endocytosis	30	2	1.75E-03
regulation of peptidase activity	270	4	1.81E-03

positive regulation of catalytic activity	885	7	1.82E-03
programmed cell death	1193	8	1.82E-03

Appendix G – Expression Quantitative Trait Methylations as Risk cis-CpGs

Expression Quantitative Methylations at cis-CpGs associated with risk loci (rheumatoid arthritis, multiple sclerosis, asthma, and osteoarthritis.) were identified in CD4$^+$ T cells and B cells by Spearman's rho correlation. P-value = Spearman's rho p-value; Adj. Pval = Benjamini-Hochberg adjusted p-value across the number of transcripts tested at each cis-CpG. These tables report associations for all CpG and Transcript probes whereas numbers presented in the text (Chapter 5) refer only to unique CpG-Gene associations.

Cis-expression quantitative trait methylation (eQTM) at genes within ±500kb of risk-associated cis-meQTL CpGs in CD4+ T cells

Rheumatoid arthritis (CD4$^+$ T cells)					
CpG	**IlluminaID**	**Gene**	**P-value**	**Adj. Pval**	**Rho**
cg21124310	ILMN_1798947	ANKRD55	4.26E-12	2.13E-11	-0.623
cg21124310	ILMN_2341724	ANKRD55	1.49E-11	3.73E-11	-0.611
cg23343972	ILMN_1798947	ANKRD55	7.54E-11	1.88E-10	-0.594
cg23343972	ILMN_2341724	ANKRD55	4.41E-11	1.88E-10	-0.599
cg10404427	ILMN_1798947	ANKRD55	1.16E-10	5.81E-10	-0.589
cg15667493	ILMN_1798947	ANKRD55	5.90E-10	1.47E-09	-0.570
cg15667493	ILMN_2341724	ANKRD55	3.01E-10	1.47E-09	-0.578
cg10404427	ILMN_2341724	ANKRD55	6.90E-09	1.73E-08	-0.540
cg07522171	ILMN_1682727	JAZF1	4.47E-09	1.79E-08	-0.545
cg01045635	ILMN_1691693	FCRL3	1.91E-08	3.82E-08	-0.526
cg17134153	ILMN_1691693	FCRL3	1.98E-08	3.95E-08	-0.526
cg11187739	ILMN_1682727	JAZF1	4.96E-08	2.48E-07	-0.513
cg15431103	ILMN_1798947	ANKRD55	6.23E-08	3.12E-07	-0.509
cg15431103	ILMN_2341724	ANKRD55	1.83E-07	4.58E-07	-0.493
cg21721331	ILMN_1691693	FCRL3	5.60E-07	1.12E-06	-0.476
cg01045635	ILMN_1797428	FCRL3	2.54E-06	2.54E-06	-0.451
cg17134153	ILMN_1797428	FCRL3	4.22E-06	4.22E-06	-0.442
cg19602479	ILMN_1691693	FCRL3	4.95E-06	9.89E-06	-0.439
cg10909506	ILMN_1662174	ORMDL3	6.69E-07	1.07E-05	-0.473
cg08786003	ILMN_1691693	FCRL3	8.88E-06	1.78E-05	-0.428
cg19602479	ILMN_1797428	FCRL3	1.85E-05	1.85E-05	-0.414
cg21721331	ILMN_1797428	FCRL3	2.05E-05	2.05E-05	-0.412
cg18711369	ILMN_1662174	ORMDL3	3.16E-06	5.06E-05	-0.447
cg23343972	ILMN_1849013	IL6ST	3.48E-05	5.80E-05	-0.401
cg21124310	ILMN_1849013	IL6ST	3.60E-05	6.01E-05	-0.401
cg10404427	ILMN_1849013	IL6ST	6.25E-05	1.04E-04	-0.389
cg00184826	ILMN_1682727	JAZF1	5.82E-05	1.16E-04	0.391
cg08786003	ILMN_1797428	FCRL3	1.97E-04	1.97E-04	-0.364
cg18711369	ILMN_1666206	GSDMB	2.94E-05	2.35E-04	-0.405
cg15667493	ILMN_1849013	IL6ST	3.39E-04	5.64E-04	-0.351
cg15431103	ILMN_1849013	IL6ST	3.53E-04	5.88E-04	-0.350
cg25259754	ILMN_1691693	FCRL3	3.53E-04	7.06E-04	-0.350
cg16130019	ILMN_1682727	JAZF1	1.48E-04	7.39E-04	0.371
cg10909506	ILMN_1666206	GSDMB	1.02E-04	8.16E-04	-0.379
cg25259754	ILMN_1797428	FCRL3	1.33E-03	1.33E-03	-0.317
cg16213375	ILMN_1786759	C11orf10	1.39E-04	1.39E-03	-0.372
cg18707136	ILMN_1797428	FCRL3	7.57E-04	1.51E-03	0.331

cg11187739	ILMN_2374770	TAX1BP1	6.14E-04	1.53E-03	0.337
cg15431103	ILMN_1797861	IL6ST	1.57E-03	1.96E-03	-0.312
cg08519799	ILMN_1682727	JAZF1	9.94E-04	3.98E-03	0.324
cg18707136	ILMN_1691693	FCRL3	5.43E-03	5.43E-03	0.276
cg07522171	ILMN_2374770	TAX1BP1	4.36E-03	8.72E-03	0.283

Multiple sclerosis (CD4$^+$ T cells)

CpG	IlluminaID	Gene	P-value	Adj. Pval	Rho
cg21124310	ILMN_1798947	ANKRD55	4.26E-12	2.13E-11	-0.623
cg21124310	ILMN_2341724	ANKRD55	1.49E-11	3.73E-11	-0.611
cg23343972	ILMN_1798947	ANKRD55	7.54E-11	1.88E-10	-0.594
cg23343972	ILMN_2341724	ANKRD55	4.41E-11	1.88E-10	-0.599
cg10404427	ILMN_1798947	ANKRD55	1.16E-10	5.81E-10	-0.589
cg15667493	ILMN_1798947	ANKRD55	5.90E-10	1.47E-09	-0.570
cg15667493	ILMN_2341724	ANKRD55	3.01E-10	1.47E-09	-0.578
cg10404427	ILMN_2341724	ANKRD55	6.90E-09	1.73E-08	-0.540
cg07522171	ILMN_1682727	JAZF1	4.47E-09	1.79E-08	-0.545
cg05575058	ILMN_1789558	FAM164A	1.53E-08	7.67E-08	-0.529
cg03983883	ILMN_1789558	FAM164A	1.80E-08	9.02E-08	-0.527
cg11187739	ILMN_1682727	JAZF1	4.96E-08	2.48E-07	-0.513
cg15431103	ILMN_1798947	ANKRD55	6.23E-08	3.12E-07	-0.509
cg15431103	ILMN_2341724	ANKRD55	1.83E-07	4.58E-07	-0.493
cg21140145	ILMN_1789558	FAM164A	2.54E-07	1.27E-06	-0.488
cg02511570	ILMN_1703301	MRPL45P2	5.13E-07	3.08E-06	-0.477
cg10909506	ILMN_1662174	ORMDL3	6.69E-07	1.07E-05	-0.473
cg25492364	ILMN_1811933	SHMT1	1.19E-06	1.66E-05	-0.464
cg16006841	ILMN_1696828	RGS14	1.12E-06	1.79E-05	-0.465
cg07654569	ILMN_1789558	FAM164A	7.61E-06	3.80E-05	-0.431
cg20676602	ILMN_1703301	MRPL45P2	7.10E-06	4.26E-05	-0.432
cg25875191	ILMN_1696828	RGS14	2.72E-06	4.35E-05	-0.449
cg18711369	ILMN_1662174	ORMDL3	3.16E-06	5.06E-05	-0.447
cg23343972	ILMN_1849013	IL6ST	3.48E-05	5.80E-05	-0.401
cg21124310	ILMN_1849013	IL6ST	3.60E-05	6.01E-05	-0.401
cg03983883	ILMN_2057981	FAM164A	2.68E-05	6.71E-05	-0.407
cg10404427	ILMN_1849013	IL6ST	6.25E-05	1.04E-04	-0.389
cg00184826	ILMN_1682727	JAZF1	5.82E-05	1.16E-04	0.391
cg18711369	ILMN_1666206	GSDMB	2.94E-05	2.35E-04	-0.405
cg05575058	ILMN_2057981	FAM164A	1.61E-04	4.02E-04	-0.369
cg15667493	ILMN_1849013	IL6ST	3.39E-04	5.64E-04	-0.351
cg02116225	ILMN_1811933	SHMT1	4.16E-05	5.82E-04	-0.398
cg15431103	ILMN_1849013	IL6ST	3.53E-04	5.88E-04	-0.350
cg09689469	ILMN_1714393	RAB24	4.46E-05	7.14E-04	0.396
cg16130019	ILMN_1682727	JAZF1	1.48E-04	7.39E-04	0.371
cg10909506	ILMN_1666206	GSDMB	1.02E-04	8.16E-04	-0.379
cg11598255	ILMN_1696828	RGS14	7.14E-05	1.14E-03	-0.386
cg23071186	ILMN_1761764	ALKBH7	8.90E-05	1.16E-03	0.382
cg11187739	ILMN_2374770	TAX1BP1	6.14E-04	1.53E-03	0.337
cg06060754	ILMN_1696828	RGS14	1.05E-04	1.68E-03	-0.378
cg21140145	ILMN_2057981	FAM164A	7.59E-04	1.90E-03	-0.331
cg15431103	ILMN_1797861	IL6ST	1.57E-03	1.96E-03	-0.312
cg08519799	ILMN_1682727	JAZF1	9.94E-04	3.98E-03	0.324
cg01030110	ILMN_3245559	CDK2AP1	3.67E-04	5.88E-03	-0.349
cg07654569	ILMN_2057981	FAM164A	2.81E-03	7.01E-03	-0.296
cg12550541	ILMN_1723846	METTL21B	3.87E-04	7.73E-03	-0.348
cg07654569	ILMN_1762262	PKIA	6.22E-03	7.84E-03	-0.272
cg07654569	ILMN_2337974	PKIA	6.27E-03	7.84E-03	-0.272
cg24044988	ILMN_1812877	ZNF688	2.26E-04	7.92E-03	0.361
cg07522171	ILMN_2374770	TAX1BP1	4.36E-03	8.72E-03	0.283

Asthma (CD4$^+$ T cells)

CpG	IlluminaID	Gene	P-value	Adj. Pval	Rho
cg13899648	ILMN_1676924	CD247	4.37E-06	5.24E-05	-0.441
cg13924073	ILMN_1676924	CD247	6.99E-06	5.59E-05	-0.433
cg10375409	ILMN_1676924	CD247	9.05E-06	1.09E-04	-0.428
cg00080417	ILMN_1676924	CD247	2.27E-05	2.72E-04	-0.410
cg20970810	ILMN_1676924	CD247	3.60E-05	4.32E-04	-0.401
cg26880239	ILMN_1676924	CD247	4.69E-05	5.63E-04	-0.395
cg06674732	ILMN_1676924	CD247	7.63E-04	9.16E-03	-0.331
cg01097872	ILMN_2112301	DRAP1	3.16E-04	8.84E-03	0.353
cg18711369	ILMN_1666206	GSDMB	2.94E-05	2.35E-04	-0.405
cg10909506	ILMN_1666206	GSDMB	1.02E-04	8.16E-04	-0.379
cg00272070	ILMN_1781700	IL18R1	2.44E-06	7.31E-06	-0.451
cg07522171	ILMN_1682727	JAZF1	4.47E-09	1.79E-08	-0.545
cg11187739	ILMN_1682727	JAZF1	4.96E-08	2.48E-07	-0.513
cg00184826	ILMN_1682727	JAZF1	5.82E-05	1.16E-04	0.391
cg16130019	ILMN_1682727	JAZF1	1.48E-04	7.39E-04	0.371
cg08519799	ILMN_1682727	JAZF1	9.94E-04	3.98E-03	0.324
cg24839871	ILMN_1718129	MAP2K5	1.21E-03	6.04E-03	-0.319
cg10909506	ILMN_1662174	ORMDL3	6.69E-07	1.07E-05	-0.473
cg18711369	ILMN_1662174	ORMDL3	3.16E-06	5.06E-05	-0.447
cg00112517	ILMN_1662174	ORMDL3	1.21E-04	1.81E-03	-0.375
cg16484858	ILMN_1802380	RERE	6.32E-05	3.16E-04	-0.389
cg14004768	ILMN_1802380	RERE	1.64E-04	8.22E-04	-0.368
cg15732724	ILMN_1802380	RERE	1.14E-03	3.42E-03	-0.321
cg19712600	ILMN_1747857	SMARCE1	3.47E-04	2.43E-03	0.351
cg11187739	ILMN_2374770	TAX1BP1	6.14E-04	1.53E-03	0.337
cg07522171	ILMN_2374770	TAX1BP1	4.36E-03	8.72E-03	0.283
cg13109634	ILMN_1777487	ZNF839	1.56E-03	7.80E-03	-0.312

Osteoarthritis (CD4$^+$ T cells)

CpG	IlluminaID	Gene	P-value	Adj. Pval	Rho
cg04226788	ILMN_2393693	LRRC37A4	2.03E-14	1.22E-13	0.672
cg17117718	ILMN_2393693	LRRC37A4	5.70E-14	6.27E-13	-0.663
cg18815117	ILMN_2393693	LRRC37A4	6.08E-13	6.69E-12	0.642
cg18228076	ILMN_2393693	LRRC37A4	2.35E-12	1.65E-11	0.629
cg07368061	ILMN_2393693	LRRC37A4	2.27E-11	1.13E-10	0.607
cg20059597	ILMN_2393693	LRRC37A4	1.32E-10	1.45E-09	-0.587
cg06537391	ILMN_2393693	LRRC37A4	3.31E-10	2.98E-09	-0.577
cg01882395	ILMN_2393693	LRRC37A4	4.13E-10	4.54E-09	-0.574
cg03238273	ILMN_2393693	LRRC37A4	2.78E-09	2.51E-08	-0.551
cg04226788	ILMN_1784428	MGC57346	1.29E-08	3.87E-08	-0.531
cg17117718	ILMN_1784428	MGC57346	8.51E-09	4.68E-08	0.537
cg18815117	ILMN_1784428	MGC57346	2.24E-08	1.23E-07	-0.524
cg18228076	ILMN_1784428	MGC57346	9.62E-08	3.37E-07	-0.503
cg09793084	ILMN_2393693	LRRC37A4	6.87E-08	6.18E-07	0.508
cg18753072	ILMN_2393693	LRRC37A4	9.97E-08	1.10E-06	-0.503
cg16520312	ILMN_2393693	LRRC37A4	1.87E-07	1.31E-06	0.493
cg07817266	ILMN_2393693	LRRC37A4	2.66E-07	2.93E-06	-0.488
cg07368061	ILMN_1784428	MGC57346	1.42E-06	3.54E-06	-0.461
cg14598846	ILMN_2370872	GRINA	1.91E-07	5.53E-06	0.493
cg04892187	ILMN_2370872	GRINA	2.54E-07	7.38E-06	-0.488
cg22822867	ILMN_2370872	GRINA	4.15E-07	1.20E-05	-0.481
cg11117266	ILMN_2393693	LRRC37A4	1.96E-06	1.37E-05	0.455
cg24891660	ILMN_2370872	GRINA	5.03E-07	1.46E-05	-0.478
cg15295732	ILMN_2393693	LRRC37A4	2.98E-06	2.09E-05	0.448
cg03238273	ILMN_1784428	MGC57346	4.86E-06	2.19E-05	0.439
cg21900799	ILMN_2370872	GRINA	1.03E-06	2.99E-05	-0.466

cg07531549	ILMN_2370872	GRINA	1.17E-06	3.39E-05	-0.464
cg18878992	ILMN_2393693	LRRC37A4	7.80E-06	5.46E-05	-0.430
cg01882395	ILMN_1784428	MGC57346	1.16E-05	6.39E-05	0.423
cg16520312	ILMN_1784428	MGC57346	2.38E-05	8.33E-05	-0.409
cg09793084	ILMN_1784428	MGC57346	2.21E-05	9.93E-05	-0.411
cg06537391	ILMN_1784428	MGC57346	2.32E-05	1.05E-04	0.410
cg15295732	ILMN_1784428	MGC57346	3.09E-05	1.08E-04	-0.404
cg20784950	ILMN_2370872	GRINA	5.10E-06	1.48E-04	-0.438
cg24044478	ILMN_2370872	GRINA	5.16E-06	1.50E-04	-0.438
cg04757492	ILMN_2370872	GRINA	5.39E-06	1.56E-04	-0.437
cg07817266	ILMN_1784428	MGC57346	3.76E-05	2.07E-04	0.400
cg25475366	ILMN_2370872	GRINA	7.22E-06	2.09E-04	-0.432
cg20059597	ILMN_1784428	MGC57346	9.16E-05	5.04E-04	0.381
cg18753072	ILMN_2200636	KIAA1267	1.14E-04	6.28E-04	0.376
cg07531549	ILMN_1744268	PLEC	7.64E-05	1.11E-03	-0.385
cg14913216	ILMN_2370872	GRINA	4.11E-05	1.19E-03	-0.398
cg08116092	ILMN_1668743	RILPL2	7.89E-05	1.34E-03	-0.384
cg08090367	ILMN_2370872	GRINA	4.73E-05	1.37E-03	-0.395
cg15847845	ILMN_2370872	GRINA	5.28E-05	1.53E-03	0.393
cg02623114	ILMN_1744268	PLEC	5.53E-05	1.60E-03	-0.392
cg08116092	ILMN_1678490	RILPL2	2.25E-04	1.92E-03	-0.361
cg01934064	ILMN_1784428	MGC57346	3.71E-04	2.23E-03	0.349
cg11117266	ILMN_1784428	MGC57346	1.32E-03	4.60E-03	-0.317
cg15633388	ILMN_1680353	NSF	1.26E-03	4.62E-03	-0.318
cg15633388	ILMN_2330845	NSF	1.85E-03	4.62E-03	-0.308
cg11117266	ILMN_1709549	PLEKHM1	1.98E-03	4.62E-03	-0.306
cg08161931	ILMN_2370872	GRINA	1.78E-04	5.17E-03	-0.366
cg00891649	ILMN_2393693	LRRC37A4	7.80E-04	5.46E-03	0.331
cg14913216	ILMN_1744268	PLEC	3.80E-04	5.51E-03	-0.349
cg04892187	ILMN_1744268	PLEC	3.92E-04	5.68E-03	-0.348
cg01030110	ILMN_3245559	CDK2AP1	3.67E-04	5.88E-03	-0.349
cg01934064	ILMN_2393693	LRRC37A4	2.34E-03	7.03E-03	-0.301
cg13389508	ILMN_1744268	PLEC	3.16E-04	9.15E-03	-0.353

Cis-expression quantitative trait methylation (eQTM) at genes within ±500kb of risk-associated cis-meQTL CpGs in B cells

Rheumatoid arthritis (B cells)					
CpG	**IlluminaID**	**Gene**	**P-value**	**Adj. Pval**	**Rho**
cg16429190	ILMN_1687213	FAM167A	2.92E-14	2.92E-13	0.647
cg16429190	ILMN_3248511	FAM167A	1.93E-13	9.67E-13	0.631
cg12749226	ILMN_1662174	ORMDL3	8.88E-12	1.60E-10	-0.595
cg21721331	ILMN_1691693	FCRL3	2.91E-11	1.75E-10	-0.583
cg19602479	ILMN_1691693	FCRL3	3.89E-11	2.33E-10	-0.580
cg01045635	ILMN_1699599	FCRL3	1.73E-10	1.04E-09	-0.564
cg09528494	ILMN_1687213	FAM167A	4.91E-10	2.46E-09	0.552
cg09528494	ILMN_3248511	FAM167A	3.93E-10	2.46E-09	0.555
cg01045635	ILMN_1691693	FCRL3	1.30E-09	2.60E-09	-0.541
cg01045635	ILMN_1797428	FCRL3	1.30E-09	2.60E-09	-0.541
cg01383082	ILMN_3248511	FAM167A	2.61E-10	2.61E-09	-0.559
cg21497594	ILMN_1687213	FAM167A	5.57E-10	4.27E-09	0.551
cg21497594	ILMN_3248511	FAM167A	9.50E-10	4.27E-09	0.544
cg01383082	ILMN_1687213	FAM167A	1.42E-09	7.12E-09	-0.539
cg19602479	ILMN_1699599	FCRL3	3.46E-09	9.43E-09	-0.529
cg19602479	ILMN_1797428	FCRL3	4.72E-09	9.43E-09	-0.525
cg21721331	ILMN_1699599	FCRL3	3.58E-09	1.07E-08	-0.528
cg21721331	ILMN_1797428	FCRL3	1.50E-08	3.00E-08	-0.510

cg15602298	ILMN_1797428	FCRL3	1.70E-08	1.02E-07	-0.508
cg16429190	ILMN_1668277	BLK	4.44E-08	1.48E-07	-0.495
cg12749226	ILMN_1666206	GSDMB	3.48E-08	3.14E-07	-0.498
cg04986849	ILMN_3248511	FAM167A	7.81E-08	7.81E-07	0.487
cg15602298	ILMN_1691693	FCRL3	2.85E-07	8.56E-07	-0.468
cg04986849	ILMN_1687213	FAM167A	2.82E-07	1.41E-06	0.468
cg15602298	ILMN_1699599	FCRL3	3.58E-06	7.16E-06	-0.427
cg11944933	ILMN_1687213	FAM167A	1.59E-06	1.01E-05	-0.441
cg11944933	ILMN_3248511	FAM167A	2.01E-06	1.01E-05	-0.437
cg01383082	ILMN_1668277	BLK	5.37E-06	1.79E-05	0.420
cg15222091	ILMN_1690907	CCR6	8.58E-06	2.57E-05	-0.412
cg16523158	ILMN_1690907	CCR6	1.04E-05	3.13E-05	-0.408
cg23507676	ILMN_1687213	FAM167A	9.79E-06	4.40E-05	0.410
cg23507676	ILMN_3248511	FAM167A	9.65E-06	4.40E-05	0.410
cg01527115	ILMN_1687213	FAM167A	5.45E-06	5.21E-05	-0.420
cg01527115	ILMN_3248511	FAM167A	1.04E-05	5.21E-05	-0.408
cg13200575	ILMN_1662174	ORMDL3	6.22E-06	1.12E-04	-0.418
cg25259754	ILMN_1691693	FCRL3	2.00E-05	1.20E-04	-0.396
cg19954286	ILMN_1690907	CCR6	4.42E-05	1.33E-04	-0.381
cg18691862	ILMN_1662174	ORMDL3	8.07E-06	1.45E-04	-0.413
cg14348996	ILMN_1662174	ORMDL3	9.07E-06	1.54E-04	0.411
cg25259754	ILMN_1797428	FCRL3	7.24E-05	2.17E-04	-0.371
cg21497594	ILMN_1668277	BLK	8.38E-05	2.51E-04	-0.368
cg17134153	ILMN_1797428	FCRL3	4.32E-05	2.59E-04	-0.381
cg25259754	ILMN_1699599	FCRL3	1.32E-04	2.65E-04	-0.358
cg05094429	ILMN_1690907	CCR6	9.52E-05	2.86E-04	-0.365
cg17134153	ILMN_1691693	FCRL3	1.06E-04	2.96E-04	-0.363
cg17134153	ILMN_1699599	FCRL3	1.48E-04	2.96E-04	-0.356
cg09528494	ILMN_1668277	BLK	1.49E-04	4.97E-04	-0.355
cg23507676	ILMN_1668277	BLK	1.71E-04	5.14E-04	-0.352
cg14348996	ILMN_1666206	GSDMB	6.67E-05	5.67E-04	0.372
cg03002059	ILMN_1687213	FAM167A	1.01E-04	6.97E-04	0.364
cg03002059	ILMN_3248511	FAM167A	1.55E-04	6.97E-04	0.355
cg04986849	ILMN_1668277	BLK	2.14E-04	7.14E-04	-0.347
cg00288844	ILMN_1771862	TXNDC11	6.84E-05	8.20E-04	-0.372
cg21794222	ILMN_1690907	CCR6	4.67E-04	1.40E-03	-0.330
cg12816198	ILMN_1670576	IRF5	1.81E-04	1.63E-03	-0.351
cg21497594	ILMN_1715680	NEIL2	8.75E-04	1.97E-03	0.314
cg01527115	ILMN_1668277	BLK	6.27E-04	2.09E-03	0.322
cg12655416	ILMN_1666206	GSDMB	1.82E-04	3.28E-03	-0.351
cg09528494	ILMN_1724762	XKR6	1.39E-03	3.49E-03	-0.302
cg18711369	ILMN_1662174	ORMDL3	2.33E-04	3.50E-03	-0.346
cg18711369	ILMN_1666206	GSDMB	3.88E-04	3.50E-03	-0.334
cg21473142	ILMN_2200917	SLC4A7	1.96E-03	3.92E-03	0.293
cg14348996	ILMN_3245973	MSL1	7.83E-04	4.44E-03	0.317
cg18691862	ILMN_2300695	IKZF3	9.64E-04	8.68E-03	0.312
cg18691862	ILMN_1707448	CDK12	1.49E-03	8.95E-03	0.301

Multiple sclerosis (B cells)

CpG	IlluminaID	Gene	P-value	Adj. Pval	Rho
cg03983883	ILMN_2057981	FAM164A	4.25E-21	2.55E-20	-0.752
cg21140145	ILMN_2057981	FAM164A	7.83E-21	4.70E-20	-0.749
cg03983883	ILMN_1789558	FAM164A	2.45E-19	7.34E-19	-0.729
cg21140145	ILMN_1789558	FAM164A	2.47E-18	7.41E-18	-0.715
cg01951420	ILMN_1708798	EAF2	1.75E-18	1.23E-17	-0.717
cg12032497	ILMN_1708798	EAF2	4.96E-14	3.47E-13	-0.643
cg24574508	ILMN_1708798	EAF2	7.82E-14	5.47E-13	-0.639

cg07654569	ILMN_2057981	FAM164A	1.53E-11	9.15E-11	-0.590
cg12749226	ILMN_1662174	ORMDL3	8.88E-12	1.60E-10	-0.595
cg07654569	ILMN_1789558	FAM164A	1.48E-10	4.43E-10	-0.566
cg12749226	ILMN_1666206	GSDMB	3.48E-08	3.14E-07	-0.498
cg05575058	ILMN_1789558	FAM164A	2.30E-07	1.38E-06	-0.471
cg02586212	ILMN_1656011	RGS1	4.00E-07	4.00E-06	-0.463
cg09871101	ILMN_2057981	FAM164A	1.03E-06	6.20E-06	0.448
cg05575058	ILMN_2057981	FAM164A	4.12E-06	1.23E-05	-0.425
cg25492364	ILMN_1811933	SHMT1	9.09E-07	1.36E-05	-0.450
cg01030110	ILMN_1812721	HIP1R	9.27E-07	1.67E-05	0.450
cg09871101	ILMN_1789558	FAM164A	1.07E-05	3.22E-05	0.408
cg10605766	ILMN_1708798	EAF2	1.25E-05	8.77E-05	-0.405
cg12032497	ILMN_1747935	GOLGB1	1.41E-04	4.94E-04	0.357
cg01007589	ILMN_1682781	TEAD2	2.05E-05	7.16E-04	-0.396
cg00599273	ILMN_1767481	XRCC6BP1	5.00E-05	1.00E-03	-0.378
cg12032497	ILMN_2316104	IQCB1	7.29E-04	1.70E-03	-0.319
cg21140145	ILMN_1762262	PKIA	1.28E-03	2.56E-03	-0.305
cg01951420	ILMN_2316104	IQCB1	8.16E-04	2.86E-03	-0.316
cg03983883	ILMN_1762262	PKIA	1.60E-03	3.19E-03	-0.299
cg12655416	ILMN_1666206	GSDMB	1.82E-04	3.28E-03	-0.351
cg00599273	ILMN_1723846	METTL21B	4.28E-04	4.28E-03	0.332
cg10024583	ILMN_1703301	MRPL45P2	3.52E-04	4.57E-03	0.336
cg02189760	ILMN_1682781	TEAD2	1.70E-04	5.60E-03	-0.353
cg07418126	ILMN_1682781	TEAD2	1.97E-04	6.49E-03	-0.349
cg11428475	ILMN_1662174	ORMDL3	4.22E-04	8.02E-03	-0.332
cg01007589	ILMN_2375825	CD37	4.59E-04	8.04E-03	-0.330

Asthma (B cells)					
CpG	**IlluminaID**	**Gene**	**P-value**	**Adj. Pval**	**Rho**
cg12749226	ILMN_1662174	ORMDL3	8.88E-12	1.60E-10	-0.595
cg12749226	ILMN_1666206	GSDMB	3.48E-08	3.14E-07	-0.498
cg24910161	ILMN_1662174	ORMDL3	1.40E-07	2.25E-06	-0.479
cg26162295	ILMN_1662174	ORMDL3	2.25E-07	3.59E-06	-0.472
cg23202472	ILMN_1666206	GSDMB	3.05E-06	5.49E-05	-0.430
cg11817230	ILMN_1662174	ORMDL3	5.39E-06	9.70E-05	-0.420
cg13200575	ILMN_1662174	ORMDL3	6.22E-06	1.12E-04	-0.418
cg19758448	ILMN_1662174	ORMDL3	1.48E-05	1.33E-04	-0.402
cg19758448	ILMN_1805636	PGAP3	1.21E-05	1.33E-04	-0.406
cg18691862	ILMN_1662174	ORMDL3	8.07E-06	1.45E-04	-0.413
cg14348996	ILMN_1662174	ORMDL3	9.07E-06	1.54E-04	0.411
cg11817230	ILMN_1805636	PGAP3	3.83E-05	3.44E-04	-0.384
cg11817230	ILMN_1666206	GSDMB	6.65E-05	3.99E-04	-0.372
cg23202472	ILMN_1662174	ORMDL3	5.79E-05	5.21E-04	-0.375
cg14348996	ILMN_1666206	GSDMB	6.67E-05	5.67E-04	0.372
cg24910161	ILMN_1666206	GSDMB	1.11E-04	8.90E-04	-0.362
cg26162295	ILMN_1666206	GSDMB	1.41E-04	1.13E-03	-0.357
cg16600909	ILMN_2089875	TNFSF4	7.05E-04	1.41E-03	-0.320
cg04317648	ILMN_1802380	RERE	6.65E-04	2.66E-03	-0.321
cg12655416	ILMN_1666206	GSDMB	1.82E-04	3.28E-03	-0.351
cg18711369	ILMN_1662174	ORMDL3	2.33E-04	3.50E-03	-0.346
cg18711369	ILMN_1666206	GSDMB	3.88E-04	3.50E-03	-0.334
cg12183861	ILMN_1703301	MRPL45P2	2.79E-04	3.63E-03	-0.341
cg14348996	ILMN_3245973	MSL1	7.83E-04	4.44E-03	0.317
cg24211550	ILMN_1747857	SMARCE1	5.88E-04	5.29E-03	0.324
cg11428475	ILMN_1662174	ORMDL3	4.22E-04	8.02E-03	-0.332
cg14004768	ILMN_1802380	RERE	2.11E-03	8.45E-03	-0.291
cg18691862	ILMN_2300695	IKZF3	9.64E-04	8.68E-03	0.312
cg18691862	ILMN_1707448	CDK12	1.49E-03	8.95E-03	0.301
cg02551532	ILMN_1666206	GSDMB	5.26E-04	9.47E-03	0.327

Osteoarthritis (B cells)					
CpG	**IlluminaID**	**Gene**	**P-value**	**Adj. Pval**	**Rho**
cg17117718	ILMN_2393693	LRRC37A4	3.41E-16	4.10E-15	-0.682
cg18228076	ILMN_2393693	LRRC37A4	1.95E-15	1.36E-14	0.669
cg18815117	ILMN_2393693	LRRC37A4	5.73E-15	6.88E-14	0.660
cg09793084	ILMN_2393693	LRRC37A4	2.07E-14	1.86E-13	0.650
cg04226788	ILMN_2393693	LRRC37A4	1.34E-13	8.04E-13	0.634
cg07368061	ILMN_2393693	LRRC37A4	6.47E-12	3.24E-11	0.598
cg20059597	ILMN_2393693	LRRC37A4	5.69E-11	6.82E-10	-0.576
cg04757492	ILMN_2370872	GRINA	9.33E-11	2.80E-09	-0.571
cg24891660	ILMN_2370872	GRINA	1.51E-10	4.52E-09	-0.565
cg06537391	ILMN_2393693	LRRC37A4	1.41E-08	1.27E-07	-0.511
cg18878992	ILMN_2393693	LRRC37A4	3.29E-08	2.30E-07	-0.499
cg04226788	ILMN_1784428	MGC57346	1.01E-07	3.02E-07	-0.484
cg00916973	ILMN_2393693	LRRC37A4	4.76E-08	5.71E-07	0.494
cg01882395	ILMN_2393693	LRRC37A4	6.32E-08	7.58E-07	-0.490
cg03238273	ILMN_2393693	LRRC37A4	1.11E-07	1.00E-06	-0.482
cg07817266	ILMN_2393693	LRRC37A4	1.15E-07	1.26E-06	-0.482
cg16520312	ILMN_2393693	LRRC37A4	2.27E-07	1.59E-06	0.472
cg25999728	ILMN_1754121	CSK	1.71E-07	2.39E-06	-0.476
cg15411667	ILMN_2393693	LRRC37A4	4.24E-07	2.54E-06	-0.462
cg14598846	ILMN_2370872	GRINA	1.20E-07	3.59E-06	0.481
cg00916973	ILMN_1784428	MGC57346	8.53E-07	5.12E-06	-0.451
cg07368061	ILMN_1784428	MGC57346	3.24E-06	8.10E-06	-0.429
cg17117718	ILMN_1784428	MGC57346	1.67E-06	1.00E-05	0.440
cg18815117	ILMN_1784428	MGC57346	2.54E-06	1.52E-05	-0.433
cg01030110	ILMN_1812721	HIP1R	9.27E-07	1.67E-05	0.450
cg20059597	ILMN_1784428	MGC57346	4.38E-06	2.63E-05	0.424
cg20784950	ILMN_2370872	GRINA	9.73E-07	2.92E-05	-0.449
cg23659289	ILMN_2393693	LRRC37A4	3.26E-06	3.92E-05	-0.429
cg14772590	ILMN_1754121	CSK	3.34E-06	4.67E-05	-0.429
cg03238273	ILMN_1784428	MGC57346	1.10E-05	4.93E-05	0.407
cg15847845	ILMN_2370872	GRINA	1.74E-06	5.21E-05	0.440
cg14838715	ILMN_1754121	CSK	4.02E-06	5.63E-05	-0.425
cg06537391	ILMN_1784428	MGC57346	1.81E-05	8.13E-05	0.398
cg18228076	ILMN_1784428	MGC57346	2.38E-05	8.32E-05	-0.393
cg05301556	ILMN_2393693	LRRC37A4	2.55E-05	1.79E-04	-0.392
cg02331830	ILMN_2370872	GRINA	7.10E-06	2.13E-04	0.415
cg09214591	ILMN_1738239	RBM6	1.04E-05	2.19E-04	0.408
cg25475366	ILMN_2370872	GRINA	8.73E-06	2.62E-04	-0.412
cg19124816	ILMN_1815205	LYZ	4.55E-05	2.73E-04	-0.380
cg15633388	ILMN_1680353	NSF	6.00E-05	3.60E-04	-0.375
cg12396344	ILMN_1815205	LYZ	8.01E-05	4.81E-04	0.369
cg09860564	ILMN_1680353	NSF	8.09E-05	4.86E-04	-0.368
cg21900799	ILMN_2370872	GRINA	2.12E-05	6.37E-04	-0.395
cg01640727	ILMN_1784428	MGC57346	2.13E-04	6.38E-04	-0.348
cg01640727	ILMN_2393693	LRRC37A4	1.43E-04	6.38E-04	0.356
cg09793084	ILMN_1784428	MGC57346	1.44E-04	6.47E-04	-0.356
cg16131304	ILMN_1745152	UQCC	3.70E-05	8.15E-04	-0.384
cg15633388	ILMN_2330845	NSF	4.16E-04	1.25E-03	-0.332
cg07817266	ILMN_1784428	MGC57346	2.38E-04	1.31E-03	0.345
cg01527957	ILMN_2393693	LRRC37A4	1.47E-04	1.76E-03	0.356
cg22375663	ILMN_1815205	LYZ	3.45E-04	2.07E-03	0.337
cg03954353	ILMN_2393693	LRRC37A4	1.73E-04	2.08E-03	-0.352
cg05301556	ILMN_1784428	MGC57346	6.56E-04	2.29E-03	0.321
cg19124816	ILMN_1801387	YEATS4	7.77E-04	2.33E-03	-0.317
cg04892187	ILMN_2370872	GRINA	9.63E-05	2.89E-03	-0.365
cg02331830	ILMN_1721411	PARP10	2.12E-04	3.18E-03	0.348

cg09764761	ILMN_1784428	MGC57346	9.66E-04	3.86E-03	-0.312
cg11117266	ILMN_2393693	LRRC37A4	5.75E-04	4.03E-03	0.325
cg23659289	ILMN_1784428	MGC57346	7.85E-04	4.71E-03	0.317
cg12520615	ILMN_2370872	GRINA	1.75E-04	4.89E-03	0.352
cg07298766	ILMN_1784428	MGC57346	1.58E-03	6.30E-03	-0.299
cg18878992	ILMN_1784428	MGC57346	2.05E-03	7.18E-03	0.292
cg16520312	ILMN_1784428	MGC57346	2.12E-03	7.43E-03	-0.291
cg02322039	ILMN_2330845	NSF	1.32E-03	7.94E-03	-0.304
cg09860564	ILMN_2330845	NSF	2.79E-03	8.37E-03	-0.284
cg04255391	ILMN_2370872	GRINA	3.15E-04	9.44E-03	0.339

Appendix H – Causal Inference Testing

Causal Inference Test (CIT) results for all triplets (SNP, CpG probe, Transcript probe) at cis-meQTLs/cis-eQTMs associated with risk loci (rheumatoid arthritis, multiple sclerosis, asthma, and osteoarthritis) in CD4$^+$ T cells and B cells (previous page). Pval CIT = CIT 'omnibus' p-value to test DNA methylation as a mediator of gene expression levels at risk loci (see Chapter 2.); FDR CIT = False discovery rate (FDR)-corrected CIT p-value generated using 1000 permutations of the data. FDR values are also given for the four component tests: EaL - Expression (Transcript Levels) is associated with Locus (Risk SNP); EaMgvL - Expression (E) is associated with Methylation (M) given the Locus (L); MaLgvE - Methylation is associated with the Locus given Expression; LiEgvM - The Locus is independent of Expression given Methylation.

Results from Causal Inference Testing at all Risk cis-meQTL/eQTM sites in CD4$^+$ T cells

Rheumatoid arthritis (CD4$^+$ T cells)									
SNP	**CpG**	**IlluminaID**	**Gene**	**Pval CIT**	**FDR CIT**	**FDR EaL**	**FDR EaMgvL**	**FDR MaLgvE**	**FDR LiEgvM**
rs6859219	cg21124310	ILMN_1798947	ANKRD55	1.45E-04	7.06E-04	1.32E-05	8.73E-05	1.30E-04	4.76E-04
rs6859219	cg21124310	ILMN_2341724	ANKRD55	1.11E-04	7.06E-04	1.32E-05	8.73E-05	1.30E-04	4.76E-04
rs2189966	cg07522171	ILMN_1682727	JAZF1	3.97E-04	3.45E-03	1.32E-05	7.44E-04	2.27E-05	3.16E-03
rs12946510	cg18711369	ILMN_1662174	ORMDL3	4.46E-04	3.45E-03	1.32E-05	1.79E-04	1.00E-04	3.16E-03
rs12946510	cg10909506	ILMN_1662174	ORMDL3	1.64E-03	4.39E-03	1.32E-05	8.73E-05	1.14E-03	3.16E-03
rs2210913	cg17134153	ILMN_1691693	FCRL3	2.66E-03	4.40E-03	1.32E-05	5.59E-03	2.27E-05	7.60E-03
rs4722758	cg11187739	ILMN_1682727	JAZF1	3.49E-03	4.40E-03	1.32E-05	6.67E-03	2.27E-05	1.01E-02
rs6859219	cg10404427	ILMN_1798947	ANKRD55	5.67E-03	4.40E-03	1.32E-05	8.73E-05	3.38E-03	9.29E-04
rs6859219	cg10404427	ILMN_2341724	ANKRD55	3.11E-03	4.40E-03	1.32E-05	1.79E-04	2.06E-03	3.16E-03
rs6859219	cg23343972	ILMN_1798947	ANKRD55	6.24E-03	6.93E-03	1.32E-05	1.79E-04	3.59E-03	3.16E-03
rs6859219	cg23343972	ILMN_2341724	ANKRD55	7.23E-03	7.22E-03	1.32E-05	8.73E-05	3.97E-03	3.16E-03
rs2210913	cg01045635	ILMN_1691693	FCRL3	1.20E-02	2.96E-02	1.32E-05	1.50E-02	2.27E-05	1.88E-02
rs61897793	cg16213375	ILMN_1786759	C11orf10	1.63E-02	2.96E-02	7.56E-04	1.62E-02	2.27E-05	1.29E-02
rs6859219	cg15431103	ILMN_1849013	IL6ST	1.16E-02	2.96E-02	1.32E-05	1.43E-02	7.83E-04	1.88E-02
rs6859219	cg15431103	ILMN_1797861	IL6ST	9.95E-03	2.96E-02	1.03E-03	1.45E-02	3.04E-04	1.88E-02
rs6859219	cg15667493	ILMN_1849013	IL6ST	1.39E-02	2.96E-02	1.32E-05	1.50E-02	1.91E-03	1.98E-02
rs2210913	cg17134153	ILMN_1797428	FCRL3	2.62E-02	3.05E-02	1.32E-05	2.28E-02	2.27E-05	2.03E-02
rs4722758	cg11187739	ILMN_2374770	TAX1BP1	4.70E-02	3.05E-02	1.04E-03	3.48E-02	2.27E-05	2.42E-02
rs6859219	cg10404427	ILMN_1849013	IL6ST	2.16E-02	3.05E-02	1.32E-05	2.00E-02	2.27E-05	2.08E-02
rs6859219	cg21124310	ILMN_1849013	IL6ST	3.49E-02	3.05E-02	1.32E-05	2.70E-02	2.27E-05	2.15E-02
rs6859219	cg15431103	ILMN_2341724	ANKRD55	4.96E-02	3.05E-02	1.32E-05	8.73E-05	2.69E-02	3.56E-03
rs6859219	cg23343972	ILMN_1849013	IL6ST	3.52E-02	3.05E-02	1.32E-05	2.70E-02	4.35E-05	2.39E-02
rs12946510	cg18711369	ILMN_1666206	GSDMB	2.77E-02	3.05E-02	1.32E-05	1.50E-02	2.27E-05	2.53E-02
rs12946510	cg10909506	ILMN_1666206	GSDMB	4.48E-02	3.05E-02	1.32E-05	2.48E-02	6.25E-05	3.65E-02
rs6859219	cg15431103	ILMN_1798947	ANKRD55	5.36E-02	3.19E-02	1.32E-05	8.73E-05	2.83E-02	3.56E-03
rs917117	cg16130019	ILMN_1682727	JAZF1	5.29E-02	3.19E-02	1.32E-05	3.83E-02	2.27E-05	3.67E-02
rs2210913	cg01045635	ILMN_1797428	FCRL3	6.93E-02	8.22E-02	1.32E-05	4.74E-02	2.27E-05	3.65E-02
rs7522061	cg18707136	ILMN_1797428	FCRL3	6.80E-02	8.22E-02	1.32E-05	4.22E-02	9.35E-04	4.60E-02
rs2189966	cg08519799	ILMN_1682727	JAZF1	9.37E-02	1.08E-01	1.32E-05	6.20E-02	1.91E-03	4.75E-02
rs2210913	cg21721331	ILMN_1691693	FCRL3	1.09E-01	1.11E-01	1.32E-05	6.77E-02	2.27E-05	4.60E-02
rs2893312	cg00184826	ILMN_1682727	JAZF1	1.09E-01	1.12E-01	1.32E-05	6.77E-02	2.27E-05	4.75E-02
rs2210913	cg08786003	ILMN_1691693	FCRL3	2.07E-01	1.21E-01	1.32E-05	1.24E-01	2.27E-05	7.21E-02
rs6859219	cg15667493	ILMN_1798947	ANKRD55	2.33E-01	1.21E-01	1.32E-05	8.73E-05	1.18E-01	3.16E-03
rs2189966	cg07522171	ILMN_2374770	TAX1BP1	1.28E-01	1.21E-01	4.35E-03	7.78E-02	2.27E-05	4.44E-02
rs6859219	cg15667493	ILMN_2341724	ANKRD55	2.59E-01	1.30E-01	1.32E-05	8.73E-05	1.28E-01	3.16E-03
rs2210913	cg21721331	ILMN_1797428	FCRL3	2.61E-01	2.25E-01	1.32E-05	1.53E-01	2.27E-05	8.45E-02
rs7522061	cg18707136	ILMN_1691693	FCRL3	2.84E-01	2.46E-01	1.32E-05	1.62E-01	6.72E-04	9.98E-02
rs2210913	cg19602479	ILMN_1691693	FCRL3	3.09E-01	2.47E-01	1.32E-05	1.71E-01	2.27E-05	9.16E-02
rs2210913	cg19602479	ILMN_1797428	FCRL3	3.31E-01	2.57E-01	1.32E-05	1.78E-01	2.27E-05	9.57E-02
rs7522061	cg25259754	ILMN_1691693	FCRL3	4.19E-01	3.12E-01	1.32E-05	2.19E-01	2.27E-05	1.19E-01
rs2210913	cg08786003	ILMN_1797428	FCRL3	4.59E-01	3.26E-01	1.32E-05	2.35E-01	2.27E-05	1.19E-01
rs7522061	cg25259754	ILMN_1797428	FCRL3	4.73E-01	3.32E-01	1.32E-05	2.36E-01	2.27E-05	1.26E-01

Multiple sclerosis (CD4$^+$ T cells)

SNP	CpG	IlluminaID	Gene	Pval CIT	FDR CIT	FDR EaL	FDR EaMgvL	FDR MaLgvE	FDR LiEgvM
rs6859219	cg21124310	ILMN_1798947	ANKRD55	1.45E-04	6.53E-04	1.32E-05	9.19E-05	6.76E-05	4.81E-04
rs6859219	cg21124310	ILMN_2341724	ANKRD55	1.11E-04	6.53E-04	1.32E-05	9.19E-05	6.76E-05	4.81E-04
rs2189966	cg07522171	ILMN_1682727	JAZF1	3.97E-04	4.93E-03	1.32E-05	1.15E-03	1.56E-05	4.61E-03
rs12946510	cg18711369	ILMN_1662174	ORMDL3	4.46E-04	4.93E-03	1.32E-05	1.92E-04	5.71E-05	4.66E-03
rs4722758	cg11187739	ILMN_1682727	JAZF1	3.49E-03	5.27E-03	1.32E-05	8.95E-03	1.56E-05	1.35E-02
rs6859219	cg10404427	ILMN_1798947	ANKRD55	5.67E-03	5.27E-03	1.32E-05	9.19E-05	3.28E-03	1.88E-03
rs6859219	cg10404427	ILMN_2341724	ANKRD55	3.11E-03	5.27E-03	1.32E-05	1.92E-04	1.84E-03	4.61E-03
rs12946510	cg10909506	ILMN_1662174	ORMDL3	1.64E-03	5.27E-03	1.32E-05	9.19E-05	1.02E-03	4.61E-03
rs6859219	cg23343972	ILMN_1798947	ANKRD55	6.24E-03	8.23E-03	1.32E-05	1.92E-04	3.44E-03	4.61E-03
rs6859219	cg23343972	ILMN_2341724	ANKRD55	7.23E-03	8.64E-03	1.32E-05	9.19E-05	3.95E-03	4.61E-03
rs1021156	cg03983883	ILMN_2057981	FAM164A	1.75E-02	3.16E-02	8.54E-05	2.24E-02	1.56E-05	1.97E-02
rs4722758	cg11187739	ILMN_2374770	TAX1BP1	4.70E-02	3.16E-02	1.39E-03	4.05E-02	1.56E-05	2.71E-02
rs6859219	cg10404427	ILMN_1849013	IL6ST	2.16E-02	3.16E-02	9.78E-05	2.46E-02	1.56E-05	2.64E-02
rs6859219	cg21124310	ILMN_1849013	IL6ST	3.49E-02	3.16E-02	9.78E-05	3.11E-02	1.56E-05	2.69E-02
rs6556313	cg09689469	ILMN_1714393	RAB24	2.00E-02	3.16E-02	1.06E-02	1.14E-02	1.56E-05	1.81E-02
rs6859219	cg15431103	ILMN_2341724	ANKRD55	4.96E-02	3.16E-02	1.32E-05	9.19E-05	2.67E-02	4.91E-03
rs6859219	cg15431103	ILMN_1849013	IL6ST	1.16E-02	3.16E-02	9.78E-05	1.81E-02	6.10E-04	2.50E-02
rs6859219	cg15431103	ILMN_1797861	IL6ST	9.95E-03	3.16E-02	1.29E-03	1.81E-02	1.25E-04	2.38E-02
rs7206971	cg02511570	ILMN_1703301	MRPL45P2	2.95E-02	3.16E-02	1.32E-05	2.87E-02	1.56E-05	2.71E-02
rs6859219	cg15667493	ILMN_1849013	IL6ST	1.39E-02	3.16E-02	9.78E-05	1.93E-02	1.74E-03	2.64E-02
rs6859219	cg23343972	ILMN_1849013	IL6ST	3.52E-02	3.16E-02	9.78E-05	3.11E-02	2.94E-05	2.71E-02
rs2605141	cg25492364	ILMN_1811933	SHMT1	1.53E-02	3.16E-02	1.32E-05	2.11E-02	1.56E-05	2.64E-02
rs12946510	cg18711369	ILMN_1666206	GSDMB	2.77E-02	3.16E-02	1.32E-05	2.10E-02	1.56E-05	2.83E-02
rs12946510	cg10909506	ILMN_1666206	GSDMB	4.48E-02	3.16E-02	1.32E-05	2.87E-02	2.94E-05	3.80E-02
rs7220935	cg02511570	ILMN_1703301	MRPL45P2	1.82E-02	3.16E-02	1.32E-05	2.24E-02	1.56E-05	2.64E-02
rs7220935	cg20676602	ILMN_1703301	MRPL45P2	4.86E-02	3.16E-02	1.32E-05	4.06E-02	1.56E-05	3.65E-02
rs1132812	cg24044988	ILMN_1812877	ZNF688	5.22E-02	3.34E-02	2.65E-02	1.81E-02	7.89E-05	1.81E-02
rs6859219	cg15431103	ILMN_1798947	ANKRD55	5.36E-02	3.34E-02	1.32E-05	9.19E-05	2.85E-02	4.91E-03
rs917117	cg16130019	ILMN_1682727	JAZF1	5.29E-02	3.34E-02	1.32E-05	4.29E-02	1.56E-05	3.83E-02
rs12654812	cg25875191	ILMN_1696828	RGS14	5.93E-02	7.52E-02	1.32E-05	4.67E-02	1.56E-05	3.80E-02
rs12654812	cg16006841	ILMN_1696828	RGS14	6.13E-02	7.52E-02	1.32E-05	4.68E-02	1.56E-05	2.97E-02
rs2605141	cg02116225	ILMN_1811933	SHMT1	6.67E-02	9.06E-02	1.32E-05	4.94E-02	1.56E-05	4.33E-02
rs7206971	cg20676602	ILMN_1703301	MRPL45P2	8.03E-02	9.81E-02	1.32E-05	5.79E-02	1.56E-05	4.27E-02
rs1021156	cg05575058	ILMN_2057981	FAM164A	9.91E-02	1.08E-01	8.54E-05	6.79E-02	1.56E-05	4.33E-02
rs2189966	cg08519799	ILMN_1682727	JAZF1	9.37E-02	1.08E-01	1.32E-05	6.59E-02	1.74E-03	4.87E-02
rs1021156	cg03983883	ILMN_1789558	FAM164A	1.09E-01	1.13E-01	1.32E-05	7.12E-02	1.56E-05	4.51E-02
rs2893312	cg00184826	ILMN_1682727	JAZF1	1.09E-01	1.17E-01	1.32E-05	7.12E-02	1.56E-05	4.92E-02
rs1021156	cg05575058	ILMN_1789558	FAM164A	2.27E-01	1.22E-01	1.32E-05	1.42E-01	1.56E-05	7.83E-02
rs6859219	cg15667493	ILMN_1798947	ANKRD55	2.33E-01	1.22E-01	1.32E-05	9.19E-05	1.18E-01	4.61E-03
rs2189966	cg07522171	ILMN_2374770	TAX1BP1	1.28E-01	1.22E-01	4.49E-03	8.20E-02	1.56E-05	4.36E-02
rs6859219	cg15667493	ILMN_2341724	ANKRD55	2.59E-01	1.32E-01	1.32E-05	9.19E-05	1.28E-01	4.61E-03
rs12654812	cg06060754	ILMN_1696828	RGS14	2.76E-01	2.41E-01	1.32E-05	1.69E-01	1.56E-05	8.70E-02
rs1021156	cg21140145	ILMN_2057981	FAM164A	3.19E-01	2.62E-01	8.54E-05	1.86E-01	1.56E-05	9.40E-02
rs1021156	cg21140145	ILMN_1789558	FAM164A	4.69E-01	3.72E-01	1.32E-05	2.67E-01	1.56E-05	1.43E-01
rs3899796	cg07654569	ILMN_1762262	PKIA	5.10E-01	3.86E-01	1.32E-05	2.83E-01	1.56E-05	1.43E-01
rs703842	cg12550541	ILMN_1723846	METTL21B	6.24E-01	4.61E-01	1.32E-05	3.39E-01	1.25E-04	1.84E-01
rs641760	cg01030110	ILMN_3245559	CDK2AP1	6.85E-01	4.90E-01	1.32E-05	3.65E-01	1.56E-05	1.96E-01
rs12654812	cg11598255	ILMN_1696828	RGS14	8.00E-01	5.28E-01	1.32E-05	4.07E-01	1.56E-05	2.03E-01
rs3899796	cg07654569	ILMN_2057981	FAM164A	7.15E-01	5.28E-01	1.32E-05	3.73E-01	1.56E-05	3.25E-01
rs3899796	cg07654569	ILMN_1789558	FAM164A	8.47E-01	5.31E-01	1.32E-05	1.73E-01	1.56E-05	4.34E-01
rs3899796	cg07654569	ILMN_2337974	PKIA	9.25E-01	5.93E-01	1.32E-05	4.62E-01	1.56E-05	2.44E-01

Asthma (CD4$^+$ T cells)

SNP	CpG	IlluminaID	Gene	Pval CIT	FDR CIT	FDR EaL	FDR EaMgvL	FDR MaLgvE	FDR LiEgvM
rs2189966	cg07522171	ILMN_1682727	JAZF1	3.97E-04	4.77E-03	3.57E-05	3.89E-04	2.78E-05	4.35E-03
rs12946510	cg18711369	ILMN_1662174	ORMDL3	4.46E-04	4.77E-03	3.57E-05	2.96E-04	9.52E-05	4.35E-03
rs12946510	cg10909506	ILMN_1662174	ORMDL3	1.64E-03	5.55E-03	3.57E-05	2.96E-04	8.80E-04	4.35E-03
rs4722758	cg11187739	ILMN_1682727	JAZF1	3.49E-03	1.09E-02	3.57E-05	4.11E-03	2.78E-05	1.00E-02
rs1773542	cg00080417	ILMN_1676924	CD247	8.61E-03	1.09E-02	6.20E-03	3.89E-04	2.78E-05	4.35E-03

rs1617333	cg10375409	ILMN_1676924	CD247	8.10E-03	1.09E-02	6.20E-03	3.89E-04	2.78E-05	4.35E-03
rs1617333	cg13899648	ILMN_1676924	CD247	8.10E-03	1.09E-02	6.20E-03	1.48E-03	2.78E-05	5.27E-03
rs1617333	cg26880239	ILMN_1676924	CD247	8.10E-03	1.09E-02	6.20E-03	4.23E-04	2.78E-05	4.35E-03
rs1617333	cg20970810	ILMN_1676924	CD247	8.10E-03	1.09E-02	6.20E-03	8.80E-04	2.78E-05	5.27E-03
rs2988279	cg13924073	ILMN_1676924	CD247	3.47E-02	2.45E-02	1.99E-02	2.96E-04	2.78E-05	4.35E-03
rs1773542	cg06674732	ILMN_1676924	CD247	8.61E-03	2.45E-02	6.20E-03	4.45E-03	5.00E-05	1.42E-02
rs2517953	cg00112517	ILMN_1662174	ORMDL3	2.33E-02	2.45E-02	1.03E-03	1.76E-02	3.86E-04	3.09E-02
rs12946510	cg18711369	ILMN_1666206	GSDMB	2.77E-02	2.45E-02	3.57E-05	1.15E-02	2.78E-05	3.09E-02
rs12946510	cg10909506	ILMN_1666206	GSDMB	4.48E-02	6.28E-02	3.57E-05	2.37E-02	5.00E-05	4.00E-02
rs4722758	cg11187739	ILMN_2374770	TAX1BP1	4.70E-02	6.81E-02	3.28E-03	3.52E-02	2.78E-05	3.09E-02
rs1420101	cg00272070	ILMN_1781700	IL18R1	4.92E-02	7.42E-02	3.57E-05	3.52E-02	3.91E-04	4.00E-02
rs917117	cg16130019	ILMN_1682727	JAZF1	5.29E-02	7.58E-02	3.57E-05	3.72E-02	2.78E-05	4.00E-02
rs71421262	cg13109634	ILMN_1777487	ZNF839	5.78E-02	8.37E-02	6.98E-03	3.88E-02	2.78E-05	4.00E-02
rs2189966	cg08519799	ILMN_1682727	JAZF1	9.37E-02	1.09E-01	3.57E-05	6.17E-02	1.52E-03	4.85E-02
rs2893312	cg00184826	ILMN_1682727	JAZF1	1.09E-01	1.13E-01	3.57E-05	6.77E-02	2.78E-05	4.85E-02
rs2189966	cg07522171	ILMN_2374770	TAX1BP1	1.28E-01	1.23E-01	6.20E-03	7.62E-02	2.78E-05	4.43E-02
rs72743461	cg24839871	ILMN_1718129	MAP2K5	1.88E-01	1.48E-01	4.18E-02	4.45E-03	2.78E-05	1.07E-01
rs479844	cg01097872	ILMN_2112301	DRAP1	1.96E-01	1.84E-01	8.53E-02	1.76E-03	2.78E-05	1.07E-01
rs7223136	cg19712600	ILMN_1747857	SMARCE1	2.12E-01	2.09E-01	3.57E-05	1.20E-01	1.52E-03	9.95E-02
rs3795310	cg15732724	ILMN_1802380	RERE	3.38E-01	2.87E-01	3.57E-05	1.84E-01	6.67E-04	1.26E-01
rs301806	cg16484858	ILMN_1802380	RERE	5.90E-01	4.11E-01	3.57E-05	3.05E-01	2.78E-05	1.52E-01

Osteoarthritis (CD4$^+$ T cells)

SNP	CpG	IlluminaID	Gene	Pval CIT	FDR CIT	FDR EaL	FDR EaMgvL	FDR MaLgvE	FDR LiEgvM
rs28466887	cg08116092	ILMN_1678490	RILPL2	5.46E-03	2.88E-01	4.50E-06	1.23E-01	2.69E-04	1.87E-01
rs28466887	cg08116092	ILMN_1668743	RILPL2	5.80E-03	2.90E-01	9.51E-04	1.33E-01	5.83E-05	1.87E-01
rs56328224	cg18753072	ILMN_2200636	KIAA1267	6.65E-03	2.90E-01	2.97E-03	1.23E-01	8.10E-05	1.87E-01
rs35524223	cg18753072	ILMN_2200636	KIAA1267	1.06E-02	2.91E-01	4.77E-03	1.23E-01	5.39E-05	1.87E-01
rs62061818	cg20059597	ILMN_2393693	LRRC37A4	3.03E-02	4.02E-01	4.50E-06	1.23E-01	1.63E-03	3.17E-01
rs1724409	cg16520312	ILMN_1784428	MGC57346	3.07E-02	4.07E-01	4.50E-06	2.39E-01	5.32E-06	3.17E-01
rs1724390	cg15633388	ILMN_1680353	NSF	3.35E-02	4.07E-01	1.60E-02	2.39E-01	5.32E-06	2.08E-01
rs62522556	cg02623114	ILMN_1744268	PLEC	3.50E-02	4.43E-01	4.50E-06	2.50E-01	1.02E-05	3.17E-01
rs62057151	cg18753072	ILMN_2200636	KIAA1267	3.38E-02	4.43E-01	1.13E-03	2.39E-01	5.32E-06	3.17E-01
rs2668668	cg04226788	ILMN_2393693	LRRC37A4	5.27E-02	4.43E-01	4.50E-06	2.39E-01	5.32E-06	3.17E-01
rs113093579	cg15633388	ILMN_1680353	NSF	5.48E-02	4.43E-01	1.33E-02	3.06E-01	5.32E-06	1.87E-01
rs62073157	cg16520312	ILMN_1784428	MGC57346	5.47E-02	4.43E-01	4.50E-06	3.06E-01	1.52E-05	3.17E-01
rs58579887	cg13389508	ILMN_1744268	PLEC	1.26E-01	4.93E-01	4.50E-06	4.43E-01	5.32E-06	3.17E-01
rs56406407	cg20059597	ILMN_2393693	LRRC37A4	8.58E-02	4.93E-01	4.50E-06	2.70E-01	6.12E-04	3.17E-01
rs56026524	cg04226788	ILMN_2393693	LRRC37A4	1.11E-01	4.93E-01	4.50E-06	4.29E-01	5.32E-06	3.17E-01
rs112836774	cg16520312	ILMN_1784428	MGC57346	7.32E-02	4.93E-01	4.50E-06	3.41E-01	5.32E-06	3.17E-01
rs62057151	cg18753072	ILMN_2393693	LRRC37A4	1.26E-01	4.93E-01	4.50E-06	2.39E-01	2.39E-02	3.17E-01
rs62073157	cg16520312	ILMN_2393693	LRRC37A4	9.84E-02	4.93E-01	4.50E-06	3.82E-01	5.72E-04	3.17E-01
rs2696559	cg11117266	ILMN_2393693	LRRC37A4	9.26E-02	4.93E-01	4.50E-06	2.39E-01	2.65E-02	3.17E-01
rs1724390	cg15633388	ILMN_2330845	NSF	1.11E-01	4.93E-01	3.30E-02	4.29E-01	5.32E-06	3.06E-01
rs113093579	cg15633388	ILMN_2330845	NSF	1.47E-01	5.80E-01	3.30E-02	4.81E-01	5.32E-06	3.17E-01
rs2668665	cg07817266	ILMN_2393693	LRRC37A4	1.58E-01	5.80E-01	4.50E-06	3.82E-01	4.62E-03	3.17E-01
rs56328224	cg18753072	ILMN_2393693	LRRC37A4	1.63E-01	5.82E-01	4.50E-06	3.41E-01	7.14E-02	3.17E-01
rs6992333	cg07531549	ILMN_1744268	PLEC	1.88E-01	6.23E-01	4.50E-06	4.85E-01	5.32E-06	3.17E-01
rs2668668	cg15633388	ILMN_1680353	NSF	1.69E-01	6.23E-01	4.73E-03	4.85E-01	5.32E-06	3.17E-01
rs1724409	cg16520312	ILMN_2393693	LRRC37A4	2.01E-01	6.23E-01	4.50E-06	4.85E-01	2.52E-04	3.17E-01
rs62062803	cg07817266	ILMN_2393693	LRRC37A4	1.66E-01	6.23E-01	4.50E-06	4.81E-01	7.92E-04	3.17E-01
rs2696559	cg11117266	ILMN_1784428	MGC57346	2.01E-01	6.23E-01	4.50E-06	4.85E-01	4.73E-04	3.17E-01
rs35524223	cg18753072	ILMN_2393693	LRRC37A4	2.05E-01	6.23E-01	4.50E-06	4.29E-01	3.37E-02	3.17E-01
rs12543539	cg15847845	ILMN_2370872	GRINA	2.37E-01	6.24E-01	4.50E-06	4.85E-01	5.32E-06	3.17E-01
rs77807457	cg03238273	ILMN_1784428	MGC57346	4.91E-01	6.24E-01	4.50E-06	4.85E-01	5.32E-06	3.17E-01
rs2668668	cg15633388	ILMN_2330845	NSF	2.67E-01	6.24E-01	1.93E-02	4.85E-01	5.32E-06	3.17E-01
rs17688249	cg06537391	ILMN_2393693	LRRC37A4	5.59E-01	6.24E-01	4.50E-06	4.85E-01	5.32E-06	3.17E-01
rs6992333	cg14913216	ILMN_1744268	PLEC	2.11E-01	6.24E-01	4.50E-06	4.85E-01	5.32E-06	3.17E-01
rs56026524	cg04226788	ILMN_1784428	MGC57346	2.82E-01	6.24E-01	4.50E-06	4.85E-01	5.32E-06	3.17E-01
rs112836774	cg16520312	ILMN_2393693	LRRC37A4	4.06E-01	6.24E-01	4.50E-06	4.85E-01	2.84E-04	3.17E-01
rs112836774	cg11117266	ILMN_2393693	LRRC37A4	2.37E-01	6.24E-01	4.50E-06	4.43E-01	2.22E-02	3.17E-01
rs112836774	cg11117266	ILMN_1784428	MGC57346	2.60E-01	6.24E-01	4.50E-06	4.85E-01	2.03E-04	3.17E-01

rs112836774	cg11117266	ILMN_1709549	PLEKHM1	5.95E-01	6.24E-01	2.73E-01	2.12E-01	5.32E-06	3.43E-01
rs75743061	cg15295732	ILMN_1784428	MGC57346	4.09E-01	6.24E-01	4.50E-06	4.85E-01	5.32E-06	3.17E-01
rs56328224	cg18815117	ILMN_2393693	LRRC37A4	5.57E-01	6.24E-01	4.50E-06	4.85E-01	5.32E-06	3.17E-01
rs56328224	cg18815117	ILMN_1784428	MGC57346	2.12E-01	6.24E-01	4.50E-06	4.85E-01	5.32E-06	3.17E-01
rs56146262	cg01934064	ILMN_1784428	MGC57346	3.67E-01	6.24E-01	4.50E-06	4.85E-01	8.10E-05	3.17E-01
rs62056877	cg11117266	ILMN_2393693	LRRC37A4	2.26E-01	6.24E-01	4.50E-06	4.85E-01	6.31E-03	3.17E-01
rs62056877	cg11117266	ILMN_1784428	MGC57346	4.59E-01	6.24E-01	4.50E-06	4.85E-01	5.39E-05	3.17E-01
rs17760733	cg15295732	ILMN_1784428	MGC57346	5.09E-01	6.24E-01	4.50E-06	4.85E-01	5.32E-06	3.17E-01
rs17660865	cg18228076	ILMN_2393693	LRRC37A4	3.03E-01	6.24E-01	4.50E-06	4.85E-01	5.32E-06	3.17E-01
rs56026524	cg18815117	ILMN_2393693	LRRC37A4	3.59E-01	6.24E-01	4.50E-06	4.85E-01	5.32E-06	3.17E-01
rs56026524	cg18815117	ILMN_1784428	MGC57346	2.30E-01	6.24E-01	4.50E-06	4.85E-01	5.32E-06	3.17E-01
rs150592114	cg07368061	ILMN_2393693	LRRC37A4	5.00E-01	6.24E-01	4.50E-06	4.85E-01	5.32E-06	3.17E-01
rs62062803	cg07817266	ILMN_1784428	MGC57346	5.26E-01	6.24E-01	4.50E-06	4.85E-01	1.02E-05	3.17E-01
rs9891103	cg03238273	ILMN_2393693	LRRC37A4	5.27E-01	6.24E-01	4.50E-06	4.85E-01	5.32E-06	3.17E-01
rs9891103	cg03238273	ILMN_1784428	MGC57346	3.08E-01	6.24E-01	4.50E-06	4.85E-01	5.32E-06	3.17E-01
rs2668668	cg04226788	ILMN_1784428	MGC57346	4.85E-01	6.24E-01	4.50E-06	4.85E-01	5.32E-06	3.17E-01
rs56026524	cg17117718	ILMN_2393693	LRRC37A4	3.99E-01	6.24E-01	4.50E-06	4.85E-01	5.32E-06	3.17E-01
rs77804065	cg07368061	ILMN_2393693	LRRC37A4	5.29E-01	6.24E-01	4.50E-06	4.85E-01	5.32E-06	3.17E-01
rs1724390	cg00891649	ILMN_2393693	LRRC37A4	5.14E-01	6.24E-01	4.50E-06	4.85E-01	6.96E-03	3.17E-01
rs451737	cg04226788	ILMN_2393693	LRRC37A4	2.44E-01	6.24E-01	4.50E-06	4.85E-01	5.32E-06	3.17E-01
rs451737	cg04226788	ILMN_1784428	MGC57346	2.88E-01	6.24E-01	4.50E-06	4.85E-01	5.32E-06	3.17E-01
rs2950015	cg00891649	ILMN_2393693	LRRC37A4	3.95E-01	6.24E-01	4.50E-06	4.85E-01	2.84E-03	3.17E-01
rs3110331	cg15295732	ILMN_1784428	MGC57346	3.11E-01	6.24E-01	4.50E-06	4.85E-01	5.32E-06	3.17E-01
rs35524223	cg06537391	ILMN_2393693	LRRC37A4	5.87E-01	6.24E-01	4.50E-06	4.85E-01	5.32E-06	3.17E-01
rs2668665	cg07817266	ILMN_1784428	MGC57346	3.44E-01	6.24E-01	4.50E-06	4.85E-01	4.00E-05	3.17E-01
rs1819040	cg17117718	ILMN_2393693	LRRC37A4	3.36E-01	6.24E-01	4.50E-06	4.85E-01	5.32E-06	3.17E-01
rs1819040	cg17117718	ILMN_1784428	MGC57346	4.79E-01	6.24E-01	4.50E-06	4.85E-01	5.32E-06	3.17E-01
rs2532239	cg18815117	ILMN_2393693	LRRC37A4	2.48E-01	6.24E-01	4.50E-06	4.85E-01	5.32E-06	3.17E-01
rs2532239	cg18815117	ILMN_1784428	MGC57346	2.58E-01	6.24E-01	4.50E-06	4.85E-01	5.32E-06	3.17E-01
rs62522556	cg14598846	ILMN_2370872	GRINA	7.08E-01	6.38E-01	4.50E-06	4.85E-01	5.32E-06	3.17E-01
rs11136342	cg04892187	ILMN_2370872	GRINA	9.13E-01	6.38E-01	4.50E-06	4.85E-01	5.32E-06	3.17E-01
rs6992333	cg20784950	ILMN_2370872	GRINA	7.01E-01	6.38E-01	4.50E-06	4.85E-01	5.32E-06	3.82E-01
rs11136342	cg24044478	ILMN_2370872	GRINA	8.09E-01	6.38E-01	4.50E-06	4.85E-01	5.32E-06	3.17E-01
rs641760	cg01030110	ILMN_3245559	CDK2AP1	6.85E-01	6.38E-01	4.50E-06	4.85E-01	5.32E-06	3.17E-01
rs6992333	cg24891660	ILMN_2370872	GRINA	6.00E-01	6.38E-01	4.50E-06	4.85E-01	5.32E-06	3.17E-01
rs6992333	cg04757492	ILMN_2370872	GRINA	8.59E-01	6.38E-01	4.50E-06	4.85E-01	5.32E-06	3.43E-01
rs6992333	cg25475366	ILMN_2370872	GRINA	7.79E-01	6.38E-01	4.50E-06	4.85E-01	5.32E-06	3.60E-01
rs7819099	cg22822867	ILMN_2370872	GRINA	7.71E-01	6.38E-01	4.50E-06	4.85E-01	5.32E-06	3.17E-01
rs17573447	cg07368061	ILMN_1784428	MGC57346	7.15E-01	6.38E-01	4.50E-06	4.85E-01	5.32E-06	3.83E-01
rs7819099	cg21900799	ILMN_2370872	GRINA	7.04E-01	6.38E-01	4.50E-06	4.85E-01	5.32E-06	3.17E-01
rs6992333	cg08090367	ILMN_2370872	GRINA	8.35E-01	6.38E-01	4.50E-06	4.85E-01	5.32E-06	4.29E-01
rs56356641	cg00891649	ILMN_2393693	LRRC37A4	6.17E-01	6.38E-01	4.50E-06	4.85E-01	1.26E-03	3.17E-01
rs77807457	cg03238273	ILMN_2393693	LRRC37A4	9.55E-01	6.38E-01	4.50E-06	4.85E-01	5.32E-06	3.43E-01
rs6992333	cg07531549	ILMN_2370872	GRINA	9.38E-01	6.38E-01	4.50E-06	4.85E-01	5.32E-06	3.17E-01
rs7003580	cg08161931	ILMN_2370872	GRINA	8.84E-01	6.38E-01	4.50E-06	4.85E-01	5.32E-06	3.43E-01
rs1724409	cg07817266	ILMN_2393693	LRRC37A4	9.56E-01	6.38E-01	4.50E-06	4.85E-01	5.32E-06	3.17E-01
rs1724409	cg07817266	ILMN_1784428	MGC57346	6.05E-01	6.38E-01	4.50E-06	4.85E-01	5.32E-06	3.17E-01
rs17688249	cg06537391	ILMN_1784428	MGC57346	7.69E-01	6.38E-01	4.50E-06	4.85E-01	5.32E-06	4.05E-01
rs6992333	cg14913216	ILMN_2370872	GRINA	6.70E-01	6.38E-01	4.50E-06	4.85E-01	5.32E-06	3.17E-01
rs4627402	cg18878992	ILMN_2393693	LRRC37A4	7.48E-01	6.38E-01	4.50E-06	4.85E-01	5.32E-06	3.97E-01
rs56026524	cg18228076	ILMN_2393693	LRRC37A4	6.16E-01	6.38E-01	4.50E-06	4.85E-01	5.32E-06	3.17E-01
rs56026524	cg18228076	ILMN_1784428	MGC57346	8.53E-01	6.38E-01	4.50E-06	4.85E-01	5.32E-06	3.43E-01
rs56406407	cg20059597	ILMN_1784428	MGC57346	8.12E-01	6.38E-01	4.50E-06	4.85E-01	5.32E-06	4.24E-01
rs62056931	cg01882395	ILMN_2393693	LRRC37A4	9.29E-01	6.38E-01	4.50E-06	4.85E-01	5.32E-06	3.17E-01
rs62056931	cg01882395	ILMN_1784428	MGC57346	9.34E-01	6.38E-01	4.50E-06	4.85E-01	5.32E-06	3.17E-01
rs56328224	cg17117718	ILMN_2393693	LRRC37A4	8.19E-01	6.38E-01	4.50E-06	4.85E-01	5.32E-06	3.17E-01
rs56328224	cg17117718	ILMN_1784428	MGC57346	6.07E-01	6.38E-01	4.50E-06	4.85E-01	5.32E-06	3.17E-01
rs62054760	cg09793084	ILMN_2393693	LRRC37A4	9.34E-01	6.38E-01	4.50E-06	4.85E-01	5.32E-06	3.43E-01
rs62054760	cg09793084	ILMN_1784428	MGC57346	9.18E-01	6.38E-01	4.50E-06	4.85E-01	5.32E-06	3.43E-01
rs17576954	cg01934064	ILMN_1784428	MGC57346	7.18E-01	6.38E-01	4.50E-06	4.85E-01	1.02E-05	3.17E-01
rs17576954	cg01934064	ILMN_2393693	LRRC37A4	9.42E-01	6.38E-01	4.50E-06	3.57E-01	5.32E-06	4.76E-01
rs75743061	cg15295732	ILMN_2393693	LRRC37A4	8.30E-01	6.38E-01	4.50E-06	4.85E-01	5.32E-06	4.29E-01
rs56146262	cg01934064	ILMN_2393693	LRRC37A4	9.56E-01	6.38E-01	4.50E-06	3.06E-01	1.02E-05	4.79E-01
rs62056877	cg11117266	ILMN_1709549	PLEKHM1	6.85E-01	6.38E-01	2.73E-01	2.12E-01	5.32E-06	3.82E-01
rs17760733	cg15295732	ILMN_2393693	LRRC37A4	8.99E-01	6.38E-01	4.50E-06	4.43E-01	5.32E-06	4.59E-01
rs17660865	cg18228076	ILMN_1784428	MGC57346	6.40E-01	6.38E-01	4.50E-06	4.85E-01	5.32E-06	3.17E-01

rs62061820	cg18878992	ILMN_2393693	LRRC37A4	7.19E-01	6.38E-01	4.50E-06	4.85E-01	5.32E-06	3.17E-01
rs56328224	cg09793084	ILMN_2393693	LRRC37A4	8.89E-01	6.38E-01	4.50E-06	4.85E-01	5.32E-06	3.43E-01
rs56328224	cg09793084	ILMN_1784428	MGC57346	8.20E-01	6.38E-01	4.50E-06	4.85E-01	5.32E-06	3.43E-01
rs150592114	cg07368061	ILMN_1784428	MGC57346	6.97E-01	6.38E-01	4.50E-06	4.85E-01	5.32E-06	3.82E-01
rs56026524	cg17117718	ILMN_1784428	MGC57346	6.27E-01	6.38E-01	4.50E-06	4.85E-01	5.32E-06	3.17E-01
rs1060105	cg01030110	ILMN_3245559	CDK2AP1	6.71E-01	6.38E-01	4.50E-06	4.85E-01	5.32E-06	3.17E-01
rs77804065	cg07368061	ILMN_1784428	MGC57346	9.34E-01	6.38E-01	4.50E-06	4.85E-01	5.32E-06	3.17E-01
rs9891103	cg01882395	ILMN_2393693	LRRC37A4	7.03E-01	6.38E-01	4.50E-06	4.85E-01	5.32E-06	3.17E-01
rs9891103	cg01882395	ILMN_1784428	MGC57346	6.48E-01	6.38E-01	4.50E-06	4.85E-01	5.32E-06	3.17E-01
rs2696559	cg11117266	ILMN_1709549	PLEKHM1	6.35E-01	6.38E-01	2.91E-01	2.12E-01	5.32E-06	3.60E-01
rs3110331	cg15295732	ILMN_2393693	LRRC37A4	6.40E-01	6.38E-01	4.50E-06	4.85E-01	5.32E-06	3.60E-01
rs35524223	cg09793084	ILMN_2393693	LRRC37A4	9.13E-01	6.38E-01	4.50E-06	4.85E-01	5.32E-06	3.19E-01
rs35524223	cg09793084	ILMN_1784428	MGC57346	8.63E-01	6.38E-01	4.50E-06	4.85E-01	5.32E-06	3.17E-01
rs35524223	cg06537391	ILMN_1784428	MGC57346	8.72E-01	6.38E-01	4.50E-06	4.85E-01	5.32E-06	3.43E-01
rs56328224	cg18878992	ILMN_2393693	LRRC37A4	7.10E-01	6.38E-01	4.50E-06	4.85E-01	5.32E-06	3.83E-01
rs56328224	cg18228076	ILMN_2393693	LRRC37A4	8.71E-01	6.38E-01	4.50E-06	4.85E-01	5.32E-06	3.43E-01
rs56328224	cg18228076	ILMN_1784428	MGC57346	8.50E-01	6.38E-01	4.50E-06	4.85E-01	5.32E-06	3.43E-01
rs62061818	cg20059597	ILMN_1784428	MGC57346	6.90E-01	6.38E-01	4.50E-06	4.85E-01	5.32E-06	3.82E-01
rs11136342	cg04892187	ILMN_1744268	PLEC	9.67E-01	6.51E-01	4.50E-06	4.87E-01	5.32E-06	3.19E-01

Results from Causal Inference Testing at all Risk cis-meQTL/eQTM sites in B cells

Rheumatoid arthritis (B cells)									
SNP	**CpG**	**IlluminaID**	**Gene**	**Pval CIT**	**FDR CIT**	**FDR EaL**	**FDR EaMgvL**	**FDR MaLgvE**	**FDR LiEgvM**
rs2210913	cg19602479	ILMN_1691693	FCRL3	4.69E-04	4.20E-02	8.33E-06	1.43E-05	3.44E-02	1.01E-02
rs2210913	cg19602479	ILMN_1699599	FCRL3	4.62E-04	4.20E-02	8.33E-06	1.43E-05	3.44E-02	1.01E-02
rs7522061	cg01045635	ILMN_1699599	FCRL3	5.49E-04	4.20E-02	8.33E-06	1.43E-05	3.44E-02	7.88E-03
rs2210913	cg19602479	ILMN_1797428	FCRL3	1.37E-03	4.82E-02	8.33E-06	1.43E-05	3.44E-02	1.80E-02
rs7522061	cg01045635	ILMN_1797428	FCRL3	2.11E-03	4.82E-02	8.33E-06	1.43E-05	3.84E-02	1.01E-02
rs7522061	cg01045635	ILMN_1691693	FCRL3	4.69E-03	6.64E-02	8.33E-06	1.43E-05	4.92E-02	1.80E-02
rs2210913	cg21721331	ILMN_1691693	FCRL3	5.34E-03	8.05E-02	8.33E-06	1.43E-05	4.92E-02	3.29E-02
rs2210913	cg21721331	ILMN_1699599	FCRL3	6.43E-03	8.21E-02	8.33E-06	1.43E-05	4.92E-02	3.45E-02
rs3093025	cg15222091	ILMN_1690907	CCR6	1.01E-02	9.66E-02	8.33E-06	1.42E-03	7.62E-02	2.07E-02
rs2061831	cg16429190	ILMN_1668277	BLK	5.20E-02	1.33E-01	8.33E-06	1.43E-05	1.21E-01	1.16E-01
rs2210913	cg21721331	ILMN_1797428	FCRL3	2.23E-02	1.33E-01	8.33E-06	1.43E-05	9.61E-02	6.90E-02
rs7522061	cg15602298	ILMN_1797428	FCRL3	3.14E-02	1.33E-01	8.33E-06	1.43E-05	1.12E-01	6.90E-02
rs3093025	cg16523158	ILMN_1690907	CCR6	6.63E-02	1.33E-01	8.33E-06	1.01E-03	1.31E-01	1.32E-01
rs3093025	cg19954286	ILMN_1690907	CCR6	2.58E-02	1.33E-01	8.33E-06	5.34E-04	1.01E-01	6.64E-02
rs3093025	cg21794222	ILMN_1690907	CCR6	5.14E-02	1.33E-01	8.33E-06	9.69E-04	1.31E-01	1.04E-01
rs3093025	cg05094429	ILMN_1690907	CCR6	3.47E-02	1.33E-01	8.33E-06	4.43E-03	1.13E-01	6.90E-02
rs1008723	cg14348996	ILMN_3245973	MSL1	6.92E-02	1.33E-01	3.55E-02	1.43E-05	3.44E-02	6.90E-02
rs11557466	cg12749226	ILMN_1662174	ORMDL3	2.49E-02	1.33E-01	8.33E-06	1.43E-05	1.01E-01	6.90E-02
rs9903250	cg18691862	ILMN_2300695	IKZF3	2.49E-02	1.33E-01	8.33E-06	1.43E-05	1.01E-01	7.01E-02
rs2061831	cg09528494	ILMN_1724762	XKR6	1.27E-01	1.75E-01	1.97E-04	1.43E-05	1.31E-01	1.32E-01
rs7522061	cg25259754	ILMN_1691693	FCRL3	1.41E-01	1.75E-01	8.33E-06	1.43E-05	1.46E-01	1.32E-01
rs7522061	cg25259754	ILMN_1797428	FCRL3	1.36E-01	1.75E-01	8.33E-06	1.43E-05	1.46E-01	1.32E-01
rs7522061	cg17134153	ILMN_1797428	FCRL3	8.18E-02	1.75E-01	8.33E-06	8.51E-04	1.37E-01	1.32E-01
rs2618476	cg04986849	ILMN_1668277	BLK	7.43E-02	1.75E-01	8.33E-06	1.05E-04	1.31E-01	1.32E-01
rs2061831	cg21497594	ILMN_1715680	NEIL2	9.87E-02	1.75E-01	2.58E-04	1.43E-05	1.31E-01	1.32E-01
rs2618476	cg01383082	ILMN_3248511	FAM167A	9.22E-02	1.75E-01	8.33E-06	4.60E-03	1.44E-01	1.16E-01
rs2618476	cg01383082	ILMN_1687213	FAM167A	1.34E-01	1.75E-01	8.33E-06	1.39E-03	1.46E-01	1.32E-01
rs2618476	cg01383082	ILMN_1668277	BLK	1.08E-01	1.75E-01	8.33E-06	1.43E-05	1.31E-01	1.32E-01
rs11557466	cg12749226	ILMN_1666206	GSDMB	1.11E-01	1.75E-01	8.33E-06	1.43E-05	1.31E-01	1.32E-01
rs9903250	cg13200575	ILMN_1662174	ORMDL3	1.48E-01	1.75E-01	8.33E-06	3.08E-03	1.46E-01	1.32E-01
rs9903250	cg18691862	ILMN_1707448	CDK12	1.55E-01	1.75E-01	7.99E-02	1.43E-05	8.11E-02	2.44E-02
rs9903250	cg18691862	ILMN_1662174	ORMDL3	1.76E-01	2.64E-01	8.33E-06	2.46E-03	1.49E-01	1.32E-01
rs7522061	cg25259754	ILMN_1699599	FCRL3	1.80E-01	2.83E-01	8.33E-06	1.43E-05	1.46E-01	1.60E-01
rs2618476	cg04986849	ILMN_3248511	FAM167A	1.80E-01	2.83E-01	8.33E-06	3.22E-02	1.49E-01	1.32E-01
rs1579258	cg00288844	ILMN_1771862	TXNDC11	1.85E-01	2.83E-01	8.33E-06	1.43E-05	1.46E-01	1.60E-01
rs2618476	cg01527115	ILMN_1687213	FAM167A	2.03E-01	3.04E-01	8.33E-06	5.23E-02	1.54E-01	1.32E-01
rs7522061	cg17134153	ILMN_1699599	FCRL3	2.31E-01	3.04E-01	8.33E-06	3.17E-04	1.46E-01	1.86E-01
rs7522061	cg15602298	ILMN_1691693	FCRL3	2.36E-01	3.04E-01	8.33E-06	1.43E-05	1.46E-01	1.86E-01

rs2618476	cg04986849	ILMN_1687213	FAM167A	2.20E-01	3.04E-01	8.33E-06	1.23E-02	1.59E-01	1.65E-01
rs2618476	cg01527115	ILMN_3248511	FAM167A	2.22E-01	3.04E-01	8.33E-06	5.54E-02	1.59E-01	1.32E-01
rs2061831	cg21497594	ILMN_1668277	BLK	2.50E-01	3.11E-01	8.33E-06	1.43E-05	1.46E-01	1.93E-01
rs2061831	cg21497594	ILMN_1687213	FAM167A	2.87E-01	3.25E-01	8.33E-06	7.17E-04	1.59E-01	2.01E-01
rs2618476	cg11944933	ILMN_1687213	FAM167A	2.76E-01	3.25E-01	8.33E-06	3.62E-03	1.59E-01	2.01E-01
rs2618476	cg11944933	ILMN_3248511	FAM167A	2.69E-01	3.25E-01	8.33E-06	4.80E-03	1.60E-01	2.01E-01
rs2061831	cg23507676	ILMN_1687213	FAM167A	2.91E-01	3.25E-01	8.33E-06	3.30E-02	1.73E-01	2.01E-01
rs2618476	cg01527115	ILMN_1668277	BLK	3.00E-01	3.25E-01	8.33E-06	3.45E-04	1.49E-01	2.06E-01
rs7522061	cg15602298	ILMN_1699599	FCRL3	3.24E-01	3.34E-01	8.33E-06	1.43E-05	1.53E-01	2.13E-01
rs1008723	cg14348996	ILMN_1666206	GSDMB	3.18E-01	3.34E-01	8.33E-06	1.43E-05	1.54E-01	2.13E-01
rs4728142	cg12816198	ILMN_1670576	IRF5	4.12E-01	3.75E-01	8.33E-06	1.43E-05	1.59E-01	2.57E-01
rs2061831	cg23507676	ILMN_3248511	FAM167A	3.34E-01	3.75E-01	8.33E-06	3.58E-02	1.83E-01	2.15E-01
rs2061831	cg16429190	ILMN_1687213	FAM167A	4.14E-01	3.81E-01	8.33E-06	1.43E-05	1.70E-01	2.57E-01
rs2061831	cg09528494	ILMN_1668277	BLK	4.70E-01	3.81E-01	8.33E-06	1.43E-05	1.60E-01	2.63E-01
rs2061831	cg21497594	ILMN_3248511	FAM167A	4.65E-01	3.81E-01	8.33E-06	5.78E-04	1.77E-01	2.63E-01
rs11557466	cg12655416	ILMN_1666206	GSDMB	4.64E-01	3.81E-01	8.33E-06	3.95E-04	1.73E-01	2.63E-01
rs2618473	cg03002059	ILMN_1687213	FAM167A	4.48E-01	3.81E-01	8.33E-06	2.03E-02	2.02E-01	2.63E-01
rs7522061	cg17134153	ILMN_1691693	FCRL3	4.73E-01	3.85E-01	8.33E-06	1.50E-04	1.65E-01	2.63E-01
rs3806624	cg21473142	ILMN_2200917	SLC4A7	6.18E-01	4.21E-01	1.45E-01	1.43E-05	3.10E-01	1.80E-02
rs1008723	cg14348996	ILMN_1662174	ORMDL3	6.06E-01	4.21E-01	8.33E-06	1.43E-05	1.96E-01	3.22E-01
rs2061831	cg23507676	ILMN_1668277	BLK	5.86E-01	4.21E-01	8.33E-06	1.28E-04	1.84E-01	3.17E-01
rs9916765	cg18711369	ILMN_1666206	GSDMB	5.82E-01	4.21E-01	8.33E-06	1.43E-05	1.83E-01	3.17E-01
rs2061831	cg09528494	ILMN_1687213	FAM167A	6.36E-01	4.64E-01	8.33E-06	8.11E-05	1.97E-01	3.33E-01
rs9916765	cg18711369	ILMN_1662174	ORMDL3	7.37E-01	5.34E-01	8.33E-06	1.43E-05	3.67E-01	2.63E-01
rs2061831	cg16429190	ILMN_3248511	FAM167A	8.01E-01	5.44E-01	8.33E-06	1.43E-05	2.23E-01	4.13E-01
rs2618473	cg03002059	ILMN_3248511	FAM167A	8.49E-01	5.80E-01	8.33E-06	1.03E-02	2.54E-01	4.31E-01
rs2061831	cg09528494	ILMN_3248511	FAM167A	9.01E-01	5.80E-01	8.33E-06	2.78E-05	2.36E-01	4.51E-01
rs2210913	cg19602479	ILMN_1691693	FCRL3	4.69E-04	4.20E-02	8.33E-06	1.43E-05	3.44E-02	1.01E-02

Multiple sclerosis (B cells)									
SNP	CpG	IlluminaID	Gene	Pval CIT	FDR CIT	FDR EaL	FDR EaMgvL	FDR MaLgvE	FDR LiEgvM
rs2015125	cg03983883	ILMN_2057981	FAM164A	2.57E-02	1.64E-01	1.79E-05	9.44E-02	2.63E-05	7.68E-02
rs2015125	cg03983883	ILMN_1789558	FAM164A	1.96E-02	1.64E-01	1.79E-05	9.44E-02	2.63E-05	7.68E-02
rs9804163	cg02586212	ILMN_1656011	RGS1	1.91E-02	1.64E-01	1.79E-05	9.44E-02	2.63E-05	7.68E-02
rs921985	cg25492364	ILMN_1811933	SHMT1	1.55E-02	1.64E-01	1.79E-05	9.44E-02	3.27E-03	7.68E-02
rs11557466	cg12749226	ILMN_1662174	ORMDL3	2.49E-02	1.64E-01	1.79E-05	9.44E-02	4.00E-05	7.77E-02
rs1441850	cg21140145	ILMN_2057981	FAM164A	1.56E-02	1.64E-01	1.79E-05	9.44E-02	2.63E-05	7.68E-02
rs1790123	cg01030110	ILMN_1812721	HIP1R	1.84E-02	1.64E-01	1.79E-05	9.44E-02	2.63E-05	7.68E-02
rs1441850	cg21140145	ILMN_1789558	FAM164A	3.65E-02	1.65E-01	1.79E-05	9.44E-02	2.63E-05	7.77E-02
rs7642303	cg12032497	ILMN_1747935	GOLGB1	4.00E-02	1.65E-01	2.26E-04	9.44E-02	2.63E-05	7.77E-02
rs7642303	cg12032497	ILMN_1708798	EAF2	5.79E-02	2.14E-01	1.79E-05	9.44E-02	4.00E-05	1.32E-01
rs1809476	cg07654569	ILMN_1789558	FAM164A	9.71E-02	2.53E-01	1.79E-05	1.39E-01	2.63E-05	1.32E-01
rs11557466	cg12749226	ILMN_1666206	GSDMB	1.11E-01	2.53E-01	1.79E-05	1.39E-01	2.63E-05	1.32E-01
rs1465697	cg07418126	ILMN_1682781	TEAD2	1.13E-01	2.61E-01	1.79E-05	1.39E-01	3.60E-03	1.42E-01
rs703842	cg00599273	ILMN_1767481	XRCC6BP1	1.30E-01	2.61E-01	1.79E-05	1.39E-01	4.00E-05	1.42E-01
rs1809476	cg07654569	ILMN_2057981	FAM164A	1.50E-01	2.61E-01	1.79E-05	1.39E-01	2.63E-05	1.42E-01
rs7642303	cg01951420	ILMN_1708798	EAF2	1.58E-01	2.61E-01	1.79E-05	1.39E-01	2.63E-05	1.42E-01
rs1441850	cg21140145	ILMN_1762262	PKIA	1.42E-01	2.61E-01	3.15E-03	1.39E-01	2.63E-05	1.42E-01
rs2015125	cg03983883	ILMN_1762262	PKIA	1.67E-01	2.67E-01	3.02E-03	1.39E-01	2.63E-05	1.46E-01
rs4676756	cg10605766	ILMN_1708798	EAF2	1.86E-01	2.78E-01	1.79E-05	1.39E-01	1.12E-02	1.52E-01
rs12983800	cg01007589	ILMN_1682781	TEAD2	2.97E-01	3.55E-01	1.79E-05	2.47E-01	2.63E-05	1.42E-01
rs2941522	cg11428475	ILMN_1662174	ORMDL3	3.06E-01	4.02E-01	1.79E-05	2.47E-01	9.53E-03	1.98E-01
rs1809476	cg05575058	ILMN_1789558	FAM164A	3.55E-01	4.15E-01	1.79E-05	2.75E-01	6.38E-04	1.93E-01
rs1465697	cg02189760	ILMN_1682781	TEAD2	4.36E-01	4.39E-01	1.79E-05	3.05E-01	4.00E-05	1.93E-01
rs1466526	cg09871101	ILMN_2057981	FAM164A	4.58E-01	4.39E-01	1.79E-05	3.05E-01	1.15E-04	1.93E-01
rs11557466	cg12655416	ILMN_1666206	GSDMB	4.64E-01	4.39E-01	1.79E-05	3.05E-01	2.22E-04	1.93E-01
rs1809476	cg05575058	ILMN_2057981	FAM164A	5.07E-01	4.56E-01	1.79E-05	3.22E-01	6.38E-04	1.98E-01
rs6438652	cg24574508	ILMN_1708798	EAF2	6.10E-01	4.87E-01	1.79E-05	3.59E-01	2.63E-05	1.99E-01
rs1466526	cg09871101	ILMN_1789558	FAM164A	6.79E-01	5.12E-01	1.79E-05	3.85E-01	4.00E-05	2.06E-01
rs4793836	cg10024583	ILMN_1703301	MRPL45P2	6.87E-01	5.76E-01	1.79E-05	3.54E-01	2.63E-05	3.43E-01
rs7642303	cg12032497	ILMN_2316104	IQCB1	8.23E-01	5.82E-01	8.33E-05	4.58E-01	2.63E-05	2.28E-01
rs703842	cg00599273	ILMN_1723846	METTL21B	9.76E-01	6.17E-01	1.79E-05	4.88E-01	4.00E-05	2.52E-01

rs12983800	cg01007589	ILMN_2375825	CD37	9.13E-01	6.17E-01	1.79E-05	4.88E-01	2.63E-05	2.73E-01
rs7642303	cg01951420	ILMN_2316104	IQCB1	9.48E-01	6.17E-01	8.33E-05	4.88E-01	2.63E-05	2.55E-01

Asthma (B cells)

SNP	CpG	IlluminaID	Gene	Pval CIT	FDR CIT	FDR EaL	FDR EaMgvL	FDR MaLgvE	FDR LiEgvM
rs903504	cg24910161	ILMN_1662174	ORMDL3	1.15E-02	7.06E-02	2.17E-05	5.18E-03	5.68E-03	6.04E-02
rs903504	cg24910161	ILMN_1666206	GSDMB	9.99E-03	7.06E-02	9.62E-05	3.61E-02	9.44E-04	6.04E-02
rs1495100	cg11817230	ILMN_1662174	ORMDL3	9.86E-02	7.64E-02	2.17E-05	9.02E-02	2.38E-05	6.48E-02
rs9893132	cg24211550	ILMN_1747857	SMARCE1	3.19E-02	7.64E-02	2.50E-04	4.85E-02	1.74E-04	6.04E-02
rs1008723	cg14348996	ILMN_3245973	MSL1	6.92E-02	7.64E-02	4.29E-02	4.85E-02	2.38E-05	5.05E-02
rs301805	cg04317648	ILMN_1802380	RERE	5.51E-02	7.64E-02	1.66E-03	6.83E-02	2.38E-05	6.04E-02
rs2517952	cg19758448	ILMN_1662174	ORMDL3	4.35E-02	7.64E-02	2.17E-05	6.83E-02	2.38E-05	6.04E-02
rs2517952	cg19758448	ILMN_1805636	PGAP3	1.57E-02	7.64E-02	2.17E-05	4.85E-02	2.38E-05	6.04E-02
rs35123741	cg26162295	ILMN_1662174	ORMDL3	5.38E-02	7.64E-02	2.17E-05	6.83E-02	2.38E-05	6.48E-02
rs35123741	cg26162295	ILMN_1666206	GSDMB	5.34E-02	7.64E-02	2.17E-05	6.83E-02	2.38E-05	6.37E-02
rs11557466	cg12749226	ILMN_1662174	ORMDL3	2.49E-02	7.64E-02	2.17E-05	4.85E-02	2.38E-05	6.04E-02
rs11557466	cg12749226	ILMN_1666206	GSDMB	1.11E-01	7.64E-02	2.17E-05	9.52E-02	2.38E-05	6.78E-02
rs12944882	cg02551532	ILMN_1666206	GSDMB	1.10E-01	7.64E-02	2.17E-05	8.90E-02	5.08E-03	8.84E-02
rs9903250	cg18691862	ILMN_2300695	IKZF3	2.49E-02	7.64E-02	4.17E-05	4.85E-02	2.38E-05	6.04E-02
rs4141183	cg12183861	ILMN_1703301	MRPL45P2	1.71E-02	7.64E-02	1.11E-04	4.81E-02	2.29E-04	6.04E-02
rs12944882	cg23202472	ILMN_1666206	GSDMB	2.95E-02	7.64E-02	2.17E-05	4.81E-02	2.16E-03	6.04E-02
rs35123741	cg19758448	ILMN_1662174	ORMDL3	5.52E-02	7.64E-02	2.17E-05	4.85E-02	5.68E-03	6.48E-02
rs35123741	cg19758448	ILMN_1805636	PGAP3	1.11E-01	7.64E-02	6.37E-02	2.91E-03	4.55E-05	1.07E-02
rs1495100	cg11817230	ILMN_1666206	GSDMB	1.54E-01	1.68E-01	8.00E-05	1.15E-01	2.38E-05	7.12E-02
rs6503526	cg24910161	ILMN_1666206	GSDMB	1.44E-01	1.68E-01	2.17E-05	1.13E-01	2.38E-05	7.74E-02
rs9903250	cg13200575	ILMN_1662174	ORMDL3	1.48E-01	1.68E-01	2.17E-05	9.02E-02	2.83E-03	1.13E-01
rs9903250	cg18691862	ILMN_1707448	CDK12	1.55E-01	1.68E-01	8.38E-02	3.38E-02	2.38E-05	6.04E-02
rs9903250	cg18691862	ILMN_1662174	ORMDL3	1.76E-01	2.19E-01	2.17E-05	1.13E-01	2.34E-03	1.18E-01
rs6503526	cg24910161	ILMN_1662174	ORMDL3	2.51E-01	2.76E-01	2.17E-05	1.79E-01	2.38E-05	1.18E-01
rs1495100	cg11817230	ILMN_1805636	PGAP3	2.86E-01	2.91E-01	2.17E-05	1.96E-01	2.38E-05	1.18E-01
rs1008723	cg14348996	ILMN_1666206	GSDMB	3.18E-01	3.02E-01	2.17E-05	2.01E-01	2.38E-05	1.26E-01
rs2941522	cg11428475	ILMN_1662174	ORMDL3	3.06E-01	3.02E-01	2.17E-05	1.98E-01	9.09E-03	1.72E-01
rs10158467	cg16600909	ILMN_2089875	TNFSF4	3.93E-01	3.35E-01	5.00E-04	2.39E-01	2.38E-05	1.26E-01
rs11557466	cg12655416	ILMN_1666206	GSDMB	4.64E-01	3.85E-01	2.17E-05	2.63E-01	3.60E-04	1.66E-01
rs9916765	cg18711369	ILMN_1666206	GSDMB	5.82E-01	4.31E-01	2.17E-05	3.13E-01	2.38E-05	1.72E-01
rs301805	cg14004768	ILMN_1802380	RERE	5.96E-01	4.32E-01	1.66E-03	3.13E-01	2.38E-05	1.72E-01
rs1008723	cg14348996	ILMN_1662174	ORMDL3	6.06E-01	4.40E-01	2.17E-05	3.13E-01	2.38E-05	1.85E-01
rs12944882	cg23202472	ILMN_1662174	ORMDL3	7.05E-01	4.78E-01	2.17E-05	3.52E-01	7.12E-04	1.94E-01
rs9916765	cg18711369	ILMN_1662174	ORMDL3	7.37E-01	5.35E-01	2.17E-05	2.63E-01	2.38E-05	3.70E-01

Osteoarthritis (B cells)

SNP	CpG	IlluminaID	Gene	Pval CIT	FDR CIT	FDR EaL	FDR EaMgvL	FDR MaLgvE	FDR LiEgvM
rs62056835	cg18228076	ILMN_2393693	LRRC37A4	5.99E-03	2.58E-01	4.03E-06	8.45E-02	5.62E-06	2.37E-01
rs2668668	cg18228076	ILMN_2393693	LRRC37A4	3.34E-03	2.58E-01	4.03E-06	4.48E-02	5.62E-06	2.37E-01
rs35524223	cg09793084	ILMN_2393693	LRRC37A4	8.31E-03	2.58E-01	4.03E-06	2.68E-02	8.64E-04	2.37E-01
rs1406947	cg16131304	ILMN_1745152	UQCC	9.17E-03	2.70E-01	1.20E-05	7.93E-02	5.97E-04	2.37E-01
rs4627402	cg09793084	ILMN_2393693	LRRC37A4	2.72E-02	2.70E-01	4.03E-06	4.34E-02	2.34E-04	2.37E-01
rs1790123	cg01030110	ILMN_1812721	HIP1R	1.84E-02	2.70E-01	4.03E-06	1.26E-01	5.62E-06	2.37E-01
rs1724390	cg09793084	ILMN_2393693	LRRC37A4	1.18E-02	2.70E-01	4.03E-06	4.34E-02	2.55E-04	2.37E-01
rs2316771	cg15633388	ILMN_2330845	NSF	2.02E-02	2.70E-01	1.19E-03	1.26E-01	5.62E-06	2.37E-01
rs2668668	cg18878992	ILMN_2393693	LRRC37A4	2.93E-02	2.78E-01	4.03E-06	5.36E-02	7.46E-04	2.37E-01
rs79730878	cg02322039	ILMN_2330845	NSF	2.76E-02	2.78E-01	1.24E-03	1.53E-01	4.51E-04	2.37E-01
rs1724390	cg15633388	ILMN_2330845	NSF	2.84E-02	2.78E-01	1.69E-03	9.03E-02	5.62E-06	2.37E-01
rs11136336	cg02331830	ILMN_1721411	PARP10	3.16E-02	3.47E-01	2.39E-03	1.53E-01	5.62E-06	2.37E-01
rs451737	cg18228076	ILMN_2393693	LRRC37A4	3.32E-02	3.47E-01	4.03E-06	1.53E-01	5.62E-06	2.37E-01
rs56328224	cg18878992	ILMN_2393693	LRRC37A4	4.06E-02	3.47E-01	4.03E-06	8.86E-02	1.21E-03	2.82E-01
rs62071573	cg03238273	ILMN_1784428	MGC57346	3.82E-02	3.47E-01	4.03E-06	1.53E-01	1.08E-05	2.82E-01
rs317685	cg12396344	ILMN_1815205	LYZ	6.24E-02	3.55E-01	4.03E-06	2.22E-01	5.62E-06	2.82E-01
rs2668668	cg15633388	ILMN_2330845	NSF	6.05E-02	3.55E-01	8.72E-04	2.22E-01	5.62E-06	2.37E-01
rs4627402	cg18878992	ILMN_2393693	LRRC37A4	8.09E-02	3.55E-01	4.03E-06	8.86E-02	1.43E-03	2.91E-01

rs62062786	cg09860564	ILMN_2330845	NSF	6.14E-02	3.55E-01	1.98E-03	2.22E-01	5.62E-06	2.37E-01
rs9891103	cg03238273	ILMN_1784428	MGC57346	6.92E-02	3.55E-01	4.03E-06	2.22E-01	1.08E-05	2.82E-01
rs2316771	cg15633388	ILMN_1680353	NSF	6.57E-02	3.55E-01	1.18E-03	2.22E-01	5.62E-06	2.37E-01
rs1724390	cg15633388	ILMN_1680353	NSF	5.06E-02	3.55E-01	1.24E-03	2.08E-01	5.62E-06	2.37E-01
rs12898997	cg14838715	ILMN_1754121	CSK	9.78E-02	3.60E-01	4.03E-06	2.41E-01	5.62E-06	2.82E-01
rs56328224	cg00916973	ILMN_2393693	LRRC37A4	9.42E-02	3.60E-01	4.03E-06	2.41E-01	1.08E-05	2.97E-01
rs77804065	cg03238273	ILMN_1784428	MGC57346	8.22E-02	3.60E-01	4.03E-06	2.41E-01	5.62E-06	2.82E-01
rs2668668	cg09860564	ILMN_1680353	NSF	1.07E-01	3.60E-01	4.25E-04	2.41E-01	5.62E-06	2.82E-01
rs62062786	cg09860564	ILMN_1680353	NSF	1.09E-01	3.60E-01	1.18E-03	8.86E-02	5.62E-06	2.97E-01
rs79730878	cg00916973	ILMN_2393693	LRRC37A4	1.08E-01	3.60E-01	4.03E-06	2.41E-01	5.62E-06	2.97E-01
rs62061820	cg06537391	ILMN_2393693	LRRC37A4	1.02E-01	3.60E-01	4.03E-06	2.37E-01	1.03E-03	2.97E-01
rs35524223	cg00916973	ILMN_2393693	LRRC37A4	8.46E-02	3.60E-01	4.03E-06	2.08E-01	7.73E-05	2.92E-01
rs56328224	cg00916973	ILMN_1784428	MGC57346	1.29E-01	3.61E-01	4.03E-06	2.58E-01	5.62E-06	2.82E-01
rs317685	cg19124816	ILMN_1815205	LYZ	1.13E-01	3.61E-01	4.03E-06	2.43E-01	5.62E-06	2.82E-01
rs317685	cg19124816	ILMN_1801387	YEATS4	1.18E-01	3.61E-01	2.80E-03	2.46E-01	5.62E-06	2.82E-01
rs17576954	cg01640727	ILMN_1784428	MGC57346	1.10E-01	3.61E-01	4.03E-06	2.41E-01	7.00E-04	2.97E-01
rs9891103	cg03238273	ILMN_2393693	LRRC37A4	1.13E-01	3.61E-01	4.03E-06	8.86E-02	4.97E-02	2.97E-01
rs79730878	cg00916973	ILMN_1784428	MGC57346	1.09E-01	3.61E-01	4.03E-06	2.41E-01	5.62E-06	2.82E-01
rs62061820	cg06537391	ILMN_1784428	MGC57346	1.24E-01	3.61E-01	4.03E-06	2.53E-01	5.62E-06	2.82E-01
rs62071573	cg03238273	ILMN_2393693	LRRC37A4	1.33E-01	3.61E-01	4.03E-06	4.48E-02	6.79E-02	2.82E-01
rs35524223	cg00916973	ILMN_1784428	MGC57346	1.10E-01	3.61E-01	4.03E-06	2.41E-01	5.62E-06	2.82E-01
rs77804065	cg03238273	ILMN_2393693	LRRC37A4	1.39E-01	3.86E-01	4.03E-06	1.53E-01	1.34E-02	3.19E-01
rs62055717	cg06537391	ILMN_1784428	MGC57346	1.93E-01	3.86E-01	4.03E-06	3.20E-01	5.62E-06	2.97E-01
rs138397226	cg09214591	ILMN_1738239	RBM6	1.82E-01	3.86E-01	4.03E-06	3.10E-01	5.62E-06	2.97E-01
rs451737	cg01527957	ILMN_2393693	LRRC37A4	1.88E-01	3.86E-01	4.03E-06	2.41E-01	1.59E-02	3.29E-01
rs111905143	cg04226788	ILMN_1784428	MGC57346	2.05E-01	3.86E-01	4.03E-06	3.20E-01	5.62E-06	3.00E-01
rs111541901	cg07817266	ILMN_2393693	LRRC37A4	1.37E-01	3.86E-01	4.03E-06	1.62E-01	5.72E-02	3.19E-01
rs56356641	cg20059597	ILMN_2393693	LRRC37A4	1.55E-01	3.86E-01	4.03E-06	2.90E-01	5.62E-06	3.19E-01
rs56356641	cg20059597	ILMN_1784428	MGC57346	2.24E-01	3.86E-01	4.03E-06	3.21E-01	5.62E-06	3.09E-01
rs113093579	cg01527957	ILMN_2393693	LRRC37A4	2.02E-01	3.86E-01	4.03E-06	2.41E-01	8.99E-03	3.29E-01
rs192252295	cg09860564	ILMN_1680353	NSF	2.37E-01	3.86E-01	1.62E-03	8.43E-02	5.62E-06	3.29E-01
rs192252295	cg09860564	ILMN_2330845	NSF	1.60E-01	3.86E-01	1.20E-03	2.90E-01	5.62E-06	2.82E-01
rs2532417	cg20059597	ILMN_2393693	LRRC37A4	2.25E-01	3.86E-01	4.03E-06	2.90E-01	7.73E-05	3.29E-01
rs2532417	cg20059597	ILMN_1784428	MGC57346	2.09E-01	3.86E-01	4.03E-06	3.20E-01	5.62E-06	3.22E-01
rs12898997	cg25999728	ILMN_1754121	CSK	2.82E-01	4.70E-01	4.03E-06	3.21E-01	5.62E-06	3.23E-01
rs317685	cg22375663	ILMN_1815205	LYZ	2.50E-01	4.70E-01	4.03E-06	3.21E-01	5.62E-06	3.22E-01
rs150592114	cg05301556	ILMN_2393693	LRRC37A4	2.60E-01	4.70E-01	4.03E-06	3.20E-01	5.39E-02	3.32E-01
rs150592114	cg05301556	ILMN_1784428	MGC57346	2.71E-01	4.70E-01	4.03E-06	3.21E-01	1.13E-03	3.29E-01
rs56328224	cg05301556	ILMN_1784428	MGC57346	2.81E-01	4.70E-01	4.03E-06	3.21E-01	1.54E-03	3.29E-01
rs2668665	cg07817266	ILMN_2393693	LRRC37A4	2.82E-01	4.70E-01	4.03E-06	8.86E-02	1.41E-01	3.23E-01
rs2668668	cg15633388	ILMN_1680353	NSF	2.84E-01	5.15E-01	4.25E-04	3.21E-01	5.62E-06	3.23E-01
rs62055717	cg06537391	ILMN_2393693	LRRC37A4	3.00E-01	5.15E-01	4.03E-06	3.21E-01	3.02E-04	3.29E-01
rs1724390	cg20059597	ILMN_1784428	MGC57346	3.23E-01	5.15E-01	4.03E-06	3.36E-01	5.62E-06	3.29E-01
rs2668668	cg09860564	ILMN_2330845	NSF	3.03E-01	5.15E-01	8.72E-04	3.21E-01	5.62E-06	3.22E-01
rs17573447	cg05301556	ILMN_2393693	LRRC37A4	3.12E-01	5.15E-01	4.03E-06	3.21E-01	2.79E-02	3.59E-01
rs12150048	cg05301556	ILMN_2393693	LRRC37A4	3.25E-01	5.15E-01	4.03E-06	3.21E-01	4.05E-02	3.59E-01
rs2696559	cg07298766	ILMN_1784428	MGC57346	3.06E-01	5.15E-01	4.03E-06	3.21E-01	1.25E-04	3.29E-01
rs439945	cg15411667	ILMN_2393693	LRRC37A4	3.49E-01	5.15E-01	4.03E-06	1.26E-01	1.73E-01	3.29E-01
rs17760733	cg09764761	ILMN_1784428	MGC57346	3.07E-01	5.15E-01	4.03E-06	3.21E-01	1.49E-04	3.29E-01
rs56328224	cg05301556	ILMN_2393693	LRRC37A4	2.96E-01	5.15E-01	4.03E-06	3.21E-01	8.90E-02	3.59E-01
rs56328224	cg04226788	ILMN_1784428	MGC57346	2.97E-01	5.15E-01	4.03E-06	3.21E-01	5.62E-06	3.22E-01
rs11136336	cg02331830	ILMN_2370872	GRINA	7.59E-01	5.42E-01	4.03E-06	3.84E-01	5.62E-06	4.79E-01
rs11136336	cg04255391	ILMN_2370872	GRINA	7.72E-01	5.42E-01	4.03E-06	3.45E-01	5.62E-06	4.79E-01
rs7003580	cg14598846	ILMN_2370872	GRINA	9.58E-01	5.42E-01	4.03E-06	4.91E-01	5.62E-06	4.55E-01
rs7003580	cg21900799	ILMN_2370872	GRINA	7.99E-01	5.42E-01	4.03E-06	3.43E-01	5.62E-06	4.79E-01
rs62056835	cg18228076	ILMN_1784428	MGC57346	9.09E-01	5.42E-01	4.03E-06	3.20E-01	5.62E-06	4.90E-01
rs11780978	cg15847845	ILMN_2370872	GRINA	9.28E-01	5.42E-01	4.03E-06	2.82E-01	5.62E-06	4.90E-01
rs1378942	cg14772590	ILMN_1754121	CSK	6.06E-01	5.42E-01	4.03E-06	4.78E-01	5.62E-06	3.59E-01
rs7819099	cg20784950	ILMN_2370872	GRINA	9.62E-01	5.42E-01	4.03E-06	2.41E-01	5.62E-06	4.90E-01
rs6992333	cg04757492	ILMN_2370872	GRINA	6.93E-01	5.42E-01	4.03E-06	4.91E-01	5.62E-06	4.55E-01
rs6992333	cg25475366	ILMN_2370872	GRINA	7.96E-01	5.42E-01	4.03E-06	3.45E-01	5.62E-06	4.79E-01
rs7819099	cg24891660	ILMN_2370872	GRINA	9.51E-01	5.42E-01	4.03E-06	4.91E-01	5.62E-06	4.55E-01
rs6992333	cg12520615	ILMN_2370872	GRINA	8.22E-01	5.42E-01	4.03E-06	3.21E-01	5.62E-06	4.80E-01
rs58879558	cg07298766	ILMN_1784428	MGC57346	4.00E-01	5.42E-01	4.03E-06	3.45E-01	1.49E-04	3.29E-01
rs1635298	cg07817266	ILMN_2393693	LRRC37A4	6.48E-01	5.42E-01	4.03E-06	4.91E-01	1.49E-04	3.72E-01
rs1635298	cg07817266	ILMN_1784428	MGC57346	9.48E-01	5.42E-01	4.03E-06	4.91E-01	5.62E-06	4.55E-01

rs4627402	cg18878992	ILMN_1784428	MGC57346	7.97E-01	5.42E-01	4.03E-06	4.91E-01	5.62E-06	4.55E-01
rs7819099	cg04892187	ILMN_2370872	GRINA	7.78E-01	5.42E-01	4.03E-06	3.84E-01	5.62E-06	4.79E-01
rs1724390	cg20059597	ILMN_2393693	LRRC37A4	3.59E-01	5.42E-01	4.03E-06	3.40E-01	5.85E-05	3.43E-01
rs80209523	cg04226788	ILMN_2393693	LRRC37A4	9.84E-01	5.42E-01	4.03E-06	1.53E-01	5.62E-06	4.96E-01
rs80209523	cg04226788	ILMN_1784428	MGC57346	4.56E-01	5.42E-01	4.03E-06	3.84E-01	5.62E-06	3.29E-01
rs56328224	cg17117718	ILMN_2393693	LRRC37A4	9.61E-01	5.42E-01	4.03E-06	2.22E-01	5.62E-06	4.90E-01
rs56328224	cg17117718	ILMN_1784428	MGC57346	8.11E-01	5.42E-01	4.03E-06	4.91E-01	5.62E-06	4.55E-01
rs17661141	cg07368061	ILMN_2393693	LRRC37A4	9.17E-01	5.42E-01	4.03E-06	2.90E-01	5.62E-06	4.90E-01
rs17661141	cg07368061	ILMN_1784428	MGC57346	8.72E-01	5.42E-01	4.03E-06	4.91E-01	5.62E-06	4.55E-01
rs17573447	cg05301556	ILMN_1784428	MGC57346	3.60E-01	5.42E-01	4.03E-06	3.40E-01	7.69E-04	3.29E-01
rs62055936	cg16520312	ILMN_2393693	LRRC37A4	9.47E-01	5.42E-01	4.03E-06	4.91E-01	7.73E-05	4.55E-01
rs62055936	cg16520312	ILMN_1784428	MGC57346	8.23E-01	5.42E-01	4.03E-06	4.91E-01	5.62E-06	4.55E-01
rs56328224	cg18815117	ILMN_2393693	LRRC37A4	7.24E-01	5.42E-01	4.03E-06	4.91E-01	5.62E-06	4.55E-01
rs56328224	cg18815117	ILMN_1784428	MGC57346	7.63E-01	5.42E-01	4.03E-06	4.91E-01	5.62E-06	4.55E-01
rs62062294	cg16520312	ILMN_2393693	LRRC37A4	5.18E-01	5.42E-01	4.03E-06	4.33E-01	1.49E-04	3.59E-01
rs62062294	cg16520312	ILMN_1784428	MGC57346	9.74E-01	5.42E-01	4.03E-06	4.91E-01	5.62E-06	4.55E-01
rs4627402	cg09793084	ILMN_1784428	MGC57346	7.12E-01	5.42E-01	4.03E-06	4.91E-01	5.62E-06	4.55E-01
rs1724390	cg09793084	ILMN_1784428	MGC57346	7.30E-01	5.42E-01	4.03E-06	4.91E-01	5.62E-06	4.55E-01
rs2532276	cg09764761	ILMN_1784428	MGC57346	3.83E-01	5.42E-01	4.03E-06	3.45E-01	1.08E-05	3.29E-01
rs56328224	cg03954353	ILMN_2393693	LRRC37A4	6.75E-01	5.42E-01	4.03E-06	4.00E-01	1.56E-03	4.55E-01
rs12150048	cg05301556	ILMN_1784428	MGC57346	5.37E-01	5.42E-01	4.03E-06	4.33E-01	1.73E-04	3.59E-01
rs56328224	cg01882395	ILMN_2393693	LRRC37A4	6.45E-01	5.42E-01	4.03E-06	4.33E-01	5.62E-06	4.55E-01
rs56026524	cg07368061	ILMN_2393693	LRRC37A4	7.17E-01	5.42E-01	4.03E-06	4.33E-01	5.62E-06	4.55E-01
rs56026524	cg07368061	ILMN_1784428	MGC57346	7.19E-01	5.42E-01	4.03E-06	4.91E-01	5.62E-06	3.83E-01
rs17576954	cg01640727	ILMN_2393693	LRRC37A4	6.80E-01	5.42E-01	4.03E-06	4.33E-01	6.27E-04	4.55E-01
rs111905143	cg04226788	ILMN_2393693	LRRC37A4	9.43E-01	5.42E-01	4.03E-06	2.41E-01	5.62E-06	4.90E-01
rs62057151	cg03954353	ILMN_2393693	LRRC37A4	8.49E-01	5.42E-01	4.03E-06	3.20E-01	1.16E-04	4.90E-01
rs62055717	cg11117266	ILMN_2393693	LRRC37A4	9.68E-01	5.42E-01	4.03E-06	4.91E-01	7.52E-03	4.55E-01
rs451737	cg23659289	ILMN_2393693	LRRC37A4	9.54E-01	5.42E-01	4.03E-06	4.91E-01	1.47E-03	4.55E-01
rs451737	cg23659289	ILMN_1784428	MGC57346	7.73E-01	5.42E-01	4.03E-06	4.91E-01	5.62E-06	4.55E-01
rs80209523	cg01882395	ILMN_2393693	LRRC37A4	7.45E-01	5.42E-01	4.03E-06	4.91E-01	5.62E-06	4.55E-01
rs111541901	cg07817266	ILMN_1784428	MGC57346	5.67E-01	5.42E-01	4.03E-06	4.33E-01	5.62E-06	3.59E-01
rs113093579	cg18815117	ILMN_2393693	LRRC37A4	8.17E-01	5.42E-01	4.03E-06	4.91E-01	5.62E-06	4.55E-01
rs113093579	cg18815117	ILMN_1784428	MGC57346	7.78E-01	5.42E-01	4.03E-06	3.85E-01	5.62E-06	4.79E-01
rs2532276	cg11117266	ILMN_2393693	LRRC37A4	8.86E-01	5.42E-01	4.03E-06	4.91E-01	3.86E-03	4.55E-01
rs56026524	cg17117718	ILMN_2393693	LRRC37A4	8.71E-01	5.42E-01	4.03E-06	3.21E-01	5.62E-06	4.90E-01
rs56026524	cg17117718	ILMN_1784428	MGC57346	7.19E-01	5.42E-01	4.03E-06	4.91E-01	5.62E-06	4.55E-01
rs451737	cg18228076	ILMN_1784428	MGC57346	8.93E-01	5.42E-01	4.03E-06	3.21E-01	5.62E-06	4.90E-01
rs2668668	cg18228076	ILMN_1784428	MGC57346	8.47E-01	5.42E-01	4.03E-06	3.40E-01	5.62E-06	4.90E-01
rs2668668	cg18878992	ILMN_1784428	MGC57346	9.66E-01	5.42E-01	4.03E-06	4.91E-01	5.62E-06	4.55E-01
rs35524223	cg09793084	ILMN_1784428	MGC57346	8.49E-01	5.42E-01	4.03E-06	4.91E-01	5.62E-06	4.55E-01
rs56328224	cg16520312	ILMN_2393693	LRRC37A4	9.53E-01	5.42E-01	4.03E-06	4.91E-01	1.65E-04	4.55E-01
rs56328224	cg16520312	ILMN_1784428	MGC57346	8.46E-01	5.42E-01	4.03E-06	4.91E-01	5.62E-06	4.55E-01
rs56328224	cg18878992	ILMN_1784428	MGC57346	9.13E-01	5.42E-01	4.03E-06	4.91E-01	5.62E-06	4.55E-01
rs1724390	cg07368061	ILMN_2393693	LRRC37A4	8.15E-01	5.42E-01	4.03E-06	3.40E-01	5.62E-06	4.80E-01
rs1724390	cg07368061	ILMN_1784428	MGC57346	6.47E-01	5.42E-01	4.03E-06	4.91E-01	5.62E-06	3.72E-01
rs56328224	cg04226788	ILMN_2393693	LRRC37A4	9.95E-01	5.42E-01	4.03E-06	8.86E-02	5.62E-06	4.98E-01
rs112480703	cg01882395	ILMN_2393693	LRRC37A4	5.49E-01	5.42E-01	4.03E-06	4.33E-01	1.85E-04	3.72E-01
rs35524223	cg03954353	ILMN_2393693	LRRC37A4	7.21E-01	5.42E-01	4.03E-06	3.84E-01	1.35E-03	4.55E-01
rs2668665	cg07817266	ILMN_1784428	MGC57346	3.57E-01	5.42E-01	4.03E-06	3.40E-01	9.69E-05	3.29E-01
rs62059008	cg23659289	ILMN_1784428	MGC57346	8.01E-01	5.42E-01	4.03E-06	4.91E-01	5.62E-06	4.55E-01
rs1819040	cg17117718	ILMN_2393693	LRRC37A4	9.14E-01	5.42E-01	4.03E-06	2.90E-01	5.62E-06	4.90E-01
rs1819040	cg17117718	ILMN_1784428	MGC57346	6.79E-01	5.42E-01	4.03E-06	4.91E-01	5.62E-06	4.55E-01
rs2668668	cg18815117	ILMN_2393693	LRRC37A4	6.10E-01	5.42E-01	4.03E-06	4.78E-01	5.62E-06	3.72E-01
rs2668668	cg18815117	ILMN_1784428	MGC57346	8.61E-01	5.42E-01	4.03E-06	4.91E-01	5.62E-06	4.55E-01