$CD19^+$ B cells in early rheumatoid arthritis

Nishanthi Thalayasingam
MA, BM BCh, MRCP

A thesis submitted in partial fulfilment of the requirements for the degree of Doctor of Philosophy

Translational and Clinical Research Institute
Faculty of Medical Sciences
Newcastle University

December 2019

Abstract

Background

Rheumatoid Arthritis (RA) is a genetically complex disease which causes inflammation primarily affecting synovial joints. The therapeutic success of B cell depletion in RA has confirmed the clinical relevance of B cells in disease, but their specific role in pathogenesis remains unclear.

Aims

1. Identify differences in the transcriptome of $CD19^{+}$ B cells between RA samples and disease controls
2. Carry out an expression quantitative trait locus (eQTL) analysis of confirmed RA genetic risk loci
3. Identify RA disease-specific eQTLs
4. Immunophenotype peripheral blood B cells

Method

242 patients were recruited, RNA and DNA was extracted for subsequent analyses and parallel flow cytometry data obtained.

Results

A list of differentially expressed genes, without multiple test correction (MTC), was identified between the transcriptome in RA and disease controls. Web-based analysis tools identified downregulation of B cell receptor (BCR) signalling and pathways involved in transcription and RNA processing in the RA group. A list of differentially expressed genes (with MTC) was identified when samples were divided based on chronological age and inflammatory status, not diagnosis. The eQTL analysis at RA risk loci identified 10 cis eQTLs in B cells and a further 21 potential RA-specific eQTLs which lay outside the known RA risk loci. The RA group had an increased frequency of $CD19^{+}CD24^{hi}CD38^{hi}$ cells, a postulated regulatory subset

Conclusions

Age and inflammatory status have a greater influence on the $CD19^{+}$ B cell transcriptome than RA in this cohort. The genetic component to gene expression is highlighted by the eQTL findings and will aid the prioritisation of genes for downstream functional work. The disease-specific eQTLs identified may indicate novel mechanisms of disease. The absence of a robust diagnostic gene signature between the disease groups examined may relate to the heterogeneity of the B cells examined, as highlighted by differences in the frequency of $CD19^{+}CD24^{hi}CD38^{hi}$ cells.

Acknowledgements

I am grateful to the patients and healthy volunteers who participated in this study, without whom this study would not be possible.

My supervisors Prof John Isaacs, Dr Arthur Pratt and Dr Amy Anderson have each provided me with guidance, encouragement and their endless patience over the years. Their support and optimism has been crucial in this research and the writing of this thesis.

The experimental teaching in the laboratory from Dr Amy Anderson has been invaluable to me; always with endless good humour and remarkable patience. I have appreciated the support of the group as whole throughout this project but must thank Lisa Tait for her reassurance and wise words.

The MSD data presented here was carried out by Dr Amy Anderson and Dr Arthur Pratt.

I would like to thank Andrew Skelton for the advice provided for the bioinformatics analyses of the gene expression data presented here and have valued our discussions of the topic. Andrew Skelton completed the eQTL analysis in parallel with that from $CD4^+$ T cells for the published work shown here and the interaction analysis.

I am grateful to Prof Anne Barton and her colleagues at the University of Manchester for the opportunity to collaborate with them and visit the laboratory. The microarray work and genotyping for this project was carried out by Dr Nisha Nair at the University of Manchester.

Funding for this work was provided by the Wellcome Trust through a clinical research fellowship.

Lastly, an utterly inadequate thank you to my parents, daughters and Ruairidh for their patience.

Declaration

I declare that the material submitted in this thesis is my own work, other than where otherwise stated.

Nishanthi Thalayasingam
December 2019

Table of Contents

List of figures

List of tables

List of abbreviations

ACPA	anti- citrullinated protein/peptide antibodies
ACR	American College of Rheumatology
AID	activation-induced cytidine deaminase
ANCA	Antineutrophil cytoplasmic antibodies
anti-CCP	anti-cyclic citrullinated peptide
APC	antigen presenting cell
ASC	antibody secreting cell
BAFF	B cell activating factor
BCR	B cell receptor
BLNK	B cell linker protein
Breg	regulatory B cell
BTK	Bruton's tyrosine kinase
CARD11	caspase recruitment domain-containing protein 11
CD	cluster differentiation
CCL	chemokine (C-C) ligand
CHD	coronary heart disease
CIN85	Cbl-interacting protein of 85kDa
CLP	common lymphoid progenitor
CRP	C reactive protein
CTD	connective tissue disease
CXCL	chemokine (C-X-C motif) ligand
CXCR	chemokine (C-X-C motif) receptor
DAG	diacylglycerol
DAS	disease activity score
DEG	differentially expressed gene
DMARD	disease modifying anti-rheumatic drug
DNA	deoxyribonucleic acid
eQTL	expression quantitative trait locus
EAE	autoimmune encephalomyelitis
EIF2	eukaryotic initiation factor 2
ELISA	enzyme-linked immunosorbent assay
ELS	ectopic lymphoid-like structure

ER	endoplasmic reticulum
ERK	extracellular signal regulated kinase
ES	enrichment score
ESR	erythrocyte sedimentation rate
estRA	established rheumatoid arthritis
EULAR	European League against Rheumatism
FACS	fluorescence activated cell sorting
FC	fold change
FCRL	Fc receptor like
FCS	foetal calf serum
FDCs	follicular dendritic cells
FDR	false discovery rate
FLOCK	Flow clustering without K
FLS	fibroblast like synoviocytes
FSC-A	forward scatter area
GC	germinal centre
GM-CSF	Granulocyte-macrophage colony-stimulating factor
GP	general practitioner
GSEA	gene set enrichment analysis
GWAS	genome wide association studies
GxD	genotype x disease
HC	healthy control
HLA	human leucocyte antigen
IBD	inflammatory bowel disease
IFN	interferon
Ig	immunoglobulin
IKK	inhibitor of NF-κB kinase
IL	interleukin
IP_3	inositol trisphosphate
IPA	ingenuity pathway analysis
ITAM	immunoreceptor tyrosine-based activation motifs
JAK	janus kinase
LD	linkage disequilibrium
MAF	minor allele frequency

MALT1	mucosa-associated lymphoid tissue lymphoma transition protein 1
MAPK	mitogen-activated protein kinase pathways
Mb	megabase
MHC	major histocompatibility complex
MSigDB	Molecular Signatures Database
MSD	MesoScale Discovery
MTC	multiple test correction
mTOR	mammalian target of rapamycin
NEAC	Northeast early arthritis clinic
NES	normalised enrichment score
NGS	next generation sequencing
NF-κB	nuclear factor-kappa B
NFAT	nuclear factor of activated T cells
NPV	negative predictive value
NRES	National Research Ethics Service
OA	osteoarthritis
PAD	peptidyl arginine deiminases
PBMC	peripheral blood mononuclear cell
PI3K	phosphoinositide 3-kinase
PIP_2	phosphotidylinositol-4,5-bisphosphate
PIP_3	phosphotidylinositol-3,4,5-triphosphate
PKCβ	protein kinase C beta
PLCγ	phospholipase C gamma
PPV	positive predictive value
RA	rheumatoid arthritis
RAG	recombination-activating genes
RANKL	receptor activator of nuclear factor kappa-B ligand
REC	research ethics committee
RF	rheumatoid factor
RIN	RNA integrity number
RNA	ribonucleic acid
SF	synovial fluid
SFK	Src family of protein kinases
SE	shared epitope

SJC	swollen joint count
SLC	surrogate light chain
SLE	systemic lupus erythematosus
SMR	standard mortality ratio
SNP	single nucleotide polymorphism
SSC-A	side scatter area
SSC-W	side scatter width
STAT	signal transducer and activator of transcription
Tfh	T follicular helper
TGFβ	transforming growth factor beta
Th	T helper
TJC	tender joint count
TNF	tumour necrosis factor
Treg	T regulatory cell
UA	undifferentiated arthritis
UPR	unfolded protein response

1. Introduction

1.1 Rheumatoid Arthritis

1.1.1 Rheumatoid Arthritis Overview

Rheumatoid Arthritis (RA) is a chronic autoimmune condition characterised by inflammation at the synovial joints. It classically presents as a symmetrical, small joint polyarthritis. It is a systemic disease which can cause anaemia, osteoporosis, fatigue and organ specific extra-articular features such as lung fibrosis. Unchecked chronic inflammation leads to progressive, irreversible joint damage and is associated with an increased mortality (standard mortality ratio (SMR) 1.27), primarily due to an increase in cardiovascular disease (SMR 1.49)[1].

RA affects 0.5-1% of adults in Western populations and there are estimated to be 26,000 new diagnoses in England annually[2]. Its prevalence is higher in females than males, with a male:female ratio of 1:3 typically quoted, but this influence decreases with age [3, 4]. There are certainly genetic predisposing factors to the condition as demonstrated by the 15-30% disease concordance in monozygotic twins and observations from genome wide association studies[5, 6]. Additional risk factors include smoking, obesity and lower levels of formal education as a measure of socioeconomic status [7, 8].

RA is a debilitating condition and around three quarters of patients are of working age at the time of diagnosis. A third of sufferers stop work within 2 years of disease onset[9]. In addition to the personal cost of the disease to the individual, the annual financial costs are estimated to be £560 million to the NHS and £1.8 billion in sick leave and work-related disability[2].

In autoimmune conditions, such as RA, the inflammatory immune processes which activate in response to pathogens or danger signals, react to self (autoantigen); perhaps due to a break down in self-tolerance. In RA the synovial joint in particular becomes the target. The perceived ‘danger’ or autoantigen which activates the inflammatory process and indeed which parts of the immune system are most important in disease pathogenesis, remain to be fully elucidated. Autoantibodies, namely Rheumatoid factor (RF) and anti- citrullinated protein antibodies (ACPAs), are detected in the circulation

of the majority of patients with RA and their presence is prominent in the classification criteria of the condition[10, 11].

The therapeutic management of RA aims to alleviate symptoms and prevent joint damage. The standard approach centres around the early introduction of immune-modulatory medications (disease modifying anti-rheumatic drugs, DMARDs) and regular monitoring of response to enable appropriate drug titration. In cases where this is insufficient to control disease then more focussed treatments are introduced including biologic therapies and janus kinase (JAK) inhibitors[9].

1.1.2 Current management of RA

The last two decades have seen substantial changes in the management strategies used in RA as there has been an increased appreciation that early diagnosis enables the prompt introduction of disease modifying agents and, as a consequence of good clinical control of disease, the potential to prevent or minimise disease related damage. The aim is for early referral of patients to secondary care enabling prompt assessment and diagnosis in order to begin and induce clinical remission.

However, despite an increased awareness of the genetic and environmental factors which predispose a person to developing RA, the diagnosis remains a clinical one. The increased emphasis on the need for diagnostic biomarkers comes from evidence that the early initiation of treatment regimes has been shown to reduce subsequent joint damage[12]. This has led to the concept of the 'window of opportunity' for commencing treatment: ideally within 12 weeks of the first symptoms[13, 14].

1.1.3 Treat to target

In 2010 an international task force provided a consensus on recommendations to improve the management of RA with the aim of treating to a given target. The primary goal was to maximise long term outcomes by treating the disease to a defined target by measuring disease activity on a regular basis and adjusting treatment to optimise outcomes. The recommendations have been incorporated widely into clinical practice. They rely on the definition of a treatment 'target' and adaptations made to the clinical pathway to achieve this. The primary target, based on expert opinion, is a state of

clinical remission, which was felt to be achievable in a significant proportion of patients, especially those with early RA. Clinical remission is defined as the "absence of the signs and symptoms of significant inflammatory disease activity" and low disease activity, particularly in those with established RA. Measures of disease activity should be carried out and documented on a regular basis, and treatments adjusted at least every 3 months to reach the target, with the aim of a sustained remission throughout the disease course (figure 1.1) [15].

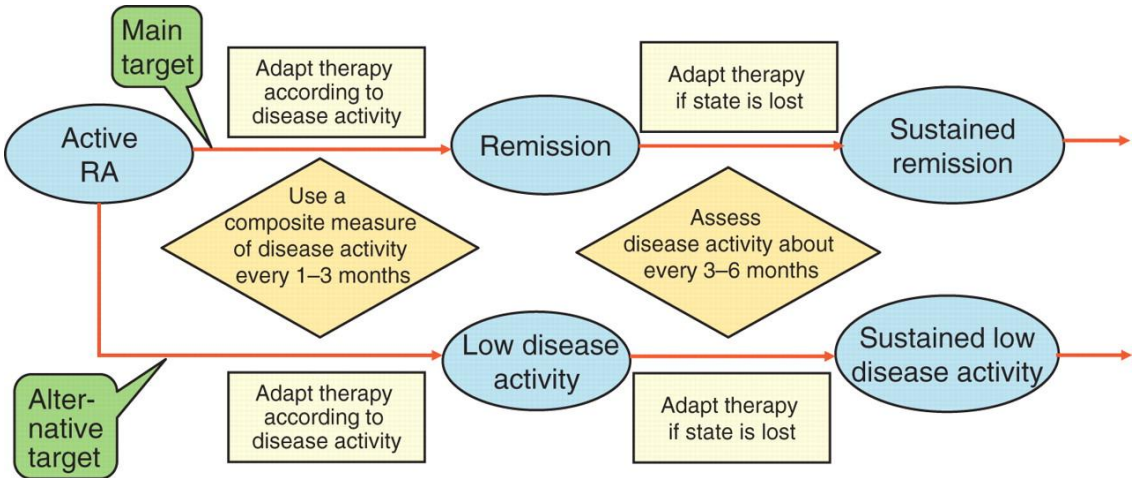


Figure 1.1 Algorithm for treating rheumatoid arthritis (RA) to target

The top thread indicates the main target: remission and sustained remission. The lower thread indicates the alternative target of sustained low disease activity in patients with long term disease. The approaches to attain and sustain the targets are essentially identical. Figure adapted from Smolen et al 2010[15].

1.1.4 Diagnosis and classification criteria

There is no specific test or diagnostic criteria for RA, however, classification criteria exist which do inform clinical practice.

In 2010 the American College of Rheumatology / European League Against Rheumatism (ACR/EULAR) classification criteria were introduced (table 1.1). The working group was focussed on developing classification criteria to facilitate the study of patients with early RA; classification criteria do not aim to capture all patients with RA. It replaced the 1987 ACR criteria which had been based on a cohort of patients with established disease, and was intended to differentiate between those with RA and those with other rheumatological conditions[11].

Target population Patients who: 1) have at least joint with definite clinical synovitis 2) with the synovitis not better explained by another disease	
Classification criteria for RA. Add scores for categories (A-D) a score of ≥6/10 is required for classification as definite RA	
A. Joint involvement	
1 large joint	0
2-10 large joints	1
1-3 small joints (with or without involvement of large joints)	2
4-10 small joints (with or without involvement of large joints)	3
>10 joints (at least one small joint)	5
B. Serology (at least 1 test is needed for classification)	
Negative RF and negative ACPA	0
Low positive RF or low positive ACPA	2
High positive RF or high positive ACPA	3
C. Acute phase reactants (at least 1 test is needed for classification)	
Normal CRP and normal ESR	0
Abnormal CRP or abnormal ESR	1
D. Duration of symptoms	
<6 weeks	0
≥6 weeks	1

Table 1.1 The 2010 Rheumatoid Arthritis classification criteria

Joint involvement refers to any swollen or tender joint on examination, distal interphalangeal joints, first carpometacarpal joints, and first metatarsophalangeal joints are excluded from assessment. "Large joints" refers to shoulders, elbows, hips, knees, and ankles. "Small joints" refers to the metacarpophalangeal joints, proximal interphalangeal joints, second through fifth metatarsophalangeal joints, thumb interphalangeal joints, and wrists. Duration of symptoms refers to patient self-reporting of the duration of signs or symptoms of synovitis in joints that are clinically involved at the time of assessment. Rheumatoid factor (RF); anti-citrullinated protein antibody (ACPA); C-reactive protein (CRP); erythrocyte sedimentation rate (ESR), Adapted from Aletaha et al 2010[11]

The diagnosis is based on the clinical opinion of the Rheumatologist using a combination of clinical features and autoantibody status. In practice, patients with a suspected inflammatory arthritis are referred for assessment and may be diagnosed with

definite RA, an alternative rheumatological condition or an undifferentiated arthritis (UA). Around 40% of patients with a new onset inflammatory arthritis have disease that cannot be classified at their first appointment and so are described as having an undifferentiated arthritis (UA)[16]. It is varyingly estimated that a third of these patients will develop RA but, due to the diagnostic uncertainty, their diagnosis is delayed, placing them at risk of irreversible joint damage[17]. Although aggressive treatment regimes offer the opportunity to reduce subsequent disability, if they are applied to all patients there is a risk of over-treatment of those with more benign disease and exposure to potential drug related toxicities.

1.2 Pathogenesis of RA

1.2.1 The synovial joint in RA

In the normal joint the synovial membrane is a thin layer lining the non-articular surfaces of the joint and is just a few cells thick. The resident cells are mesenchymal-derived, fibroblast like synoviocytes (FLS) and macrophages. In RA the synovium becomes infiltrated with immune cells. Leucocytes migrate to the inflamed joint in response to chemokines and this is further facilitated by changes in the synovial microvasculature, including the increased expression of adhesion molecules. The adaptive immune system is thought to be crucial to both disease initiation and persistence but the exact mechanisms of this are not yet known.

The synovium becomes thickened to form the pannus which is able to invade and destroy the adjacent articular cartilage and subchondral bone (figure 1.2) [18]. In the disease state the FLS display a different phenotype, becoming more resistant to apoptosis which is, in part, mediated by the upregulation of oncogenes such as p53[19]. The FLS and infiltrating immune cells release cytokines, chemokines and metalloproteinases which contribute to cartilage damage and disease persistence[20].

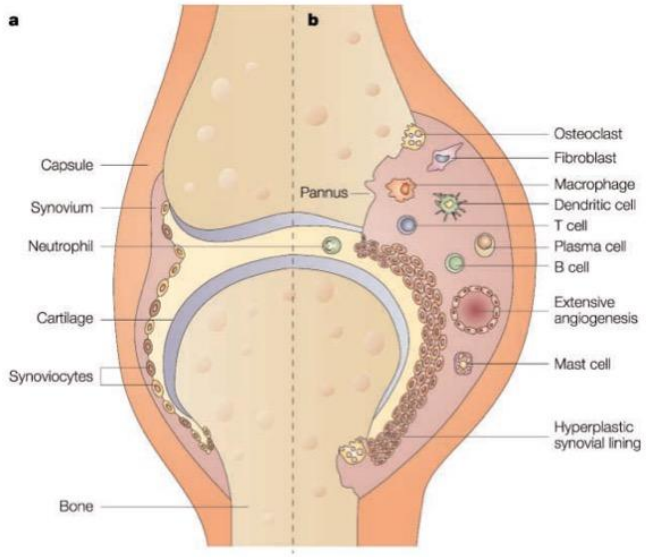


Figure 1.2 Schematic view of a normal synovial joint and in rheumatoid arthritis

a) The normal synovial joint with two bony ends each covered with an even layer of cartilage, separated by a joint space and surrounded by the synovial membrane and joint capsule. The synovial lining consists of a thin layer of synoviocytes. b) The inflamed joint seen in rheumatoid arthritis (RA) is depicted with evidence of synovial inflammation, inflammatory cell infiltrate, new vessel formation, cartilage loss and bony destruction. The synovial lining becomes hyperplastic and the membrane expands. The hallmark of RA is bone destruction which mostly starts at the cartilage-bone-synovial membrane junction, the destructive cellular element is the osteoclast. Enzymes released by neutrophils, chondrocytes and synoviocytes lead to cartilage degradation. Figure adapted from Smolen and Steiner 2003.[21].

The articular cartilage is limited in its ability to grow and repair. Thus, cartilage damage is irreversible and so increased stress and wear occurs at the bone surface. Bone erosions are a characteristic of RA which can be identified radiologically and underline the further structural damage occurring at the joint. The presence of bone erosions is related to the extent of synovitis and elevated inflammatory markers. Receptor activator of nuclear factor kappa-B ligand (RANKL) and the pro-inflammatory cytokines which are abundant in the inflamed synovium, including tumour necrosis factor (TNF) and interleukin-6 (IL-6), promote osteoclast differentiation which mediates bone resorption[22].

The chronic inflammation and tissue damage in RA is mediated by several cell subtypes including T cells, B cells, macrophages, osteoclasts and fibroblasts. No single cell type has been identified as primarily responsible for the initiation and maintenance of

pathogenic inflammation. The dominant pathogenic cell type(s) may differ between patients explaining the observed heterogeneity of disease.

1.2.2 T cells and the pathogenesis of RA

Rheumatoid Arthritis is traditionally thought of as primarily a T cell driven disease. Indeed, T cells are abundant in the synovium of RA patients and $CD4^+$ T cells predominate[23]. $CD4^+$ T cell activation is dependent on recognition by the T cell receptor of antigen fragments presented on the major histocompatibility complex class II (MHC II) molecule, the presence of co-stimulatory molecules (CD80/CD86) on the antigen presenting cell (APC) and the cytokine milieu. The established association between the HLA-DRB1 locus and RA is the cornerstone of the evidence for $CD4^+$ T cells in RA pathogenesis. However, the functional role of T cells is not fully understood. There is indirect evidence of their importance through the success of treatments such as Abatacept, a co-stimulation modulator, which prevents full T cell activation and has been shown to ameliorate disease[24].

There is now an increasing focus on pathogenic T helper type 17 (Th17) cells which can produce IL-17A, IL-22, IL-23 and TNF. These cytokines subsequently activate fibroblasts and chondrocytes implicating Th17 cells in the pathogenesis of RA and a range of autoimmune conditions. The pro-inflammatory milieu of the synovium promotes the differentiation of Th17 cells and suppresses that of regulatory T cells, shifting the balance towards an inflammatory phenotype[20].

1.2.3 B cells and the pathogenesis of RA

There has been an increased focus on the role of B cells in the pathogenesis of disease following the effectiveness of selective B cell depletion by the chimeric monoclonal antibody against CD20, rituximab[25]. B cells have a multi-faceted role in the adaptive immune system and their potential roles in disease pathogenesis include: self-reactive B cells initiating and perpetuating disease though the production of auto-antibodies, the release of pro-inflammatory cytokines (TNF, IL-1, IL-6 and interferon gamma (IFNγ)) and, as antigen-presenting cells (APCs), activating or amplifying autoreactive T cell responses. Ectopic germinal centres have been identified in the synovium in RA, providing a local source of autoantibody production [26]. The traditional focus has been

on the 'help' antigen-specific $CD4^+$ T cells provide to B cells, via CD40/CD40L interactions, to promote B cell receptor affinity maturation and class switch recombination, but such interactions are likely to be bidirectional.

1.3 B cells

1.3.1 B cell overview

B cells are continuously produced in the bone marrow from haematopoetic stem cells but the size of the peripheral pool remains relatively stable. There are two key requirements for survival: the formation of a functional B cell receptor and the presence of B cell activating factor (BAFF) which is essential for B cell maturation and survival in the periphery. Autoreactive B cells can be generated during the immunoglobulin variable gene rearrangements in the bone marrow and also by somatic hypermutations at the germinal centre stage. There are multiple methods of maintaining B cell tolerance which include deletion and anergy.

1.3.2 B cell receptor

The B cell receptor (BCR) is the membrane form of the antibody and is non-covalently linked to a signal transducing heterodimer, CD79A and CD79B, which both have immunoreceptor tyrosine-based activation motifs (ITAMs) in their intracellular tails. It carries out the dual function of initiating a signalling cascade in response to antigen and can also internalise and process it to deliver antigenic peptide to the cell to bind to MHC II molecules.

Engagement of foreign antigen and cross-linking of the receptor promotes ITAM phosphorylation and the overall effect is activation of the nuclear factor of activated T cells (NFAT), nuclear factor-kappa B (NF-κB) and mitogen-activated protein kinase (MAPK) pathways (figure 1.3)[27]. CD19 functions as a BCR co-receptor, decreasing the threshold for receptor activation. In the presence of antigen binding it is phosphorylated by the Src family of protein kinases (SFKs) associated with the BCR, recruiting and activating phosphoinositide 3-kinase (PI3K) and downstream Akt kinases.

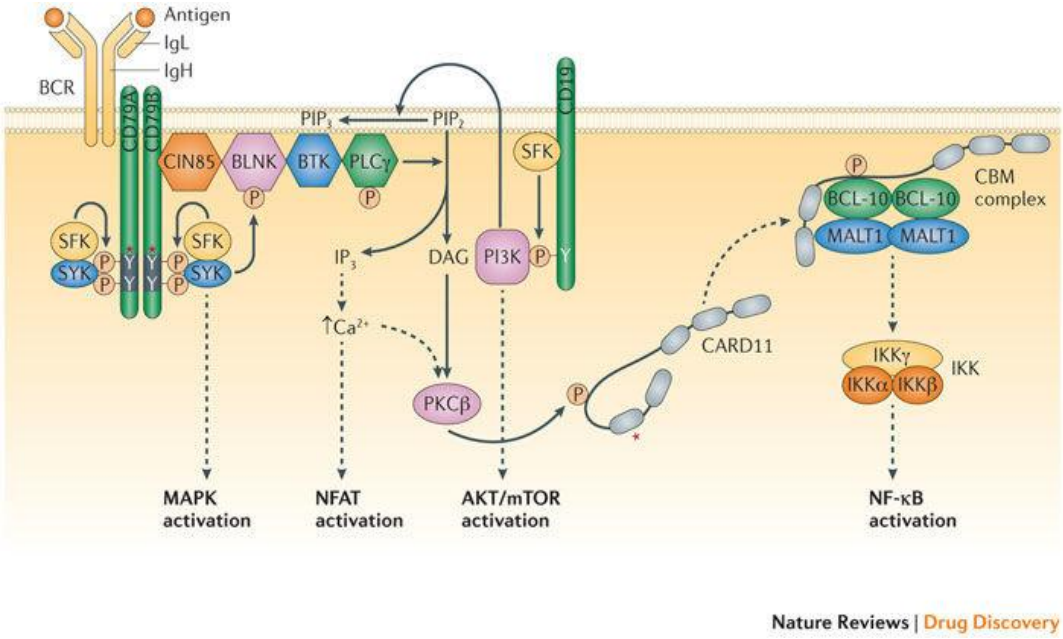


Figure 1.3 Overview of B cell receptor and CD19 signalling

Binding of antigen to the B cell receptor (BCR) leads to the engagement of several downstream pathways leading to the activation of MAPK, NFAT, AKT/mTOR and NF-κB. BLNK, B-cell linker protein; BTK, Bruton tyrosine kinase; CARD11, caspase recruitment domain-containing protein 11; CBM, CARD11–BCL-10–MALT1; CIN85, Cbl-interacting protein of 85 kDa; DAG, diacylglycerol; IKK, inhibitor of NF-κB kinase; IgH, immunoglobulin heavy chain; IgL, immunoglobulin light chain; IP_3, inositol trisphosphate; MALT1, mucosa-associated lymphoid tissue lymphoma translocation protein 1; MAPK, mitogen-activated protein kinase; mTOR, mammalian target of rapamycin; NF-κB, nuclear factor-κB; NFAT, nuclear factor of activated T cells; PI3K, phosphoinositide 3-kinase; PIP_2, phosphatidylinositol-4,5-bisphosphate; PIP_3, phosphatidylinositol-3,4,5-trisphosphate; PKCβ, protein kinase Cβ; PLCγ, phospholipase Cγ; SFK, SRC family kinase. Adapted from Young and Staudt 2013 [27]

In the absence of antigen, phosphorylation of signalling molecules can be detected in B cells, driven by tonic signalling by the BCR. This is essential for cell development and survival. In mice, the inducible deletion of the BCR leads to rapid mature B cell loss, with a half-life of 3-6 days, due to the loss of B cell signalling rather than simply the loss of expression of the BCR on the surface[28, 29]. Low level activation of the PI3K pathway has been shown to rescue mature B cells in which the BCR had been deleted and so to be critical to tonic signalling[30].

In the mature cell, as tonic signalling is required for survival in the periphery, the B cell is continuously selected to survive. In immature cells, tonic signalling promotes developmental progression while BCR ligation promotes developmental arrest and receptor editing.

1.3.3 Generation of a functional BCR

The generation of the antibody repertoire relies on random V(D)J recombination in early development and random somatic mutations in the periphery to generate diverse, high affinity antibodies. A consequence of this process is the generation of autoantibodies which are found in healthy individuals as well as in autoimmune conditions. In health, autoreactive B cells are regulated by processes to decrease their frequency, affinity or function. These processes take place centrally, in the bone marrow when the cells are not yet mature, and peripherally, for example in the spleen and lymph nodes, where the cells develop further and gain the capacity to be activated[31]. Defects in the tolerance process have been implicated in autoimmune conditions.

B cells develop from the common lymphoid progenitor in the bone marrow. The first stage of commitment to the B cell lineage is the pro B cell stage which begins with the ordered rearrangement of the immunoglobulin heavy chain segments, initiated by the expression of the recombination activating genes, *RAG1* and *RAG2*. The D (diverse) segments are first rearranged to the J (joining) segment, followed by the V (variable) segment to the DJ rearrangement creating a unique locus. If a functioning heavy chain is not produced at this stage then the cell may undergo repeated rearrangements. A complete pre-BCR is transiently expressed on the cell surface using a heterodimeric surrogate light chain (λ5 and Vpre-B) and the cell is now termed a large pre-B cell.

Spontaneous signalling through the pre-BCR downregulates *RAG* expression and stimulates the proliferative expansion of B cells with a functional heavy chain to produce numerous small pre-B cells. These re-express RAG proteins and so rearrange the light chain V and J segments. Once both chains have been successfully rearranged the cell is termed an immature B cell expressing a complete, functional IgM molecule which signals tonically. Defects in the receptor itself would prohibit tonic signalling and so the cell would not mature further at this stage. Tonic tyrosine phosphorylation of components of the BCR complex promotes PI3K signalling: suppressing RAG protein expression and promoting cell survival (figure 1.4).

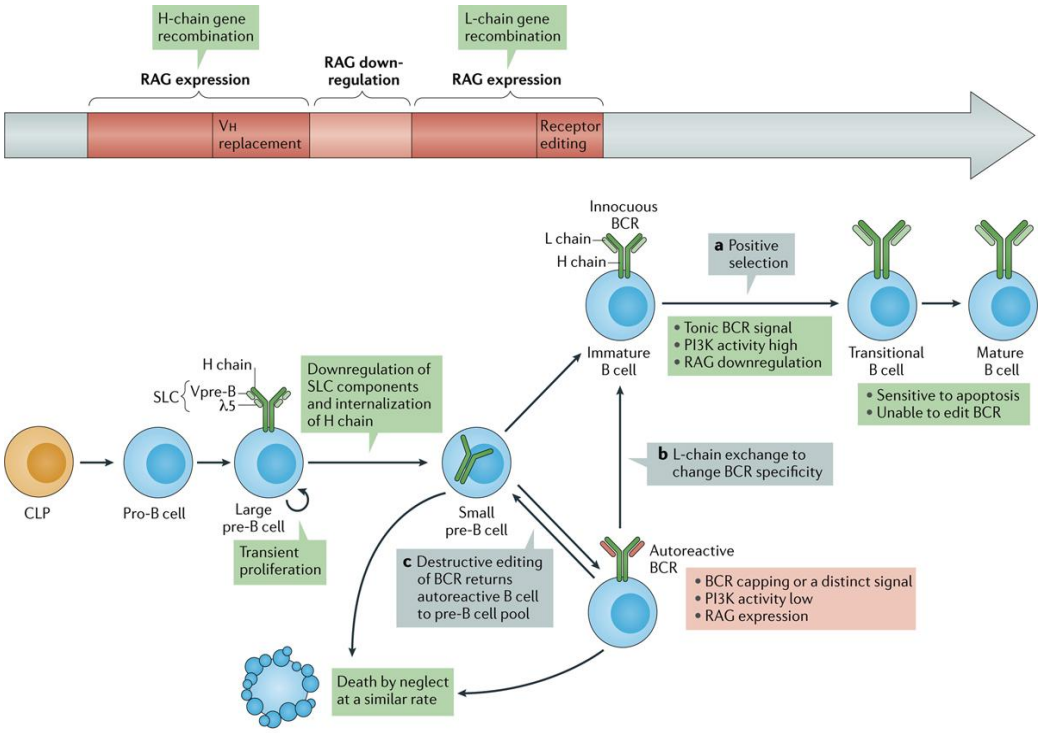


Figure 1.4 B cell development in the bone marrow

*Pro-B cells in the bone marrow are derived from common lymphoid progenitor (CLP) cells. They initiate heavy (H)-chain gene rearrangement through expression of recombination activating genes (*RAG1 *and* RAG2*; collectively known as* RAG*) and epigenetic modifications of the H-chain loci that promote accessibility. Productive H-chain assembly leads to the association of the IgM H-chain (μ-chain) protein with surrogate light-chain (SLC) components λ5 (known as IGLL1) and Vpre-B, and surface expression of the pre-B cell receptor (pre-BCR) in large pre-B cells. Spontaneous, antigen-independent triggering of the pre-BCR promotes progression to the large pre-B cell stage, with downregulation of* RAG *expression and transient proliferation. Differentiation to the small pre-B cell stage follows; at this stage SLC components are downregulated,* RAG *is re-expressed and* RAG *activity is redirected to the light (L)-chain genes. L-chains pair with H-chains and trigger tonic BCR signalling which promotes positive selection when the BCR is non autoreactive (a) or receptor editing when the BCR is autoreactive or if tonic signalling is impaired (b and c). Editing can lead to exchange of one functional L-chain for another, which can allow developmental progression (b) or to secondary rearrangements that prevent L-chain expression (c), such as out-of-frame joins that destroy the original L-chain gene but fail to replace it, which returns the cell to the pre-B cell compartment. Cells that go through positive selection enter the transitional B cell stage, at which stage they are sensitive to apoptosis. Adapted from Nemazee 2017 [31].*

It is at this stage that the cell is first tested for auto-reactivity, a process termed ‘central tolerance’. The fate for the B cell now depends on the signals from the functional receptor complex. If there is no strong reactivity to self antigens (the BCR is not ligated) then the B cell is selected and the cell migrates from the bone marrow, as an immature/transitional B cell to undergo further maturation steps in the secondary lymphoid organs.

If the new BCR reacts to an antigen in the bone marrow, and hence to self the possible outcomes include: (i) receptor editing, thought to be the dominant process, (ii) deletion, and (iii) anergy [32]. During receptor editing the cell undergoes further antibody gene recombination to generate a functional BCR of a different specificity, if unsuccessful it will undergo apoptosis. The cell signalling processes involved in this remain unclear and postulated models include: reduced cellular expression of the BCR on ligation, and therefore reduced tonic signalling, or the production of a distinct intracellular signal e.g by the activation of BLNK, B cell linker protein [31].

Central tolerance will only eliminate the antigens that are detectable in the bone marrow and so tolerance is incomplete, there must be an additional mechanism in the periphery. There is no option for further receptor editing and so any B cell that encounters a self antigen after exit from the bone marrow will undergo deletion, anergy or its activity may be limited by regulatory lymphocytes or the lack of a cognate autoreactive T cell[32].

The proportion of autoreactive B cells in healthy donors has been shown to be as much as 75% of recombinant antibodies cloned from early immature B cells with a pre-B surface phenotype and expressing functional light chain transcripts; this then reduces during development to just over 40% in newly emigrated cells in the blood, and reduces further to 20% in mature B cells. The majority of these early, autoreactive cells are polyreactive but the proportion that are polyreactive drops dramatically in the newly emigrated cells and mature naive cells, suggesting that polyreactive cells are ‘counter-selected’ and that there are checkpoints during development to remove autoreactive cells [33].

1.3.4 Germinal centres and the potential to develop autoreactive B cells

Germinal centres (GC) form around follicular dendritic cells (FDCs) within the peripheral lymphoid organs and it is here that B cells undergo the next stage of receptor diversification, class switching, and differentiate into high affinity plasma cells and memory B cells[34]. It is also a further stage at which autoreactive B cells may be generated[35]. Formation of the GCs is initiated when a naive B cell is activated by antigen and interacts with its corresponding antigen-specific T cell to become fully activated. A subset of these activated B cells moves to the medullary cords of the lymph node to differentiate into plasmablasts which secrete antibodies, although these are of lower affinity than plasma cells. They produce IgM antibody, are still dividing, are able to continue to respond to antigenm, and can present antigen on the MHC on the cell surface. Their lifespan is relatively short.

The B cells in the germinal centre rapidly proliferate in the dark zone, an area densely populated with B cell blasts, and undergo somatic hypermutation, whereby point mutations occur in the immunoglobulin variable region genes, altering the affinity of the antibody and generating cells with a range of affinities[36]. The affinity maturation is driven by competition for antigen presented on immune complexes by FDCs and the later receipt of help from the corresponding T cell[37]. T follicular helper (Tfh) cells are motile and continuously scan B cells in the GC and form the strongest interactions with cells with a high surface density of cognate peptide on MHC complexes, leading to their positive selection[38, 39]. Higher BCR affinity is associated with better antigen capture and so a higher density of peptide-MHC complexes. The positively selected B cells may recirculate to the dark zone and undergo successive rounds of mutation and selection.

In addition to the help from Tfh cells, it has been suggested that access to antigen on follicular dendritic cells is limited by antibodies secreted early in the GC reaction which, if bound to antigen presented by the FDC, shield it from the BCR so that only high affinity BCRs are able to compete to bind antigen. In this way, the selection process is accelerated and becomes progressively more stringent by ‘antigen masking’ and antibody related feedback [40] [41].

BCRs which are of lower reactivity, or those that are potentially autoreactive, are eliminated by apoptosis. Elimination of autoreactive GC B cells is probably most

effective if the self-antigen is expressed within or near the GC microenvironment. If there is low local expression of the self-antigen, or if its expression is tissue specific, then cells may be positively selected in error[42].

B cells from the germinal centre reaction exit as memory B cells or plasma cells. The precise mechanisms that determine cell fate remain unclear. In the case of memory B cells, no specific transcriptional regulator has been identified to date. The differentiation into plasma cells involves major changes in cell morphology and function to enable the production of a large amount of secreted antibody which can be up to 20% of all the protein secreted by the cell. In order for plasma cell differentiation to occur key transcription factors which determine B cell identity are down-regulated (for example *PAX5* and *BACH2*), and transcription factors crucial for the development into antibody secreting cells are upregulated (*IRF4*, *BLIMP1* and *XBP1*). *BLIMP1* is a transcriptional repressor in B cells, exclusively expressed in antibody secreting cells although at a lower level in plasmablasts than plasma cells, and it inhibits pathways required for B cell proliferation, class switching and affinity maturation. Plasma cells then home to the bone marrow, driven by *CXCL12* and its receptor *CXCR4*, where they are the source of the long lasting high affinity antibody; others can migrate to lymph nodes [43, 44].

1.3.5 B cell subsets in peripheral blood

B cells in the circulation comprise a variety of subsets including transitional, naive, mature, memory and plasmablasts, with differing states of maturation and activation. Plasma cells are rarely found in the peripheral blood, localising to the lymphoid tissues and bone marrow.

In mice, B cells are generated in the bone marrow, exiting as immature transitional B cells to complete maturation in the spleen. There is less evidence about the development of human B cells in the periphery but an analogous human transitional B cell subset has been defined as $CD19^{+}CD24^{hi}CD38^{hi}$ B cells[45]. Human transitional B cells make up just 5% of the peripheral blood B cell compartment, although this is known to be higher in cord blood and after bone marrow transplantation[46]. The biology and function of human transitional B cells is not clearly established and, although it was initially suggested that the transitional B cell subset could be subdivided into two major groups, 2

further subgroups have been identified which are functionally and phenotypically distinct[47].

Naive B cells, which express low levels of IgM, high levels of IgD and do not express CD27, make up approximately 50-60% of blood B cells in adults. Upon recognition of an antigen they may differentiate into early antibody producing cells or enter the germinal centre reaction[46, 48]

Memory B cells are traditionally defined by the presence of CD27 on their cell surface and have been broadly divided into IgD^+ and IgD^- class switched cells, which preferentially enter the germinal centres or differentiate into plasmablasts respectively upon stimulation[48]. However, there is certainly greater complexity and heterogeneity within this broad grouping with functional overlap between the two groups. In addition, smaller subsets have been defined within the $CD27^+$ B cell population[43, 49].

Plasmablasts found in blood and lymph nodes are defined phenotypically by the high expression of CD27 and CD38, they are produced in the extrafollicular foci and provide a short lived antibody response after the primary antigen contact. Like their fellow antibody producing plasma cells they express BLIMP1, albeit to a lesser extent, and they continue to express MHC II. It remains unclear whether these proliferating, short-lived cells are the precursors of the terminally differentiated plasma cells or if plasma cells can be derived from an earlier plasma cell-committed stage[44, 46].

1.3.6 Regulatory B cells

B cells are critical to generating a successful immune response but a small subset of B cells, regulatory B cells (Bregs), are able to suppress inflammation. There is a growing body of evidence that Bregs play an immunoregulatory role in autoimmune diseases and transplant tolerance, but progress has been hampered by the lack of consensus on their immunophenotype and the absence of a unique transcription factor to define this subset [50, 51].

In mouse models of experimental autoimmune encephalomyelitis and collagen induced arthritis the inflammatory response can be suppressed by the presence of B cells in an

effect that is IL-10 dependent [52, 53]. In addition, transforming growth factor β (TGFβ), IL-35, and direct cell to cell contact, with CD80/CD86 playing a critical role, have been shown to also contribute to Breg-dependent immunoregulation[54]. They suppress the proliferation of effector T cells (Th1 and Th17), modulate the T cell production of pro-inflammatory cytokines, promote the differentiation of regulatory T cells and have recently been shown to decrease the production of interferon alpha (IFNα) by plasmacytoid dendritic cells[55, 56].

The majority of the current work in humans relies on either isolating B cells, subjecting them to *in vitro* stimulation to identify IL-10 producing cells and subsequently phenotyping these cells; or focussing on a chosen immunophenotype. The use of intracellular staining to identify IL-10 further hampers functional assessments of this subset.

IL-10 producing cells are enriched within established B cell subsets, including the transitional, memory and plasmablast subsets, but within these groups they form only a minority of that population[57-59].

The most established Breg subsets in humans are $CD19^{+}CD24^{hi}CD27^{+}$ and $CD19^{+}CD24^{hi}CD38^{hi}$ B cells. As, *in vitro,* B cells require stimulation in order to identify IL-10 producing cells it has been suggested that they arise in response to inflammation to restrain the immune response; it is the environment rather than, for example, a specific Breg lineage factor which stimulates their differentiation. This theory is corroborated by work in mice where B cell differentiation and IL-10 production is promoted by the induction of arthritis, and also by alterations to the gut microbiome, which lead to inflammatory signals, and this effect is reduced by blocking the IL-6 and IL-1 receptors on B cells[60].

In autoimmune disease, Mauri's group have shown that $CD19^{+}CD24^{hi}CD38^{hi}$ B cells are reduced in number in established RA patients with active disease when compared to those with inactive disease and healthy individuals[61]. In addition, they have shown that $CD19^{+}CD24^{hi}CD38^{hi}$ B cells from healthy individuals are able to convert $CD4^{+}$ T cells into suppressive regulatory T cells (Tregs) and limit Th17 development, unlike

$CD19^{+}CD24^{hi}CD38^{hi}$ B cells from RA patients. This suggests that, in patients with RA, this subset is unable to prevent the differentiation of Th17 cells in Th17 polarising conditions or convert naive T cells to regulatory T cells and so are unable to prevent the development of autoreactive inflammation. In this study, the Bregs maintained their ability to inhibit Th1 cell differentiation. This presents an insight into potential mechanisms of disease in RA. Similarly, in systemic lupus erythematosus (SLE), $CD19^{+}CD24^{hi}CD38^{hi}$ B cells have been shown to be numerically deficient and functionally impaired. *In vitro* $CD19^{+}CD24^{hi}CD38^{hi}$ B cells from SLE patients produce less IL-10 in response to CD40 stimulation than those from healthy controls and are unable to suppress pro-inflammatory cytokine production by T cells[54]. In addition, in SLE the increased levels of IFNα promotes B cell differentiation into plasmablasts over regulatory B cells[55].

There is currently little work regarding the frequency or function of Bregs in early RA. A recent study looked at Bregs in this setting but focussed on $CD5^{+}CD1d^{+}$ B cells (a phenotype which is more established in mouse models of disease) and, combined with intracellular IL-10 staining, found them to be fewer in number in RA patients that healthy controls. This cell population was also shown to negatively correlate with disease activity score 28 (DAS-28)[62].

The importance of B cells in the maintenance of tolerance is highlighted in the transplant literature where the adoptive transfer of B cells from tolerant animals in a rat model of long term allograft tolerance resulted in the transfer of allograft tolerance[63]. These B cells may be described as displaying a regulatory phenotype.

Operationally tolerant renal transplant patients have been shown to have an increased frequency of transitional B cells and higher levels of this subset have been associated with protection from rejection, providing further evidence that this subset of immature cells is important for the maintenance of tolerance[51, 64]. Cherukuri *et al* have shown that the regulatory effect is based on the cytokine polarisation profile of this subset, as patients with graft rejection displayed a reduced IL-10/TNFα ratio compared to those with stable graft function (IL-10 expression levels alone were similar between the two groups)[65]. There is increasing evidence that the regulatory role of B cells is not solely

due to IL-10 production. This is highlighted by the ability of IL-21 dependent B cells from tolerant patients to inhibit the effector T cell response in a contact and granzyme B dependent pathway in operationally tolerant patients. The effect was independent of IL-10 and TGFβ [60].

1.4 B cells and Rheumatoid Arthritis

1.4.1 Overview of B cells and Rheumatoid Arthritis

A potential pathogenic role for B cells in RA has been long suggested following the identification of Rheumatoid Factor (RF) and later, additional autoantibodies, including anti-citrullinated protein/peptide antibodies (ACPA), in patients. The question has always been whether the humoral immune response is the driver of disease or a consequence of the overall breakdown in tolerance to self. B cells exert their effects through the production of autoantibodies, inflammatory and regulatory cytokines, acting as antigen presenting cells to provide co-stimulation to T cells and are also able to generate ectopic lymphoid structures.

1.4.2 Autoantibodies

The presence of autoantibodies prior to the onset of disease predicts the development of RA, with anti-cyclic citrullinated peptide (anti-CCP) having the highest predictive value[66]. Anti-CCP antibodies are detectable in the blood a median of 4.5 years before the clinical development of disease[67]. The presence of these auto antibodies is used to define the seropositive subset of RA, representing around two thirds of patients, and associated with more severe disease, extra-articular manifestations, joint destruction, genetic and environmental risk factors and a better response to B cell depletion with rituximab and co-stimulation blockade with abatacept [68, 69].

Anti-CCP2 is the current commercially available assay with a sensitivity of 70-75% and specificity of 95-99% in RA. In contrast to RF it is found in less than 2% of healthy individuals and just at low levels in other inflammatory conditions[69, 70]. In RA, RF and ACPA are the clinically used autoantibodies but there is evidence for antibodies specific for other post-translational modifications, such as carbamylation and acetylation, with cross-reactivity present between the modified antigens[71].

It is possible that the presence of ACPAs is simply a reflection of the dysregulation of the immune system but the data for a pathogenic role is increasing[69]. In addition to the clinical associations detailed above we know that citrullination, the post-translational conversion of arginine to citrulline by peptidyl arginine deiminases (PAD), increases the affinity of peptides for the MHC II molecules with the 'shared epitope', an established genetic risk factor for RA[72]. This change may lead to the increased presentation of potential autoantigens by APCs to autoreactive T cells, initiating disease. The citrullinated antigens: fibrinogen, vimentin, collagen type II and α-enolase are targets of ACPAs and can be found in the articular joint which may lead to the persistence of inflammation via, for example, immune complex formation [73].

1.4.3 Synovial infiltrate and ectopic lymphoid neogenesis

RA is characterised by synovial inflammation and subsequent joint damage but the histological features seen in the RA synovium are heterogeneous and are broadly defined as: pauci-immune (fibroblast type), diffuse (myeloid type) or follicular synovitis with ectopic lymphoid-like structures (ELS) (lymphoid type), which is found in 40% of RA patients [74]. In the lymphoid subset aggregates of T and B cells, often displaying T/B segregation, are found with GC reactions identified in approximately half of these[74, 75]. The presence of lymphoid neogenesis has not been shown to correlate with RA clinical phenotype, although it is associated with high inflammatory markers[76]. ELS development requires lymphotoxin, CXCL13, CCL19, CCL21, CXCL12, and is positively regulated by the inflammatory cytokines IL-17, IL-21, IL-22, IL-23, TNF and negatively regulated by IL-27[74].

The host factors which determine the development of ELS in the RA synovium are not yet known but it has been shown that these structures support the production of somatically hypermutated, class switched autoantibodies[26]. Synovial B cells may therefore, lead to joint damage by the production of local antibodies and inflammatory cytokines including IL-6, TNF, and RANKL inducing osteoclast activation, the release of destructive metalloproteinases and the further recruitment of immune cells[77].

Next generation sequencing (NGS) of the BCR heavy chain mRNA has identified multiple dominant clones in the synovial tissue of patients with active, seropositive RA in

early, disease modifying antirheumatic drug (DMARD)-naive, and established RA patients[78]. The presence of dominant BCR clones in the peripheral blood of patients is also predictive of the development of RA in an autoantibody positive at risk group. At the onset of RA these clones were no longer detectable peripherally but could be found in the synovial tissue, suggesting that the activated clones may migrate to the target tissue as disease develops – a hypothesis supported by the identification of the same clones at different joints[78, 79]. The phenotype and antigen specificity of the BCR clones were not determined as cell lysis was required for NGS.

Therefore, in a subset of RA, B cell clonal expansion occurs in the synovium providing a local source of class switched autoantibodies but the site of the initial generation of autoreactive B cells remains to be elucidated. Interestingly, ELS are also found in conditions which are not associated with B cell pathology; including spondylarthritis and osteoarthritis leading to the suggestion that lymphoid neogenesis is not disease specific and may be related to the degree rather than the type of inflammation[80].

1.4.4 B cell depletion in RA

Rituximab is a monoclonal antibody against CD20, licensed for the treatment of patients with active RA who have not responded or been able to tolerate TNF inhibitors or in whom their use in contra-indicated[25, 81]. The clinical effectiveness of rituximab in RA has provided an insight into disease pathogenesis but its exact mechanism of action, and therapeutic biomarkers of response, have yet to be established. A meta-analysis of four placebo-controlled randomised trials demonstrated an additional treatment benefit with rituximab in patients who were seropositive (as defined by the presence of RF and/or ACPA autoantibodies)[82]. The effect seen was modest but suggests that seropositive patients may have more B cell driven disease.

CD20 is believed to act as a calcium channel in the cell membrane and is involved in cell activation and growth. It is expressed on the surface of pre-B cells through the different stages of development to memory B cells, but it is not found on haematopoetic stem cells, pro-B cells or plasma cells (figure 1.5)[83]. Rituximab, therefore, almost completely depletes the peripheral B cell compartment by antibody-dependent cellular cytotoxicity,

complement-dependent cytotoxicity and apoptosis, with variable depletion in the synovium, lymphoid organs and bone marrow[83, 84].

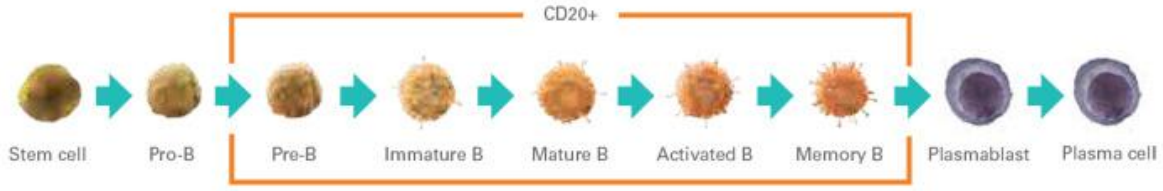


Figure 1.5 B cell subsets displaying CD20 are depleted by rituximab

CD20 is not found on stem cells, pro-B cells, plasmablasts or plasma cells[85]

Potential biomarkers have been identified ranging from genetic polymorphisms (including in FC-gamma receptor type IIIA and a promoter polymorphism in the BAFF gene), B cell subset proportions, transcriptomic profiles, levels of Type I interferon and cytokine levels, but, to date, none have entered clinical practice[84, 86-88].

1.4.5 Rituximab and B cell subsets

Highly sensitive flow cytometry, a technique established for the detection of minimal residual disease in haematological malignancies, has shown that patients have variable depletion of circulating B cells after rituximab infusions and that response rates are significantly better in patients who experience a complete depletion of their B cells after the first infusion of rituximab[89]. Repopulation of the peripheral B cell compartment occurs, on average, at 8 months; initially with immature and naive cells and, later, by memory B cells but levels of memory cells can remain reduced for more than 2 years[90-92]. Disease relapse can occur at varying time points after B cell repopulation.

There is evidence to suggest that the persistence of particular B cell subsets is associated with lack of clinical response to rituximab. A low baseline frequency of $CD27^+$ memory B cells in patients is associated with an improved clinical response to rituximab, but the absence of a specific threshold of effect limits the clinical utility of this observation [93]. In smaller studies, early clinical relapse in responders has been associated with a higher proportion of $CD27^+$ memory B cells before therapy and repopulation with a higher number of memory B cells[91, 94]. Low numbers of peripheral blood switched memory (IgD^-CD27^+) cells after treatment with rituximab have been shown to correlate with a

good response at 6, 9 and 12 months post treatment. This effect was not seen when looking at non-switched memory B cells or naive cells at the same time points[95].

Higher baseline levels of $CD20^-$ pre-plasma cells, which are not depleted by rituximab, are associated with incomplete B cell depletion and inferior response. In patients in whom B cells persist after treatment these are mainly $CD20^-$ pre-plasma cells [89, 96]. Pre-plasma cells are short-lived in the circulation and are generated from $CD20^+$ B cells so their persistence may indicate ongoing B cell activity at other sites where B cells are protected from depletion.

In contrast to the peripheral B cell compartment there is only partial depletion in the bone marrow. In a small study, rituximab responders had a significant decrease in $CD27^+$ memory cells in the peripheral blood and bone marrow at three months but this was not seen in non-responders[97].

In synovial tissue B cell depletion is again variable. The small numbers of patients in such studies has led to difficulties in reaching firm conclusions regarding the relationship between rituximab and synovial depletion, although it has been suggested response is related to better synovial depletion, specifically of plasma cells[98, 99].

1.4.6 Rituximab and transcriptomic biomarkers

Owczarczyk *et al* showed that mean baseline mRNA expression of IgJ, a marker of plasmablasts, was elevated in rituximab non-responders. This was confirmed in three additional cohorts of patients undergoing B cell depletion (2 cohorts using rituximab and the third using ocrelizumab). Application of a combination biomarker of IgJ^{hi} and $FCRL5^{lo}$ (a splice variant expressed predominantly on mature B cells) demonstrated that the ACR50 response rate (a 50% improvement in a composite measure) was 28% for the biomarker negative group and 9% for the $IgJ^{hi}FCRL5^{lo}$ group (OR 3.6 95%, CI 1.8-8.4). Plasmablasts are not believed to be present in significant levels in healthy individuals outside the context of infection or immunisation but are elevated in autoimmune conditions such as lupus[100]. Ozczarczyk and colleagues suggest that plasmablasts escape depletion with rituximab and may contribute to disease persistence by localising to

sites of inflammation. Plasmablasts express high affinity BAFF receptors and BAFF levels rise after B cell depletion, potentially promoting the survival of these cells[101].

Type I interferon regulated genes have been previously shown to be elevated in a subset of RA patients compared to healthy controls[102]. A type I IFN signature in peripheral blood mononuclear cells (PBMCs) has also been identified as a response-predictor for rituximab; with IFN-low scores (based on levels of selected type 1 IFN-response genes) associated with a better response as determined by DAS-28 score and EULAR response[103, 104].

Whole blood transcriptome profiling of RA patients commencing rituximab identified a 143 gene signature which successfully identified non-responders. Genes downregulated in the responder group were predominantly in the interferon pathway while the inflammatory genes (such as for IL-33 and *STAT5A*) and those related to NFκB were upregulated[105]. The transcriptomic data for IL-33 has been confirmed at a protein level by ELISA, as detectable IL-33 was predictive of clinical response to rituximab[106].

1.5 Genetics of RA

1.5.1 RA as a complex trait

RA is a complex, polygenic disease and the presence of strong genetic factors predisposing to RA was suggested by disease concordance in monozygotic twins and confirmed by genome wide association studies (GWAS) which have identified over 100 risk loci for RA[5, 6, 107].

The risk alleles identified by GWAS studies are single nucleotide polymorphisms (SNPs), the variant allele having a frequency of greater than 1% in the population, the most common form of genetic variation in humans. There are approximately 10 million throughout the human genome which are used as biological markers[108]. Alleles are considered to be in linkage disequilibrium (LD) if they are associated together due to infrequent recombination between them, and regions inherited together are termed haplotype blocks.

GWAS studies in RA have identified many SNPs associated with disease but the majority lie in non-coding regions and so their functional effect has been difficult to elucidate. This is for a combination of reasons: the current limitations in our understanding of non-coding regions of the genome, difficulty in identifying the causal SNP from those in LD with it, and challenges identifying which clinical context or cellular subset is affected by a variant. Each risk allele implicates multiple candidate genes and the functional effects at a cellular level remain to be established.

The human leukocyte antigen (HLA) alleles have been estimated to contribute between 10 and 40 % of the genetic risk associated with RA, while the non-HLA alleles combined explain around 5% of the association[109, 110]. In common with complex traits such as height, there remains a discrepancy between the effect of the detected genetic variants identified and the estimated heritability of the trait, approximately 60% in RA, and this is often described as the missing heritability of such traits [107, 111, 112]. The possible explanations for this include: missed variants with small effects, rare variants with larger effects and additional factors which may modify the effect of risk variants such as chromatin architecture and environment[111, 113].

The majority of work on the genetic risk for RA has focused on the ACPA positive group. The genetic contribution to the seronegative form of RA is smaller; as are the number of polymorphisms associated with it[114]. In the Swedish population, the heritability of ACPA positive RA has been estimated at around 50% compared to 20% in ACPA negative RA[115]. The seronegative subset is less well understood, possibly as it represents a more heterogenous group as there are no serological markers to define the group [112, 116]. There are shared non-HLA risk loci between the two groups for example *PTPN22*, *BLK*, *ANKRD55/IL6ST*, *STAT4*, *TNFAIP3 locus 1* and *C5orf30*. The effect size of *BLK*, *ANKRD55*, *STAT4* and *C5orf30* is the same for both subsets, while that of *PTPN22* is greater in the ACPA positive group[117]. The RA risk loci have been established in ACPA positive cohorts but ACPA negative specific risk loci, *PRL* and *NFIA*, have also been identified[117, 118].

1.5.2 HLA and the 'shared epitope' hypothesis

The strongest genetic association with RA comes from a group of alleles within the HLA region. Three amino acid positions in HLA-DRβ1 (positions 11, 71 and 74), HLA-B position 9 and HLA-DPβ1 position 9 have been shown to explain the majority of this association in seropositive RA[119]. The alleles, termed the shared epitope (SE), encode amino acid sequences that lead to structural similarities within the binding groove of the MHC class-II heterodimer[120].

Changes in this region of the 'shared epitope' are thought to influence the interaction between the MHCII complex and $CD4^+$ T helper cells and, in mice, the conversion of arginine to citrulline increases the binding affinity between the peptide and SE, leading to activation of $CD4^+$ T cells[72]. Although a well-established risk factor, the exact mechanism and stage of disease at which the SE contributes to the development RA is not established. As it is more strongly associated with ACPA positive RA and associated with higher titres of autoantibodies, the SE may indeed preferentially present citrullinated antigens initiating disease in the correct setting [121]. It has recently been shown, in twin studies, that the presence of the SE may be particularly important in determining which ACPA positive individuals develop clinical RA[122].

Although it was previously suggested that this region had little effect in the ACPA negative group, two amino acid positions (HLA-DRβ1 position 11 and HLA-B position 9) have been associated with ACPA negative RA[114, 116]. These HLA alleles are also associated with ACPA positive disease but the amino acid residues conferring risk were distinct between groups[123].

1.5.3 PTPN22

The functional consequences from the inheritance of the risk alleles is, as yet, unclear but there is evidence from functional work implicating intrinsic defects in B cells in disease pathogenesis. One of the strongest additional risk factors for RA lies in the protein tyrosine phosphatase nonreceptor type 22 gene, *PTPN22*. This polymorphism is associated with numerous autoimmune disorders including RA where risk is conferred by the minor allele at rs2476601[124]. It is notable in its association with disorders with a strong autoantibody component. The protein tyrosine phosphatases remove phosphate

residues from tyrosines on intracellular proteins, and so regulate the signal transduction threshold. In T cells it is established as a negative regulator of T cell signalling. The role in B cells is less clear; it is thought to have an inhibitory effect on the BCR, as shown by reduced calcium mobilization in response to stimulation[125]. Activation of the BCR by soluble anti-IgM leads to decreased phosphorylation of downstream signaling proteins in the presence of the disease risk allele, and this effect is reversed by inhibition of PTPN22[126]. In healthy donors it has been shown that the *PTPN22* RA risk allele is associated with an increase in polyreactive, newly emigrant cells from the bone marrow, suggesting an alteration in B cell central tolerance[127]. The presence of the risk allele alters BCR signal transduction and so, as discussed earlier, this has the potential to affect B cell development as well as function. The risk allele may, therefore, lead to a hyporesponsive state, preventing the generation of the required level of signal in response to binding self antigen and the subsequent survival of autoreactive B cells which are released into the periphery.

1.5.4 Genetics of gene expression

The primary ambition behind GWAS studies in complex disease such as RA was to improve the understanding of disease susceptibility and pathogenesis. However, the lead SNPs identified are frequently in non-coding regions of the genome or tagging LD blocks containing more than one gene and so their functional effect has been difficult to elucidate. It has not been as straightforward as a genetic variant being identified in a given gene and, by extension, the implication that the gene product is disrupted in RA, leading to insights into pathogenesis and potential therapeutic targets.

In many cases, such as at the *FAM167A/BLK* locus, work has focussed on the most likely causal gene; in this case B lymphocyte kinase (*BLK*) rather than the uncharacterised *FAM167A*. One method to establish a functional link and identify causal genes is to look for downstream consequences of this genetic variation, such as gene expression.

Gene expression levels have a heritable component and expression quantitative trait locus (eQTL) mapping analyses test for associations between genetic variants and total gene expression levels. It is used to prioritise the variants identified from GWAS studies. The gene expression level, a quantitative trait, is compared between individuals with different

genotypes at the polymorphic locus under analysis. The assumption being that the quantitative trait is modified by a genetic variant.

The genetic influences on gene expression may act in *cis* or *trans.* A *cis* eQTL acts locally within a defined distance, typically set at 1Mb but can be up to 5Mb, while a *trans* acting eQTL shows an association with more distant gene expression, typically set at 5Mb, and can act on a different chromosome (figure 1.6) [128, 129].

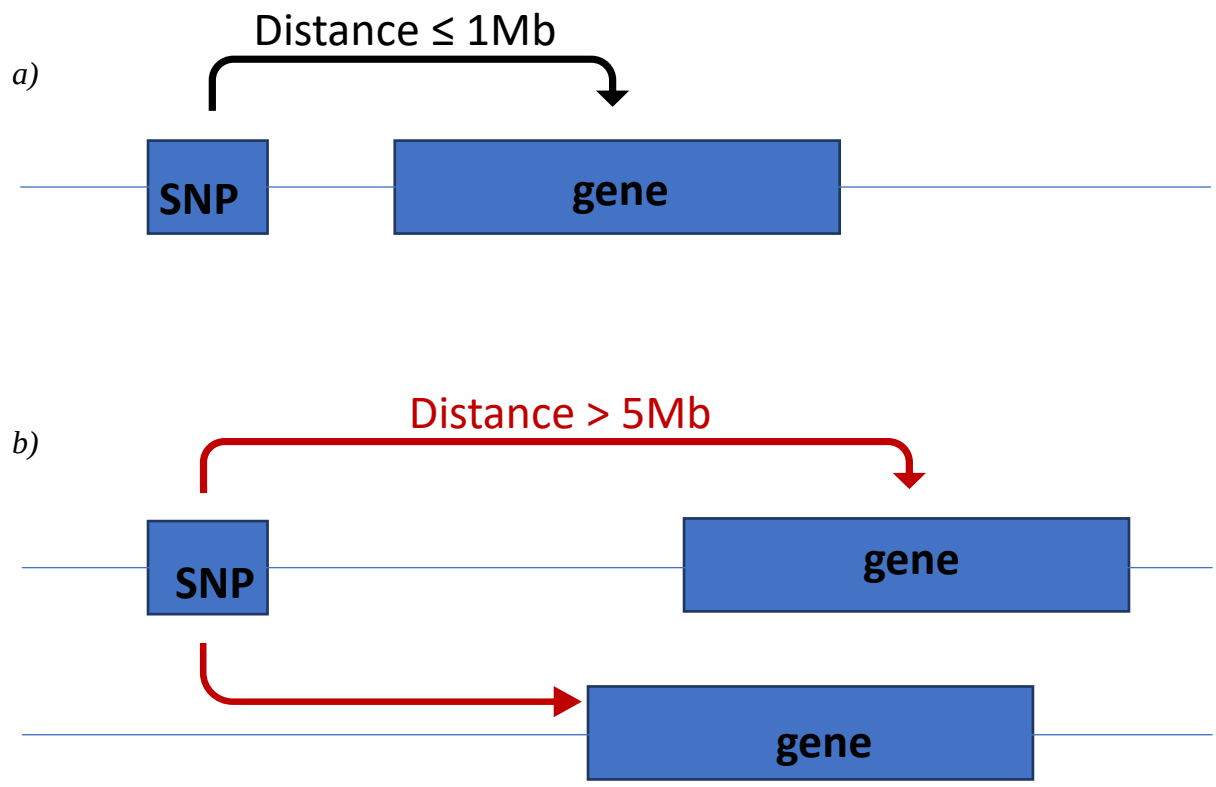


Figure 1.6 Schematic representation of expression quantitative trait loci

Expression quantitative trait loci (eQTL) are single nucleotide polymorphisms (SNPs) associated with gene transcription levels a) Cis eQTLs alter the expression of local genes shown here as within 1 megabase (Mb) b) Trans eQTLs act on distant genes, shown here at over 5Mb, and genes on different chromosomes.

Although heritable, gene expression varies across cell types and is also influenced by the environment. Examining *cis*-regulatory variation in lymphoblastoid cell lines, skin and fat, the MuTHER study showed that eQTLs can be shared between tissues or be tissue

specific. However, even in the case of shared eQTLs, the magnitude of the effect can differ between tissues[130].

Analysis of cell subsets has demonstrated the cell specificity of eQTLs in immune cells, although it has been suggested that early studies had overestimated this by directly comparing lists of eQTLs generated by separate analyses and failing to account for incomplete power in tissue by tissue analyses[131]. Indeed, a comparison of *cis* eQTLs identified between positively selected $CD19^+$ and $CD14^+$ cells from 280 healthy volunteers found an overlap of just 21.8%[132]. More recent studies using a joint analysis framework have shown that this methodology improves the identification of shared eQTLs and found that 45% of eQTLs were shared between $CD19^+$, $CD4^+$, $CD8^+$, $CD16^+$ and $CD14^+$ cells[133]. Interestingly, a subset of eQTLs shared between different immune cells exert opposing effects on gene expression[132, 133]. The difficulty in replicating identified eQTLs in PBMCs highlights the importance of cell subset considerations to avoid missing cell-specific eQTLs, due to lack of power in studies on whole blood or PBMCs[132, 133]. Deconvolution strategies have been designed to explore the relative proportions of different immune subsets and their activation status within microarray datasets. However, this relies on the prior knowledge of the expression signature for each cell subset to be established; either from an established database or the isolation of individual subsets for the construction of expression signatures, which is challenging for the rare subsets[134, 135].

Beyond cell type, environmental factors have been shown to affect eQTLs *in vitro* and their clinical relevance has been highlighted by Peters *et al* who used linear modelling with a 'genotype x disease' interaction term to demonstrate eQTLs which are only present in the setting of inflammatory disease, in this case inflammatory bowel disease (IBD)[136]. The same group showed that the eQTL for CTDP1 found in the IBD cohort and a cohort of patients with ANCA associated vasculitis disappeared in the latter group with treatment. This finding may indicate evidence of the effect of inflammation on the eQTL, although the observation may be secondary to drug effects[133].

GWAS trait-associated SNPs are enriched at eQTLs, hence eQTL mapping has the potential to aid the biological understanding of RA GWAS loci and prioritise potential disease-causing genes [137]. In RA, an example of this is the GWAS SNP rs3761847,

which is in LD with two plausible inflammatory genes: *TRAF1* and *C5*, in B cells, this SNP has been shown to be an eQTL for *TRAF1*[133]. Fairfax *et al* showed that 14 of the *cis* eQTLs (HLA loci excluded) identified in B cells were shared with GWAS SNPs associated with RA, 6 of these were shared eQTLs with monocytes and an additional 7 such RA SNPS were specific to monocytes[132]. Although, this study only looked at two immune subsets it does highlight the possibility that a subset of RA GWAS loci may be specific to B cells.

1.6 Biomarker discovery in RA

1.6.1 Definition of a biomarker

A biomarker, as defined by the Biomarkers Definitions Working Group, is "a characteristic that is objectively measured and evaluated as an indicator of normal biological processes, pathogenic processes, or pharmacological responses to a therapeutic intervention"[138]. In RA there remains a lack of diagnostic, prognostic and therapeutic biomarkers.

1.6.2 The need for biomarkers in RA

RA is also a clinically heterogeneous condition with patients experiencing different patterns of disease. Although aggressive treatment regimens offer the opportunity to reduce subsequent disability, if applied to all patients, this will result will in the over-treatment of those with more benign disease and exposure to potential drug related toxicities. In addition, not all patients respond to the same medications and so therapeutic biomarkers are required to personalise or stratify treatment regimes. The current step up treatment regime, universally applied, can result in a delay in attaining disease control[68].

1.6.3 Transcriptional biomarkers

Gene expression analyses using high-throughput microarray technologies allow the study of the expression of thousands of genes in one experiment, providing quantitative information on gene expression in a given condition[139]. The protein-coding RNAs detected reflect the proteins required for cellular function and so may provide an insight into disease pathogenesis. The use of leucocyte subsets for microarray analyses is well

established and has the advantage of obtaining data which is not confounded by the relative proportions of different cell subsets present between experimental groups[140, 141].

These technologies generate a vast amount of data but, using bioinformatics technology, it is possible to group differentially expressed genes to identify cellular pathways which may provide insights into disease pathogenesis and aid the identification of potential new therapeutic targets. Alternatively, they could be used to identify subsets of disease, leading to personalised or stratified patient care[142]. This area has been most fruitful in oncology where, for example, gene expression profiling of breast tumours can help determine prognosis[143, 144].

1.6.4 Transcriptional profiling in autoimmune diseases

Transcriptional profiling has provided insights into autoimmune diseases and has often been focussed on whole blood samples. Work on whole blood identified an interferon gene signature which predicts the development of RA in at risk individuals[145, 146]. The increased expression of B cell related genes such as *CD79A*, *CD79B*, *CD19*, *CD20* and *FCRL5* was protective against progression to arthritis. This finding was unexpected given that the increased expression of *CD79A*, *CD79B* and *CD19* suggests upregulation of the B cell receptor and increased humoral activity[146]. It has been suggested that the interferon gene signature is present in a subset of RA patients and, rather than being universal and translatable as a diagnostic biomarker, it may have greater utility as a therapeutic biomarker[104, 147].

Microarray analyses in patients presenting to an early arthritis clinic have identified a $CD4^+$ T cell-derived, 12 gene transcriptional signature which predicts the later development of RA in patients with UA. This also provided insights into potential disease mechanisms, as signal transducer and activator of transcription 3 (STAT3) inducible genes were shown to be over-represented[16]. Furthermore, transcription profiling in $CD8^+$ T cells in autoimmune vasculitis has identified two distinct subgroups predictive of long term prognosis. The subgroups could be identified by measuring the expression of just three genes, which adds great potential for its translation into clinical practice[148].

The majority of leucocyte subset studies have focused on T cell subsets, which may be a reflection on the relative paucity of B cells in PBMCs as they make up approximately 1-7% of PBMCs. There is, to date, one published study which assessed differences in gene expression in B cells between healthy controls and RA patients but pooled samples of RNA were used. The authors identified 305 genes which were upregulated in the disease group and 231 genes which were downregulated. The genes identified were involved in the cell-cycle, proliferation and apoptosis [149]. However, more recently, using RNA sequencing, Imgenberg-Kreuz *et al* have identified over 4000 differentially expressed genes in the $CD19^+$ transcriptome of a small cohort of female patients with untreated Sjogren's syndrome and healthy controls; the top upregulated, validated gene being *CX3CR1*[150].

1.6.5 Transcriptomic evidence for the role of B cells in tolerance

The transplantation literature has provided a novel insight into tolerance. The improvements in immunosuppression have led to improved graft survival but patients are committed to long term treatment with immunosuppressive regimes that come with potential side effects. The ideal situation is a state of 'tolerance' as defined as 'patients with stable graft function without continuous immunosuppression'; in the majority of cases this relates to patients who are non-compliant with the relevant medications.

The role of B cells in transplant tolerance was demonstrated by Newell *et al* who identified a B cell gene signature following a microarray study comparing the whole blood gene expression profiles of renal transplant tolerant patients to those with stable graft function on immunosuppressive treatment and healthy controls. This showed that, of 30 genes upregulated by two fold or more in the tolerant group, 22 were B cell specific. There was no significant difference in gene expression when comparing tolerant patients to healthy controls. A set of just three genes (*IGKV4-1*, *IGLL1* and *IGKV1D-13*) distinguished between tolerant and non-tolerant transplant recipients (PPV 83%, NPV 84%). The 3 identified classifier genes all encode for the immunoglobulin light chain[51]. The same group identified an increased number of naive and transitional B cells in tolerant patients, which combined with the expression data adds weight to the importance of B cells in tolerance and, in particular, the potential role of individual B cell subsets.

Sagoo *et al* similarly described a predominance of B cell related genes in the renal tolerance gene signature but, as is often the case with transcriptomic studies, the gene lists between the two groups showed limited overlap[151]. In light of the conflicting results in the literature a meta-analysis combining the different data sets from 5 separate studies looking at a total of 596 samples identified and validated a gene signature to discriminate between tolerant and stable patients. The top 20 markers, the majority of which were B cell centred (including upregulation of *BLK*, *MS4A1*, *CD79B* in the tolerant group), successfully discriminated between tolerant patients and patients stable on standard immunosuppression, with little or no difference between the tolerant group and healthy volunteers, highlighting the central role of B cells in tolerance[152].

1.7 Summary

There is increasing evidence of the importance of B cells in RA, beyond the presence of autoantibodies, but their specific role in its pathogenesis remains unknown. It may be that the disease is initiated by a particular subset of B cells; for example, the presence of pathogenic B cells, a relative deficiency of Bregs or an abnormal Breg population.

Transcriptomic profiling has the potential to provide insights into disease pathogenesis. The clinical utility of any cell signature identified in B cell subsets will be determined by the identification of a small number of classifier genes and its subsequent, further validation in whole blood RNA samples to enable translation into clinical practice.

The insights gained from immunophenotyping and gene expression data may allow the future targeting of a subset of B cells, rather than blanket depletion in the treatment of RA.

In this thesis, I will examine B cells from patients in a DMARD-naive early arthritis cohort to explore changes in gene expression, B cell subsets, and the influence of genetic variants on gene expression in this cohort.

1.8 Hypothesis, aims and objectives

Hypothesis:

I hypothesise that the pathogenic role of B cells in RA is associated with a distinct transcriptional profile and changes in the proportions of B cell subsets in the peripheral blood.

In addition, I hypothesise that an eQTL analysis, focussed on B cells, will reveal eQTLs which are both specific to this cell subset and RA.

Aims:

To improve the early diagnosis of RA and identify causative pathways which may lead to new treatments and insights into pathogenesis.

Objectives:

1) Examine the B cell transcriptome in an early arthritis cohort, comparing RA and non-RA patients to identify disease related changes
2) Examine the influence of age and inflammation on the B cell transcriptome in an early arthritis cohort
3) Perform an eQTL analysis of non-HLA RA risk loci in B cells and an interaction analysis to identify RA specific eQTLs
4) Analyse B cell subsets in an early arthritis cohort

2. Methods

2.1 Samples

2.1.1 Ethics and Sponsorship

The NRES Committee North East - County Durham & Tees Valley provided ethical approval for the project: *Prognostic and Therapeutic Biomarkers in an Inception cohort: the Northeast Early Arthritis Clinic*, REC number12/NE/0251 (2012). Additional approval was provided by the South West 3 Research ethics committee for the project: *Newcastle Autoimmune Inflammatory Rheumatic Disease Research Biobank;* REC reference 10/H0106/30 (2012) for the recruitment of healthy volunteers and patients with established rheumatoid arthritis. Sponsorship was provided by Newcastle upon Tyne Hospitals NHS foundation Trust.

2.1.2 Patient recruitment

Patients are referred to the Newcastle Early Arthritis clinic (NEAC) from primary care with a suspected inflammatory arthritis. The primary aim of this NHS clinic is to promptly identify and initiate treatment for patients presenting with a new inflammatory arthritis. General Practitioners (GPs) are encouraged to refer those they clinically suspect to have an inflammatory arthritis rather than waiting for supporting investigations prior to referral. In light of this, the clinic assesses a proportion of patients who are later shown to have a non-inflammatory arthritis.

Patients are given two linked appointments, 1-2 weeks apart. The first appointment is intended to be within 2 weeks of receipt of the referral letter. Patients are also sent information sheets and consent forms related to research prior to their appointment.

At the first appointment (visit 1) patients are seen by a nurse specialist for a clinical assessment including a DAS-28 score, a musculoskeletal ultrasound, X-rays of hands, feet and chest, urine dipstick and blood tests for routine clinical purposes. The DAS-28 is a composite score derived from a count of the number of swollen joints (out of 28), tender joints (out of 28), patient global health assessment and inflammatory markers (CRP or ESR can be used). At this appointment, the patient has the opportunity to discuss the research study.

At the second appointment (visit 2) the patient is assessed by a consultant Rheumatologist and a diagnosis made based on the clinical picture and investigations. This diagnosis

assigned at this stage is the baseline diagnosis. Visit 2 will be termed the baseline visit in this thesis. The diagnosis is coded to one of 12 categories for later database entry (*see Appendix A.1*). If the diagnosis is unclear then the patient is described as having an undifferentiated arthritis (UA).

After this visit, the patient may be discharged or assigned an NHS follow up appointment depending on clinical need. The diagnosis may change depending on the clinical picture, particularly in the case of UA patients. The working diagnosis reflects the most up to date diagnosis assigned to a patient and may be different to that made at the baseline visit. The working diagnosis is updated on the clinical database after follow up clinic appointments (figure 2.1).

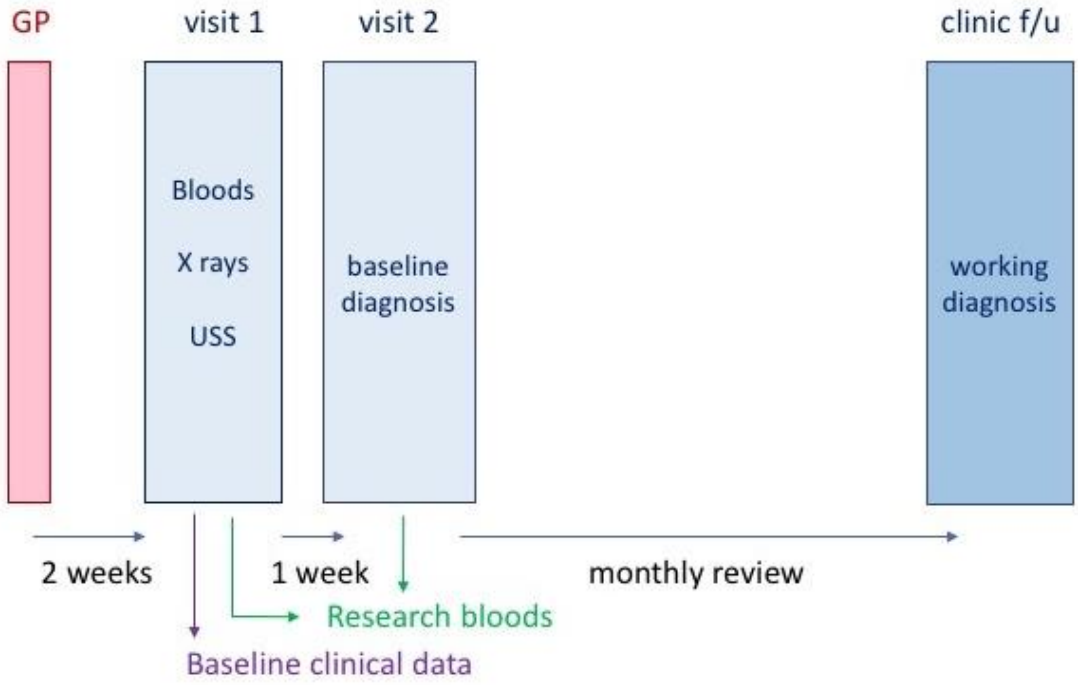


Figure 2.1 Newcastle Early Arthritis Clinic patient pathway

Patients are referred to the Newcastle Early Arthritis Clinic by the General Practitioners (GPs). At the first appointment (visit 1) patients are seen by a nurse specialist and clinical assessments and investigations carried out. At the second appointment (visit 2) patients a baseline diagnosis is made. Patients may be recruited to the study at visit 1 or visit 2. Patients are subsequently seen on a monthly basis and the updated diagnosis are termed the working diagnosis for the purpose of this study.

Patients were recruited for this study at visit 1 or visit 2. The inclusion criteria were:

a) Referred with suspected inflammatory arthritis

b) ≥ 16 years of age

c) Able and willing to give informed consent.

d) Naive to disease modifying anti-rheumatic drugs (DMARDs)

e) No steroid therapy of any kind for ≥ 2 months.

There were no exclusion criteria if the inclusion criteria were met.

2.1.3 Newcastle Early Arthritis Clinic patient cohorts

The samples were divided in to groups determined by baseline diagnosis. The primary analysis, to identify disease specific, differentially expressed genes, compared the gene expression profile of RA samples to non-RA samples in a discovery cohort. The patients who were labelled with UA at initial presentation were omitted from this analysis, with the intention of using this group as a prediction cohort to validate any potential discriminatory gene signature identified (figure 2.2).

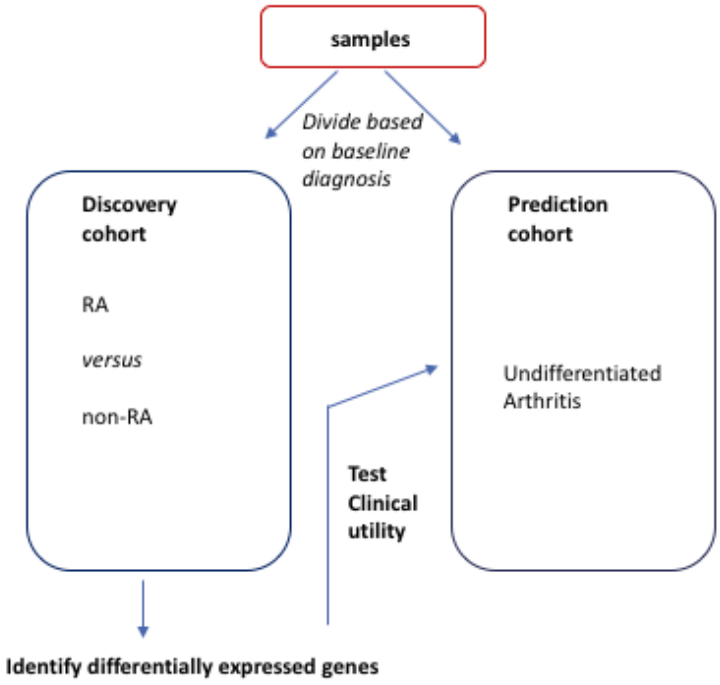


Figure 2.2 Patient cohorts for gene expression analysis

The samples were divided in to two cohorts based on baseline diagnosis. Rheumatoid arthritis (RA) and non-RA samples were included in the discovery cohort to identify any differentially expressed genes. Samples from undifferentiated arthritis samples were included in the prediction cohort.

2.1.4 Healthy volunteers

Healthy volunteers were recruited from the Musculoskeletal Research Group and donated blood for whole blood flow cytometry studies.

2.1.5 Established rheumatoid arthritis patients

Patients with established rheumatoid arthritis were recruited from the Freeman Hospital and the samples used for whole blood flow cytometry studies. Patients were all awaiting treatment with rituximab for active rheumatoid arthritis.

2.1.6 Blood sampling

Blood samples from patients attending the NEAC were taken at the same time as for routine tests where possible. All samples were taken in the morning. The time between blood sampling and processing was between 1 and 4 hours, depending on the time of the patient's clinic appointment.

2.1.7 Clinical database

The clinical data including relevant aspects of the patient history, examination findings and investigations are stored on a Microsoft Access database.

2.2 B cell isolation

2.2.1 Peripheral blood mononuclear cell (PBMC) isolation

Principle

PBMCs can be removed from whole blood by density centrifugation at room temperature

Method

Whole blood (median volume 45ml) was collected in EDTA tubes (containing K_2EDTA) (Greiner Bio-one, Germany). Blood was diluted 1:1 with Hank's balanced salt solution (Hanks) (Ca^{2+} and Mg^{2+} free) (Lonza, Switzerland) containing 2mM endotoxin free EDTA (Fisher Scientific UK). Density centrifugation was performed by layering 15-20ml of the diluted blood over 15ml of Lymphoprep^{TM} (Axis-Shield Diagnostics, Norway) in each 50ml centrifuge tube and centrifuged at 895g at room temperature for 30 minutes. After centrifugation the PBMCs form a band at the interface between the sample and medium. The PBMCs were recovered from this interface using a Pasteur pipette and

transferred to a new 50ml centrifuge tube. Cells from 2 lymphoprep tubes were pooled into each new centrifuge tube. Each tube was filled with cold Hank's balanced salt solution containing 1% foetal calf serum (FCS) (Sigma Aldrich, UK) to make a total volume of 50ml. The sample was centrifuged at 600g for 7 minutes at 4^{o}C to remove any contaminating Lymphoprep^{TM}. Lymphoprep^{TM} is toxic to cells if left in contact with them. The supernatant was aspirated and the cells resuspended. The cells were pooled again into one centrifuge tube and diluted with cold Hank's balanced salt solution containing 1% FCS to a total volume of 50ml and spun at 250g for 7 minutes at 4^{o}C, to remove any contaminating platelets. The supernatant was aspirated and cells resuspended in 20ml of cold Hank's balanced salt solution containing 1% FCS. The cell suspension was strained through a 70µm nylon filter to remove any debris or clumps. The number of PBMCs was counted using a Burker counting chamber. $2x10^5$ PBMCs were transferred to a 96 v-bottom well plate and stored at 4˚C for use in a cell purity check by flow cytometry. The remaining cells were used for $CD19^+$ B cell isolation.

2.2.2 Positive selection of $CD19^+$B cells

Principle

CD19 is a transmembrane glycoprotein expressed during all phases of B cell development, until terminal differentiation into plasma cells. CD19 is also expressed on follicular dendritic cells; these cells are found in the primary and secondary lymphoid structures. CD19 MicroBeads are 50nm superparamagnetic particles conjugated with monoclonal CD19 antibodies (isotype: mouse IgG1).

A positive cell selection means the target cell type is magnetically labelled, in this case with CD19 MicroBeads, and the cell suspension is passed through a magnetic column. The column contains a matrix of ferromagnetic spheres. The column is held in a magnetic separator and the spheres amplify the magnetic field, inducing a high gradient within the column. The unlabelled cells pass through but labelled cells are retained in the column. The column is then removed from the magnetic field and the target, positively selected, cells eluted (figure 2.3).

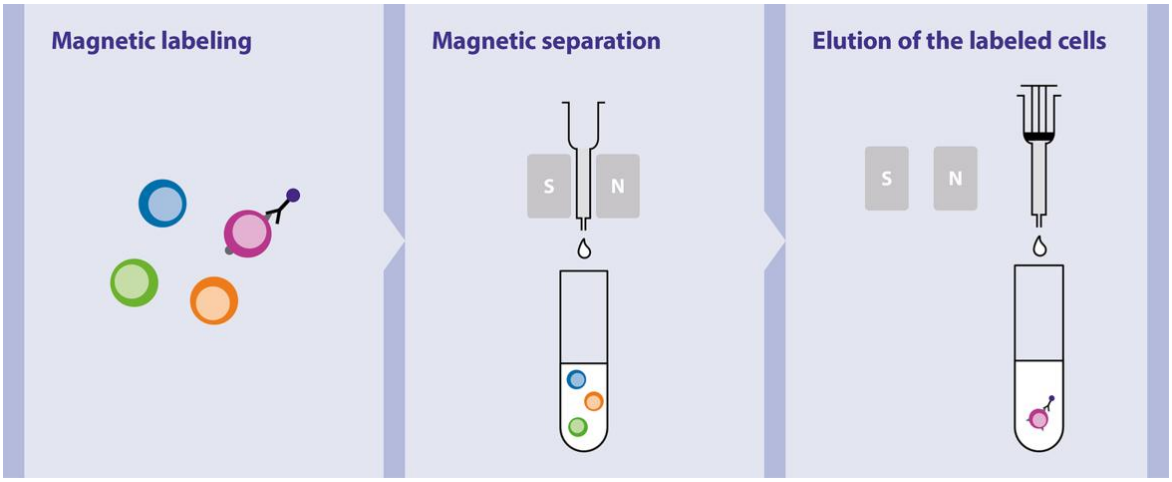


Figure 2.3 Positive cell selection using MicroBead technology

The target cell is magnetically labelled, retained within the column during separation while unlabelled cells flow through. After a washing step the column is removed from the magnetic field and target cells eluted. www.miltenyibiotech.com[153]

Method

The isolations were carried out using pre-cooled solutions and the cells kept on ice to prevent capping of antibodies on the cell surface and non-specific cell labelling. The volumes described below are for the magnetic labelling of up to $1x10^7$ total cells. When working with higher cell numbers the reagent volumes and total volumes were scaled up accordingly.

MACS buffer contains 500 ml PBS (Ca^{2+}/Mg^{2+} free) (Lonza, Switzerland), 2.5ml FCS and 2 ml Endotoxin free 0.5M EDTA, filtered through a 0.2 µm filter and cooled on ice prior to use.

Hanks containing 1% FCS was added to fill the tube containing the remaining PBMCs, after cells were removed for flow cytometry. The tube was spun at 400g at 4°C for 7 minutes. The supernatant was aspirated and cells resuspended in MACS buffer (80µl MACS buffer per $1x10^7$ PBMCs) and 20µl MACS CD19 MicroBeads (Miltenyi Biotech, Germany) were added per $1x10^7$ cells. The cells were mixed and incubated at 4°C for 15 minutes in a refrigerator. The cells were washed by adding 10ml of ice-cold MACS buffer and spun at 400g at 4°C for 7 minutes. The supernatant was aspirated completely and cells resuspended in 500µl of MACS buffer (500µl per $1x10^8$ cells).

A positive selection column (LS) (Miltenyi Biotech, Germany) was placed in the MACS separator and the column prepared with a pre-rinse of 3ml of ice-cold MACS buffer. The cell suspension was added to the column. The unlabelled cells which passed through were collected in a centrifuge tube and the column washed 3 times by adding 3ml of MACS buffer. The MACS column was removed from the magnetic separator and placed in a 30ml universal tube. 5ml of MACS buffer was added to the column and the plunger pushed firmly in to the column to flush out the magnetically labelled cells. The tube was filled with MACS buffer and spun at 400g at 4^{o}C to wash the cells. The supernatant was aspirated and the cells resuspended in 1ml of Hanks containing 1% FCS. The cells were then counted. If there were $\geq$ 1.2x10^{6} CD19^{+} B cells then 2x10^{5} cells were transferred into one well of the 96 v-bottom well plate and stored at 4^{o}C for use in the purity check by flow cytometry. The remaining cells were lysed.

2.2.3 Freezing homogenised CD19^{+} B cells in Qiagen RLT buffer

Principle

RNA stabilisation is required after the harvesting of samples to prevent unwanted changes in the transcriptome due to RNA degradation or transcriptional induction. This can be done by disrupting and homogenising the cells in the presence of RNase-inhibiting or denaturing reagents. The disruption of cell walls and plasma membranes allows the release of all the RNA from the samples and homogenisation reduces the viscosity of the lysate produced by the disruption. Incomplete homogenisation can lead to inefficient binding of RNA to the spin column used for the RNA extraction. Samples can then be stored in lysis buffer at –80°C for months.

Method

RNase, DNase-free filter tips and RNase, DNase, pyrogen-free microcentrifuge tubes were used.

The universal tube containing the positively selected cells was filled with Hanks containing 1% FCS and spun at 400g at 4^{o}C for 7 minutes. The supernatant was aspirated and 350µl of lysis buffer added to disrupt the cells. The lysis buffer used was Qiagen Buffer RLT from the AllPrep DNA/RNA mini kit (Qiagen, Germany) to which β-

mercaptomethanol (Sigma Aldrich, USA) had been added at a ratio of 1:100. If the cell count was over $5x10^6$ cells then 600µl of lysis buffer was added. This was mixed thoroughly by pipetting and vortexed. The lysate was homogenised by adding it to a QIA-shredder column (Qiagen, Germany) in a 2ml collection tube and spun at maximum speed in a microcentrifuge for 2 minutes at 4°C. The QIA-shredder column was removed, a cap placed on the 2ml collection tube and the sample stored immediately at -80°C until RNA and DNA extraction.

2.2.4 CD19$^+$ B cell purity check by flow cytometry

Principle

The purity of the isolated $CD19^+$ B cells can be checked by flow cytometry. PBMC and isolated $CD19^+$ B cells are labelled with fluorescently conjugated antibodies recognising different surface markers and the proportions of $CD19^+$ B cells measured using flow cytometry. The starting population of PBMCs is assessed to determine the recovery of the $CD19^+$ B cells. The Miltenyi MACS MicroBead technology used for the cell isolation requires only minimal cell labelling and so sufficient epitopes remain available for fluorescent staining on the magnetically labelled cells.

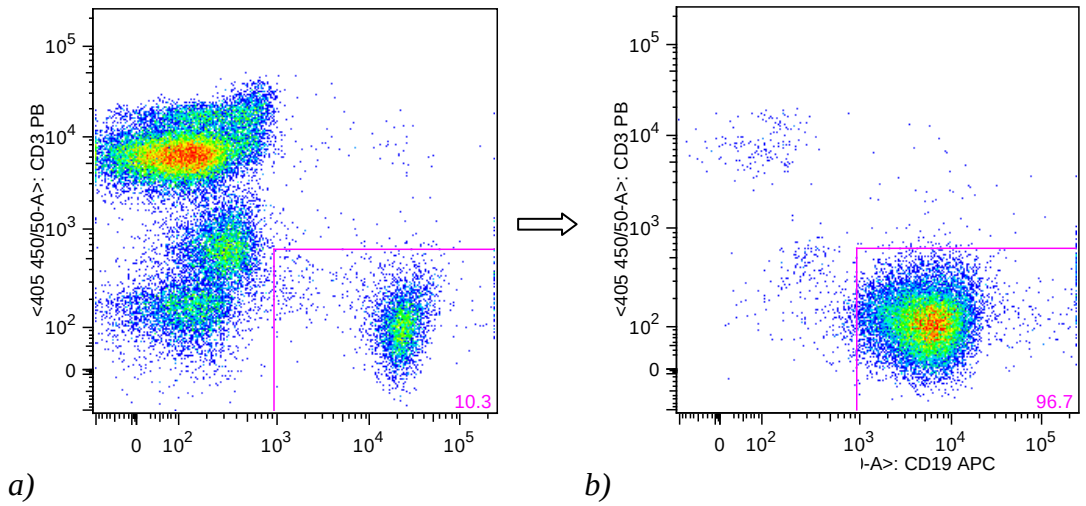
Method

The PBMCs and $CD19^+$ B cells were centrifuged at 400g for 3 minutes at 4 °C. The supernatant was removed and the pellet resuspended in flow cytometry staining buffer (Dulbecco's phosphate buffered saline (DPBS) (Mg^{2+} and CA^{2+} free) (Lonza, Switzerland) containing 0.01% sodium azide (Sigma-Aldrich, UK), 0.5% bovine serum albumin (Sigma-Aldrich, UK) and 1mM EDTA), human IgG at 4µg/ml (Octagam, Octapharma Limited), and fluorophore labelled surface antibodies as per the table below (table 2.1) to a final volume of 50µl per well.

Surface marker	Fluorophore	Dilution	Clone	Supplier
CD3	PB	1:50	UCHT1	BD, Biosciences
CD4v4	FITC	1:200	L120	BD, Biosciences
CD14	PE	1:20	M5E2	Beckton Dickinson,
CD19	APC	1:10	LT19	Miltenyi Biotec

Table 2.1 Fluorophore labelled antibodies used to assess purity of positively selected CD19$^+$ B cells

The samples were incubated at 4 ^{o}C in the dark for 30 minutes and then washed twice in flow cytometry staining buffer by centrifugation at 400g for 3 minutes at 4°C, resuspended in 200µl of staining buffer and acquired on a FACSCanto II (Beckton Dickinson, USA). Data was analysed using FlowJo software (TreeStar, USA). Figure 2.4 shows a representative example.



a) *b)*

Figure 2.4 An example of a CD19$^+$ B cell purity check

The starting PBMC population a) and the purified CD19$^+$ B cell population b) were stained for cell surface markers and assessed by flow cytometry. The population in the pink box are CD19$^+$ B cells and the number in pink in the lower right corner indicates the percentage of these cells in the entire population. T cells (CD3$^+$) can be seen in the top left area of plot b).

2.3 Nucleic acid extraction

2.3.1 Extraction of RNA and DNA from frozen homogenised CD19$^+$ cells

Principle

Total RNA and genomic DNA can be simultaneously purified from frozen homogenised CD19$^+$ B cell lysates using spin technology (AllPrep DNA/RNA mini kit) (figure 2.5). The thawed lysate is first passed through a DNA spin column which, in combination, with a high salt buffer, allows the selective binding of DNA to the column membrane from which DNA can then be eluted. Ethanol is added to the flow through from this process to create appropriate binding conditions for RNA and this is then applied to an RNeasy spin column where the total RNA binds to the column and contaminants washed away. An on-column DNase digestion step is carried out to eliminate any contaminating genomic DNA. The RNA can then be eluted from the column.

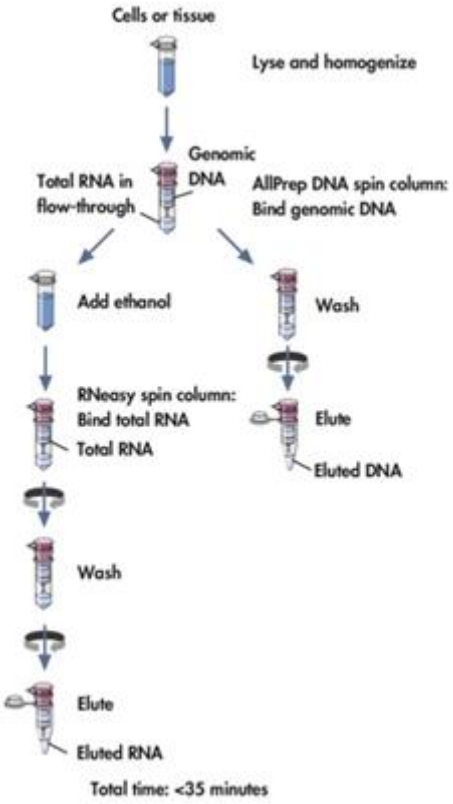


Figure 2.5 Overview of AllPrep RNA/RNA procedure

Qiagen, Germany[154]

Method

The AllPrep DNA/RNA mini kits were used (Qiagen, Germany) to extract DNA and RNA from the homogenised lysates as per the manufacturer's instructions (AllPrep DNA/RNA mini handbook, Qiagen, Germany).

The sample was removed from the freezer and allowed to thaw to room temperature. The homogenised lysate was added to an AllPrep DNA spin column placed in a 2ml collection tube and centrifuged for 30 seconds at 8000g. The DNA spin column was then placed in a new 2ml collection tube and stored at room temperature for later purification after RNA purification.

RNA purification: 350 µl of 70% ethanol was added to the flow through from the DNA spin step and mixed well by pipetting. This was then transferred to an RNeasy spin column placed in a 2ml collection tube and spun for 15 seconds at 8000g. The flow through was discarded. The on-column DNase digestion was then carried out using the RNase-Free DNase Set (Qiagen, Germany). 350µl Buffer RW1 was added to the RNeasy spin column and centrifuged for 15 seconds at 8000g to wash the spin column membrane. The flow through was discarded. 10µl DNase I was added to 70µl Buffer RDD and mixed gently. 80µl of the DNase incubation mix was added directly to the RNeasy spin column membrane, where the RNA was bound, and incubated at room temperature for 15 minutes. 350µl Buffer RW1 was added to the RNeasy spin column and centrifuged for 15 seconds at 8000g column to remove the DNase I. The flow through was discarded, completing the DNase digestion.

The column was washed by adding 500µl Buffer RPE to the RNeasy spin column and centrifuged for 15 seconds at 8000g. The flow through was discarded, a further 500µl Buffer RPE was added and the column centrifuged for 2 minutes at 8000g to dry the column membrane and ensure no ethanol was carried over to the elution stage. The RNeasy column was placed in a new 2ml collection tube and centrifuged at full speed for 1 minute to eliminate carry over of Buffer RPE. The RNeasy column was placed in a new 1.5ml collection tube to elute the RNA. 30µl RNase-free water was added directly to the spin column membrane and the column centrifuged for 1 minute at 8000g. The sample was then stored at -80^{o}C.

DNA purification: The AllPrep DNA spin column which was stored at room temperature was now washed by adding 500µl Buffer AW1 and centrifuged for 15s at 8000g. A further wash was carried out by adding 500µl Buffer AW2 and the column centrifuged at full speed for 2 minutes to dry the spin column and prevent carry over of ethanol to the DNA elution. The AllPrep DNA spin column was placed in a new 1.5ml collection tube and DNA eluted by adding 100µl Buffer EB directly to the spin column, incubating at room temperature for 1 minute. The column was then centrifuged for 1 minute at 8000g and the eluted DNA stored at -80^{o}C.

2.4 Illumina Human HT12-v4 Expression BeadChip

290 RNA samples were sent to the laboratory of Professor Anne Barton at the University of Manchester for whole genome microarray analysis. The sample preparation, quality control and microarray work described in section 2.4 was carried out by Dr Nisha Nair at the University of Manchester, unless otherwise stated.

2.4.1 RNA Quality assessment

Principle

The use of intact RNA is critical to successful gene expression analyses. The Agilent bioanalyser system uses electrophoretic separation and the subsequent detection of RNA samples to provide a visual display of the quality of the RNA sample and the software is used to generate a standardised assessment: RNA integrity number (RIN) for each sample. The RIN range produced is 0-10, RIN:0 describes the most degraded sample, RIN:10 the most intact RNA.

Method

The RNA quality of the samples was assessed using the Agilent 2100 bioanalyzer (Agilent Technologies, USA), the Agilent RNA 6000 Nano kit and Agilent 2100 expert software as per the manufacturer's instructions to obtain an RNA integrity number (RIN) for each sample. An example of a sample electropherogram trace is shown in figure 2.6.

The median RIN for the samples was 10 (range 7.9-10).

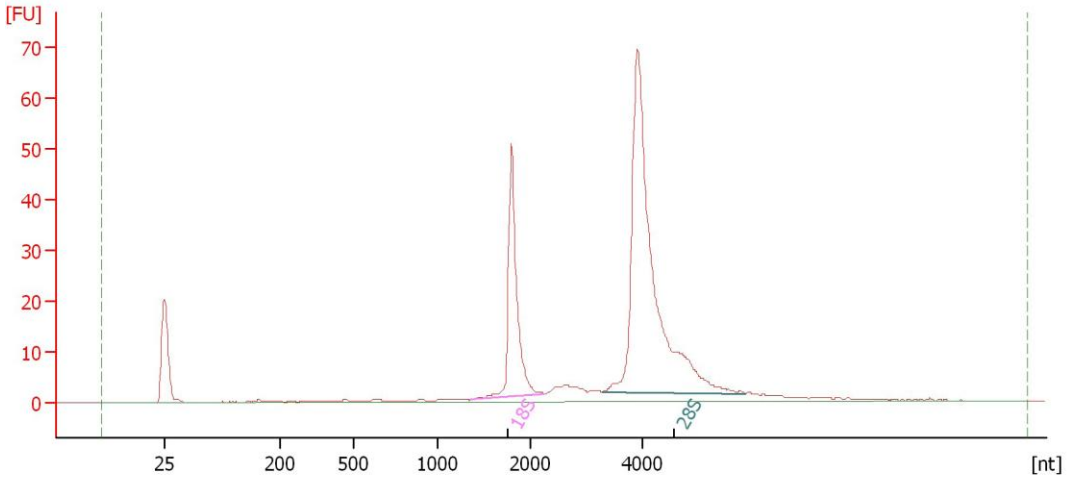


Figure 2.6 Agilent bioanalyser electropherogram summary of one RNA sample

Fluorescence is shown on the y axis, fragment size on the x axis. A 25 nucleotide marker peak is shown. 2 distinct ribosomal peaks are seen, sample with a RIN 10, indicative of intact RNA (results and image from Cambridge Genomics Services, UK, produced as part of my MRes project).

2.4.2 Illumina Human HT12-v4 Expression BeadChip

Principle

The human HT12-v4 expression beadchip uses a direct hybridisation assay where gene specific probes are used to detect labelled cRNA. Each bead has a sequence specific oligo probe. Each bead is covered with hundreds of thousands of copies this probe, that act as the capture sequence (figure 2.7).

The HT12-v4 expression beadchip allows the measurement of the expression levels of 47,000 transcripts and splice variants from RefSeq Database Release 38 and other sources.

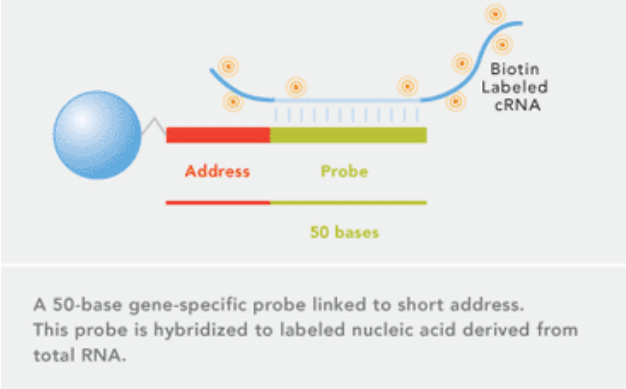


Figure 2.7 Direct hybridisation of labelled cRNA from the sample to probe

Each silica bead on the array has hundreds of thousands of copies of a specific oligonucleotide attached to it, a single probe is depicted here for ease of visualisation (Image, Illumina, USA[155])

Method

The Illumina TotalPrep 96-RNA amplification kit from Ambion (Life Technologies, USA) was used to generate biotinylated, amplified RNA for hybridisation to the microarrays. In 3 batches the RNA was converted to complementary DNA (cDNA) by reverse transcription. The cDNA underwent second strand synthesis followed by a single in vitro transcription amplification step to generate biotin-labelled cRNA. The cRNA was purified and quantified before hybridisation to the beadchip. 42 samples were noted to be contaminated at this stage with ethanol and underwent a reconcentration and clean up procedure (see *Methods 2.4.3*).

The Illumina iScan was used to detect fluorescence emission and was carried out in 7 batches. The raw data was processed using the *beadarray* package in RStudio and was transferred to Newcastle University in the form of IDAT files with the a Sentrix id sheet containing positional information to enable identification of the arrays.

2.4.3 Ethanol precipitation of RNA for re-concentration and clean up

Principle

Contamination of the sample can occur during the RNA conversion process. The sample can be precipitated out of the aqueous solution using ethanol and then resuspended for downstream applications.

Method

During the conversion process 42 samples were noted to be contaminated with ethanol. To each sample the following were added: 2 µl glycogen, 0.5 of the starting volume of 7.5M ammonium acetate, 2.5 of the starting volume of ice cold 100% ethanol. The sample was mixed thoroughly by vortexing, precipitated at -20^{o}C for 1 hour then spun at full speed (16,000g at 4^{o}C for 30 minutes) and the supernatant removed. The pellet was washed twice with 0.5ml of ice cold 80% ethanol (16,000g at 4^{o}C for 10 minutes) and the pellet air dried for 10 minutes before resuspending in 50µl of nuclease-free water.

2.5 Microarray data analysis

The microarray data analysis was carried out with the support of Andrew Skelton from the Bioinformatics Support Unit. The analysis was performed in RStudio using Bioconductor libraries[156]. The workflow pathway can be summarised as: importing the data, pre-processing the data, fitting the linear models, making the comparisons required to test the hypotheses, and lastly visualising the results.

2.5.1 Pre-processing

The sample probe profiles were read into RStudio. Background correction and normalisation was carried out using the limma package[157]. Normalisation is carried out to remove systematic effects due to technical differences which are unrelated to the biological differences between the samples. It aims to place the expression values for all the samples on the same measurement scale prior to analysis based on phenotype. The neqc function was used to perform background correction followed by quantile normalization. This function is used for Illumina BeadChip data utilising the control probes specific to these arrays: negative control probes for background correction and both negative and positive controls for normalisation[158].

The illuminaHumanv4.db package was used to map annotation of the array probes, providing information on the NuID, IlluminaID, gene symbol, ensemble annotation and probe sequence.

2.5.2 Quality Control

Quality control is carried out to optimise the downstream analyses: the methods used were based on detection P-values, the removal of technical failures (samples which had failed on the array) and principle components analyses. The pre-processing steps were repeated after the removal of probes or samples identified in this way.

The detection p-value, is the confidence that a given transcript is expressed above the background level and this data is read into RStudio. The probes were filtered in RStudio to keep probes which met the P-value detection threshold of $p<0.01$ in 25% or more samples. This removed 31,676 probes. An annotation filter was used to identify probes which do not represent unique capture sequences, removing 5560 probes. 10974 probes were taken forward after probe filtering to the further analyses.

The technical failures are identified by checking the proportion of microarray probes expressed in each array and 3 samples were identified as technical failures and subsequently removed. A principal components analysis identified 1 outlier which was removed from all further analyses.

The principal components analysis was employed to visually examine the data to identify problems potentially arising from sample processing. This enables the removal of any systematic bias prior to further analysis. The samples were colour coded to look in turn for effects based on: conversion batch, scanning batch, chip ID, samples which had undergone a clean up procedure and $CD19^+$ B cell purity. The samples were then coded by clinical factors: gender, smoking, diagnosis, seropositivity and based on levels of inflammation. The samples clustered together based on conversion batch alone (figure 2.8). Conversion batch was, therefore, added to the linear model.

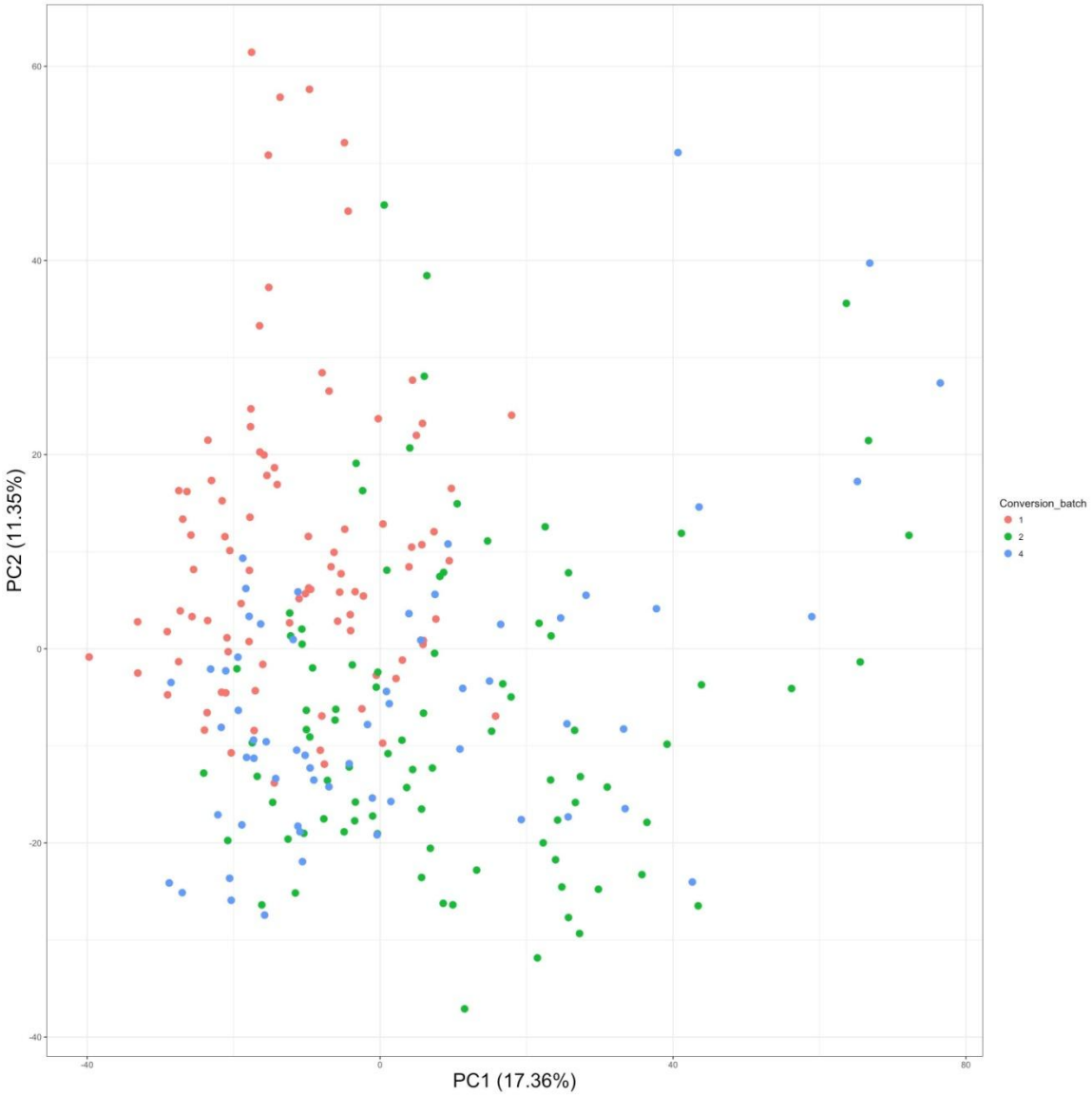


Figure 2.8 Principal Components Analysis by conversion batch

Visualisation of the normalised data in 2 principal components. Each point represents one array and the samples are colour coded by conversion batch. The proportion of the total variance attributable to each component is indicated on the axis.

2.5.3 Differential gene expression

Differentially expressed genes were sought using linear modelling in the *limma* package. This method allowed the flexible modelling of experimental features such as batch and additional clinical factors such as inflammatory markers.

A design matrix was first created indicating which RNA samples have been applied to which array. A contrast matrix was created to define the contrasts to be used to test the

hypotheses of interest. The *lmfit* and *contrasts.fit* functions were used for assessing differential expression. *Limma* computes the $\log_2$-fold-changes and *t*-statistics for the contrasts of interest. An empirical Bayesian method was used to moderate the standard errors of the estimated fold changes.

To prioritise the results, a $\log_2$-fold change threshold was set at 1.2 to identify genes with a fold change equal or greater than the set threshold, rather than different to zero. This fold change threshold has been shown to clinically relevant in $CD4^+$ T cells[16]. The Benjamini-Hochberg false discovery rate (FDR) corrected p-value of <0.05 was used to adjust for multiple testing. The *toptable* function was used to produce a spreadsheet of the filtered results including gene annotation. The results of the individual contrasts carried out were visualised using volcano plots.

2.5.4 Ingenuity pathway analysis

Ingenuity® Pathway Analysis (IPA) (Qiagen, Germany) is a web-based software application used to interpret high-throughput biological data. IPA uses the Ingenuity Knowledge Base for its analysis; a manually curated, maintained and frequently updated repository of relevant biological and chemical data based on the literature which includes data on the directional changes in any reported relationship between molecules[159].

A gene list table for each comparison to be examined was created containing: gene symbol, logFC, P-value and FDR. The lists were uploaded to the IPA platform for analysis. IPA integrates the dataset with previously observed relationships in the literature. The data was examined to identify canonical pathways, upstream regulators and molecular networks within the dataset with the aim of providing a biological insight to the observed changes.

There are 2 primary statistical tests used in the core IPA analysis: the p-value of overlap and the Z score. The null hypothesis employed is that the molecule in the dataset does not overlap with those present in a given biological pathway, function or disease. The p-value of overlap is calculated using a right tailed Fisher's exact test, indicating the chance of getting a given result, or a more extreme result, if the null hypothesis is true. A p-value of <0.05 is generally deemed significant and is generated for all the analyses in IPA. The Z

score makes a prediction as to the activation status of a pathway. It takes directional expression data to compare the expression patterns of your dataset to information in the ingenuity database. A Z score of ≤-2 predicts inhibition and ≥2 activation. If there is insufficient information in the literature curated database then a Z score is not provided. The molecular network analysis does not use directional data.

2.5.5 Gene set enrichment analysis

The Gene Set Enrichment Analysis (GSEA) software evaluates the microarray data at the level of gene sets rather than isolated changes in single genes[160, 161]. The gene sets are determined by prior biological knowledge. The Molecular Signatures Database (MSigDB) is a collection of manually curated, annotated gene sets designed for use in gene set enrichment analyses and link with the GSEA software[162]. By examining gene sets rather than individual genes the aim is to aid the interpretation of data where long lists of genes may not appear to have a unifying biological function and their interpretation is reliant on the knowledge of the assessor. By using gene sets it increases the signal relative to noise and improves statistical power. It is useful in the setting of noisy data due to human heterogeneity.

GSEA is able to detect modest co-ordinated differences in gene expression. Cellular processes often affect groups of genes together and single gene analyses may miss such effects; the level of change in a single gene may be small but a change in a group genes in concert may have a meaningful, relevant effect on a cellular pathway. GSEA aims to identify if members of a gene set tend to occur at the top or bottom of the list of genes from the experiment.

The GSEA desktop application was downloaded for use. Phenotype and expression files for the analysis were produced in R. For the expression file, the GSEA algorithm does not filter the dataset and recommends that an unfiltered dataset is uploaded. The detection P-value filter was, therefore, removed. The annotation filter remained in place for creating the expression file.

The genes are ranked based on their expression and the user determined phenotypic class to create a list of genes (L). The aim of GSEA is to determine if members of the prior

defined gene set (S) are randomly distributed through L, or primarily found at the bottom or top of this list.

An enrichment score (ES) is first calculated to reflect the degree gene set S is overrepresented at one extreme of the list L. The ES is calculated by walking down the list L and increasing a running-sum statistic if a gene in S is identified or decreasing it, if the gene identified is not in S. The enrichment score is the maximum deviation from zero in the random walk and corresponds to a weighted Kolmogorov–Smirnov-like statistic (figure 2.9). The statistical significance of the ES is calculated using a phenotype-based permutation test. The ES is normalised to account for the size of the gene sets, to provide a normalised enrichment score (NES). The proportion of false positives is controlled for by calculating the FDR. GSEA highlights the gene sets with an FDR of less than 25%. The term leading-edge subset refers to the genes in S that appear in list L before the running sum statistic reaches its maximum deviation from zero.

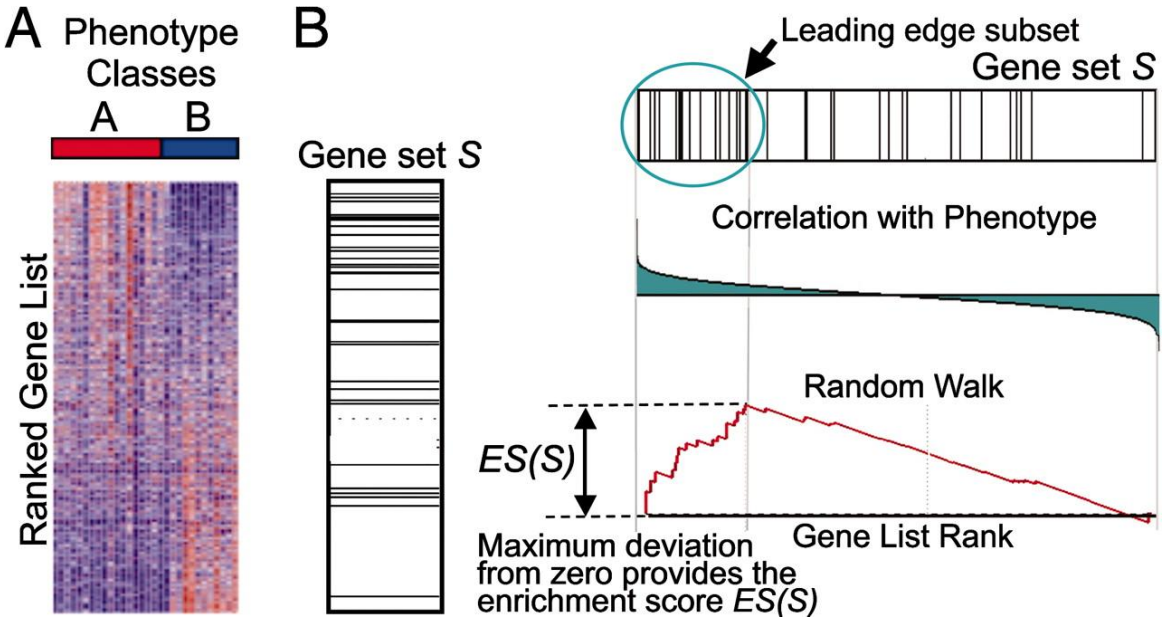


Figure 2.9 GSEA overview

(A) A heat map of an expression data set sorted by correlation with phenotype. (B) the "gene tags," i.e., location of genes from a set S within the sorted list and plot of the running sum for set S in the data set, including the location of the maximum enrichment score (ES) and the leading-edge subset[160].

2.6 Genotyping (University of Manchester, UK)

Genotyping was carried out at the laboratory of Professor Anne Barton at the University of Manchester by Dr Nisha Nair using Illumina Human CoreExome-24 version 1-0 arrays, following the manufacturer's protocol.

The quality control and data processing steps were carried out by Andrew Skelton at the Bioinformatics Unit, Newcastle University. Samples and SNPs with a call rate of <98% were excluded. SNPs with an Illumina GenomeStudio cluster separation of <0.4 were excluded from further analysis. Data were pre-phased using SHAPEIT2 and imputed to the 1000 Genomes Phase 1, version 3, reference panel using IMPUTE2. Imputed SNPs with INFO scores of <0.8 were excluded.

2.7 B cell phenotyping

2.7.1 Cell surface protein expression

Principle

Cell surface markers can be used in combination to identify different B cell subsets in whole blood using flow cytometry.

Method

200µl of whole blood collected in EDTA tubes was added to 2ml microcentrifuge tubes. Surface marker antibodies were added as per table 2.2.

Surface marker	Fluorophore	Dilution	Clone	Supplier
CD19	APC	1:200	HIB19	BD, Biosciences
CD24	APC-eFluor 780	1:50	SN3	eBioscience
CD27	V450	1:200	M-T271	BD, Biosciences
CD38	PerCP-Cy5.5	1:50	HIT2	BD, Biosciences

Table 2.2 Fluorophore labelled antibodies used for B cell phenotyping in whole blood

The microcentrifuge tube was then incubated at 37 ^{o}C in a water bath for 15 minutes following a protocol which had been optimised within the group, as a proportion of B cell markers are not robust to detection after fixation. The cells were fixed by adding 10 volumes of warmed FACS lysing solution (Beckton Dickinson, Biosciences, USA). The

sample was mixed by vortexing to ensure erythrocyte lysis and incubated at 37 ^{o}C in a water bath for 12 minutes. The sample was removed and centrifuged at 600g for 8 minutes at room temperature, the supernatant removed, the tube vortexed to disrupt the cell pellet and then cells washed with FACS buffer by centrifuging at 600g for 6 minutes at room temperature. Samples were resuspended in 200µl of FACS buffer and acquired on a FACSCanto II (Beckton Dickinson, USA). Data was analysed using FlowJo software (TreeStar, USA).

2.7.2 Flow cytometry analysis

The data acquired was analysed using FlowJo software v10.4.2 (TreeStar, USA). Doublets were first excluded by examining the side scatter area (SSC-A) and side scatter width (SSC-W) plot (figure 2.10a). The singlets were taken forward from this gate. The SSC-A against forward scatter area (FSC-A) plot was used to identify the lymphocyte and leucocyte populations (figure 2.10b).

The lymphocyte gate was used for the identification of $CD19^{+}CD27^{+}$ memory B cells and $CD19^{+}CD27^{-}$ naive B cells (figure 2.10d).

The leucocyte gate was taken forward and, plotting SSC-A against CD19, used to identify the $CD19^{+}$ B cell population (figure 2.10c). Using this $CD19^{+}$ B cell population the markers from table 2.2 were used in combination, using contour plots to identify the regulatory B cell subsets $CD19^{+}CD24^{hi}CD38^{hi}$ (figure 2.10e) and $CD19^{+}CD24^{hi}CD27^{+}$ (figure 2.10f) and plasmablasts ($CD19^{+}CD27^{hi}CD38^{hi}$) (figure 2.10g) as a percentage of $CD19^{+}$ B cells.

To identify the $CD19^{+}CD24^{hi}CD38^{hi}$ population the gate was drawn to the right of the inner contour of the CD38 positive population (CD38 on the x axis) and at the bottom of the inner contour of the CD24 population (CD24 on the y axis). To identify the $CD19^{+}CD24^{hi}CD27^{+}$ population the $CD24^{hi}$ contours were used as before and $CD27^{+}$ co-ordinates taken from the previous plots described. For plasmablasts ($CD19^{+}CD27^{hi}CD38^{hi}$) identification, the CD38 contours were used as before and $CD27^{hi}$ population gate drawn at the top of the inner contour of the positive population. In cases where $CD27^{hi}$ population could not be clearly defined on the B cell population, the co-

ordinates were taken from the $CD27^{hi}$ population defined on the lymphocyte plot. If required, the co-ordinates for the $CD24^{hi}$ and $CD38^{hi}$ gates from the $CD19^{+}CD24^{hi}CD38^{hi}$ plot were used to define the $CD24^{hi}$ and $CD38^{hi}$ populations in the $CD19^{+}CD24^{hi}CD27^{+}$ and $CD19^{+}CD27^{hi}CD38^{hi}$ plots.

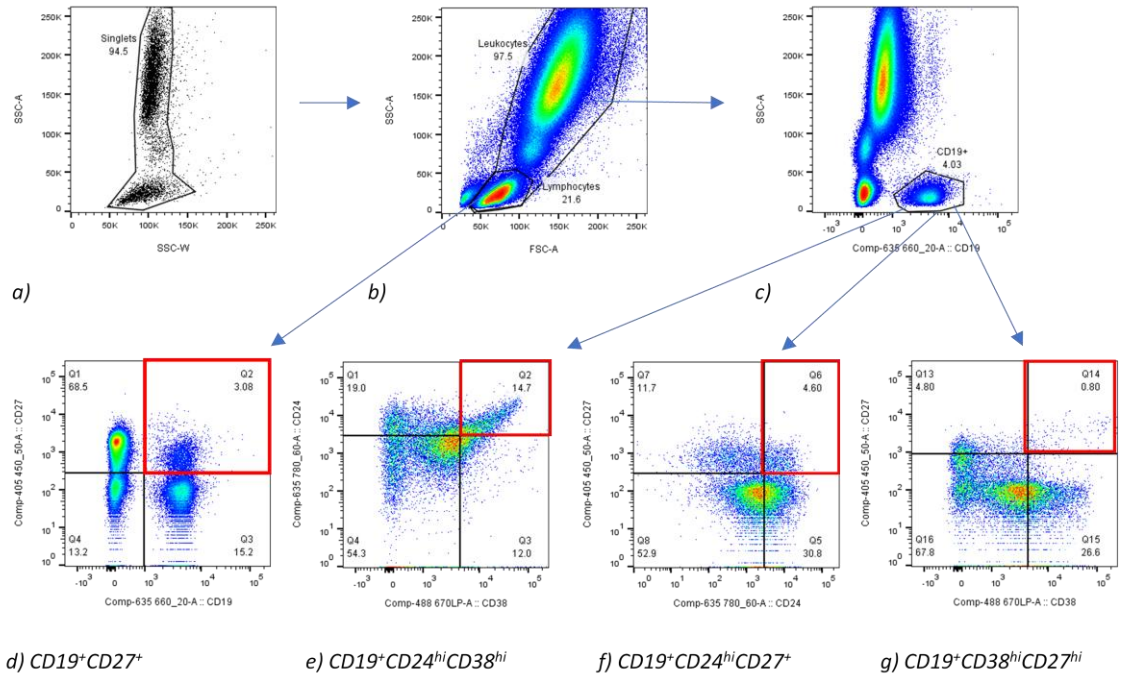


Figure 2.10 Gating strategy for B cell subsets

The singlet gate was identified (a) and carried forward to identify the leucocyte and lymphocyte populations (b). The lymphocyte gate was taken forward to identify the memory ($CD19^{+}CD27^{+}$) and naive ($CD19^{+}CD27^{+}$) B cell populations (d). The leucocyte gate was taken forward to identify the $CD19^{+}$ B cell population (d). The $CD19^{+}$ B cell population was taken forward to first identify the $CD19^{+}CD24^{hi}CD38^{hi}$ subset (e), followed by the $CD19^{+}CD24^{hi}CD27^{+}$ population (f) and $CD19^{+}CD27^{hi}CD38^{hi}$ population (g) Red box identifies the gate of interest for each plot.

2.8 Serum IL-6 detection by MesoScale Discovery (MSD) assay

Serum IL-6 was measured using the MesoScale Discovery (MSD) assay by Dr Amy Anderson and Dr Arthur Pratt following the manufacturer's instructions.

Principle

The MSD assay allows the quantitative measurement of cytokines, using electrochemiluminescent labels conjugated to detection antibodies. The MSD plates are pre-coated with capture antibodies. The sample is added to the plates with a solution

containing the detection antibody conjugated to the electrochemiluminescent labels. The cytokine to be measured binds to the capture antibody and the detection antibody then binds to the cytokine of interest. The MSD buffer is added to create the appropriate chemical environment for electrochemiluminesce. Voltage is applied to the plate electrodes, causing the bound electrochemiluminescent labels to emit light. The light intensity emitted is measured, providing a quantitative measure of the cytokine in each sample.

Method

Blood from patients was collected in Serum Separator tubes (Greiner). The serum was separated by centrifuging the tube at 1800g for 10 minutes at room temperature. Aliquots of serum were stored in labelled 1.5ml microcentrifuge tubes and frozen at -80°C. Samples were defrosted for use in a MSD assay.

The MSD analysis of serum was carried out according to the manufacturer's instructions using a V-PLEX Human IL-6 Kit diluent (Meso Scale Discovery; Maryland, USA).

2.9 Statistical methods

The microarray analyses, eQTL and interaction analyses were carried out using RStudio and the appropriate packages as detailed in the appropriate sections. The statistical tests presented for clinical data and flow cytometry data were performed using GraphPad Prism (version 8.0) and IBM SPSS (version 24).

3. The gene expression profile of peripheral blood B cells in patients with early RA

3.1 Background

Oligonucleotide microarrays provide an opportunity to analyse the simultaneous expression of thousands of genes in one experiment. This has the potential to identify individual genes or gene expression patterns important in RA disease pathogenesis.

The majority of transcriptomic analyses in RA have used whole blood or PBMCs, focussed on comparing RA samples to healthy controls and sample numbers have been relatively small[102, 140, 163, 164]. The analysis of combined cell subsets, in whole blood and PBMCs, may mask meaningful biological signals from the individual cell subsets and those detected may reflect the relative abundance of a different subsets. The majority of differentially expressed genes identified in cell subset analyses are not replicated in PBMCs, and this is true of both relatively rare subsets, such as B cells, and the more abundant $CD4^+$ and $CD8^+$ T cells[165].

The choice of subset for analysis is more complex in autoimmune conditions than in oncology where transcriptomic analyses have, in some cases, translated into clinical practice[166, 167]. In RA, the primary site of disease (the synovium) is relatively difficult to access, biopsies are not routinely required for clinical practice and the pathogenesis of RA is less clearly defined. The analysis of T cell subsets has been successfully used to identify diagnostic and prognostic signatures for RA and other autoimmune conditions[16, 148, 165, 168].

As discussed in *Chapter 1*, B cells are potentially pathogenic in early RA. The sole publication on the B cell transcriptome in RA did detect differentially expressed genes but the 8 samples in each group were pooled to produce an RA sample and a healthy control sample for comparison. This was done to overcome the inter individual variability in gene expression and to maximise the amount of RNA for the microarray. The group identified over 500 differentially expressed genes and pathway analyses were carried out. The top four functional groups related to i) B cell activation and proliferation, ii) autoimmunity, iii) neuro-immune modulators and iv) angiogenesis [149]. However, this interesting result may be influenced by the use of steroid in the RA group (each patient was on a stable dose of oral prednisolone, <10mg) and pooling of the RNA may mask heterogeneity between samples limiting interpretation of this result. Although $CD19^+$ B

cells can be successfully isolated from peripheral blood samples, the lack of microarray work on this subset may also relate to the relatively low yield of $CD19^+$ B cells, and so RNA, from a given peripheral blood sample when compared to $CD4^+$, $CD8^+$ and $CD14^+$ cells[141].

The use of healthy controls as a control group in analyses provides insights into disease pathogenesis but a comparison with disease controls, particularly those presenting to an early arthritis clinic, may provide clinically useful biomarkers of disease. By looking at untreated patients with early disease I aim to identify diagnostic biomarkers and avoid the influence of DMARDs on the transcriptome. The importance of looking at early disease is highlighted by the identification of significant differences between the synovial transcriptome of patients with untreated RA and those with established RA, which may be secondary to treatments and the impact of prolonged inflammation[169].

Microarray experimental technology is well established, although the analysis remains complex and must control for factors including technical variation and batch effects, which may otherwise confound the identification of meaningful biological variation between samples[170]. The results have been shown to be reproducible and the lack of translation to clinical utility in autoimmune conditions may relate to the tissue analysed, sample numbers and inter-individual variation that is not disease related[171, 172]. In this chapter, I will examine the $CD19^+$ B cell transcriptome in a large cohort of DMARD naive patients, which has the potential to provide insights into B cell mediated mechanisms of disease pathogenesis. This approach attempts to circumvent the challenges of previous studies which have used small cohorts and mixed cell populations. The benefit of using peripheral blood is that circulating B cells are accessible, allowing for easier translation into clinical practice.

3.2 Hypothesis and aims

3.2.1 Hypothesis

In RA, the B cells are the source of autoantibody production, can produce pro-inflammatory cytokines and act as antigen presenting cells activating or amplifying auto-reactive T cells. Therefore, B cells potentially have a crucial role in disease initiation.

Examination of the B cell transcriptome will provide insights into the pathogenesis of RA, potentially highlighting cellular pathways which are altered in disease and identifying diagnostic biomarkers of disease.

3.2.2 Aims

1. To identify a list of differentially expressed genes in the transcriptome of circulating $CD19^+$ B cells of DMARD-naive RA patients *versus* disease controls.
2. To test the clinical utility of any potential diagnostic gene signature in a separate cohort of patients with undifferentiated arthritis.
3. To undertake a pathway analysis of differentially expressed genes (DEGs) with a view to gaining new insight into B cell mediated pathobiology of early RA.

3.3 Results

3.3.1 Patient cohort overview

Microarray data with accompanying phenotype clinical data were available for 240 samples in total. Based on the diagnoses at the baseline visit, (see *Methods 2.1.2* for details regarding patient recruitment) the samples comprised 59 RA patients, 50 patients with an undifferentiated arthritis (UA) and 131 patients with a rheumatological condition other than RA, referred to as the non-RA group.

The non-RA group comprised 59 patients with other inflammatory conditions and 72 patients with non-inflammatory conditions. The strategy of grouping inflammatory and non-inflammatory samples together has been successfully used in a $CD4^+$ T cell study and increases the sample size of the control group[16]. In addition, the UA cohort in which we aim to test any diagnostic signature includes those who will subsequently be given both inflammatory and non-inflammatory diagnoses[16]. The patients with UA were not included in the primary analysis, as this group included patients who would go on to be diagnosed with RA and it was intended that this sub-cohort would provide a means of validating a putative discriminatory gene signature arising from my primary comparison.

3.3.2 Demographics

In total 240 samples have been used in the analysis described in this chapter. A $CD19^+$ B cell purity check was carried out for 187 of these samples (see *Methods 2.2.4).* The first analysis used all the available samples and the second analysis used only the samples where the $CD19^+$ B cell purity was known to be $\geq 90\%$.

The clinical data for the samples used in the analyses is shown in table 3.1. The RA group is older, with more evidence of inflammation (measured by inflammatory markers ESR and CRP), and a greater number of swollen and tender joints than the disease control group.

	All samples			$CD19^+$ B cell Purity ≥90.0%		
	RA	**Non-RA**	**P-value**	**RA**	**Non-RA**	**P-value**
Sample Number	59	131	-	39	101	-
Age (yrs)	61 (21-89)	52 (18-92)	0.0002	58 (21-85)	51 (18-92)	0.0156
Gender (%F)	76.3	72.5	ns	76.9	70.3	ns
ESR (mm/hr)	25 (1-91)	10 (1-111)	0.0002	21 (1-91)	8 (1-100)	0.0013
CRP (mg/L)	10 (0-91)	7 (0-171)	0.0009	9 (0-80)	5 (0-53)	0.0120
SJC	1 (0-25)	0 (0-11)	<0.0001	1 (0-25)	0 (0-11)	0.0023
TJC	6 (0-22)	2 (0-28)	0.0006	5 (0-22)	3 (0-28)	0.0149
DAS28	4.59 (1.26-8.46)	3.44 (0.71-6.99)	<0.0001	4.31 (1.26-8.46)	3.56 (0.16-5.67)	<0.0001

Table 3.1 Demographics for gene expression analyses based on baseline diagnosis for RA and non-RA samples

The analysis was initially carried out for 'all samples' (n=190) which refers to samples analysed without knowledge of the $CD19^+$ B cell purity of all the samples used. The analysis was repeated using the samples where the $CD19^+$ B cell purity was known to be ≥90.0% (n=140). In each analysis, the rheumatoid arthritis (RA) samples were compared to the non-RA samples. Median (range) shown. P-values were calculated using the unpaired T test (age), Mann-Whitney test (CRP, ESR, SJC, TJC, DAS28), or Fisher's exact test (gender).

A third analysis included the UA cohort and samples were divided into RA and non-RA based on the current diagnosis from clinic letters. The samples which continued to be labelled as UA were excluded from this analysis. The RA group is older, with higher levels of inflammation (measured by inflammatory markers ESR and CRP), and a greater number of swollen and tender joints than the disease control group (table 3.2*).*

	Samples based on current diagnosis		
	RA	**Non-RA**	**P-value**
Sample Number	73	159	-
Age (yrs)	60 (21-89)	52 (18-92)	<0.0001
Gender (%F)	75.3	74.2	ns
ESR (mm/hr)	23 (1-91)	11 (1-111)	0.0008
CRP (mg/L)	10 (0-91)	5 (0-171)	0.0004
SJC	2 (0-25)	0 (0-12)	<0.0001
TJC	5 (0-22)	3 (0-28)	0.0009
DAS28	4.40 (1.26-8.46)	3.45 (0.16-6.99)	<0.0001

Table 3.2 Demographics for gene expression analyses based on current diagnosis for RA and non-RA samples

The undifferentiated arthritis cohort was combined with the original sample group and samples divided into RA and non-RA groups based on the current diagnosis. Median (range) shown. P-values were calculated using the unpaired T test (age), Mann-Whitney test (CRP, ESR, SJC, TJC, DAS28), or Fisher's exact test (gender).

At the first consultant visit, the baseline visit, patients are allocated 1 of 12 clinical diagnoses, including RA and UA (*see Appendix A.1*). The term 'non-RA' combines the remaining diagnoses: osteoarthritis, other non-inflammatory conditions, psoriatic arthritis, crystal arthritis, reactive arthritis, ankylosing spondyloarthritis, undifferentiated spondyloarthritis, enteropathic arthritis, Lupus/other CTD related condition or other inflammatory arthritis. The number of samples for each diagnosis within the non-RA group is shown in table 3.3. The 10 diagnoses can also be broadly grouped into other inflammatory and non-inflammatory conditions, the non-inflammatory category includes those with osteoarthritis and other non-inflammatory conditions (shown in blue in table 3.3) and the remainder are inflammatory conditions.

The demographic data for the 'other inflammatory' and 'non-inflammatory' samples provides information on the composition of the non-RA group (table 3.4).

Diagnosis	All samples	$CD19^+$ Purity ≥90.0%
Other non-inflammatory condition	39	34
Osteoarthritis	33	27
Psoriatic arthritis	20	14
Other inflammatory arthritis	12	7
Crystal arthritis	9	7
Reactive arthritis	9	4
Lupus/other CTD related condition	3	3
Undifferentiated spondyloarthritis	3	2
Ankylosing spondylitis	2	2
Enteropathic arthritis	1	1

Table 3.3 Diagnoses for samples within the non-RA group

Number of samples for each diagnosis within the non-RA group. Diagnoses categorised as non-inflammatory conditions shown in blue, inflammatory conditions in black.

	All samples			$CD19^+$ Purity ≥90.0%		
	Other inflammatory	Non Inflammatory	P-value	Other inflammatory	Non Inflammatory	P-value
Sample Number	59	72	-	40	61	-
Age (yrs)	51 (18-92)	52 (22-87)	ns	50 (18-92)	51 (22-72)	ns
Gender (%F)	64.4	79.2	ns	60.0	77.1	ns
ESR (mm/hr)	15.5 (1-111)	7 (1-100)	0.0201	11 (1-66)	7 (1-100)	ns
CRP (mg/L)	7 (0-171)	5 (0-49)	0.0166	5 (0-53)	5 (0-49)	ns
SJC	0 (0-9)	0 (0-11)	<0.0001	2 (0-9)	0 (0-11)	<0.0001
TJC	3 (0-28)	2 (0-19)	ns	2 (0-19)	3 (-28)	ns
DAS28	3.78 (0.71-6.99)	3.03 (0-5.46)	0.0156	3.44 (0.90-5.67)	3.31 (0.16-5.46)	ns

Table 3.4 Demographics for samples with other inflammatory and non-inflammatory diagnoses within the non-RA group

Median (range) shown. P-values were calculated using the unpaired T test (age), Mann-Whitney test (CRP, ESR, SJC, TJC, DAS28), or Fisher's exact test (gender).

The RA group used in the first analysis using all samples based on baseline diagnosis, comprised 13 patients (22.0%) who were seronegative and the remainder (46 patients)

were seropositive for rheumatoid factor (RF) and/or anti-cyclic citrullinated protein (anti-CCP).

The demographic data for the UA group at baseline visit are shown in table 3.5a and the current diagnoses made for patients within this group in table 3.5b. 14 of the 50 UA patients have been subsequently been diagnosed with RA, 6 of whom were seropositive.

	Undifferentiated Arthritis
Sample Number	50
Age (yrs)	53.5 (20-83)
Gender (%F)	78
ESR (mm/hr)	14 (1-99)
CRP (mg/L)	8.5 (0-76)
SJC	0 (0-15)
TJC	3 (0-22)
DAS28	3.76 (0.97-6.54)

a)

Working Diagnosis	**Number of patients**
Rheumatoid Arthritis	14
Other non-inflammatory condition	5
Osteoarthritis	5
Psoriatic arthritis	3
Other inflammatory arthritis	1
Crystal arthritis	0
Reactive arthritis	12
Lupus/other CTD related condition	1
Undifferentiated spondyloarthritis	0
Ankylosing spondylitis	1
Enteropathic arthritis	0
Undifferentiated arthritis	8

b)

Table 3.5 Demographics and current diagnoses for undifferentiated arthritis group

a) demographic data for the undifferentiated arthritis (UA) group (n=50). Median (range) shown. b) current diagnoses for patients within the initial UA group. CTD, connective tissue disease.

3.3.3 CD19$^+$ B cell quality

The normal range for B cells in peripheral blood is 1-10%. In samples where greater than 1.2 x10^6 CD19$^+$ B cells were isolated then 0.2 x10^6 cells were removed to check the purity of the isolate by flow cytometry (see *Methods 2.2.4)*. Purity checks were carried out for 187 of the 240 samples used in this analysis.

There were no significant differences in the CD19$^+$ B cell purity between the diagnostic groups. This was true when comparing the RA group to the non-RA group as a whole (as

used in the initial gene expression analysis) and when the non-RA group was subdivided into inflammatory, non-inflammatory and undifferentiated arthritis groups (figure 3.1a). The median cell purity for all samples was 95.1% (range 81.3 - 98.8%). There were no significant differences in the number of $CD19^+$ B cells isolated for each group (figure 3.1b)

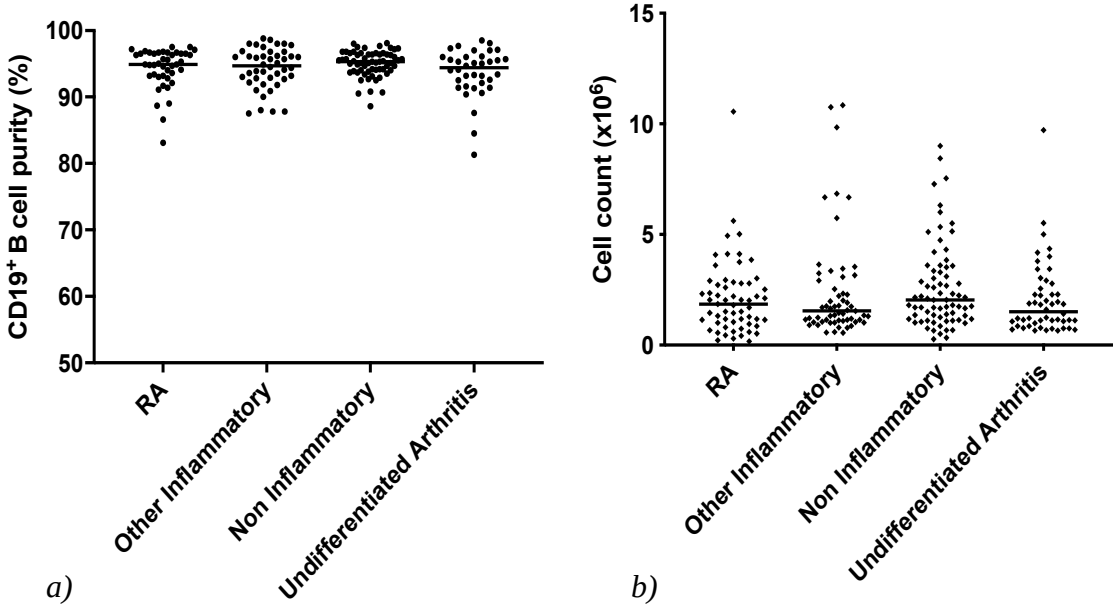


Figure 3.1 $CD19^+$ B cell purity and total $CD19^+$ B cell counts for samples from patients in the early arthritis cohort

a) $CD19^+$cell purity, RA (n=43), other inflammatory (n=44), non-inflammatory (n=62), undifferentiated arthritis (n=36) b) $CD19^+$cell count RA (n=59), other inflammatory (n=59), non-inflammatory (n=72), undifferentiated arthritis (n=50). Diagnostic categories based on baseline diagnosis. No significant differences were detected between the groups. Kruskal Wallis test.

3.3.4 Differential gene expression between RA patients and non-RA patients

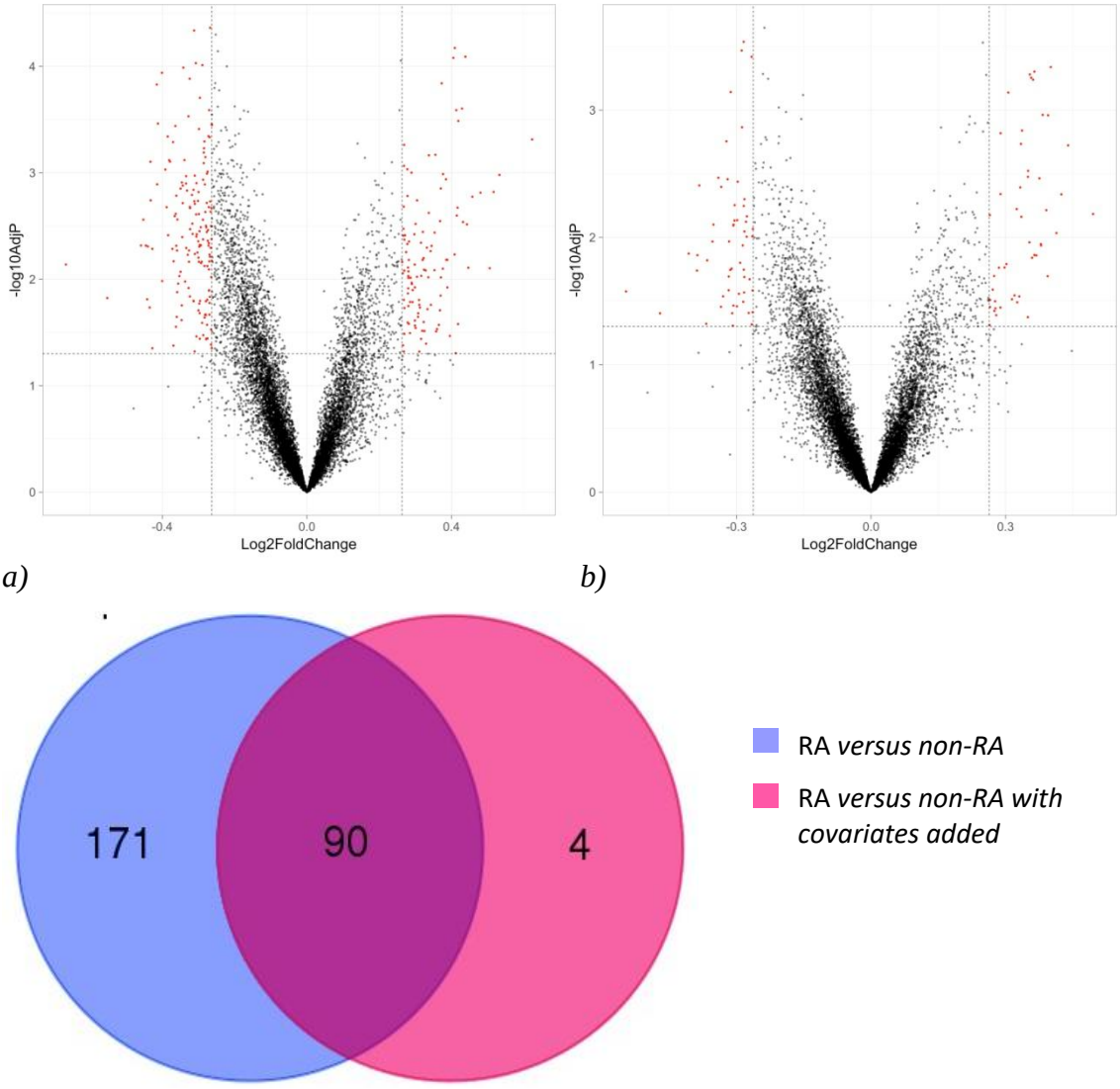
I hypothesised that, in a cohort of DMARD-naive patients presenting to an early arthritis clinic, DEGs could be identified between the $CD19^+$ B cell transcriptome of patients with RA and those diagnosed other forms of arthritis. The patients who were labelled with UA at initial presentation were omitted from this analysis, with the intention of using this group to validate DEGs.

59 RA samples were compared to 131 non-RA samples. I first compared the profiles with a multiple test correction in place (Benjamini-Hochberg, FDR adjusted p-value <0.05) looking for a fold change of ≥1.2. No differentially expressed genes were identified and I

repeated the analysis after removing the multiple test correction and this revealed 279 differentially expressed probes between the two groups, representing 261 unique genes (figure 3.2a) *(see Appendix A.2 for full list of DEGs)*

The RA group were significantly older with higher levels of inflammation and so age, ESR and CRP were then added as covariates to the linear model. There were no DEGs with multiple test correction in place but once removed, 104 probes were identified at a FC ≥1.2 and p-value <0.05, representing 94 unique genes (figure 3.2b) *(see Appendix A.3 for full list of DEGs)*

There was an overlap of 90 genes between the two lists of DEGs between RA and non-RA samples with and without the inclusion of clinical covariates (figure 3.2c). 4 genes were unique to the list generated with clinical covariates included: *DSTYK*, *MFGE2*, *CCR9* and *MAL*.



a) b) c)

Figure 3.2 Differentially expressed genes identified between the CD19$^+$ B cell transcriptome of RA and non-RA samples

CD19$^+$ B cells were positively isolated from patients presenting to the early arthritis clinic, RNA extracted and transcriptome analysed using microarray technology. Samples from DMARD-naive RA patients (n=59) and non-RA patients (n=131) were compared. a) Volcano plot of RA against non-RA samples, 279 differentially expressed probes were identified (no multiple test correction in place). b) Volcano plot of RA against non-RA samples with covariates age, ESR and CRP added to the linear model, 104 differentially expressed probes were identified (no multiple test correction in place). Vertical, dotted lines denote fold change (FC) 1.2. The x axis represents log2 of the fold change, y axis represents the $-\log_{10}$ adjusted p-value. Horizontal, dotted lines denote p-value 0.05. Red dots indicate probes which are differentially expressed between the comparator groups. c) Venn diagram of the unique differentially expressed genes (DEGs) identified comparing RA and non-RA samples with (261 unique DEGs) and without covariates added (94 unique DEGs). There is an overlap of 90 DEGs between the two lists of DEGs.

3.3.5 Ingenuity pathway analysis – choice of analysis

Differentially expressed genes were analysed using the web-based functional analysis program ingenuity pathway analysis (IPA) to identify canonical pathways, upstream regulators and molecular networks within the DEG list. The aim of IPA is to carry out a functional analysis of the dataset and to explore the biological relevance of the DEGs. IPA uses the Ingenuity Knowledge Base for its analysis, a manually curated and maintained repository of relevant biological and chemical data based on the literature.

There are 2 primary statistical tests used in the core IPA analysis: the p-value of overlap and the Z score. The Z score makes a prediction as to the activation status of a pathway. A Z score of ≤-2 predicts inhibition and ≥2 activation. If there is insufficient information in the literature curated database then a Z score is not provided. The molecular network analysis does not use directional data.

The supporting information provided by IPA advises that the gene list used should be between 200 and 700 genes long for an optimal analysis. Two lists of DEGs were generated from the comparison between RA and non-RA samples: 279 DEGs identified without clinical covariates and 104 DEGs with clinical covariates added to the model.

The list of DEGs generated with clinical covariates potentially allows a focus on the RA associated genes alone; without the confounders of genes associated with age and inflammation. However, when this shorter list of 104 DEGs was put through the IPA software, no canonical pathways, upstream regulators or molecular networks were identified within the DEG list which met the filtering criteria ($p < 0.05$ or Z score ≤-2 or ≥+2). Therefore, in light of this and the IPA supporting information, I focussed on the DEGs from the RA versus non-RA comparison without a consideration of clinical covariates for further analysis and exploratory purposes.

3.3.6 Ingenuity pathway analysis results

Canonical pathways

The canonical pathways analysis generated a total of 256 pathways. 165 of the pathways identified had just 1 or 2 molecules from the list of DEGs for each pathway. In order to

prioritise the pathways identified by the IPA software, I applied filtering criteria of ≥ 5 molecules from the DEG list found in the pathway, a Z-score of ≤ -2 or ≥ +2 and p-value <0.05. The 6 pathways that met the filtering criteria are all down regulated in the RA group (table 3.6).

There is a noticeable overlap in the molecules listed for each of the 5 pathways, for example *CD79A* and *CD79B*, transmembrane proteins which form a complex with the BCR and mediate signalling on antigen detection, are found in 4 of the 6 pathways.

Canonical Pathway	**-log (p-value)**	**z-score**	**Genes**
Phospholipase C Signaling	1.66	-2.449	*HDAC9*, *FYN*, *CD79B*, *GNG2*, *RALB*, *PRKCE*, *CD79A*
EIF2 Signaling	1.36	-2.449	*RPL32*, *WARS*, *RPL8*, *ATF5*, *IRS2*, *RPLP0*
p70S6K Signaling	1.73	-2.236	*JAK1*, *CD79B*, *PRKCE*, *CD79A*, *IRS2*
Tec Kinase Signaling	1.33	-2	*FYN*, *JAK1*, *GNG2*, *PRKCE*, *IRS2*
Role of NFAT in Regulation of the Immune Response	1.19	-2	*FYN*, *CD79B*, *GNG2*, *CD79A*, *IRS2*
B Cell Receptor Signaling	1.15	-2	*PTPRC*, *CD79B*, *BCL10*, *CD79A*, *IRS2*

Table 3.6 Canonical pathways identified from the list of differentially expressed genes between RA samples and non-RA samples

IPA software was used to analyse the list of DEGs identified by comparing the CD19^{+} B cell transcriptome from RA patients (n=59) to non-RA patients (n=131) without clinical variables added. The pathways shown have a minimum of 5 molecules from the pathway identified in the list of DEGs. Z score ≤ -2 or ≥ +2, p-val <0.05.

BCR signalling is downregulated in the RA group and the phospholipase C (PLC) signalling, p70S6K signalling and NFAT pathways operate downstream of BCR signalling. Phospholipase C signalling is one of the principal BCR activation pathways and the subsequent increase in intracellular calcium upregulates the transcription factor NFAT. p70S6K signalling is downstream of another principal pathway downstream of

the BCR, PI3K signalling. The pathway "PI3K signalling in B lymphocytes" was identified in the canonical pathway analysis with 7 molecules within the dataset (PTPRC, FYN, CD79B, ATF5, BCL10, CD79A, IRS2), but the pathway did not meet the Z score threshold selected (Z score -1.633). TEC kinase signalling refers to a family of non receptor tyrosine kinases, of which Bruton's tyrosine kinase is the best investigated in B cells and is a key component of the BCR signalling pathway.

Eukaryotic initiation factor 2 (EIF2) signalling is downregulated in response to the unfolded protein response (UPR), a process that is upregulated in antibody producing B cells.

Upstream regulators

A separate analysis stream in IPA identifies upstream regulators; cytokines, transcription factors or other molecules that explain the observed changes in gene expression in the dataset. The IPA software predicts if the regulator is activated or inhibited.

The overlap p-value measures whether there is a statistically significant overlap between the dataset genes and the genes known to be regulated by an upstream regulator, using Fisher's Exact Test, and significance is generally attributed to p-values < 0.01. 14 upstream regulators were identified, 4 predicted to be inhibited and 10 activated from the DEG list for RA versus non-RA samples (table 3.7).

The upstream regulator analysis can also be displayed as a network to allow an exploration of the relationship between the regulator and the relevant molecules in the dataset (figure 3.3).

IFNL1, a type III interferon, and IFNA2, a type I interferon, are identified as inhibited upstream regulators. Each has 5 molecules linked to it from the dataset, but 4 overlap, highlighting the importance of context when interpreting the results. The network display shows that the interactions of IFNL1 and IFNA2 with the molecules they are connected to are indirect (figure 3.3a).

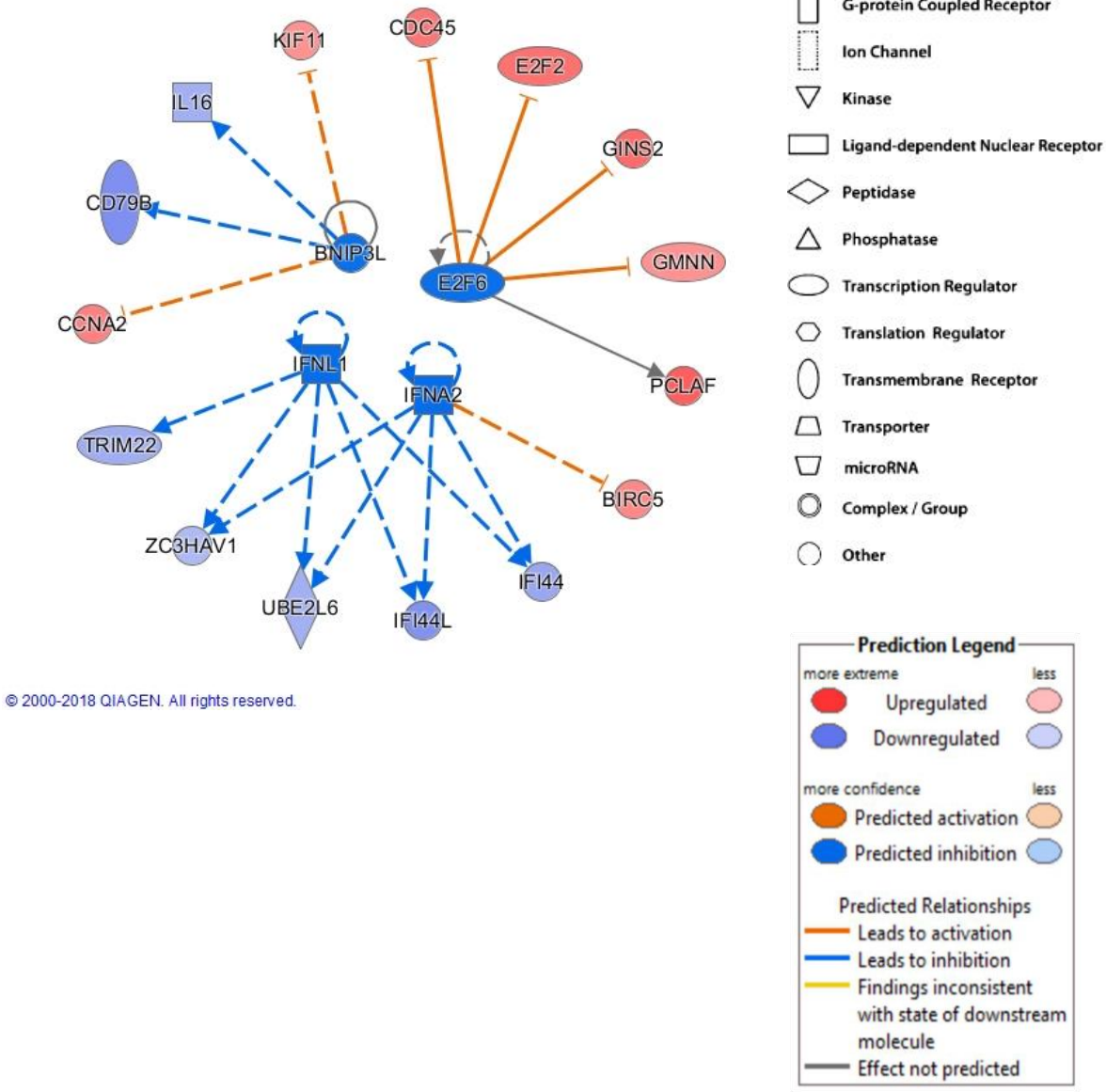
The upstream regulators predicted to be activated include the cytokine pathways for GM-CSF, IL-5 and IL-6. CD38, a B cell surface protein with ADP-ribosyl cyclase activity,

used as a marker for B cell subsets, including transitional B cells is identified as activated. In *Chapter 6* the frequency of the transitional B cell subset, $CD19^{+}CD24^{hi}$ $CD38^{hi}$ B cells, is indeed elevated in the RA group. The majority of connections between regulators and molecules can be seen to be indirect, with the exception of ESR1 and ATF4 (figure 3.3).

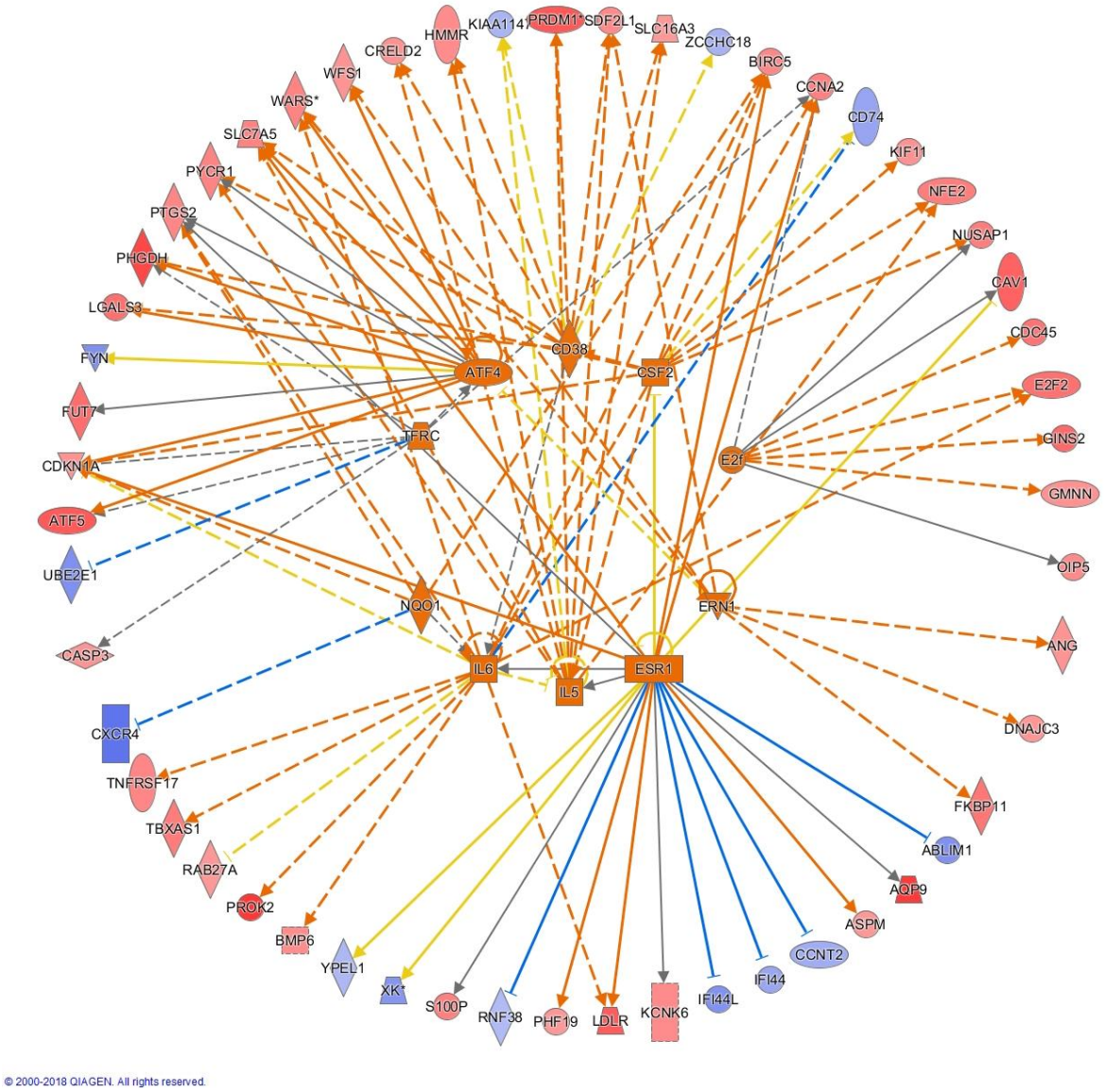
Upstream regulator	Molecule type	Name	z-score	p-value	Target molecules in the dataset
IFNL1	cytokine	Interferon lambda-1	-2.236	0.000734	IFI44, IFI44L, TRIM22, UBE2L6, ZC3HAV1
BNIP3L	other	BCL2 interacting protein 3 like	-2	0.00415	CCNA2, CD79B, IL16, KIF11
E2F6	transcription regulator	E2F transcription factor 6	-2	0.00058	CDC45, E2F2, GINS2, GMNN, PCLAF
IFNA2	cytokine	Interferon alpha 2	-2.207	0.00384	BIRC5, IFI44, IFI44L, UBE2L6, ZC3HAV1
E2f	transcription factor	E2f family	2	0.0000194	CAV1, CCNA2, CDC45, E2F2, GINS2, GMNN, NUSAP1
IL5	cytokine	IL-5	2.353	0.000292	CRELD2, HMMR, KIAA1147, NFE2, PRDM1, PYCR1
TFRC	transporter	Transferrin receptor 1	2	0.00044	ATF5, CASP3, CCNA2, CDKN1A, PHGDH, UBE2E1
NQO1	enzyme	NAD(P)H quinone dehydrogenase 1	2.169	0.0000639	BIRC5, CDKN1A, CXCR4, PTGS2, TP63
CSF2	cytokine	GMCSF	2.1	0.00749	BIRC5, CCNA2, CD74, CDKN1A, KIF11, NFE2, NUSAP1
CD38	enzyme	CD38	2.01	0.0000115	CRELD2, HMMR, KIAA1147, LGALS3, PRDM1, PYCR1
ERN1	enzyme	inositol-requiring enzyme 1	2.401	0.00245	ANG, DNAJC3, FKBP11, SDF2L1, WARS, WFS1
ATF4	transcription regulator	activating transcription factor 4	2.18	0.0000004	ATF5, CDKN1A, FUT7, FYN, LGALS3, PHGDH
IL6	cytokine	IL-6	2.451	0.000442	BIRC5, BMP6, CD74, CDKN1A, E2F2, LDLR, PROK2
ESR1	nuclear receptor	Estrogen receptor 1	2.486	0.0000037	ABLIM1, AQP9, ASPM, BIRC5, CAV1, CC

Table 3.7 Upstream regulators for the differentially expressed genes between RA and non-RA samples

IPA software was used to identify upstream regulators for the list of DEGs identified by comparing RA patients (n=59) to non-RA patients (n=131) without clinical variables added. Z score ≤ -2 predicts an inhibited state (shown in blue); ≥ 2 predicts an activated state for the regulator.



a)



b)

Figure 3.3 Network display of upstream regulators for the differentially expressed genes between RA and non-RA samples

IPA software was used to predict activated upstream regulators for the list of DEGs identified by comparing the CD19^{+} B cell transcriptome from RA patients (n=59) to non-RA patients (n=131) without clinical variables added. a) inhibited upstream regulators; b) activated upstream regulators. Upstream regulators are displayed centrally with lines connecting each to the molecules in the dataset which they are predicted to effect. Dashed lines represent indirect interactions and solid lines, direct interactions.

The E2F transcription factors appear twice in table 3.7 with contrasting activation states predicted. The E2F family consists of activators (E2F1, E2F2, E2F3a) and repressors (E2F3b to E2F8) of the cell cycle. The repressor E2F6 is predicted to be an inhibited upstream regulator of this dataset, while the activator E2F2 is upregulated. E2F2 is shown

in figure 3.3b connected to the activated upstream regulator E2f, the family of transcription factors rather than a specific member of it. This highlights the complexity of the regulators required for cellular homeostasis.

Molecular networks

In addition to the analyses above, the IPA software identifies molecular networks within the list of DEGs. The networks are assembled based on the connections between molecules apparent from the Ingenuity Knowledge Base. The presumption is that highly interconnected networks are more likely to represent a meaningful biological function. From the list of DEGs, the molecules which interact with each other, and those in the IPA knowledge base, are termed Network Eligible and act as the "seeds" for network generation. Molecules from the Ingenuity Knowledge Base may be added to fill or join areas lacking connectivity. The networks are annotated with functional categories.

The top 2 networks assembled from the list of DEGs generated from the comparison between RA and non-RA samples are shown in figures 3.4 and 3.5. The diseases and functions associated with network 1 were cellular development, cellular growth and proliferation, haematological system development and function. Network 2 is related to cell-to-cell signalling and interaction, skeletal and muscular system development and function and the cell cycle.

In network 1 it is notable that the connections are dense around the BCR complex and molecules downstream of this, including the ERK1/2 complex, which has been added from the Ingenuity Knowledge Base (figure 3.4).

Network 2 relates to cell to cell signalling and the cell cycle and the most dense connections surround the NF-kB complex which was not present in the dataset (figure 3.5). In the canonical pathway analysis the pathway termed 'NF-kB signalling' was downregulated but did not reach the Z score threshold set (Z score -1.342); the molecules from the DEG list in this pathway were BCL10, PELI1, IRS2, TRAF5 and TNFRSF17.

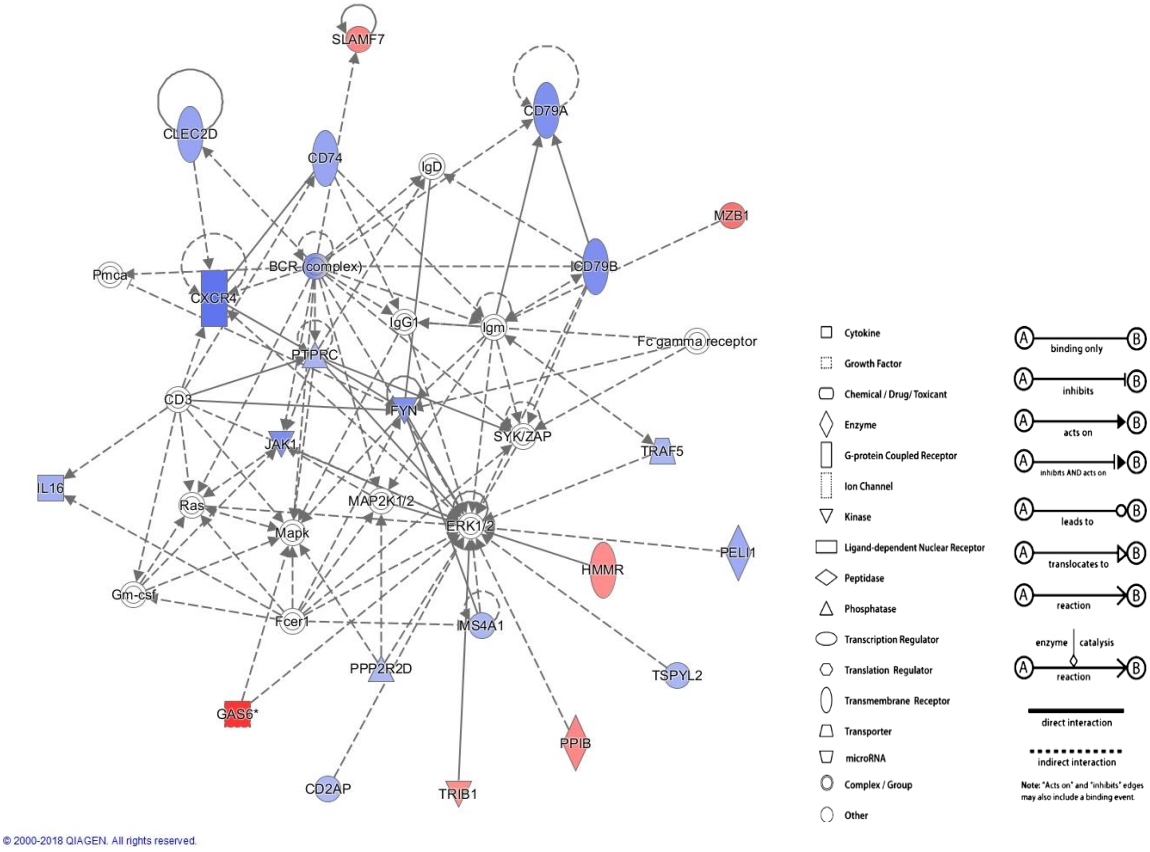


Figure 3.4 Molecular network 1 with functions related to cellular development, cellular growth and proliferation, haematological system development and function

Analysis using IPA software to identify molecular networks from the list of DEGs from the comparison of RA (n=59) to non-RA (n=131) samples. The IPA software assembled a network of 34 molecules in network 1, 21 of which were identified in the DEG list. Molecules not coloured are from the Ingenuity Knowledge Base and not present in the dataset. The shape of the gene represents the functional class of the gene product (see key).

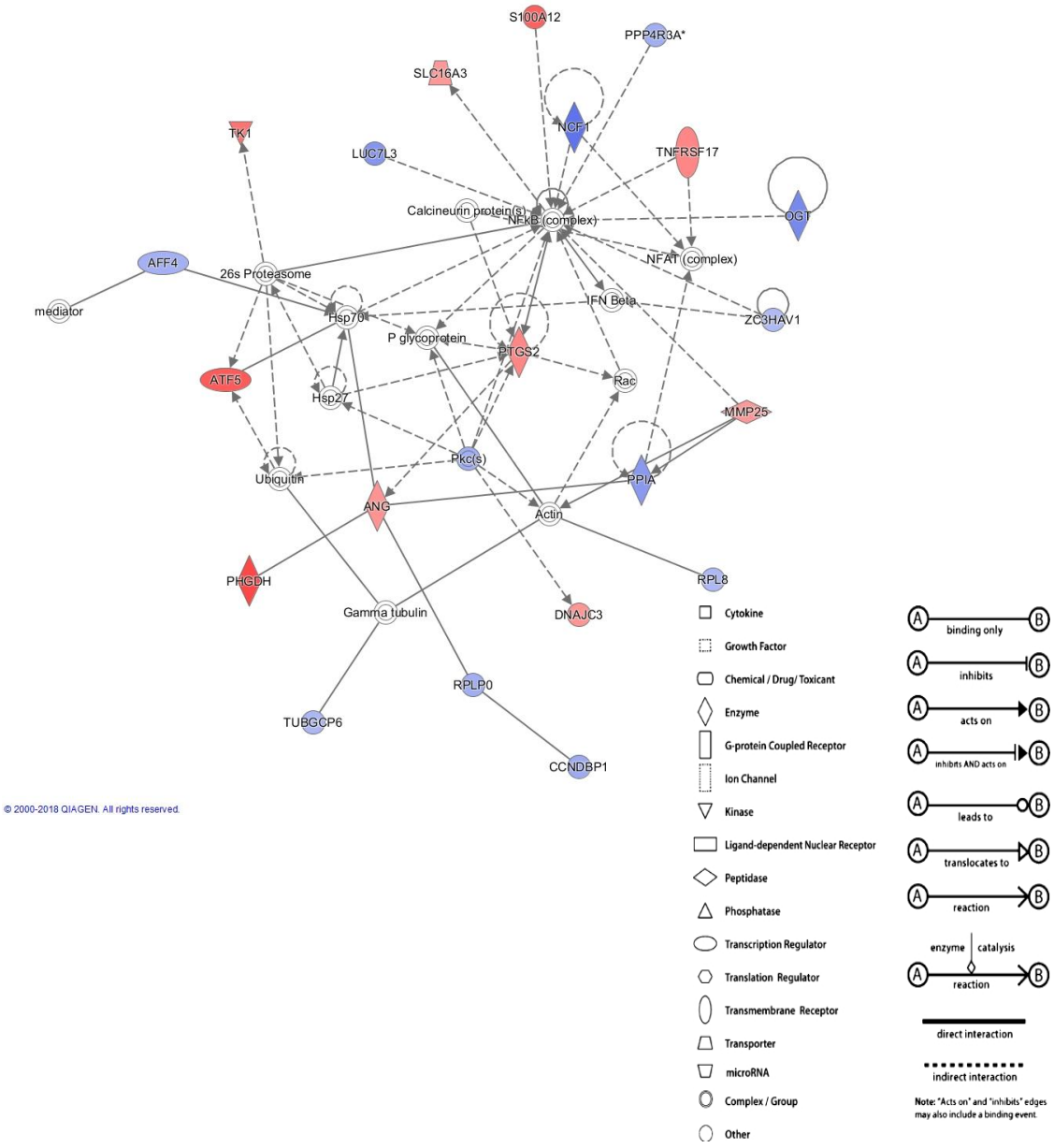


Figure 3.5 Molecular network 2 with functions related to cell-to-cell signalling and interaction, skeletal and muscular system development and function, cell cycle

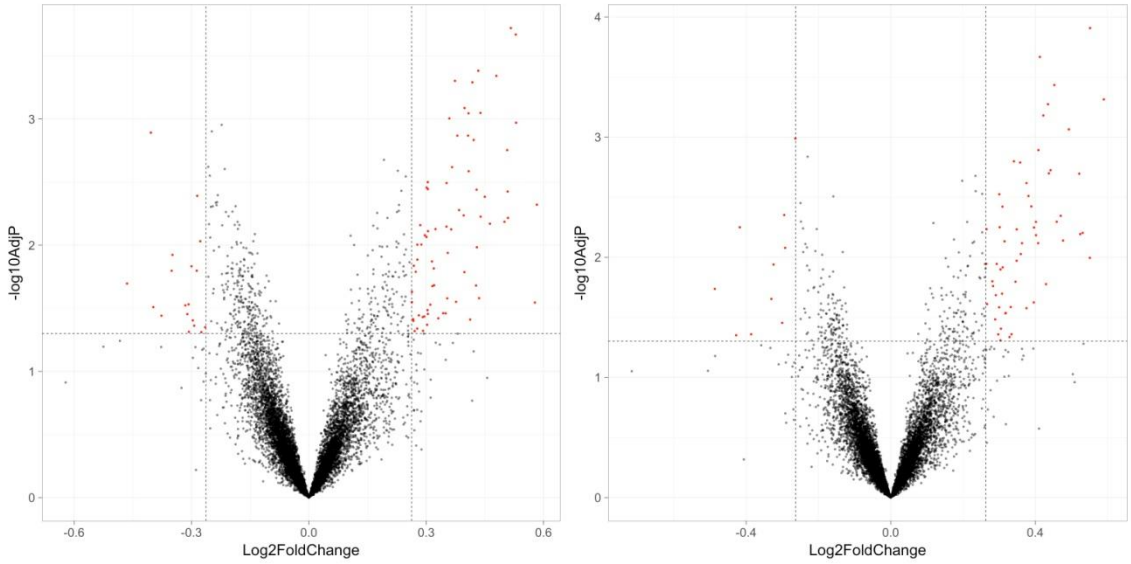
Analysis using IPA software to identify molecular networks from the list of DEGs from the comparison of RA (n=59) to non-RA (n=131) samples. The IPA software assembled a network of 35 molecules, 22 of which were identified in the list of DEGs. Molecules not coloured are from the Ingenuity Knowledge Base and not present in the dataset. The shape of the gene represents the functional class of the gene product (see key).

3.3.7 Differential gene expression between RA patients and non-RA patients using samples of known CD19$^+$ cell purity ≥90%

The analysis described in section *3.3.4* did not identify any differentially expressed probes which stood up to multiple test correction. I considered different methods for optimising the analysis. The first analysis, using all samples, had the benefit of maximising the number of samples available but the extent of contamination with other cells was unknown for the 50 samples where CD19$^+$ B cell purity was not examined.

The available data on cell purity does not demonstrate a difference in the proportion of contaminating cells between the sample groups (figure 3.1) and the proportion of contaminating cells may, indeed, not differ in the samples where the cell purity was not measured. However, contamination with other cell subsets may be a potential source of noise in the data and a potential explanation for the absence of a robust signal. I chose to repeat the analysis using only samples with a known CD19$^+$ B cell purity of ≥ 90%. The analysis was carried out using the same methods and analysis procedures.

The RA group (n=39) were compared to the non-RA group (n=101) and no DEGs were identified with the multiple test correction in place at FC ≥1.2. The multiple test correction was removed and 93 probes, representing 79 unique genes, were identified with a FC ≥1.2 (pval <0.05) (figure 3.6a). The analysis was repeated with the clinical covariates (age, CRP and ESR) added to the linear model and 68 probes, representing 58 unique genes, were identified as differentially expressed (figure 3.6b). There was an overlap of 49 genes between the two lists of DEGs.



a) *b)*

Figure 3.6 Differentially expressed genes identified between the CD19$^+$ B cell transcriptome of RA and non RA samples with known CD19$^+$B cell purity ≥90%

CD19$^+$ B cells were positively isolated from patients presenting to the early arthritis clinic, RNA extracted and transcriptome analysed using microarray technology. Samples from DMARD-naive RA patients (n=39) and non-RA patients (n=101) were compared. a) Volcano plot of RA against non-RA samples, 93 differentially expressed probes were identified (no multiple test correction in place). b) Volcano plot of RA against non-RA samples with covariates age, ESR and CRP added to the linear model, 68 differentially expressed probes were identified (no multiple test correction in place). Vertical, dotted lines denote FC 1.2. The x axis represents log2 of the fold change, y axis represents the –log10 adjusted p-value. Horizontal, dotted lines denote p-value 0.05. Red dots indicate probes which are differentially expressed between the comparator groups.

Ingenuity pathway analysis

The list of DEGs from RA versus non-RA samples without clinical covariates added to the linear model was analysed using IPA software. The canonical pathway analysis was uninformative, identifying 127 pathways none of which met the filtering criteria applied in the previous analysis described; 97 of these pathways listed just 1 molecule from the list of DEGs used and none of the remainder reached a Z score of ≤ -2 or ≥ +2. This may reflect the difficulties which arise when the length of the list of DEGs used is shorter than that recommend by the IPA supporting information.

The upstream regulator analysis identified two regulators with a predicted upregulated state: IL-4 (Z score +2.156, p-value 0.00122) and IL-6 (Z score +2.19, p-value 0.00538). In combination, IL-4 and IL-6 have connections to 12 molecules in the list of DEGs, the

interactions are indirect. The cytokines lead to activation of the downstream molecules, except in the case of the effects of IL-4 on FUT7 and HIPK2 where the program is unable to predict the expected direction of change; the molecules are upregulated in the dataset (figure 3.7).

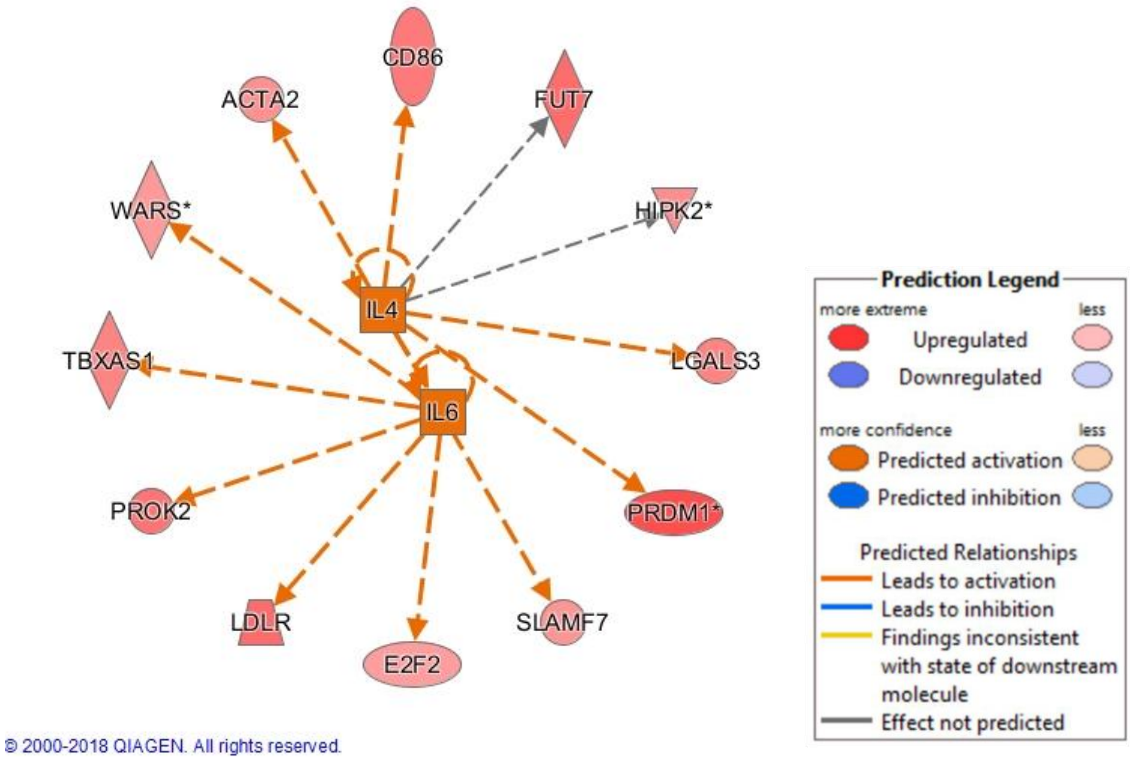


Figure 3.7 Network display of activated upstream regulators for the differentially expressed genes between RA and non RA samples with known CD19^{+}B cell purity ≥90%

*IPA software was used to predict activated upstream regulators for the list of DEGs identified by comparing the CD19^{+} B cell transcriptome from RA patients (n=39) to non-RA patients (n=101) without clinical variables added. Upstream regulators are displayed centrally with lines connecting each to the molecules in the dataset which they are predicted to effect. * indicates that multiple identifiers were found in the dataset to map to single gene. There were 3 probes for PRDM1, 2 for WARS and 2 for HIPK2 in the dataset.*

Of the 12 molecules shown here, 10 of these were present in the list of DEGs generated from the linear model with clinical covariates added (*E2F2* and *SLAMF7* were not present in the list with clinical covariates added). The 12 downstream molecules connected to IL4 and IL6 are shown in more detail in table 3.8. *PRDM1* encodes the transcriptional repressor, BLIMP1 which is important for plasma cell differentiation.

Gene	Gene name	Function
ACTA2	smooth muscle alpha (α)-2 actin	Actins are highly conserved proteins involved in cell motility, structure and intercellular signalling. Described as a smooth muscle actin. Expressed in lymph nodes.
CD86	CD86	Expressed on antigen presenting cells, ligand for CD28 and CTLA4 on T cells.
FUT7	fucosyltransferase 7	Golgi membrane protein involved in the synthesis of sialyl-Lewis X antigens which can bind selectins.
HIPK2	Homeodomain-interacting protein kinase 2	Serine/threonine nuclear kinase interacts with transcription factors such as p53. Can function as a corepressor or coactivator depending on the context.
LGALS3	Galectin-3	Member of a family of carbohydrate binding proteins, localises to the extracellular matrix, the cytoplasm and the nucleus, plays a role in numerous cellular functions including apoptosis, innate immunity, cell adhesion and T-cell regulation.
PRDM1	PR/SET domain 1	Protein also known as BLIMP-1, transcriptional repressor
SLAMF7	Signaling Lymphocyte Activation Molecule family member 7	Also known as CD319
E2F2	E2F transcription factor 2	Transcription factor, part of the E2F family which are involved in the control of cell cycle and action of tumour suppressor proteins
LDLR	Low density lipoprotein receptor	Cell surface protein involved in receptor-mediated endocytosis of specific ligands. Low density lipoprotein (LDL) is normally bound at the cell membrane and taken into the cell ending up in lysosomes where the protein is degraded and the cholesterol is made available to repress HMG CoA reductase the rate limiting step in cholesterol synthesis
PROK2	Prokineticin 2	Expressed in the uprachiasmatic nucleus (SCN) circadian clock. Biased expression in the bone marrow.

TBXAS1	thromboxane A synthase 1	Encodes a member of the cytochrome P450 enzyme family, on the basis of sequence similarity rather than functional similarity. This ER membrane protein catalyzes the conversion of prostglandin H2 to thromboxane A2
WARS	Tryptophanyl-tRNA synthetase	Tryptophanyl-tRNA synthetase (WARS) catalyzes the aminoacylation of tRNA(trp) with tryptophan and is induced by interferon. Tryptophanyl-tRNA synthetase belongs to the class I tRNA synthetase family.

Table 3.8 Gene names and function of the molecules with connections with the upstream regulators IL-4 and IL-6

Data from the NCBI Gene database[173].

3.3.8 A comparison of the DEGs from both analyses

The first analysis described in section *3.3.4*, using all samples, has the benefit of 50 extra samples but the extent of contamination with other cells is unknown. The first analysis (section *3.3.4*) has the advantage of an increased sample size but the second analysis (section *3.3.7*), with known purity, may reduce the interference of gene expression from other cell types. To combine the advantages of sample size and sample quality I identified a list of 52 individual genes which were present in the DEG lists from the two separate analyses.

The IPA software analysis was repeated using this list of 52 genes; the upstream regulator analysis identified IL-6 as an activated regulator of the dataset with an indirect effect on *LDLR*, *E2F2*, *PROK2*, *TBZAS1* and *WARS* (Z score 2.19, p-value 0.000843) (figure 3.8). Interferon gamma was the next potential upstream regulator identified; however, it did not reach the Z-score filtering criteria (Z score -1.915, p-value 0.0017). This pathway was inhibited affecting the genes: *AQP9*, *CAV1*, *IFI44*, *KCNMA1*, *NCALD*, *PRDM1*, *TBXAS1*. There were no significant canonical pathways or molecular networks identified by the IPA software.

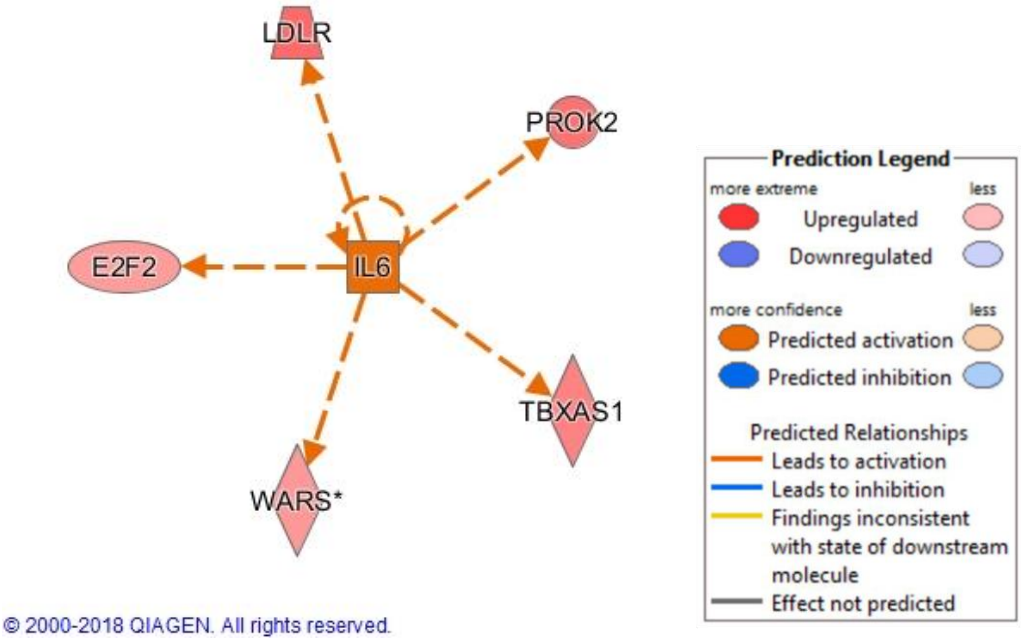


Figure 3.8 Network display for the activated upstream regulator IL-6

*IPA software was used to predict upstream regulators for the shared list of DEGs found in both the comparisons of the CD19$^+$ B cell transcriptome from RA patients to non-RA patients using i) all samples and ii) samples with known CD19$^+$ B cell purity ≥ 90%. IL-6 was the sole upstream regulator identified. IL-6 is shown centrally with lines connecting to each of the molecules in the dataset which it is predicted to effect. * indicates that multiple identifiers were found in the dataset to map to single gene.*

3.3.9 Validation using top 25% of probes from comparison of RA against non-RA with clinical variables included.

I originally hypothesised that examining the differences between the CD19$^+$ B cell transcriptome from patients presenting with RA and a disease control group would reveal potential diagnostic biomarkers and these biomarkers could be used to identify patients with RA within an undifferentiated arthritis (UA) group. DEGs identified were to be validated in the UA cohort. However, no DEGs were identified with multiple test correction in place, limiting this approach.

I took the top 25% of DEGs, 26 probes, from the comparison of RA *versus* non-RA samples (with clinical covariates included in the linear model) and looked for differences in the expression of the 26 probes in the UA cohort who, over time, acquired a definitive diagnosis of RA *versus* another type of arthritis. There were 14 samples in the RA group and 28 samples in the non RA group for this "validation" cohort. In the non-RA group 10

had osteoarthritis or other non-inflammatory conditions and the remainder another type of inflammatory arthritis, 12 of which were reactive arthritis.

The same methods were used as for the other expression analyses in this chapter and no differences were identified between the two groups in the validation cohort when the subset of 26 selected probes were examined.

3.3.10 Differential gene expression between RA patients and non-RA based on current working diagnosis

The comparisons described so far in this chapter divided the samples based on the baseline diagnosis made at the patient's first consultant appointment. At baseline, 50 patients were labelled as having an undifferentiated arthritis. I did not include this group in the disease control, non-RA group, for the analyses described as, within this group, there were patients who were subsequently diagnosed with RA and I had planned to use this UA group as a validation cohort.

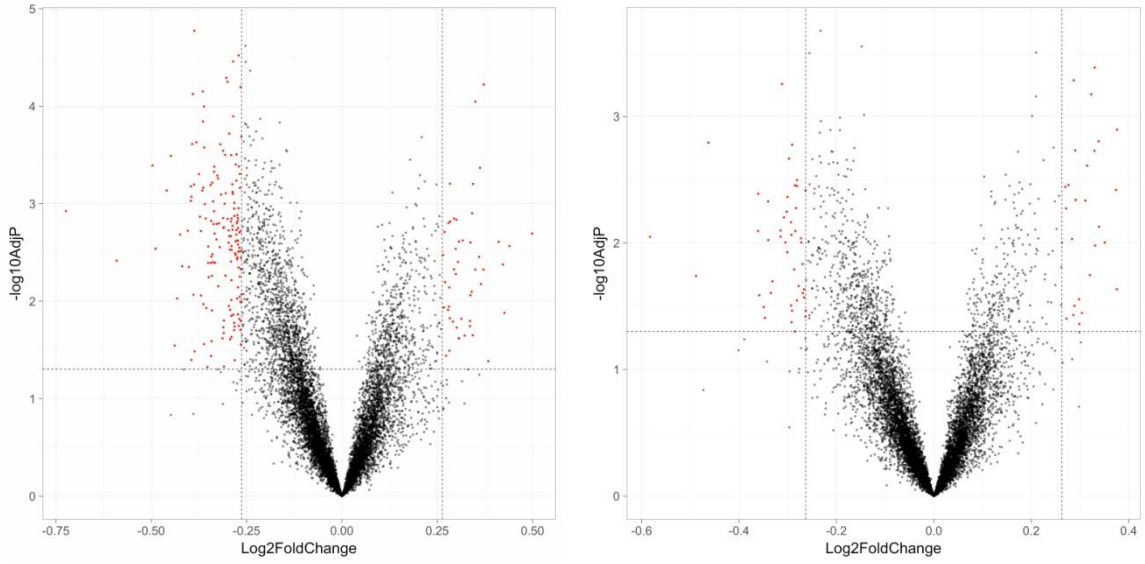
However, the analyses based on baseline diagnosis did not reveal clear potential biomarkers for the diagnosis of RA. The possible reasons that this was not the case are discussed later but one reason may be that sample groups were not large enough to detect subtle, meaningful differences which were robust to multiple test correction.

I carried out a third analysis using the most up to date, current diagnosis from clinic follow up letters to increase the number of samples. The samples were then divided as before into RA and non-RA, increasing the sample sizes to 73 RA and 159 non-RA samples. Those who remained without a definitive diagnosis were excluded from the analysis. The samples were not restricted based on cell purity in order to further optimise group sizes; there were no differences in the cell purity between the RA, non-RA and original UA groups (figure 3.1).

No DEGs were identified in this comparison with the multiple test correction in place, and when the multiple test correction was removed 213 differentially expressed probes, representing 201 unique genes, were identified with a FC $\geq$1.2 (p-value $<$0.05) (figure

3.9a). 159 of the genes from this list were also identified in the original analysis using baseline diagnosis (section *3.3.4*).

71 differentially expressed probes were identified when clinical covariates (age, ESR and CRP were added to the linear model (figure 3.9b).



a) *b)*

Figure 3.9 Differentially expressed genes identified between the CD19$^+$ B cell transcriptome of RA and non-RA samples based on current diagnosis

CD19$^+$ B cells were positively isolated from patients presenting to the early arthritis clinic, RNA extracted and transcriptome analysed using microarray technology. Samples from DMARD-naive RA patients (n=73) and non-RA patients (n=159) were compared. a) Volcano plot of RA against non-RA samples, 213 differentially expressed probes were identified (no multiple test correction in place). b) Volcano plot of RA against non-RA samples with covariates age, ESR and CRP added to the linear model, 71 differentially expressed probes were identified (no multiple test correction in place). Vertical, dotted lines denote FC 1.2. The x axis represents log2 of the fold change, y axis represents the –log$_{10}$ adjusted p-value. Horizontal, dotted lines denote p-value 0.05. Red dots indicate probes which are differentially expressed between the comparator groups.

Ingenuity pathway analysis

The IPA canonical pathway analysis of the list 213 DEGs identified 8 canonical pathways, all downregulated in the RA group, which met the filtering criteria (table 3.9). The top two pathways in this comparison were leucocyte extravasation signalling and integrin signalling. 5 of the 8 canonical pathways were also identified in the IPA analysis of the DEG list for the comparison using the diagnoses at baseline (section 3.3.6). In the

case of the leucocyte extravasation signalling pathway, this was also identified in the analysis using baseline diagnoses, with 6 molecules in the pathway, however the pathway did not meet the Z score criteria (Z score -1.633).

Canonical Pathways	**-log(p-value)**	**z-score**	**Genes**
Leukocyte Extravasation Signaling	1.93	-2.449	*NCF1*, *CXCR4*, *CD44*, *ABL1*, *PRKCE*, *IRS2*
Integrin Signaling	3.1	-2.236	*NCK2*, *FYN*, *TSPAN3*, *MPRIP*, *RALB*, *ABL1*, *IRS2*, *NEDD9*
p70S6K Signaling	2.18	-2.236	*JAK1*, *CD79B*, *PRKCE*, *CD79A*, *IRS2*
EIF2 Signaling	1.84	-2.236	*RPL32*, *RPL8*, *EIF4G2*, *ATF5*, *IRS2*, *RPLP0*
B Cell Receptor Signaling	1.55	-2.236	*CD79B*, *BCL10*, *ABL1*, *CD79A*, *IRS2*
Tec Kinase Signaling	1.74	-2	*FYN*, *JAK1*, *GNG2*, *PRKCE*, *IRS2*
Ephrin Receptor Signaling	1.69	-2	*NCK2*, *FYN*, *CXCR4*, *GNG2*, *ABL1*
Role of NFAT in Regulation of the Immune Response	1.59	-2	*FYN*, *CD79B*, *GNG2*, *CD79A*, *IRS2*
PI3K Signaling in B Lymphocytes	3.77	-1.633	*FYN*, *CD79B*, *ATF5*, *BCL10*, *ABL1*, *CD79A*, *IRS2*
Phospholipase C Signaling	2.22	-1.633	*FYN*, *MPRIP*, *CD79B*, *GNG2*, *RALB*, *PRKCE*, *CD79A*

Table 3.9 Canonical pathways identified from the list of differentially expressed genes between RA samples and non-RA samples based on current diagnosis

IPA software was used to analyse the list of DEGs identified by comparing the CD19^{+} B cell transcriptome from RA patients (n=73) to non-RA patients (n=159) without clinical variables added. The pathways shown have a minimum of 5 molecules from the pathway identified in the list of DEGs, Z score ≤ -2 or ≥ +2, p-val <0.05. In red pathways with Z scores which miss the filtering criteria

The three, additional, downregulated canonical pathways which emerge through this comparison: leucocyte extravasation signalling, integrin signalling and ephrin receptor signalling; relate to cell to cell communication and cell migration. Leucocyte extravasation signalling describes the process of leucocyte migration from blood to

tissues during inflammation. Integrins are cell surface glycoproteins involved in cell to cell and cell to extracellular matrix interactions. These interactions are crucial for the structural changes involved in cell migration and signal transduction which can affect cell survival, differentiation and proliferation. Ephrin receptors are tyrosine kinases; the receptors and their corresponding ephrin ligands are membrane-bound proteins that require direct cell-cell interactions for ephrin receptor activation.

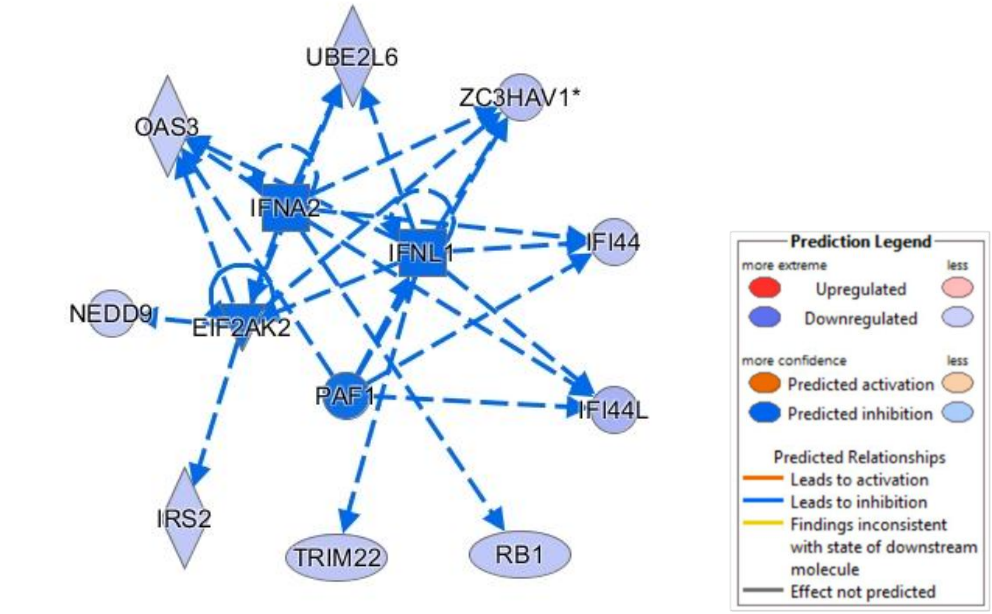
BCR signalling is again shown to be downregulated and the two major signalling pathways (PLC and PI3K) downstream of the receptor are the next pathways identified by IPA but do not meet the Z-score threshold.

IPA software identified 7 upstream regulators for this dataset. It is notable that the upstream regulator seen consistently across the analyses is IL-6 (table 3.10). The network displays for the upstream regulators demonstrate that the effects of the inhibited upstream regulators (*PAF1*, *IFNL1*, *IFNA2* and *EIF2AK*) are indirect on their target molecules in the dataset (figure 3.10a). The effects of IL-6 are also indirect on its target molecules, while the effects of ESR1 are predicted to be direct (figure 3.10b).

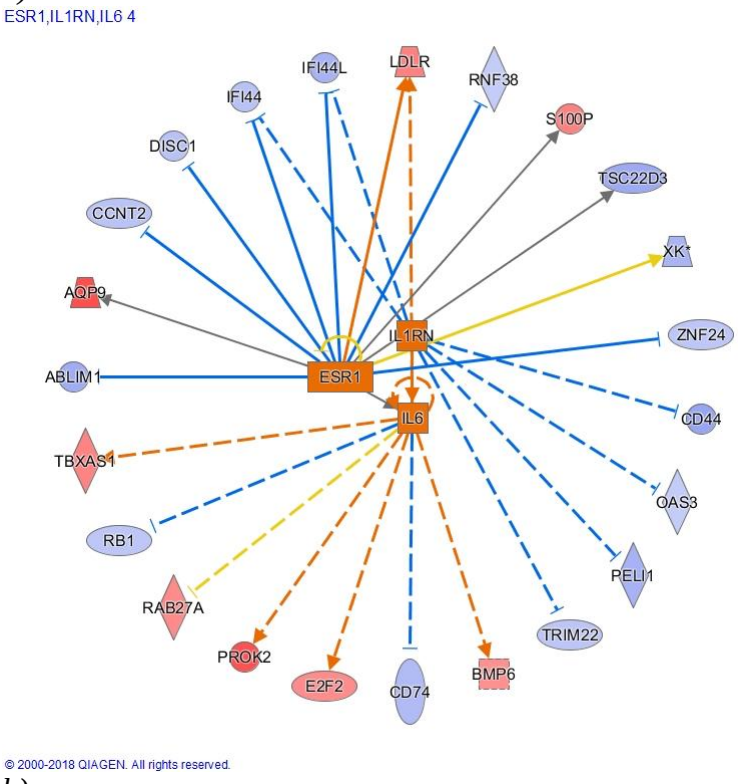
Upstream regulator	Molecule type	Name	z-score	p-value	Target genes in dataset
PAF1	other	Polymerase associated factor 1	-2	0.00106	*IFI44, IFI44L, OAS3, ZC3HAV1*
IFNL1	cytokine	Interferon lambda 1	-2.449	0.0000185	*IFI44, IFI44L, OAS3, TRIM22, UBE2L6, ZC3HAV1*
IFNA2	cytokine	Interferon alpha 2	-2.403	0.000152	*IFI44, IFI44L, OAS3, RB1, UBE2L6, ZC3HAV1*
EIF2AK2	kinase	Interferon-induced, dsRNA-activated protein kinase	-2.219	0.000601	*IRS2, NEDD9, OAS3, UBE2L6, ZC3HAV1*
IL1RN	cytokine	IL-1 receptor antagonist	2.433	0.000219	*CD44, IFI44, IFI44L, OAS3, PELI1, TRIM22*
IL6	cytokine	IL-6	2.158	0.00923	*BMP6, CD74, E2F2, LDLR, PROK2, RAB27A, RB1, TBXAS1*
ESR1	ligand-dependent nuclear receptor	Estrogen receptor	2.333	0.00136	*ABLIM1, AQP9, CCNT2, DISC1, IFI44, IFI44L, LDLR, RNF38, S100P, TSC22D3, XK, ZNF24*

Table 3.10 Upstream regulators for the differentially expressed genes between RA and non-RA samples based on up to date diagnosis

IPA software was used to identify upstream regulators for the list of DEGs identified by comparing the CD19^{+} B cell transcriptome from RA patients (n=73) to non-RA patients (n=159) without clinical variables added. Z score $\leq$ -2 predicts an inhibited state (shown in blue) and $\geq$ +2 predicts an activated state for the regulator.



a)



b)

Figure 3.10 Network display of upstream regulators for the differentially expressed genes between RA and non-RA samples based on current diagnosis

*IPA software was used to predict activated upstream regulators for the list of DEGs identified by comparing the CD19$^+$ B cell transcriptome from RA patients (n=73) to non-RA patients (n=159) without clinical variables added. Upstream regulators are displayed centrally with lines connecting each to the molecules in the dataset which they are predicted to effect. * indicates that multiple identifiers were found in the dataset to map to single gene.*

3.3.11 IL-6

MSD data were available for a subset of the microarray cohort (data from Dr Amy Anderson) and serum IL-6 levels were compared between the 2 sample groups: RA and non-RA in the three datasets: all samples (using baseline diagnosis), samples with $CD19^+$ B cell purity ≥ 90% (using baseline diagnosis) and all samples using current diagnosis (figure 3.11). This consistently shows that serum IL-6 levels are higher in the RA than non-RA group for all three comparisons.

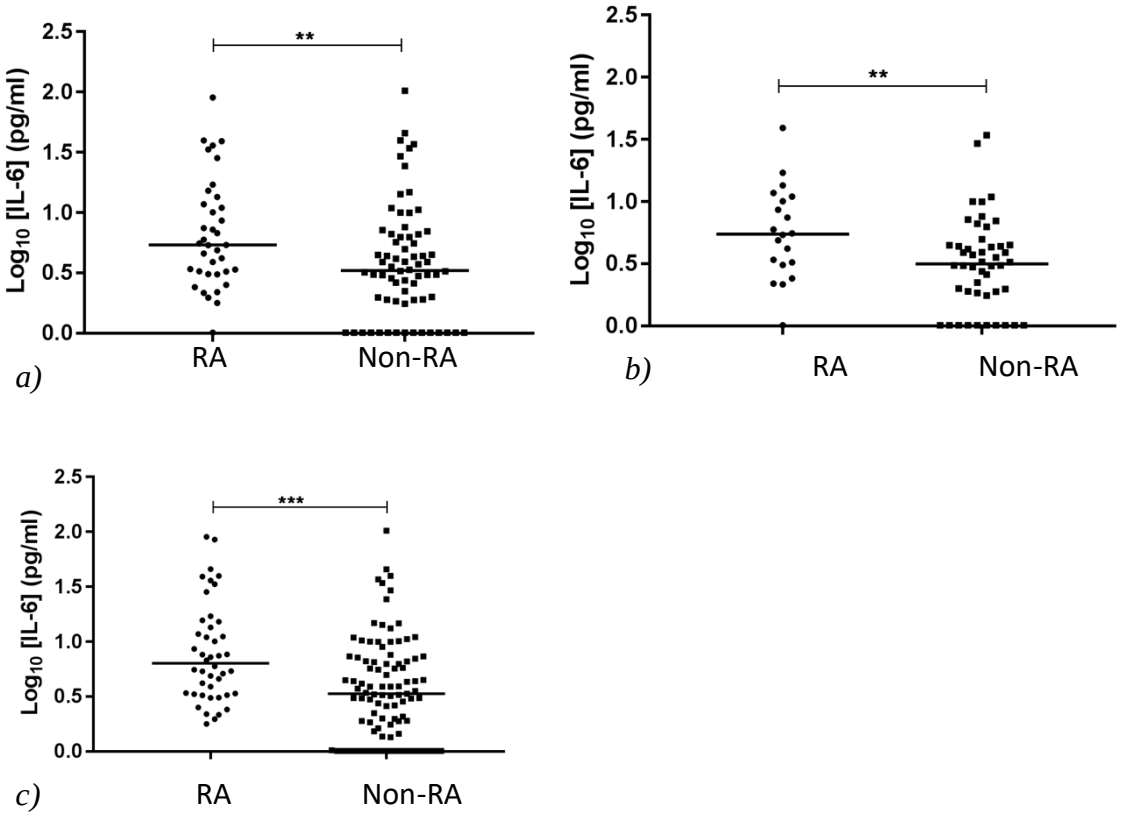


Figure 3.11 Log_{10} IL-6 serum levels by diagnostic group

a) Samples grouped by baseline diagnosis, RA n=37, non-RA n=70, pval 0.0081 b) samples with $CD19^+$ B cell purity ≥ 90% grouped by baseline diagnosis, RA n=20, non-RA n=48, pval 0.0086 c) Samples grouped by current diagnosis RA n=44, non-RA n=97 pval 0.0001. Mann Whitney tests.

3.3.12 Gene set enrichment analysis (GSEA)

GSEA interrogates data at the level of sets of genes, grouped based on prior experimental knowledge from the literature, rather than looking at individual genes. Cellular processes would usually affect a group of molecules, rather than in isolation. The study of gene sets

may allow the detection of changes which are subtle but co-ordinated; overcoming the challenges of vast microarray data where differences in expression can be modest when compared to the background 'noise', and interpretation of any list of DEGs can be challenging to identify a meaningful unifying, biological explanation for the change

The web-based GSEA from the Broad Institute was used to look for co-ordinated changes in gene sets between two user defined phenotypes [160]. The gene sets are divided into 8 'collections' and the Gene Ontology collection of annotated sets was used.

The phenotypes used were RA (n=59) versus non-RA samples (n=131) and the program ranked the genes in the data based on the differential expression between the two phenotypes. An enrichment score (ES) was first calculated to reflect the degree that any gene set being examined was over-represented at the extremes (the top or bottom) of the ranked gene list, the statistical significance assessed next by a permutation test and lastly adjusted for multiple hypothesis testing. The ES was normalised for each gene set to account for the size of the set, providing a normalised enrichment score (NES). An FDR cut off was applied to control for the proportion of false positives, the estimated probability that an NES represents a false positive finding.

4187 sets from the Gene Ontology collection were tested and 37 pathways were differentially expressed between the two groups (FDR <25%). To render a more intuitive summary of functional information, pathways were grouped by function (table 3.11).

The relative expression of the 37 gene sets identified is higher in the disease control group, non-RA, than in the RA group, hence the negative values for the enrichment scores shown. Given that the rheumatological conditions represented in the disease control group are not considered to be B cell mediated, these results may indicate pathway downregulation in the RA group.

Category	Gene ontology set	ES	NES	FDR
DNA repair	Global genome nucleotide excision repair	-0.5607982	-1.6132796	0.23832387
	Nucleotide excision repair DNA duplex unwinding	-0.61677194	-1.5758291	0.24963197
Transcription, translation and silencing	DNA directed RNA polymerase II holoenzyme	-0.5048356	-1.6432542	0.2334442
	Transcriptional repressor complex	-0.55602115	-1.7730718	0.23473078
	dsRNA fragmentation	-0.64933866	-1.6547841	0.24210277
	Posttranscriptional gene silencing DNA directed RNA polymerase II holoenzyme	-0.51306003	-1.6551133	0.24704707
Nucleosome	Nucleosome binding	-0.5144523	-1.6157382	0.24781571
Histone modification	Histone methyltransferase complex	-0.5257227	-1.613204	0.2348409
	Histone deubiquitination	-0.6148687	-1.614706	0.24252507
	Histone methyltransferase activity	-0.53560984	-1.6328591	0.24889699
	SIN3 type complex	-0.59872967	-1.5815475	0.24807833
	MLL1_2 complex	-0.6282426	-1.5775148	0.24864604
RNA processing	snRNA binding	-0.562579	-1.6486899	0.2403679
	mRNA processing	-0.4083914	-1.6045289	0.24474731
	mRNA metabolic process	-0.41401324	-1.6226012	0.24723203
	RNA splicing via transestertification reactions	-0.3926336	-1.5797857	0.24923289
Post translational modifications	Acetyltransferase complex CUL3 rung ubiquitin ligase complex	-0.6357071	-1.9409733	0.22092754
	Lysine N-methyltransferase activity	-0.5501833	-1.6126041	0.23269609

	CUL3 ring ubiquitin ligase complex	-0.5345131	-1.644695	0.2347206
	Cullin ring ubiquitin ligase complex	-0.43626913	-1.6089464	0.23777463
	Peptidyl lysine trimethylation	-0.6279962	-1.6467234	0.24000877
	N-methyltransferase activity	-0.4840115	-1.6258838	0.24353907
	Transferase complex transferring phosphorus containing groups	-0.45952913	-1.6154255	0.24463683
	Deacetylase activity	-0.5190539	-1.6170225	0.24895547
	Peptide N acetyltransferase activity	-0.522617	-1.594338	0.24906248
	Negative regulation of dephosphorylation	-0.4655819	-1.5970217	0.24921104
	Protein acylation	-0.45356375	-1.6290699	0.24925643
Protein/lipid processing	Bloc complex	-0.614468	-1.6040541	0.24233998
	Golgi organisation	-0.47892088	-1.5749071	0.24921449
Cytoskeleton	Microtubule nucleation	-0.69583064	-1.6451253	0.23860565
	Lamellipodium organisation	-0.5539983	-1.6530507	0.24102226
Cell signalling	Inositol phosphate mediated signalling	-0.5248304	-1.6280522	0.24268949
	Protein kinase A regulatory subunit binding	-0.59867424	-1.5967791	0.24643724
	Phosphatidylinositol 3 kinase signalling	-0.6323521	-1.6199005	0.24984726
Growth factor response	PDGF signalling pathway	-0.54785293	-1.6505347	0.24134766
	Response to nerve growth factor	-0.46626347	-1.6287607	0.24545705
B cell differentiation	B cell differentiation pathway	-0.44574007	-1.6144383	0.2393389

Table 3.11 Gene sets differentially expressed between RA and non-RA samples using Gene set enrichment analysis

$CD19^+$ B cell gene expression data for RA (n=59) and non RA (n=131) were analysed using the GSEA program from the Broad Institute. 37 gene sets from the Gene Ontology collection of gene sets were differentially expressed in the RA samples compared to non-RA samples (FDR <0.25). The negative values for the enrichment scores shows that the gene sets are upregulated in the non-RA groups, or downregulated in the RA group. Gene sets are grouped by cellular function. ES; enrichment score, NES; normalised enrichment score, FDR, false discovery rate.

The downregulated pathways in *table 3.11* include basic cellular functions, in particular the generation and modification of proteins. This suggesting an overall reduction in cellular activity. B cell differentiation and signalling pathways downstream of the BCR are also downregulated which would be in agreement with the IPA canonical pathway analysis described. The B cell differentiation gene set is described in the annotation as 'the process in which a precursor cell type acquires the specialised features of a B cell.' This is an interesting finding and downregulation of this pathway in the RA group may reflect a change in the B cells of RA patients to promote their differentiation into antibody producing cells

No gene sets were significantly enriched at an FDR of <25% using the data solely from samples with known $CD19^+$ B cell purity ≥ 90%.

3.4 Discussion

My original hypothesis stated that DEGs would be identified between the RA group and those with other forms of arthritis. This was not the case with a multiple test correction in place, however, a thorough examination of the data using different comparisons and analysis programs has shown themes which emerge from the data: an alteration in BCR signalling, changes in B cell differentiation, the role of IL-6 and the potential role for pathways related to the oestrogen receptor in B cells in early RA.

IPA, in particular the canonical pathways and molecular network analyses, and GSEA concur that BCR signalling and its downstream cell signalling pathways are downregulated in the RA group compared to the non-RA group. When interpreting the results this assumes that the disease control group exhibits a normal level of gene expression for these pathways. Only 3 of 131 non-RA samples used in the first analysis had a baseline diagnosis of lupus or CTD associated conditions, diseases which are B cell driven. These samples were not excluded as the project was aimed at identifying a diagnostic gene signature.

As discussed in *Chapter 1*, the level of BCR signalling is critical in determining the effect on the cell and is crucial to B cell development, proliferation and activation. Sustained

BCR signalling is required for cell survival in the periphery through induction of NFκB and PI3K dependent signalling and an intermediate signal may be optimal for survival. The removal of autoreactive B cells relies on a strong BCR signal to promote apoptosis, so, in RA, it could be that the BCR signal is constitutively downregulated, and binding of the BCR to self-antigens does not lead to the removal of autoreactive cells. The downregulation of the signalling pathways may therefore lead to tolerance breakdown and the initiation of autoimmune disease, with the sustained presence of autoreactive cells in the periphery.

The *PTPN22* disease RA risk allele alters BCR signalling. It is believed to be a gain of function mutation in humans; an increase in the encoded protein tyrosine phosphatase activity impairs B cell signal transduction, as measured by phosphorylation of Syk, PLCγ2 and Akt in response to activation by soluble anti-IgM in heterozygotes and this is more marked in homozygotes[126]. A dysregulated, impaired BCR signal may affect B cell selection and maturation, resulting in disease initiation and maintenance by allowing the survival of autoreactive cells. This would fit with the increase in polyreactive, newly emigrant cells from the bone marrow in healthy donors carrying the risk allele and the observation that transitional B cells (although the immunophenotypes used differ) are increased in the RA group in this cohort (C*hapter 6*)[127]. Given that not all RA patients carry this risk allele there may be different potential routes to downregulation of BCR signalling including those related to co-receptors of the BCR.

The GSEA results suggest a downregulation of the fundamental processes of transcription, translation and protein modification alongside decreased B cell differentiation in the RA group. This raises the possibility that, in early RA, the B cells are in an anergic or exhausted state similar to that described in T cells. A discrete, exhausted subset of T cells have been identified which are thought to be induced by a combination of continuous antigen exposure, inflammatory cytokines and the local microenvironment. They represent a distinct lineage expressing altered transcription factors and inhibitory receptors, leading to cells with impaired effector functions and a limited proliferative potential[174, 175].

The relative downregulation of B cell differentiation in the RA group may reflect the transition into a terminally differentiated phase to produce antibody secreting cells. This requires fundamental changes in gene expression and the silencing of the transcription factors which control B cell identity, allowing the focus to shift to the production of large amounts of immunoglobulin[44]. This change may also explain the relative change in the cellular functions identified by the GSEA analysis. The shift to antibody secreting cells is supported by the observations regarding PRDM1, EIF2 signalling and CD38. *PRDM1*, a widely recognized plasma cell related gene is upregulated in the list of DEGs (3 probes for the gene are upregulated). The EIF2 signalling pathway was shown to be downregulated in the RA group in the IPA analysis, which may be secondary to the unfolded protein response (UPR) in antibody producing cells. CD38 was identified as an upstream regulator in the RA group and daratumumab (a monoclonal antibody against CD38) is currently under investigation as a potential treatment for RA and SLE, acting by depleting plasmablasts/plasma cells[176].

The RA peripheral blood $CD19^+$ B cells examined here may, therefore, be at different stages than those in non-RA as they differentiate towards antibody producing cells. The findings, therefore, hint at potential changes in the B cell compartment of RA patients. The parallel flow data for the cohort described in *Chapter 6* did not, however, identify a difference in the proportion of plasmablasts between the groups and plasma cell levels were not assessed.

The IPA analysis identified IL-6 as a potential upstream regulator to explain the changes seen in gene expression, a cytokine critical to plasma cell differentiation and the humoral response. Serum IL-6 was consistently higher in the RA group and the monoclonal antibody tocilizumab is an effective treatment for RA. Interestingly, B cell depletion has been shown to alleviate central nervous system autoimmunity through ablation of an IL-6 producing pathogenic subset of B cells, which potentially acts by promoting Th17 responses in mouse models with experimental autoimmune encephalomyelitis [177]. The IL-6 dependent mechanism was independent of antibody production and the subset was not required for disease initiation but shown to exacerbate ongoing disease. The IPA results in this chapter show an indirect effect of IL-6. We cannot assume that the B cells are producing IL-6 themselves, simply that IL-6 is potentially influencing the gene expression changes which are seen in the RA group. The pathogenic role of B cells in RA

may be, in part, related to cytokine production rather than the traditional focus of humoral immunity. One mechanism may be breakdown in tolerance occurring as a defect in an IL-10 dependent regulatory function of B cells and this change is maintained by an IL-6 producing B cell subset[52].

The oestrogen receptor, ESR1, is identified as an activated upstream regulator in the RA group in the analyses using both baseline and outcome diagnoses to determine sample groups; despite no difference in the gender balance between the groups. The role of oestrogens in RA development remains poorly understood particularly in disease development but it has been suggested that oestradiol promotes autoimmune disease development via an action on B cells, including an increase in the survival of autoreactive B cells and, interestingly, RA male patients have higher levels of oestradiol than healthy controls[178, 179].

Type I interferons are an area of particular interest in rheumatic diseases. An interferon gene signature (IGS) is a feature of RA development in subset of patients[145]. It initially appears counter-intuitive that my data are suggestive of inhibition of this pathway in circulating early RA B cells. However, it should be remembered that studies on the type I IGS have generally used whole blood or PBMC samples, and a study in patients in SLE demonstrated marked differences in IGS expression between individual cell subsets[180]. The literature used in the IPA database to identify relevant pathways is likely to be based on literature related to mixed cell subsets.

The absence of a robust list of DEGs in this study may be related to technical considerations or the heterogeneity of the cell populations, samples, and sample groups examined.

The inter-individual variability in the $CD19^+$ B cell transcriptome was highlighted in the small pilot study I carried out prior to this project. Comparing the $CD19^+$ B cell profiles of 12 healthy controls and 11 RA patients with established RA, the overall variability in gene expression in $CD19^+$ B cells, independent of disease state, was notable. There are many potential sources of this overall variability including genetic factors, recent infections or immunisations, age and sex. I attempted to maximise the opportunity to

identify significant changes in the transcriptome by using the current diagnoses of patients in the third analysis described, however, a clear DEG list did not emerge.

In this study, I attempted to reduce the issues that arise from the study of mixed cell populations; nonetheless, the $CD19^+$ B cell profile analysed is comprised of several different B cell subsets including regulatory, memory, and naive cells for example. If one particular subset is dysregulated in RA, such as regulatory B cells, then any change in the dysregulated subset may be masked by the transcriptional profiles of the other subsets. In addition, the relative proportions of immune cell subsets has a heritable component and is also strongly influenced by additional non-heritable factors in healthy individuals such as age and the environment. For example transitional B cell proportions decrease linearly with age, adding increased complexity to the data beyond the contribution of disease[181, 182]. A further possible explanation for the lack of a diagnostic signal may be that the $CD19^+$ B cell isolation used did not include plasma cells. Whilst plasma cells are not depleted by rituximab, which is an effective treatment for RA, this subset may still provide insights into disease pathogenesis. An alternative approach would be to sort individual B cell subsets by flow cytometry prior to downstream transcriptomic work; this would be reliant on the prior identification of a subset of interest, however, and would likely be restricted in terms of cell numbers.

The RA group comprises a heterogeneous population, with differing levels of disease activity and autoantibody status. In addition, given that autoantibodies can be identified in the serum of patients years before the development of clinical symptoms, examining patient samples once symptoms have developed could miss the window during which pathogenic changes are present in the $CD19^+$ B cell transcriptome [67]. The downregulation of the canonical pathway related to leucocyte extravasation signalling and integrin signalling may indicate the migration of B cells critical to pathogenesis to sites of inflammation prior to presentation to secondary care[183].

Beyond the heterogeneity in the patient cohort and B cell subsets there are technical considerations which may have affected the outcome of this study. The number of patients recruited inevitably meant that the samples were processed in separate batches for different stages of the experiment. Examination of the data demonstrated a batch effect related to conversion of RNA to cDNA and the generation of biotin-labelled cRNA

(*Methods 2.5.2*), which was carried out in 3 batches. The 'batch effect' was incorporated into the linear model for analysis and is a common issue for experiments on this scale which is difficult to overcome outside large facilities. The second major technical consideration is cell contamination, the data for which are not available for 50 samples. The cell purity in these cases was not measured to maximize the number of cells available for analysis.

This dataset of 240 $CD19^+$ B cell transcriptomes represents a great potential resource but the interpretation of the data relies on the analysis programs available, their maintenance and the curated literature used to underpin them. The literature used to build the associations identified are from a range of tissues and cell subsets which may not always be relevant to the dataset under analysis and this must be borne in mind when interpreting the results. The advantage is that, with this caveat, the results can be hypothesis generating and novel.

There are multiple factors potentially affecting the gene expression profile in $CD19^+$ B cells beyond disease and so, in *Chapter 4*, I will go on to examine the effects of age and inflammation on the $CD19^+$ B cell transcriptome.

3.5 Future work

- Determine if BCR signal transduction is impaired in RA compared to controls by measuring phosphoproteins by flow cytometry
- Determine whether the $CD19^+$ B cells, as a broad group, from RA patients are anergic in early disease and if they become activated and proliferate in response to stimulation by the BCR *in vitro*.
- Quantify IL-6 production in response to stimulation in B cells from RA patients compared to controls, and examine whether this is increased as in the paper by Barr *et al*[177].

4. Factors affecting the gene expression profile of peripheral blood B cells

4.1 Background

In *Chapter 3* the $CD19^+$ B cell gene expression profile was examined to identify changes related to disease pathology. Beyond the presence of arthritis there are multiple factors potentially affecting the gene expression profile in $CD19^+$ B cells including genetic and environmental factors[130, 133, 184-186]. These may explain the absence of a clear, multiple test corrected gene signature distinguishing RA samples from non-RA samples discussed in *Chapter 3.*

Age related changes in gene expression in immune cell subsets have been described, although not in $CD19^+$ B cells[187]. Aging is associated with telomere attrition, epigenetic changes and genomic instability which in turn lead to gene expression changes[188, 189]. There is increasing interest in the molecular changes associated with ageing to provide insights into the age-related increase in the incidence of autoimmune diseases[190].

When discussing aging and the immune system "immunosenescence" and "inflamm-aging" are frequently used terms. Immunosenescence refers to the age-related decline in the immune system which leads to the increased morbidity and mortality observed in the elderly from infection and the reduced response to vaccines in this population. Inflamm-aging describes the persistent, low grade increase in inflammation observed with age which is associated with an increase in proinflammatory cytokines such as IL-6. The underlying cause of the latter state remains to be identified, but it has been suggested that this is secondary to chronic antigen stimulation, the presence of persistent, latent infections or the transfer of microbes across the gut epithelium[191, 192].

Our understanding of the changes in the human B cell population with aging are based largely on the quantification of circulating peripheral cell subsets. Work in mouse models has shown that there is a decrease in lymphopoeisis with increasing age. Production shifts toward the generation of myeloid progenitors over lymphoid progenitors and the B cell population is increasingly comprised of antigen-experienced cells at the expense of naive cells. In addition, novel subsets emerge, such as age related B cells, and the immunoglobulin pool is dominated by immunoglobulins that have undergone somatic hypermutation such that diversity is reduced.[193, 194]

The response to influenza vaccination in the elderly is reduced with a decreased frequency of vaccine specific antibody secreting cells (ASCs), reduced expression of BLIMP-1, a transcriptional repressor critical for the terminal differentiation of B cells, and a reduction in the quantity of vaccine specific antibodies. It has been suggested that memory cells are less able to respond to antigen, cells are less able to differentiate into ASCs and clonal expansion may also be reduced[195, 196].

In this chapter, the samples will be divided based on age and levels of inflammation, regardless of diagnosis, to examine the influence of these environmental covariates on the $CD19^+$ B cell transcriptome. The analysis of the dataset in this way was prompted by results from a comparison of RA and non-inflammatory samples and this will be described first.

4.2 Hypothesis and Aims

4.2.1 Hypothesis

In an older population, the peripheral B cell transcriptome will reveal upregulation of pro-inflammatory pathways as a consequence of inflamm-ageing. In samples with higher levels of systemic inflammation, measured by ESR and CRP, upregulation of pro-inflammatory pathways will be identified.

4.2.2 Aims

To determine the relationship between the transcriptome of circulating B cells with both chronological age and systemic inflammation in a cross-sectional study of patients attending an early arthritis clinic; specifically:

1. To identify a list of differentially expressed genes in the transcriptome of circulating $CD19^+$ B cells from older *versus* younger people.
2. To undertake a pathway analysis of these DEGs with a view to gaining new insight in to age related changes in B cells
3. To identify a list of differentially expressed genes in the transcriptome of circulating $CD19^+$ B cells based on inflammation

- To undertake a pathway analysis of DEGs with a view to gaining new insight in to changes in B cells due to increased levels of inflammation

4.3 Results

4.3.1 The effect of age and inflammation on a comparison between RA and non-inflammatory samples

In *Chapter 3* my data were examined to identify changes to the $CD19^+$ B cell transcriptome related to disease. Further analyses divided the original control group into subsets to enable comparisons between RA samples and samples from those with other inflammatory conditions alone (n=59) and then those with non-inflammatory conditions alone (n=72).

No DEGs robust to multiple test correction were identified in the comparison of RA *versus* other inflammatory conditions.

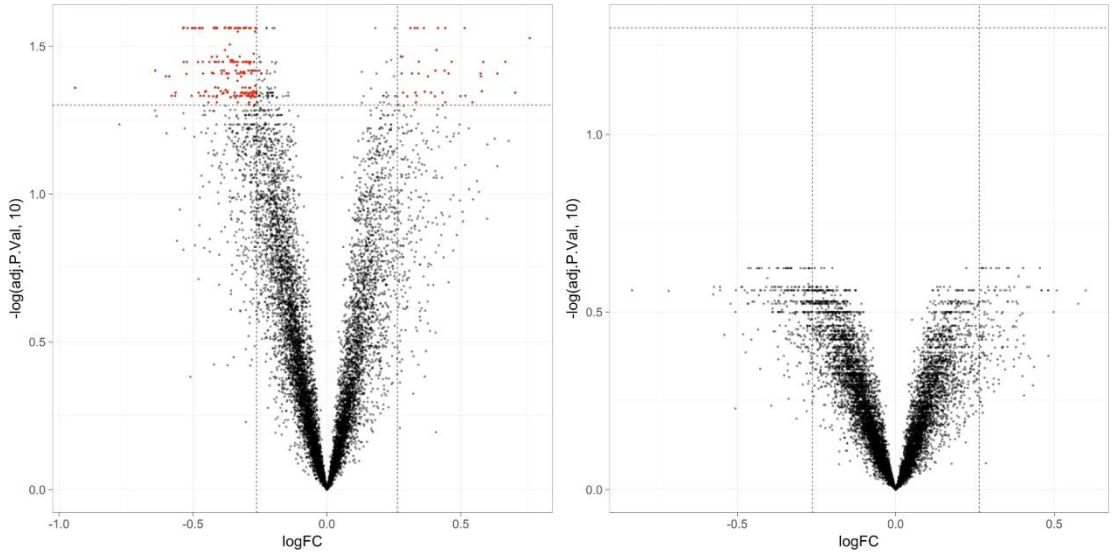
A list of DEGs, which withstood multiple test correction, was identified in the comparison between RA samples and non-inflammatory samples (using the baseline diagnosis and all samples available) (figure 4.1a). This identified a list of 225 differentially expressed probes, FC ≥ 1.2 *(see Appendix A.4 and A.5 for the list of DEGs and the results of the IPA analysis*).

A comparison of the clinical characteristics for the RA and non-inflammatory groups revealed no significant differences in gender but significant differences in age, CRP, ESR, swollen joint count, tender joint count and DAS28 score between the two groups (table 4.1). In light of this the clinical variables CRP, ESR and age were added to the linear model and the list of DEGs no longer withstood multiple test correction (figure 4.1b).

	RA	Non-inflammatory	P-value
Sample number	59	72	-
Age (yrs)	61 (21-89)	52 (22-87)	0.0008
Gender (%F)	76.27	79.17	0.83
ESR (mm/hr)	25 (1-91)	7 (1-100)	0.0002
CRP (mg/L)	10 (0-91)	5 (0-49)	<0.0001
SJC	1 (0-25)	0 (0-11)	<0.0001
TJC	6 (0-22)	3 (0-28)	0.0028
DAS28	4.59 (1.26-8.46)	3.03 (0-5.46)	<0.0001

Table 4.1 Demographics for gene expression analyses based on baseline diagnosis for RA and non-inflammatory samples

Median (range) shown. P-values were calculated using the unpaired T test (age), Mann Whitney (CRP, ESR, SJC, TJC, DAS28), or Fisher's exact test (gender).



a) *b)*

Figure 4.1 Differentially expressed genes identified between the CD19$^+$ B cell transcriptome of RA and non-inflammatory samples

CD19$^+$ B cells were positively isolated from patients presenting to the early arthritis clinic, RNA extracted and transcriptome analysed using microarray technology. a) Volcano plot of RA (n=59) against non-inflammatory samples (n=72) identified 225 differentially expressed probes (FC ≥ 1.2, p 0.05, MTC Benjamini-Hochberg). b) Volcano plot of RA (n=59) against non-inflammatory (n=72) samples with covariates age, ESR and CRP added to the linear model, no differentially expressed probes were identified (FC ≥ 1.2, p 0.05, MTC Benjamini-Hochberg). Vertical, dotted lines denote FC 1.2. The x axis represents log2 of the fold change, y axis represents the –log$_{10}$ adjusted p-value. Horizontal, dotted lines denote p-value 0.05. Red dots indicate probes which are differentially expressed between the comparator groups.

Adding the clinical variables separately (age, ESR and CRP) to the linear model led to the disappearance of the list of DEGs in each instance (*Appendix A.6).*

This result suggests that chronological age and/or systemic inflammation, rather than clinical diagnosis, may explain the majority of the variability in $CD19^+$ B cell gene expression in this population. In order to further examine the effect of these factors on the $CD19^+$ B cell transcriptome I compared the samples based on age and inflammatory status. To maximise power I used all the samples which had passed the quality control regardless of diagnosis and used the median values for each chosen clinical variable to divide the samples into two groups for the subsequent analyses.

A total of 243 samples were available for analysis. The median age of the patients at recruitment was 54 years old and this was selected as the cut off age between the sample groups. The comparator groups were aged under 55 years (n=131) and 55 years and over (n=112).

To examine changes in the transcriptome related to inflammation the samples were divided using median CRP and ESR levels. The low inflammation group being less than or equal to the median value for the population. For ESR, the median value was 14mm/hr and so the two groups were low inflammation (ESR <15mm/hr) and high inflammation (ESR ≥15mm/hr), giving groups of 126 and 117 samples respectively. In the case of CRP, the median was 7mg/L and so the two groups were low inflammation (CRP <8mg/L) and high inflammation (CRP ≥ 8mg/L), giving groups of 128 and 115 samples respectively.

4.3.2 Relationship between clinical covariates

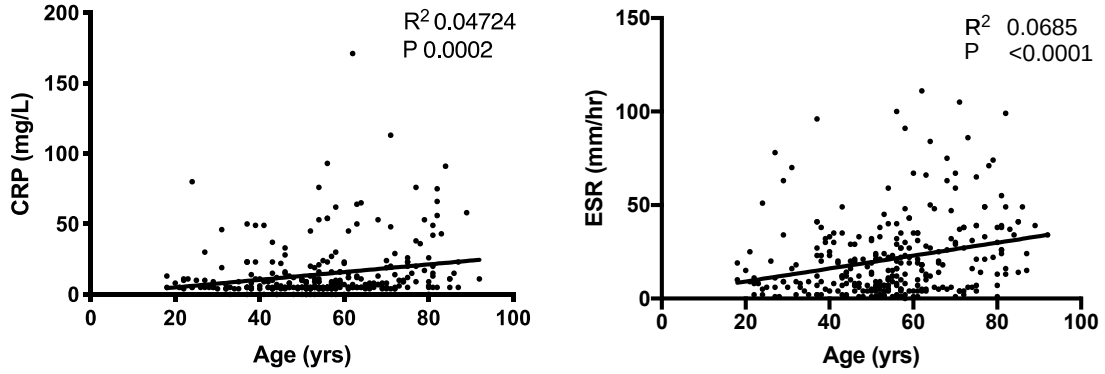
The <55 years group had significantly lower levels of inflammatory markers and a lower proportion of patients diagnosed with RA, based on their outcome diagnoses (table 4.2).

	< 55 years	≥ 55 years	P-value
Sample number	131	112	-
ESR mm/hr	9 (0-96)	21 (0-111)	<0.0001
CRP mg/L	5 (0-80)	8 (0-171)	0.0074
Gender (% F)	67.0	80.9	0.0178
Outcome diagnosis (% RA)	20.6	41.1	0.0007

Table 4.2 Clinical data for gene expression analyses based on age at baseline diagnosis

Median (range) shown. P-values were calculated using the unpaired T test (age), Mann Whitney test (CRP, ESR), or Fisher's exact test (gender and diagnosis).

Further analysis demonstrated that age is significantly correlated with CRP and ESR (figure 4.2). This is in keeping with the concept of inflamm-ageing. 41.1% of the samples in the ≥55 years age group have an outcome diagnosis of RA compared to 20.6% in the younger group. When the RA group is combined with other inflammatory conditions there are again more samples from patients with a current diagnosis of an inflammatory arthritis in the ≥55 years age group; 70.5% versus 55.0% (P 0.0167, Fisher's exact test).



a) *b)*

Figure 4.2 Linear regression of age against inflammatory markers

a) Age and CRP b) Age and ESR using samples from all recruited patients. Age positively correlates with CRP and ESR.

Age positively correlates with both ESR and CRP, however the proportion of variance of the CRP which is explained by age (as expressed by the R^2 value) is only 0.05, and for ESR is only 0.07. Further, the samples in the ≥ 55 years age group and each high inflammation group are not the same. Of the samples in the ≥55 years age group 59.8% overlapped with the high ESR group and 52.7% overlapped with the high CRP group.

This demonstrates that higher levels of inflammation do not simply reflect older age. 72.6% of the high ESR samples overlapped with the high CRP group (figure 4.3).

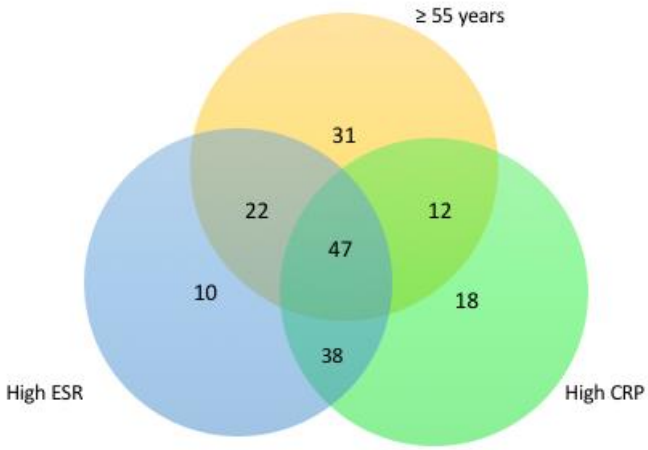


Figure 4.3 Venn diagram of samples in the ≥ 55 years age group, high ESR and high CRP groups

4.3.3 The effect of age on the CD19^{+} cell transcriptome

The samples were divided into two sample groups: <55 years old and ≥ 55 years old. The principal components analysis did not show a clear difference based on age (figure 4.4).

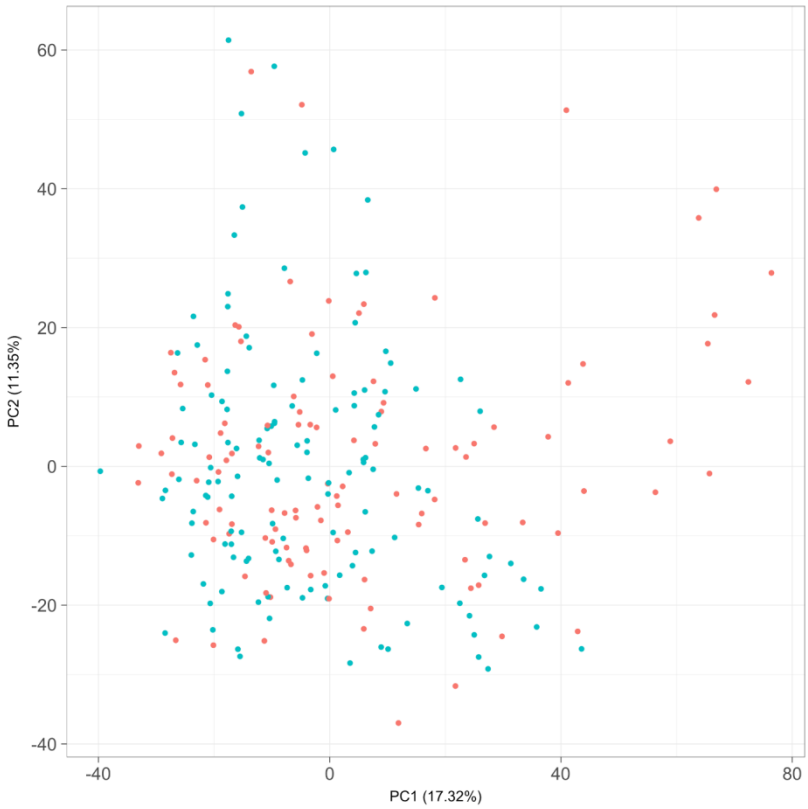


Figure 4.4 Principal component analysis (PCA) of samples used in the analysis between age groups

The younger age group (age <55 years) samples are shown in blue and older age group (age ≥55 years) in red. The samples in each group do not clearly cluster together.

A differential expression analysis was carried out to compare the two age groups with ESR and CRP, the additional factors under review, added to the linear model. 136 differentially expressed probes which stood up to multiple test correction at a FC ≥ 1.2 were identified (figure 4.5).

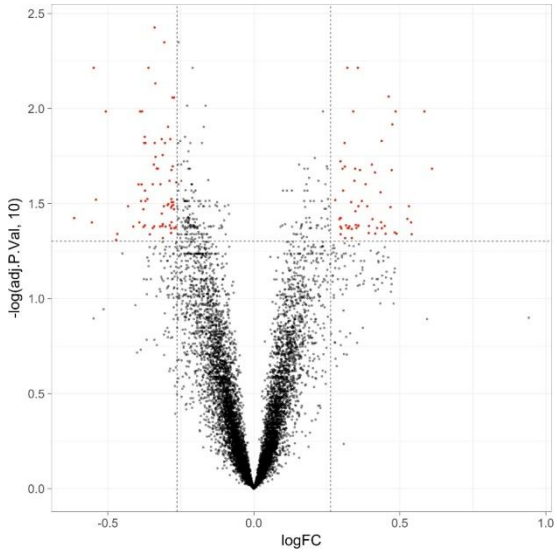


Figure 4.5 Differentially expressed genes identified in the CD19^{+} B cell transcriptome in a comparison based on chronological age with ESR and CRP added to the linear model

CD19^{+} B cells were positively isolated from patients presenting to the early arthritis clinic, RNA extracted and transcriptome analysed using microarray technology. The samples were divided by median age at presentation. Volcano plot of comparison between ≥55 years old age group (n=112) and <55 years old group (n=131). ESR and CRP added to the linear model. MTC; Benjamini-Hochberg, FC ≥ 1.2, p 0.05. 136 differentially expressed probes identified. Vertical, dotted lines denote FC 1.2. The x axis represents log2 of the fold change, y axis represents the –log$_{10}$ adjusted p-value. Horizontal, dotted lines denote p-value 0.05. Red dots indicate probes which are differentially expressed between the comparator groups.

Ingenuity pathway analysis

The 136 probes identified, representing 129 unique genes, were analysed using IPA. No significant canonical pathways were identified which met the filtering criteria previously described. The upstream regulator analysis identified 5 potential regulators (table 4.3). 14 out of 129 molecules are identified as regulated by TNF. TNF, IL-2, TGM2 and NOS2 are predicted to be activated in the ≥55 years group and APOE is inhibited in the older age group.

The network display of the upstream regulators demonstrates that the vast majority of predicted interactions, with molecules in the DEG list, are indirect (figure 4.6).

Upstream Regulator	Molecule Type	Name	z-score	p-value	Target molecules in the dataset
APOE	Transporter	Apolioprotein E	-2.219	0.000335	ADGRG1, IL1RN, JAK1, LCN2, LTBR, NCF2, TIMP1, TNFRSF1A
IL2	Cytokine	IL-2	2.167	0.000977	BCCIP, CX3CR1, EPHA4, GART, IP6K2, PRF1, TIMP1, TNFRSF1A
TNF	Cytokine	TNF	2.149	0.00033	CSF1R, GBP2, HK3, IL1RN, IRAK3, ITGA4, JAK1, LCN2, MAFF, NCF2, NLRP3, PIK3CB, TIMP1, TNFRSF1A
TGM2	Enzyme	Transglutaminase 2	2.646	0.00027	AQP9, CX3CR1, HK3, LILRA5, NCF2, PADI4, SIRPA
NOS2	Enzyme	Nitric oxide synthase 2	2.219	0.00294	ATP2A2, IL1RN, LCN2, PRF1, TIMP1

Table 4.3 Upstream regulators for the differentially expressed genes between < 55 years and ≥ 55 year samples

IPA software was used to identify upstream regulators for the list of DEGs identified by comparing the CD19^{+} B cell transcriptome between ≥55 years old age group (n=112) and <55 years old group (n=131) with ESR and CRP added to the linear model. Z score ≤ -2 predicts an inhibited state (blue) and ≥ 2 predicts an activated state for the regulator

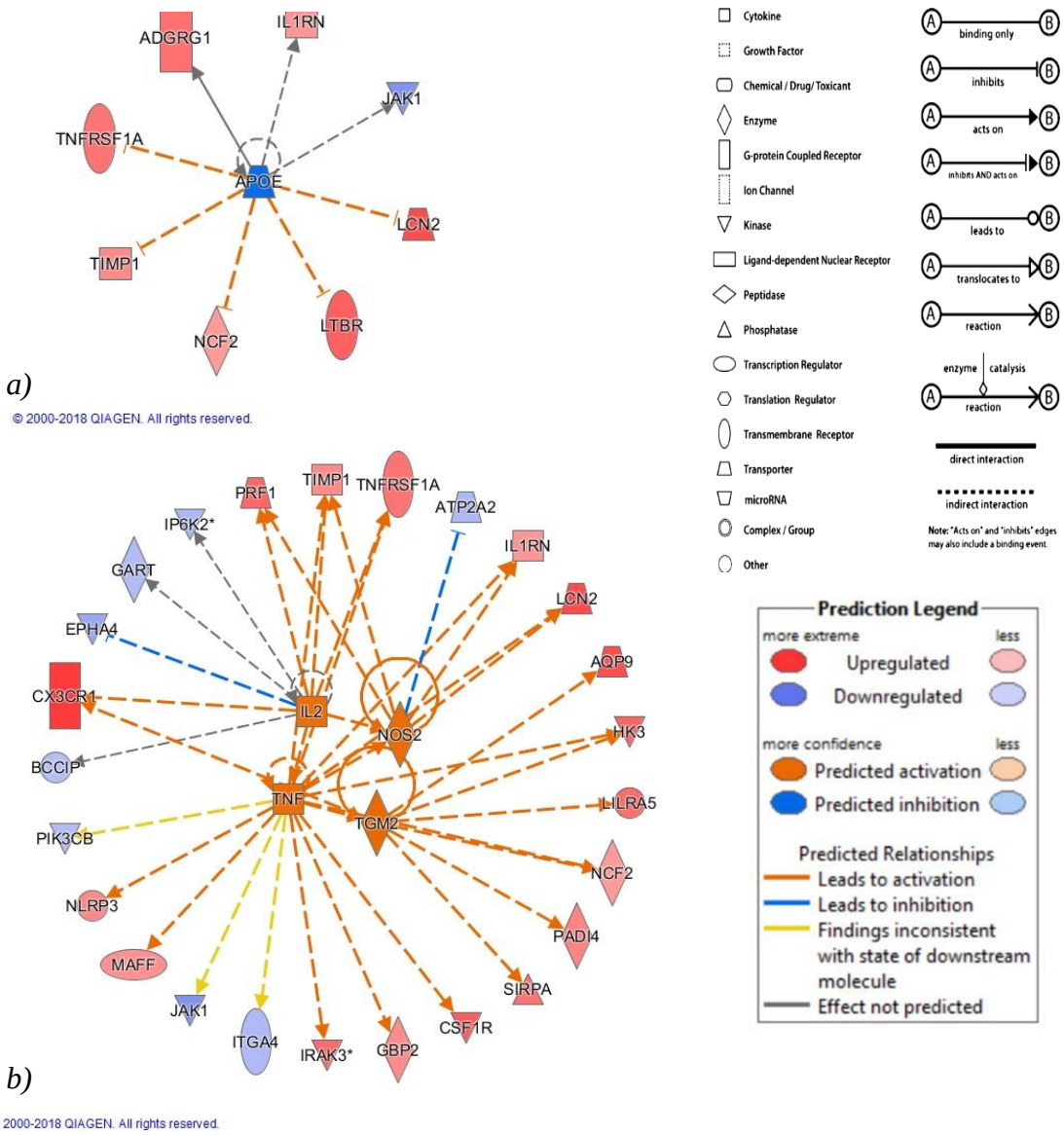


Figure 4.6 Network display of upstream regulators for the differentially expressed genes between ≥55 years and <55 years old samples

IPA software was used to predict activated upstream regulators for the list of DEGs identified by comparing the CD19^{+} B cell transcriptome from ≥55 years old patients (n=112) and <55 years old patients (n=131) with ESR and CRP added to the linear model. a) Inhibited upstream regulator, APOE, predicted to be inhibited in the ≥ 55 years old group. b) Activated upstream regulators, activated in the ≥55 years old group. Upstream regulators are displayed centrally with lines connecting each to the molecules in the dataset which they are predicted to effect.
** indicates that multiple identifiers were found in the dataset to map to single gene.*

The top two networks assembled by IPA were: network 1, described as relating to infectious diseases, inflammatory response and cellular movement (figure 4.7) and network 2 which related to cellular movement, cancer and endocrine system disorders (figure 4.8).

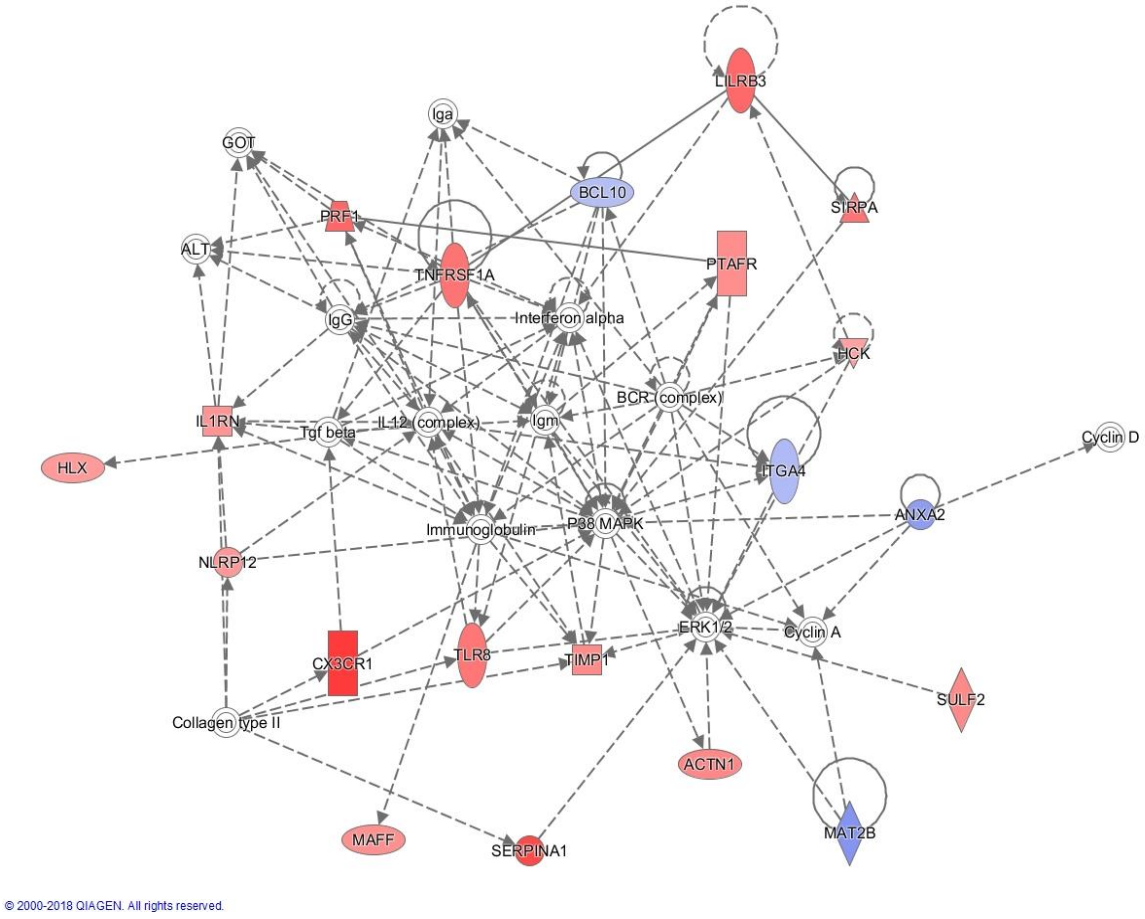


Figure 4.7 Network 1 with function related to infectious diseases, inflammatory response and Cellular Movement

Analysis using IPA software to identify molecular networks from the list of DEGs from the comparison of the ≥55 years (n=112) to <55 years (n=131) samples. Molecules not coloured are from the Ingenuity Knowledge Base and not present in the dataset. The shape of the gene represents the functional class of the gene product (key as for figure 4.6)

.

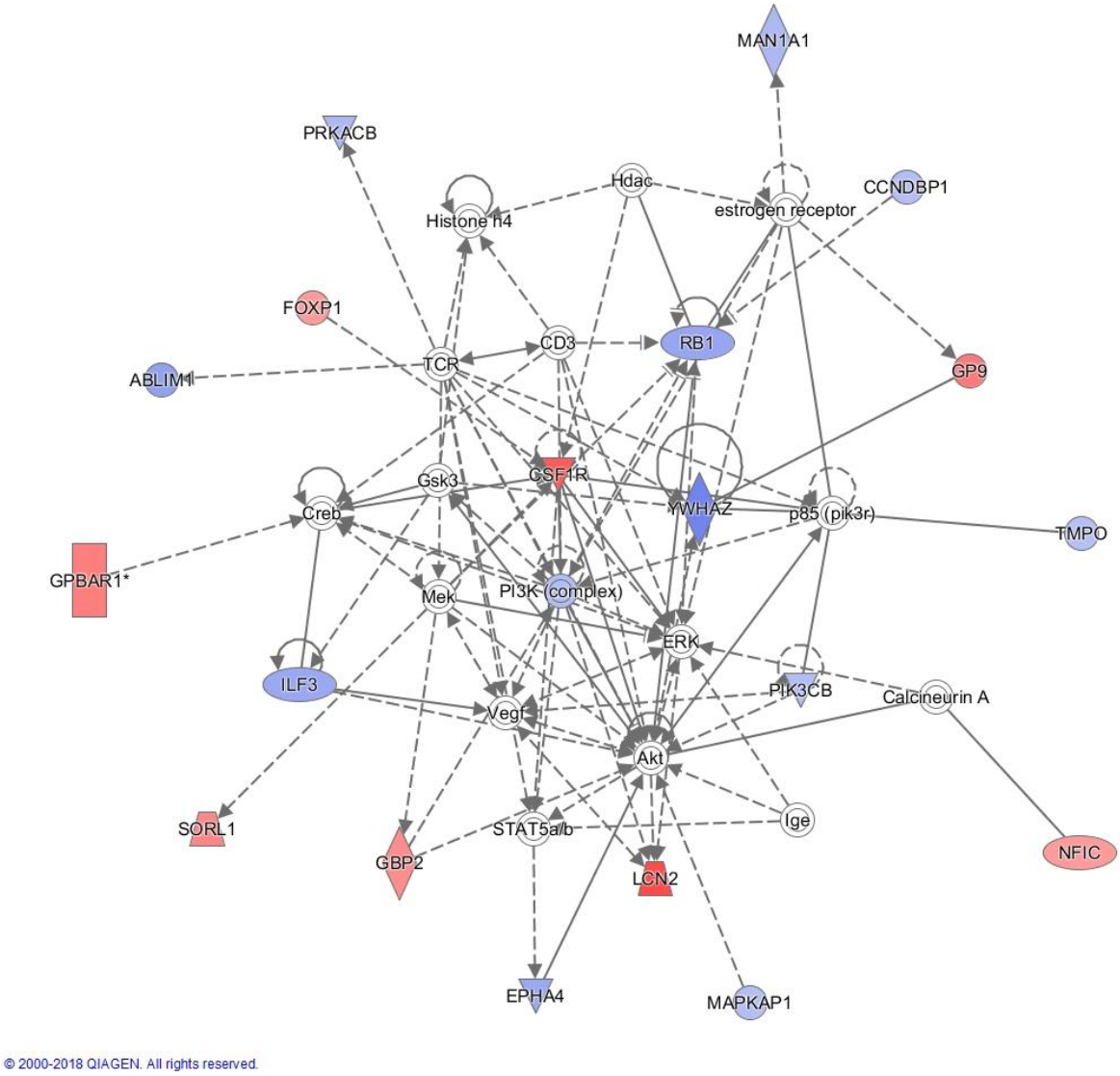


Figure 4.8 Network 2 with function related to cellular movement, cancer and endocrine system disorders

Analysis using IPA software to identify molecular networks from the list of DEGs from the comparison of the ≥55 years (n=112) to <55 years (n=131) samples. Molecules not coloured are from the Ingenuity Knowledge Base and not present in the dataset. The shape of the gene represents the functional class of the gene product (key as for figure 4.6)

4.3.4 The effect of age on the CD19^{+} cell transcriptome without consideration of inflammatory status

The analysis described in section *4.3.3* was repeated without including ESR and CRP in the linear model. The concept of inflamm-ageing describes the persistent, low grade increase in inflammation observed with age. In a comparison between samples based on

age, interesting pathways related to inflamm-ageing may be missed if inflammatory status is incorporated within the linear model. When the analysis was repeated, 127 probes were differentially expressed between the older and younger sample groups, representing 120 unique genes (figure 4.9).

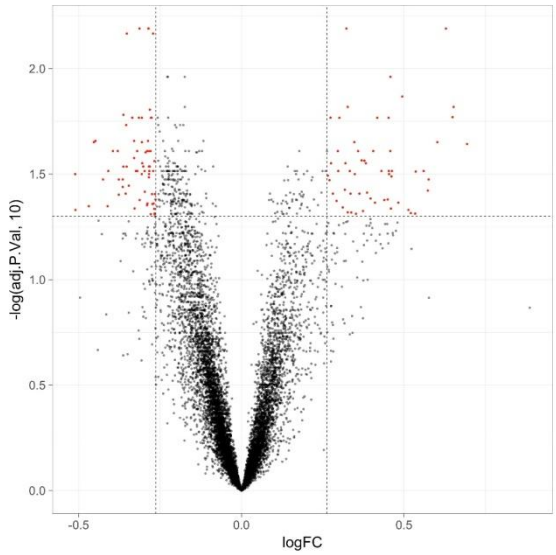


Figure 4.9 Differentially expressed genes identified in the CD19^{+} B cell transcriptome in a comparison based on chronological age

CD19^{+} B cells were positively isolated from patients presenting to the early arthritis clinic, RNA extracted and transcriptome analysed using microarray technology. The samples were divided by median age at presentation. Volcano plot of groups age ≥55 years (n=112) against age < 55 years old (n=131). MTC; Benjamini-Hochberg, FC ≥ 1.2, p 0.05. 127 differentially expressed probes identified. Vertical, dotted lines denote FC 1.2. The x axis represents log2 of the fold change, y axis represents the –log$_{10}$ adjusted p-value. Horizontal, dotted lines denote p-value 0.05. Red dots indicate probes which are differentially expressed between the comparator groups.

Ingenuity pathway analysis

The list of differentially expressed genes was subjected to IPA. The canonical pathway analysis identified one pathway, the ERK/MAPK pathway, which met the filtering criteria of ≥ 5 molecules in the dataset and a Z-score ≤-2 or ≥2.

The ERK/MAPK pathway (Z-score -2, log10 Pval 2.4) was downregulated in the ≥ 55years age group in comparison to the <55years age group. The molecules from the dataset in the pathway were: ITGA4, DUSP6, PRKACB, PIK3CB and YWHAZ. The ERK/MAPK pathway is involved in many cell signalling pathways including that of the BCR.

The upstream regulator analysis identified 22 potential upstream regulators, of which 7 were inhibited in the ≥ 55 years group (table 4.4). The pathways inhibited in the ≥ 55 years group include APOE which was seen in the analysis with ESR and CRP added to the linear model. The pathways activated in the ≥ 55 years old group include TNF and 5 additional cytokines, including the proinflammatory cytokines IL-1A, IL-17A and the receptor IL-17RA.

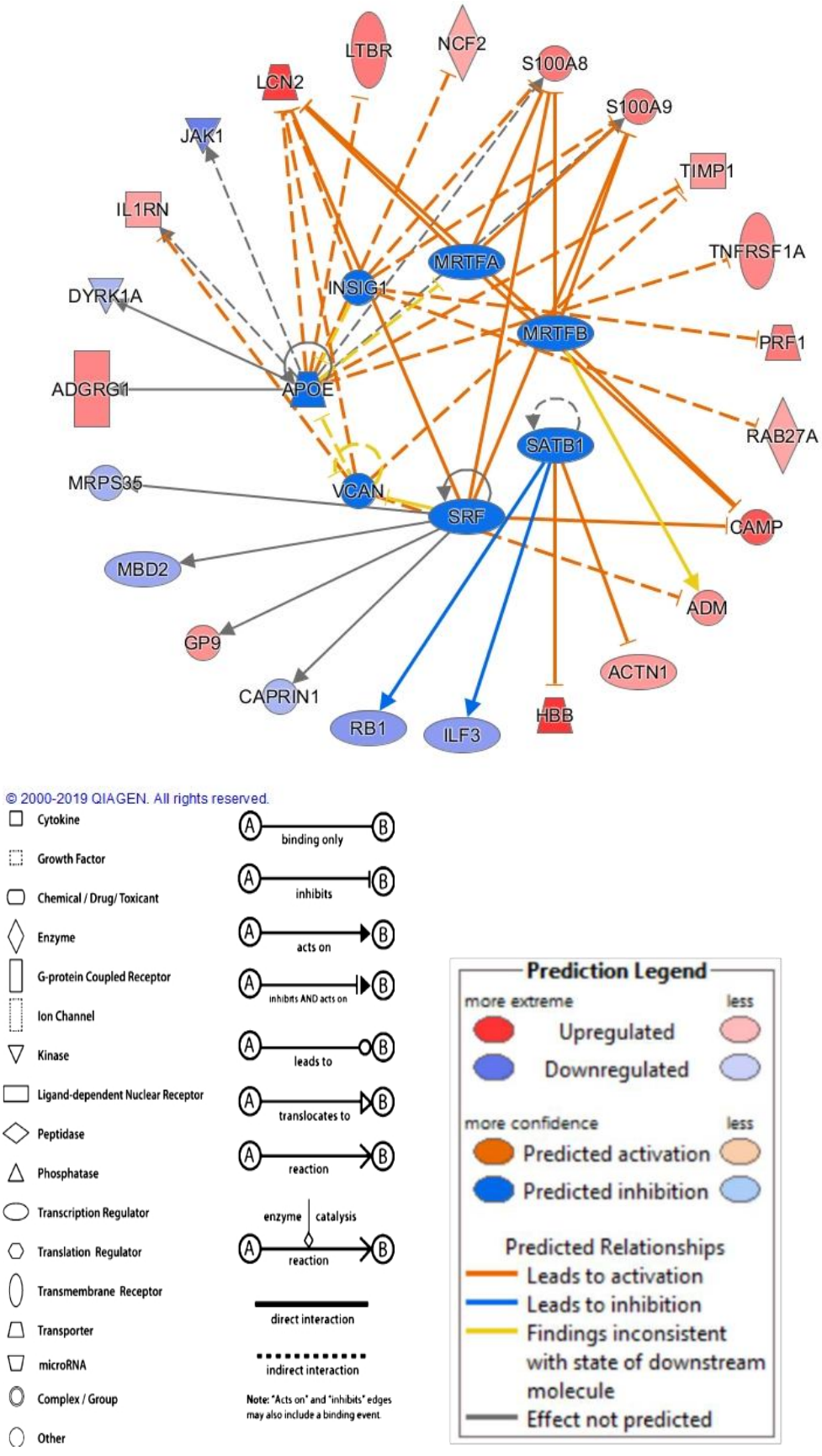
The upstream regulator analysis can also be displayed as a network to allow an exploration of the relationship between the regulator and the relevant molecules in the dataset (figure 4.10).

Upstream regulator	Molecule type	Name	Z -core	P-value	Target molecules in dataset
VCAN	other	Versican	-2.236	0.000475	ADM, IL1RN, LCN2, TIMP1, VCAN
INSIG1	other	Insulin induced gene 1	-2.236	0.000151	LCN2, PRF1, RAB7A, S100A8, S100A9
APOE	transporter	Apolioprotein E	-2.219	0.000000814	ADGRG1, DYRK1A, IL1RN, JAK1, LCN2, LTBR, NCF2, S100A8, S100A9, TIMP1, TNFRSF1A
MRTFB	transcription regulator	Myocardin related transcription factor B	-2	0.00107	CAMP, LCN2, S100A8, S100A9
SRF	transcription regulator	Serum response factor	-2	0.0000424	CAMP, CAPRIN1, GP9, LCN2, MBD2, MRPS35, S100A8, S100A9
MRTFA	transcription regulator	Myocardin related transcription factor A	-2	0.0045	CAMP, LCN2, S100A8, S100A9
SATB1	transcription regulator	SATB homeobox 1	-2	0.0105	ACTN1, HBB, ILF3, RB1
TNFSF12	cytokine	TNF superfamily member 12	2	0.00107	CCR1, S100A8, S100A9, TIMP1
IL5	cytokine	IL-5	2	0.0163	CCR1, DUSP6, GBP2, NFE2
mir-223	Micro RNA	microRNA 223	2	0.00437	IL1RN, MMP25, PADI4, S100A9
IL17RA	transmembrane receptor	IL-17A receptor	2	0.000015	CCR1, S100A8, S100A9, TIMP1
EHF	transcription regulator	ETS homologous factor	2	0.000936	IL1RN, S100A12, S100A8, S100A9
TP63	transcription regulator	Tumour protein p63	2.186	0.0187	ADM, DUSP6, ITGA4, S100A8, TNFRSF1A
IKBKG	kinase	inhibitor of nuclear factor kappa B kinase regulatory subunit gamma	2.213	0.000594	CCR1, DUSP6, GBP2, IL1RN, LCN2
NOS2	enzyme	Nitric oxide synthase 2	2.219	0.00245	IL1RN, LCN2, PRF1, S100A8, TIMP1

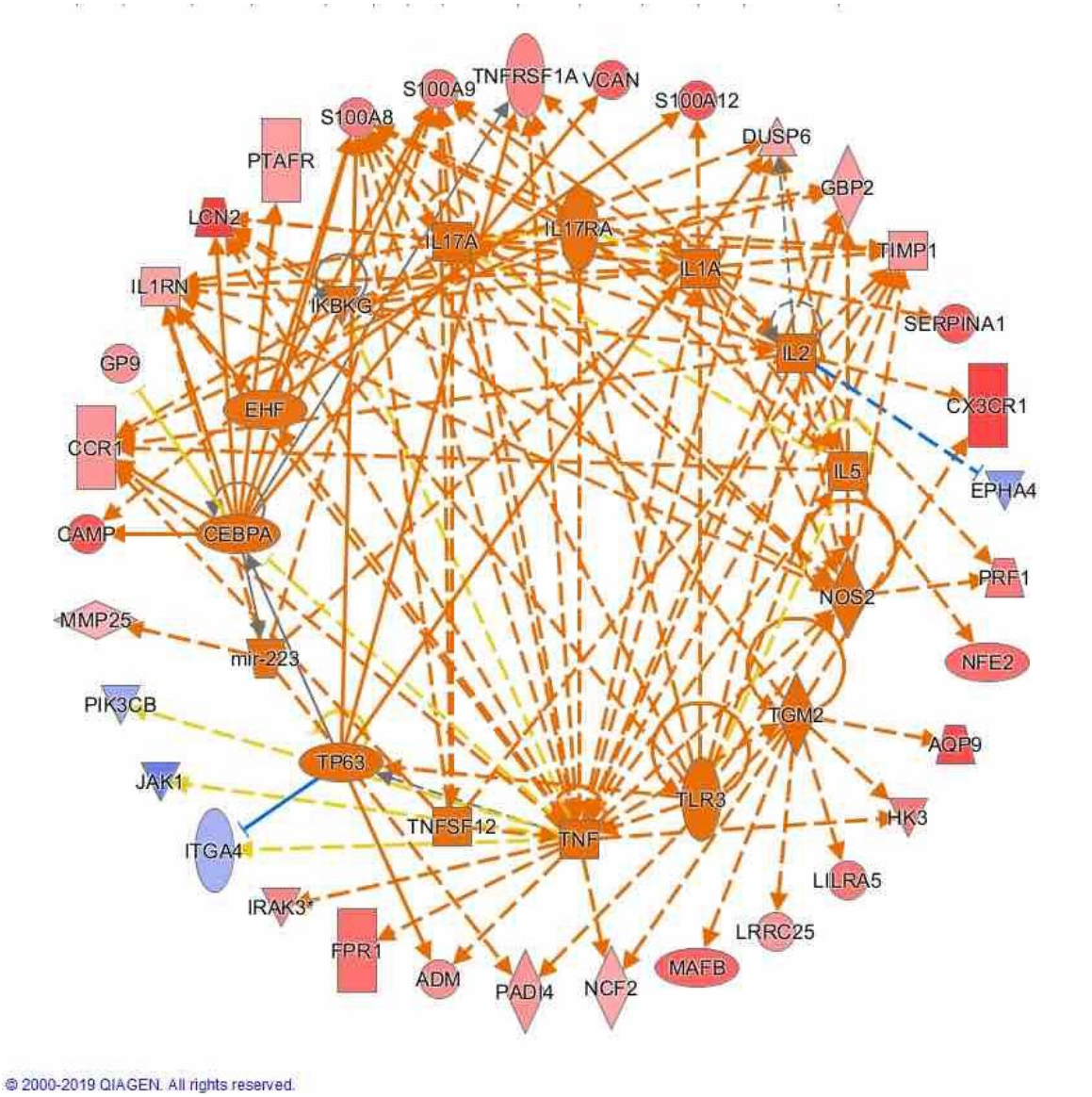
TLR3	transmembrane receptor	Toll like receptor 3	2.219	0.00561	GBP2, LCN2, S100A8, TIMP1, TNFRSF1A
CEBPA	transcription regulator	CCAAT enhancer binding protein alpha	2.395	0.0000166	CAMP, CCR1, GP9, IL1RN, LCN2, PTAFR, S100A8, S100A9, TNFRSF1A, VCAN
IL1A	cytokine	IL-1A	2.433	0.0000534	IL1RN, LCN2, S100A12, S100A8, S100A9, SERPINA1
IL17A	cytokine	IL-17A	2.527	0.00000591	CAMP, GBP2, IL1RN, LCN2, S100A8, S100A9, TIMP1
TNF	cytokine	TNF	2.558	0.00000343	ADM, CCR1, DUSP6, FPR1, GBP2, HK3, IL1RN, IRAK3, ITGA4, JAK1, LCN2, NCF2, PIK3CB, S100A8, S100A9, TIMP1, TNFRSF1A
IL2	cytokine	IL-2	2.569	0.000741	CCR1, CX3CR1, DUSP6, EPHA4, PRF1, S100A8, TIMP1, TNFRSF1A
TGM2	enzyme	Transglutaminase 2	3	0.00000339	AQP9, CX3CR1, HK3, LILRA5, LRRC25, MAFB, NCF2, PADI4, S100A8

Table 4.4 Upstream regulators for the differentially expressed genes between ≥ 55 years and < 55 years samples

IPA software was used to identify upstream regulators for the list of DEGs identified by comparing the CD19^{+} B cell transcriptome from patients ≥55 years (n=112) to <55 years (n=131). Z score ≤ -2 predicts an inhibited state and ≥ 2 predicts an activated state for the regulator. Inhibited regulators shown in blue


LTBR
NCF2
S100A8
LCN2
S100A9
JAK1
TIMP1
IL1RN
TNFRSF1A
MRTFA
INSIG1
DYRK1A
MRTFB
PRF1
ADGRG1
APOE
RAB27A
SATB1
MRPS35
VCAN
SRF
CAMP
MBD2
ADM
GP9
ACTN1
CAPRIN1
RB1
ILF3
HBB
© 2000-2019 QIAGEN. All rights reserved.
Cytokine
Growth Factor
Chemical / Drug/ Toxicant
Enzyme
G-protein Coupled Receptor
Ion Channel
Kinase
Ligand-dependent Nuclear Receptor
Peptidase
Phosphatase
Transcription Regulator
Translation Regulator
Transmembrane Receptor
Transporter
microRNA
Complex / Group
Other
binding only
inhibits
acts on
inhibits AND acts on
leads to
translocates to
reaction
enzyme
catalysis
reaction
direct interaction
indirect interaction
Note: "Acts on" and "inhibits" edges may also include a binding event.
Prediction Legend
more extreme
less
Upregulated
Downregulated
more confidence
less
Predicted activation
Predicted inhibition
Predicted Relationships
Leads to activation
Leads to inhibition
Findings inconsistent with state of downstream molecule
Effect not predicted

a)



b)

Figure 4.10 Network display of upstream regulators for the differentially expressed genes between ≥55 years old and < 55 years old patient samples

IPA software was used to predict activated upstream regulators for the list of DEGs identified by comparing the CD19^{+} B cell transcriptome from <55 years old patients (n=131) and ≥55 years old patients (n=112) a) Inhibited upstream regulators, inhibited in the ≥55 years old group. b) Activated upstream regulators, activated in the ≥55 years old group. Upstream regulators are displayed centrally with lines connecting each to the molecules in the dataset which they are predicted to effect.

In a molecular network analysis of this dataset the top network is described as relating to: Cellular Movement, Haematological System Development and Function and Immune Cell Trafficking (figure 4.11). The second network related to connective tissue disorders, developmental disorder, haematological disease (*Appendix A.7*).

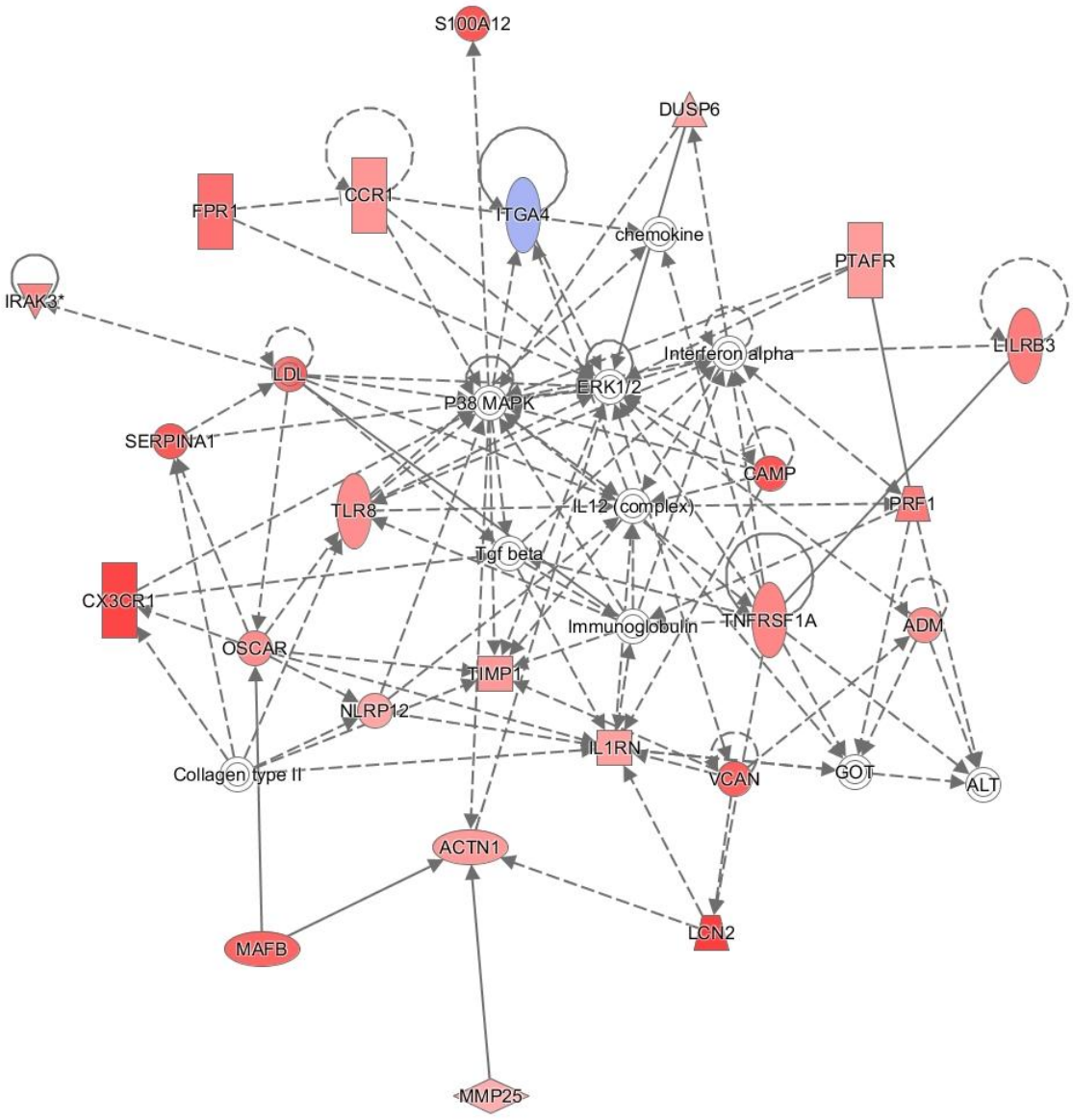


Figure 4.11 Network 1 with function related to Cellular Movement, Haematological System Development and Function and Immune Cell Trafficking

Analysis using IPA software to identify molecular networks from the list of DEGs from the comparison of the ≥55 years old patients (n=112) to ≥ 55 years old patients (n=131) samples. Molecules not coloured are from the Ingenuity Knowledge Base and not present in the dataset. The shape of the gene represents the functional class of the gene product (Key as for figure 4.10).

4.3.5 The effect of inflammation on the CD19$^+$ cell transcriptome

The addition of ESR and CRP, individually, to the linear model used to compare samples from RA patients and those with non-inflammatory arthritis, led to the loss of the differentially expressed genes. In order to explore the effect of inflammation on the transcriptome, the samples were divided into two groups based on the median values for the ESR and CRP.

ESR

The high ESR group was significantly older, this is not unexpected given that ESR positively correlated with age. There was a significant difference in the proportion of RA patients in each group, with a higher percentage of RA patients in the high ESR group (table 4.5).

	ESR low	**ESR high**	**P-value**
Sample number	126	117	-
Gender (%F)	73.8	75.2	ns
Age (yrs)	51 (18-85)	57 (18-92)	0.0003
Outcome diagnosis (% RA)	22.2%	38.5%	0.0076

Table 4.5 Clinical data for gene expression analyses based on ESR at baseline diagnosis

Median (range) shown. Unpaired T test (age), Fisher's exact test (gender and diagnosis).

The samples in the high and low ESR groups did not separate out on a principal components analysis (figure 4.12).

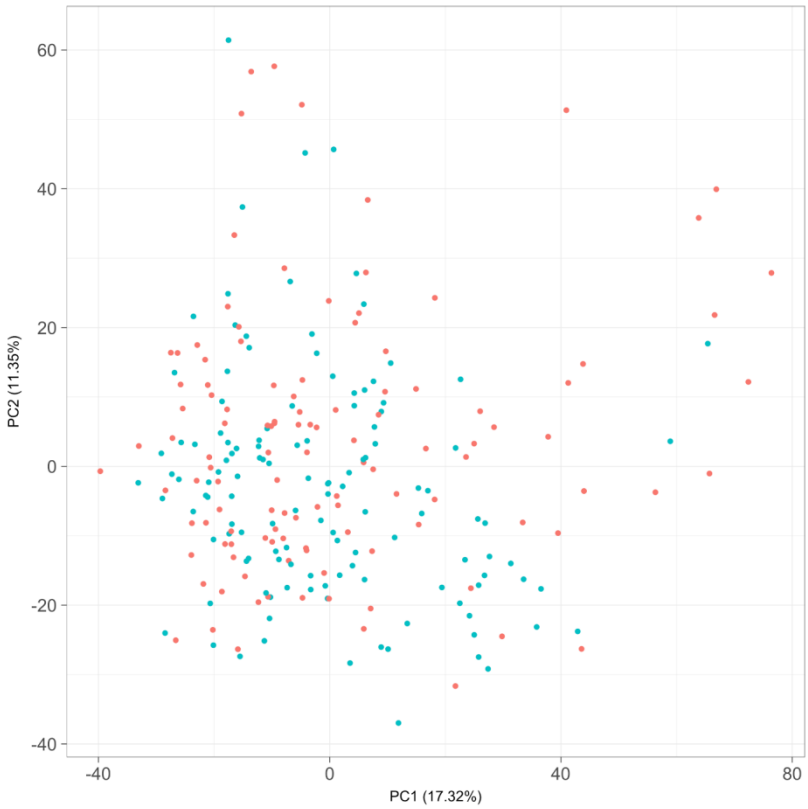
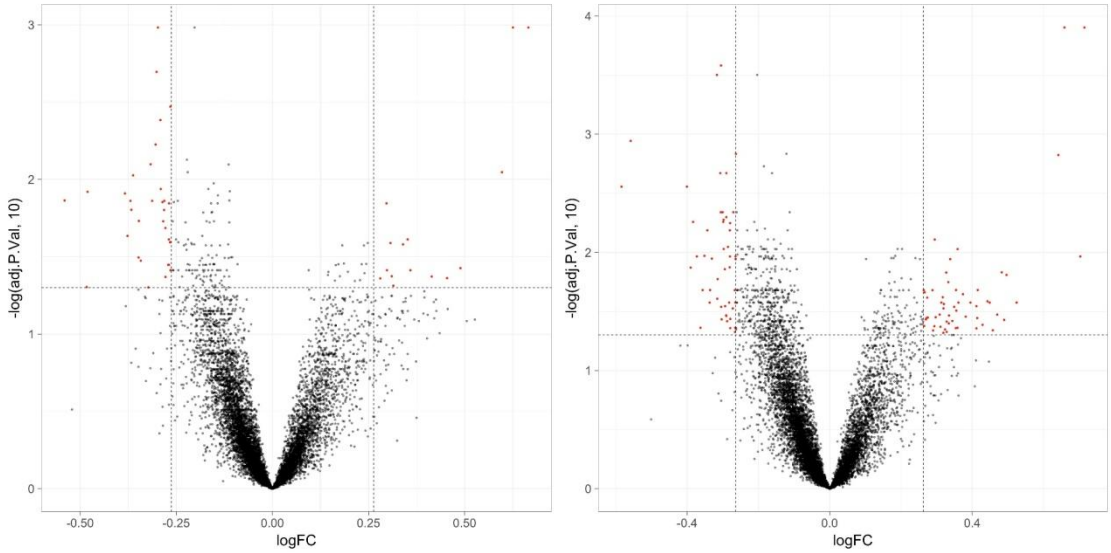


Figure 4.12 Principal component analysis of samples used in the analysis between low ESR and high ESR samples

There are no demonstrable differences between the two groups (red denotes high ESR samples and blue denotes low ESR samples). The samples in each group do not cluster together.

An initial comparison of the high ESR and low ESR groups was carried out with age added to the linear model. This identified 47 probes which were diffentially expressed (FC ≥1.2, p-value 0.05) with multiple test correction in place (figure 4.13a)

The analysis was repeated after removing age from the linear model and 104 differentially expressed probes were identified (figure 4.13b). The 104 probes identified represented 96 unique genes and just 8 of these were found in the list of differentially expressed genes in the comparison of samples based on age without consideration for clinical variables.



a) *b)*

Figure 4.13 Differentially expressed genes identified between the CD19$^+$ B cell transcriptome of samples with high ESR and low ESR

CD19$^+$ B cells were positively isolated from patients presenting to the early arthritis clinic, RNA extracted and transcriptome analysed using microarray technology. a) Volcano plot of samples with high ESR (n=117) against samples with low ESR (n=126) with age added to the linear model. 47 differentially expressed probes were identified. (FC ≥ 1.2, p ≤ 0.05, MTC Benjamini-Hochberg). b) Volcano plot of samples with high ESR (n=117) against samples with low ESR (n=126), no clinical covariates added to the linear model. 104 differentially expressed probes were identified. (FC ≥ 1.2, p ≤ 0.05, MTC Benjamini-Hochberg). Vertical, dotted lines denote FC 1.2. The x axis represents log2 of the fold change, y axis represents the –log$_{10}$ adjusted p-value. Horizontal, dotted lines denote p-value 0.05. Red dots indicate probes which are differentially expressed between the comparator groups.

Ingenuity pathway analysis

The list of 47 DEGs from the high ESR *versus* low ESR analysis with age added to the linear model was analysed using IPA and did not yield any significant pathways or upstream regulators. This may be related to the length of the list of DEGs.

The list of DEGs identified from the analysis of high ESR and low ESR samples without additional variables added to the linear model was subjected to IPA. There were no canonical pathways that met the filtering criteria set. 12 upstream regulators were identified by the IPA software in the list of 104 DEGs (table 4.6).

Upstream Regulator	Molecule Type	Name	z-score	p-value	Number of molecules
BNIP3L	other	BCL2 interacting protein 3 like	-2.236	0.00000362	ALAS2, CCNA2, CDKN3, CHEK1, KIF11
TOB1	transcription regulator	transducer of ERBB2, 1	-2.236	0.000000925	CCNA2, CDT1, CHEK1, SON, UBE2T
XBP1	transcription regulator	X-box binding protein 1	2	0.00136	DNAJC3, FKBP11, PPIB, PRDM1, SDF2L1
E2F3	transcription regulator	E2F transcription factor 3	2	0.000372	CCNA2, CDC45, CDK1, PCLAF, TK1
ESR1	ligand-dependent nuclear receptor	Estrogen receptor 1	2.087	0.0000317	ASPM, CCNA2, CDK1, CDKN3, CENPM, LCN2, NCAPG, UBE2T, XK, ZNF24
TAL1	transcription regulator	TAL bHLH transcription factor 1	2.236	0.0000388	ASPM, CDKN3, HBB, IL10RA, MELK, NCAPG, OIP5
EGFR	kinase	epidermal growth factor receptor	2.423	0.000318	CCNA2, CDK1, FKBP11, LCN2, SDF2L1, TIMM10
CSF2	cytokine	colony stimulating factor 2	2.438	0.000544	CCNA2, CD38, CDCA5, CDK1, KIF11, PHGDH
RABL6	other	RAB, RAS oncogene family like 6	2.449	0.000000334	CCNA2, CHEK1, MCM10, MELK, NCAPG, TTK
ERBB2	kinase	erb-b2 receptor tyrosine kinase 2	2.449	2.4E-10	ASPM, CCNA2, CDC45, CDCA5, CDK1, CDKN3, CDT1, CHEK1, CKS2, GAS6, LCN2, MCM10, NCAPG, TK1, TOP1MT, ZWINT

PTGER2	g-protein coupled receptor	prostaglandin E receptor 2	2.646	0.000000172	ASPM, CCNA2, CDKN3, CKS2, KIF11, MELK, TTK
RARA	ligand-dependent nuclear receptor	retinoic acid receptor alpha	3.095	0.000000168	ASPM, CCNA2, CD38, CDK1, CDKN3, CENPM, NCAPG, UBE2T, XK, ZNF24

Table 4.6 Upstream regulators for the differentially expressed genes between high ESR and low ESR samples

IPA software was used to identify upstream regulators for the list of DEGs identified by comparing the $CD19^{+}$ B cell transcriptome from patients with a high ESR (n=117) to low ESR patients (n=126). Z score ≤ -2 predicts an inhibited state (shown in blue) and ≥ 2 predicts an activated state for the regulator.

The upstream regulator analysis can also be displayed as a network to allow an exploration of the relationship between the regulator and the relevant molecules in the dataset (figure 4.14).

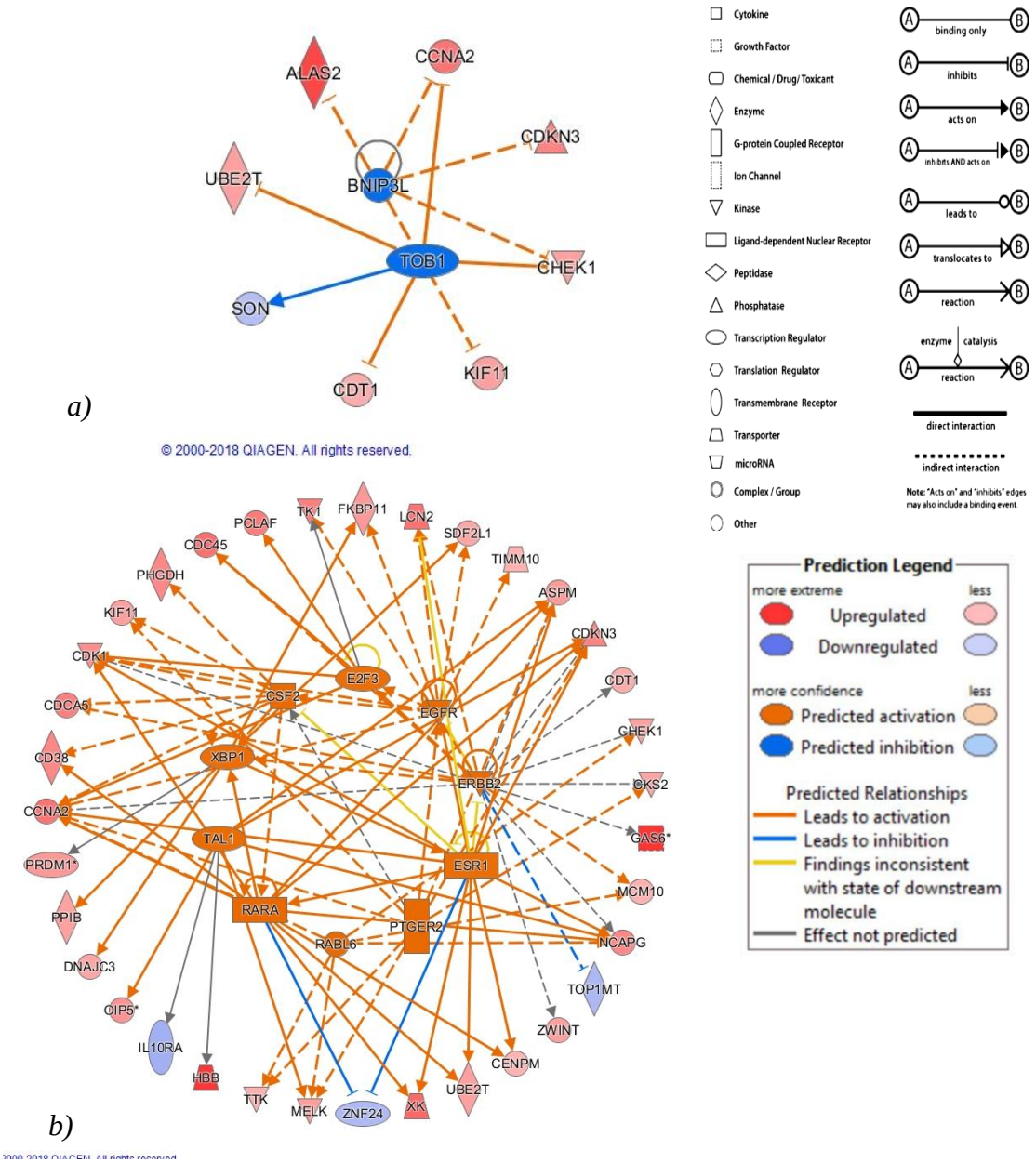


Figure 4.14 Network display of upstream regulators for the differentially expressed genes between high ESR and low ESR patient samples

IPA software was used to predict activated upstream regulators for the list of DEGs identified by comparing the CD19^{+} B cell transcriptome from high ESR patients (n=117) and low ESR samples (n=126). a) Inhibited upstream regulators, predicted to be inhibited in the high ESR group. b) Activated upstream regulators, activated in the high ESR group. Upstream regulators are displayed centrally with lines connecting each to the molecules in the dataset which they are predicted to effect.

The top two networks identified in the network analysis were network 1: DNA Replication, Recombination, and Repair, Cell Cycle, Cellular Movement and Cell Death and network 2: Survival, Cancer, Haematological Disease which is in agreement with the identification of several proto-oncogenes, nuclear receptors and growth factors in the list of potential upstream regulators. The top network is shown in figure 4.15.

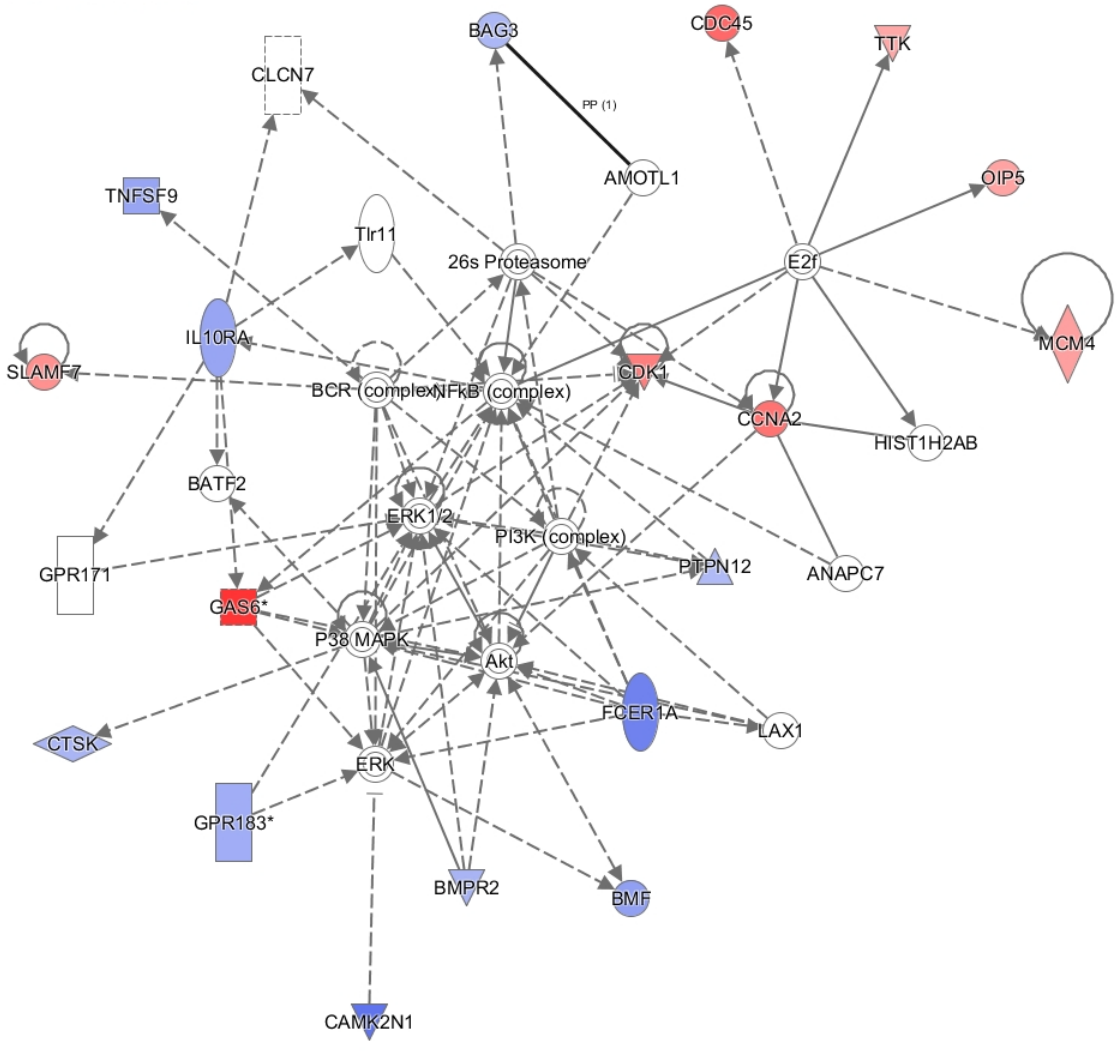


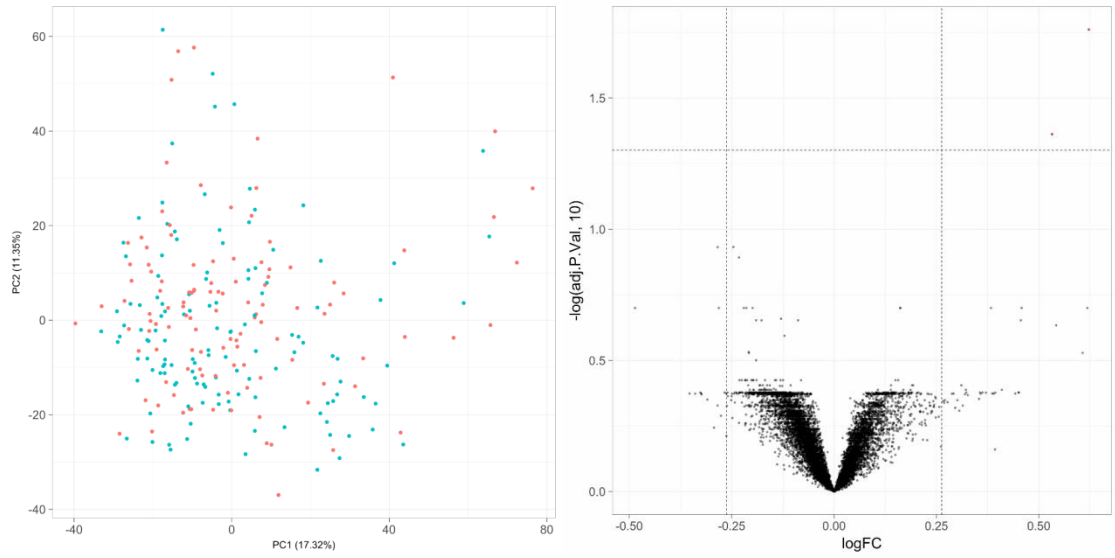
Figure 4.15 Network 1 with function related to DNA Replication, Recombination, and Repair, Cell Cycle, Cellular Movement

Network analysis using IPA software to identify molecular networks from the list of DEGs from the comparison of the high ESR group (n=117) to low ESR group(n=126) samples. Molecules not coloured are from the Ingenuity Knowledge Base and not present in the dataset. The shape of the gene represents the functional class of the gene product (key as for figure 4.14).

CRP

The samples in the high and low CRP groups did not separate out on a principal components analysis (figure 4.16). The high CRP group was significantly older (P 0.017, unpaired T test) than the low inflammation group but there was no significant difference in gender or the proportion of patients with RA between the groups.

The differential gene expression analysis of the samples in the low CRP and high CRP groups with age added to the linear model did not identify any differentially expressed probes which stood up to multiple test correction. When age was removed from the linear model, and the anlaysis repeated, 2 probes which were diffentially expressed with a FC ≥1.2 and met the multiple test correction criteria were identified (figure 4.16b). The probes upregulated in the high CRP group were both for GAS6, growth arrest specific 6, which encodes a gamma-carboxyglutamic acid (Gla)-containing protein believed to be involved in the stimulation of cell proliferation.



a) *b)*

Figure 4.16 Prinicpal components analysis and volcano plot for high and low CRP groups

a) PCA plot displaying samples identified based on high or low levels of CRP. There are no demonstrable differences between the two groups (red high inflammation, CRP ≥8mg/L, blue low inflammation, CRP <8mg/L). b) Volcano plot for high CRP (n=115) versus low CRP (n=128), FC1.2, MTC; Benjamini-Hochberg, no clinical variables added to the linear model. 2 differentially expressed probes identified. Vertical, dotted lines denote FC1.2. Horizontal, dotted lines denote p-value 0.05. Red dots indicate probes which are differentially expressed between the comparator groups.

4.4 Discussion

Age and the CD19^{+} B cell transcriptome

Large scale meta-analyses have confirmed the variation in gene expression with age and identified overexpression of inflammation and immune response genes, which is in keeping with the upstream regulators identified here and the network analyses[185, 197]. Published transcriptomic analyses focussing on the effects of aging in specific immune cell subsets are restricted to T cells and monocytes and so this is a novel analysis[187, 198]. The identification of the transcriptional changes which occur with age has the potential to identify altered molecular pathways, improving our understanding of age-related immune modulation and so identify therapeutic targets. The dissection of the pathways underlying the aging process is complicated by the combination of varying genetic influences and the accumulated environmental influences.

The results described in this exploratory experiment have identified several interesting upstream regulators for the DEG lists generated. In keeping with the concept of inflamm-ageing, pro-inflammatory cytokines have been identified as activated upstream regulators in the ≥ 55 years age group; in particular in the comparison of the < 55 years old group to the ≥ 55 years old group without the addition of ESR and CRP to the linear model. The pro inflammatory cytokines IL-1A, IL17A and its receptor IL17RA were identified as potential activated upstream regulators, alongside TNF in this comparison. In this discussion, I will focus on upstream regulators identified in both analyses between the two age groups.

APOE is an inhibited transcriptional regulator in the ≥55 years group which is identified in both analyses. *APOE* has an established association with aging and age-related diseases such as cardiovascular disease and Alzheimer's disease. Along with *FOXO3*, it is associated in multiple, independent genome wide association studies with longevity, the longevity phenotype being survival to ≥90 years old[199, 200].

The *APOE* gene encodes the protein apolipoprotein E, synthesised primarily in the liver but also produced in other tissues. It combines with lipids to form lipoproteins, playing an essential role in cholesterol metabolism. The *Apoe* deficient mouse (*Apoe*$^{-/-}$) is a widely used mouse model for human atherosclerosis as it displays poor lipoprotein clearance,

and develops hypercholesterolaemia, which promotes the development of atherosclerotic plaques[201]. A splicing analysis using RNA-seq data has recently shown that the expression of *APOE* exons and links change with age in skin tissue. This results in the production of different isoforms of the gene, favouring an isoform skipping the third exon of the gene in older individuals and this may alter protein function with age[188].

There is increasing interest in B cell function in the field of cardiovascular research and coronary heart disease, in particular in the use of B cell modulation in the treatment of atherosclerosis[202, 203]. In mouse models, it has been shown that functional B cells are protective of atherosclerosis[204, 205]. A systems biology approach using whole blood gene expression profiles from the Framingham Heart Study (using coronary heart disease (CHD) cases and age and sex matched controls) identified a module enriched for B cell activation, which demonstrated strong co-expression in controls but not CHD cases, suggesting a potential role for B cell dysregulation in atherosclerotic CHD[206]. There is, therefore, a potential link between dysregulated B cell function and atherosclerosis which is in turn linked to *APOE*.

APOE itself is not differentially expressed between the two sample groups in this experiment. The effects ascribed here to *APOE* by IPA are largely indirect associations, thus we may be seeing the effects of *APOE* from other sources influencing the B cell transcriptome. Changes in the isoform of *APOE* expressed in the two age groups may differ and as such the differences may not be adequately detected in a microarray experiment. Changes in *APOE* expression in B cells could be further investigated using RNA-seq experiments to look for differential expression of splicing variants and measuring paired serum *APOE* to examine alternative explanations for the gene expression changes presented here.

NOS2 encodes an inducible nitric oxide synthase which generates the second messenger nitric oxide. *NOS2* is not differentially expressed in the dataset but is shown to be an activated upstream regulator in the ≥ 55 years group. This is in keeping with the hypothesis that oxidative stress is a major factor in the aging process and age related diseases such as atherosclerosis[207, 208].

TGM2, transglutaminase 2, is a widely expressed, multifunctional enzyme which catalyses the crosslinking of proteins and has been identified as a potential biomarker of frailty. It is induced by pro inflammatory cytokines and accumulates in atherosclerotic plaques, playing a role in the atherosclerotic pathway by NFkB activation, TNF and NOS expression[209]. It has been associated with the physiological response to stress, apoptosis, inflammation and fibrosis, where cross linking of proteins in the extracellular matrix by *TGM2* increases resistance to breakdown[210, 211]. The encoded protein of *TGM2* is the autoantigen implicated in coeliac disease.[212].

IL-2 is primarily produced by T cells in the secondary lymphoid organs following antigen-mediated activation and is important for the maintenance of regulatory T cells and the differentiation of effector T cells subsets; the overall effect is finely tuned to the strength and duration of the IL-2 signal. The IL-2 receptor, IL-2Rα, is found on immature B cells and promotes the differentiation of primed B cells into plasma cells[213, 214]. Given that IL-2 production is increased in the setting of an immune response its identification here as an activated upstream regulator suggests a heightened activation state in the immune system as we age, although the trigger for this is unknown. This may also be relevant to the increase in autoantibody formation with age.

The identification of the pro-inflammatory cytokine TNF as an activated upstream regulator in the ≥55 years group is in keeping with the literature that serum TNF is elevated in older people and is one potential explanation of the chronic inflammatory state described[215]. A study on healthy volunteers found that unstimulated $CD19^+$ B cells isolated from older individuals (≥60 years old) produced more TNFα mRNA and this was positively correlated with serum TNFα. There was an observed reduction in activation-induced cytidine deaminase (AID) after stimulation with CpG in this older age group, which negatively correlated with pre-stimulation TNFα mRNA levels; leading the group to suggest that the intrinsic change to B cells in the older group prior to stimulation rendered them unable to respond optimally to antigen[215]. The cytokine changes with aging are likely to be complex as there is also evidence that IL-10 producing B cells decrease with age and an additional factor, in the observed effect of TNF, may be the altered balance between the production of IL-10 and TNFα production by B cells[216].

It is interesting to note that two of the upstream regulators TNFα and *NOS2* are both considered evolutionarily preserved mediators to the cellular response to stress and *TGM2* is upregulated in response to stress. The changes seen in aging have been described as a reduction in the ability to cope with stress. The trigger that leads to this activated stress response is unknown but antigen has been suggested in this role[217].

Inflammation and the CD19^{+}B cell transcriptome

In the analyses described, dividing the samples by ESR measurement, not CRP, led to the identification of a number of differentially expressed genes. This is despite a large overlap in the samples in both high inflammation groups. In this population, inflammation as measured by ESR is an indicator of changes in gene expression. The IPA analysis for the comparison between groups based on ESR with age added to the linear model yielded a list of DEGs, however, the analysis using IPA was uninformative and so I will focus on the analysis of high ESR *versus* low ESR without clinical covariates added to the linear model.

The IPA analysis identified two inhibited upstream regulators *BNIP3L* and *TOB1*, in the high ESR group, however, there is limited data in the literature regarding these genes. Their inhibition promotes proliferation and cell survival as *BNIP3L* is pro-apoptotic and *TOB1* anti-proliferative. *BNIP3L* belongs to the pro-apoptotic subfamily within the Bcl-2 family. *TOB1* is a member of a family of anti-proliferative factors and may function as a tumour suppressor, it has been shown to be highly expressed in unstimulated T cells and is downregulated on activation ([173],[218]). *TOB1* knockout mice show an increase in B cell proliferation both when unstimulated and when stimulated with LPS[219]. The inhibition of *TOB1* in the high inflammation group here is in keeping with the findings in T and B cells, suggesting it may be a negative regulator of the immune response and important in maintenance of quiescence in immune cells[218, 219].

The activated regulators in this sample group include transcription factors, several of which have been implicated in tumorigenesis: *ERBB2*, *EGFR*, *RABL6*, *RARA*, *ESR1* and *E2F2*. *ERBB2*, also known as *HER2*, is connected to 16 molecules in the dataset; like *EGFR*, it is a receptor tyrosine kinase and dysregulation of their downstream signalling is associated with certain cancers. Overexpression of *ERBB2* is, in particular, associated

with breast cancer associated with a poor prognosis and clinically targeted by trastuzumab (Herceptin)[220].

In B cells the transcription factor *XBP1* is important for the development of plasma cells, and is considered to act downstream of *BLIMP1*, its expression increases the secretory apparatus and protein synthesis, thereby co-ordinating the cellular changes in structure and function required for plasma cells[221, 222]. *XBP1* is also involved in the unfolded protein response in other tissues. It has been identified in GWAS as a risk factor for IBD and work in the IBD field has shown that deletion of *XBP1* results in ER stress [223]. The activation of its downstream pathway reflects an activated immune response and response to stress.

CSF2, also known as *GM-CSF*, is a growth and differentiation factor which acts on cell types of different lineages via its receptor but the expression of this on lymphocytes is debated. There are published data demonstrating that *GM-CSF* is produced by B cells; this increases on activation and promotes cell survival, in keeping with an activated immune response[224, 225].

The IPA analysis indicates that, in the high ESR group, $CD19^+$ B cells are activated and more able to survive and proliferate. The molecular network analysis includes the BCR complex, NFkB complex and PI3K complex to link molecules within the network which suggests there may be an alteration in B cell signalling in the setting of inflammation.

In summary, the identification of multiple test corrected DEGs between groups divided by both age and level of inflammation highlights the influence of both factors on the $CD19^+$ B cell transcriptome. The analysis based on age was the most productive, providing an interesting insight into the potential additional effect of *APOE* on B cells as we age, beyond the established focus on cardiovascular and neurodegenerative conditions. The upstream regulators identified also confirm the concept of a heightened inflammatory state with age, with the identification of activated pathways downstream of pro inflammatory cytokines and pathways related to stress.

In this analysis ESR was a more effective discriminator between samples than CRP, at the thresholds chosen. The clinical utility of ESR and CRP varies with clinical context and

this result potentially marks ESR as more likely to reflect changes in the $CD19^+$ B cell population than CRP. The upstream regulators identified in this analysis are indicative of pathways promoting cell survival and increased cellular activity, which would be expected during an immune response. In keeping with this, the top network identified relates to DNA Replication, Recombination, and Repair, Cell Cycle, Cellular Movement.

In this chapter, my analyses have shown that age and inflammation, here measured by ESR, are important factors determining gene expression in $CD19^+$ B cells and may be contributory factors to help explain the absence of a clear DEG signature in *Chapter 3*. The results have also provided novel insights into the influence of age on the transcriptome and confirmed the pro-inflammatory state present as we age.

The results described here are exploratory and an extension of the efforts to identify differentially expressed genes related to diagnosis in the early arthritis clinic.

4.5 Future work

- Further examine the data with a more relaxed FC cut off, for example FC ≥1.1 for CRP to identify possible DEGs for further analysis

APOE

- Identify genes from the literature, alongside those from this study, downstream of APOE and confirm changes in *APOE* related genes in $CD19^+$ B cells with RT-PCR.
- Measure *APOE* expression in B cells and, if present, identify splice variants of *APOE* in B cells.
- Measure serum levels of *APOE* and correlate with changes in gene expression

TNF

- Investigate TNF production by $CD19^+$ B cells further by looking specifically at which B cell subsets produce TNF in healthy controls
- Pair serum TNF measurements and correlate with changes in gene expression in putatively regulated genes

IL-2

- Measure serum IL-2 levels and relate this to age with parallel measurements of IL-2 receptor expression on B cells subsets.

5. Integration of genotype, expression profiling of peripheral blood B cells and clinical data

5.1 Background

There is an established genetic component to RA identified through GWAS and twin studies. The strongest association is with the MHC region and, more specifically, genetic variation in the HLA-DRB1 locus, but also the HLA-DPB and HLA-B loci; accounting for a large proportion of the risk in cases of seropositive RA[119]. The additional 101 RA risk loci identified from GWAS studies are frequently in non-coding regions of the genome or tag LD blocks containing more than one gene. Identifying the most likely candidate gene, or genes, at each of these loci is challenging as, on average, each has 4 genes within the region of LD[107]. Indeed, the mechanisms by which genetic variation predisposes people to a given disease remain poorly understood. Given that the majority of GWAS variants are non-coding, it is postulated that their effects are likely to be regulatory and so understanding the mechanisms by which they regulate the genome is also essential to deciphering the significance of these variants[226].

The number of genes potentially implicated in complex traits limits the use of gene knock down or overexpression studies to identify causal variants without the earlier prioritisation of likely candidates. To do this, and potentially establish the biological consequences of the genetic variation, expression quantitative trait locus (eQTL) mapping has been used to analyse the effect of genetic variants on total gene expression levels. However, the heritable component of gene expression also varies across cell types and is context specific[132]. The context specificity highlights the potential flaws in using cell lines to investigate genetic variants as they may display different influences compared to *in vivo.* In addition, cell specific effects may be undetectable in an eQTL analysis of a heterogenous population such as PBMCs where B cells make up 5-10% of the cells examined and so transcriptional changes may be masked by those from other cell types.

Cis eQTLs have been shown to have a larger effect size than *trans* eQTLs in humans and so relatively small sample sizes have been used to detect *cis* eQTLs[184, 227]. In B cells, Fairfax *et al*, using 283 samples from healthy volunteers (a cohort of comparable size to my dataset) identified over 40,000 *cis* eQTLs (eSNPs) (minor allele frequency (MAF) 0.01, permuted P <0.001). In this study, paired monocyte data were available which demonstrated that the majority of eSNPs were unique to B cells (26,549 eSNPs unique to B cells and 17,962 shared with monocytes). Interestingly, some shared eSNPs displayed opposing directional effects on gene expression in a cell-specific manner, emphasising

the importance of examining cell subsets to interpret GWAS studies. The group further examined eQTLs involving SNPs identified as disease risk alleles for common traits (based on Catalog of Published Genome-Wide Association Studies accessed in 2011) and showed that, after excluding the HLA loci, 14 of the cis eQTLs identified in B cells were shared with GWAS SNPs associated with RA. Of these, 6 were shared eQTLs with monocytes (*ANAPC4*, *C5*, *FAM119B*, *MEGF9*, *VARS2*, *XRCC6BP1*) and 8 unique to B cells (*ABCG1*, *BLK*, *CCR6*, *COL8A2*, *DIP2A*, *FAM167A*, *GIN1*, *MSRA*). An additional 7 RA SNPs were specific to monocytes[132].

Environmental conditions were first shown to affect a proportion of eQTLs in model organisms such as yeast[228, 229]. The identification of similar, context-specific eQTLs in the setting of complex diseases is more challenging, as the environmental change is less clearly defined and accurate measurement of exposure levels of a given factor difficult to measure, in contrast to *in vitro* work. However, analysis of combined datasets confirms the effects of sex, age, treatments such as glucocorticoids and exposure to immune stimuli on gene regulation[133, 136, 229-231]. This raises the possibility that a proportion of eQTLs may only be evident in the disease state.

Peters *et al* examined this *in vivo,* looking for eQTLs which differed between the disease state, in this case inflammatory bowel disease (IBD), and health using an interaction analysis. They used a linear model with a genotype x disease interaction term (GxD) to look for differences in the effect of genotype on expression in the disease and healthy volunteer cohorts. A significant GxD interaction term for a SNP-gene pair indicated that the effect of the given genotype on expression was significantly different in disease (IBD in this case) and health. The eQTL may have a greater or lesser effect in disease, be absent in health but present in disease, or vice versa. In essence, they asked if there was a significant difference in the slope of the genotype-expression regression line between the two groups.

The group identified 13 eQTLs with a GxD interaction effect across different cell subsets: $CD4^+$ and $CD8^+$ T cells, monocytes and neutrophils. There were no B cell data available due to the relatively low number of available samples for this subset. The approach described by Peters *et al* has the potential to identify novel pathways for disease pathogenesis and the genes identified are not all immune related[133].

The same group also showed that a proportion of eQTLs identified in an ANCA associated vasculitis cohort disappear with treatment, highlighting the importance of studying, not only a disease cohort but also untreated disease to gain insights into pathogenesis. In common with other complex diseases the missing heritability may also be unravelled by this approach, identifying novel pathways for disease development[111].

A similar analysis of a multi-tissue RNA-sequencing dataset identified 16 *cis* genotype x body mass index (GxBMI) interactions, all of which were specific to adipose tissue and a subset were replicated in an independent dataset. The genes identified were enriched for inflammatory and metabolic genes and may provide insights into the potential mechanisms for the heterogeneity in obesity related co-morbidities[232].

In this thesis, the eQTL landscape, at known RA risk loci, in $CD19^+$ B cells from early arthritis patients was explored. The published findings can be found at the end of this thesis (*Appendix C)* and an overview of the findings is provided in this chapter.

The principle used by Peters *et al* will be applied and adapted to integrate disease state data for the $CD19^+$ B cell samples in this cohort and thereby identify disease specific eQTLs. This chapter will focus on work undertaken to seek differences in eQTL effects between RA and non-RA cohorts.

5.2 Hypothesis and Aims

5.2.1 Hypothesis

By integrating genotype, B cell gene expression and clinical data in untreated RA patients, I hypothesise that novel insights into mechanisms of disease could be revealed. Specifically, an eQTL analysis of confirmed genetic risk loci in RA will provide insights into the mechanism of genetic risk in RA. Characterising the disease related specificity of variations in gene expression may provide insights into the biological significance of earlier discoveries from traditional GWAS approaches and alternative mechanisms of disease pathogenesis.

5.2.2 Aims

1. Carry out an eQTL analysis of confirmed RA genetic risk loci in $CD19^+$ B cells from patients presenting to an early arthritis clinic to gain further insight into the potential mechanisms of genetic risk in RA
2. To identify a list of *cis* eQTLs in B cells which exert different effects in DMARD-naive RA patients compared to a non-RA patients

5.3 Analytical Methodology

5.3.1 eQTL analysis of RA risk loci

Cis and *trans* eQTLs were sought, focussing on variants in LD (defined as $r^2 \geq 0.8$) with non-HLA RA SNPs confirmed in Caucasians and described by Okada *et al* [107]. The analysis was limited to non-HLA variants due to cross-hybridisation of expression probes and confounding effects of copy number variants in the HLA region. The bioinformatics analysis, using the MATRIX eQTL package in R, was completed by Andrew Skelton in the Bioinformatics Unit, in parallel with that from $CD4^+$ T cells from the NEAC patients.

5.3.2 Interaction analysis

The principle used by Peters *et al* was based on a two step procedure; in step 1 gene-SNP pairs were identified by linear regression of expression on genotype (no covariates included) and SNPs reaching the selected criteria were taken forward and a full model fitted with regression of genotype (0,1,2), disease (0,1) and genotype x disease interaction term. The aim of step 1 was to reduce the number of tests and improve the ability to detect significant changes. However, by analysing the cohort in one run, without stratifying the data in step 1 into disease and non-disease cohorts there is the potential to miss a proportion of interaction effect SNPs, as they may not appear as eQTLs. In light of this concern, the method used for this analysis was the LINEARCROSS function in MATRIX eQTL; as used by Glastonbury *et al*[232]. The function in matrix eQTL 'modellinear_cross' was employed to indicate that MATRIX eQTL include the interaction of SNP and covariate (disease) in the model and test for its significance.

The MAF was set at 0.05 and FDR threshold was set at 5%. Peters *et al* used an FDR threshold of 15% and MAF 0.1.

For the genome wide interaction analysis the disease group was made up of all the RA samples, combining seropositive and seronegative patients, and the control group comprised the remaining samples; excluding those with a diagnosis of UA. The choice of 'RA' and 'non-RA' mirrored the earlier analyses described in *Chapter 3*, and had the advantage of maximising the number of samples available. This comparison has been the most informative in the previous analyses of $CD19^+$ B cell gene expression in this cohort. Alternative control groups were considered, in particular, non-inflammatory samples with normal CRP and ESR (n=32), however, given that this is an exploratory analysis, group size was prioritised for the choice of analysis and consistency of analysis groups.

5.4 Results

5.4.1 eQTL analysis of RA risk loci in $CD19^+$ B cells

The *cis*-eQTL analysis of non-HLA RA risk loci in $CD19^+$ B cells identified 194 *cis*-eQTLs and this corresponded to 10 genes at 7 loci: *FAM167A*, *FADS1*, *ORMDL3*, *FADS2*, *FCRL3*, *GSDMB*, *SYNGR1*, *BLK*, *PPIL3* and *CD83* (figure 5.1). At the 8p23 locus *FAM167A* and *BLK* were both subject to *cis* regulation, however, the *FAM167A* eQTL was specific to $CD19^+$ B cells and not seen in the $CD4^+$ T cells.

The incorporation of disease phenotype (RA *versus* non-RA) to the linear model did not alter the final eQTL list. There was no evidence of disease-specific eQTLs at RA risk loci.

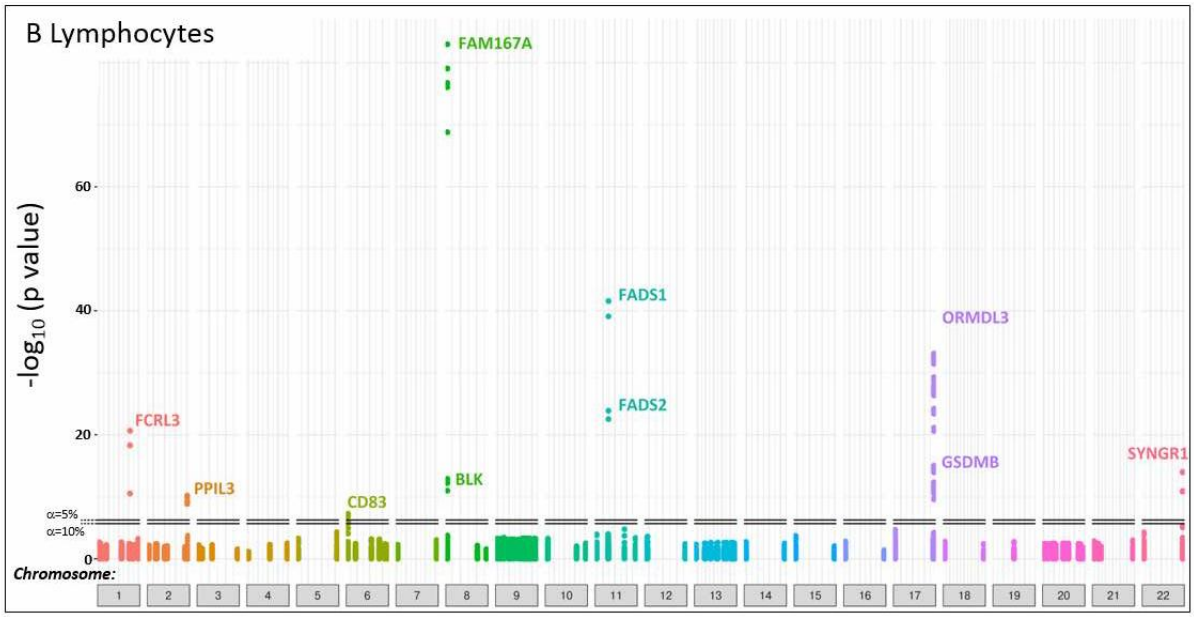


Figure 5.1 Manhattan plot of the 101 RA risk loci analysed

The 101 RA risk loci analysed are shown and P-values for the single nucleotide polymorphism (SNP)-probe pairs in CD19^{+} B cells in early arthritis patients. Human Genome Organisation gene symbols for SNP-probe pairs that reached, or approached, experiment-wise significance (at thresholds of α= 5% and α=10%, horizontal lines) are indicated.

5.4.2 Demographics for interaction analysis RA and Non-RA samples

177 samples were used in the interaction analysis. The RA group was significantly older, with significantly higher levels of inflammation as measured by CRP and ESR (table 5.1). The age differences are not unexpected given the known age profile of RA patients and the inclusion of patients with non-inflammatory conditions within the control group, non-RA. The non-RA group includes 55 patients diagnosed with other inflammatory conditions and 65 patients with non-inflammatory conditions. In the RA group 37 of the 57 patients were ACPA positive.

	RA	Non-RA	P-value
Sample number	57	120	-
Age (yrs)	61 (21-89)	51.5 (18-92)	0.0004
Gender (%F)	75.4	70.8	ns
ESR (mm/hr)	25 (1-91)	9.5 (0-111)	0.0001
CRP (mg/L)	10 (0-91)	5 (0-171)	0.0008
SJC	1 (0-25)	0 (0-11)	<0.0001
TJC	6 (0-22)	2 (0-28)	0.0008
DAS28	4.59 (1.26-8.56)	3.36 (0-6.99)	<0.0001

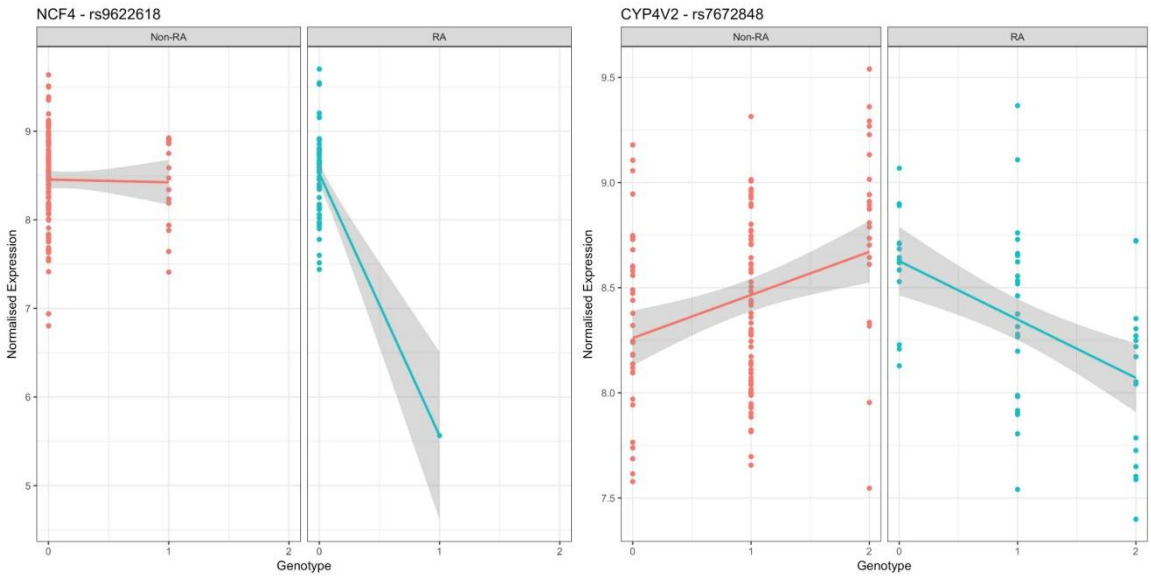
Table 5.1 Demographics for interaction analysis for RA and non-RA samples

Median and range shown. P-values were calculated using unpaired T-test (age and DAS28), Fisher's exact test (gender), or Mann-Whitney test (ESR, CRP, SJC, TJC).

5.4.3 Interaction analysis – RA and non-RA

The RA and non-RA data were analysed to identify genome wide *cis* eQTLs which differed between the two groups; meaning the effect of genotype on expression is significantly different in the RA and non-RA cohort.

The initial analysis identified SNPs related to 64 genes where such an effect was seen. However, visualisation of the data demonstrated a marked imbalance in the genotype counts in a proportion of these results, such as the absence of samples homozygous for the minor allele at the SNP examined. Figure 5.2 depicts one example of a spurious finding at rs9622618, *NCF4*, alongside an example of a valid plot for rs7672848, *CYP4V2*. Indeed, observation of the plots in the paper by Peters *et al* demonstrate the potential flaws. The absence of filtering for genotype counts can lead to skewed genotype-expression regression lines in one cohort, potentially due to one sample, and the identification of a false positive interaction effect. This is particularly an issue in studies with small cohorts such as this.



a) *b)*

Figure 5.2 Exemplar plots of disease specific eQTLs in the absence of filtering for genotype counts

a) disease specific eQTL for NCF4 demonstrates a skewed genotype expression regression line in the RA group and *b) the disease specific eQTL for CYP4V2 is an example of balanced genotype counts between groups. Plots of expression of a given gene on the vertical axis and genotype on the horizontal axis. Each data point represents one sample. The coloured lines drawn on the plots represent the estimated best fit line of the regression of expression on genotype in each group. The non-RA group is shown in red and RA in blue.*

There are various possible approaches to reduce the number of false positive results in this setting: increasing the MAF or adding genotype count filters either pre-analysis or post-analysis. The published interaction analyses do not describe this issue in detail in their methods sections.

To address this issue the MAF was first increased from 0.05 to 0.1, the level used by Peters *et al.* In contrast to the addition of a genotype count, this approach did not require the inclusion of an additional step to the analysis pathway. However, this decreased the number of genes identified in the analysis to 4 (*CSNK1G3, LRRC58, P2RX4* and *POLR3H*, which were also identified in the subsequent analyses) and so this option was rejected, given that this analysis is primarily for discovery purposes. In addition, when rerunning the analyses with control groups with smaller sample sizes no genes with interaction effects were identified, highlighting the need to adapt the methodology for discovery approaches.

Next, a genotype count filter was included prior to running the analysis. A minimum count of 5 samples for each genotype was tested and found to be too stringent; no disease related *cis* eQTLs were identified.

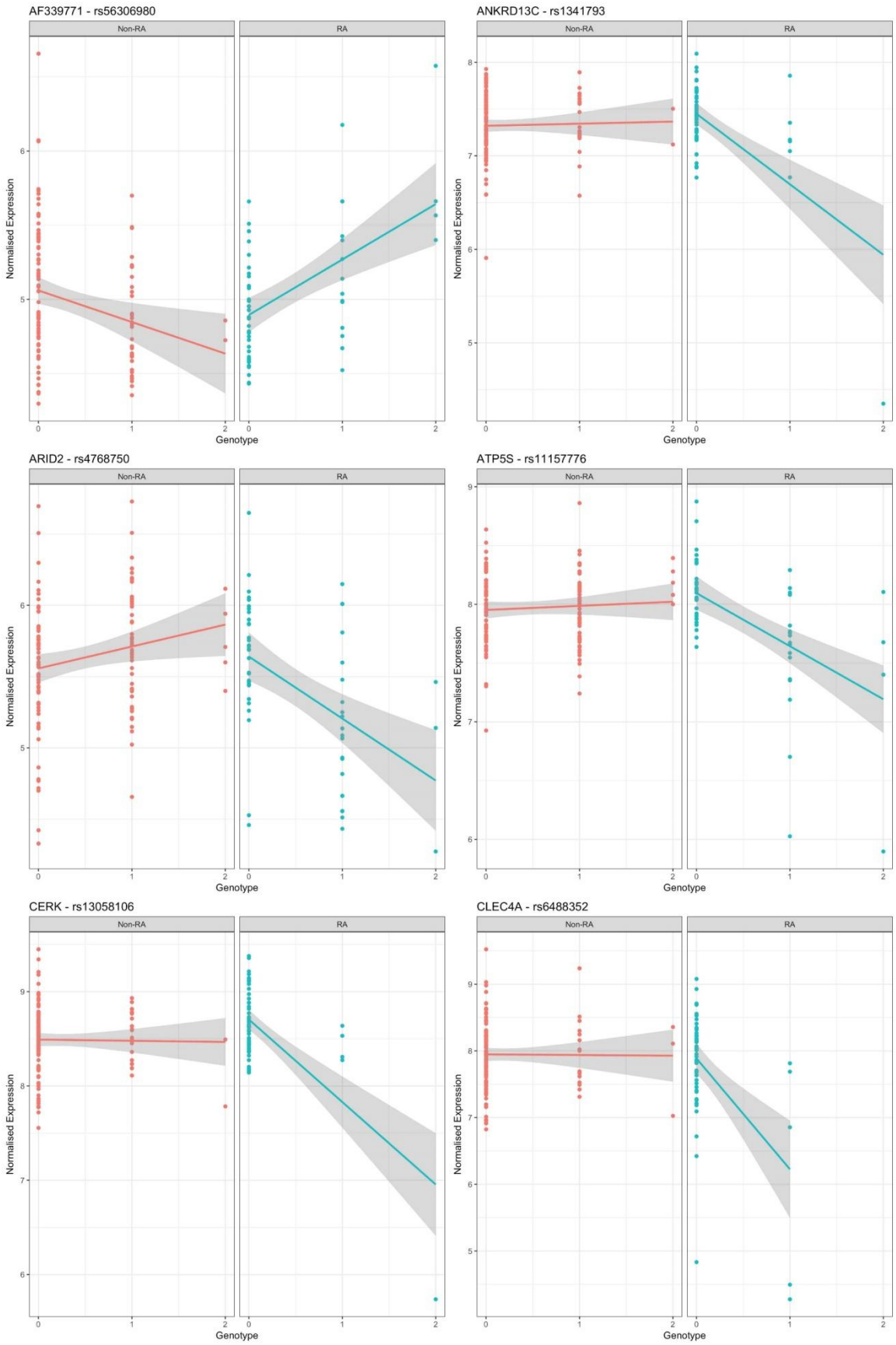
The third option, a post analysis genotype filter, was selected for discovery purposes. The genotype count filter of ≥3 samples in each of the three genotype groups (prior to stratification by diagnosis) was selected to balance the considerations of false positives and false negatives in the analysis. After applying this post hoc filter to the data, 21 individual genes, subject to a GxD interaction effect in $CD19^{+}$ B cells, were identified (table 5.2), in each case the SNP with the lowest p-value is shown. The individual plots are shown in figure 5.3 allowing visualisation of the plots which is useful for the interpretation of results.

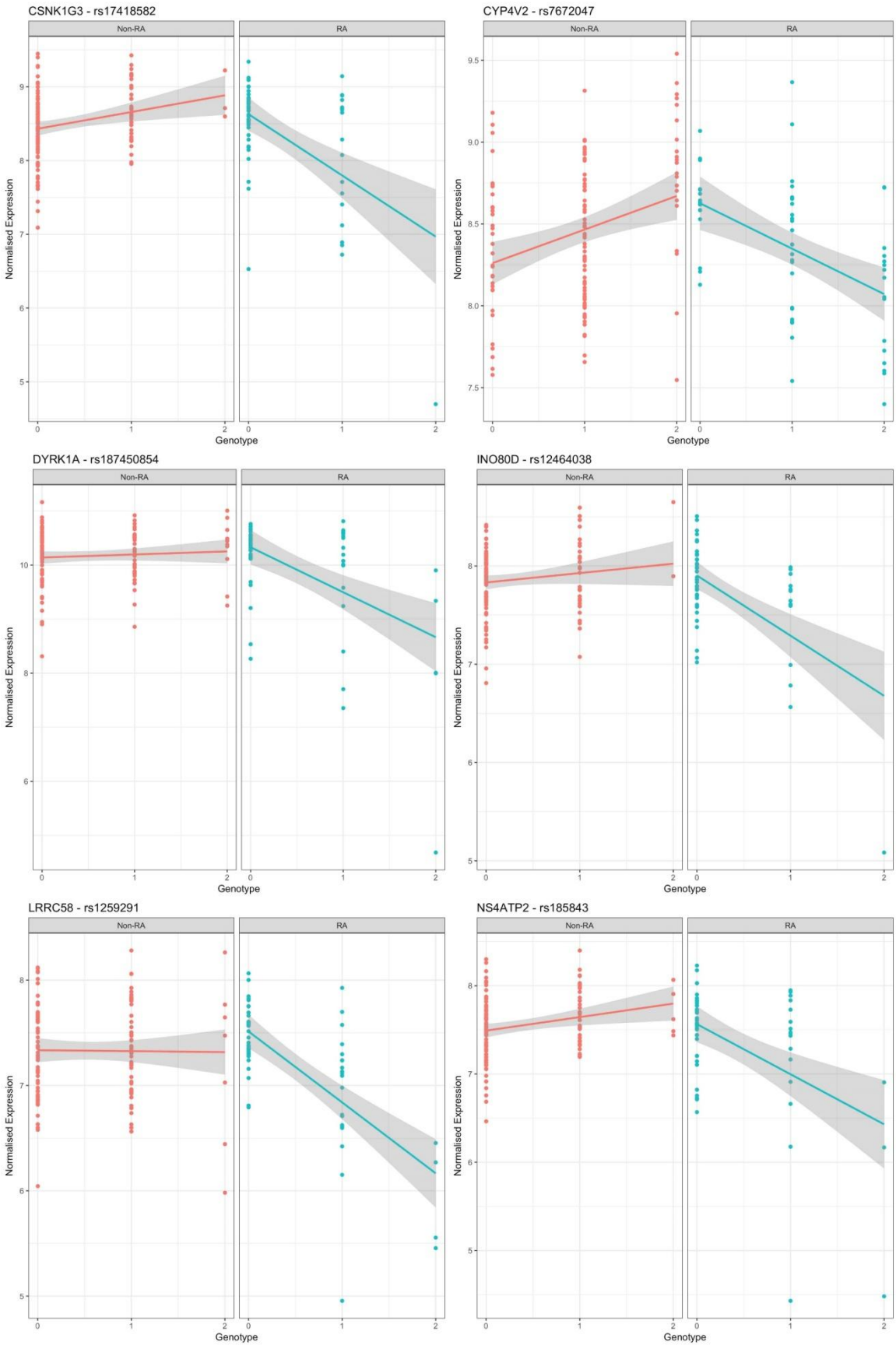
The plots in figure 5.3 all met the criteria set for statistical significance and genotype counts, however, the *ANKRD13C* and *CERK* plots may represent false positives, skewed by the low expression of a RA minor allele homozygote sample in each case. In the case of *INO80D*, there is again, a RA minor allele homozygote with low expression but the plot appears more convincing as the heterozygotes in the RA sample demonstrate lower levels of expression than the non-RA group. It is, of course, easier to be convinced by plots where the spread of samples is more balanced between the homozygotes and heterozygotes in both the RA and non-RA groups as for *CYPV42*.

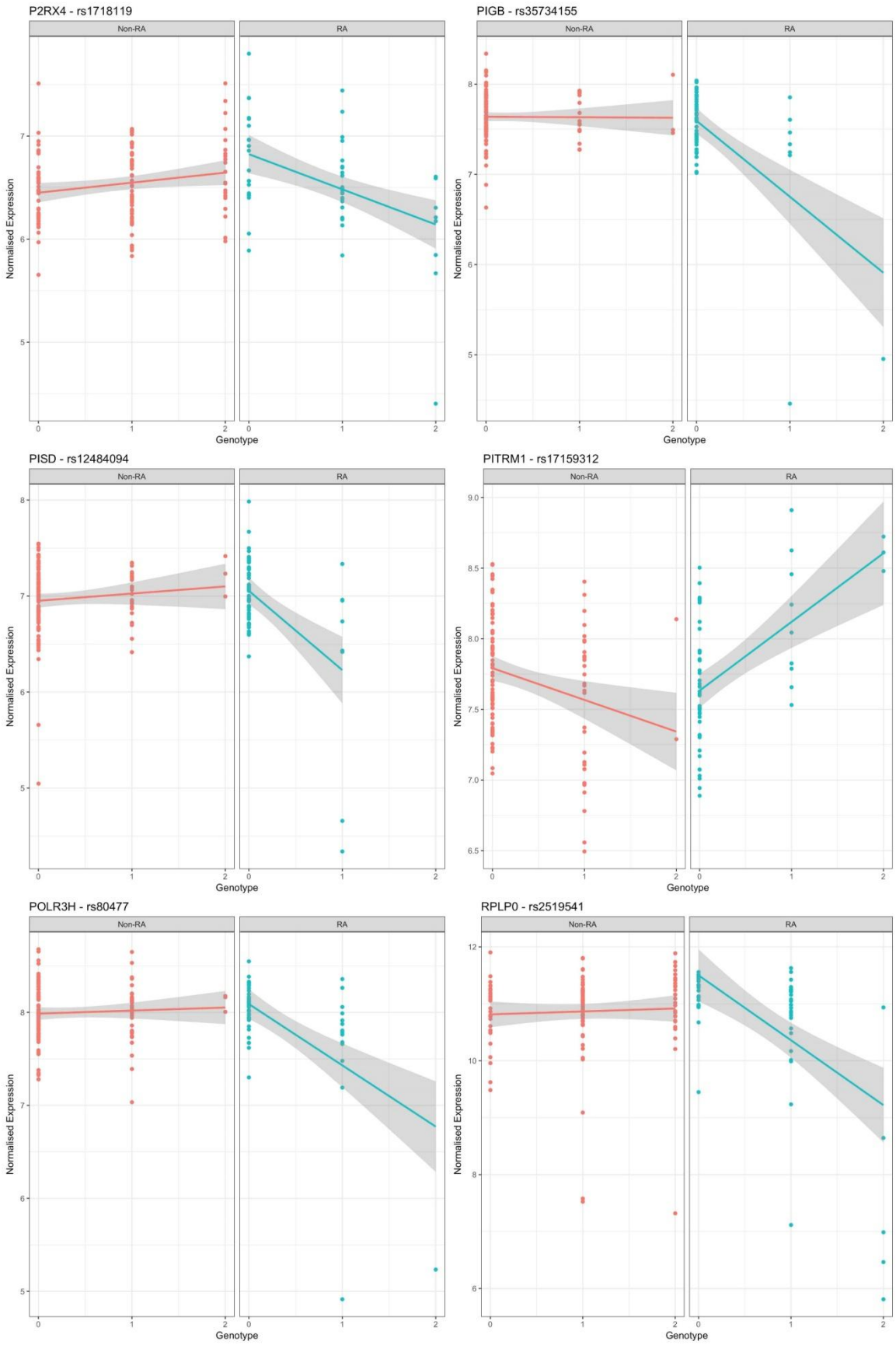
Gene Symbol	SNP	P-value	FDR corrected p-value	Gene Description
AF339771	Rs56306980	7.15478E-07	0.024485652	DOCK9 DT (divergent transcript)
ANKRD13C	Rs1341793	1.20149E-07	0.006552875	Ankyrin repeat domain-containing protein 13C
ARID2	Rs4768750	6.35435E-07	0.022313361	AT rich interaction domain 2 (DNA binding protein)
ATP5S	Rs11157776	2.42387E-07	0.010610813	distal membrane arm assembly complex 2 like (mitochondria)
CERK	Rs13058106	5.56412E-08	0.004017125	ceramide kinase (sphingolipid metabolism)
CLEC4	Rs6488352	9.13765E-07	0.029844095	C-type lectin domain family 4 member A
CSNK1G3	Rs17418582	1.15783E-08	0.002223959	casein kinase 1 gamma 3
CYP4V2	Rs7672047	8.52922E-08	0.005207385	cytochrome P450 family 4 subfamily V member 2
DYRK1A	Rs187450854	3.63226E-07	0.014744806	dual specificity tyrosine phosphorylation regulated kinase 1A
INO80D	Rs12464038	9.02861E-08	0.005355195	INO80 complex subunit D
LRRC58	Rs1259291	2.70393E-07	0.011464193	leucine rich repeat containing 58
NS4APT2	Rs185843	1.14945E-07	0.006372081	SAP30 like
P2RX4	Rs1718119	8.9594E-07	0.029387677	purinergic receptor P2X 4
PIGB	Rs35734155	3.40315E-09	0.001011888	phosphatidylinositol glycan anchor biosynthesis class Bp
PISD	Rs12484094	7.7126E-07	0.026121367	phosphatidylserine decarboxylase
PITRM1	Rs17159312	1.32716E-07	0.006619933	pitrilysin metallopeptidase 1 (mitochondria)
POLR3H	Rs80477	3.24247E-08	0.003258653	RNA polymerase III subunit H
RPLP0	Rs2519541	3.44591E-07	0.014088288	ribosomal protein lateral stalk subunit P0
SUN1	Rs28548330	2.1397E-07	0.009402815	Sad1 and UNC84 domain containing 1
TMEM62	Rs1991324	7.50871E-07	0.025506298	transmembrane protein 62
TTC31	Rs1474335	8.78238E-07	0.028931371	tetratricopeptide repeat domain 31

Table 5.2 cis eQTLs displaying different effects in RA and non-RA groups

For each individual gene the SNP with the lowest P-value is shown. Gene description from NCBI Gene database[173].


AF339771 - rs56306980
Non-RA
RA
Normalised Expression
Genotype
ANKRD13C - rs1341793
Non-RA
RA
Normalised Expression
Genotype
ARID2 - rs4768750
Non-RA
RA
Normalised Expression
Genotype
ATP5S - rs11157776
Non-RA
RA
Normalised Expression
Genotype
CERK - rs13058106
Non-RA
RA
Normalised Expression
Genotype
CLEC4A - rs6488352
Non-RA
RA
Normalised Expression
Genotype


CSNK1G3 - rs17418582
CYP4V2 - rs7672047
DYRK1A - rs187450854
INO80D - rs12464038
LRRC58 - rs1259291
NS4ATP2 - rs185843
Non-RA
RA
Normalised Expression
Genotype


P2RX4 - rs1718119
PIGB - rs35734155
PISD - rs12484094
PITRM1 - rs17159312
POLR3H - rs80477
RPLP0 - rs2519541
Non-RA
RA
Normalised Expression
Genotype

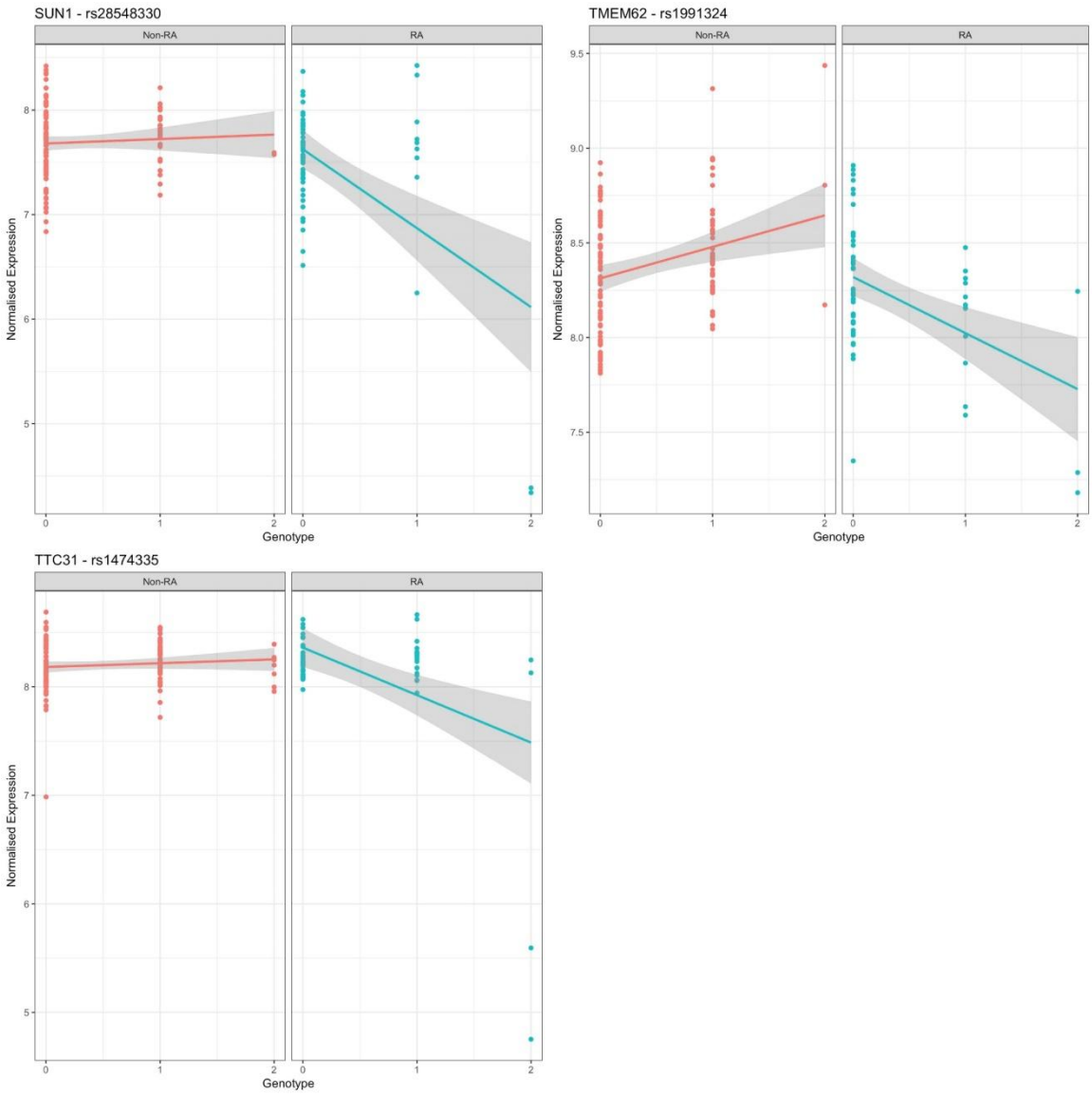


Figure 5.3 cis eQTLs with distinct effects related to disease state

Individual eQTL plots for each gene identified from the analysis using RA and non-RA sample groups, plots shown represent SNP with lowest P-value for the interaction effect for each gene. Each plot shows expression against genotype for the SNP. Genotype 0 major allele homozygote, 1 heterozygote, 2 minor allele homozygote. Non-RA shown in red, RA shown in blue. Lines demonstrate the best fit linear regression line of expression on genotype for each group.

20 of the 21 genes identified are protein coding. *AF33977*, also known as *DOCK9* divergent transcript, is a validated non-coding RNA (ncRNA). The term ncRNA refers to RNA which is not translated to protein. There are several different categories of ncRNA, including long non-coding RNA and micro RNAs, which have been shown to potentially have functional roles in the cell, particularly in the regulation of gene expression. The remaining genes are protein coding and will be summarised in turn, the functional annotation, is from the NCBI Gene website unless otherwise stated[173].

ANKRD13C, ankyrin repeat domain 13C. Ankyrin repeats are widely found protein motifs which mediate protein-protein interactions. Understanding of the functional role of the *ANKRD13* family is limited, it consists of 4 members, A, B, C and D. *ANKRD13C* is the only one of the 4 members not to contain an ubiquitin-interacting motif[233]. It localises to the cytosolic side of the endoplasmic reticulum and is thought to act as a molecular chaperone for G protein-coupled receptors, regulating their expression and exit from the endoplasmic reticulum[234]. Examination of the plot demonstrates one sample which may be skewing the data, there is a single minor allele homozygote in the RA group and its expression level is markedly lower than for the other samples.

ARID2 AT-rich interaction domain 2. *ARID2* is a DNA-binding protein which functions as a subunit of the PBAF (SWI/SNF-B) chromatin remodelling complex that regulates chromatin accessibility for transcription factors. The PBAF complex is a tumour suppressor and mutations in its subunits, including *ARID2*, are associated with human cancers including melanoma and hepatocellular carcinomas. *ARID2* has been shown to suppress the expression of interferon-γ responsive genes. The PBAF complex reduces chromatin accessibility for IFN-γ inducible genes, increasing resistance to T cell–mediated cytotoxicity; targeting the PBAF complex is under investigation as a potential mechanism to overcome this resistance in tumour cells [235].

ATP5S, official symbol *DMAC2L*, distal membrane arm assembly complex 2 like. *ATP5S* encodes a subunit of the mitochondrial ATP synthase which catalyses ATP synthesis, this subunit is necessary for the energy transduction activity of the ATP synthase complexes.

CERK ceramide kinase. *CERK* phosphorylates ceramide to the sphingolipid metabolite, and lipid mediator, ceramide 1-phosphate (C1P). Both *CERK* and C1P have been implicated in various cellular processes and are thought to be involved in autocrine and paracrine signalling, and cell growth and survival[236]. The exact physiological role of ceramide and C1P have yet to be elucidated but there is increasing interest in their role in cancer and inflammation. Examination of the plot again demonstrates one sample which may be skewing the data. There is a single minor allele homozygote in the RA group and its expression level is markedly lower than for the other samples.

CLEC4A C-type lectin domain family 4 member A. A member of the C-type lectin/C-type lectin-like domain (CTL/CTLD) superfamily, transmembrane proteins which can bind Ca^{2+} and carbohydrate ligands which are derived from pathogen or self. They may have a potential role in both innate and adaptive immunity and in differentiating between self and non self.

CSNK1G3 casein kinase 1 gamma 3. It encodes a member of a family of serine/threonine protein kinases that are regulators of signal transduction pathways.

CYP4V2 cytochrome P450 family 4 subfamily V member 2. A member of the cytochrome P450 hemethiolate protein superfamily which is involved in oxidizing various substrates in the metabolic pathway. It is implicated in the metabolism of fatty acid precursors into polyunsaturated fatty acids but its exact function is unclear. The individual plot for this is convincing, showing that genotype has an effect in opposing directions in the RA and non-RA groups. The functional consequences of potential differences in energy generation are not clear.

DYRK1A dual specificity tyrosine phosphorylation regulated kinase 1A. *DYRK1A* contains a nuclear targeting signal sequence, a protein kinase domain and a leucine zipper motif. It is localized in the Down syndrome critical region of chromosome 21, and is considered to be a strong candidate gene for learning difficulties associated with Down syndrome, hence the focus in the literature in the field of intellectual disability. It has a critical role in lymphopoeisis, where *DYRK1A* has a major role in the transition from the highly proliferative state of pre-B and pre-T cells to the quiescence of later development (where cells exit the cell cycle to undergo further differentiation), by negatively

regulating cyclin D3 levels. Loss of *DYRK1A* therefore impairs cell cycle exit and lymphocyte differentiation[237]. Expression of *DYRK1A* is not related to genotype in the non-RA group, however, an eQTL can be seen in the RA group.

INO80D INO80 complex subunit D. This is a putative regulatory component of the chromatin remodelling INO80 complex which regulates nucleosomes and is important for responses to DNA damage. Dynamic modifications in chromatin are vital for maintaining DNA repair [238]

LRRC58 leucine rich repeat containing 58. The protein is validated but its functional role is unknown.

NS4APT2, official symbol *SAP30L,* Sin3A associated protein 30 like. The protein is a component of the histone deacetylase complex, deacetylation favours a closed chromatin structure and inhibition of transcription.

P2RX4 purinergic receptor P2X4. The product of this gene belongs to the family of purinoceptors for ATP. This receptor functions as a ligand-gated ion channel with high calcium permeability. The main pharmacological distinction between the members of the purinoceptor family is their relative sensitivity to the antagonists suramin and PPADS. The product of this gene has the lowest sensitivity for these antagonists.

PIGB phosphatidylinositol glycan anchor biosynthesis class B. This transmembrane protein is located in the endoplasmic reticulum and is involved in glycosylphosphatidylinositol GPI-anchor biosynthesis, anchoring proteins to the cell surface. There are 2 RA samples, one heterozygote and one minor allele homozygote, each associated with lower gene expression than the remaining samples, which are skewing this plot.

PISD phosphatidylserine decarboxylase. The enzyme is involved in phospholipid metabolism, catalysing the conversion of phosphatidylserine to phosphatidylethanolamine in the inner mitochondrial membrane.

PITRM1 pitrilysin metallopeptidase 1**.** This ATP-dependent metalloprotease degrades post-cleavage mitochondrial transit peptides. The encoded protein binds zinc and can also degrade amyloid beta A4 protein, suggesting a possible role in Alzheimer's disease.

POLR3H RNA polymerase III subunit H. Pol III is responsible for the transcription of housekeeping genes, it is tightly regulated during the cell cycle and its activity changes during cell differentiation[239]. There are 2 RA samples, one heterozygote and one minor allele homozygote associated with lower gene expression than the remaining samples which may be skewing this plot.

RPLP0 ribosomal protein lateral stalk subunit P0**.** This encodes a ribosomal protein that is a component of the 60S subunit.

SUN1 Sad1 and UNC84 domain containing 1**.** This encodes a nuclear envelope protein involved in nuclear anchorage and migration. It has been shown that *SUN1* overexpression prevents human immunodeficiency virus 1 nuclear entry[240].

TMEM62 transmembrane protein 62. Its functional role remains to be elucidated.

TTC31 tetratricopeptide repeat domain 31. Its functional role remains to be elucidated.

5.5 Discussion

Interaction analysis – RA and non-RA

The results shown here demonstrate the presence of disease specific *cis* eQTLs in $CD19^{+}$ B cells from RA patients compared to the non-RA cohort. Interestingly, there is no overlap between the genes identified in this analysis and the RA risk SNPs described by Okada *et al[107]*. The results of this interaction analysis may, therefore, offer new insights into disease pathogenesis and identify potential explanations for the 'missed heritability' of RA.

Interpretation of analyses such as this, which are not hypothesis driven, is complicated by the identification of genes without clear associations with the disease under examination,

genes that encode ncRNAs which may regulate more than one gene and genes whose function remains to elucidated, such as *LRRC58, TMEM62, TTC31*.

It is possible to prioritise the findings based on their perceived relevance to RA and so focus on immune function genes such as *CLEC4* and *DYRK1A.* Alternatively, direct visualisation of the eQTL plots allows a focus on results where the distribution of the samples between genotypes is balanced, such as *CYP4V2*. However, both methods potentially ignore results which may provide insights into pathogenesis or identify genetic factors relevant to disease progression or severity.

DYRK1A is not an eQTL in the non-RA population but, in RA patients, the presence of the minor allele is associated with decreased expression of the gene. DYRK1A has been shown to phosphorylate many proteins with diverse functions, including cyclin D3 which itself plays a critical role in the regulation of B cell development and proliferation[237]. Work in the *Dyrk1A*$^{-/-}$ knockout mouse has shown that loss of *DYRK1A* during haematopoesis stabilises Cyclin D3, reducing its degradation and so impairs the appropriate exit of pre-B cells from the cell cycle which is required for normal differentiation and development[237]. The functional consequences to the peripheral B cell compartment of a reduction in *DYRK1A* are yet to be investigated, but the identification of this disease specific eQTL raises the possibility that B cell development may be altered in a subgroup of RA patients.

CLEC4A, C-type lectin domain family 4 member A, also known as the dendritic cell immunoreceptor (*DCIR*), is a member of the C-type lectin receptors which have diverse functions including acting as pattern recognition receptors. *CLEC4A* has been associated with ACPA-negative RA in Asian populations and is expressed on antigen presenting cells, including B cells[241, 242]. The extracellular domain contains the carbohydrate recognition domain with a cytoplasmic immunoreceptor tyrosine-based inhibitory motif, which transduces negative signals into cells; leading to the suggestion that it has a potential immunoregulatory role [243].

A small study in humans identified *DCIR* expression on cells from synovial fluid and synovial tissue from rheumatoid joints and this was downregulated following corticosteroid treatment. *DCIR* expression was not identified in healthy synovium[244].

In mouse models, *Dcir*$^{-/-}$ mice display a more severe collagen induced arthritis phenotype and develop serum autoantibodies over time, alongside an excessive expansion of the dendritic cell population; suggesting *DCIR* is a negative regulator of dendritic cell (DC) expansion and may be important in the development of autoimmunity[245]. Engagement of the *DCIR* has been shown to inhibit BCR-mediated Ca^{2+} mobilisation, therefore potentially influencing B cell activation[246]. The precise role of *DCIR* in B cells is not established but differences in its expression in RA patients may influence antigen presentation or the B cell response to antigen binding through modulation of the strength of signalling through the BCR.

The plot in figure 5.3 for *CLEC4A* shows the gene expression is not related to genotype in the non-RA group but expression is reduced in the RA group in the presence of the minor allele, of note there are no minor allele homozygotes in the RA group and the result should be confirmed in a larger cohort and at a protein level.

The eQTL effect seen at *CYP4V2* is striking for the convincing opposing effects seen between the RA and non-RA groups. The exact function of this *CYP4V2* is not known although it is involved in fatty acid metabolism. Alterations in the expression of *CYP4V2*, if reflected at the protein level, may relate to activation state of the cell or altered effective B cell function if cellular metabolism is altered.

In this analysis, the genes related to the regulation of gene expression, cell signalling and metabolism predominate over immunological function. *ARID2, INO080D* and *NS4APT2* are important in chromatin remodelling thereby influencing gene expression and also DNA repair. Gene expression and the cellular processing of proteins is potentially influenced by *ANKRD13C, DYRK1A, PIGB, POLR3H* and *RPLP0*.

ATP5S, CYP4V2, PISD and, *PITRM1* are related to mitochondrial function and cell metabolism, which may link to the altered B cell activation state identified in the gene set analysis for the RA group in *Chapter 3*. The other potential functional category is cell signalling influenced by *CERK, CSNK1G3* and *P2RX4*.

The results in table 5.2 display the SNPs with the most siginifcant p-value for each gene. In most cases, several SNPs were shown to display an interaction effect for each of the 21

genes. A proportion of these SNPs are likely to be in LD with each other. In this setting clumping might optimise power and so may lead to the identification of additional interaction effects. Clumping identifies the most significant SNP in each LD block and this is used in subsequent analyses, reducing the number of tests to be carried out and ensuring there remains one representative SNP for each region of the genome. The analysis pipeline presented here has been shown to be effective but additions such as clumping may lead to the discovery of further disease specific eQTLs.

The demographics of the two comparator groups used for this analysis differed significantly in terms of age and inflammation, as measured by ESR and CRP. The gene expression analysis previously described, in *Chapter 4*, provides evidence that both inflammation and age influence the B cell transcriptome. The differences identified between the RA and non-RA groups may, therefore, reflect differences attributable to age or inflammation rather than to disease *per se*. The next step with these data would be to add these variables to the model and so examine the influence of clinical covariates on the results shown.

In summary, the principle behind this interaction analysis was to identify disease specific eQTLs and the methodology selected was chosen for discovery purposes, hence the decision to use all the RA samples and the non-RA group as control, to optimise the sample sizes used. The results shown here highlight the potential to gain valuable insights into disease pathogenesis using this methodology.

The individual results shown here should be interpreted with caution give the sample sizes and the potential confounders of age and levels of inflammation. The identification of relatively few disease specific *cis* eQTLs is not surprising given that Peters *et al* identified 3 such eQTLs in $CD4^+$ T cells, 3 in neutrophils, none in $CD8^+$ T cells and 7 in monocytes, with the FDR relaxed to 15%. The lack of immunological bias in the results is also in keeping with their work where the genes identified did not solely have immunological functions, concluding that analyses may identify novel, potentially clinically relevant eQTLs only present in disease. Their paper used a healthy volunteer cohort and so, in this context, levels of inflammation may not have been corrected for.

There are many potential avenues to be explored prompted by these findings if replicated. The disease specific eQTLs related to chromatin state highlight the potential importance of the regulation of gene expression in the development of disease. eQTLs related to cell metabolism may indicate that B cell activity is altered in RA as suggested in my expression analysis.

In the first instance I would like to establish the influence of age and inflammation on the analyses. Initially, I would repeat the analyses adding age, ESR and CRP to the model. I would then like to recruit a healthy volunteer cohort to repeat the analysis prior to carrying out downstream work.

5.6 Future work

- Repeat the analyses incorporating age, ESR and CRP into the model
- Repeat the analysis using a larger RA cohort and a healthy volunteer age-matched control group

6. B cell subsets in an early arthritis cohort

6.1 Background

Lymphopaenia is observed in many autoimmune conditions including SLE where it is included in the diagnostic criteria. An analysis of a small cohort of British established RA patients showed that 15% of patients had a persistent lymphopaenia, a lymphocyte count of $\leq 1.00 \times 10^9$/L on at least 2 occasions, over the course of 1 year[247]. The observed persistent lymphopaenia was secondary to a decrease in circulating T cells, with no significant change in the CD4:CD8 T cell ratio. There was no change observed in the number of circulating B cells.

The French ESPOIR cohort of over 800 DMARD naive early arthritis patients showed that 6.2% of patients were lymphopaenic at baseline and this was associated with inflammation, RF positivity and disease activity[248]. The most common diagnosis in the lymphopaenic group (at 3 years) was RA (54%). The study concluded that lymphopaenia, in early inflammatory arthritis, may be a feature of RA but the proportion of lymphopaenic patients was rare in the early RA group overall. The lymphocyte subsets were not examined.

There is little evidence to suggest that the overall proportion of B cells within the lymphocyte population differs between healthy controls (HC) and the early RA population[247, 249, 250]. There is no clear consensus on the frequency of different B cell subsets in early RA, reflecting the paucity of studies in this area and the small samples sizes used. In early RA patients have been shown to have a higher percentage of circulating total $CD19^+$ B cells, a lower proportion of plasmablasts and an expanded proportion of antigen inexperienced naive B cells compared to those with established disease. This may reflect the influence of disease duration or the effects of medications[250]. In addition, patients with early, untreated disease have a higher percentage of circulating naive B cells and a lower percentage of total memory B cells compared to healthy controls [249, 251]. The increase in the proportion of circulating naive cells described may be secondary to the migration of the memory subset to the synovium where the percentage of memory B cells has been shown to be higher than in the peripheral blood in established disease [252].

Analysis of the B cell subsets in early, untreated disease may provide information related to the pathogenesis and initiation of disease. Insights into the potential importance of

individual B cell subsets comes from the rituximab response literature, where preplasma cells, and the plasmablast marker, IgJ, have been associated with response to B cell depletion[96, 101]. The proportion of naive and memory B cells was also higher in a rituximab non-responder group, although this did not reach statistical significance[96].

Hypotheses regarding disease pathogenesis include the expansion of an autoreactive clone of B cells and loss of regulation of the immune system. The subset from which any autoreactive cells arise remains to be established, but the naive cell subset, which is expanded in early RA, have been postulated as a potential source[250, 253].

The regulatory capacity of B cells has gained increasing interest in recent years. In mouse models of autoimmune diseases, including collagen induced arthritis, B cells which produce high levels of IL-10 can suppress the inflammatory response[52]. IL-10 producing B cells, termed regulatory B cells (Bregs), can be found in humans and they play a crucial role in autoimmune diseases and transplant tolerance[56, 254]. As there is no consensus on the exact phenotype of B cells with a regulatory capacity in mice or humans, regulatory B cells have been identified by the expression and release of IL-10 but it is unlikely that suppression is entirely mediated by IL-10. Transforming growth factor β (TGFβ), IL-35, and direct cell to cell contact, with co-stimulatory molecules CD80/CD86, have been shown to also contribute to Breg dependent immunoregulation[54].

In mice, different phenotypes have been shown to have regulatory capability with the consensus that Bregs express high levels of CD1d, CD24 and CD21 but variable levels of CD5, CD10 and CD23[255]. They do not fall into a single homogenous population.

In humans, regulatory B cells again do not fall into a single, clearly defined population and have been shown to be enriched in memory, plasmablast and transitional B cell subsets[59, 61, 256]. As a consequence of the lack of a definitive phenotype, the majority of current human work relies on B cell isolation, *in vitro* stimulation to identify IL-10 producing cells and then phenotyping of these cells. An alternative approach is to focus on a group's favoured phenotype, for example $CD19^{+}CD24^{hi}CD38^{hi}$ or $CD19^{+}CD24^{hi}CD27^{hi}$ B cells[59, 61].

In autoimmune disease, Mauri's group has shown that $CD19^{+}CD24^{hi}CD38^{hi}$ B cells are reduced in number in established RA (estRA) patients with active disease when compared to those with inactive disease and healthy individuals[61]. In addition, they have shown that $CD19^{+}CD24^{hi}CD38^{hi}$ B cells from healthy individuals are able to convert $CD4^{+}$ T cells into suppressive regulatory T cells (Tregs) and limit Th17 development, unlike $CD19^{+}CD24^{hi}CD38^{hi}$ B cells from RA patients. This suggests that, in RA, immature B cells are unable to prevent the differentiation of Th17 cells under Th17 polarising conditions, or convert naive T cells to regulatory T cells and so are less able to restrict the development of autoreactive inflammation. However, in RA, $CD19^{+}CD24^{hi}CD38^{hi}$ B cells maintained their ability to inhibit Th1 cell differentiation. This presents an exciting insight into potential mechanisms of disease in RA.

Mauri's group have examined the $CD19^{+}CD24^{hi}CD38^{hi}$ B cell subset in other autoimmune conditions and demonstrated a functional impairment in SLE. Cells from SLE patients produced less IL-10 in response to CD40 stimulation than those from HCs, and were unable to suppress pro-inflammatory cytokine production from $CD4^{+}$ T cells[54].

There is currently little work regarding Bregs in early RA. One study looked at postulated Bregs in early RA, focussing on $CD5^{+}CD1d^{+}$ B cells, a phenotype more established in mouse models of disease. This subset was reduced in RA patients compared to HCs and negatively correlated with DAS28 score[62]. It is interesting to note that Mauri's group have shown that the majority (71%) of $CD19^{+}CD5^{+}CD1d^{hi}$ B cells are found within the $CD19^{+}CD24^{hi}CD38^{hi}$ B cell subset .[54] This suggests that there may be concordance with the Breg phenotype found in mice.

The importance of B cells in the maintenance of tolerance is highlighted in the transplant literature where the adoptive transfer of B cells from tolerant animals, in a rat transplantation model, resulted in the transfer of allograft tolerance[63]. These B cells could be described as displaying a regulatory phenotype. Newell *et al* identified an increased frequency of transitional $CD19^{+}CD24^{+}CD38^{+}$ B cells in renal transplant patients who did not require continuous drug immunosuppressive therapy; providing further evidence that this subset of immature cells may be important for the maintenance of tolerance[51].

Simon *et al* recently highlighted the complexity of the regulatory function of B cells[47]. In $CD19^{+}CD24^{hi}CD38^{hi}$ B cells, multicolour flow cytometry, bioinformatics and functional studies were used to identify multiple subsets within this population; each subset demonstrating differing regulatory properties. A 10-colour flow cytometry analysis using Flow clustering without K (FLOCK) software identified 8 clusters within the $CD19^{+}CD24^{hi}CD38^{hi}$ B cell population. Three markers: CD27, IgM and IgD were the most useful in differentiating subsets within the clusters. These markers were used to congregate the 8 clusters into 4 groups, subsequently named T1 (immature), T2 (intermediate) , T3 (resting state) and $CD27^{+}$ transitional (activated memory). The transitional $CD27^{+}$ population could be identified as $CD27^{+}$ within the $CD19^{+}CD24^{hi}CD38^{hi}$ group and T1, T2 and T3 groups were identified within the $CD19^{+}CD24^{hi}CD38^{hi}CD27^{-}$ subset, based on the relative expression of IgD and IgM, and this strategy was used for functional work on the subsets.

The transitional $CD27^{+}$ subset was able to suppress proinflammatory cytokines and produce high levels of IL-10. Interestingly, the percentage of $CD27^{+}$ cells within the $CD19^{+}CD24^{hi}CD38^{hi}$ B cell population was significantly increased in Sjogren's syndrome and SLE compared to HC. This may be secondary to disease activity or drug treatments, which may influence the composition of the B cell compartment. Alternatively, the observed increase in the transitional $CD27^{+}$ subset may represent a compensatory response to control the inflammatory process. The patients with Sjogren's syndrome and SLE had a significantly higher overall percentage of $CD19^{+}CD24^{hi}CD38^{hi}$ B cells (as a % of $CD19^{+}$ cells) compared to HCs.

The T3 subset was consistently able to suppress T cell proliferation, in contrast to the T1 and $CD27^{+}$ subsets. This work demonstrates that the regulatory properties of B cells may not lie solely within one subset. The group did not look at RA transitional B cell subsets in this manner[47].

IL-10 is produced by B cells on stimulation, but production is not limited to a single, distinct subpopulation. IL-10 producing cells are enriched in both the $CD19^{+}CD27^{+}$ and $CD19^{+}CD38^{+}$ B cell compartments [58, 59]. In a $CD19^{+}CD27^{+}$ B cell population, the frequency of IL-10 producing cells was 2-4 times that in the $CD19^{+}CD27^{-}$ B cell population, suggesting a correlation with the subsets identified by Simon *et al*[47, 58].

More specifically, IL-10 producing cells have been shown to be enriched in a $CD19^{+}CD24^{hi}CD27^{+}$ B cell population[59]

These data highlight the potential importance of B cell subsets in RA disease pathogenesis and the need to further examine subsets in a DMARD-naive, early RA cohort. As highlighted above, there is now substantial evidence linking the $CD19^{+}CD24^{hi}CD38^{hi}$ B cell subset to regulation of the immune system, and autoimmunity, but data are lacking on this population in DMARD naive RA patients. This transitional B cell subset, newly emerged from the bone marrow and so at a key stage in B cell differentiation, may hold further clues to RA pathogenesis and play a crucial role in the development of autoreactivity.

6.2 Hypothesis and Aims

6.2.1 Hypothesis

The peripheral B cell compartment has been shown to be altered in established and early RA but there is a lack of data comparing DMARD naive RA patients to control groups. Given the assumed role of autoantibody producing cells in RA pathogenesis I would predict the plasmablast subset to be elevated in early disease, and memory population reduced, as this latter subset migrates to the synovium and other sites of disease activity; the naive population will consequentially expand in proportion. I also predict that the regulatory B cell populations will be expanded in the RA group, in an attempt to control the inflammatory process.

6.2.2 Aims

In this chapter I will look for differences in subsets within the $CD19^{+}$ B cell compartment in whole blood, to answer the following questions:

1. Do the proportions of naive and memory B cells differ between patients with RA and those with other inflammatory or non-inflammatory conditions?
2. Does the proportion of plasmablasts differ between patients with RA and those with other inflammatory or non-inflammatory conditions?
3. Can I identify differences in the main three postulated regulatory B cell subsets in early RA?

6.3 Results

6.3.1 Patient cohort

Patients were recruited from the Newcastle Early Arthritis Clinic for the microarray study (*Chapter 3*) and an aliquot of whole blood was used for phenotyping $CD19^+$ B cell subsets by flow cytometry. The patients recruited were DMARD and glucocorticoid naive. A small group of HC and patients with estRA were recruited for further work on the purported regulatory B cell subset, $CD19^+CD24^{hi}CD38^{hi}$ cells. The patients with estRA had seropositive RA and were all due to receive their first dose of rituximab but had differing treatment regimes, including glucocorticoids, in the weeks prior to recruitment.

The initial comparisons were between the early RA group and non-RA groups (table 6.1). The RA group were older, with higher levels of inflammation as measured by ESR and CRP. The non-RA group included patients diagnosed with both inflammatory and non-inflammatory conditions. Following the initial analysis, the non-RA group was split into 'other inflammatory' and 'non-inflammatory' conditions and the three groups differed significantly in terms of gender, age, CRP and ESR (table 6.2 and figure 6.1*)*. The comparisons in figure 6.1 are all based on comparisons with the RA group. The RA group are older, with more evidence of inflammation as measured by ESR.

	RA	**Non-RA**	**P-value**
Number of samples	56	124	-
Gender (% female)	78.6	67.2	ns
Age (yrs)	61 [21-73]	51 [18-92]	<0.0001
ESR (mm/hr)	26 [1-91]	12 [1-113]	0.0008
CRP (mg/L)	11 [4-91]	5 [0-129]	0.0007

Table 6.1 Demographics and clinical characteristics for RA and non-RA samples

Median values and range shown. P-values were calculated by the Mann-Whitney or unpaired T tests, except in the case of gender where the individual Fisher's exact test was used.

	RA	Non-RA		P-value
		Non inflammatory	Other inflammatory	
Number of samples	56	68	56	-
Gender (% female)	78.6	78.3	53.6	0.0032
Age (yrs)	61 [21-73]	51 [22-87]	51.5 [18-92]	0.0001
ESR (mmHg)	26 [1-91]	11 [1-100]	13 [1-113]	0.002
CRP (mg/L)	11 [4-91]	5 [0-49]	6.5 [0-189]	0.0009

Table 6.2 Demographics and clinical characteristics for RA, other inflammatory and non-inflammatory samples

Median values and range shown. One way ANOVA or Kruskal Wallis tests were used to compare groups except in the case of gender where Fisher's exact test was used.

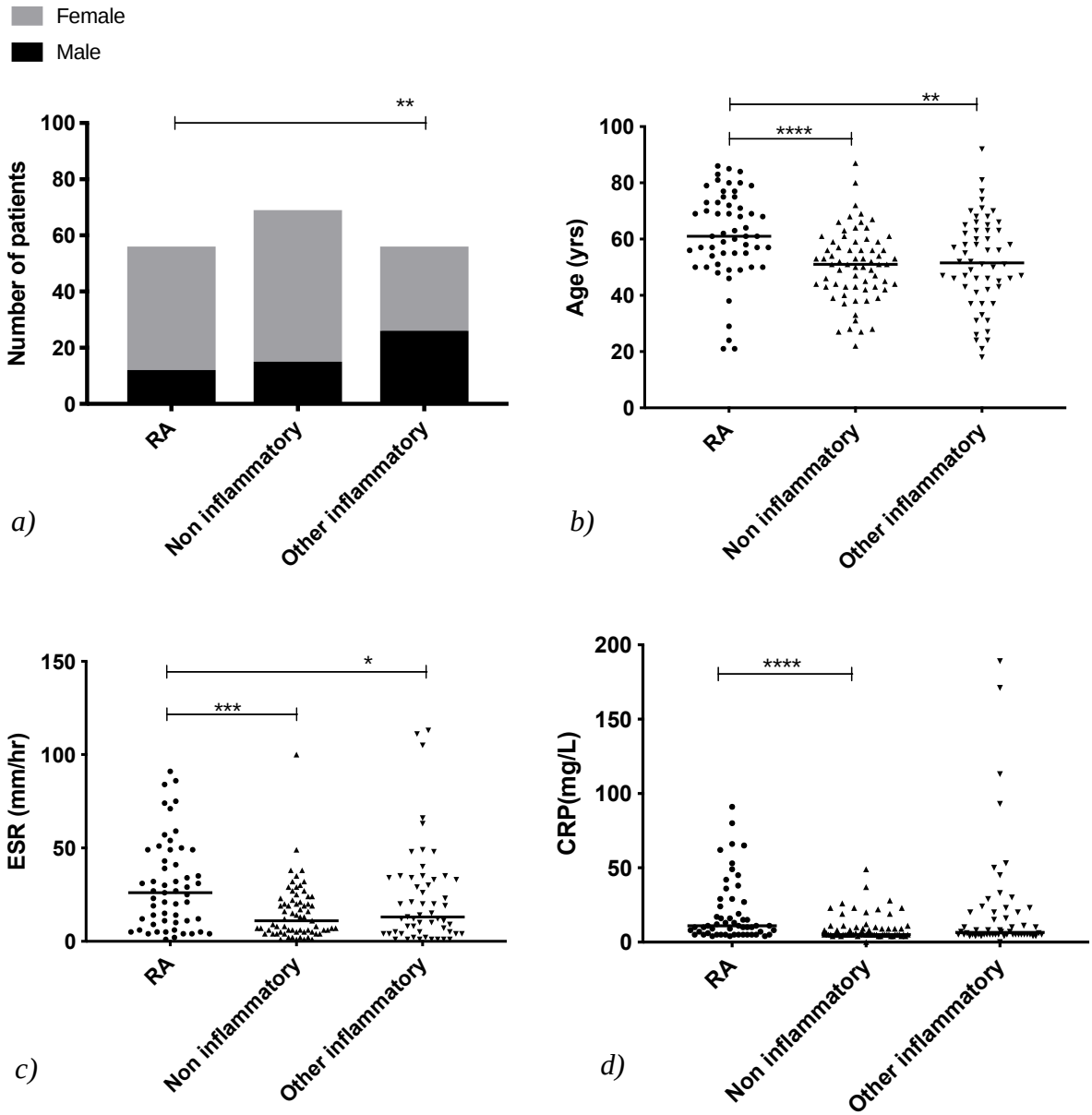


Figure 6.1 Demographics and clinical characteristics: The Newcastle Early Arthritis cohort

*The clinical characteristics for the rheumatoid arthritis (RA) group (n=56), non-inflammatory arthritis group (n=68) and other inflammatory group (n=56) were compared. The groups were compared by a) Gender, b) Age, c) ESR and d) CRP. Horizontal lines depict the median values in plots b)-d). Individual T tests/Mann Whitney tests were carried out to compare the RA group separately to the non-inflammatory group and then to the other inflammatory group. Fisher's exact test was used for gender. *P≤0.05, **P≤0.01, ***P≤0.001, ****P≤0.0001.*

The group termed non-inflammatory includes patients with a final diagnosis of osteoarthritis (OA) and also other non-inflammatory conditions, as attributed by their consultant; this includes a mixture of diagnoses including fibromyalgia. When the non-inflammatory group is divided into OA (n=28) and other non-inflammatory conditions

(n=40), the OA group is significantly older; the CRP is also lower in the OA group although this difference does not reach statistical significance (table 6.3). The OA group is similar to the RA group in terms of gender distribution and age, but not levels of inflammation (table 6.4).

	Osteoarthritis patients	**Other non-inflammatory patients**	**P-value**
Number	28	40	-
Gender (%F)	78.6	80	ns
Age (yrs)	55 [37-87]	45.5 [22-69]	0.0005
ESR (mm/hr)	7 [1-38]	15.5 [1-100]	0.269
CRP (mg/L)	5 [4-26]	6.5 [5-49]	0.055

Table 6.3 Demographics and clinical characteristics for OA and other non-inflammatory samples

Median values and range are shown. P-values were calculated by the Mann-Whitney or unpaired T test, except in the case of gender where the individual Fisher's exact test was used.

	RA	**Osteoarthritis**	**P-value**
Number of samples	56	28	-
Gender (%F)	78.6	75.9	ns
Age (yrs)	61 [21-73]	55 [37-87]	0.1682
ESR (mm/hr)	26 [1-91]	7 [1-38]	0.0006
CRP (mg/L)	11 [4-91]	5 [4-26]	<0.0001

Table 6.4 Demographics and clinical characteristics for RA and OA samples

Median values and range are shown. P-values were calculated by the Mann-Whitney or unpaired T test, except in the case of gender where the individual Fisher's exact test was used.

6.3.2 Gating strategy

As described in *Methods 2.7.2* , the $CD19^+$ B cell subset was identified by first using a singlet gate (SSC-A against SSC-W) to remove doublets and this was carried forward to identify the leucocyte and lymphocyte populations. The percentage of $CD19^+$ B cells was then determined as a proportion of total leucocytes. The proportions of postulated regulatory B cell subsets ($CD19^+CD24^{hi}CD38^{hi}$, $CD19^+CD24^{hi}CD27^+$) and plasmablasts ($CD19^+CD27^{hi}CD38^{hi}$) were then measured as percentages of the $CD19^+$ B cell gate. The lymphocyte gate was used for the identification of $CD19^+CD27^+$ memory B cells and $CD19^+CD27^-$ naive B cells.

6.3.3 Total B cells

The percentage of $CD19^+$ B cells was significantly lower in the early RA group than the main comparator group, non-RA (figure 6.2a). When the non-RA group was split there was a significant difference in the percentage of $CD19^+$ B cells between the early RA group and those with a non-inflammatory arthritis, but not those with other inflammatory arthritides (figure 6.2b). Comparing the other inflammatory arthritis group to the non-inflammatory group the observed difference did not reach statistical significance ($p = 0.064$). In summary, a significant difference was identified in the percentage of $CD19^+$ B cells between the early RA group and all the other conditions combined and with the non-inflammatory group but this was not true when comparing the early RA group to patients with other inflammatory conditions.

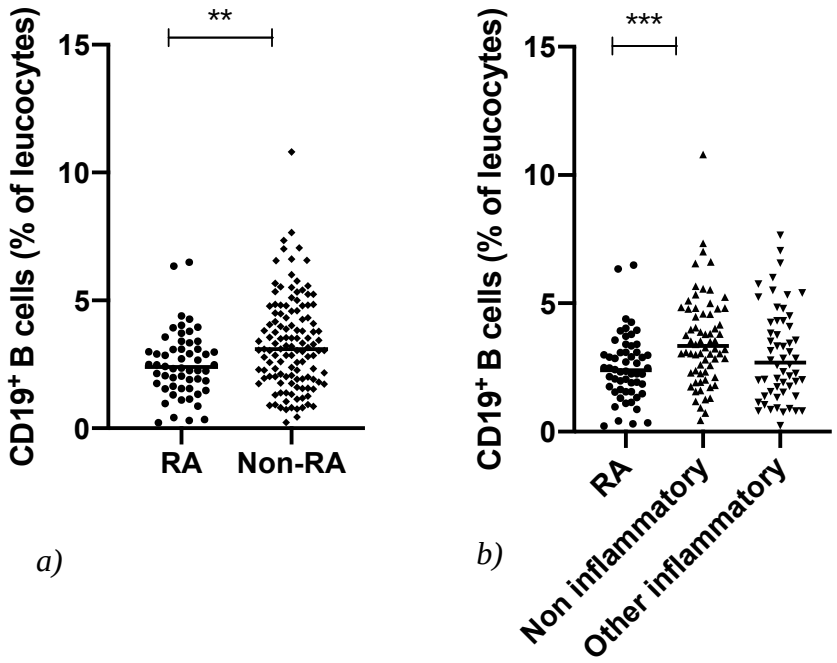


Figure 6.2 The frequency of CD19$^+$ B cells in different disease groups

*Whole blood from early arthritis patients was stained for flow cytometry with a panel of antibodies to detect CD19$^+$ B cells. CD19$^+$ B cells were identified from leucocytes following exclusion of doublets using SSC-A v SSC-W. a) Data from RA (n=56) and non-RA (n=124) individual donors were plotted. b) The non-RA group is split into two groups: non-inflammatory (n=68) and other inflammatory (n=56) and compared to the RA group. The horizontal bars represent the median value, significance was determined by Mann-Whitney tests to compare the RA group to each group separately. ** p<0.01, ***p<0.001.*

Given the differences in age and acute phase response (as measured by CRP and ESR) between the early RA and non-RA group a logistic regression analysis was performed. This demonstrated that, in addition to age, the percentage of CD19$^+$ B cells is independently associated with a diagnosis of RA in patients presenting to the early arthritis clinic. In contrast, markers of the acute phase response (ESR and CRP) are not independently associated with a diagnosis of RA (table 6.5).

	Exp (B)	P-value	95% CI
$CD19^+$ B cells (% of leucocytes)	0.769	0.036	0.602-0.983
Age (yrs)	1.035	0.014	1.007-1.063
ESR (mm/hr)	1.019	0.075	0.988-1.040
CRP (mg/L)	0.989	0.241	0.972-1.007

Table 6.5 Multivariate analysis of clinical variables and the frequency of $CD19^+$ B cells as predictors of clinical outcome

Logistic regression analysis was carried out using SPSS. RA (n=56) and non-RA (n=124) samples. Multivariate analysis indicates that age and frequency of $CD19^+$ B cells are independently associated with a diagnosis of RA.

The relationship between the percentage of $CD19^+$ B cells and clinical characteristics was next examined using all available samples; including samples without a clear diagnosis at the baseline assessment, to improve the sample size. There was a negative correlation between the percentage of $CD19^+$ B cells and age, ESR and CRP (figure 6.3).

In the RA group alone (n=75) the percentage of $CD19^+$ B cells negatively correlated with age (R=-0.2981, P 0.0094) and CRP (R=-0.2800, P 0.0157) but not with ESR or disease activity as measured by DAS28 score (data not shown).

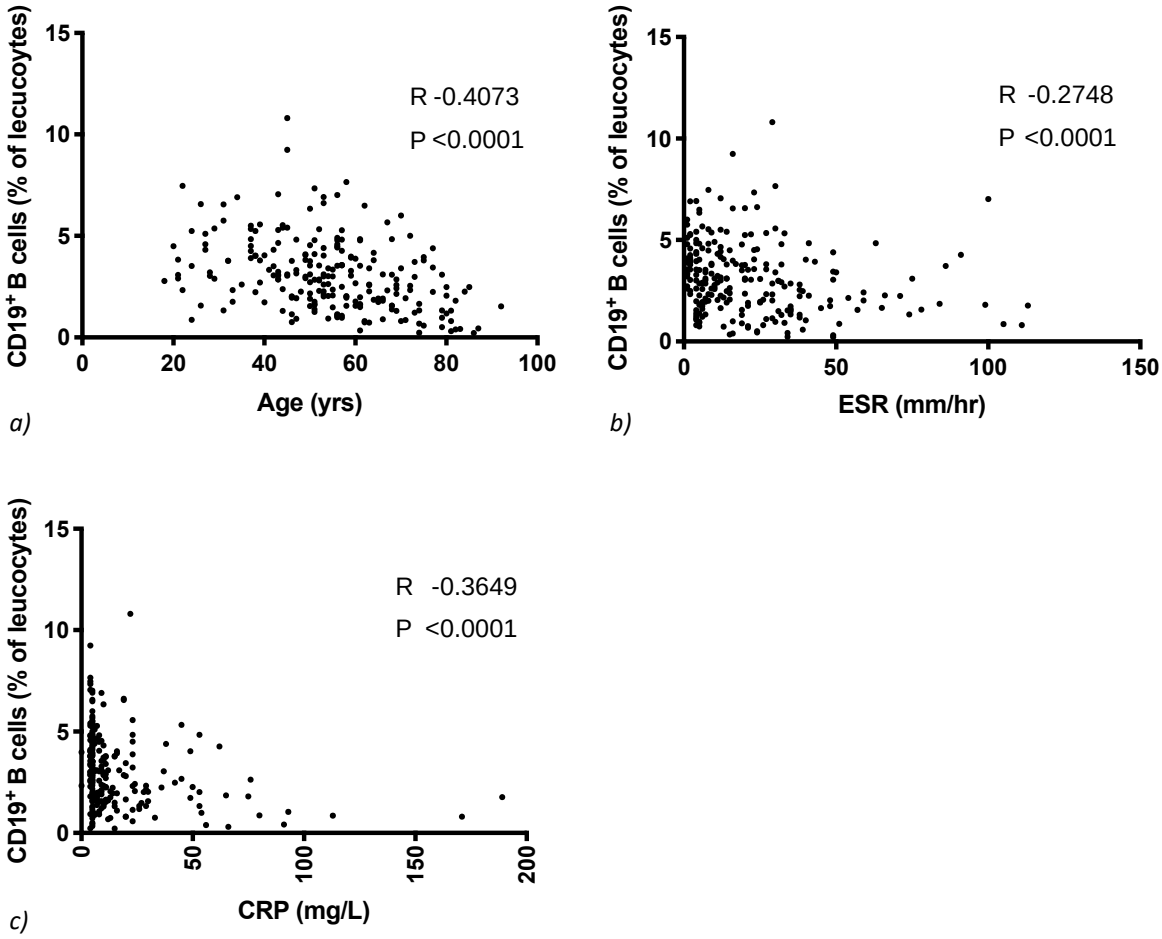


Figure 6.3 Relationship between the percentage of CD19$^+$ B cells and clinical characteristics

Whole blood from early arthritis patients was stained for flow cytometry with a panel of antibodies to detect CD19$^+$ B cells. CD19$^+$ B cells were identified from leucocytes following exclusion of doublets using SSC-A v SSC-W. The relationship between the frequency of CD19$^+$ B cells and a) age (n=233), b) ESR (n=230) and c) CRP (n=232) was tested using Spearman's correlation.

6.3.4 Naive and memory B cells

There were no differences detected in the frequency of naive ($CD19^+CD27^-$ cells) or memory ($CD19^+CD27^+$ cells) B cells between the RA group and the non-RA group. The non-RA group was split into a non-inflammatory and other inflammatory group and these two groups were compared, in turn, to the RA group. No difference in the frequency of naive and memory B cells was detected between the groups (figure 6.4).

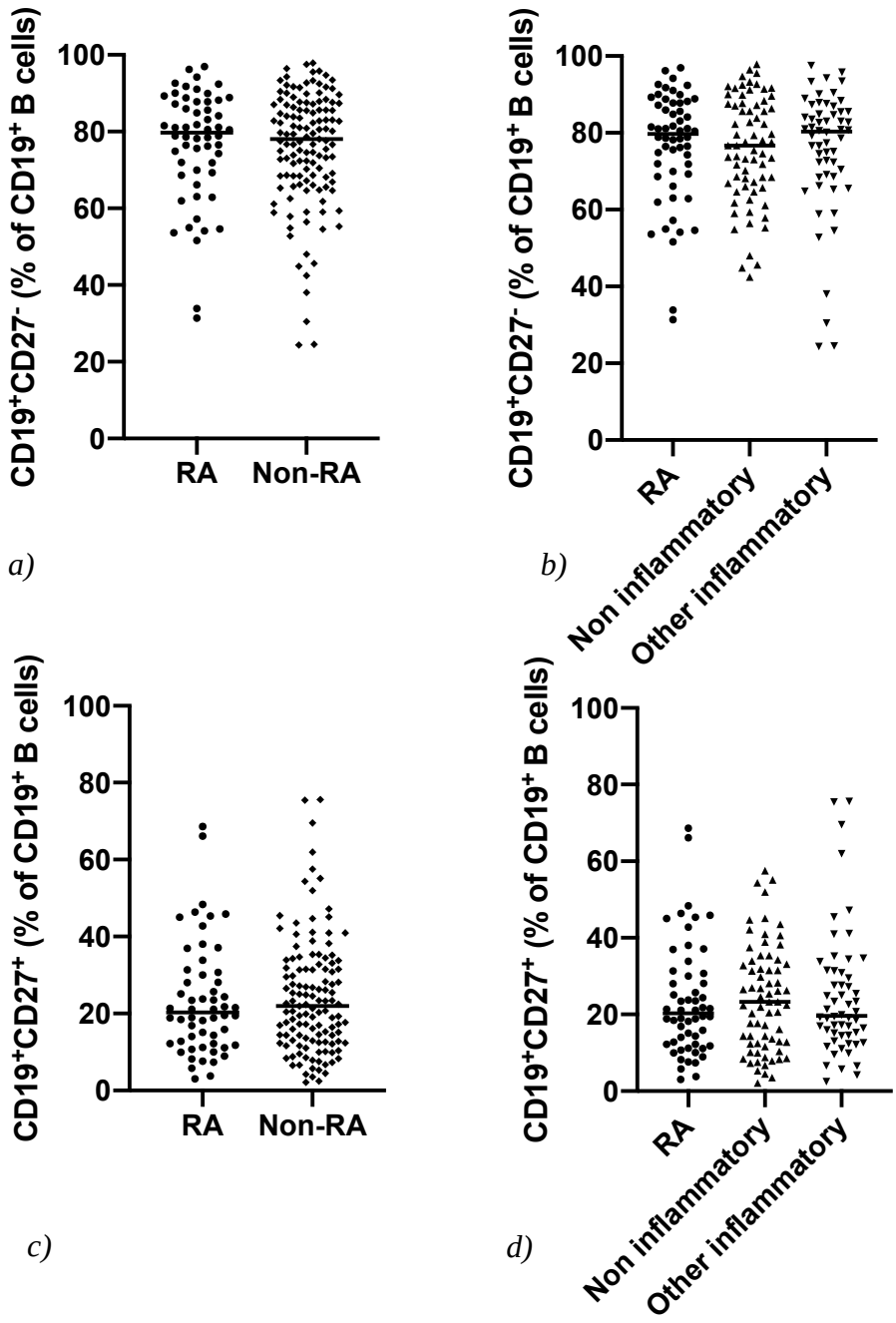


Figure 6.4 The frequency of naive and memory B cells

Whole blood from early arthritis patients was stained for flow cytometry with a panel of antibodies to detect $CD19^+CD27^-$ B cells (naive) and $CD19^+CD27^+$ B cells (memory). Naive and memory B cells were identified from lymphocytes following exclusion of doublets using SSC-A v SSC-W. Plots a) and b) show data for $CD19^+CD27^-$ (naive) B cells from a) RA (n=56) and non-RA (n=124) individual donors, b) the non-RA group is split into two groups: non-inflammatory (n=68) and other inflammatory (n=56) and compared to the RA group. Plots c) and d) show data for $CD19^+CD27^+$ (memory) B cells from c) RA (n=56) and non-RA (n=124) individual donors, d) the non-RA group is split into two groups: non-inflammatory (n=68) and other inflammatory (n=56) and compared to the RA group. The horizontal bars represent the median value. There was no significant difference between the groups using the Mann-Whitney test.

6.3.5 Regulatory B cells

$CD19^{+}CD24^{hi}CD27^{+}$ B cells

One postulated regulatory B cell subset is part of the memory B cell compartment and the frequency of this subset, $CD19^{+}CD24^{hi}CD27^{+}$ B cells, was assessed within the $CD19^{+}$ B cell population. There was no detectable difference between the percentage of this subset in the RA group compared to the other clinical categories (figure 6.5).

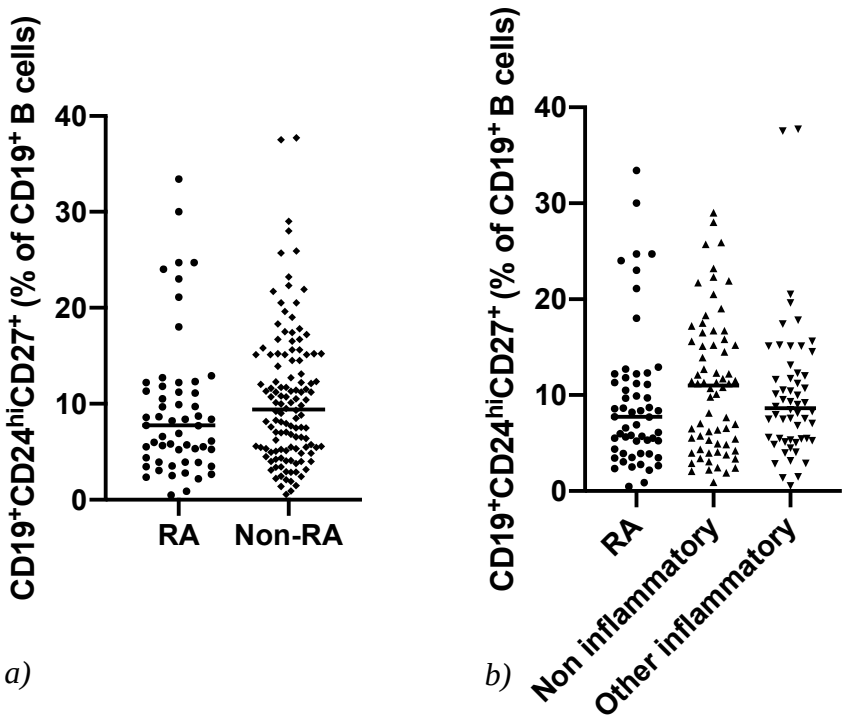


Figure 6.5 The frequency of $CD19^{+}CD24^{hi}CD27^{+}$B cells

Whole blood from early arthritis patients was stained for flow cytometry with a panel of antibodies to detect $CD19^{+}CD24^{hi}CD27^{+}$B cells. $CD19^{+}$ B cells were identified from leucocytes following exclusion of doublets using SSC-A v SSC-W. The $CD19^{+}CD24^{hi}CD27^{+}$ population was identified within the $CD19^{+}$ gate. a) Data from RA (n=56) and non-RA (n=124) individual donors were plotted. b) The non-RA group is split into two groups: non-inflammatory (n=68) and other inflammatory (n=56) and compared to the RA group. The horizontal bars represent the median value. There was no significant difference between the groups using the Mann-Whitney test.

Plasmablasts

Plasmablasts ($CD19^{+}CD27^{hi}CD38^{hi}$ B cells) have also been described as potential regulatory B cells based on their ability to produce IL-10. The frequency of this subset, $CD19^{+}CD27^{hi}CD38^{hi}$ B cells, was assessed in the $CD19^{+}$ B cell population. There was no

detectable difference in the percentage of this subset in the RA group compared to the other clinical categories (figure 6.6).

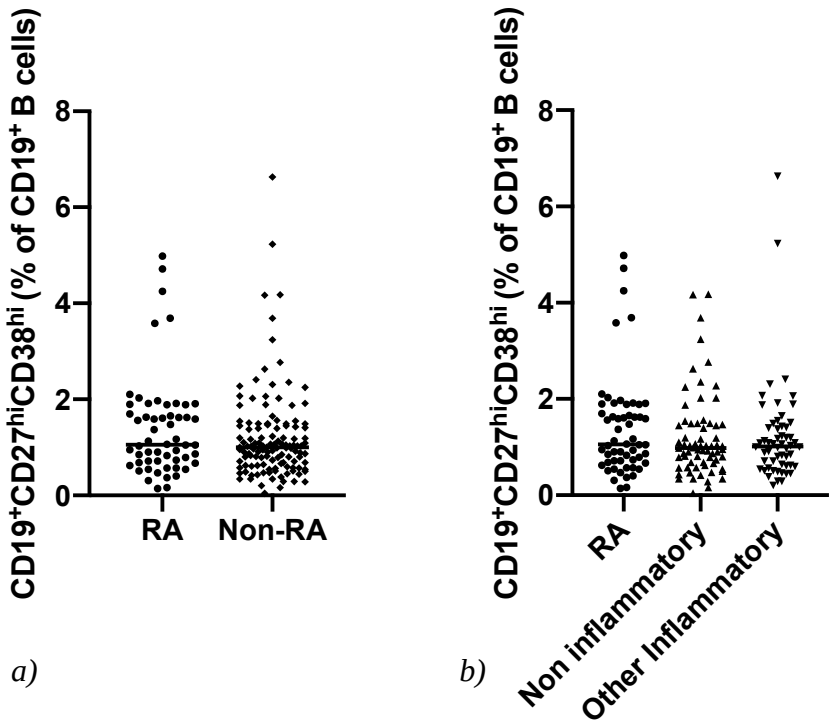
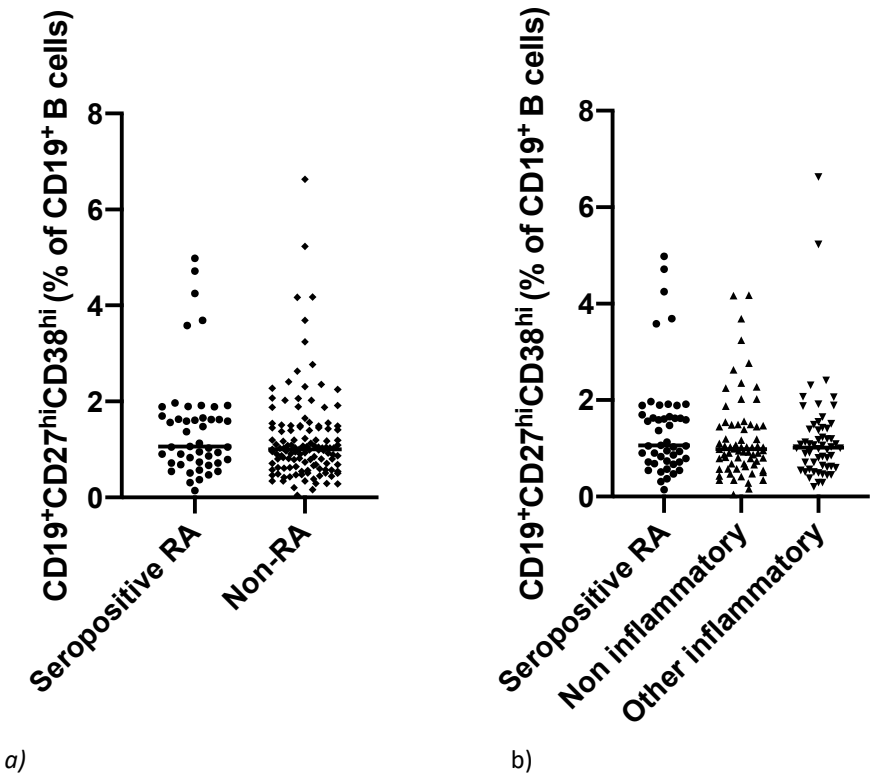


Figure 6.6 The frequency of $CD19^{+}CD27^{hi}CD38^{hi}$ B cells

Whole blood from early arthritis patients was stained for flow cytometry with a panel of antibodies to detect $CD19^{+}CD27^{hi}CD38^{hi}$ B cells. $CD19^{+}$ B cells were identified from leucocytes following exclusion of doublets using SSC-A v SSC-W. The $CD19^{+}CD27^{hi}CD38^{hi}$ population was identified within the $CD19^{+}$ gate. a) Data from RA (n=56) and non-RA (n=124) individual donors were plotted. b) The non-RA group is split into two groups: non-inflammatory (n=68) and other inflammatory (n=56) and compared to the RA group. The horizontal bars represent the median value. There was no significant difference between the groups using the Mann-Whitney test.

Given that plasmablasts are a potential source of autoantibodies and, therefore, may play a role in disease pathogenesis beyond any potential regulatory capacity, the same comparisons were repeated but using only seropositive (RF and/or anti-CCP positive) RA patients (n=47) in the RA group. No difference was detected in percentage of plasmablasts between the seropositive RA and the disease control groups (figure 6.7).



a) b)

Figure 6.7 The frequency of $CD19^{+}CD27^{hi}CD38^{hi}$ B cells in seropositive RA and different disease groups

Whole blood from early arthritis patients was stained for flow cytometry with a panel of antibodies to detect $CD19^{+}CD27^{hi}CD38^{hi}$ B cells. $CD19^{+}$ B cells were identified from leucocytes following exclusion of doublets using SSC-A v SSC-W. The $CD19^{+}CD27^{hi}CD38^{hi}$B cell population was identified within the $CD19^{+}$ gate. a) Data from seropositive RA (n=47) and non-RA (n=124) individual donors were plotted. b) The non-RA group is split into two groups: non-inflammatory (n=68) and other inflammatory (n=56) and compared to the RA group. The horizontal bars represent the median value. There was no significant difference between the groups using the Mann-Whitney test.

Transitional B cells

B cells with a transitional phenotype, $CD19^{+}CD24^{hi}CD38^{hi}$ B cells, are one of the most studied of the postulated Breg cell populations. There was a significant increase in the percentage of $CD19^{+}CD24^{hi}CD38^{hi}$ B cells in the RA group compared to the non-RA group (figure 6.8a). When the non-RA group was split, the observed difference remained between the RA and other inflammatory group but did not reach statistical significance between the RA and non-inflammatory group (figure 6.8b). However, a significant increase in the percentage of $CD19^{+}CD24^{hi}CD38^{hi}$ cells was observed between the RA group and those with OA (figure 6.8c).

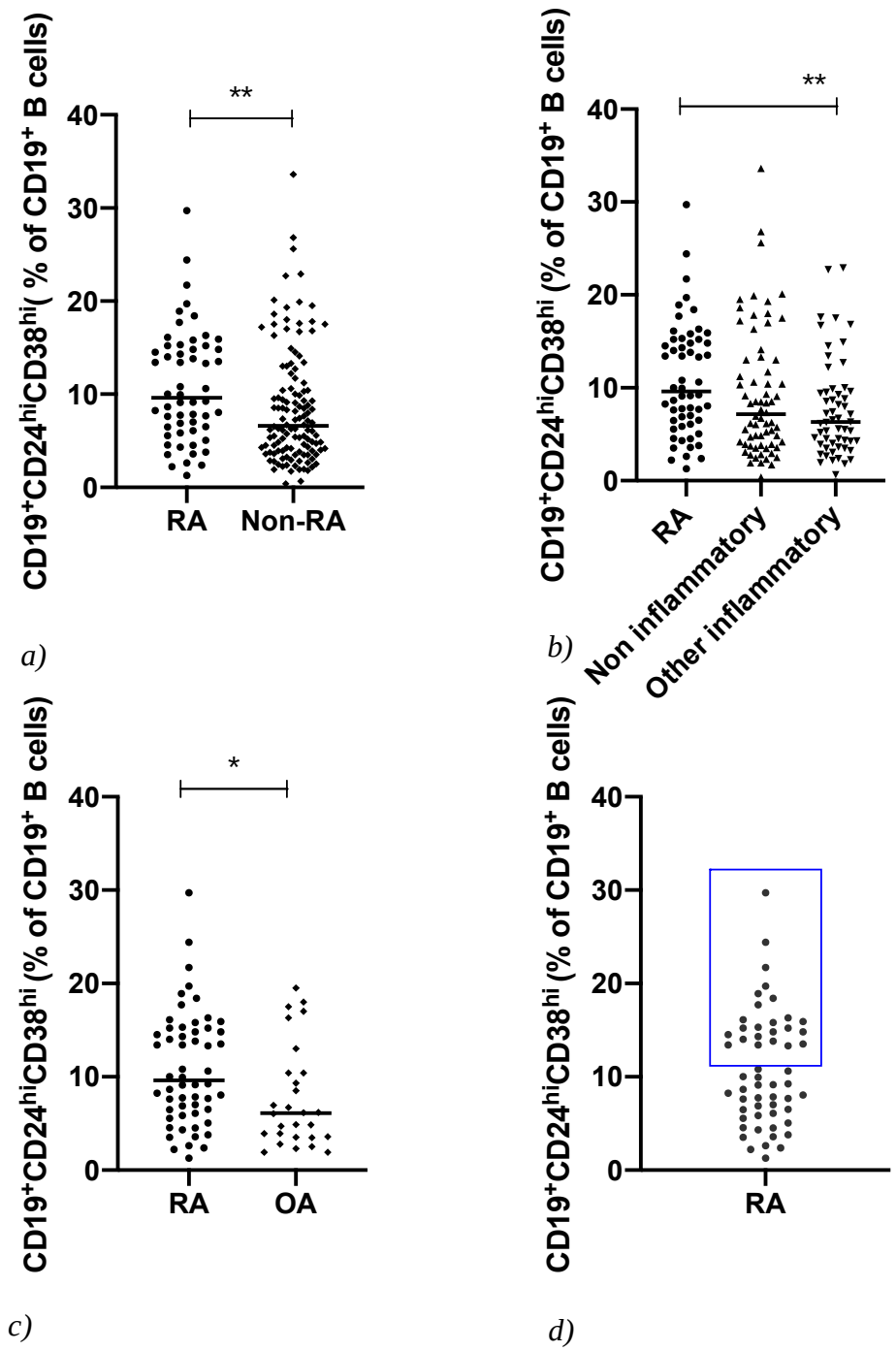


Figure 6.8 The frequency of $CD19^{+}CD24^{hi}CD38^{hi}$ B cells

*Whole blood from early arthritis patients was stained for flow cytometry with a panel of antibodies to detect $CD19^{+}CD24^{hi}CD38^{hi}$ B cells. $CD19^{+}$ B cells were identified from leucocytes following exclusion of doublets using SSC-A v SSC-W. The $CD19^{+}CD24^{hi}CD38^{hi}$ population was identified within the $CD19^{+}$ gate. a) Data from RA (n=56) and non-RA (n=124) individual donors were plotted. b) The non-RA group is split into two groups: non-inflammatory (n=68) and other inflammatory (n=56) and compared to the RA group. c) RA and OA (n=28) donors, d) RA donors alone, blue box to identify the subgroup of RA donors with relatively high frequency of $CD19^{+}CD24^{hi}CD38^{hi}$ B cells. The horizontal bars represent the median value, significance was determined by Mann-Whitney tests to compare the RA group to each other group. *p<0.05,** p<0.01.*

The RA group, shown on the plots in figure 6.8, has the appearance of a bimodal distribution, with groups with relatively high (RA_high) and low (RA_low) percentages of $CD19^{+}CD24^{hi}CD38^{hi}$ cells (figure 6.8d). There was no significant difference in autoantibody status, gender, age, ESR or DAS28 score between the two subgroups (table 6.6). The RA_low subgroup had a significantly higher CRP than the RA_high subgroup but there was a broad spread of values (figure 6.9).

	RA_low	**RA_high**	**P-value**
Number of samples	32	24	-
% CCP positive	65.63	66.7	0.999
% RF positive	78.1	66.7	0.375
% seropositive	87.5	79.2	0.476
Gender (%F)	75	83.3	0.525
Age (years)	59.5 [21-86]	63.5 [21-85]	0.983
ESR (mm/hr)	21 [1-91]	26.5 [4-84]	0.853
CRP (mg/L)	14 [5-91]	9 [5-80]	0.0383
DAS 28	4.31 [1.26-7.15]	4.63 [2.62-8.46]	0.220

Table 6.6 Demographics and clinical characteristics of rheumatoid arthritis

The RA_low subgroup refers to rheumatoid arthritis patients with low levels of $CD19^{+}CD24^{hi}CD38^{hi}$ B cells and the RA_high subgroup refers to rheumatoid arthritis patients with a higher frequency of this subset based on flow cytometry data. Median values and ranges are shown. P-values were calculated by Mann-Whitney or unpaired T tests, except in the case of gender and serological status where Fisher's exact test was used.

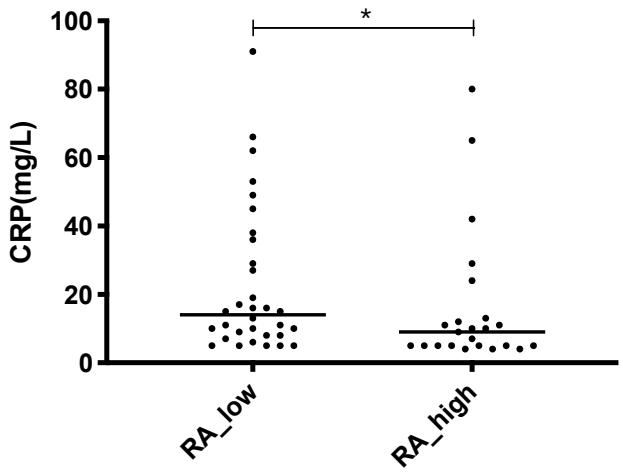


Figure 6.9 CRP levels in rheumatoid arthritis patients

*The RA patients were divided into two groups based on whether they were identified as having a low (RA_low) (n=32) or a high (RA_high) (n=23) frequency of $CD19^{+}CD24^{hi}CD38^{hi}$ cells in figure 6.8. The CRP levels for patients in the RA_low and RA_high groups were compared. The horizontal bars represent the median value, significance was determined by the Mann-Whitney test. *p<0.05*

The RA group has a significantly higher frequency of $CD19^{+}CD24^{hi}CD38^{hi}$B cells than the other inflammatory group, but this was not seen in the comparison with those with non-inflammatory conditions. The RA and other inflammatory conditions group differ in age, ESR level and gender (figure 6.1). However, a multivariate analysis indicates that, in addition to age and gender, the frequency of $CD19^{+}CD24^{hi}CD38^{hi}$ B cells, but not ESR, is independently associated with a diagnosis of RA amongst individuals presenting to an early arthritis clinic with inflammatory arthritis (table 6.7).

	Exp(B)	**P-value**	**95% CI**
Frequency of $CD19^{+}CD24^{hi}CD38^{hi}$ cells	1.092	0.031	1.008-1.184
Gender	2.771	0.029	1.113-6.902
Age	1.043	0.005	1.013-1.074
ESR	0.999	0.907	0.982-1.1017

Table 6.7 Multivariate analysis of clinical variables and frequency of $CD19^{+}CD24^{hi}CD38^{hi}$ B cells as predictors of clinical outcome

Logistic regression analysis was carried out using SPSS demonstrating that frequency of $CD19^{+}CD24^{hi}CD38^{hi}$ B cells, age and gender are independently associated with a diagnosis of RA.

There was no correlation between the frequency of $CD19^{+}CD24^{hi}CD38^{hi}$ B cells and age, ESR or CRP when using all the available samples from patients recruited (n=230-233

depending on analyses, difference in numbers due to missing clinical data) (data not shown).

Looking at the RA population in isolation there was also no correlation between $CD19^{+}CD24^{hi}CD38^{hi}$ B cell frequency and age, ESR, CRP or DAS28 score (n=56). The frequency of $CD19^{+}CD24^{hi}CD38^{hi}$ B cells did correlate with the percentage of $CD19^{+}$ B cells when looking at all samples (figure 6.10), but this did not reach statistical significance in the RA group alone ($p = 0.0586$, $r = 0.2543$).

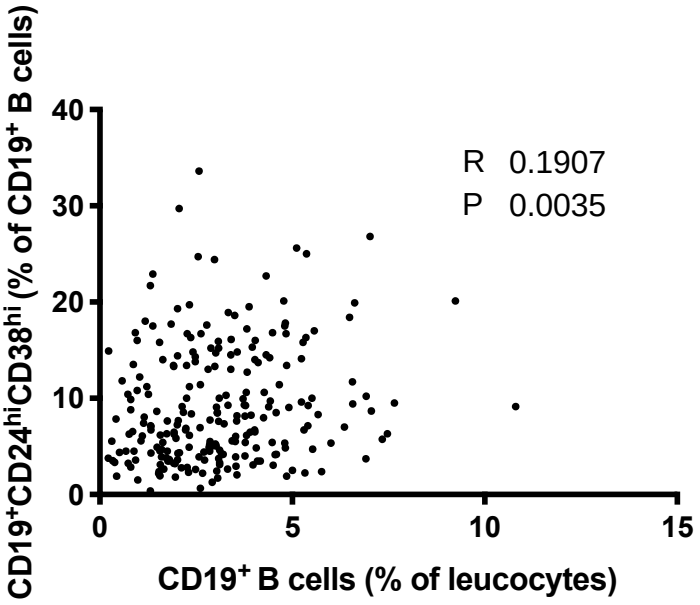


Figure 6.10 Relationship between the frequency of $CD19^{+}$ B cells and $CD19^{+}CD24^{hi}CD38^{hi}$ cells

Whole blood from early arthritis patients was stained for flow cytometry with a panel of antibodies to detect $CD19^{+}$ B cells and $CD19^{+}CD24^{hi}CD38^{hi}$ cells. The relationship between the frequency of $CD19^{+}$ B cells and $CD19^{+}CD24^{hi}CD38^{hi}$ cells (n=233) was tested using Spearman's correlation.

In order to look at this transitional B cell subset in greater detail the $CD27^{+}$ population was identified within the $CD19^{+}CD24^{hi}CD38^{hi}$ subset as described by Simon *et al [47]* but no significant differences were identified between the disease groups (data not shown).

To provide a more in-depth assessment of $CD19^{+}CD24^{hi}CD38^{hi}$ transitional B cells in health and disease, small cohorts of HC (n=6) and patients with estRA (n=7) were recruited. There were statistically fewer $CD19^{+}CD24^{hi}CD38^{hi}$ B cells in established RA

compared with early RA (figure 6.11). There was no difference between HC and early RA or HC and estRA but statistical power was low due to the small numbers in HC and estRA cohorts.

The estRA cohort had a significantly higher DAS28 score than the early RA cohort, with a median DAS28 of 6.18 in estRA and 4.59 in early RA (P 0.037). Patients in the estRA cohort were awaiting their first treatment with rituximab and so had high disease activity, but there was no correlation between DAS28 and the frequency of $CD19^{+}CD24^{hi}CD38^{hi}$ B cells. Complete clinical data were not available but 50% of the estRA cohort had received oral or intramuscular steroids within 3 months of recruitment. There was no significant difference in gender, age, ESR or CRP between the estRA and early RA groups (data not shown).

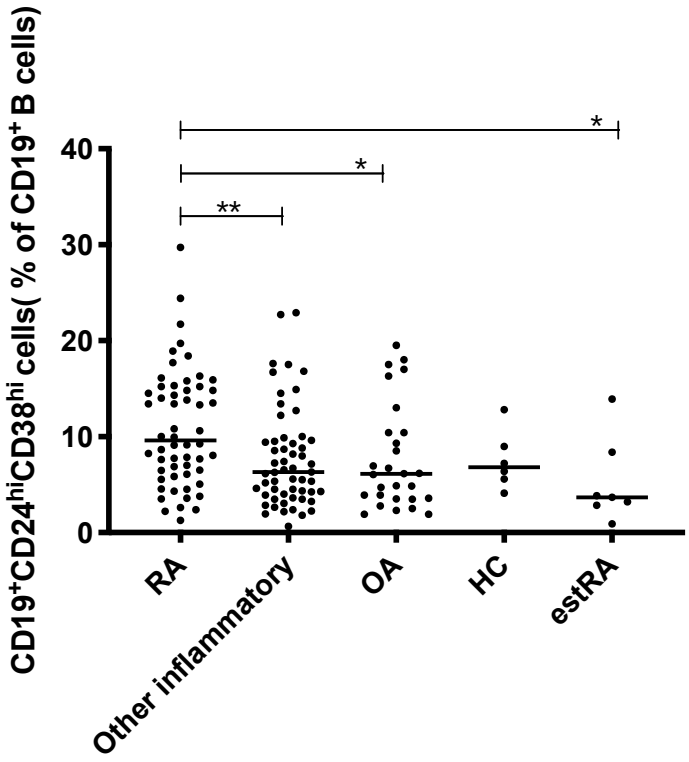


Figure 6.11 The frequency of $CD19^{+}CD24^{hi}CD38^{hi}$ B cells in different disease groups, healthy controls and established RA

*Whole blood from early arthritis patients, healthy controls (HC) and patients with established RA (estRA) was stained for flow cytometry with a panel of antibodies to detect $CD19^{+}CD24^{hi}CD38^{hi}$ B cells. $CD19^{+}$ B cells were identified from leucocytes following exclusion of doublets using SSC-A v SSC-W. The $CD19^{+}CD24^{hi}CD38^{hi}$ population was identified within the $CD19^{+}$ gate. Data from RA (n=56), other inflammatory (n=56), OA (n=28), HC (n=6) donors, estRA (n=7) patients were compared. The horizontal bars represent the median value, significance was determined by the Mann-Whitney test to compare the RA group to each other group. *p<0.05,** p<0.01.*

6.4 Discussion

In contrast to the literature I have shown the that the percentage of $CD19^+$ B cells in peripheral blood is reduced in early RA compared to patients in the disease control (non-RA) group. However, when the non-RA group was split, the difference only remained significant between the RA and non-inflammatory groups. The reduction in $CD19^+$ B cells may be secondary to the effects of systemic inflammation. TNF, a key cytokine linked to RA pathogenesis, promotes B cell mobilisation from the bone marrow and promotes granulopoesis at the potential expense of bone marrow lymphopoesis, alongside a corresponding increase in developing B cells in spleen[257, 258]. This initially leads to an increased release of immature B cells into the periphery, but reduced bone marrow lymphopoesis may ultimately result in decreased circulating B cells over time. Alternatively, systemic inflammation may cause migration to the extramedullary tissues, or local sites of inflammation, leading to an apparent decrease in circulating B cells. Memory B cells accumulate in the synovium of RA patients, expressing relatively higher levels of *CXCR1*, *CXCR2*, *CXCR4* and *CCR2* than those remaining in the peripheral blood[252].

Data from the whole cohort demonstrate a negative correlation between the frequency of $CD19^+$ B cells and levels of inflammation, measured by ESR and CRP, strengthening the possibility that $CD19^+$ B cells have migrated to sites of inflammation or germinal centres. In the RA group alone the frequency of $CD19^+$ B cells negatively correlated with CRP, but not ESR.

The RA group was significantly older and, as expected, had higher levels of inflammation than the non-RA group. However, a logistic regression analysis showed that older age and lower $CD19^+$ B cell percentages were predictors of a diagnosis of RA, in contrast to measures of the acute phase response. Age negatively correlated with the proportion of $CD19^+$ B cells.

The immune system is known to change with age: there is a decline in response to vaccination and an increased susceptibility to cancer and infection[191, 259-261]. The B cell compartment changes in terms of function, with a decline in the ability to generate a sustained humoral response; the relative proportions of the different B cell subsets also

change with age. However, the data are conflicting regarding a decline in the proportion of total B cells in peripheral blood with age[181, 262, 263]. This may reflect a lack of data for people at the extremes of age, particularly those over 80 years old. Interestingly, the proportion of transitional B cells has been shown to decrease with increasing age in healthy volunteers, possibly reflecting an age related reduction in bone marrow activity[181]. This has not been confirmed by my data, perhaps due to the fact that I have studied a disease cohort.

I did not identify a difference in the proportion of circulating memory and naive B cells between disease groups. This contrasts with changes reported in the literature but those reports compared early RA and HCs rather than disease controls. Given that synovial inflammation, and indeed ectopic lymphogenesis, is not specific to the RA synovium, common memory B cell migration patterns may have existed in both my early RA and disease control cohorts, contributing to a lack of difference between these groups.

I have studied, for the first time, the proportion of circulating $CD19^{+}CD24^{hi}CD38^{hi}$ transitional B cells in early RA, cells which are postulated to have a regulatory function. In contrast to the work by the Mauri group, who examines estRA, I found this population to be elevated rather than reduced in early RA. However, the frequency of $CD19^{+}CD24^{hi}CD38^{hi}$ B cells in the small group of patients with estRA receiving DMARD treatment was significantly lower than found in the early RA cohort and, whilst not statistically significant, numerically lower than the level in healthy and disease controls. Therefore, the data from the estRA group here is in agreement with the findings from the Mauri group.

Although the number of HC is low in my cohort, the scatter plots do appear to show that the median frequency of the $CD19^{+}CD24^{hi}CD38^{hi}$ B cell subset is similar between the healthy group and patients with other inflammatory causes of arthritis, in keeping with the literature.

The Mauri group showed that $CD19^{+}CD24^{hi}CD38^{hi}$ B cell frequency was inversely correlated with disease activity. At first sight this appears counterintuitive for a regulatory subset, which might increase in frequency to combat inflammation, especially considering its potentially impaired function in RA[61]. On the other hand, similar to my

arguments for $CD19^+$ B cells overall, it is possible that these cells migrate to sites of inflammation, explaining their disappearance from blood in active disease. Indeed, work from the Mauri group has demonstrated that $CD19^+CD24^{hi}CD38^{hi}$ B cell frequency was significantly higher in the synovial fluid (SF) of patients with active disease than in matched blood. As CCR5, which is associated with Treg migration, was elevated on the SF B cells it was suggested that the reduction in circulating cells was a consequence of their migration. However, the frequency was not measured in SF from inactive disease or other diseases. It would be interesting to investigate the proportion of the $CD19^+CD24^{hi}CD38^{hi}$ B cell subset in the SF of an early disease cohort.

The increased frequency of circulating transitional $CD19^+CD24^{hi}CD38^{hi}$ B cells in early RA may also reflect mobilisation from the bone marrow in response to recent onset inflammation[258]. Given the suggested regulatory potential of this subset, such early mobilisation seems logical from a disease perspective, with subsequent migration to inflamed tissues as suggested above. Alternatively, transitional B cells potentially include more autoreactive cells than other peripheral B cell subsets as they have not been subjected to peripheral tolerance checkpoints; in this regard their higher frequency in early RA may reflect the disease process.

The RA group can be divided into two groups based on the frequency of $CD19^+CD24^{hi}CD38^{hi}$ B cells and there is a significant difference in the levels of CRP between the two groups. The RA patients with a lower frequency of $CD19^+CD24^{hi}CD38^{hi}$ B cells have a higher CRP, which may be secondary to the migration of these cells to the inflamed synovium. Additionally, it may be that there are functional differences in the $CD19^+CD24^{hi}CD38^{hi}$ B cells between the two groups which have not been examined here.

These data provide a snapshot of the peripheral B cell compartment in patients at presentation. Given the regulatory role of this subset, long term follow up data may address questions such as whether patients with higher levels of $CD19^+CD24^{hi}CD38^{hi}$ B cells at presentation have a better prognosis, given the postulated regulatory role of this subset. The duration of symptoms prior to presentation to the early arthritis clinic may also influence the observed frequency of $CD19^+CD24^{hi}CD38^{hi}$ B cells in early RA, but the available clinical data did not allow me to test this possibility. It is also possible that

the observed bimodal transitional B cell distribution in early RA patients is an artefact and may disappear with a larger cohort size.

The apparent increase in frequency of $CD19^{+}CD24^{hi}CD38^{hi}$ B cells in early RA is interesting but requires work, such as assessment of IL-10 production and the influence of this subset on T cell responses, to determine whether it is functionally suppressive in this setting. It would also be valuable to determine if $CD19^{+}CD24^{hi}CD38^{hi}$ B cell function as well as frequency changes with disease duration.

6.5 Future work

- Increase the number of donors in the HC and estRA cohorts to more robustly compare the frequency of B cell subsets with the early arthritis cohort.
- Obtain longitudinal samples from the RA patients in the early arthritis cohort to look for changes in Breg frequency over the time course of the disease.
- Relate Breg frequency at presentation and disease progression as measured by disease activity over time, glucocorticoid use and erosive changes, to assess whether Breg frequency at presentation influences disease prognosis.
- Measure the frequency of $CD19^{+}CD24^{hi}CD38^{hi}$ B cells in synovial fluid samples from early RA patients, and disease controls (such as OA), and the relationship with the percentage observed in peripheral blood.
- Functional assays in early RA, estRA and HC:
    - sort cell subsets using flow cytometry: regulatory B cells ($CD19^{+}CD24^{hi}CD38^{hi}$), naive B cells ($CD19^{+}CD27^{-}IgD^{+}$), memory B cells ($CD19^{+}CD27^{+}IgD^{-}$) and T cells ($CD3^{+}CD4^{+}CD25^{-}$)
    - stimulate the B cell subsets and measure IL-10 secretion to confirm $CD19^{+}CD24^{hi}CD38^{hi}$ B cells are able to produce high levels of IL-10 in response to inflammation. Cell sorting is required because the markers, in particular CD24, are altered in response to cell stimulation limiting later identification of the subset in a mixed population of cells, such as PBMCs, for a secretion assay.
    - Co-culture experiments of B cell subsets and allogenic T cells to look at stimulation of T cells (measured by assessment of proliferation and CD25 induction), as well as assessment of Treg induction.

7. General Discussion

7.1 General Discussion

Interest in the role of B cells in the pathogenesis of RA was renewed by the success of B cell depletion for the treatment of RA. B cells are the source of autoantibody production, produce pro-inflammatory cytokines, act as antigen presenting cells activating or amplifying auto-reactive T cells, and a subset may have a regulatory function within the immune system. Therefore, B cells are likely to have a critical role in disease initiation and persistence but their specific role in RA pathogenesis remains unclear.

In this thesis, I have examined $CD19^{+}$ B cells from patients in a DMARD-naive early arthritis cohort to explore differences in gene expression and peripheral blood B cell subsets between patients with RA and disease controls. Transcriptomic, genetic and clinical data have been combined to examine the influence of genetic variants on gene expression at RA risk loci and to identify disease specific eQTLs.

The summary presented in this chapter will highlight the findings from my study and the potential insights obtained into B cell function in RA, including the genetic and environmental factors that influence the B cell transcriptome.

7.2 Gene expression data

In the primary comparison of RA samples to non-RA samples, a list of differentially expressed genes (DEGs) which was robust to multiple test correction (MTC) was not identified. However, pathway analyses of the DEGs obtained, after removal of the multiple test correction, revealed interesting themes which may provide an insight into the peripheral B cell compartment in RA.

'BCR signalling' is identified as a pathway which is downregulated in RA and is shown as a central component in the top molecular network fitted to the DEGs by IPA. Pathways downstream of the BCR: 'phospholipase C signalling' and 'the role of NFAT in the regulation of the immune response' are also shown to be downregulated. The GSEA results, which utilise a different methodology, add weight to this finding by identifying PI3K signalling and IP3 signalling as downregulated in RA, both pathways downstream of the BCR.

In RA, one might expect the BCR signalling to be upregulated, representing activation of the cell and promoting cell survival. However, BCR signalling is critical to B cell development, survival and activation, and so defects may promote the development of an autoimmune response, perhaps through the persistence of autoreactive clones if an adequate BCR signal is not generated at tolerance checkpoints. Intrinsic defects in B cell signalling have been suggested as a potential pathogenic mechanism in RA as, for example, carriers of the PTPN22 risk allele demonstrate impaired BCR signalling in response to stimulation[125, 126]. The PTPN22 risk allele has been linked to high frequencies of autoreactive clones within the transitional B cell subset, which have newly emerged from the bone marrow[127]. In this setting, carriers of the PTPN22 allele may be unable to generate an adequate BCR signal to self antigen to induce the required B cell tolerance mechanisms, leading to the persistence of autoreactive cells and disease initiation.

Alternatively, the downregulation of pathways downstream of the BCR may represent exhaustion of the peripheral B cells, analogous to that seen in T cells[174]. The GSEA results indicate that RA $CD19^+$ B cells have downregulated DNA repair mechanisms and RNA and protein processing, which may reflect an 'exhausted' or inactive cell state. A reduced BCR signal may also impair B cell survival, which has been shown to be dependent on PI3K signalling, a pathway identified as downregulated in the GSEA analysis, [30].

The samples used in this analysis are from peripheral blood and it is possible that, at the onset of disease, an activated set of B cells has already migrated to the joints or germinal centres[79]. The downregulation in the RA group of 'leucocyte extravasation signalling' and 'integrin signalling' may indicate that the cells able to migrate from the peripheral compartment have done so, hence the development of clinical symptoms.

In examining the $CD19^+$ B cells from peripheral blood I have not analysed the transcriptome of plasma cells and it is interesting to note that the GSEA analysis has also shown the 'B cell differentiation' pathway to be downregulated in the RA group which, in combination with the upregulation in *PRDM1*, may represent the shift in cell function to antibody producing cells which represents a major shift in the cellular machinery.

A further theme to emerge relates to IL-6, where the combination of IPA results and MSD data confirm the influence of IL-6 on the $CD19^+$ B cell transcriptome in RA. IL-6 is critical to plasma cell differentiation and so has an established effect on B cells in RA. This finding is not unexpected given the clinical success of anti-IL-6 treatments in RA, and underlines the influence of the cytokine on this cell type.

The pathway analyses of the DEG list without multiple test correction raises interesting themes regarding the state of the peripheral B cell compartment in RA but, importantly, a robust list of DEGs was not identified in the three analyses discussed in *Chapter 3*. Subsetting the data did identify a list of DEGs which withstood MTC between the RA and non-inflammatory samples but the addition of the clinical variables age, CRP and ESR led to the loss of any DEGs. This suggests that chronological age and inflammatory status, rather than a clinical diagnosis of RA, have a greater influence on the $CD19^+$ B cell transcriptome.

In *Chapter 4*, examination of the transcriptome using sample groups based on age, demonstrated a list of DEGs robust to multiple test correction. The IPA platform identified upregulation of pathways related to TNF, IL-2 and oxidative stress, *NOS2*, in the older age group. This is in keeping with the concept of inflamm-ageing, a chronic inflammatory state that develops with increasing chronological age. When measures of inflammatory status, CRP and ESR, are removed from the model, IL-17 and IL-1 were also identified as potential activated upstream regulators in the older age group.

The identification of *APOE* as an inhibited upstream regulator in the older age group raises the possibility that the influence of *APOE* in the ageing phenotype may be mediated by $CD19^+$ B cells. *APOE* variants are associated with longevity in GWAS studies, the $Apoe^{-/-}$ mouse is a model for atherosclerosis and, additionally, B cell dysregulation has been linked to atherosclerosis and so this finding provides a potential insight into the biological significance of the association of *APOE* and ageing[199, 201, 206].

Dividing the samples using ESR levels at the time of recruitment shows that ESR, rather than CRP, reflects changes in the $CD19^+$ B cell transcriptome. In samples with a higher ESR the activated upstream regulators in the DEG list (with MTC) include *XBP1*,

indicating potential differentiation into plasma cells, *GMCSF*, indicating B cell activation, and transcription factors associated with tumorigenesis. The $CD19^+$ B cells from samples with a higher ESR may thus possess a profile favouring survival, proliferation and antibody secretion.

The results described here are intriguing, particularly in light of the multiple test corrected list of DEGs based on age or inflammatory status. This identifies these factors as a greater influence on the $CD19^+$ B cell transcriptome than disease in this cohort. These factors, and additional inter-individual variability, may explain the absence of robust gene signature differences between disease groups in this cohort. If this study were repeated we would need large sample groups and careful matching of the samples for age and ESR. An alternative explanation for the absence of a robust diagnostic signature is that analysis of the $CD19^+$ B cells may miss a signal from a relatively small subset of B cells, which may be masked in the heterogeneous B-cell population.

The pathway analyses reported here require further validation. IPA and GSEA are well established analytical procedures but they reflect the literature, the quality of the manual curation and rely on regular maintenance. Transcriptomic data from B cells is not as plentiful as that from other cell types, hence the novelty of this study, and the pathways identified may be reflective of findings in different immune subsets, PBMCs and tumour samples. As transcriptomic data from $CD19^+$ B cells is increasingly published and shared, the pathway analyses employed here may become more fruitful. A further caveat is that the pathways identified by IPA require interpretation in the context of the disease under study. In the canonical pathways identified, for example, there is marked overlap in the molecules between pathways and so not all may be relevant.

7.3 eQTL analysis

The eQTL analysis at the 101 RA risk loci explored the downstream correlates of known genetic variation and highlights the importance of studying appropriate cell subsets when analysing the cellular mechanisms underlying genetic risk. The eQTL analysis in $CD19^+$ B cells identified 194 *cis* acting significant SNP-transcript associations, corresponding to 10 unique genes. After validation of these findings at the protein level, this will provide a

foundation for the prioritisation of genes for further downstream, functional studies in B cells.

At the 8p23 locus *FAM167A* and *BLK* were both subject to eQTLs in $CD19^+$ B cells. The focus at the locus has predominantly been on *BLK* which phosphorylates tyrosine residues in the ITAM of the BCR. The RA risk haplotype at this region has been associated with decreased expression of *BLK* in B cells and evidence of lower basal BCR signalling activity yet the cells have been shown to be hyperactive after crosslinking of the BCR, with enhanced B cell-T cell interactions[264]. Simpfendorfer *et al* have also demonstrated the *cis* regulatory effect on *BLK* and confirmed this at the protein level but this finding was restricted to early B cells, leading to the suggestion that the influence of this risk variant may be critical in the early stages of B cell development[265]. Interestingly, these findings highlight the potential significance of an alteration in BCR signalling in the initiation of RA and further eQTL analyses of subsets within the peripheral $CD19^+$ B cell population may shed light on the functional consequences of genetic variants. FCRL3, which was also identified as an eQTL in $CD19^+$ B cells here, has also been shown to potentially inhibit BCR signalling, altering the activation threshold and promoting the breakdown of tolerance[266].

It is noteworthy that the eQTL effects seen at the 101 RA risk loci did not differ between the RA and non-RA groups, and incorporation of additional clinical covariates to the linear model did not alter the eQTL list generated. There were, therefore, no disease specific eQTLs identified at the RA risk loci.

21 disease specific eQTLs were successfully identified using an genome wide interaction analysis and these lie outside the known 101 RA risk loci. This finding is in keeping with the IBD literature where the IBD-specific eQTLs were also outside the currently known disease risk loci[133]. The description here, in $CD19^+$ B cells, is novel and confirms the success of the methodology. One limitation, however, is the imbalance of genotypes between the disease groups and I would next repeat the analysis with a change to the genotype filter to a minimum of 3 samples of each genotype by disease group, prior to attempting to validate the data at the protein level.

This methodology has the potential to provide new insights into the inherited risk of RA, given the lack of overlap with known risk loci. The themes which emerge from the genes identified through the interaction analysis relate to transcription, chromatin remodelling and metabolic function, which may reflect changes in the activation status of the B cell.

A fundamental challenge to the interpretation of RA risk loci from GWAS has been the number of genes implicated. However, I have shown here that gene prioritisation can be addressed using eQTL analyses and that disease specific eQTLs can be identified using interaction analyses, providing potential new insights into disease mechanisms. Furthermore, improvements in cell sorting by flow cytometry will, in the future, enable similar analyses of subsets of $CD19^+$ B cells, which may provide further insights into pathogenesis.

7.4 B cell subsets

The frequency of the transitional, postulated regulatory, $CD19^+CD24^{hi}CD38^{hi}$ subset was increased in the RA cohort in my study, when compared to the cohorts with other inflammatory conditions. Alongside age and gender, the frequency of this subset was an independent predictor of RA. The results are in contrast to the data in the literature from an established RA cohort[61]. This difference may be related to the stage of disease: the increase in this postulated regulatory subset in DMARD-naive RA may be an attempt to control the autoimmune process at this early clinical stage or a rebound increase given the previously described functional impairment of this subset in RA[61].

The $CD19^+CD24^{hi}CD38^{hi}$ subset represents transitional B cells and so, alternatively, an increase may reflect an increase in autoreactive cells within this newly emigrant population or a compensatory response to migration of $CD19^+$ B cells from the blood to the joints. The frequency of $CD19^+$ B cells overall (as a % of leucocytes) is reduced in the RA population here compared to non-RA samples.

My data represent a snapshot of the peripheral B cell compartment and I would be most interested to obtain follow up samples to determine if the frequency of the $CD19^+CD24^{hi}CD38^{hi}$ subset changes with disease duration/progression and if it is related

to prognosis or treatment response. Functional work is required to determine if the observation represents an increase in autoreactive cells, functional regulatory cells or functionally impaired regulatory cells.

In light of the increased frequency of this subset in RA compared to other inflammatory conditions, that it is an independent predictor of RA, and the identification of CD38 as a potential activated upstream regulator in RA in the gene expression data, it would be interesting to study this cell subset more closely through functional and transcriptomic work.

7.5 Strengths and weaknesses

The strengths and weaknesses of my study are summarised in table 7.1.

Strengths	***Weaknesses***
Large cohort of DMARD naive samples	No $CD19^+$ B cell purity checks for a proportion of samples
Parallel genetic, transcriptomic, clinical and flow cytometry data	No validation of transcriptomic or eQTL findings
Transcriptomic analysis using 2 methods: IPA and GSEA	Absence of longitudinal flow cytometry data
Examination of the transcriptomic profile to identify the influence of environmental factors beyond disease: age and inflammation	No functional analyses of $CD19^+CD24^{hi}CD38^{hi}$ subset
eQTL analysis in a clinically relevant cell subset	No parallel data for transcriptomics or flow cytometry from synovia
Methodology for interaction analysis tested in $CD19^+$ B cells	Small established RA and healthy control groups for flow cytometry data comparisons
Flow cytometry data for B cell subsets in a DMARD and steroid naive cohort	Interaction analysis not repeated with the inclusion of clinical covariates

Table 7.1 Experimental strengths and weaknesses

The primary strength of this study comes from the recruitment of a large cohort of DMARD-naive patients with early arthritis and the parallel clinical, transcriptomic, genetic and flow cytometry data. Studies in $CD19^+$ B cells at a transcriptomic level have been limited due to problems with cell purity and isolating sufficient cells, so the transcriptomic data here are a useful, clinically relevant resource. The data have been rigorously analysed and I am confident that a diagnostic gene signature has not been missed.

The absence of a clear gene signature may be due to the heterogeneity of the $CD19^+$ B cell population, migration of a pathogenic subset to sites of inflammation, cell contamination during the $CD19^+$ B cell isolation, differences in genetic backgrounds between patients and the potential influence of additional environmental factors.
The heterogeneity of $CD19^+$ B cells may mean that changes in a small, potentially pathogenic subset, such as transitional B cells, may be masked by other subsets. An additional consideration is that peripheral plasma cells were not examined; they are rarely found in the peripheral blood but this subset may be informative in a disease such as RA. In a small group of patients at risk of RA the presence of ≥ 5 dominant BCR clones in peripheral blood was associated with the development of RA, but at the onset of RA these clones were no longer detectable in the periphery and appeared in the synovial tissue[79].

The clinical diagnosis of RA encompasses a broad population; including patients who with differing autoantibody status, clinical patterns of disease, genetic backgrounds, synovial findings and IFN scores in a single group[74, 109, 147]. The RA cohort studied here was heterogeneous and this may be a contributing factor in the absence of a robust DEG list between the disease comparator groups.

The RA cohort includes seropositive and seronegative patients and it has become increasingly clear, since this project started, that the two groups may represent different diseases. It is established that the heritability of RA is greater in the ACPA positive group and the ‘shared epitope’ hypothesis is more strongly associated with ACPA positive RA[112, 113, 116, 120]. Indeed, the ACPA positive and ACPA negative RA may be described as genetically different disease subsets, which share only a small proportion of genetic susceptibility factors[118]. In addition, the environmental risk factors smoking and the periodontal pathogen, Porphyromonas gingivalis, are associated with the

development of RA in ACPA positive disease in particular [267, 268]. In genetically predisposed individuals, citrullination induced by these factors may promote the development of autoantibodies [268, 269]. There is also further evidence from therapeutics studies that seropositivity is a prognostic factor for response to treatment, not just to rituximab, but also to adalimumab and abatacept [82, 270].

In light of the developments in the understanding of seronegative and seropositive RA over the course of this project, the gene expression data was analysed *post hoc* using first ACPA positive samples and, secondly, seropositive (RF and/or ACPA positive) samples as the disease group. These analyses were carried out using the procedures and comparisons described in *Chapter 3* and did not identify lists of differentially expressed genes which stood up to multiple test correction. The reduced sample size in the disease group for these analyses may, of course, be a factor in the absence of differentially expressed genes.

There is further heterogeneity in the comparator non-RA group, which includes patients with inflammatory and non-inflammatory conditions. The influence of inflammation, as measured by ESR, on the $CD19^+$ B cell transcriptome is highlighted in *Chapter 4.* The additional analyses referred to in *Chapter 4*, divided the original non-RA group into subsets to allow separate comparisons between RA samples and other inflammatory samples alone and with non-inflammatory alone. No DEGs robust to multiple test correction were identified in the comparison between RA samples and other inflammatory samples. A list of DEGs robust to multiple test correction was identified in the comparison between RA samples and non-inflammatory samples but no DEGs remained when the clinical variables ESR, CRP and age were added to the linear model. The work here indicates that in future projects, it would be beneficial for samples in the comparator groups to be matched for age and inflammatory status to improve the likelihood of detecting relevant, disease specific differentially expressed genes.

The interpretation of big datasets, such as the one described in this thesis, is reliant on the technology of pathway analysis platforms which are, in turn, dependent on regular maintenance and manual curation. I attempted to use platforms that are maintained; however, the documentation that informs the pathways identified often requires manual curation. The output from such platforms must always be interpreted in the clinical

context and with knowledge of the relevant literature. However, this unsupervised approach does have the advantage of potentially identifying novel discoveries.

7.6 Conclusion

This study has not demonstrated differential changes in gene expression in the $CD19^+$ B cell transcriptome attributable to a diagnosis of RA. However, it is notable that a robust signature is identified when the samples are divided based on the median chronological age and inflammatory status, measured by ESR. In combination with the eQTL data this confirms the marked influence of genetic and environmental factors on gene expression in $CD19^+$ B cells in peripheral blood.

The mechanisms of RA disease initiation and persistence are no doubt complex but the findings presented here may indicate a downregulation of BCR signalling leading to a breakdown in tolerance and an increase in autoreactive cells, seen here as an increase in the transitional B cell subset. Closer examination into the functional and transcriptomic profile of transitional B cells may provide further insights into this theory.

References

1. Young, A., et al., *Mortality in rheumatoid arthritis. Increased in the early course of disease, in ischaemic heart disease and in pulmonary fibrosis.* Rheumatology (Oxford), 2007. **46**(2): p. 350-7.
2. NAO, *Services for people with Rheumatoid Arthritis.* 2009.
3. Kvien, T.K., et al., *Epidemiological aspects of rheumatoid arthritis: the sex ratio.* Ann N Y Acad Sci, 2006. **1069**: p. 212-22.
4. Humphreys, J.H., et al., *The incidence of rheumatoid arthritis in the UK: comparisons using the 2010 ACR/EULAR classification criteria and the 1987 ACR classification criteria. Results from the Norfolk Arthritis Register.* Ann Rheum Dis, 2013. **72**(8): p. 1315-20.
5. Wellcome Trust Case Control, C., *Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls.* Nature, 2007. **447**(7145): p. 661-78.
6. Silman, A.J., et al., *Twin concordance rates for rheumatoid arthritis: results from a nationwide study.* Br J Rheumatol, 1993. **32**(10): p. 903-7.
7. Symmons, D.P., et al., *Blood transfusion, smoking, and obesity as risk factors for the development of rheumatoid arthritis: results from a primary care-based incident case-control study in Norfolk, England.* Arthritis Rheum, 1997. **40**(11): p. 1955-61.
8. Bergstrom, U., et al., *Smoking, low formal level of education, alcohol consumption, and the risk of rheumatoid arthritis.* Scand J Rheumatol, 2013. **42**(2): p. 123-30.
9. NICE, *Rheumatoid Arthritis in adults: management (NG100).* 2018.
10. Sebbag, M., et al., *Clinical and pathophysiological significance of the autoimmune response to citrullinated proteins in rheumatoid arthritis.* Joint Bone Spine, 2004. **71**(6): p. 493-502.
11. Aletaha, D., et al., *2010 Rheumatoid arthritis classification criteria: an American College of Rheumatology/European League Against Rheumatism collaborative initiative.* Arthritis Rheum, 2010. **62**(9): p. 2569-81.
12. Lard, L.R., et al., *Early versus delayed treatment in patients with recent-onset rheumatoid arthritis: comparison of two cohorts who received different treatment strategies.* Am J Med, 2001. **111**(6): p. 446-51.
13. Grigor, C., et al., *Effect of a treatment strategy of tight control for rheumatoid arthritis (the TICORA study): a single-blind randomised controlled trial.* Lancet, 2004. **364**(9430): p. 263-9.
14. van der Linden, M.P., et al., *Long-term impact of delay in assessment of patients with early arthritis.* Arthritis Rheum, 2010. **62**(12): p. 3537-46.
15. Smolen, J.S., et al., *Treating rheumatoid arthritis to target: recommendations of an international task force.* Ann Rheum Dis, 2010. **69**(4): p. 631-7.
16. Pratt, A.G., et al., *A CD4 T cell gene signature for early rheumatoid arthritis implicates interleukin 6-mediated STAT3 signalling, particularly in anti-citrullinated peptide antibody-negative disease.* Ann Rheum Dis, 2012. **71**(8): p. 1374-81.
17. van der Helm-van Mil, A.H., et al., *Validation of a prediction rule for disease outcome in patients with recent-onset undifferentiated arthritis: moving toward*

individualized treatment decision-making. Arthritis Rheum, 2008. **58**(8): p. 2241-7.
18. Chang, S.K., Z. Gu, and M.B. Brenner, *Fibroblast-like synoviocytes in inflammatory arthritis pathology: the emerging role of cadherin-11.* Immunol Rev, 2010. **233**(1): p. 256-66.
19. Aupperle, K.R., et al., *Regulation of synoviocyte proliferation, apoptosis, and invasion by the p53 tumor suppressor gene.* Am J Pathol, 1998. **152**(4): p. 1091-8.
20. McInnes, I.B. and G. Schett, *The pathogenesis of rheumatoid arthritis.* N Engl J Med, 2011. **365**(23): p. 2205-19.
21. Smolen, J.S. and G. Steiner, *Therapeutic strategies for rheumatoid arthritis.* Nat Rev Drug Discov, 2003. **2**(6): p. 473-88.
22. Schett, G. and E. Gravallese, *Bone erosion in rheumatoid arthritis: mechanisms, diagnosis and treatment.* Nat Rev Rheumatol, 2012. **8**(11): p. 656-64.
23. Pratt, A.G., J.D. Isaacs, and D.L. Mattey, *Current concepts in the pathogenesis of early rheumatoid arthritis.* Best Pract Res Clin Rheumatol, 2009. **23**(1): p. 37-48.
24. Cope, A.P., H. Schulze-Koops, and M. Aringer, *The central role of T cells in rheumatoid arthritis.* Clin Exp Rheumatol, 2007. **25**(5 Suppl 46): p. S4-11.
25. Edwards, J.C., et al., *Efficacy of B-cell-targeted therapy with rituximab in patients with rheumatoid arthritis.* N Engl J Med, 2004. **350**(25): p. 2572-81.
26. Humby, F., et al., *Ectopic lymphoid structures support ongoing production of class-switched autoantibodies in rheumatoid synovium.* PLoS Med, 2009. **6**(1): p. e1.
27. Young, R.M. and L.M. Staudt, *Targeting pathological B cell receptor signalling in lymphoid malignancies.* Nat Rev Drug Discov, 2013. **12**(3): p. 229-43.
28. Lam, K.P., R. Kuhn, and K. Rajewsky, *In vivo ablation of surface immunoglobulin on mature B cells by inducible gene targeting results in rapid cell death.* Cell, 1997. **90**(6): p. 1073-83.
29. Kraus, M., et al., *Survival of resting mature B lymphocytes depends on BCR signaling via the Igalpha/beta heterodimer.* Cell, 2004. **117**(6): p. 787-800.
30. Srinivasan, L., et al., *PI3 kinase signals BCR-dependent mature B cell survival.* Cell, 2009. **139**(3): p. 573-86.
31. Nemazee, D., *Mechanisms of central tolerance for B cells.* Nat Rev Immunol, 2017. **17**(5): p. 281-294.
32. Bashford-Rogers, R.J.M., K.G.C. Smith, and D.C. Thomas, *Antibody repertoire analysis in polygenic autoimmune diseases.* Immunology, 2018.
33. Wardemann, H., et al., *Predominant autoantibody production by early human B cell precursors.* Science, 2003. **301**(5638): p. 1374-7.
34. De Silva, N.S. and U. Klein, *Dynamics of B cells in germinal centres.* Nat Rev Immunol, 2015. **15**(3): p. 137-48.
35. Tiller, T., et al., *Autoreactivity in human IgG+ memory B cells.* Immunity, 2007. **26**(2): p. 205-13.
36. Jacob, J., et al., *Intraclonal generation of antibody mutants in germinal centres.* Nature, 1991. **354**(6352): p. 389-92.
37. Schaerli, P., et al., *CXC chemokine receptor 5 expression defines follicular homing T cells with B cell helper function.* J Exp Med, 2000. **192**(11): p. 1553-62.

38. Victora, G.D., et al., *Germinal center dynamics revealed by multiphoton microscopy with a photoactivatable fluorescent reporter.* Cell, 2010. **143**(4): p. 592-605.
39. Shulman, Z., et al., *Dynamic signaling by T follicular helper cells during germinal center B cell selection.* Science, 2014. **345**(6200): p. 1058-62.
40. Zhang, Y., et al., *Germinal center B cells govern their own fate via antibody feedback.* J Exp Med, 2013. **210**(3): p. 457-64.
41. Zhang, Y., L. Garcia-Ibanez, and K.M. Toellner, *Regulation of germinal center B-cell differentiation.* Immunol Rev, 2016. **270**(1): p. 8-19.
42. Chan, T.D., et al., *Elimination of germinal-center-derived self-reactive B cells is governed by the location and concentration of self-antigen.* Immunity, 2012. **37**(5): p. 893-904.
43. Kurosaki, T., K. Kometani, and W. Ise, *Memory B cells.* Nat Rev Immunol, 2015. **15**(3): p. 149-59.
44. Nutt, S.L., et al., *The generation of antibody-secreting plasma cells.* Nat Rev Immunol, 2015. **15**(3): p. 160-71.
45. Sims, G.P., et al., *Identification and characterization of circulating human transitional B cells.* Blood, 2005. **105**(11): p. 4390-8.
46. Bemark, M., et al., *Translational Mini-Review Series on B cell subsets in disease. Reconstitution after haematopoietic stem cell transplantation - revelation of B cell developmental pathways and lineage phenotypes.* Clin Exp Immunol, 2012. **167**(1): p. 15-25.
47. Simon, Q., et al., *In-depth characterization of CD24(high)CD38(high) transitional human B cells reveals different regulatory profiles.* J Allergy Clin Immunol, 2016. **137**(5): p. 1577-1584 e10.
48. Klein, U., K. Rajewsky, and R. Kuppers, *Human immunoglobulin (Ig)M+IgD+ peripheral blood B cells expressing the CD27 cell surface antigen carry somatically mutated variable region genes: CD27 as a general marker for somatically mutated (memory) B cells.* J Exp Med, 1998. **188**(9): p. 1679-89.
49. Fecteau, J.F., G. Cote, and S. Neron, *A new memory CD27-IgG+ B cell population in peripheral blood expressing VH genes with low frequency of somatic mutation.* J Immunol, 2006. **177**(6): p. 3728-36.
50. Mauri, C. and A. Bosma, *Immune regulatory function of B cells.* Annu Rev Immunol, 2012. **30**: p. 221-41.
51. Newell, K.A., et al., *Identification of a B cell signature associated with renal transplant tolerance in humans.* J Clin Invest, 2010. **120**(6): p. 1836-47.
52. Mauri, C., et al., *Prevention of arthritis by interleukin 10-producing B cells.* J Exp Med, 2003. **197**(4): p. 489-501.
53. Fillatreau, S., et al., *B cells regulate autoimmunity by provision of IL-10.* Nat Immunol, 2002. **3**(10): p. 944-50.
54. Blair, P.A., et al., *CD19(+)CD24(hi)CD38(hi) B cells exhibit regulatory capacity in healthy individuals but are functionally impaired in systemic Lupus Erythematosus patients.* Immunity, 2010. **32**(1): p. 129-40.
55. Menon, M., et al., *A Regulatory Feedback between Plasmacytoid Dendritic Cells and Regulatory B Cells Is Aberrant in Systemic Lupus Erythematosus.* Immunity, 2016. **44**(3): p. 683-697.

56. Wortel, C.M. and S. Heidt, *Regulatory B cells: Phenotype, function and role in transplantation.* Transpl Immunol, 2017. **41**: p. 1-9.
57. Mauri, C. and M. Menon, *Human regulatory B cells in health and disease: therapeutic potential.* J Clin Invest, 2017. **127**(3): p. 772-779.
58. Bouaziz, J.D., et al., *IL-10 produced by activated human B cells regulates CD4(+) T-cell activation in vitro.* Eur J Immunol, 2010. **40**(10): p. 2686-91.
59. Iwata, Y., et al., *Characterization of a rare IL-10-competent B-cell subset in humans that parallels mouse regulatory B10 cells.* Blood, 2011. **117**(2): p. 530-41.
60. Chesneau, M., et al., *Tolerant Kidney Transplant Patients Produce B Cells with Regulatory Properties.* J Am Soc Nephrol, 2015. **26**(10): p. 2588-98.
61. Flores-Borja, F., et al., *CD19+CD24hiCD38hi B cells maintain regulatory T cells while limiting TH1 and TH17 differentiation.* Sci Transl Med, 2013. **5**(173): p. 173ra23.
62. Ma, L., et al., *Reduced numbers of regulatory B cells are negatively correlated with disease activity in patients with new-onset rheumatoid arthritis.* Clin Rheumatol, 2014. **33**(2): p. 187-95.
63. Le Texier, L., et al., *Long-term allograft tolerance is characterized by the accumulation of B cells exhibiting an inhibited profile.* Am J Transplant, 2011. **11**(3): p. 429-38.
64. Shabir, S., et al., *Transitional B lymphocytes are associated with protection from kidney allograft rejection: a prospective study.* Am J Transplant, 2015. **15**(5): p. 1384-91.
65. Cherukuri, A., et al., *Immunologic human renal allograft injury associates with an altered IL-10/TNF-alpha expression ratio in regulatory B cells.* J Am Soc Nephrol, 2014. **25**(7): p. 1575-85.
66. Rantapaa-Dahlqvist, S., et al., *Antibodies against cyclic citrullinated peptide and IgA rheumatoid factor predict the development of rheumatoid arthritis.* Arthritis Rheum, 2003. **48**(10): p. 2741-9.
67. Nielen, M.M., et al., *Specific autoantibodies precede the symptoms of rheumatoid arthritis: a study of serial measurements in blood donors.* Arthritis Rheum, 2004. **50**(2): p. 380-6.
68. Isaacs, J.D., *The changing face of rheumatoid arthritis: sustained remission for all?* Nat Rev Immunol, 2010. **10**(8): p. 605-11.
69. Malmstrom, V., A.I. Catrina, and L. Klareskog, *The immunopathogenesis of seropositive rheumatoid arthritis: from triggering to targeting.* Nat Rev Immunol, 2017. **17**(1): p. 60-75.
70. van Venrooij, W.J. and A.J. Zendman, *Anti-CCP2 antibodies: an overview and perspective of the diagnostic abilities of this serological marker for early rheumatoid arthritis.* Clin Rev Allergy Immunol, 2008. **34**(1): p. 36-9.
71. Juarez, M., et al., *Identification of novel antiacetylated vimentin antibodies in patients with early inflammatory arthritis.* Ann Rheum Dis, 2016. **75**(6): p. 1099-107.
72. Hill, J.A., et al., *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.* J Immunol, 2003. **171**(2): p. 538-41.

73. Wegner, N., et al., *Autoimmunity to specific citrullinated proteins gives the first clues to the etiology of rheumatoid arthritis.* Immunol Rev, 2010. **233**(1): p. 34-54.
74. Corsiero, E., et al., *Ectopic Lymphoid Structures: Powerhouse of Autoimmunity.* Front Immunol, 2016. **7**: p. 430.
75. Takemura, S., et al., *Lymphoid neogenesis in rheumatoid synovitis.* J Immunol, 2001. **167**(2): p. 1072-80.
76. Thurlings, R.M., et al., *Synovial lymphoid neogenesis does not define a specific clinical rheumatoid arthritis phenotype.* Arthritis Rheum, 2008. **58**(6): p. 1582-9.
77. Lino, A.C., et al., *Cytokine-producing B cells: a translational view on their roles in human and mouse autoimmune diseases.* Immunol Rev, 2016. **269**(1): p. 130-44.
78. Doorenspleet, M.E., et al., *Rheumatoid arthritis synovial tissue harbours dominant B-cell and plasma-cell clones associated with autoreactivity.* Ann Rheum Dis, 2014. **73**(4): p. 756-62.
79. Tak, P.P., et al., *Dominant B cell receptor clones in peripheral blood predict onset of arthritis in individuals at risk for rheumatoid arthritis.* Ann Rheum Dis, 2017. **76**(11): p. 1924-1930.
80. Cantaert, T., et al., *B lymphocyte autoimmunity in rheumatoid synovitis is independent of ectopic lymphoid neogenesis.* J Immunol, 2008. **181**(1): p. 785-94.
81. NICE, *Adalimumab, etanercept, infliximab, rituximab and abatacept for the treatment of rheumatoid arthritis after the failure of a TNF inhibitor.* . 2010.
82. Isaacs, J.D., et al., *Effect of baseline rheumatoid factor and anticitrullinated peptide antibody serotype on rituximab clinical response: a meta-analysis.* Ann Rheum Dis, 2013. **72**(3): p. 329-36.
83. Nakken, B., et al., *B-cells and their targeting in rheumatoid arthritis--current concepts and future perspectives.* Autoimmun Rev, 2011. **11**(1): p. 28-34.
84. Cohen, M.D. and E. Keystone, *Rituximab for Rheumatoid Arthritis.* Rheumatol Ther, 2015. **2**(2): p. 99-111.
85. www.rituxan.com, 2018.
86. Ruyssen-Witrand, A., et al., *Association between -871C>T promoter polymorphism in the B-cell activating factor gene and the response to rituximab in rheumatoid arthritis patients.* Rheumatology (Oxford), 2013. **52**(4): p. 636-41.
87. Ruyssen-Witrand, A., et al., *Fcgamma receptor type IIIA polymorphism influences treatment outcomes in patients with rheumatoid arthritis treated with rituximab.* Ann Rheum Dis, 2012. **71**(6): p. 875-7.
88. Juge, P.A., et al., *Variants of genes implicated in type 1 interferon pathway and B-cell activation modulate the EULAR response to rituximab at 24 weeks in rheumatoid arthritis.* RMD Open, 2017. **3**(2): p. e000448.
89. Dass, S., et al., *Highly sensitive B cell analysis predicts response to rituximab therapy in rheumatoid arthritis.* Arthritis Rheum, 2008. **58**(10): p. 2993-9.
90. Boumans, M.J., et al., *Response to rituximab in patients with rheumatoid arthritis in different compartments of the immune system.* Arthritis Rheum, 2011. **63**(11): p. 3187-94.
91. Leandro, M.J., et al., *Reconstitution of peripheral blood B cells after depletion with rituximab in patients with rheumatoid arthritis.* Arthritis Rheum, 2006. **54**(2): p. 613-20.

92. Roll, P., et al., *Regeneration of B cell subsets after transient B cell depletion using anti-CD20 antibodies in rheumatoid arthritis.* Arthritis Rheum, 2006. **54**(8): p. 2377-86.
93. Sellam, J., et al., *Blood memory B cells are disturbed and predict the response to rituximab in patients with rheumatoid arthritis.* Arthritis Rheum, 2011. **63**(12): p. 3692-701.
94. Roll, P., T. Dorner, and H.P. Tony, *Anti-CD20 therapy in patients with rheumatoid arthritis: predictors of response and B cell subset regeneration after repeated treatment.* Arthritis Rheum, 2008. **58**(6): p. 1566-75.
95. Moller, B., et al., *Class-switched B cells display response to therapeutic B-cell depletion in rheumatoid arthritis.* Arthritis Res Ther, 2009. **11**(3): p. R62.
96. Vital, E.M., et al., *Management of nonresponse to rituximab in rheumatoid arthritis: predictors and outcome of re-treatment.* Arthritis Rheum, 2010. **62**(5): p. 1273-9.
97. Nakou, M., et al., *Rituximab therapy reduces activated B cells in both the peripheral blood and bone marrow of patients with rheumatoid arthritis: depletion of memory B cells correlates with clinical response.* Arthritis Res Ther, 2009. **11**(4): p. R131.
98. Kavanaugh, A., et al., *Assessment of rituximab's immunomodulatory synovial effects (ARISE trial). 1: clinical and synovial biomarker results.* Ann Rheum Dis, 2008. **67**(3): p. 402-8.
99. Thurlings, R.M., et al., *Synovial tissue response to rituximab: mechanism of action and identification of biomarkers of response.* Ann Rheum Dis, 2008. **67**(7): p. 917-25.
100. Anolik, J.H., et al., *Rituximab improves peripheral B cell abnormalities in human systemic lupus erythematosus.* Arthritis Rheum, 2004. **50**(11): p. 3580-90.
101. Owczarczyk, K., et al., *A plasmablast biomarker for nonresponse to antibody therapy to CD20 in rheumatoid arthritis.* Sci Transl Med, 2011. **3**(101): p. 101ra92.
102. van der Pouw Kraan, T.C., et al., *Rheumatoid arthritis subtypes identified by genomic profiling of peripheral blood cells: assignment of a type I interferon signature in a subpopulation of patients.* Ann Rheum Dis, 2007. **66**(8): p. 1008-14.
103. Thurlings, R.M., et al., *Relationship between the type I interferon signature and the response to rituximab in rheumatoid arthritis patients.* Arthritis Rheum, 2010. **62**(12): p. 3607-14.
104. Raterman, H.G., et al., *The interferon type I signature towards prediction of non-response to rituximab in rheumatoid arthritis patients.* Arthritis Res Ther, 2012. **14**(2): p. R95.
105. Sellam, J., et al., *Use of whole-blood transcriptomic profiling to highlight several pathophysiologic pathways associated with response to rituximab in patients with rheumatoid arthritis: data from a randomized, controlled, open-label trial.* Arthritis Rheumatol, 2014. **66**(8): p. 2015-25.
106. Sellam, J., et al., *Serum IL-33, a new marker predicting response to rituximab in rheumatoid arthritis.* Arthritis Res Ther, 2016. **18**(1): p. 294.
107. Okada, Y., et al., *Genetics of rheumatoid arthritis contributes to biology and drug discovery.* Nature, 2014. **506**(7488): p. 376-81.
108. https://ghr.nlm.nih.gov/primer/genomicresearch/snp.

109. Deane, K.D., et al., *Genetic and environmental risk factors for rheumatoid arthritis.* Best Pract Res Clin Rheumatol, 2017. **31**(1): p. 3-18.
110. Kurko, J., et al., *Genetics of rheumatoid arthritis - a comprehensive review.* Clin Rev Allergy Immunol, 2013. **45**(2): p. 170-9.
111. Manolio, T.A., et al., *Finding the missing heritability of complex diseases.* Nature, 2009. **461**(7265): p. 747-53.
112. MacGregor, A.J., et al., *Characterizing the quantitative genetic contribution to rheumatoid arthritis using data from twins.* Arthritis Rheum, 2000. **43**(1): p. 30-7.
113. Stahl, E.A., et al., *Bayesian inference analyses of the polygenic architecture of rheumatoid arthritis.* Nat Genet, 2012. **44**(5): p. 483-9.
114. Bossini-Castillo, L., et al., *A genome-wide association study of rheumatoid arthritis without antibodies against citrullinated peptides.* Ann Rheum Dis, 2015. **74**(3): p. e15.
115. Frisell, T., et al., *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 Rheum, 2013. **65**(11): p. 2773-82.
116. Padyukov, L., et al., *A genome-wide association study suggests contrasting associations in ACPA-positive versus ACPA-negative rheumatoid arthritis.* Ann Rheum Dis, 2011. **70**(2): p. 259-65.
117. Viatte, S., et al., *Genetic markers of rheumatoid arthritis susceptibility in anti-citrullinated peptide antibody negative patients.* Ann Rheum Dis, 2012. **71**(12): p. 1984-90.
118. Viatte, S., et al., *Replication of Associations of Genetic Loci Outside the HLA Region With Susceptibility to Anti-Cyclic Citrullinated Peptide-Negative Rheumatoid Arthritis.* Arthritis Rheumatol, 2016. **68**(7): p. 1603-13.
119. Raychaudhuri, S., et al., *Five amino acids in three HLA proteins explain most of the association between MHC and seropositive rheumatoid arthritis.* Nat Genet, 2012. **44**(3): p. 291-6.
120. Gregersen, P.K., J. Silver, and R.J. Winchester, *The shared epitope hypothesis. An approach to understanding the molecular genetics of susceptibility to rheumatoid arthritis.* Arthritis Rheum, 1987. **30**(11): p. 1205-13.
121. Cui, J., et al., *Genome-wide association study of determinants of anti-cyclic citrullinated peptide antibody titer in adults with rheumatoid arthritis.* Mol Med, 2009. **15**(5-6): p. 136-43.
122. Hensvold, A.H., et al., *Environmental and genetic factors in the development of anticitrullinated protein antibodies (ACPAs) and ACPA-positive rheumatoid arthritis: an epidemiological investigation in twins.* Ann Rheum Dis, 2015. **74**(2): p. 375-80.
123. Han, B., et al., *Fine mapping seronegative and seropositive rheumatoid arthritis to shared and distinct HLA alleles by adjusting for the effects of heterogeneity.* Am J Hum Genet, 2014. **94**(4): p. 522-32.
124. Begovich, A.B., et al., *A missense single-nucleotide polymorphism in a gene encoding a protein tyrosine phosphatase (PTPN22) is associated with rheumatoid arthritis.* Am J Hum Genet, 2004. **75**(2): p. 330-7.
125. Rieck, M., et al., *Genetic variation in PTPN22 corresponds to altered function of T and B lymphocytes.* J Immunol, 2007. **179**(7): p. 4704-10.

126. Arechiga, A.F., et al., *Cutting edge: the PTPN22 allelic variant associated with autoimmunity impairs B cell signaling.* J Immunol, 2009. **182**(6): p. 3343-7.
127. Menard, L., et al., *The PTPN22 allele encoding an R620W variant interferes with the removal of developing autoreactive B cells in humans.* J Clin Invest, 2011. **121**(9): p. 3635-44.
128. Nica, A.C. and E.T. Dermitzakis, *Expression quantitative trait loci: present and future.* Philos Trans R Soc Lond B Biol Sci, 2013. **368**(1620): p. 20120362.
129. Morley, M., et al., *Genetic analysis of genome-wide variation in human gene expression.* Nature, 2004. **430**(7001): p. 743-7.
130. Nica, A.C., et al., *The architecture of gene regulatory variation across multiple human tissues: the MuTHER study.* PLoS Genet, 2011. **7**(2): p. e1002003.
131. Flutre, T., et al., *A statistical framework for joint eQTL analysis in multiple tissues.* PLoS Genet, 2013. **9**(5): p. e1003486.
132. Fairfax, B.P., et al., *Genetics of gene expression in primary immune cells identifies cell type-specific master regulators and roles of HLA alleles.* Nat Genet, 2012. **44**(5): p. 502-10.
133. Peters, J.E., et al., *Insight into Genotype-Phenotype Associations through eQTL Mapping in Multiple Cell Types in Health and Immune-Mediated Disease.* PLoS Genet, 2016. **12**(3): p. e1005908.
134. Abbas, A.R., et al., *Immune response in silico (IRIS): immune-specific genes identified from a compendium of microarray expression data.* Genes Immun, 2005. **6**(4): p. 319-31.
135. Abbas, A.R., et al., *Deconvolution of blood microarray data identifies cellular activation patterns in systemic lupus erythematosus.* PLoS One, 2009. **4**(7): p. e6098.
136. Fairfax, B.P., et al., *Innate immune activity conditions the effect of regulatory variants upon monocyte gene expression.* Science, 2014. **343**(6175): p. 1246949.
137. Nicolae, D.L., et al., *Trait-associated SNPs are more likely to be eQTLs: annotation to enhance discovery from GWAS.* PLoS Genet, 2010. **6**(4): p. e1000888.
138. Frank, R. and R. Hargreaves, *Clinical biomarkers in drug discovery and development.* Nat Rev Drug Discov, 2003. **2**(7): p. 566-80.
139. Lequerre, T., et al., *A new tool for rheumatology: large-scale analysis of gene expression.* Joint Bone Spine, 2003. **70**(4): p. 248-56.
140. Batliwalla, F.M., et al., *Peripheral blood gene expression profiling in rheumatoid arthritis.* Genes Immun, 2005. **6**(5): p. 388-97.
141. Lyons, P.A., et al., *Microarray analysis of human leucocyte subsets: the advantages of positive selection and rapid purification.* BMC Genomics, 2007. **8**: p. 64.
142. Toonen, E.J., et al., *Gene expression profiling in rheumatoid arthritis: current concepts and future directions.* Ann Rheum Dis, 2008. **67**(12): p. 1663-9.
143. Glas, A.M., et al., *Converting a breast cancer microarray signature into a high-throughput diagnostic test.* BMC Genomics, 2006. **7**: p. 278.
144. Sorlie, T., et al., *Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications.* Proc Natl Acad Sci U S A, 2001. **98**(19): p. 10869-74.
145. Lubbers, J., et al., *The type I IFN signature as a biomarker of preclinical rheumatoid arthritis.* Ann Rheum Dis, 2013. **72**(5): p. 776-80.

146. van Baarsen, L.G., et al., *Gene expression profiling in autoantibody-positive patients with arthralgia predicts development of arthritis.* Arthritis Rheum, 2010. **62**(3): p. 694-704.
147. Cooles, F.A.H., et al., *The interferon gene signature is increased in patients with early treatment-naive rheumatoid arthritis and predicts a poorer response to initial therapy.* J Allergy Clin Immunol, 2018. **141**(1): p. 445-448 e4.
148. McKinney, E.F., et al., *A CD8+ T cell transcription signature predicts prognosis in autoimmune disease.* Nat Med, 2010. **16**(5): p. 586-91, 1p following 591.
149. Szodoray, P., et al., *A genome-scale assessment of peripheral blood B-cell molecular homeostasis in patients with rheumatoid arthritis.* Rheumatology (Oxford), 2006. **45**(12): p. 1466-76.
150. Imgenberg-Kreuz, J., et al., *Transcription profiling of peripheral B cells in antibody-positive primary Sjogren's syndrome reveals upregulated expression of CX3CR1 and a type I and type II interferon signature.* Scand J Immunol, 2018. **87**(5): p. e12662.
151. Sagoo, P., et al., *Development of a cross-platform biomarker signature to detect renal transplant tolerance in humans.* J Clin Invest, 2010. **120**(6): p. 1848-61.
152. Baron, D., et al., *A common gene signature across multiple studies relate biomarkers and functional regulation in tolerance to renal allograft.* Kidney Int, 2015. **87**(5): p. 984-95.
153. Biotech, M. *Magnetic cell separation*. 2019; Available from: www.miltenyibiotech.com.
154. Qiagen. *AllPrep DNA/RNA mini handbook*. 2005 [cited 2019; Available from: www.qiagen.com.
155. Illumina. *Illumina direct hybridisation assay overview*. 2010 [cited 2010; Available from: www.illumina.com.
156. Gentleman, R.C., et al., *Bioconductor: open software development for computational biology and bioinformatics.* Genome Biol, 2004. **5**(10): p. R80.
157. Ritchie, M.E., et al., *limma powers differential expression analyses for RNA-sequencing and microarray studies.* Nucleic Acids Res, 2015. **43**(7): p. e47.
158. Shi, W., A. Oshlack, and G.K. Smyth, *Optimizing the noise versus bias trade-off for Illumina whole genome expression BeadChips.* Nucleic Acids Res, 2010. **38**(22): p. e204.
159. Kramer, A., et al., *Causal analysis approaches in Ingenuity Pathway Analysis.* Bioinformatics, 2014. **30**(4): p. 523-30.
160. Subramanian, A., et al., *Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles.* Proc Natl Acad Sci U S A, 2005. **102**(43): p. 15545-50.
161. Mootha, V.K., et al., *PGC-1alpha-responsive genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes.* Nat Genet, 2003. **34**(3): p. 267-73.
162. Liberzon, A., et al., *The Molecular Signatures Database (MSigDB) hallmark gene set collection.* Cell Syst, 2015. **1**(6): p. 417-425.
163. Teixeira, V.H., et al., *Transcriptome analysis describing new immunity and defense genes in peripheral blood mononuclear cells of rheumatoid arthritis patients.* PLoS One, 2009. **4**(8): p. e6803.

164. Burska, A.N., et al., *Gene expression analysis in RA: towards personalized medicine.* Pharmacogenomics J, 2014. **14**(2): p. 93-106.
165. Lyons, P.A., et al., *Novel expression signatures identified by transcriptional analysis of separated leucocyte subsets in systemic lupus erythematosus and vasculitis.* Ann Rheum Dis, 2010. **69**(6): p. 1208-13.
166. Slodkowska, E.A. and J.S. Ross, *MammaPrint 70-gene signature: another milestone in personalized medical care for breast cancer patients.* Expert Rev Mol Diagn, 2009. **9**(5): p. 417-22.
167. Boutros, P.C., *The path to routine use of genomic biomarkers in the cancer clinic.* Genome Res, 2015. **25**(10): p. 1508-13.
168. Lee, J.C., et al., *Gene expression profiling of CD8+ T cells predicts prognosis in patients with Crohn disease and ulcerative colitis.* J Clin Invest, 2011. **121**(10): p. 4170-9.
169. Lequerre, T., et al., *Early and long-standing rheumatoid arthritis: distinct molecular signatures identified by gene-expression profiling in synovia.* Arthritis Res Ther, 2009. **11**(3): p. R99.
170. Jaksik, R., et al., *Microarray experiments and factors which affect their reliability.* Biol Direct, 2015. **10**: p. 46.
171. Consortium, M., et al., *The MicroArray Quality Control (MAQC) project shows inter- and intraplatform reproducibility of gene expression measurements.* Nat Biotechnol, 2006. **24**(9): p. 1151-61.
172. Webb, P.M., et al., *Microarrays and epidemiology: not the beginning of the end but the end of the beginning.* Cancer Epidemiol Biomarkers Prev, 2007. **16**(4): p. 637-8.
173. NCBI. *NCBI gene webpage*. 2019 [cited 2019; Available from: https://www.ncbi.nlm.nih.gov/gene/
.
174. Wherry, E.J. and M. Kurachi, *Molecular and cellular insights into T cell exhaustion.* Nat Rev Immunol, 2015. **15**(8): p. 486-99.
175. McKinney, E.F., et al., *T-cell exhaustion, co-stimulation and clinical outcome in autoimmunity and infection.* Nature, 2015. **523**(7562): p. 612-6.
176. Cole, S., et al., *Integrative analysis reveals CD38 as a therapeutic target for plasma cell-rich pre-disease and established rheumatoid arthritis and systemic lupus erythematosus.* Arthritis Res Ther, 2018. **20**(1): p. 85.
177. Barr, T.A., et al., *B cell depletion therapy ameliorates autoimmune disease through ablation of IL-6-producing B cells.* J Exp Med, 2012. **209**(5): p. 1001-10.
178. Tengstrand, B., et al., *Abnormal levels of serum dehydroepiandrosterone, estrone, and estradiol in men with rheumatoid arthritis: high correlation between serum estradiol and current degree of inflammation.* J Rheumatol, 2003. **30**(11): p. 2338-43.
179. Grimaldi, C.M., V. Jeganathan, and B. Diamond, *Hormonal regulation of B cell development: 17 beta-estradiol impairs negative selection of high-affinity DNA-reactive B cells at more than one developmental checkpoint.* J Immunol, 2006. **176**(5): p. 2703-10.
180. Sharma, S., et al., *Widely divergent transcriptional patterns between SLE patients of different ancestral backgrounds in sorted immune cell populations.* J Autoimmun, 2015. **60**: p. 51-58.

181. Carr, E.J., et al., *The cellular composition of the human immune system is shaped by age and cohabitation.* Nat Immunol, 2016. **17**(4): p. 461-468.
182. Brodin, P., et al., *Variation in the human immune system is largely driven by non-heritable influences.* Cell, 2015. **160**(1-2): p. 37-47.
183. Kraan, M.C., et al., *Immunohistological analysis of synovial tissue for differential diagnosis in early arthritis.* Rheumatology (Oxford), 1999. **38**(11): p. 1074-80.
184. Westra, H.J. and L. Franke, *From genome to function by studying eQTLs.* Biochim Biophys Acta, 2014. **1842**(10): p. 1896-1902.
185. Peters, M.J., et al., *The transcriptional landscape of age in human peripheral blood.* Nat Commun, 2015. **6**: p. 8570.
186. Huan, T., et al., *A whole-blood transcriptome meta-analysis identifies gene expression signatures of cigarette smoking.* Hum Mol Genet, 2016. **25**(21): p. 4611-4623.
187. Reynolds, L.M., et al., *Transcriptomic profiles of aging in purified human immune cells.* BMC Genomics, 2015. **16**: p. 333.
188. Vinuela, A., et al., *Age-dependent changes in mean and variance of gene expression across tissues in a twin cohort.* Hum Mol Genet, 2018. **27**(4): p. 732-741.
189. Teschendorff, A.E., J. West, and S. Beck, *Age-associated epigenetic drift: implications, and a case of epigenetic thrift?* Hum Mol Genet, 2013. **22**(R1): p. R7-R15.
190. Prelog, M., *Aging of the immune system: a risk factor for autoimmunity?* Autoimmun Rev, 2006. **5**(2): p. 136-9.
191. Fulop, T., et al., *Immunosenescence and Inflamm-Aging As Two Sides of the Same Coin: Friends or Foes?* Front Immunol, 2017. **8**: p. 1960.
192. Nikolich-Zugich, J., *The twilight of immunity: emerging concepts in aging of the immune system.* Nat Immunol, 2018. **19**(1): p. 10-19.
193. Kogut, I., et al., *B cell maintenance and function in aging.* Semin Immunol, 2012. **24**(5): p. 342-9.
194. Rubtsov, A.V., et al., *Toll-like receptor 7 (TLR7)-driven accumulation of a novel CD11c(+) B-cell population is important for the development of autoimmunity.* Blood, 2011. **118**(5): p. 1305-15.
195. Sasaki, S., et al., *Limited efficacy of inactivated influenza vaccine in elderly individuals is associated with decreased production of vaccine-specific antibodies.* J Clin Invest, 2011. **121**(8): p. 3109-19.
196. Frasca, D., et al., *The generation of memory B cells is maintained, but the antibody response is not, in the elderly after repeated influenza immunizations.* Vaccine, 2016. **34**(25): p. 2834-40.
197. de Magalhaes, J.P., J. Curado, and G.M. Church, *Meta-analysis of age-related gene expression profiles identifies common signatures of aging.* Bioinformatics, 2009. **25**(7): p. 875-81.
198. Bektas, A., et al., *Age-associated changes in basal NF-kappaB function in human CD4+ T lymphocytes via dysregulation of PI3 kinase.* Aging (Albany NY), 2014. **6**(11): p. 957-74.
199. Broer, L., et al., *GWAS of longevity in CHARGE consortium confirms APOE and FOXO3 candidacy.* J Gerontol A Biol Sci Med Sci, 2015. **70**(1): p. 110-8.

200. Mahley, R.W., *Apolipoprotein E: from cardiovascular disease to neurodegenerative disorders.* J Mol Med (Berl), 2016. **94**(7): p. 739-46.
201. Lo Sasso, G., et al., *The Apoe(-/-) mouse model: a suitable model to study cardiovascular and respiratory diseases in the context of cigarette smoke exposure and harm reduction.* J Transl Med, 2016. **14**(1): p. 146.
202. Sage, A.P. and Z. Mallat, *Readapting the adaptive immune response - therapeutic strategies for atherosclerosis.* Br J Pharmacol, 2017. **174**(22): p. 3926-3939.
203. Tsiantoulas, D., et al., *B cells and humoral immunity in atherosclerosis.* Circ Res, 2014. **114**(11): p. 1743-56.
204. Caligiuri, G., et al., *Protective immunity against atherosclerosis carried by B cells of hypercholesterolemic mice.* J Clin Invest, 2002. **109**(6): p. 745-53.
205. Major, A.S., S. Fazio, and M.F. Linton, *B-lymphocyte deficiency increases atherosclerosis in LDL receptor-null mice.* Arterioscler Thromb Vasc Biol, 2002. **22**(11): p. 1892-8.
206. Huan, T., et al., *A systems biology framework identifies molecular underpinnings of coronary heart disease.* Arterioscler Thromb Vasc Biol, 2013. **33**(6): p. 1427-34.
207. Yang, X., et al., *Oxidative Stress-Mediated Atherosclerosis: Mechanisms and Therapies.* Front Physiol, 2017. **8**: p. 600.
208. Marrocco, I., F. Altieri, and I. Peluso, *Measurement and Clinical Significance of Biomarkers of Oxidative Stress in Humans.* Oxid Med Cell Longev, 2017. **2017**: p. 6501046.
209. Cardoso, A.L., et al., *Towards frailty biomarkers: Candidates from genes and pathways regulated in aging and age-related diseases.* Ageing Res Rev, 2018. **47**: p. 214-277.
210. Griffin, M., R. Casadio, and C.M. Bergamini, *Transglutaminases: nature's biological glues.* Biochem J, 2002. **368**(Pt 2): p. 377-96.
211. Elli, L., et al., *Transglutaminases in inflammation and fibrosis of the gastrointestinal tract and the liver.* Dig Liver Dis, 2009. **41**(8): p. 541-50.
212. Lindfors, K., K. Kaukinen, and M. Maki, *A role for anti-transglutaminase 2 autoantibodies in the pathogenesis of coeliac disease?* Amino Acids, 2009. **36**(4): p. 685-91.
213. Berglund, L.J., et al., *IL-21 signalling via STAT3 primes human naive B cells to respond to IL-2 to enhance their differentiation into plasmablasts.* Blood, 2013. **122**(24): p. 3940-50.
214. Boyman, O. and J. Sprent, *The role of interleukin-2 during homeostasis and activation of the immune system.* Nat Rev Immunol, 2012. **12**(3): p. 180-90.
215. Frasca, D., et al., *High TNF-alpha levels in resting B cells negatively correlate with their response.* Exp Gerontol, 2014. **54**: p. 116-22.
216. van der Geest, K.S., et al., *Aging-dependent decline of IL-10 producing B cells coincides with production of antinuclear antibodies but not rheumatoid factors.* Exp Gerontol, 2016. **75**: p. 24-9.
217. Franceschi, C., et al., *Inflamm-aging. An evolutionary perspective on immunosenescence.* Ann N Y Acad Sci, 2000. **908**: p. 244-54.
218. Tzachanis, D., et al., *Tob is a negative regulator of activation that is expressed in anergic and quiescent T cells.* Nat Immunol, 2001. **2**(12): p. 1174-82.

219. Didonna, A., et al., *Immune cell-specific transcriptional profiling highlights distinct molecular pathways controlled by Tob1 upon experimental autoimmune encephalomyelitis.* Sci Rep, 2016. **6**: p. 31603.
220. Nicolini, A., P. Ferrari, and M.J. Duffy, *Prognostic and predictive biomarkers in breast cancer: Past, present and future.* Semin Cancer Biol, 2018. **52**(Pt 1): p. 56-73.
221. Shaffer, A.L., et al., *XBP1, downstream of Blimp-1, expands the secretory apparatus and other organelles, and increases protein synthesis in plasma cell differentiation.* Immunity, 2004. **21**(1): p. 81-93.
222. Recaldin, T. and D.J. Fear, *Transcription factors regulating B cell fate in the germinal centre.* Clin Exp Immunol, 2016. **183**(1): p. 65-75.
223. Kaser, A., et al., *XBP1 links ER stress to intestinal inflammation and confers genetic risk for human inflammatory bowel disease.* Cell, 2008. **134**(5): p. 743-56.
224. Harris, R.J., et al., *Granulocyte-macrophage colony-stimulating factor as an autocrine survival factor for mature normal and malignant B lymphocytes.* J Immunol, 2000. **164**(7): p. 3887-93.
225. Bhattacharya, P., et al., *GM-CSF: An immune modulatory cytokine that can suppress autoimmunity.* Cytokine, 2015. **75**(2): p. 261-71.
226. Ward, L.D. and M. Kellis, *Interpreting noncoding genetic variation in complex traits and human disease.* Nat Biotechnol, 2012. **30**(11): p. 1095-106.
227. Stranger, B.E., et al., *Population genomics of human gene expression.* Nat Genet, 2007. **39**(10): p. 1217-24.
228. Smith, E.N. and L. Kruglyak, *Gene-environment interaction in yeast gene expression.* PLoS Biol, 2008. **6**(4): p. e83.
229. Gagneur, J., et al., *Genotype-environment interactions reveal causal pathways that mediate genetic effects on phenotype.* PLoS Genet, 2013. **9**(9): p. e1003803.
230. Yao, C., et al., *Sex- and age-interacting eQTLs in human complex diseases.* Hum Mol Genet, 2014. **23**(7): p. 1947-56.
231. Maranville, J.C., et al., *Interactions between glucocorticoid treatment and cis-regulatory polymorphisms contribute to cellular response phenotypes.* PLoS Genet, 2011. **7**(7): p. e1002162.
232. Glastonbury, C.A., et al., *Adiposity-Dependent Regulatory Effects on Multi-tissue Transcriptomes.* Am J Hum Genet, 2016. **99**(3): p. 567-579.
233. Tanno, H., et al., *The Ankrd 13 family of UIM-bearing proteins regulates EGF receptor endocytosis from the plasma membrane.* Mol Biol Cell, 2012. **23**(7): p. 1343-53.
234. Parent, A., et al., *ANKRD13C acts as a molecular chaperone for G protein-coupled receptors.* J Biol Chem, 2010. **285**(52): p. 40838-51.
235. Pan, D., et al., *A major chromatin regulator determines resistance of tumor cells to T cell-mediated killing.* Science, 2018. **359**(6377): p. 770-775.
236. Hait, N.C. and A. Maiti, *The Role of Sphingosine-1-Phosphate and Ceramide-1-Phosphate in Inflammation and Cancer.* Mediators Inflamm, 2017. **2017**: p. 4806541.
237. Thompson, B.J., et al., *DYRK1A controls the transition from proliferation to quiescence during lymphoid development by destabilizing Cyclin D3.* J Exp Med, 2015. **212**(6): p. 953-70.

238. Morrison, A.J., *Genome maintenance functions of the INO80 chromatin remodeller.* Philos Trans R Soc Lond B Biol Sci, 2017. **372**(1731).
239. Dumay-Odelot, H., et al., *Cell growth- and differentiation-dependent regulation of RNA polymerase III transcription.* Cell Cycle, 2010. **9**(18): p. 3687-99.
240. Luo, X., W. Yang, and G. Gao, *SUN1 Regulates HIV-1 Nuclear Import in a Manner Dependent on the Interaction between the Viral Capsid and Cellular Cyclophilin A.* J Virol, 2018. **92**(13).
241. Bates, E.E., et al., *APCs express DCIR, a novel C-type lectin surface receptor containing an immunoreceptor tyrosine-based inhibitory motif.* J Immunol, 1999. **163**(4): p. 1973-83.
242. Guo, J., et al., *Correction: A Replication Study Confirms the Association of Dendritic Cell Immunoreceptor (DCIR) Polymorphisms with ACPA - Negative RA in a Large Asian Cohort.* PLoS One, 2012. **7**(10).
243. Chiffoleau, E., *C-Type Lectin-Like Receptors As Emerging Orchestrators of Sterile Inflammation Represent Potential Therapeutic Targets.* Front Immunol, 2018. **9**: p. 227.
244. Eklow, C., et al., *Cellular distribution of the C-type II lectin dendritic cell immunoreceptor (DCIR) and its expression in the rheumatic joint: identification of a subpopulation of DCIR+ T cells.* Ann Rheum Dis, 2008. **67**(12): p. 1742-9.
245. Fujikado, N., et al., *Dcir deficiency causes development of autoimmune diseases in mice due to excess expansion of dendritic cells.* Nat Med, 2008. **14**(2): p. 176-80.
246. Kanazawa, N., et al., *DCIR acts as an inhibitory receptor depending on its immunoreceptor tyrosine-based inhibitory motif.* J Invest Dermatol, 2002. **118**(2): p. 261-6.
247. Symmons, D.P., et al., *Lymphopenia in rheumatoid arthritis.* J R Soc Med, 1989. **82**(8): p. 462-3.
248. Duquenne, C., et al., *Lymphopenia in early arthritis: Impact on diagnosis and 3-year outcomes (ESPOIR cohort).* Joint Bone Spine, 2015. **82**(6): p. 417-22.
249. Moura, R.A., et al., *Alterations on peripheral blood B-cell subpopulations in very early arthritis patients.* Rheumatology (Oxford), 2010. **49**(6): p. 1082-92.
250. Fedele, A.L., et al., *Memory B cell subsets and plasmablasts are lower in early than in long-standing rheumatoid arthritis.* BMC Immunol, 2014. **15**: p. 28.
251. McComish, J., et al., *Changes in peripheral blood B cell subsets at diagnosis and after treatment with disease-modifying anti-rheumatic drugs in patients with rheumatoid arthritis: correlation with clinical and laboratory parameters.* Int J Rheum Dis, 2015. **18**(4): p. 421-32.
252. Souto-Carneiro, M.M., et al., *Alterations in peripheral blood memory B cells in patients with active rheumatoid arthritis are dependent on the action of tumour necrosis factor.* Arthritis Res Ther, 2009. **11**(3): p. R84.
253. Duty, J.A., et al., *Functional anergy in a subpopulation of naive B cells from healthy humans that express autoreactive immunoglobulin receptors.* J Exp Med, 2009. **206**(1): p. 139-51.
254. Rosser, E.C. and C. Mauri, *Regulatory B cells: origin, phenotype, and function.* Immunity, 2015. **42**(4): p. 607-12.

255. Mauri, C. and P.A. Blair, *Regulatory B cells: Are we really ready to manipulate them for the benefit of patients with autoimmune diseases?* Arthritis Rheumatol, 2014.
256. Matsumoto, M., et al., *Interleukin-10-producing plasmablasts exert regulatory function in autoimmune inflammation.* Immunity, 2014. **41**(6): p. 1040-51.
257. Ueda, Y., et al., *Inflammation controls B lymphopoiesis by regulating chemokine CXCL12 expression.* J Exp Med, 2004. **199**(1): p. 47-58.
258. Cain, D., et al., *Effects of acute and chronic inflammation on B-cell development and differentiation.* J Invest Dermatol, 2009. **129**(2): p. 266-77.
259. Leonardi, G.C., et al., *Ageing: from inflammation to cancer.* Immun Ageing, 2018. **15**: p. 1.
260. Wagner, A., et al., *Age-related differences in humoral and cellular immune responses after primary immunisation: indications for stratified vaccination schedules.* Sci Rep, 2018. **8**(1): p. 9825.
261. Siegrist, C.A. and R. Aspinall, *B-cell responses to vaccination at the extremes of age.* Nat Rev Immunol, 2009. **9**(3): p. 185-94.
262. Paganelli, R., et al., *Changes in circulating B cells and immunoglobulin classes and subclasses in a healthy aged population.* Clin Exp Immunol, 1992. **90**(2): p. 351-4.
263. Caraux, A., et al., *Circulating human B and plasma cells. Age-associated changes in counts and detailed characterization of circulating normal CD138- and CD138+ plasma cells.* Haematologica, 2010. **95**(6): p. 1016-20.
264. Simpfendorfer, K.R., et al., *Autoimmune disease-associated haplotypes of BLK exhibit lowered thresholds for B cell activation and expansion of Ig class-switched B cells.* Arthritis Rheumatol, 2015. **67**(11): p. 2866-76.
265. Simpfendorfer, K.R., et al., *The autoimmunity-associated BLK haplotype exhibits cis-regulatory effects on mRNA and protein expression that are prominently observed in B cells early in development.* Hum Mol Genet, 2012. **21**(17): p. 3918-25.
266. Kochi, Y., et al., *FCRL3, an autoimmune susceptibility gene, has inhibitory potential on B-cell receptor-mediated signaling.* J Immunol, 2009. **183**(9): p. 5502-10.
267. Kallberg, H., et al., *Smoking is a major preventable risk factor for rheumatoid arthritis: estimations of risks after various exposures to cigarette smoke.* Ann Rheum Dis, 2011. **70**(3): p. 508-11.
268. Hajishengallis, G., *Periodontitis: from microbial immune subversion to systemic inflammation.* Nat Rev Immunol, 2015. **15**(1): p. 30-44.
269. Makrygiannakis, D., et al., *Smoking increases peptidylarginine deiminase 2 enzyme expression in human lungs and increases citrullination in BAL cells.* Ann Rheum Dis, 2008. **67**(10): p. 1488-92.
270. Sokolove, J., et al., *Impact of baseline anti-cyclic citrullinated peptide-2 antibody concentration on efficacy outcomes following treatment with subcutaneous abatacept or adalimumab: 2-year results from the AMPLE trial.* Ann Rheum Dis, 2016. **75**(4): p. 709-14.

Appendix A Additional information and data

Appendix A.1
Diagnostic categories for early arthritis patients

Diagnosis
Ankylosing spondylitis
Crystal Arthritis
Enteropathic arthritis
Lupus/other connective tissue disease associated condition
Non-inflammatory
Osteoarthritis
Other inflammatory arthritis
Psoriatic arthritis
Reactive arthritis
Rheumatoid arthritis
Undifferentiated arthritis
Undifferentiated Spondylo-Arthropathy

Appendix A.2
List of 279 differentially expressed genes between RA (n=59) and non-RA (n=131) samples without the inclusion of clinical covariates. FC ≥ 1.2. No multiple test correction in place.

Gene	**logFC**	**P.Value**
KIAA1370	-0.267571393	4.18E-05
LPIN2	-0.311827532	4.69E-05
LMNA	0.408713569	6.74E-05
KCNK12	0.405818443	7.67E-05
NRSN2	0.438024293	7.98E-05
VPS13A	-0.307437162	9.44E-05
UST	-0.342338183	9.57E-05
ZSCAN12	-0.288531476	1.04E-04
COG3	-0.401061837	1.10E-04
OXCT1	-0.324102491	1.29E-04
FUT7	0.37425993	1.34E-04
TOP1MT	-0.416663121	1.41E-04
PHLPP2	-0.29645826	1.86E-04
C10orf32-AS3MT	-0.270165949	2.52E-04
BRMS1L	-0.263534447	2.66E-04
ATF5	0.426787027	2.79E-04
HOPX	0.41510197	2.88E-04
LDLR	0.418943206	3.08E-04
NA	-0.411936432	3.32E-04
ARID2	-0.263656937	3.37E-04
LDOC1	-0.326112912	3.43E-04
SLC25A26	-0.363812764	3.65E-04

PFKFB2	-0.298716265	3.73E-04
MAPKAP1	-0.273052321	4.23E-04
YPEL2	-0.385042708	4.48E-04
DHRS9	0.624260078	4.75E-04
ZNF266	-0.283909696	4.80E-04
DYRK1A	-0.369877192	5.00E-04
SP140L	-0.267986355	5.14E-04
C17orf28	0.268506862	5.50E-04
EPB41L2	-0.285344264	6.16E-04
NCALD	0.337974229	6.46E-04
TUBGCP6	-0.283418564	6.61E-04
HIPK2	0.354602844	7.07E-04
GALNT10	-0.339482494	7.12E-04
PKD1	-0.279983257	7.36E-04
ZC3H14	-0.38113736	7.44E-04
FYN	-0.432958678	7.58E-04
CAPRIN2	-0.286116206	7.60E-04
SLTM	-0.377630759	8.03E-04
PRICKLE2	0.268333721	8.42E-04
ACAP2	-0.392382601	8.78E-04
LMNA	0.277053298	9.30E-04
UHRF1BP1L	-0.274632543	9.49E-04
YPEL1	-0.277596856	9.50E-04
NA	0.289452583	9.84E-04
GAS6	0.533495363	9.96E-04
FAM129A	0.375970682	0.001009863
SFMBT1	-0.310386721	0.001028943
RPL32	-0.334316207	0.001038074
TOX2	0.383856725	0.001153085
TRAF5	-0.287316319	0.0011654
RNF146	-0.305824705	0.001197759
LUC7L3	-0.414264525	0.001244887
JF432672	-0.333798851	0.001256027
CEP110	-0.282069774	0.001263059
APP	-0.343602749	0.001267015
SF3B1	-0.350590604	0.001321057
MBNL1	-0.343866472	0.001341491
LGALS3	0.370173628	0.001359543
PROK2	0.51810562	0.001419745
PRKCE	-0.313413299	0.001425892
CCNDBP1	-0.331839038	0.001457201
PHC2	-0.299530143	0.001535749

CCNT2	-0.318150775	0.001564744
NA	0.277577079	0.001690798
ASB16	0.457623743	0.001693588
PHGDH	0.476526102	0.001697143
NUAK2	-0.359528086	0.001722575
AY665172	0.305835514	0.001725755
UBE2E1	-0.433410974	0.001731896
PPTC7	-0.268457427	0.001790939
TAF7	-0.300951928	0.001797217
ZXDB	-0.278270006	0.001842629
KIAA1407	-0.345086912	0.001856642
ZNF614	-0.310646421	0.00190025
NFYA	-0.272256198	0.002022462
SEC24B	-0.264825751	0.002030148
MEGF6	-0.32384602	0.002040967
IL16	-0.309589077	0.002093404
DYRK1A	-0.322983155	0.002100245
TP63	0.416327753	0.002102612
OPN3	-0.387005072	0.002117143
KCNMA1	0.266936113	0.002123021
TBXAS1	0.337630976	0.002276691
PABPC3	-0.280672793	0.002323378
ARHGAP30	-0.299272177	0.002329236
MTSS1	-0.30715592	0.00232995
ADD3	-0.36257321	0.002366625
SMG7	-0.368137644	0.002397627
KIAA1468	-0.265363825	0.00250618
CAMK1G	0.413465655	0.002609081
TSPYL2	-0.28842449	0.002624534
SF1	-0.358524112	0.002649359
CD44	-0.451996761	0.002688745
E2F2	0.367115002	0.002853388
GAS6	0.44383535	0.00288804
TRIM22	-0.317396777	0.002905957
NCALD	0.300352398	0.002959512
PPP1R8	-0.280277599	0.002977531
PRDM1	0.429708956	0.002983807
CNBP	-0.265108825	0.003080317
SNRPN	-0.293511609	0.003163953
ATP2C1	-0.324318891	0.003201715
DCK	-0.274071398	0.003338666
ZNF514	-0.277606654	0.003353544

SLC16A3	0.265967552	0.003409898
PYCR1	0.321118937	0.003448543
TTPAL	-0.297050522	0.003493799
TESC	0.353216848	0.003516599
CUL4B	-0.277058647	0.003562993
SMEK1	-0.316464459	0.003595884
CR603183	0.298055545	0.003605946
AFF4	-0.30551211	0.00361092
RAB27A	0.267449779	0.003669763
CDK2AP2	0.27122797	0.003746564
HERPUD2	-0.382906925	0.003876841
UHRF2	-0.30525153	0.003966884
FAM108B1	-0.34459209	0.003994262
SIL1	0.269789607	0.004058184
PARP11	-0.263686276	0.004138222
MBTPS1	-0.287036179	0.004229787
AK8	0.315910085	0.004351976
SMEK1	-0.273831406	0.004380249
MOBKL2B	-0.300191937	0.004389479
C6orf129	0.281374892	0.004476008
EIF4E3	0.278078967	0.004513509
NA	-0.362838423	0.004582759
MAD2L1BP	-0.270852613	0.004607428
MTMR4	-0.293611023	0.004608337
OGT	-0.45964243	0.004657041
C17orf60	-0.445388148	0.004725037
MBD2	-0.295916111	0.004752199
TOP1MT	-0.272590208	0.00478497
ZCCHC18	-0.297764565	0.004816676
CD79B	-0.438995166	0.004872579
ACTRT1	0.343774197	0.00488621
GMPPB	0.33138968	0.004936865
TULP4	-0.333832079	0.004981629
ABLIM1	-0.427964004	0.005013375
C14orf43	-0.265547117	0.005183822
NDRG3	-0.358402045	0.005379781
CDC42SE1	-0.37232774	0.005388972
SAV1	-0.302695749	0.005454739
NDFIP2	0.274201117	0.005486009
COCH	0.340648269	0.005622816
PELI1	-0.329872254	0.005689638
BEX4	-0.309836086	0.005717463

BCL10	-0.269169867	0.005877756
CAV1	0.40696977	0.005953496
ADARB1	-0.324948678	0.00600355
ZNF627	-0.268328629	0.006043477
RALB	-0.29980648	0.006106776
TOX2	0.322104601	0.006310076
HPS3	-0.36979558	0.006473078
WARS	0.319350808	0.006506097
GINS2	0.387817449	0.006524673
ACACB	-0.263758331	0.00652469
DHRS9	0.386043652	0.00655558
MMP25	0.264204703	0.006578416
CRELD2	0.280819854	0.00662965
RYK	-0.270086341	0.006670262
ZFP36L2	-0.29404644	0.006706525
KIAA1147	-0.291118718	0.006727608
CDC42SE1	-0.299344346	0.006920869
TMEM123	-0.325420424	0.006974758
KCNK6	0.298176061	0.007111542
HDAC9	-0.284119031	0.007183391
CXCR4	-0.664736919	0.007312107
AQP9	0.508026974	0.007468068
NFX1	-0.274860448	0.007536333
CDKN1A	0.279596988	0.00755832
NT5DC2	0.445768644	0.007689589
CD2AP	-0.27214581	0.00770417
HCST	0.329267501	0.007897203
FAM129A	0.374861979	0.008048692
SLAMF7	0.322736289	0.008249914
FKBP11	0.351230396	0.0082924
TEX9	0.268350413	0.008313085
IFI27L1	0.280133314	0.008372077
TSPAN3	-0.348376764	0.008547858
PHF19	0.264855287	0.008635661
AK124143	-0.354277186	0.008900254
DYSF	0.317030586	0.008920023
NR1D2	-0.264636183	0.009345983
WFS1	0.279577994	0.009353554
NUSAP1	0.335416004	0.009552984
SDF2L1	0.3031163	0.009595083
PACAP	0.280384395	0.00963179
SAR1B	0.285961619	0.009638728

MZB1	0.368168341	0.010005381
SPG21	-0.336650726	0.010382804
PPIA	-0.399410314	0.010443076
CKS2	0.28914983	0.010741883
CD74	-0.354928918	0.011056475
S100A8	0.390941478	0.011412522
MANEA	0.288613321	0.011466785
PRDM1	0.298605596	0.011667343
PDE4B	-0.326069092	0.012285586
BMF	-0.272098623	0.012385876
SPOPL	-0.27111883	0.012571239
ANG	0.278397458	0.012719236
CLEC2D	-0.359030234	0.012887358
BMP6	0.291303307	0.012896742
RTEL1	0.279972988	0.013153217
G3BP2	-0.324570735	0.014036273
ZC3HAV1	-0.268828671	0.014198536
ADM	0.311588254	0.014970142
C20orf103	0.368234189	0.015021243
NCF1	-0.550374565	0.015129988
JAK1	-0.44124232	0.015261227
FLJ44054	0.268078205	0.015485242
RPLP0	-0.313467202	0.015718988
ARHGAP32	-0.27635206	0.015761235
FBLN2	-0.297466191	0.015761839
PRNP	-0.279310049	0.01587394
UBE2L6	-0.312548751	0.016103493
PHKB	-0.27006308	0.016167636
TRIB1	0.291438789	0.016214602
ILF3	-0.29814908	0.016471418
PTGS2	0.311737755	0.016777557
IFI44	-0.346858778	0.016987937
GGH	0.283576687	0.017067914
PPIB	0.312277948	0.018184177
NFE2	0.33495086	0.018296617
CD79A	-0.433637903	0.018430501
CASP3	0.279273528	0.01850669
FAM150B	0.340424185	0.01851195
C19orf10	0.268539343	0.018728382
S100A9	0.370918738	0.019037914
AGPAT5	-0.299201961	0.020063827
MOCS2	-0.297492998	0.020181372

TP63	0.308892718	0.020733423
PPP2R2D	-0.280807015	0.021108297
CTBP1	-0.268931859	0.02172329
FBLN2	-0.28162848	0.021825372
FAM190B	-0.283844502	0.022182213
HIP1R	-0.280766468	0.022208474
BIRC5	0.306644083	0.022684418
GLTSCR2	-0.358068193	0.023233231
OIP5	0.298860922	0.024066381
WAPAL	-0.345863765	0.024212042
RCBTB2	0.279857962	0.024763259
GNG2	-0.290781474	0.024933762
CHPF	0.321750383	0.025436771
GMNN	0.280911071	0.026423956
KIAA0101	0.416704758	0.026518456
TK1	0.363988564	0.026895143
CNN3	-0.275463852	0.027153685
HMMR	0.298013475	0.027175616
RNF38	-0.266291489	0.028608371
CDHR3	-0.362545602	0.028697227
CDC45	0.35922027	0.029398907
S100P	0.32554663	0.029911084
BIK	0.298215008	0.030446531
TRIM5	-0.263351982	0.03046395
SLC7A5	0.316629427	0.03164915
DNAJC3	0.269728622	0.031724095
JSRP1	0.355369823	0.03175505
WARS	0.268013977	0.032201814
KIF11	0.284634836	0.032279259
PLBD1	0.397814244	0.032695304
MS4A1	-0.287344428	0.032992802
PLEKHF2	-0.306328133	0.033571886
RPL8	-0.274569609	0.03607261
PTPRC	-0.295363151	0.036349468
IRS2	-0.276452655	0.036744682
TNFRSF17	0.313143336	0.037861688
SLC38A11	-0.276588279	0.038420951
CCNA2	0.329123109	0.040012632
NA	-0.371739094	0.040359441
C10orf26	-0.265190838	0.040739764
ASPM	0.26569874	0.041295902
ATP1A1	-0.26402067	0.043418936

IFI44L	-0.427334264	0.043865956
NINJ1	0.269118676	0.046962642
S100A12	0.413866662	0.047832257
CNPY3	-0.309100683	0.048384294
ANKDD1A	0.309176919	0.049747599

Appendix A.3
List of 104 differentially expressed genes between RA (n=59) and non-RA (n=131) samples with the inclusion of clinical covariates. FC ≥ 1.2. No multiple test correction in place.

Gene	**logFC**	**P.Value**
VPS13A	-0.29071413	3.09E-04
LPIN2	-0.282011434	3.16E-04
ZSCAN12	-0.264443548	4.21E-04
FUT7	0.356859758	4.65E-04
NRSN2	0.397817043	4.88E-04
NCALD	0.358662012	5.07E-04
LMNA	0.361686018	5.29E-04
KCNK12	0.361988351	5.36E-04
KCNMA1	0.305561942	7.26E-04
LDOC1	-0.307860228	9.38E-04
HOPX	0.390652063	9.84E-04
UST	-0.288117222	0.001286436
ATF5	0.389288157	0.001312511
MFGE8	0.291740448	0.001373339
NCALD	0.336090599	0.001416622
COG3	-0.322249016	0.001697441
TP63	0.441758184	0.001743227
HIPK2	0.330166483	0.002087204
LDLR	0.350095686	0.002874723
NA	-0.34124883	0.003214183
LGALS3	0.348949208	0.003291926
COCH	0.376145756	0.003322116
SLC25A26	-0.300588624	0.003375081
FYN	-0.38416372	0.003681543
AK8	0.333323407	0.003830051
MBNL1	-0.316618622	0.003835272
SLTM	-0.332553724	0.004090454

AY665172	0.288320216	0.004523924
ARHGAP30	-0.28235242	0.004801986
TOP1MT	-0.3083869	0.004984279
FAM150B	0.418981647	0.005210305
DYRK1A	-0.283233936	0.00554628
FAM129A	0.326352353	0.005609511
DHRS9	0.501738808	0.005888927
TESC	0.336366396	0.006454009
SF3B1	-0.300014903	0.006848611
CAMK1G	0.385062592	0.006915888
IL16	-0.275275206	0.007418065
ZC3H14	-0.303202739	0.007435935
ACAP2	-0.307962353	0.007630342
CCNDBP1	-0.277235365	0.007742287
LUC7L3	-0.350872166	0.007797459
GALNT10	-0.264894459	0.008016595
NUAK2	-0.303923585	0.008411058
PROK2	0.415584097	0.008664974
YPEL2	-0.280932695	0.009065496
APP	-0.279020219	0.009203212
RPL32	-0.264468942	0.009630459
TP63	0.358232079	0.009893684
SMG7	-0.319124612	0.01013403
UBE2E1	-0.353601288	0.010519228
KIAA1407	-0.285131964	0.011843841
ASB16	0.374106879	0.012270173
PHGDH	0.374900108	0.012652155
GAS6	0.366610487	0.012843036
C17orf60	-0.405564806	0.013291164
CD79B	-0.386605814	0.013519818
CAV1	0.368176298	0.01369187
CD44	-0.365773181	0.014613128
ADD3	-0.292545285	0.015166529
PRDM1	0.354943175	0.01537095
ACTRT1	0.300361994	0.016331658
E2F2	0.297520474	0.016880264
SF1	-0.285008334	0.017048823
HERPUD2	-0.315186147	0.017357886
NDRG3	-0.309942977	0.01753697
OGT	-0.387737435	0.017920947
TOX2	0.282376083	0.018021839
CDC42SE1	-0.323591513	0.018668189

NT5DC2	0.392744191	0.019991323
OPN3	-0.281097563	0.020638161
MANEA	0.271436833	0.02129062
ADARB1	-0.264295977	0.025603875
TULP4	-0.263606189	0.025779472
TOX2	0.26959114	0.026151073
DHRS9	0.324993865	0.026212584
AK124143	-0.295728817	0.026625114
CXCR4	-0.542884027	0.026951417
IFI44	-0.330958579	0.027843957
CNN3	-0.279924311	0.028458381
RCBTB2	0.280247091	0.029278393
FAM129A	0.316050199	0.029424038
MAL	0.328056654	0.031230166
SLAMF7	0.263352066	0.031800866
TSPAN3	-0.283608531	0.032201382
FAM150B	0.324691019	0.032303484
DSTYK	-0.308112643	0.032470359
CLEC2D	-0.306758183	0.033587921
GAS6	0.289835414	0.034004545
PPIA	-0.335233536	0.034304791
NUSAP1	0.279288174	0.034351051
CD74	-0.298128044	0.034637962
TRIB1	0.26413929	0.03504892
NA	-0.267702211	0.035675384
ABLIM1	-0.309653016	0.038463629
RPLP0	-0.272418827	0.03860245
FKBP11	0.27249962	0.038822785
NCF1	-0.467724411	0.039901685
JSRP1	0.347883881	0.04283498
NA	0.275919995	0.042845062
HPS3	-0.267652193	0.046217295
MAL	0.265110766	0.047052501
CD79A	-0.366655964	0.047253377
CCR9	-0.30961145	0.048232377

Appendix A.4

List of 225 differentially expressed genes between RA (n=59) and non-inflammatory (n=72) samples without the inclusion of clinical covariates. FC ≥ 1.2. Benjamini-Hochberg multiple test correction in place

Gene	logFC	P.Value	adj.P.Val
LPIN2	-0.378744578	9.10E-06	0.024995124
OXCT1	-0.417111446	9.42E-06	0.024995124
C10orf32-AS3MT	-0.361917695	9.98E-06	0.024995124
TOP1MT	-0.535904553	1.07E-05	0.024995124
RCSD1	-0.30326813	1.14E-05	0.024995124
NA	0.414880461	1.97E-05	0.026208951
NINJ2	0.327347907	2.48E-05	0.026208951
TRAK1	-0.342936634	2.68E-05	0.026208951
C17orf28	0.361562973	2.79E-05	0.026208951
NRSN2	0.514068698	3.31E-05	0.026208951
VPS13A	-0.364246594	3.35E-05	0.026208951
GATAD1	-0.307157012	3.48E-05	0.026208951
COG3	-0.476747029	3.74E-05	0.026208951
RBM15B	-0.30451631	4.35E-05	0.026208951
OPA1	-0.333247555	4.50E-05	0.026208951
FAM108B1	-0.536865385	4.72E-05	0.026208951
BAG5	-0.266415472	4.87E-05	0.026208951
NDRG2	-0.290652963	5.44E-05	0.026208951
YPEL2	-0.489917519	6.03E-05	0.026208951
SLTM	-0.501749294	6.18E-05	0.026208951
TSPYL2	-0.422990307	6.69E-05	0.026208951
UHRF1BP1L	-0.367426101	6.93E-05	0.026208951
ING5	-0.324299995	7.02E-05	0.026208951
ACAP2	-0.520783843	7.14E-05	0.026208951
TSHZ1	-0.307923064	7.60E-05	0.026208951
CUL4B	-0.415247053	7.78E-05	0.026208951
DYRK1A	-0.463919873	8.92E-05	0.026208951
CDC42SE1	-0.47802667	9.07E-05	0.026208951
KIAA1370	-0.284197618	9.57E-05	0.026208951
UST	-0.382035196	9.58E-05	0.026208951
PHC2	-0.409434949	9.68E-05	0.026208951
ATP2C1	-0.474506371	9.82E-05	0.026208951
RBM41	-0.327560296	9.93E-05	0.026208951
NA	-0.497739841	1.01E-04	0.026208951
NUDT2	0.338631882	1.02E-04	0.026208951
SMEK1	-0.412652111	1.03E-04	0.026208951
MAPKAP1	-0.33423213	1.09E-04	0.026208951

PHLPP2	-0.342054928	1.11E-04	0.026208951
EPB41L2	-0.358404288	1.13E-04	0.026208951
LMNA	0.441543934	1.14E-04	0.026208951
BRMS1L	-0.311015198	1.15E-04	0.026208951
ZBTB5	-0.310755244	1.23E-04	0.027508064
HECTD1	-0.26695656	1.31E-04	0.02873462
AKAP10	-0.331871004	1.37E-04	0.029387435
DHRS9	0.758726136	1.40E-04	0.029495763
PPTC7	-0.362930698	1.46E-04	0.029864878
VKORC1	0.311129258	1.47E-04	0.029864878
AY665172	0.410410729	1.56E-04	0.031061531
KIAA0430	-0.272382147	1.78E-04	0.033164801
SLC16A2	0.276304534	1.78E-04	0.033164801
PRKCE	-0.409591386	1.81E-04	0.033164801
SYN1	0.281707297	1.83E-04	0.033164801
ZXDB	-0.371639198	1.84E-04	0.033164801
LDOC1	-0.378850135	1.93E-04	0.033461916
MRPS21	-0.324321504	1.98E-04	0.033461916
CAPRIN2	-0.35258493	1.98E-04	0.033461916
JF432672	-0.428868073	1.98E-04	0.033461916
BCL7C	-0.35905934	2.08E-04	0.034533968
RNF146	-0.389506698	2.13E-04	0.034846095
GAS6	0.667130547	2.22E-04	0.035042507
ARID2	-0.302320842	2.30E-04	0.035042507
SFMBT1	-0.388093078	2.31E-04	0.035042507
LPGAT1	-0.313432899	2.35E-04	0.035042507
ZC3H14	-0.463426524	2.36E-04	0.035042507
SLC25A26	-0.418495786	2.39E-04	0.035042507
LMNA	0.342691432	2.40E-04	0.035042507
MPP5	-0.287259335	2.49E-04	0.035269816
HOPX	0.467811662	2.51E-04	0.035269816
NFYA	-0.359015829	2.55E-04	0.035269816
TXNRD1	-0.292611737	2.61E-04	0.035269816
FYN	-0.522898026	2.66E-04	0.035269816
TMEM183A	-0.294357051	2.66E-04	0.035269816
ARHGAP12	-0.272534431	2.67E-04	0.035269816
C12orf49	-0.305120059	2.68E-04	0.035269816
RPL32	-0.413884326	2.71E-04	0.035269816
WRNIP1	-0.34736794	2.73E-04	0.035269816
SMEK1	-0.438989352	2.78E-04	0.035460959
SP140L	-0.312522079	2.84E-04	0.035529891
ADD3	-0.48050213	2.95E-04	0.035529891

CCDC50	-0.34112732	2.96E-04	0.035529891
GALNT10	-0.404960835	2.96E-04	0.035529891
ZDHHC17	-0.285144533	3.03E-04	0.035529891
ZSCAN12	-0.299433215	3.05E-04	0.035529891
CCNT2	-0.404158836	3.12E-04	0.035529891
CDC42SE1	-0.535032883	3.14E-04	0.035529891
ASB16	0.58392685	3.24E-04	0.035949786
PFKFB2	-0.33594738	3.37E-04	0.036466725
KIAA1468	-0.349881135	3.44E-04	0.036466725
DIS3	-0.288071358	3.51E-04	0.036466725
TBXAS1	0.439714722	3.59E-04	0.036466725
SEC24B	-0.340964446	3.62E-04	0.036466725
ZNF32	-0.393358672	3.62E-04	0.036466725
TUBGCP6	-0.330702576	3.71E-04	0.036846925
OGT	-0.641322538	3.80E-04	0.036846925
TMEM185B	-0.283331326	3.81E-04	0.036846925
NR3C1	-0.280735237	3.82E-04	0.036846925
KCNK12	0.405838629	3.87E-04	0.036846925
PTPRA	-0.267759927	3.89E-04	0.036846925
DAZAP1	-0.292421844	4.01E-04	0.036846925
ZNF266	-0.321230324	4.02E-04	0.036846925
BCL10	-0.383945171	4.02E-04	0.036846925
PRMT6	-0.275245902	4.03E-04	0.036846925
SF3B1	-0.430137146	4.12E-04	0.037055892
WDR26	-0.309412338	4.13E-04	0.037055892
CREBZF	-0.310495934	4.22E-04	0.037055892
PROK2	0.638527199	4.22E-04	0.037055892
LDLR	0.454519663	4.53E-04	0.038017183
MTSS1	-0.393420147	4.66E-04	0.038017183
TULP4	-0.462282415	4.69E-04	0.038017183
MPHOSPH8	-0.307215861	4.73E-04	0.038017183
ZC3HAV1	-0.42348811	4.77E-04	0.038017183
ZC4H2	-0.312916172	4.80E-04	0.038017183
SEPX1	0.31712172	4.84E-04	0.038017183
SLFN11	0.279112589	4.87E-04	0.038017183
TAF7	-0.374836641	4.88E-04	0.038017183
S100P	0.574425924	4.94E-04	0.038017183
HPS3	-0.525491514	4.95E-04	0.038017183
MAD2L1BP	-0.369699088	5.07E-04	0.038017183
GAS6	0.576811324	5.09E-04	0.038017183
FCRL2	-0.283047146	5.10E-04	0.038017183
CASD1	-0.336976716	5.15E-04	0.038017183

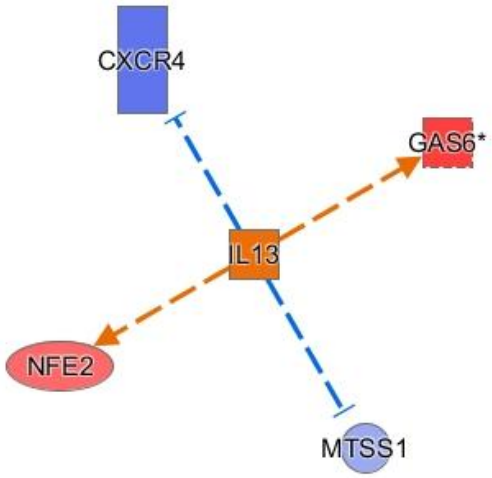
ZMYM5	-0.366073408	5.16E-04	0.038017183
ABLIM1	-0.588329994	5.22E-04	0.038017183
WDR44	-0.32601251	5.23E-04	0.038017183
NA	0.341858369	5.26E-04	0.038017183
FUT7	0.378809305	5.27E-04	0.038017183
TRIM24	-0.320506032	5.38E-04	0.038386196
CD79B	-0.600067752	5.41E-04	0.038386196
MAPKAP1	-0.369829455	5.42E-04	0.038386196
HERC3	-0.334198804	5.76E-04	0.040021247
URB2	-0.267614065	5.98E-04	0.04099923
RALB	-0.416282793	6.16E-04	0.042003274
WDR47	-0.316484035	6.25E-04	0.042090815
STXBP5	-0.265033358	6.45E-04	0.04240837
APP	-0.405917188	6.45E-04	0.04240837
SMAP1	-0.290646456	6.55E-04	0.042555468
CXCR4	-0.937614191	6.63E-04	0.04278259
ZNF614	-0.378983853	6.84E-04	0.043313477
AFF4	-0.397116463	6.85E-04	0.043313477
SF1	-0.451330991	6.90E-04	0.043313477
ARHGAP30	-0.371871659	6.93E-04	0.043313477
ZMYND8	-0.27561743	6.95E-04	0.043313477
S100A8	0.580989926	7.00E-04	0.043386972
MMP25	0.366303515	7.04E-04	0.043386972
SCARNA3	-0.263632959	7.08E-04	0.043401121
CSRNP2	-0.265759086	7.12E-04	0.043401121
ASB3	-0.285722983	7.17E-04	0.043445196
PPP2R2D	-0.455058494	7.21E-04	0.043445196
AMOT	-0.279747451	7.30E-04	0.043801864
TOMM40	-0.303004882	7.38E-04	0.044031635
MMD	-0.286784733	7.59E-04	0.044999101
ZMAT5	-0.298427971	7.69E-04	0.045191716
PJA1	-0.280585741	7.70E-04	0.045191716
CD44	-0.564312651	7.80E-04	0.04533697
CDKN2D	-0.371824181	7.99E-04	0.045606975
LGALS3	0.431506564	8.24E-04	0.045606975
YPEL1	-0.313466406	8.30E-04	0.045606975
MTMR4	-0.3858024	8.32E-04	0.045606975
AQP9	0.705625829	8.34E-04	0.045606975
SON	-0.27457028	8.43E-04	0.045606975
TRPV2	0.330040623	8.44E-04	0.045606975
FBXO21	-0.303216397	8.50E-04	0.045606975
NDRG3	-0.478132309	8.54E-04	0.045606975

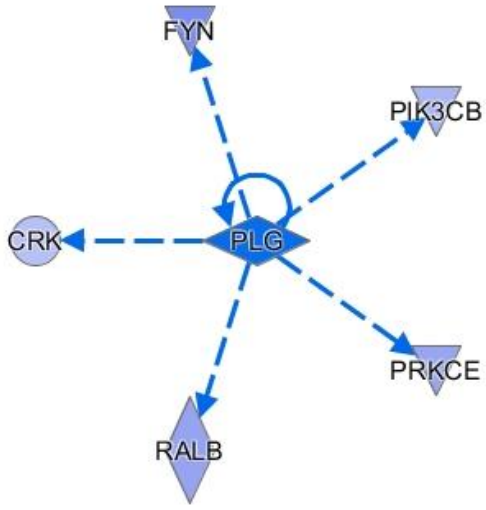
RRP1	0.284285758	8.56E-04	0.045606975
JF432672	-0.315380371	8.71E-04	0.045606975
DIDO1	-0.263542947	8.97E-04	0.045606975
DYRK1A	-0.388981878	9.03E-04	0.045606975
SLC29A2	-0.266839632	9.07E-04	0.045606975
RCL1	-0.274177445	9.07E-04	0.045606975
HIPK2	0.387764202	9.09E-04	0.045606975
SON	-0.343672396	9.14E-04	0.045606975
P2RY10	-0.275992587	9.24E-04	0.045606975
ANO6	-0.286755075	9.34E-04	0.045606975
EHBP1	-0.278782539	9.46E-04	0.045606975
PKD1	-0.306118807	9.46E-04	0.045606975
PHKB	-0.411214415	9.55E-04	0.045606975
SCAND1	-0.318774135	9.65E-04	0.045606975
FNBP4	-0.279803304	9.67E-04	0.045606975
SLC16A3	0.334167732	9.70E-04	0.045606975
NFE2	0.518633788	9.74E-04	0.045606975
ERF	-0.324861902	9.78E-04	0.045606975
C17orf60	-0.579123231	9.79E-04	0.045606975
BANF1	-0.346275149	9.89E-04	0.045606975
NEDD9	-0.291562389	9.93E-04	0.045606975
CRK	-0.282213279	0.001008228	0.045606975
REV3L	-0.299662734	0.001008663	0.045606975
GAB3	-0.281000343	0.001008942	0.045606975
SEC24B	-0.337704888	0.001024204	0.045606975
ZNF549	-0.275288278	0.001042904	0.045606975
UBE2E1	-0.506024018	0.001051699	0.045606975
UBE2L6	-0.471804254	0.001054717	0.045606975
SMG7	-0.442835084	0.001066713	0.045606975
ADARB1	-0.317836223	0.001067716	0.045606975
PARP11	-0.335337939	0.001073224	0.045606975
C12orf35	-0.272067712	0.001074967	0.045606975
KIAA1737	-0.282445863	0.001083341	0.045606975
HIAT1	-0.295463341	0.001083891	0.045606975
MAP3K9	-0.292298119	0.001087978	0.045606975
MAP4K4	-0.290448457	0.001091429	0.045606975
OPN3	-0.458736233	0.001100069	0.045606975
PPIA	-0.565806544	0.001101417	0.045606975
IRF2BP2	-0.412293135	0.001152608	0.046228448
SNRPN	-0.360549683	0.001156439	0.046228448
RALGPS2	-0.359539301	0.001161758	0.046228448
PDIK1L	-0.324268038	0.00116619	0.046228448

CAMK1G	0.497602505	0.001169313	0.046228448
FHOD1	0.299238328	0.001169323	0.046228448
AGXT2L2	-0.320971966	0.001170062	0.046228448
DIP2B	-0.270467184	0.001171194	0.046228448
MAML1	-0.283618738	0.001182312	0.046336595
ANG	0.401828073	0.001209339	0.047224479
ANKRD33	0.333365228	0.00124945	0.04798521
ZNF514	-0.340465578	0.001263193	0.04798521
FAM190B	-0.44312967	0.001268955	0.04798521
PIK3CB	-0.332760784	0.001292431	0.048181506
CHRAC1	-0.29038132	0.001293145	0.048181506
DALRD3	-0.292666875	0.001308591	0.048185136
DYSF	0.433150638	0.001327708	0.048562104
E2F2	0.440792435	0.001331353	0.048562104
PPFIA1	-0.26750883	0.001332277	0.048562104
ADARB1	-0.422849474	0.001336531	0.048562104
LUC7L3	-0.458364457	0.001378807	0.049768594
PRICKLE2	0.286884779	0.001387774	0.049773952
PLOD1	0.264586419	0.001388028	0.049773952
NUAK2	-0.409314186	0.001394775	0.049852988

Appendix A.5

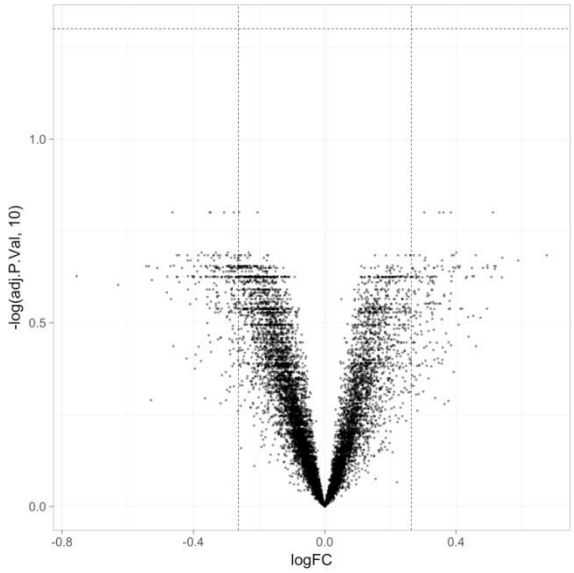
Upstream regulators identified for the list of differentially expressed genes between RA (n=59) and non-inflammatory (n=72) samples without the inclusion of clinical covariates. FC ≥ 1.2. Benjamini-Hochberg multiple test correction in place. IL-13 is predicted to be activated in the dataset. PLG is predicted to be inhibited



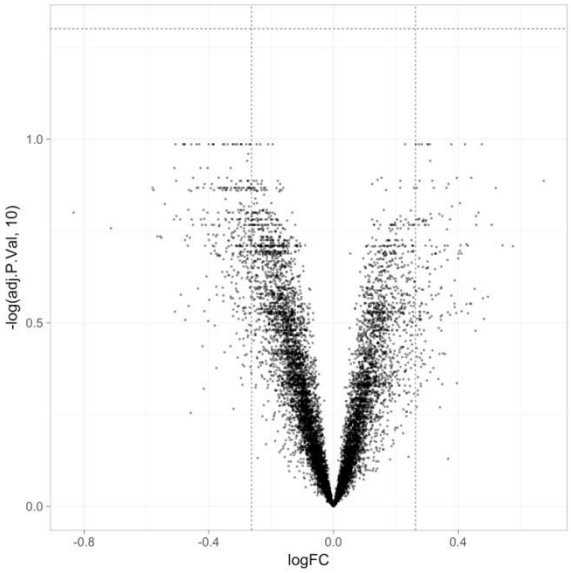


Appendix A.6

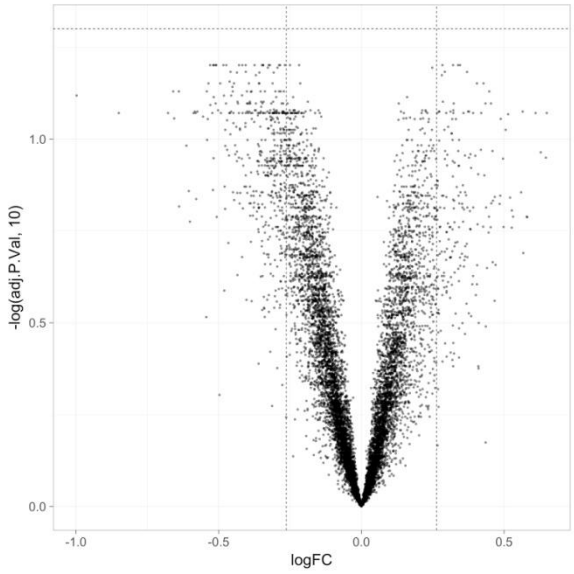
Volcano plots for Rheumatoid samples (n=59) versus non-inflammatory samples (n=72), FC1.2, MTC BH with each clinical variable added separately to the linear model. a) with age included, b) with CRP included, c) with ESR included. No DEGs were identified at a FC≥1.2 in any comparison. Vertical, dotted lines denote FC1.2. Horizontal, dotted lines denote p-value 0.05. Red dots indicate probes which are differentially expressed between the comparator groups.



a)



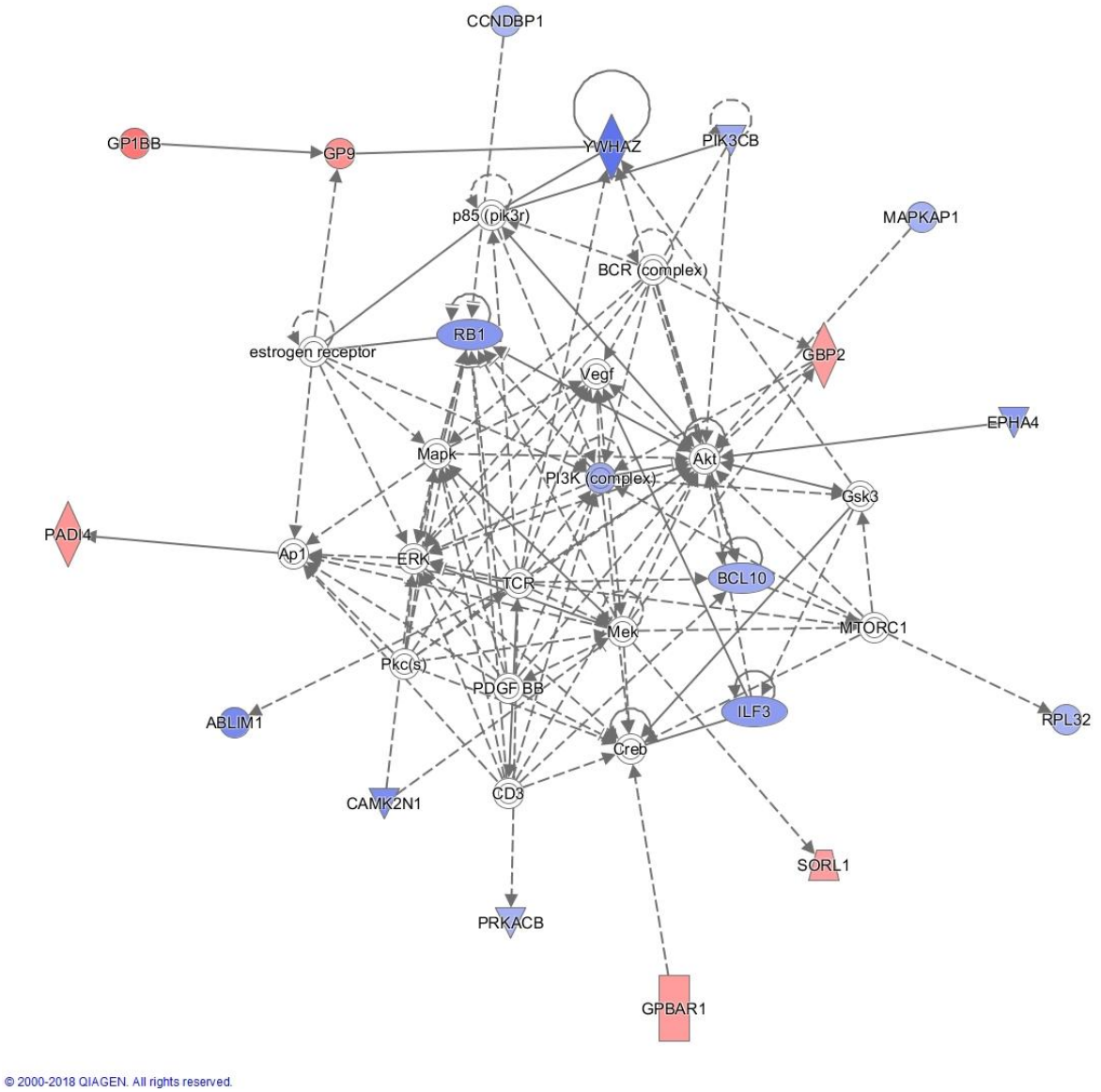
b)



c)

Appendix A.7

Network 2 identified for list of differentially expressed genes between the ≥55 years and <55 years age groups, no clinical covariates added, FC ≥1.2, Benjamini-Hochberg multiple test correction in place. The top diseases and functions for this netowrk are described as related to connective tissue disorders, developmental disorder and haematological diseases.



Appendix B Publications arising from work contained in this thesis

Thalayasingam N, Nair N, Skelton A, Massey J, Anderson AE, Clark A, Diboll J, Lendrem DW, Reynard LN, Cordell HJ, Eyre S, Isaacs JD, Barton A and Pratt AG
CD4+ and B lymphocyte expression quantitative traits at rheumatoid arthritis risk loci in untreated early arthritis: implications for causal gene identification.
Arthritis Rheumatol 2018. 70(3):361-370

McGovern A, Schoenfelder S, Martin P, Massey J, Duffus K, Plant D, Yarwood A, Pratt AG, Anderson AE, Isaacs JD, Diboll J, **Thalayasingam N**, Ospelt C, Barton A, Worthington J, Fraser P, Eyre S, Orozco G.
Capture Hi-C identifies a novel causal gene, IL20RA, in the pan-autoimmune genetic susceptibility region 6p23
Genome Biol. 2016 17(1):212

Published Abstracts:

Naamane N, **Thalayasingam N**, Nair N, Clark A, Anderson A, Lendrem D, Reynard L, Eyre S, Barton A, Isaacs JD, Pratt A
Performance of methylome over transcriptome data for discriminating rheumatoid arthritis patients in an early arthritis clinic: implications for translating “big data” into clinically useful tools
Ann Rheum Dis 2019

Clark AD, Nair N, Skelton AJ, Anderson AE, **Thalayasingam N**, Naamane N, Diboll J, Massey J, Eyre S, Barton A, Isaacs JD, Reynard LN, Pratt AG
DNA methylation in lymphocyte subsets as a mediator of genetic risk in early rheumatoid arthritis
Ann Rheum Dis 2018

Appendix C Publication: CD4+ and B lymphocyte expression quantitative traits at rheumatoid arthritis risk loci in untreated early arthritis: implications for causal gene identification

ARTHRITIS & RHEUMATOLOGY
Vol. 70, No. 3, March 2018, pp 361–370
DOI 10.1002/art.40393

CD4+ and B Lymphocyte Expression Quantitative Traits at Rheumatoid Arthritis Risk Loci in Patients With Untreated Early Arthritis

Implications for Causal Gene Identification

Nishanthi Thalayasingam,[1] Nisha Nair,[2] Andrew J. Skelton,[1] Jonathan Massey,[2] Amy E. Anderson,[1] Alexander D. Clark,[1] Julie Diboll,[1] Dennis W. Lendrem,[1] Louise N. Reynard,[1] Heather J. Cordell,[3] Stephen Eyre,[2] John D. Isaacs,[1] Anne Barton,[2] and Arthur G. Pratt[1]

Objective. **Rheumatoid arthritis (RA) is a genetically complex disease of immune dysregulation. This study sought to gain further insight into the genetic risk mechanisms of RA by conducting an expression quantitative trait locus (eQTL) analysis of confirmed genetic risk loci in CD4+ T cells and B cells from carefully phenotyped patients with early arthritis who were naive to therapeutic immunomodulation.**

Methods. **RNA and DNA were isolated from purified B and/or CD4+ T cells obtained from the peripheral blood of 344 patients with early arthritis. Genotyping and global gene expression measurements were carried out using Illumina BeadChip microarrays. Variants in linkage disequilibrium (LD) with non-HLA RA single-nucleotide polymorphisms (defined as $r^2 \geq 0.8$) were analyzed, seeking evidence of *cis*- or *trans*-eQTLs according to whether the associated probes were or were not within 4 Mb of these LD blocks.**

Results. **Genes subject to *cis*-eQTL effects that were common to both CD4+ and B lymphocytes at RA risk loci were *FADS1*, *FADS2*, *BLK*, *FCRL3*, *ORMDL3*, *PPIL3*, and *GSDMB*. In contrast, those acting on *METTL21B*, *JAZF1*, *IKZF3,* and *PADI4* were unique to CD4+ lymphocytes, with the latter candidate risk gene being identified for the first time in this cell subset. B lymphocyte–specific eQTLs for *SYNGR1* and *CD83* were also found. At the 8p23 *BLK–FAM167A* locus, adjacent genes were subject to eQTLs whose activity differed markedly between cell types; in particular, the *FAM167A* effect displayed striking B lymphocyte specificity. No *trans*-eQTLs approached experiment-wide significance, and linear modeling did not identify a significant influence of biologic covariates on *cis*-eQTL effect sizes.**

Conclusion. **These findings further refine the understanding of candidate causal genes in RA pathogenesis, thus providing an important platform from which downstream functional studies, directed toward particular cell types, may be prioritized.**

The views expressed herein are those of the authors and do not necessarily reflect those of the NHS, the NIHR, the Department of Health, or Pfizer.

Supported by the Academy of Medical Sciences, the JGW Patterson Foundation, Pfizer (unrestricted research grant), the NIHR (Newcastle Biomedical Research Centre at Newcastle Hospitals Foundation Trust and Newcastle University, and the Manchester Musculoskeletal Biomedical Research Unit), Arthritis Research UK (Centre of Excellence for RA Pathogenesis), the Wellcome Trust (clinical training fellowship to Dr. Thalayasingam), and the Medical Research Council (stratified medicine award MR/K015346/1 to Drs. Nair and Massey).

[1]Nishanthi Thalayasingam, BMCh, MA, Andrew J. Skelton, MSc, Amy E. Anderson, PhD, Alexander D. Clark, BSc, Julie Diboll, BSc, Dennis W. Lendrem, PhD, Louise N. Reynard, PhD, John D. Isaacs, MBBS, PhD, Arthur G. Pratt, MBChB, PhD: NIHR Newcastle Biomedical Research Centre, Newcastle upon Tyne Hospitals NHS Foundation Trust, and Newcastle University, Newcastle upon Tyne, UK; [2]Nisha Nair, PhD, Jonathan Massey, PhD, Stephen Eyre, PhD, Anne Barton, MBChB, PhD: Arthritis Research UK Centre for Genetics and Genomics, Centre for Musculoskeletal Research, Institute of Inflammation and Repair, University of Manchester, and NIHR Manchester Musculoskeletal Biomedical Research Unit, Central Manchester NHS Foundation Trust, Manchester, UK; [3]Heather J. Cordell, DPhil: Newcastle University, Newcastle upon Tyne, UK.

Drs. Thalayasingam and Nair contributed equally to this work. Professors Isaacs and Barton contributed equally to this work.

Address correspondence to Arthur G. Pratt, MBChB, PhD, Institute of Cellular Medicine (Musculoskeletal Research Group), Newcastle University, Newcastle upon Tyne NE2 4HH, UK. E-mail: arthur.pratt@ncl.ac.uk.

Submitted for publication July 2, 2017; accepted in revised form November 22, 2017.

Rheumatoid arthritis (RA) is a complex genetic disease in which immune tolerance becomes impaired, and an unchecked inflammatory response leads to chronic pain and damage to the synovial joints (1). Genetic variation at the *HLA–DRB1*, *HLA–DPB*, and *HLA–B* loci accounts for a large proportion of the known RA risk (2), with implications for antigen presentation to T lymphocytes (3,4). Outside of the HLA region, accumulating data now highlight an overlap between the 101 confirmed RA risk loci and cell-specific enhancer elements, which is maximal in CD4+ lymphocytes followed by B lymphocytes (5–8). Such molecular insights support a pivotal role for both CD4+ T cell and B cell lineages in the pathogenesis of RA (9–11). Mapping cellular mechanisms of genetic risk in the disease is far from straightforward, however, because lead single-nucleotide polymorphisms (SNPs) at associated loci are typically noncoding and intergenic, tagging linkage disequilibrium (LD) blocks that contain multiple genes (7,12).

To prioritize causal genes, one solution is to explore associations between genetic variants and downstream molecular quantitative traits, the most proximal of which is gene expression. Thus, with respect to a putative susceptibility gene, colocalization of an expression quantitative trait locus (eQTL) with a disease risk variant implicates the gene as a candidate for disease causation (13). Data from eQTL studies in healthy human subjects have indeed informed algorithms for prioritization of candidate genes in RA (7). Importantly, however, it is now clear that the transcriptional consequences of genetic variation can manifest as cell type specificity, with potentially profound implications for disease pathogenesis (14,15). For example, it has been observed that only 22% of *cis*-eQTLs are consistently identified in different circulating cell subsets from healthy donors; eQTLs present in a specific cell type may not be detectable in another cell type or in whole blood—and vice versa. Moreover, a number of eQTLs can be detected only under specific conditions of cell stimulation (14,16,17). This suggests that the contribution of eQTL data to inferred causality among candidate genes for a given disease must increasingly be understood at a cellular level and within a relevant biologic context (18).

The suggestion that the effect size of a risk variant's influence on gene expression may depend on the environmental parameters to which cells are exposed has potentially important implications for understanding the complexities of disease induction. In RA, for example, the unmasking of eQTL effects in relevant cell populations during a transient systemic trigger might plausibly be sufficient to break immune tolerance, permitting a transition to persistent joint inflammation. Against this backdrop, we set out to reassess the biologic landscape of candidate susceptibility genes in RA by mapping *cis*- and *trans*-eQTLs at 101 established RA risk loci in circulating CD4+ and B lymphocyte subsets sampled from a cohort of untreated patients with early arthritis. In so doing, we sought insight into potential common and/or cell-specific mechanisms of genetic risk in a highly relevant biologic context, free from the confounding influences of in vivo immune modulation or ex vivo manipulation.

PATIENTS AND METHODS

Patients. Patients with early arthritis (all of self-reported white ethnicity) who were attending the Newcastle Early Arthritis Cohort (NEAC) clinic in the UK were recruited into the study, and peripheral blood samples were obtained prior to the commencement of immunomodulatory therapy; individuals who were exposed to steroid treatment during the 3 months prior to recruitment or those whose ethnic origin, determined by genotype, was not of white Northern European descent were excluded from the analyses. This resulted in 71 patients being recruited between January 2008 and December 2009, and a further 273 during 2012 and 2013; the NEAC cohort has been described in detail elsewhere (19–22). Initial diagnoses were validated at follow-up visits over a median period of 20 months (range 13–25 months; duration of follow-up >1 year in all cases), as described previously (19,21). All patients gave their written, informed consent for inclusion into the study, which was approved by the local Regional Ethics Committee.

Measurements of gene expression, data curation, and quality control. Whole peripheral blood was stored at room temperature for ≤4 hours before processing. CD4+ lymphocytes were isolated from the peripheral blood by positive selection, as previously described (21), yielding a median cell purity of 98.9%. To obtain B lymphocytes, peripheral blood mononuclear cells were first isolated by density centrifugation using the Lymphoprep protocol (Axis-Shield Diagnostics), and then subjected to positive selection using anti-CD19 magnetic microbeads (Miltenyi Biotec). The median cell purity was 96.4%, as determined by flow cytometry (see Supplementary Figure 1, available on the *Arthritis & Rheumatology* web site at http://onlinelibrary.wiley.com/doi/10.1002/art.40393/abstract).

RNA was immediately extracted from total CD4+ T cells or B lymphocytes using an RNeasy Mini kit (prior to 2012) or AllPrep DNA/RNA Mini kit (both from Qiagen), and then subject to quality control using an Agilent 2100 Bioanalyzer (Agilent). The median RNA integrity number in the samples analyzed was 9.4. Complementary RNA generated from 250 ng total RNA (Illumina TotalPrep RNA Amplification kit) was hybridized to either an Illumina Whole Genome 6 version 3 (using CD4+ lymphocyte samples obtained prior to 2012) or a 12HT BeadChip (using CD4+ T cell samples obtained in or after 2012, and all B cell samples) (both from Illumina). The analysis was limited to probes determined to be common to both array platforms, based on unique capture sequence identifiers. Those liable to cross-hybridization according to probe-sequence BLAT analysis were then excluded.

Following normalization (robust spline normalization) and variance stability transformation (23,24), batch correction of the data from CD4+ cells by linear modeling (25), and merging of the component data sets (26), principal components analysis was carried out to confirm correction for technical bias (see Supplementary Figure 2, available on the *Arthritis & Rheumatology* web site at http://onlinelibrary.wiley.com/doi/10.1002/art.40393/abstract). The raw and processed expression data used in this study are available in the Gene Expression Omnibus database (accession nos. GSE20098, GSE80513, or GSE100648; http://www.ncbi.nlm.nih.gov/geo) (a complete list of unique identifiers is provided in Supplementary Table 1, available on the *Arthritis & Rheumatology* web site at http://onlinelibrary.wiley.com/doi/10.1002/art.40393/abstract).

Genotyping. Genomic DNA was isolated from the peripheral blood of all patients, either from the whole blood using the Wizard genomic DNA purification kit (Promega) (for samples obtained prior to 2012) or from isolated lymphocytes in parallel with RNA extraction (AllPrep DNA/RNA Mini kit; Qiagen). Genotyping was carried out using an Illumina Human CoreExome-24 version 1-0 array, following the manufacturer's protocol. Samples and SNPs with a call rate of <98% were excluded. In addition, SNPs with a minor allele frequency of <0.01 or an Illumina GenomeStudio cluster separation of <0.4 were excluded from further analysis. Data were pre-phased using SHAPEIT2 and imputed to the 1000 Genomes Phase 1, version 3, reference panel using IMPUTE2. Imputed SNPs with INFO scores of <0.8 were excluded.

Analysis of eQTLs and covariates. Analysis of eQTLs was limited to loci defined by the 101 lead disease-associated variants confirmed to be present in Caucasians, as previously described by Okada et al (7). For this analysis, linear models were fitted and residual analysis was performed to verify model assumptions using the R package; Pearson's R^2 statistics and associated *P* values were derived. Due to abundant cross-hybridization of the expression probes and the confounding effect of copy numbers within the HLA region, we limited our analysis to non-HLA variants. Permutation testing (10,000 permutation replicates) was carried out to derive experiment-wide *P* values equivalent to a predetermined α value of 5%; a more relaxed (though nonetheless robust) significance threshold was also defined at an α value of 10%. This method, utilized to correct for multiple testing, proved more stringent than the standard Benjamini-Hochberg method, which was also applied for comparison. A general linear model incorporating other potential biologic and clinical parameters, including age, sex, C-reactive protein (CRP) level, and swollen joint count, permitted evaluation of the robustness of the eQTLs in relation to inflammation markers and other potential covariates.

Comparisons with published data sets. Published eQTL data sets were identified using PubMed literature searches. Results were cross-checked and validated with reference to the GTEx Portal database (available at http://gtexportal.org) (27).

RESULTS

Mapping of eQTLs at RA risk loci in lymphocytes of treatment-naive patients with early arthritis. Expression data from primary peripheral blood lymphocytes were available for a total of 344 genotyped patients with early arthritis; available data on CD4+ lymphocytes were

Table 1. Characteristics of the patients with early arthritis*

	RA (n = 124)	Non-RA inflammatory arthritis (n = 113)	Noninflammatory arthritis (n = 107)	*P*†
Age, years	59 (48–73)	51 (39–63)	52 (44–57)	<0.001
Sex, % female	69	61	81	<0.001
Duration of symptoms, weeks	12 (8–27)	12 (6–25)	24 (8 to >52)	0.03
CRP, gm/liter	11 (5–26)	8 (5–19)	<5 (<5–8)	<0.001
ESR, mm/minute	26 (12–49)	19 (7–34)	8 (4–20)	<0.001
TJC	6 (3–12)	2 (1–6)	3 (0–8)	0.001
SJC	2 (1–6)	1 (0–3)	0 (0–0)	<0.001
DAS28	4.68 (3.5–5.5)	NA	NA	NA
RF positive, %	57	7	14	<0.001
ACPA positive, %	48	2	1	<0.001
Non-RA diagnosis, %				
Osteoarthritis	–	–	55	–
Other, noninflammatory arthritis	–	–	45	–
Spondyloarthropathy (PsA, AS, EA)	–	68	–	–
Crystal arthropathy	–	9	–	–
Other, inflammatory arthritis	–	20	–	–
Undifferentiated arthritis	–	3	–	–

* Patients with early arthritis are stratified by diagnostic category, including non–rheumatoid arthritis (RA) subclassifications. Except where indicated otherwise, values are the median (interquartile range). CRP = C-reactive protein; ESR = erythrocyte sedimentation rate; TJC = tender joint count (of 28 joints); SJC = swollen joint count (of 28 joints); DAS28 = Disease Activity Score in 28 joints; NA = not applicable; RF = rheumatoid factor; ACPA = anti–citrullinated peptide autoantibody; PsA = psoriatic arthritis; AS = ankylosing spondylitis; EA = enteropathic arthritis.

† *P* values were based on Kruskal-Wallis nonparametric analysis of variance for continuous variables, and chi-square test for dichotomous variables.

limited to 249 of the patients, data on B lymphocytes were available for 242 of the patients, and paired data were available for 147 of the patients. The baseline clinical characteristics and diagnoses of all patients are summarized in Table 1. After quality control procedures were applied, a total of 1,227 genotyped variants in LD (defined as $r^2 \geq 0.8$) with lead RA-associated SNPs were considered. Filtered expression probes whose start sites mapped to within 4 Mb of LD blocks (as defined in Patients and Methods) were initially measured to identify *cis*-acting eQTLs. In a secondary analysis of *trans*-eQTLs, those with start sites >4 Mb from the same LD blocks were evaluated in a similar manner.

Permutation testing was carried out using 10,000 permutation replicates for each analysis in each lymphocyte subset. This allowed us to account for multiple testing, in which the total number of tests for each cell type corresponded to the number of unique SNP–gene pairs in the analyses of *cis*- or *trans*-acting eQTLs across the prespecified loci, after data processing and quality control had been performed. The maximum value of the test statistic (minimum nominal *P* value) across the total number of tests in each permutation replicate was recorded, and significance thresholds exceeding 5% or 10% in each permutation replicate were determined. This procedure resulted in experiment-wide *P* value thresholds ($\alpha = 5\%$ or $\alpha = 10\%$) that were used to define evidence of eQTLs in each cell type, as summarized in Figure 1 (for *cis*-eQTL analyses) and in Supplementary Figure 3 (for *trans*-eQTL analyses; available on the *Arthritis & Rheumatology* web site at http://onlinelibrary.wiley.com/doi/10.1002/art.40393/abstract).

In total, 213 *cis*-acting significant SNP–transcript associations were identified in CD4+ lymphocytes ($\alpha = 5\%$), corresponding to 10 unique genes at 7 established RA risk loci; 194 *cis*-eQTLs were similarly identified in B lymphocytes ($\alpha = 5\%$), also corresponding to 10 unique genes at 7 loci. The *cis*-eQTL effects for *FADS1*, *FADS2*, *FCRL3*, *BLK*, *ORMDL3*, *GSDMB*, and *PPIL3* were robust in both CD4+ and B lymphocytes at RA risk loci.

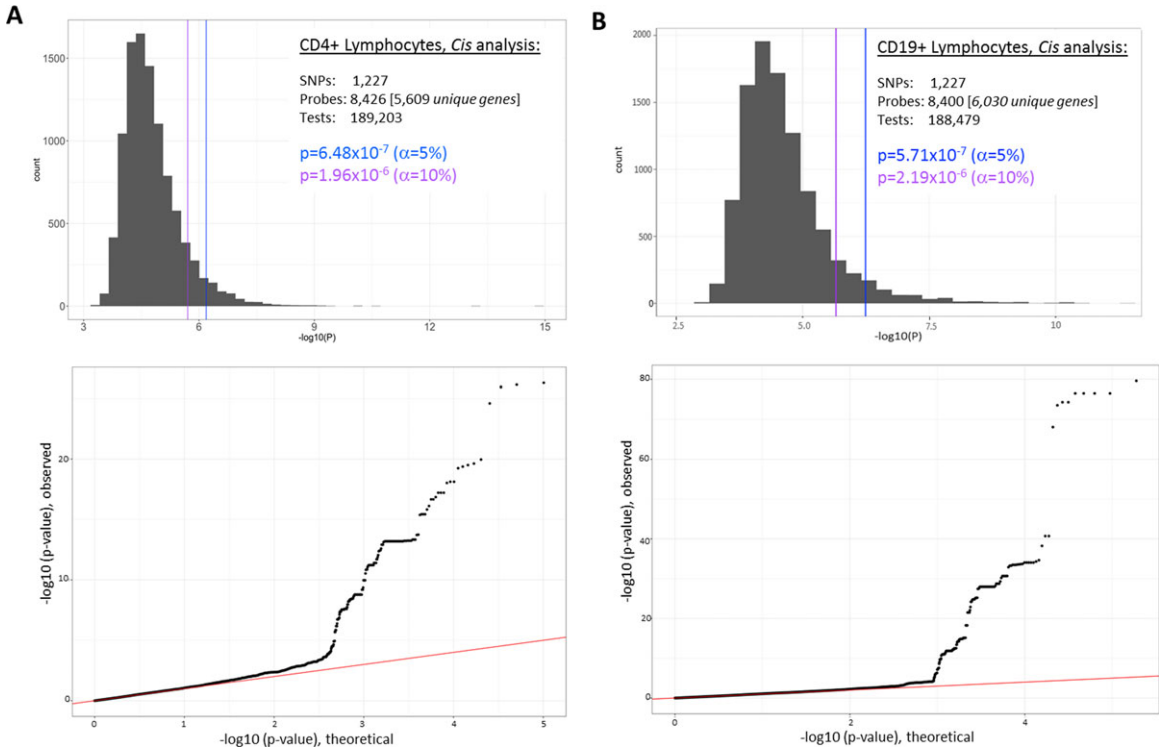


Figure 1. Determination of experiment-wide significance of *cis*-acting expression quantitative trait loci (*cis*-eQTLs) in CD4+ T lymphocytes (**A**) and CD19+ B lymphocytes (**B**). Top, Histograms summarize the data from 10,000 permutation replicates, each derived from the indicated number of single-nucleotide polymorphisms (SNPs) and expression probes, and the final number of included tests. *P* values at the $\alpha = 5\%$ and $\alpha = 10\%$ thresholds are shown. Bottom, QQ plots depict expected *P* value distributions under the null hypothesis (red line) versus observed distributions. Analogous plots for analyses of *trans*-eQTLs are shown in Supplementary Figure 3 (available on the *Arthritis & Rheumatology* web site at http://onlinelibrary.wiley.com/doi/10.1002/art.40393/abstract). Color figure can be viewed in the online issue, which is available at http://onlinelibrary.wiley.com/doi/10.1002/art.40393/abstract.

Table 2. Summary of CD4+ T lymphocyte *cis*-eQTL genes*

Gene	Lead eQTL SNP	Locus	Minor allele (MAF)	*P*†	R^2 in relation to RA index SNP	Total no. of significant SNPs for probe‡
FADS1	rs968567	11q12	T (0.177)	1.06×10^{-27}	1.0	3
BLK	rs922483	8p23	T (0.426)	1.41×10^{-20}	0.805	9
FADS2	rs968567	11q12	T (0.177)	2.40×10^{-20}	1.0	3
METTL21B	rs701006	12q13–q14	A (0.408)	3.65×10^{-19}	1.0	3
FCRL3	rs2210913	1q23	C (0.457)	6.24×10^{-16}	0.873	12
ORMDL3	rs4795397	17q12–q21	G (0.482)	1.07×10^{-14}	0.959	66
PPIL3	rs6757776	2q33	G (0.103)	8.60×10^{-11}	1.0	16
GSDMB	rs4795397	17q12–q21	G (0.314)	6.92×10^{-10}	0.969	88
IKZF3	rs1453559	17q12–q21	C (0.453)	2.52×10^{-9}	0.801	23
JAZF1	rs4722758	7p15	G (0.195)	1.70×10^{-8}	1.0	15
PADI4	rs2240339	1p36	T (0.418)	3.37×10^{-6}§	0.923	–

* Microarray probe targets are shown as Human Genome Organisation gene symbols. Lead expression quantitative trait locus (eQTL) single-nucleotide polymorphisms (SNPs) and loci are also shown, along with their minor allele and minor allele frequency (MAF). Rheumatoid arthritis (RA) index SNPs were those listed in the report by Okada et al (see ref. 7).
† The permuted significance thresholds of $\alpha = 5\%$ and $\alpha = 10\%$ equate to $P = 6.48 \times 10^{-7}$ and $P = 1.96 \times 10^{-6}$, respectively (see Figure 1).
‡ Based on a threshold of $\alpha = 10\%$.
§ Data for the PADI4 eQTL fell marginally short of the $\alpha = 10\%$ threshold.

The eQTLs acting on 3 genes (*METTL21B*, *IKZF3,* and *JAZF1*) were unique to CD4+ T lymphocytes in this population, with *PADI4* also subject to a convincing effect exclusively in this cell type despite falling marginally short of the $\alpha = 10\%$ threshold by permutation analysis; the latter gene encodes a peptidylarginine deiminase enzyme, and therefore is of interest in the pathogenesis of RA (28).

At the 8p23 locus, *FAM167A* was, in contrast to the neighboring *BLK* gene, shown to be subject to *cis* regulation only in B lymphocytes, and *SYNGR1* and *CD83* eQTLs were also specific to this cell type. These data are summarized in Tables 2 and 3 and depicted as Manhattan plots in Figure 2. No *trans*-eQTLs achieved experiment-wide significance thresholds, either in CD4+ T lymphocytes or in B lymphocytes.

Representative examples of eQTL plots are depicted in Figure 3, and a comprehensive list of all SNP–probe associations that remained significant after Benjamini-Hochberg correction for multiple testing is provided in Supplementary Tables 2 and 3 (available on the *Arthritis & Rheumatology* web site at http://onlinelibrary.wiley.com/doi/10.1002/art.40393/abstract), in which significance thresholds of $\alpha = 5\%$ and $\alpha = 10\%$ by

Table 3. Summary of CD19+ B lymphocyte *cis*-eQTL genes*

Gene	Lead eQTL SNP	Locus	Minor allele (MAF)	*P*†	R^2 with RA index SNP	Total no. of significant SNPs for probe‡
FAM167A	rs4840568	8p23	A (0.264)	2.48×10^{-80}	0.817	9
FADS1	rs968567	11q12	T (0.177)	1.96×10^{-41}	1.0	3
ORMDL3	rs9906951	17q12–q21	C (0.383)	2.29×10^{-35}	0.881	66
FADS2	rs968567	11q12	T (0.177)	6.60×10^{-25}	1.0	3
FCRL3	rs2210913	1q23	T (0.482)	3.09×10^{-22}	0.839	12
GSDMB	rs12936231	17q12–q21	G (0.434)	7.21×10^{-16}	0.862	66
SYNGR1	rs909685	22q13	A (0.31)	3.66×10^{-14}	1.0	5
BLK	rs2618476	8p23	C (0.25)	3.02×10^{-13}	0.958	9
PPIL3	rs2141331	2q33	T (0.097)	1.13×10^{-10}	0.943	8
CD83	rs78242827	6p23	C (0.058)	2.72×10^{-8}	1.0	20

* Microarray probe targets are shown as Human Genome Organisation gene symbols. Lead expression quantitative trait locus (eQTL) single-nucleotide polymorphisms (SNPs) and loci are also shown, along with their minor allele and minor allele frequency (MAF). Rheumatoid arthritis (RA) index SNPs were those listed in the report by Okada et al (see ref. 7).
† The permuted significance thresholds of $\alpha = 5\%$ and $\alpha = 10\%$ equate to $P = 5.71 \times 10^{-7}$ and 2.19×10^{-6}, respectively (see Figure 1).
‡ Based on a threshold of $\alpha = 10\%$.

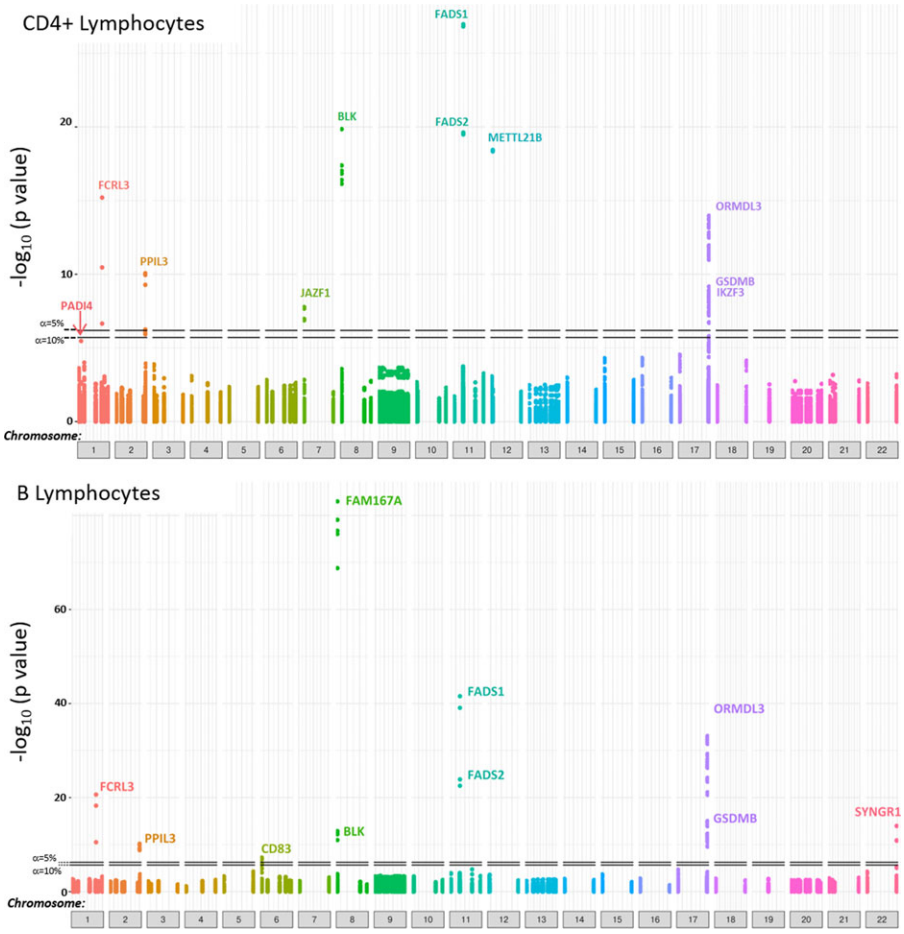


Figure 2. Manhattan plots depict the 101 rheumatoid arthritis risk loci analyzed and *P* values for the significance of single-nucleotide polymorphism (SNP)–probe pairs (denoted by different-colored dots) among CD4+ T lymphocytes (top) and B lymphocytes (bottom) in patients with early arthritis. Human Genome Organisation gene symbols for SNP–probe pairs, or groups thereof, that approached or reached experiment-wide significance (at thresholds of $\alpha = 5\%$ and $\alpha = 10\%$ [horizontal lines]) are indicated, permitting comparison of expression quantitative trait loci between cell types.

permutation testing are also indicated. Supplementary Table 4 (http://onlinelibrary.wiley.com/doi/10.1002/art.40393/abstract) summarizes this information, listing all significant eQTL SNPs (and associated genes) in relation to the index SNPs reported by Okada et al (7). Limiting any or all of the above analyses to samples

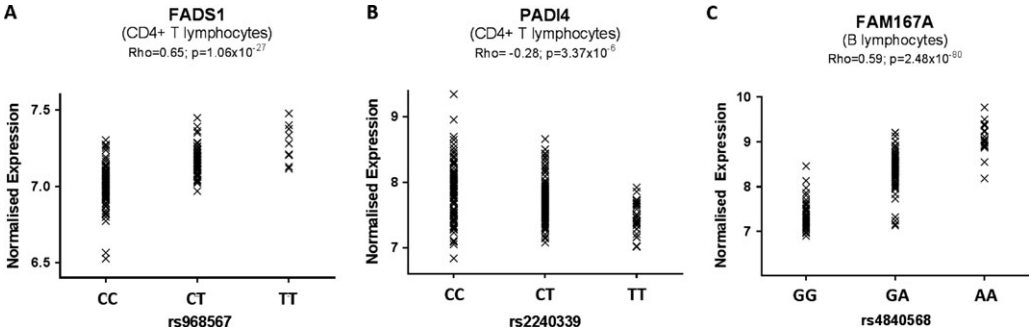


Figure 3. Representative examples of expression quantitative trait loci (eQTLs). Plots of normalized individual gene expression, along with their Spearman's rho statistics and *P* values for association, are shown for the lead eQTL single-nucleotide polymorphisms acting on *FADS1* **(A)** and *PADI4* **(B)** in CD4+ T lymphocytes and *FAM167A* **(C)** in B lymphocytes.

for which paired CD4+ and B lymphocytes were available (n = 147) had no substantial effect on the eQTL genes identified, although some associations ceased to reach experiment-wide significance due to diminished power (see Supplementary Tables 5 and 6, http://onlinelibrary.wiley.com/doi/10.1002/art.40393/abstract).

Comparison of eQTLs with published data sets. Our findings were considered in light of a number of human eQTL studies for which significant SNP–probe combinations are in the public domain. These included analyses of *cis*-eQTLs in primary CD4+ and B lymphocytes. Murphy et al studied genome-wide expression in positively selected whole CD4+ lymphocytes from 200 non-Hispanic white subjects, comprising young adults with asthma and their first-degree relatives (29). Hu et al examined paired expression limited to 270 genes in resting CD4+ lymphocytes and CD3/CD28-stimulated effector memory CD4+ lymphocytes from healthy donors; genes were selected based on their proximity to 157 SNPs with known autoimmune disease associations (including with RA) (30). Raj et al reported genome-wide eQTL data in positively selected CD45RO− (naive) CD4+ lymphocytes from 200 healthy European Americans (31), and a similar analysis, by Kasela et al, was conducted in whole CD4+ (and CD8+) T cells (32). Another study, by Fairfax et al (14), demonstrated the presence of eQTLs in primary B cells from 288 healthy Europeans. Studies by Dixon and colleagues (33,34) presented cumulative data from Epstein-Barr virus–transformed human B cells (lymphoblastoid cell lines), and a large meta-analysis was conducted to compare studies performed in the whole blood of predominantly healthy volunteers (35). Finally, our data were considered in the context of the GTEx resource database (27).

Overlap between the genes subject to *cis*-eQTLs in these studies compared with those identified in our own study is illustrated in Supplementary Figure 4 (available on the *Arthritis & Rheumatology* web site at http://onlinelibrary.wiley.com/doi/10.1002/art.40393/abstract). Reassuringly, all of the *cis*-eQTL genes identified in patients with untreated early arthritis replicated the findings reported in at least one of the comparator studies. Strong independent validation of CD4+ lymphocyte–specific associations was provided with regard to 9 of the genes. Among these, RA risk loci at 11q12 and at 17q12–21 were each observed to harbor pairs of apparently coregulated genes, *FADS1*/*FADS2* and *ORMDL3*/*GSDMB*, respectively. Moreover, at the 17q12–21 locus, *IKZF3* was confirmed to be subject to a highly significant eQTL effect in CD4+ lymphocytes (32). When a more lenient (but nonetheless robust) method for multiple test correction was employed, we highlighted, for the first time, an association between *PADI4* expression and genotype at the 1p36 locus specifically in CD4+ T cells, its having previously been identified only in whole blood. Our findings with respect to *FAM167A*, *SYNGR1*, and *CD83* corroborate those in the only other study of primary human B lymphocytes, by Fairfax et al (14), and although the same eQTLs have been noted in mixed cell populations of whole blood (35), no study (including our own) has yet replicated them in CD4+ T lymphocytes.

Lack of significant impact of clinical covariates on eQTLs. Because differential eQTL effect sizes have been observed in paired CD4+ T cells from healthy donors according to whether T cell receptor–mediated stimulation of the cells was undertaken ex vivo prior to RNA extraction (30), we hypothesized that certain clinical covariates, and/or the activation status of circulating CD4+ T cells, might have a similar influence in vivo. The clinical parameters considered included age, sex, CRP level, and erythrocyte sedimentation rate (as indicators of the systemic acute-phase response), as well as disease phenotype (RA versus non-RA). In the patient subgroup for whom CD4+ lymphocyte expression data were available, normalized transcript levels of CD25, CD69, and interferon-γ, as measured by microarray, were also considered as surrogates of the CD4+ T cell activation status. The incorporation of each of these covariates, in turn, into linear models made no difference to the final eQTL list (as shown in Tables 2 and 3), and individual regression slopes were robust to their inclusion (representative examples are depicted in Supplementary Figure 5, available on the *Arthritis & Rheumatology* web site at http://onlinelibrary.wiley.com/doi/10.1002/art.40393/abstract). Consistent with these findings, lists of genes subject to *cis*-eQTL effects did not vary substantially when patients with RA and those with alternative diagnoses were considered independently (results not shown). Thus, eQTLs were robust to clinical and biologic covariates in our study, and no evidence of early disease–specific eQTLs at RA risk loci was found.

DISCUSSION

We present the first eQTL analysis of primary lymphocytes from donors presenting with untreated, suspected inflammatory arthritis—a context highly relevant for the purpose of unravelling genetic risk mechanisms in RA. Several important observations can be made on the basis of our findings.

CD4+ and B lymphocytes in this setting exhibit distinct but overlapping eQTLs at confirmed RA risk loci (Tables 2 and 3). The specificity of an eQTL effect for

one cell type may simply be a reflection of the lack of expression of a gene by a comparator cell, but probe-level microarray data suggest that reported genes were expressed in both CD4+ and B lymphocytes in our study. Therefore, the cell-specific effects that we observed for *METTL21B*, *IKZF3*, *JAZF1*, and *PADI4* (in CD4+ lymphocytes) and *FAM167A*, *SYNGR1*, and *CD83* (in B lymphocytes) may indicate differential regulatory functions of disease risk variants between lineages.

Strikingly, at the common *BLK–FAM167A* autoimmune locus at 8p23, we found that 2 adjacent genes were subject to eQTLs whose activity differed between cell types: the *FAM167A* effect displayed robust B lymphocyte specificity and was absent in CD4+ lymphocytes, whereas the *BLK* effect that was prominent in CD4+ T lymphocytes was less prominent among B lymphocytes (compare Table 2 and Table 3, and see Figure 2). The most strongly associated SNPs differed between cell types at this locus—a finding that was maintained among patients for whom paired cell–specific data were available (as shown in Supplementary Tables 5 and 6, http://onlinelibrary.wiley.com/doi/10.1002/art.40393/abstract), potentially signifying the presence of mechanistically distinct regulatory variants in strong LD. Nonetheless, the results of our study also contribute to an emerging picture in which eQTLs can regulate the expression of more than one gene at disease-associated loci, examples being found at 11q12 and 17q12–21. This is consistent with the concept that key genetic variants may act as "master regulators" of gene expression.

Our findings provide an important platform from which downstream functional studies may be directed toward particular cell types. For example, elucidating the relevance of the *METTL21B* gene product in CD4+ T cell function would now seem a priority, given our findings confirming a pronounced eQTL effect on this gene in this cell type. Alternative causal candidate genes known, to date, to be favored at the 12q13–q14 locus are *CDK4* and *CYP27B1*, based on their respective functions in cell-cycle progression and vitamin D metabolism (7); however, since neither of these genes were shown to be subject to prominent eQTL effects, despite the growing body of literature discussing their functions, it seems justified to consider *METTL21B* as an alternative candidate gene in CD4+ T cells.

A similar case for both *CD83* and *SYNGR1* in B lymphocytes might also be made. *CD83* encodes a transmembrane member of the immunoglobulin superfamily expressed widely on dendritic cells, but also on activated lymphocytes; its important role in regulating B lymphocyte development and effector function is only now beginning to be understood (36). *SYNGR1* is an integral membrane protein associated with presynaptic vesicles in neuronal cells, and its function in lymphoid cells remains obscure. However, caution should be exercised when interpreting transcript eQTLs in isolation (37), and validation of our findings at the protein level should be prioritized. This was amply illustrated by Simpfendorfer et al, who, similar to our findings at the 8p23 locus, highlighted *BLK* transcript expression as subject to an eQTL in lymphocytes; however, the CD4+ T cell–specific effect was not sustained at the protein level in these cells. By measuring allelic expression imbalance, those authors went on to demonstrate a robust eQTL for both RNA and protein expression in naive/transitional B cell subsets isolated from umbilical cord blood, which was less evident in whole B cells, suggesting that disease risk is conferred during early B cell development rather than by CD4+ T cells (38), potentially via dysregulated B cell receptor signaling (39).

Our study is the first to provide evidence of an eQTL SNP in CD4+ lymphocytes that was in perfect LD with the RA-associated variant at the 1p36 locus, a variant that regulates *PADI4* gene expression. The *PADI4* gene has already been recognized as a strong causal candidate for the disease, encoding peptidylarginine deiminase 4, a key enzyme involved in post-translational citrullination of arginine residues that yields neoepitopes against which RA-specific anti–citrullinated peptide autoantibodies may be raised (28). However, distinct mechanisms of CD4+ lymphocyte dysregulation now warrant further investigation (40).

Similarly, the finding that *IKZF3* is subject to an eQTL in CD4+ lymphocytes is, to our knowledge, a novel observation and is intriguing, given the proven role of the transcription factor product of this gene in regulating interleukin-10 production by these cells (41).

Conceiveably, our observations with regard to *PADI4* and *IKZF3* could be interpreted as evidence that putatively common causal SNPs augment gene expression in CD4+ T cells uniquely under the particular biologic and/or environmental circumstances of early arthritis. However, our analysis of interactions between specific biologic covariates and eQTL effects did not support such an interpretation: in particular, the *IKZF* rs9916765 eQTL slope gradient was unaffected by markers of systemic inflammation, T cell activation, or clinical diagnosis (see Supplementary Figures 5D–F, http://onlinelibrary.wiley.com/doi/10.1002/art.40393/abstract). While this could be seen as surprising, given the previously reported differences between CD4+ T cell eQTLs according to activation status in vitro (30), the contrastingly cross-sectional nature of our study, which focused on unstimulated ex vivo cells from systemically

inflamed and uninflamed peripheral blood samples, render the findings complimentary rather than contradictory, in our view. Indeed, the fact that eQTL effects did not differ according to disease classification (e.g., RA versus non-RA) in our early arthritis population recalls the findings in a study by Peters et al, whereby inflammatory bowel disease–specific eQTLs resided outside of known risk loci for that condition (42). Further work is therefore needed to elucidate the mechanisms by which eQTL effects may wax or wane at a cellular level within the in vivo environment.

Our data extend the understanding of the causal candidate gene landscape in early RA, highlighting several such candidates that now deserve further investigation in defined primary lymphocyte populations. In the future, the possibility that eQTL effects may exhibit heterogeneity between subsets of CD4+ T and/or B lymphocytes should be considered, since these populations are well-known to comprise functionally diverse compartments. Moreover, it is likely that larger integrative studies, including meta-analyses of accumulating lymphocyte eQTL data sets in relevant populations, will be required to expand on this. Such work will have additional value in the identification of *trans*-eQTL effects, which, because of power considerations, we were not able to address in the present study.

ACKNOWLEDGMENTS

We are grateful to Mr. Ben Hargreaves for providing administrative support. We also acknowledge the assistance given by IT Services and the use of the Computational Shared Facility at the University of Manchester.

AUTHOR CONTRIBUTIONS

All authors were involved in drafting the article or revising it critically for important intellectual content, and all authors approved the final version to be published. Dr. Pratt had full access to all of the data in the study and takes responsibility for the integrity of the data and the accuracy of the data analysis.

Study conception and design. Thalayasingam, Isaacs, Barton, Pratt.

Acquisition of data. Thalayasingam, Nair, Massey, Anderson, Diboll, Pratt.

Analysis and interpretation of data. Thalayasingam, Nair, Skelton, Massey, Clark, Lendrem, Reynard, Cordell, Eyre, Pratt.

REFERENCES

1. McInnes IB, O'Dell JR. State-of-the-art: rheumatoid arthritis. Ann Rheum Dis 2010;69:1898–906.
2. Raychaudhuri S, Sandor C, Stahl EA, Freudenberg J, Lee HS, Jia X, et al. Five amino acids in three HLA proteins explain most of the association between MHC and seropositive rheumatoid arthritis. Nat Genet 2012;44:291–6.
3. Gregersen PK, Silver J, Winchester RJ. The shared epitope hypothesis: an approach to understanding the molecular genetics of susceptibility to rheumatoid arthritis. Arthritis Rheum 1987;30:1205–13.
4. Hill JA, Bell DA, Brintnell W, Yue D, Wehrli B, Jevnikar AM, et al. Arthritis induced by posttranslationally modified (citrullinated) fibrinogen in DR4-IE transgenic mice. J Exp Med 2008; 205:967–79.
5. 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. Nat Genet 2013;45:124–30.
6. Farh KK, 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.
7. 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;506:376–81.
8. Roadmap Epigenomics Consortium, Kundaje A, Meuleman W, Ernst J, Bilenky M, Yen A, et al. Integrative analysis of 111 reference human epigenomes. Nature 2015;518:317–30.
9. Diogo D, Okada Y, Plenge RM. Genome-wide association studies to advance our understanding of critical cell types and pathways in rheumatoid arthritis: recent findings and challenges. Curr Opin Rheumatol 2014;26:85–92.
10. Lundy SK, Sarkar S, Tesmer LA, Fox DA. Cells of the synovium in rheumatoid arthritis. T lymphocytes. Arthritis Res Ther 2007;9:202.
11. Moura RA, Graca L, Fonseca JE. To B or not to B the conductor of rheumatoid arthritis orchestra. Clin Rev Allergy Immunol 2012;43:281–91.
12. Pratt AG, Isaacs JD. Genotyping in rheumatoid arthritis: a game changer in clinical management? Expert Rev Clin Immunol 2015;11:303–5.
13. Westra HJ, Franke L. From genome to function by studying eQTLs. Biochim Biophys Acta 2014;1842:1896–192.
14. 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. Nat Genet 2012;44:502–10.
15. 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. Nat Genet 2017;49:1120–25.
16. 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.
17. Fu J, Wolfs MG, Deelen P, Westra HJ, Fehrmann RS, Te Meerman GJ, et al. Unraveling the regulatory mechanisms underlying tissue-dependent genetic variation of gene expression. PLoS Genet 2012;8:e1002431.
18. 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 Biol 2016;17:79.
19. Anderson AE, Pratt AG, Sedhom MA, 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. Ann Rheum Dis 2016; 75:466–73.
20. Pratt AG, Lorenzi AR, Wilson G, Platt PN, Isaacs JD. Predicting persistent inflammatory arthritis amongst early arthritis clinic patients in the UK: is musculoskeletal ultrasound required? Arthritis Res Ther 2013;15:R118.
21. Pratt AG, Swan DC, Richardson S, Wilson G, Hilkens CM, Young DA, et al. A CD4 T cell gene signature for early rheumatoid arthritis implicates interleukin 6-mediated STAT3 signalling, particularly in anti-citrullinated peptide antibody-negative disease. Ann Rheum Dis 2012;71:1374–81.
22. Pratt AG, Lendrem D, Hargreaves B, Aslam O, Galloway JB, Isaacs JD. Components of treatment delay in rheumatoid arthritis differ according to autoantibody status: validation of a single-

centre observation using national audit data. Rheumatology (Oxford) 2016;55:1843–8.
23. Du P, Kibbe WA, Lin SM. lumi: a pipeline for processing Illumina microarray. Bioinformatics 2008;24:1547–8.
24. Lin SM, Du P, Huber W, Kibbe WA. Model-based variance-stabilizing transformation for Illumina microarray data. Nucleic Acids Res 2008;36:e11.
25. Ritchie ME, Phipson B, Wu D, Hu Y, Law CW, Shi W, et al. limma powers differential expression analyses for RNA-sequencing and microarray studies. Nucleic Acids Res 2015;43:e47.
26. Taminau J, Meganck S, Lazar C, Steenhoff D, Coletta A, Molter C, et al. Unlocking the potential of publicly available microarray data using inSilicoDb and inSilicoMerging R/Bioconductor packages. BMC Bioinformatics 2012;13:335.
27. GTEx Consortium. The Genotype-Tissue Expression (GTEx) project. Nat Genet 2013;45:580–5.
28. Suzuki A, Yamada R, Chang X, Tokuhiro S, Sawada T, Suzuki M, et al. Functional haplotypes of PADI4, encoding citrullinating enzyme peptidylarginine deiminase 4, are associated with rheumatoid arthritis. Nat Genet 2003;34:395–402.
29. Murphy A, Chu JH, Xu M, Carey VJ, Lazarus R, Liu A, et al. Mapping of numerous disease-associated expression polymorphisms in primary peripheral blood CD4+ lymphocytes. Hum Mol Genet 2010;19:4745–57.
30. Hu X, 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 Genet 2014;10:e1004404.
31. Raj T, Rothamel K, Mostafavi S, Ye C, Lee MN, Replogle JM, et al. Polarization of the effects of autoimmune and neurodegenerative risk alleles in leukocytes. Science 2014;344:519–23.
32. 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 Genet 2017;13:e1006643.
33. Liang L, Morar N, Dixon AL, Lathrop GM, Abecasis GR, Moffatt MF, et al. A cross-platform analysis of 14,177 expression quantitative trait loci derived from lymphoblastoid cell lines. Genome Res 2013;23:716–26.
34. Dixon AL, Liang L, Moffatt MF, Chen W, Heath S, Wong KC, et al. A genome-wide association study of global gene expression. Nat Genet 2007;39:1202–7.
35. Westra HJ, Peters MJ, Esko T, Yaghootkar H, Schurmann C, Kettunen J, et al. Systematic identification of trans eQTLs as putative drivers of known disease associations. Nat Genet 2013;45:1238–43.
36. Breloer M, Fleischer B. CD83 regulates lymphocyte maturation, activation and homeostasis. Trends Immunol 2008;29:186–94.
37. 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. Nat Genet 2017;49:600–5.
38. Simpfendorfer KR, Armstead BE, Shih A, Li W, Curran M, Manjarrez-Orduño N, et al. Autoimmune disease–associated haplotypes of BLK exhibit lowered thresholds for B cell activation and expansion of Ig class-switched B cells. Arthritis Rheumatol 2015;67:2866–76.
39. Simpfendorfer KR, Olsson LM, Manjarrez Orduno N, Khalili H, Simeone AM, Katz MS, et al. The autoimmunity-associated BLK haplotype exhibits cis-regulatory effects on mRNA and protein expression that are prominently observed in B cells early in development. Hum Mol Genet 2012;21:3918–25.
40. Seri Y, Shoda H, Suzuki A, Matsumoto I, Sumida T, Fujio K, et al. Peptidylarginine deiminase type 4 deficiency reduced arthritis severity in a glucose-6-phosphate isomerase-induced arthritis model. Sci Rep 2015;5:13041.
41. Evans HG, Roostalu U, Walter GJ, Gullick NJ, Frederiksen KS, Roberts CA, et al. TNF-α blockade induces IL-10 expression in human CD4+ T cells. Nat Commun 2014;5:3199.
42. 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 Genet 2016;12:e1005908.